WORLD METEOROLOGICAL ORGANIZATION

TECHNICAL NOTE N° 143

ON THE STATISTICAL ANALYSIS OF SERIES OF OBSERVATIONS

by R. SNEYERS

WMO - N° 415

Secretariat of the World Meteorological Organization - Geneva - Switzerland

WMO

The World Meteorological Organization (WMO), of which 161 States and Territories are Members, is a specialized agency of the United Nations.

The purposes of WMO are :

- To facilitate world-wide co-operation in the establishment of networks of stations for making meteorological observations as well as hydrological and other physical observations related to meteorology, and to promote the establishment and maintenance of centres charged with the provision of meteorological and related services;
- To promote the establishment and maintenance of systems for the rapid exchange of meteorological information;
- To promote standardization of meteorological and related observations and to ensure the uniform publication of observations and statistics;
- To further the application of meteorology to aviation, shipping, water problems, agriculture and other human activities;
- To promote activities in operational hydrology and to further close co-operation between Meteorological and Hydrological Services;
- To encourage research and training in meteorology and, as appropriate, in related fields, and to assist in co-ordinating the international aspects of such research and training.

The machinery of the Organization consists of the following bodies:

The *World Meteorological Congress,* the supreme body of the Organization, brings together the delegates of all Members once every four years to determine general policies for the fulfilment of the purposes of the Organization, to adopt Technical Regulations relating to international meteorological practice and to determine the WMO programme.

The *Executive Council* is composed of 36 directors of national Meteorological or Hydrometeorological Services. It meets at least once a year to conduct the activities of the Organization, to implement the decisions taken by its Members in Congress and to study and make recommendations on any matter affecting international meteorology and related activities of the Organization.

The six *regional associations* (Africa, Asia, South America, North and Central America, South-West Pacific and Europe), which are composed of Member Governments, co-ordinate meteorological and related activities within their respective Regions and examine from the regional point of view all questions referred to them.

The eight *technical commissions,* consisting of experts designated by Members, are responsible for studying any subject within the purpose of the Organization. Technical commissions have been established for basic systems, instruments and methods of observation, atmospheric sciences, aeronautical meteorology, agricultural meteorology, marine meteorology, hydrology, and climatology.

The *Secretariat,* located at 41 Avenue Giuseppe-Motta, Geneva, Switzerland, is composed of a Secretary-General and such technical and clerical staff as are required for the work of the Organization. It undertakes to serve as the administrative, documentation and information centre of the Organization, to make technical studies as directed, to support all the bodies of the Organization, to prepare, edit or arrange for the publication and distribution of the approved publications of the Organization, and to carry out duties allocated in the Convention and the regulations and such other work as Congress, the Executive Council and the President may decide. The Secretariat works in close collaboration with the United Nations and its specialized agencies.

WORLD METEOROLOGICAL ORGANIZATION

TECHNICAL NOTE N° 143

ON THE STATISTICAL ANALYSIS OF SERIES OF OBSERVATIONS

by R. SNEYERS

U. D. C. 551.501.45

WMO - N° 415

Secretariat of the World Meteorological Organization - Geneva - Switzerland

1990

ISBN 92 - 63 - 10415-8

NOTE

This publication has been produced without editorial revisions by the WMO Secretariat. The designations employed and the presentation of material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the World Meteorological Organization concerning the legal status of any country, territory, city or area, or of its authorities, or concerning the delimitation of its frontiers or boundaries.

FOREWORD TO THE ENGLISH EDITION

Subsequent to the initial publication of WMO Technical Note No. 143 in French (with the title "Sur l'analyse statistique des séries d'observations"), the Commission for Climatology proposed that an English version also be published to make the Technical Note accessible to a wider audience in view of its particularly high value to climatologists.

Following the approval by the Executive Council of this proposal, the original French text has been translated into English through the courtesy of the Meteorological Office of the United Kingdom, and published by WMO in the present volume.

In re-iterating the expression of thanks of the World Meteorological Organization to Dr. R. Sneyers for the preparation of the original text, and additionally for his collaboration with WMO in publishing this volume, I should also like to acknowledge the valuable assistance of the UK Meteorological Office and in particular of Mr. G.H. Ross, who carried out the technical checking of the English text, and the Institut Royal Météorologique de Belgique for arranging for the production of this publication.

G.O.P. Obasi

Secretary- General

PREFACE TO THE ORIGINAL FRENCH EDITION

(published in 1975)

The Commission for Climatology (now the Commission for Special Applications of Meteorology and Climatology) set up at its fifth session a working group for considering statistical methods and the use of mathematical models in climatology. This group was charged with the task of writing a report, for publication by the WMO, in view of the increasing necessity of having a modern statistical manual available where the elementary methods would be explained simply and clearly and be illustrated by examples of their application to various types of practical problems which must be solved by the climatologist.

During its sixth session, the Commission for Special Applications of Meteorology and Climatology expressed great satisfaction with the excellent text on the statistical analysis of a series of observations presented by the working group and prepared by its chairman, Mr. R. Sneyers. The Commission considered that this work would fill a great gap in meteorology, and its members unanimously supported the proposal to publish this text in the series of WMO Technical Notes.

I should like to take this opportunity of thanking Mr. R. Sneyers on behalf of the World Meteorological Organization for this excellent report, which will be particularly useful to all those who need to use statistical methods in the applications of meteorology and climatology to solve problems in the various fields of human activities.

D.A. Davies

Secretary-General

SUMMARY

It is no longer disputed nowadays that mathematical statistics is the main tool to be used in climatology. Mathematical statistics is in fact the science of random models and the purpose of the statistical analysis of series of observations is to find, from among the models which this science makes available, the one which best represents the behaviour of the observed phenomenon.

This attitude may be justified in two ways.

From the theoretical point of view, the representation of the observed phenomenon by a model depending on a minimum number of parameters enables objective statistical conclusions to be compared with the physical theories put forward, either to confirm their validity or to reveal weaknesses; thus statistical analysis appears as a means for progress in the field of research.

From the practical point of view, due to the efficiency of the methods of statistical analysis, their utilization ensures that the best use is made of the information accumulated by the series of observations and that the maximum benefit will be derived from the capital represented by this information. If we think of the investments and operational costs which have been approved for the acquisition of the series of observations and of the economic value of these series, which is continually demonstrated in the applications of meteorology and climatology, the necessity of such a position becomes obvious.

The adjustment of a random model to a series of observations is in fact, identical with the general problem of statistical estimation and in this Technical Note this problem has been considered for the case of simple random series and for the case of persistent series.

Going into details, after examining, in the first chapter, the most appropriate tests for verifying that the simple random model fits to a given time series, the problem of statistical estimation is broached in the second chapter. This is done successively for the non-parametric case and for the parametric case, both for a continuous variate and for a discrete variate. In addition a paragraph is also devoted to estimation in the case of correlated series.

Recurrent, periodic or persistent models are then introduced in the third chapter, as an alternative solution when the simple random model is not appropriate, and various applications are considered.

The description of the methods follows the principle that a full analysis necessarily comprises the three following stages:

(1) Choice of the model;

(2) Adjustment and goodness of fit of the selected model;

(3) Calculation of the errors of estimation associated with the adjusted model;

and if, in processing the data, these three stages are not always apparent, either because the choice has been incorporated in the method of adjustment or because it has been based on a test of goodness of fit, they can always be traced in the various steps of the solution.

SUMMARY

In order to show their application as clearly as possible, the methods discussed are illustrated by means of actual examples from climatology. However, in order to avoid the work being reduced to a book of rules, it is based on a minimum of theory. This enables the reader to grasp the essentials of statistical methodology and to avoid errors which he would certainly make if the techniques were applied in an automatic fashion.

The scope of the work undoubtedly remains very limited and the omission of certain aspects may be regretted. In this connexion, we think it appropriate to remind the reader that various aspects of the statistical analysis of series of observations have already been dealt with in the Technical Notes Nos. 71, 79, 81, 98 and 100.

Thus, by considering it simply as a useful addition to preceding works, we will put this Technical Note in its rightful place among these works.

RÉSUMÉ

On ne conteste plus, à l'heure actuelle, que la statistique mathématique est la mathématique de la climatologie. La statistique mathématique est, en effet, la science des modèles aléatoires et le but de l'analyse statistique des séries d'observations est de rechercher parmi les modèles offerts par cette science celui qui représente le mieux le comportement du phénomène observé.

Cette attitude se fonde d'ailleurs sur une double justification.

Du point de vue théorique, la représentation du phénomène observé par un modèle dépendant d'un nombre minimum de paramètres rend possible la confrontation de conclusions statistiques objectives avec les théories physiques avancées, soit pour trouver la confirmation de leur validité, soit pour en déceler les faiblesses; l'analyse statistique apparaît ainsi comme une source de progrès dans le domaine de la recherche.

Du point de vue pratique, l'efficacité des méthodes de l'analyse statistique assure, par leur emploi, une valorisation optimale de l'information accumulée par les séries d'observations et garantit, de ce fait, une rentabilité maximale du capital que cette information constitue. Cette rentabilité apparaît d'ailleurs comme une nécessité évidente dès que l'on pense aux investissements et aux frais de fonctionnement qui ont été consentis pour l'acquisition des séries d'observations ainsi qu'à la valeur économique de ces séries sans cesse démontrée dans les applications de la météorologie et de la climatologie.

L'ajustement d'un modèle aléatoire à une série d'observations se confond en réalité avec le problème général de l'estimation statistique et dans la présente Note Technique ce problème a été considéré dans le cas de séries aléatoires simples et dans celui de séries persistantes.

Plus précisément, après avoir examiné au premier chapitre les tests les plus appropriés pour vérifier la conformité d'une série chronologique au modèle aléatoire simple, le problème de l'estimation statistique a été abordé au second chapitre successivement dans le cas non paramétrique et dans le cas paramétrique et ce tant pour une variable continue que pour une variable discrète; de plus, un paragraphe y a également été consacré à l'estimation dans le cas de séries en corrélation.

Les modèles récurrents, périodiques ou persistants ont ensuite été présentés au troisième chapitre comme solution de rechange lorsque le modèle aléatoire simple ne convient pas et diverses applications y ont été considérées.

L'exposé des méthodes a été conduit selon le principe qu'une analyse complète comporte nécessairement les trois étapes suivantes:

1. choix du modèle;
2. ajustement et adéquation du modèle choisi;
3. calcul des erreurs d'estimation associées au modèle ajusté;

et si dans le traitement des données ces trois étapes n'apparaissent pas toujours de façon distincte, soit que le choix ait été intégré dans la méthode d'ajustement ou qu'il se fonde sur un test d'adéquation, on les retrouvera cependant chaque fois dans la démarche des solutions.

Par ailleurs, toutes les méthodes présentées ont été illustrées au moyen d'exemples courants de la climatologie afin de rendre leur application le plus claire possible. Toutefois, pour éviter que l'ouvrage ne se réduise à un livre de recettes, on s'est appuyé sur un minimum d'éléments de théorie afin de permettre au lecteur de saisir l'essence de la méthodologie statistique et d'échapper aux erreurs qu'une application trop automatique des techniques ne manquerait pas de lui faire commettre.

Sans doute la portée de l'ouvrage reste-t-elle très limitée et l'absence de certaines matières sera-t-elle regrettée. Il nous paraît toutefois opportun de rappeler à ce sujet que divers aspects de l'analyse statistique des séries d'observations ont déjà été traités dans les Notes Techniques n^{os} 71, 79, 81, 98 et 100.

Aussi, en considérant simplement cet ouvrage comme un complément utile des précédents lui conférera-t-on sa plus juste place.

РЕЗЮМЕ

В настоящее время уже больше не обсуждается вопрос о том, что математическая статистика является отраслью математики, которая должна использоваться в климатологии. Математическая статистика является фактически наукой моделей случайных переменных и цель статистического анализа ряда наблюдений заключается в том, чтобы среди моделей, которые представляет эта наука найти такую, которая наилучшим образом представляет поведение наблюдаемого явления.

Такой подход можно объяснить двумя путями.

С теоретической точки зрения представление наблюдаемого явления моделью зависящей от минимального количества параметров дает возможность оравнивать объективные статистические выводы, с выдвинутыми физическими теориями или подтвердить их обоснованность или вскрыть их несостоятельность; таким образом статистический анализ является средством для развития в сбласти научных исследований.

С практической точки зрения использование методов статистического анализа, благодаря их эффективности, обеспечивает наилучшее использование информации, накопленной путем ряда наблюдений и получение максимальной выгоды из материала, представляемого этой информацией. Если мы будем думать о вкладах и оперативных расходах, которые утверждены для получения рядов наблюдений и об экономической ценности этих наблюдений, которая постоянно демонстрируется при применении метеорологии и климотологии, становится очевидным, что важно чтобы эти статистические данные приносили такую выгоду.

РЕЗЮМЕ

Проблема подгонки модели для случайных переменных к ряду наблюдений фактически идентична общей проблеме статистической оценки и в настоящей Технической записке эта проблема рассмотрена для случая простого случайного ряда и для случая возвратных рядов.

Переходя,после изучения в первой главе,к более подробным наиболее подходящим тестам проверки относительно того, что ряд согласуется с простой моделью случайной переменной,проблема статистической оценки рассматривается во второй главе. Это делается последовательно для непараметрического случая и для параметрического случая, как для постоянной переменной так и для дискретной переменной. Кроме того один параграф посвящен также оценке случая корреляционного ряда.

Возвратные,периодические или постоянные модели представлены в третьей главе в качестве альтернативного решения в том случае, когда простая модель для случайной переменной не подходит; и рассмотрены также различные применения.

Описание методов дано при соблюдении такого принципа, что полный анализ включает обязательно три следующие стадии:

1) Выбор модели;
2) Подгонка и соответствующая адаптация выбранной модели;
3) Расчет ошибок в оценке, связанный с выбранной моделью;

и если при обработке данных этих три стадии не всегда очевидны или из-за выбора, включенного в метод подгонки или из-за того, что подгонка основана на тесте для подходящей адаптации, все это можно всегда проследить в различных стадиях решения.

Чтобы показать их применение настолько ясно на - сколько это возможно, обсуждаемые методы иллюстрированы путем приведения фактических примеров из климотологии. Однако, чтобы избежать проведение работы путем сведения к книге правил, она основана на минимальной теории. Это дает возможность читателю схватить сущность статистической методолсгии и избежать ошибок, которые он несомнено сделает, если будет применять методы автоматически.

Объем работы несомненно остается очень ограниченным и следует сожалеть об отсутствии отдельных аспектов. В связи с этим мы думаем было бы целесообразно напомнить читателю, что различные аспекты этого статистического анализа ряда наблюдений уже рассматривалить в технических записках №№ 71, 79, 81, 98 и 100.

Таким образом настоящая техническая записка займет свое законное место среди других работ, если ее рассматривать просто в качестве полезного дополнения к предыдущим работам.

RESUMEN

Nadie niega que actualmente la estadística matemática es la matemática de la climatología. En efecto, la estadística matemática es la ciencia de los modelos aleatorios, y la finalidad del análisis estadístico de las series de observaciones no es otra que determinar, entre los modelos que esa ciencia ofrece, el que mejor representa el comportamiento del fenómeno observado.

De hecho, esta actitud tiene una doble justificación.

Desde el punto de vista teórico, la representación del fenómeno observado mediante un modelo que depende de un mínimo de parámetros hace posible la confrontación de conclusiones estadísticas objetivas con las teorías físicas avanzadas, bien para confirmar su validez o bien para revelar sus insuficiencias; así, pues, el análisis estadístico constituye una fuente de progresos en la esfera de la investigación.

Desde el punto de vista práctico, la eficacia de los métodos de análisis estadístico permite asegurar, merced a su empleo, una valorización óptima de la información acumulada por las series de observaciones y garantiza, por ese mismo hecho, una rentabilidad máxima del capital que tal información constituye. Esa rentabilidad aparece, por otra parte, como una necesidad evidente desde el mismo momento en que se tienen en cuenta las inversiones y los gastos de funcionamiento necesarios para obtener esas series de observaciones, y asimismo desde el mismo momento en que se piensa en el valor económico de esas series, constantemente demostrado en las aplicaciones de la meteorología y de la climatología.

El ajuste de un modelo aleatorio a una serie de observaciones se confunde, en realidad, con el problema general de la estimación estadística y, en la presente Nota Técnica, se ha tenido en cuenta ese problema en el caso de series aleatorias simples y en el de series recurrentes.

De manera más precisa, después de examinar en el primer capítulo las verificaciones más adecuadas para comprobar la conformidad de una serie cronológica con el modelo aleatorio simple, se ha abordado, en el segundo capítulo, el problema de la estimación estadística, sucesivamente en el caso no paramétrico y en el caso paramétrico, y ello tanto para une variable continua como para una variable discontinua; por otra parte, también se ha consagrado un párrafo a la estimación en el caso de series en correlación.

Los modelos recurrentes, periódicos o persistentes, se presentan seguidamente en el tercer capítulo como solución de recambio cuando el modelo aleatorio simple no conviene, y en el mismo se han tenido en cuenta diversas aplicaciones.

La exposición de los métodos se ha efectuado de conformidad con el principio de que un análisis completo comprende necesariamente las tres etapas siguientes:

1. selección del modelo;
2. ajuste y adecuación del modelo seleccionado;
3. cálculo de los errores de estimación asociados al modelo ajustado;

y, si en el tratamiento de los datos, esas tres etapas no aparecen siempre de manera clara, bien porque la selección no haya sido integrada en el método de ajuste o bien porque se base en una verificación de adecuación, no obstante esas etapas se reconocerán cada vez en el enfoque de las soluciones.

Por otra parte, todos los métodos presentados han sido ilustrados por medio de ejemplos corrientes de la climatología con el fin de aclarar al máximo su aplicación. A pesar de ello, para evitar que la obra se reduzca a un libro de recetas, se ha recurrido a un mínimo de elementos teóricos para que el lector pueda comprender lo esencial de la metodología estadística y soslayar los errores que una aplicación demasiado automática de las técnicas le harían cometer.

Sin duda alguna, el alcance de la obra es muy limitado y se echarán de menos ciertas materias. A pesar de ello, nos ha parecido oportuno recordar a ese respecto que diversos aspectos del análisis estadístico de las series de observaciones ya han sido tratados en las Notas Técnicas Nºs 71, 79, 81, 98 y 100.

Por ello, si se considera la presente obra simplemente como un complemento útil de las anteriores, se le habrá conferido el justo lugar que merece y debe ocupar.

TO THE READER

Except for the correction of errata and a few amendments which seemed useful, this publication is a close translation into English of the Technical Note n^{o} 143 "Sur l'analyse statistique des séries d'obervations" by Dr. R. Sneyers, published in 1975.

For the sake of consistency with the tables, which have been reproduced directly from the original French text, a comma has been used throughout this English edition in place of the more conventional decimal point.

CONTENTS

1. SIMPLE RANDOM CHARACTER OF SERIES OF OBSERVATIONS

TEST OF HYPOTHESIS

The simple random character of a series of observations is a basic assumption for the statistical analysis of such a series. Many statistical methods which are used to verify the particular properties of series of observations or to estimate certain parameters typical of the processes which produce these series are only applicable on condition that this first assumption is actually realized. Verification of this is a problem which must be solved before proceeding to any further statistical analysis.

However, as the verification of this hypothesis enters into the general field of the technique of significance tests, it seems useful to us to recall a few connected concepts.

1.1 TEST OF HYPOTHESIS - SIGNIFICANCE TEST

The application of a significance test to a series of observations can arise from the need to respond to two distinct types of question : whether we wish to ensure that the distribution law for the observations possesses well-determined properties in order to make certain estimates or statistical predictions, or whether on the other hand, we want to show to what extent the process which produces these observations is modified by a specific action which has affected it.

In both cases, the test is reduced to the comparison of two hypotheses : one, *the null hypothesis*, which assumes that the distribution function for the observations fulfils certain well-defined conditions; the other, the *alternative hypothesis*, which states that all or part of these conditions are different from those of the null hypothesis. The technique used to separate these two hypotheses consists of the calculation of a statistical function of the observations (*test statistic*) and the rejection or acceptance of the null hypothesis depends on whether the value obtained for the test statistic is or is not inside a critical region (*rejection region*) determined in advance.

Application of a test can lead to two types of error. The first consists of rejecting the null hypothesis when it is true (*Type I error*). It occurs when the null hypothesis is true but the value found for the statistic of the test lies in the region of rejection. This event has a certain probability that can be calculated in certain cases or limited in such a way that it does not exceed a value α_0 chosen beforehand. This value α_0 is the *significance level* of the test.

The second type of error is that which consists of accepting the null hypothesis when it is false (*Type II error*). This occurs when the value found for the test statistic does not lie in the rejection region when the alternative hypothesis is true. If β is the probability of an error of the second kind, the probability for accepting the alternative assumption when it is true is thus $(1 - \beta)$.

The latter probability $(1 - \beta)$ is called the *power* of the test relative to the alternative hypothesis.

As a rule a good test is one which makes the probability of both the Type I error and the Type II error a minimum at the same time, which leads one to suspect that a good test cannot be constructed with any statistic. In addition, as the significance level of a test is always fixed in advance, the best test will be that wich makes the power $(1 - \beta)$ a maximum.

Another property which a good test must possess is that of being *unbiased (without systematic error)* relative to the alternative hypothesis. This condition is achieved as soon as $\alpha_0 \leqslant 1 - \beta$, that is to say when the probability of rejecting the null hypothesis when it is true is no greater than when it is false, which is an obvious necessity.

The null hypothesis is generally accepted when the value of the statistic is close to the median of its probability distribution under this null hypothesis. On the other hand, the rejection region of the test is formed by values of the statistic which differ most from this median. The test is called *one-sided* when the rejection region of the test is formed by values which are all situated on the same side of the median value. It is called *two-sided* when this region is formed by values situated on either side of the median.

The choice of the one-sided or two-sided test is, as a rule, conditioned by the power of one or other form relative to the alternative hypothesis. Note, however, that, by suitably modifying the statistic of a two-sided test, an equivalent one-sided test can be defined. If, for example, in a two-sided test, the acceptance region of the statistic X is formed by the interval $(-X_0, X_0)$, use of the quantity X^2 as the statistic leads to an equivalent one-sided test for which the rejection region is the whole series of values of X^2 such that $X^2 > X_0^2$.

One of the most usual significance levels is $\alpha_0 = 0{,}05$. This amounts to limiting, under the null hypothesis, the frequency of a Type I error to one time out of twenty on average, which seems satisfactory in the majority of cases. Although, in the examples, the most often used level is $\alpha_0 = 0{,}05$, the calculation of the test statistic X is generally completed by calculating the level α_1 from which point the statistic X is significant. This level is directly deduced from the probability which the distribution function of the statistic assigns to the value found for X under the null hypothesis. The advantage of calculating α_1 is that the test can then be carried out simply by comparing α_1 with the α_0 level selected. Furthermore, when the null hypothesis is rejected, α_1 gives, in probability terms, an objective measurement of the degree of incompatibility of the observed situation with the null hypothesis.

Moreover, it will be seen later that the calculation of the level α_1 associated with the value obtained for X is indispensable when we wish to calculate the probability associated with a set of independent tests.

1.2 PARAMETRIC – NON-PARAMETRIC TESTS

There is a wide variety of tests of hypothesis precisely because of the diversity of hypotheses which can be considered. In fact, we can first of all distinguish *parametric* from *non-parametric tests* according to whether, under the null hypothesis, the probability distribution for the test statistic depends on the shape of the distribution underlying the observations or not. Furthermore, in the case of a parametric test, the hypotheses compared can be either *simple* or *composite*. A hypothesis is simple when the values of all parameters of the distribution function for the observations are fixed in advance, which implies that in this hypothesis, a well-determined probability can be associated with a given series of observations. On the other hand, a hypothesis is composite as soon as this condition is not fulfilled, that is, as soon as the value of at least one parameter of the probability distribution for the observations remains indeterminate. In a parametric test, both the null hypothesis and the alternative hypothesis can be either simple or composite. For example, a test for normality is a parametric test in which the null hypothesis is composite. The latter consists, in fact, of

assuming that the observations are distributed independently following a normal distribution of which both the mean and the standard deviation are indeterminate. In this case, on the other hand, all the moments of the distribution function higher than the second order are linked to the mean and the standard deviation by wellknown relationships.

As has been said, non-parametric tests are tests in which the form of the distribution for the observations is not specified. In particular, in such tests the null hypothesis is limited to stating that the observations are distributed independently with the same distribution, whereas the alternative hypothesis casts doubt on either the independence of the observations or the identity of the distribution.

We can see therefore, that non-parametric tests are those whose conclusions have the greatest generality since these conclusions are not linked to a particular form of the distribution function which governs the distribution of the observations.

Hence, the use of such tests seems very appropriate each time that the distribution of the observations is unknown or when there is doubt about its actual form. In addition, this use is recommended in the case of a test having a high *efficiency*, that is to say one whose power is almost as large as that of the most powerful corresponding parametric test. An indication of the efficiency of a non-parametric test is generally given by its *asymptotic efficiency*. In the case of large samples, this gives the ratio by which the sample size must be changed so that the corresponding most powerful parametric test has the same power.

In the same context, it is useful to recall that a parametric test is said to be *robust* when the probability distribution for the test statistic is modified only slightly by a change in the underlying distribution for the observations. A robust test is therefore "almost" non-parametric.

1.3 SIGNIFICANCE LEVEL OF A MULTIPLE TEST

The interpretation of the results of tests of hypothesis applied simultaneously to several series of independent observations is generally made difficult by the fact that it contains an extra risk of a Type II error if each individual test leads to acceptance of the null hypothesis, when this is false; or of a Type I error if some of these tests give significant values of the statistic when the null hypothesis is true.

This raises the problem of selecting a method which will enable us to attribute to the set of results obtained a probability by value of which this set can be considered as globally significant or not. The method chosen will vary, depending on whether, in the absence of individually significant results at the usual level, we wish to ensure that these results, as a whole, are also not significant, or whether, on the other hand, faced with a few individually significant results, we wish to pick out the one or more which can actually be accepted as significant.

1.3.1 The Fisher test

A method which resolves the first case is that of Fisher. It consists of noting that if k is the number of independent series and if α_i, $i = 1, 2, \ldots, k$, with $0 \leqslant \alpha_i \leqslant 1$, are the probabilities (significance levels) associated with each value found for the test statistic, the quantities $-2 \log \alpha_i$ possess, under the null hypothesis, a χ^2 distribution with two degrees of freedom.

It follows that a global test can be based on the fact that, under these conditions, the expression :

$$X = -2 \sum \log \alpha_i \tag{1}$$

has a χ^2 distribution with $2k$ degrees of freedom, the small values of α_i giving rise to large values of X.

1.3.2 The test based on the binomial distribution

The second method is based on the statement that, by applying the same test to a series of independent trials, the probability under the null hypothesis of obtaining at least one significant test for a given level is increasingly large as the number of tests is increased. If follows that a particular test, significant at the level α, can be considered as significant at the level α_0 in the series of k tests only if the probability of obtaining at least one significant test at the level α amongst the k tests does not exceed α_0 .

By virtue of the binomial distribution, this means that the upper limit of α is given by

$$1 - (1 - \alpha)^k = \alpha_0 \tag{1}$$

Furthermore, the second method can be generalized and it can be considered also that if k_1 significant tests have been obtained at the level α, these tests are significant as a whole at the level α_0 , if the probability of obtaining at least k_1 tests at the level α does not exceed α_0.

From the binomial distribution, the upper limit of α is given in this case by the relation

$$\binom{k}{k_1}\alpha^{k_1}(1-\alpha)^{k-k_1} + \binom{k}{k_1+1}\alpha^{k_1+1}(1-\alpha)^{k-k_1-1} + \cdots + \alpha^k = \alpha_0 \tag{2}$$

The calculation of α can in general be carried out using the equation

$$\alpha = \frac{\nu_1 v^2_{\alpha_0}}{\nu_2 + \nu_1 v^2_{\alpha_0}} \tag{3}$$

where $v^2{}_{\alpha_0}$ is the quantile of probability α_0 , for the variance ratio v^2 (or the F variate of Snedecor) with $\nu_1 = 2k_1$ and $\nu_2 = 2(k - k_1 + 1)$ degrees of freedom.

In particular, for $k_1 = 5$, $k = 12$, we have : $\nu_1 = 10$, $\nu_2 = 16$ and for $\alpha_0 = 0{,}05$ we find

$$v^2_{0,05}(\nu_1, \nu_2) = 1/v^2_{0,95}(\nu_2, \nu_1) = 1/2{,}83 \tag{4}$$

from which, with (3), we find

$$\alpha = 0{,}181 \tag{5}$$

1.3.3 Comparison between the Fisher test and the test based on the binomial

It is perhaps useful to stress that the Fisher test and the test based on the binomial are rather different tests. This results from the fact that the set of alternative hypotheses opposed to the null hypothsesis is not the same from one test to another. In the first test, the critical domain can, in effect, be defined by the relation

$$\alpha_1 \alpha_2 \ldots \alpha_k \leqslant \exp\left[-\frac{\chi^2_{(1-\alpha_0)}}{2}\right] \tag{1}$$

whereas in the second, this definition for k_1 probabilities randomly chosen from the k probabilities α_i becomes

$$\alpha_i \leqslant \alpha \tag{2}$$

where α is taken from equation 1.3.2(2).

It thus seems that the group of alternative hypotheses considered is larger in the first test than in the second and that, for this reason, the second test is generally more powerful than the first. It follows that use of the second test is required when we wish to proceed to the selection of significant tests in a series of independent tests forming a multiple test.

Note that this selection can be carried out by combining the two methods. In fact, when the Fisher statistic takes a significant value, the quantities $-2\log \alpha_i$ can be successively removed in decreasing order until a non-significant remainder is obtained. However, in order that the significant test on this remainder is unbiased, the level of this test must be the level α' taken from the equation

$$1 - (1 - \alpha')^{k/(k - k_1)} = \alpha_0 \tag{3}$$

where α_0 is the initial level of the Fisher test and k_1 the number of α_i quantities selected. The significant character of the selected α_i quantities is then established if the last of these is significant at the level α obtained from equation 1.3.2(1). Otherwise, the selections which do not satisfy this condition have only a random significance.

The generalization of this method to the case of a joint test statistic on k parameters is immediate. Note in this connexion that, in the case of a statistic distributed under the null hypothesis, following a χ^2 distribution, it is found from tables that for $\alpha = 0{,}05$, the successive critical values after selection are obtained approximately, by subtracting at least one unit per degree of freedom from the initial critical value. The significant character of the last selection is then established from the last reduction undergone by the test statistic.

This method of selection is also applicable in the case of a multiple test formed of several independent tests applied to the same series of observations.

1.3.3.1 **Example 1.** *Significance level of a series of twelve independent tests*

A trend test was applied to twelve series of maximum values of wind speed obtained at Uccle (Belgium) for each of the twelve months ($i = 1, 2, \ldots, 12$) of the year during 30 years of observation. The values found for the test statistic lead to the probabilities α_i given in Table 1.

TABLE 1. – Values of the probabilities α_i associated with the values of the statistic of the trend test

i	1	2	3	4	5	6	7	8	9	10	11	12
α_i	0,340	0,108	0,134	0,022	0,088	0,116	0,740	0,600	0,560	0,240	0,920	0,042

Applying the Fisher test gives :

$$X = -2 \sum \log \alpha_i = 39{,}58$$

As the critical value of the χ^2 variate with 24 degrees of freedom is equal to 36,4 for the level $\alpha_0 = 0{,}05$, we should reject the null hypothesis for the series of tests.

In order to apply the test based on the binomial distribution, the values of α for $\alpha_0 = 0{,}05$, $k = 12$ and $k_1 = 1, 2, \ldots, 12$, obtained from 1.3.2(2) by means of 1.3.2(3), are first calculated, and the values of α_i are then arranged in increasing order, i.e. α_i'. The Table 2 is then obtained :

TABLE 2 – Ordered values α_i' of the probabilities α_i. Corresponding critical values α

k_1	1	2	3	4	5	6	7	8	9	10	11	12
α	0,004	0,031	0,072	0,123	0,181	0,245	0,316	0,391	0,473	0,562	0,661	0,779
α_i'	0,002	0,042	0,088	0,108	0,116	0,134	0,240	0,340	0,560	0,600	0,740	0,920

In this way we see that the inequality $\alpha_i' < \alpha$ is satisfied for $k_1 = 4, 5, 6, 7$ and 8, which leads to the same conclusion as above. This shows in particular that the statistics obtained in Table 1 for $i = 2, 4, 5$ and 12 that is, for the months of February, April, May and December – form a group of four globally significant values at the 0,05 level and the same applies to the groups of 5, 6, 7 and 8 values which are obtained by adding successively to the four previous values those for $i = 5, 3, 10$ and 1. Although the conclusion here is the same as that obtained by means of the Fisher test, it is however, more informative as regards the nature of the disagreement with the null hypothesis. On the other hand, as the relation $\alpha_i' < \alpha$ is not satisfied for $i = 1$, the test does not enable a specific value significant at the 0,05 level to be pointed out amongst the twelve values of α_i in Table 1.

For selection starting out from the Fisher statistic, the smallest value of the probabilities α_i' gives

$$-2 \log 0{,}022 = 7{,}63$$

from which we get the new value of the statistic after a first selection

$$X = 39{,}58 - 7{,}63 = 31{,}95$$

for which, since the number of degrees of freedom relative to the selected quantity is 2, the critical value becomes

$$36{,}4 - 2 = 34{,}4$$

As the last value of the statistic is less than its critical value, there is no need to carry on with the selection. Furthermore, as the selected probability is greater than the smallest value of α in Table 2, a value taken from equation 1.3.2(1) for $k = 12$ and $\alpha_0 = 0{,}05$, the selection carried out has only a chance significance.

In conclusion, the trend disclosed by the multiple test is too slight to be able to be localized with any accuracy during the year.

1.4 TEST FOR THE SIMPLE RANDOM CHARACTER OF A SERIES OF OBSERVATIONS

The null hypothesis in the specific case consists of assuming that all observations of the series come from the same population and that they are all independent amongst themselves. It is then said that such a series is a *simple random* series.

It is clear that the domain of the alternative hypotheses which can be opposed to the null hypothesis is very wide since it contains all the relations which could arbitrarily be imagined between any term of the series and each of the others. It follows that, in order to verify the purely simple random character of a series, either the null hypothesis can be compared with a very wide range of alternative hypotheses and a not very powerful test accepted for use, or on the contrary the null hypothesis can be compared with some fully specified alternative hypotheses, which then enables more powerful tests to be used.

Since the possible alternative hypotheses are relatively well known in practice, it is generally the second method which is best to use. In particular, in the case of a series of meteorological observations, it seems that the simple

random character of a series can be considered as sufficiently well established if the application of a serial correlation test and a trend test both lead to acceptance of the null hypothesis. This choice seems all the more justifiable as either test possibly enables two extreme persistance forms to be shown, one of very short range (from one term to the other) and the other of long range. However, as the existence of a serial correlation can introduce a systematic error into the trend test and that conversely the serial correlation test is asymptotically independent of any trend test, it is the serial correlation test which must be applied first. In addition, in the absence of any certainty as regards the nature of the distribution function for the observations, it is the non-parametric tests which are most appropriate here.

Furthermore, combining the results from the two tests is done by using either the Fisher test or the test based on the binomial distribution, depending on whether the overall result of the two tests is important by itself or whether one wishes to know which of the two alternatives, trend or serial correlation, must be used in place of the null hypothesis.

1.4.1 Serial correlation test

The non-parametric test suitable for use here is the Wald-Wolfowitz test. If $x_1, x_2, \ldots, x_n$ are the values of the series being considered on which a change of origin has been carried out in such a way that

$$\Sigma x_i = 0 \tag{1}$$

and if also we put

$$x_{n+1} = x_1 \tag{2}$$

the test statistic is the quantity

$$R = \sum_{i=1}^{n} x_i x_{i+1} \tag{3}$$

whose distribution under the null hypothesis, is approximately normal for large values of n. The mean and the variance of the distribution are fairly complex. However, if it is limited to terms of order $1/n$, this becomes

$$E(R) = -S_2/(n-1) \text{ and } \operatorname{var} R = S_2^2/(n-1) \tag{4}$$

where we have put

$$S_2 = \Sigma x_i^2 \tag{5}$$

In addition, it is clear that if the statistic R is replaced by the function

$$r = R/S_2 \tag{6}$$

the mean and the variance of the asymptotic distribution become

$$E(r) = -1/(n-1) \text{ and } \operatorname{var} r = 1/(n-1) \tag{7}$$

As the statistic r is also that of R.L. Anderson's parametric test, in which this statistic has the same asymptotic distribution, it can be deduced both that the asymptotic efficiency of the non-parametric test is equal to 1 and that the R.L. Anderson test is a robust one.

It must be stressed that in the serial correlation test considered, the alternative hypothesis includes only the contingency of a positive serial correlation (persistence). In order to ensure maximum power to the test, therefore, its one-sided form is used, and the null hypothesis is rejected for large values of r.

It follows that the test is reduced to calculation of the quantity

$$u(r)=[(n-1)\,r+1]\,\sqrt{n-1} \tag{8}$$

and of the probability α_1, determined from a standard normal distribution table, such that

$$\alpha_1 = \mathrm{P}\,(u > u(r)) \tag{9}$$

If α_0 is the significance level of the test, the null hypothesis is accepted or rejected at the level α_0 depending on whether $\alpha_1 > \alpha_0$ or $\alpha_1 < \alpha_0$.

Yet another possibility is to adopt the simplified form of the serial correlation test based on the statistic (3), replacing each x_i in the x_i series by

$$+1, \text{ if } x_i > 0 \text{ or } -1, \text{ if } x_i < 0 \tag{10}$$

It can be shown that in this case, the test based on the simplified statistic R becomes equivalent to the *runs test* whose statistic is the number R' of runs obtained by counting the uninterrupted partial series of $+1$ or -1 elements contained in the total series transformed according to (10), on condition than we decide to reject the null hypothesis for small values of R' (one-sided test).

To calculate the level α_1 associated with the value obtained for R', it is noted that, under the null hypothesis, this statistic has an approximately normal distribution of mean and variance :

$$\mathrm{E(R')} = \frac{2n_1n_2}{n} + 1 \quad \text{and} \quad \mathrm{var\,R'} = \frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)} \tag{11}$$

with

$$n_1 + n_2 = n \tag{12}$$

where n_1 and n_2 denote the number of $+1$ and -1 elements respectively in the series.

It is clear, however, that this simplification of the serial correlation test leads to a reduction in its power.

1.4.1.1 Example 2 — *Application of the serial correlation test*

In the series of annual means of the wind speed at Uccle from 1933 to 1969, the original values have been replaced by the rank (order number) which is given to them when they are ordered in increasing magnitude. The numbers y_i obtained in this way are shown in Table 3. In particular, the first observation $y_1 = 20$ occupies the twentieth place in the series arranged in increasing order of magnitude. The only aim of this procedure is to simplify calculations to a maximum.

Consequently, by putting

$$y'_i = y_i - \overline{y} \text{ with } \overline{y} = \Sigma y_i/37 = 19 \tag{1}$$

condition 1.4.1(1) is satisfied.

In order to calculate R, we have

$$\Sigma(y'_i - y'_{i+1})^2 = \Sigma y'^2_i + \Sigma y'^2_{i+1} - 2\Sigma y'_i y'_{i+1} = 2S_2 - 2R \tag{2}$$

TABLE 3. — Rank y_i of the annual means of the wind speed at Uccle from 1933 to 1969 in the series of these means arranged in increasing order. Auxiliary values y'_i, $y'_i - y'_{i+1}$, $y_i - i$, n_i, n'_i necessary for calculating the statistics r, r_s and t. Progressive values of $u(t)$ of the trend test.

i	y_i	y'_i	$y'_i - y'_{i+1}$	$y_i - i$	n_i	n'_i	t	$u(t)$	i'
1	20	1	−11	19	0	19	0	0	37
2	31	12	−6	29	1	29	1	1	36
3	37	18	20	34	2	34	3	1,57	35
4	17	−2	10	13	0	16	3	0	34
5	7	−12	−19	2	0	6	3	−0,98	33
6	26	7	17	20	3	22	6	−0,56	32
7	9	−10	−20	2	1	7	7	−1,05	31
8	29	10	27	21	5	23	12	−0,49	30
9	2	−17	−14	−7	0	1	12	−1,25	29
10	16	−3	−11	6	3	12	15	−1,34	28
11	27	8	−7	16	7	19	22	−0,86	27
12	34	15	10	22	10	23	32	−0,14	26
13	24	5	−9	11	6	17	38	−0,12	25
14	33	14	12	19	11	21	49	0,38	24
15	21	2	−9	6	6	14	55	0,25	23
16	30	11	20	14	11	18	66	0,51	22
17	10	−9	−22	−7	3	6	69	0,08	21
18	32	13	−3	14	14	17	83	0,49	20
19	35	16	23	16	17	17	100	1,01	19
20	12	−7	8	−8	4	7	104	0,58	18
21	4	−15	−32	−17	1	2	105	0,00	17
22	36	17	25	14	20	15	125	0,54	16
23	11	−8	−7	−12	5	5	130	0,19	15
24	18	−1	−10	−6	9	8	139	0,05	14
25	28	9	15	3	15	12	154	0,13	13
26	13	−6	−6	−13	7	5	161	−0,07	12
27	19	0	5	−8	11	7	172	−0,14	11
28	14	−5	6	−14	8	5	180	−0,36	10
29	8	−11	−17	−21	3	4	183	−0,75	9
30	25	6	20	−5	17	7	200	−0,63	8
31	5	−14	4	−26	2	2	202	−1,04	7
32	1	−18	−22	−31	0	0	202	−1,49	6
33	23	4	8	−10	18	4	220	−1,36	5
34	15	−4	−7	−19	12	2	232	−1,44	4
35	22	3	16	−13	19	2	251	−1,32	3
36	6	−13	3	−30	4	1	255	−1,63	2
37	3	−16	−17	−34	2	0	257	−1,99	1

from which we obtain

$$R = S_2 - \Sigma (y'_i - y'_{i+1})^2/2 \tag{3}$$

Since

$$\Sigma(y'_i - y'_{i+1})^2 = 8{,}670, \qquad S_2 = 4{,}218 \tag{4}$$

from (3), 1.4.1(6) and 1.4.1(8), with $n = 37$ this becomes

$$R = -117, \quad r = -0{,}02774, \quad u(r) = 0{,}00023$$

from which we get

$$\alpha_1 = P(u > 0{,}00023) \cong 0{,}5 \tag{5}$$

We conclude that at the level $\alpha_0 = 0{,}05$, we have $\alpha_1 > \alpha_0$ which allows the null hypothesis to be accepted.

For the application of the runs test, if only the series of + and − signs are retained, the series y'_i becomes

$$+\ +\ +\ -\ -\ +\ -\ +\ -\ -\ +\ +\ +\ +\ +\ +\ -\ +\ +\ -\ -\ +\ -\ -\ +\ -\ \text{o}\ -\ -\ +\ -\ -\ +\ -\ +\ -\ -$$

If the median 0 is counted as 0,5 in the number of + signs and − signs, using 1.4.1(11) we find

$$n_1 = n_2 = 18{,}5, \quad \text{E(R')} = 19{,}5 \quad \text{and} \quad \text{var R'} = 8{,}9931$$

Since we find that R' = 20 or 22, depending on whether 0 is replaced by − or + in the series considered, by taking the average R' = 21, we obtain

$$\alpha_1 = \text{P(R'} < 21) = \text{P}(u < (21 - 19{,}5)/\sqrt{8{,}9931})$$
$$= \text{P}(u < 0{,}5) = 0{,}69.$$

We conclude that, as in the previous test, at the level $\alpha_0 = 0{,}05$, we have $\alpha_1 > \alpha_0$. Thus the null hypothesis may be accpeted.

1.4.2 Trend tests

Two non-parametric tests can be used to demonstrate the possible existence of a trend : the first is based on the Spearman coefficient r_s, the second on the rank correlation statistic t from Kendall (Mann test).

1.4.2.1 Calculation of the Spearman coefficient r_s

For the first test, the original observations x_i, $i = 1, 2, \ldots, n$ are replaced by the rank y_i which is given to them when they are arranged in increasing order of magnitude and the test statistic is the correlation coefficient r_s between the i and y_i series; this coefficient can be calculated by means of the formula

$$r_s = 1 - \frac{6}{n(n^2 - 1)} \Sigma (y_i - i)^2 \tag{1}$$

Using the null hypothesis, the distribution of this quantity is asymptotically normal with

$$\text{E}(r_s) = 0 \quad \text{and} \quad \text{var}\, r_s = \frac{1}{n - 1} \tag{2}$$

It is clear that in the absence of any assumptions regarding the existence of a trend in a given direction, the test is correct only if its two-sided form is adopted, that is to say if the null hypothesis is rejected for large values of $|r_s|$. In these conditions, after having calculated r_s, it is useful to determine the probability α_1, by means of a standard normal distribution table, such that

$$\alpha_1 = \text{P}(|u| > |u(r_s)|) \tag{3}$$

where

$$u(r_s) = r_s \sqrt{n - 1} \tag{4}$$

and the null hypothesis is accepted or rejected at the level α_0, depending on whether $\alpha_1 > \alpha_0$ or $\alpha_1 < \alpha_0$.

In the case of significant values of $|r_s|$, an increasing or decreasing trend is observed depending on whether $r_s > 0$ or $r_s < 0$.

1.4.2.2 Calculation of the Kendall coefficient t (Mann test)

In the second test, for each element x_i or, what amounts to the same thing, for each element y_i, the number n_i of elements y_j preceding it $(i > j)$ is calculated such that $y_i > y_j$.

The test statistic t is then given by equation

$$t = \sum_i n_i \tag{1}$$

and its distribution function, under the null hypothesis, is asymptotically normal, with mean and variance

$$\mathrm{E}(t) = \frac{n(n-1)}{4} \quad \text{and} \quad \operatorname{var} t = \frac{n(n-1)(2n+5)}{72} \tag{2}$$

It is clear that, as for the test based on the Spearman coefficient r_s, in the absence of any assumption regarding the existence of a trend in a given direction, the test is correct only in its two-sided form. The null hypothesis must therefore be rejected for high values of $|u(t)|$ with :

$$u(t) = [t - \mathrm{E}(t)] / \sqrt{\operatorname{var} t} \tag{3}$$

In particular, if the probability α_1 is determined using a standard normal distribution table such that

$$\alpha_1 = \mathrm{P}(|u| > |u(t)|) \tag{4}$$

the null hypothesis is accepted of rejected at the level α_0, depending on whether we have $\alpha_1 > \alpha_0$ or $\alpha_1 < \alpha_0$.

When the values of $u(t)$ are significant, an increasing or decreasing trend can be observed depending on whether $u(t) > 0$ or $u(t) < 0$.

1.4.2.3. Comparison between the r_s and t statistics

It must be said that both tests are very efficient and that, in particular, they have the same asymptotic efficiency relative to the most powerful corresponding parametric test in the case of a normal linear regression.

However, when a series shows a significant trend and when we wish to locate the period from which the trend is demonstrated, the second test statistic lends itself better to the sequential calculation required for this purpose.

1.4.2.4. **Example 3.** *Application of trend tests*

If we return to the series in Table 3 where the numbers y_i are the ranks given to the values x_i of the original series arranged in increasing magnitudes, for the first test, we obtain

$$\sum (y_i - i)^2 = 11.372 \tag{1}$$

from which, using 1.4.2.1(1) and (4), it is deduced that

$$r_s = -0{,}348, \quad u(r_s) = -2{,}088$$

and

$$\alpha_1 = \mathrm{P}(|u| > 2{,}088) = 0{,}037 \tag{2}$$

It follows that at the level $\alpha_0 = 0{,}05$, the null hypothesis must be rejected in favour of the existence of a decreasing trend. For the second test, we find

$$t = \Sigma n_i = 257 \tag{3}$$

from which, with 1.4.2.2(2) and (3), we obtain

$$u(t) = -1{,}988 \quad \text{and} \quad \alpha_1 = P(|u| > 1{,}988) = 0{,}047 \tag{4}$$

The same conclusion is drawn from this at the level $\alpha_0 = 0{,}05$ as with the first test.

1.4.2.5 Determination of the simple random character of the series of annual wind speed averages at Uccle from 1933 to 1969

If we consider the fact that we have associated the serial correlation test and the trend test in order to verify the simple random character of a series, the calculation of the overall level of significance is still required in order to bring the analysis to a conclusion.

Reconsidering the methods given in 1.3, and using the values of α_1 given in 1.4.1.1(5) and 1.4.2.4(2), we find for the Fisher test

$$X = -2\,(\log 0{,}5 + \log 0{,}037) = 7{,}98$$

a value for which the table of χ^2 with four degreees of freedom allows us to write

$$\alpha_1 = P(\chi^2 > 7{,}98) = 0{,}094$$

that is to say, for $\alpha_0 = 0{,}05$, $\alpha_1 > \alpha_0$.

In the same way, the test based on the binomial distribution for $k = 2$ and $\alpha_0 = 0{,}05$ leads to the data in the following table :

k_1	1	2
α	0,025	0,223
α_i'	0,037	0,5

from which we find that in no case do we have $\alpha_i' < \alpha$.

Thus we observe that the null hypothesis is rejected neither by the Fisher test nor by the test based on the binomial distribution.

1.4.2.6 Progressive analysis of a series by means of the statistic u(t)

As has been stated above, the statistic of the second trend test lends itself better to calculation when, in the case of a significant trend, we wish to locate the start of the phenomenon by means of a sequential analysis.

This comes immediately from the fact that the value of t for the series formed of the first i terms is none other than the sum

$$t_i = n_1 + n_2 + \ldots + n_i \tag{1}$$

that is, the sum 1.4.2.2(1) stopped at the i^{th} term. The corresponding value of $u(t_i)$ is obtained by putting $n = i$ in the equations 1.4.2.2(2) and (3).

In the example in Table 3, this procedure shows, in particular, that only the last value of $u(t_i)$ exceeds the absolute value of the critical value at the 0,05 level, which is 1,96.

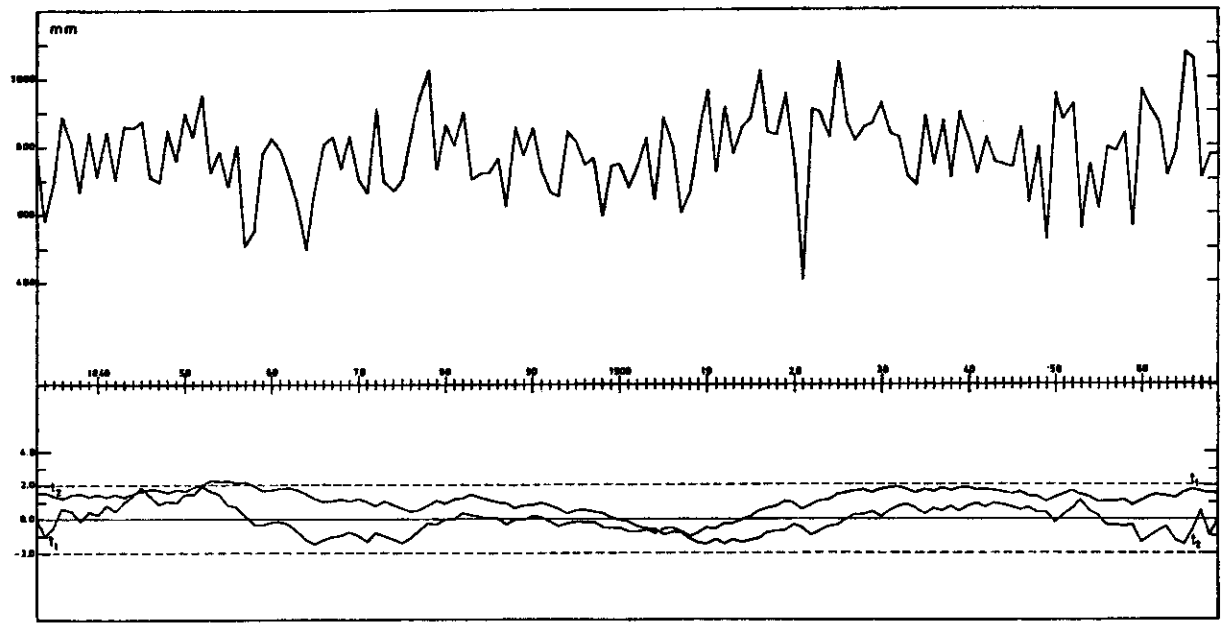

Figure 1 – Annual rainfall at Brussels-Uccle from 1833 to 1969. Sequential values of the statistics $u(t)$ and $u'(t)$

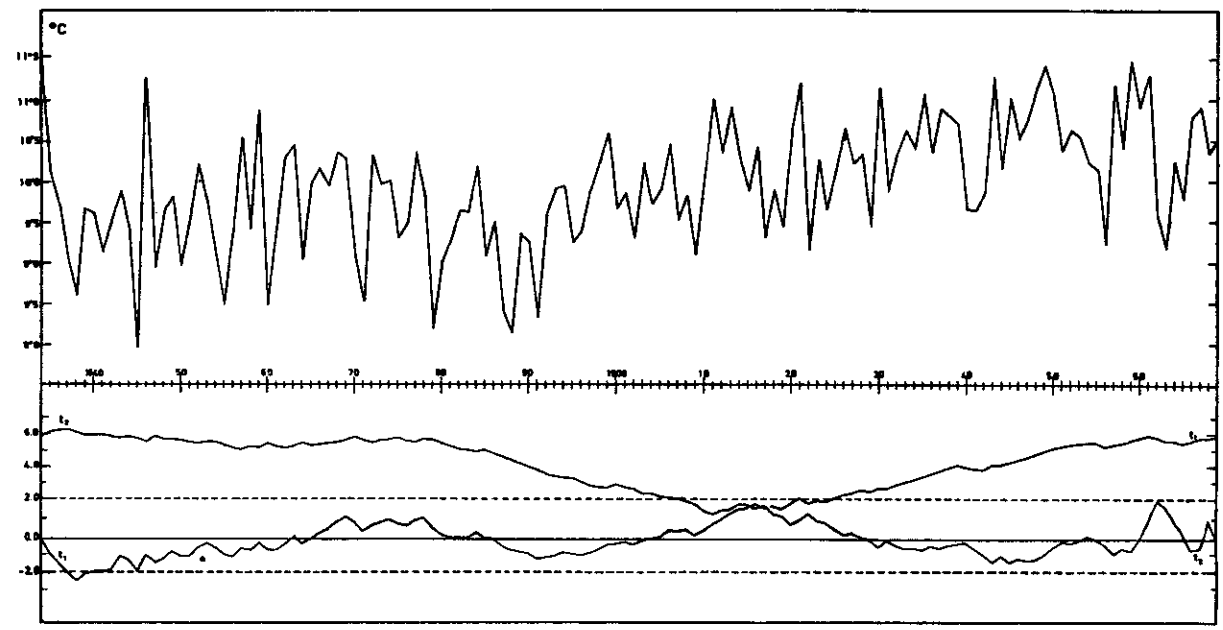

Figure 2 – Annual averages of the air temperature at Brussels-Uccle from 1833 to 1969. Sequential values of the statistics $u(t)$ and $u'(t)$

We may therefore assume that, if the trend is real, its effect is only very recent and it is advisable to await later observations to obtain confirmation from them.

Note that this principle can be usefully extended to the backward series.

In this case, we calculate the number n_i' of y_j terms for each y_i term, such that $y_i > y_j$ with $i < j$, which gives a check on the first calculation, since we have

$$n_i + n_i' = y_i - 1 \tag{2}$$

If we then put

$$i' = (n + 1) - i \quad \text{and} \quad n_i' = n_{i'}, \tag{3}$$

the values of u_i' for the backward series are given by equation

$$u_i' = -u(t_{i'}) \tag{4}$$

which in particular leads to

$$u_1' = u_n \tag{5}$$

In the absence of any trend in the series, the graphical representation of u_i and u_i' in terms of i generally gives curves which overlap several times, whereas in the case of a significant trend, the intersection of these curves enables the start of the phenomenon to be located approximately.

Figures 1 and 2 give typical examples of both cases.

1.4.2.7 Case of series containing equal terms

It occurs fairly frequently that in a series of observations two or more observations are identical. This can arise either from the natural or technical limits of the observation accuracy (reading accuracy, rounding off) as is often the case for continuous variables such as temperature, pressure, etc., or from the very nature of the distribution governing these observations, notably when it concerns a discrete quantity such as the number of frosty or stormy days, etc.

It is clear that, in all cases, the calculation of the statistics of the previous tests must be suitably modified to avoid systematic errors.

In the case of the determination of the ranks y_i of equal terms, this modification consists of giving to each of these terms the mean of the ranks which they would occupy if they followed one another directly in the ordered series. For example, if rank 16 is occupied by two or three equal terms, the rank y_i to give to each of these would be 16,5 or 17 respectively.

For the calculation of r_s it is useful to return to the original equation, which is, with $\Sigma i = \Sigma y_i$,

$$r_s = \frac{\Sigma i y_i - (\Sigma i)^2/n}{s_i s_{y_i}} \tag{1}$$

where

$$s_i^2 = \Sigma i^2 - (\Sigma i)^2/n \quad \text{and} \quad s_{y_i}^2 = \Sigma y_i^2 - (\Sigma i)^2/n \tag{2}$$

Finally, for calculation of the statistic t, it will be sufficient to increase the number of relations $y_i > y_j$ for $i > j$ by half of the number of relations $y_i = y_j$ to obtain a suitable value of n_i.

1.4.3 Test for stability of the variance

In the previous sections, the test for the simple random character of a series of observations has been reduced to the joint application of a serial correlation test and a trend test because of the fact that, amongst the alternative hypotheses which can occur, the only ones considered were those for correlation between succeeding terms of the series and for a systematic deviation from the central value of the series during the observation period.

By proceeding in this way, however we neglect another form of possible perturbation, that of the variation in the dispersion of the values of the series about the central value during the course of the observation period.

When a perturbation of this type is suggested, for example after considering the graphical representation of the observed values, the method of verifying whether it is true or not consists in calculating the series of absolute values of the deviations from the mean of the original series and in applying a trend test to the new series. If the effect is real, the grouping at the start or end of the series of the largest deviations leads to significant values of the statistic of the test, whereas the calculation of its values sequentially allows us to check that such a grouping has not occured inside the series.

In addition, as is easily verified for the y_i ranks of the observations in the ordered series, we have, using the null hypothesis of a simple random series,

$$\operatorname{cov}\left(y_t, \left| y_t - \frac{n}{2} \right|\right) = 0 \tag{1}$$

where n is the length of the series. It can be concluded from this that, under the null hypothesis of a simple random series, the trend test applied to the original series and the same test applied to the series of absolute values of deviations from the mean are independent whenever the probability distribution for the observations is approximately symmetrical.

In this case, it follows that the overall probability associated with the two trend tests can be calculated either by means of the Fisher statistic or by means of the binomial distribution.

In particular, by virtue of (1), this condition is achieved if the trend tests are applied to the series of ranks y_i and the series of absolute values

$$\left| y_i - \frac{n}{2} \right|.$$

1.5 REFERENCE WORKS CONSULTED

The method for determining the simple random character of a time series is that described in [12]. The definitions relating to the tests of hypothesis are generally in agreement with those found in general textbooks on mathematical statistics.

Otherwise, the instructions as to the use and efficiency of the tests described come mainly from [5], [8] and [10], the calculation of α by means of equation 1.3.2(3) is taken from page 674 of [6] and the examples are taken from [13] or from the Uccle climatology archives. Figure 1 and 2 are taken from [14].

Note, however, that the list of general sources quoted in the bibliography is a simple working list which professes to be neither exhaustive nor exclusive.

2. ON STATISTICAL ESTIMATION

Statistical estimation can be considered as the major part of the statistical analysis of series of observations. As a matter of fact, the adoption of probabilistic hypotheses for these observations enables us to link probabilities with any given observable value (whether it is a properly observed value or not) or to proceed with the inverse operation, namely to determine the value of the observed element corresponding to a given probability. This is generally done by assigning values (estimates) best suited to the constants (parameters) which determine the distribution function for the observations, but it can also be done without this intermediate step. Furthermore, the operation each time provides an interval, called the confidence interval, which sets the accuracy of the estimates obtained.

It follows that the statistical estimation is a rational method for comparing observation series which, without this, would retain only a purely documentary value. It should be stressed, moreover that it is the only method which makes the practical application of climatological data possible in problems where the user must, for example, introduce means whose accuracy is known or extreme values where the risk is specified.

In the following part, we first of all examine the case of statistical estimation based on series of observations forming simple random samples. This shows that checking this hypothesis by means of the tests developed in Chapter 1 is a necessary preliminary step before applying the method of estimation.

Furthermore, we have made the distinction between parametric and empirical estimation, knowing that one takes account of the shape of the distribution (which is assumed to be known in advance), and that the other proceeds without knowledge of it. In the first case, only the parameters of the distribution function for the observations are unknown, whereas in the second it is assumed simply that the distribution has the analytical properties necessary for mathematical statistical calculation.

It immediately follows that parametric estimation is more precise than empirical estimation which, on the other hand, is more general. This is the reason why in practice, the problems of estimation are resolved by adjusting a distribution function to the observations, with the empirical estimation being kept possibly as a last resort to verify the compatibility of the series of observations with the fitted distribution.

It is as well to note, however, that because of its greater simplicity, the use of empirical estimation is still to be recommended when extreme precision is not required.

The case of continuous variables, that is to say those where the distribution function is a continuous function of the random variable, is examined first, and is followed by that of discrete distributions.

2.1 EMPIRICAL ESTIMATION

As we have just pointed out, this estimation method is based on elementary properties which link the series of observations to the distribution function of these observations, or to the simplest parameters of this function whose analytical shape, however, remains entirely unknown.

In the case of a continuous variable, these properties are related either to the order statistics and to the empirical distribution function of the series of observations, or to the empirical mean and variance of the series.

2.1.1 Distribution functions for the order statistics. Confidence intervals

It is known that if $0 \leqslant F(x) \leqslant 1$ is the distribution function of the observed variable x and if x_0 is any observation whatsoever, by definition

$$F(x) = P(x_0 < x) \tag{1}$$

Since $F(x)$ is a monotonic increasing function of x, then $x_1 < x_2$ means that $F(x_1) < F(x_2)$.

If a simple random series is then available with n observations $x_1, x_2, \ldots, x_n$ and if $\Phi_n(x)$ denotes the distribution function of the largest value x'_n amongst these n values, we have, by the joint combination of the probabilities,

$$\Phi_n(x) = P(x'_n < x) = P(x_1 < x)P(x_2 < x) \ldots P(x_n < x) \tag{2}$$

that is to say, with (1).

$$\Phi_n(x) = [F(x)]^n \tag{3}$$

Furthermore, if these n observations are arranged in increasing order of magnitude in such a way that we have

$$x'_1 \leqslant x'_2 \leqslant \ldots \leqslant x'_m \leqslant \ldots \leqslant x'_n \tag{4}$$

values which are then called *order statistics* of the series, by writing F for $F(x)$, the distribution function $\Phi_m(x)$ of the m^{th} value or of the order statistic x'_m is given in general by the equation

$$\Phi_m(x) = \binom{n}{m} F^m (1-F)^{n-m} + \binom{n}{m+1} F^{m+1} (1-F)^{n-m-1} + \cdots + F^n \tag{5}$$

Thus the function which has already been introduced in the test based on the binomial distribution is found again, which, for calculation of F, when $\Phi_m(x)$, m and n are given, enables us to refer to the model given in 1.3.2 by means of equations (2) and (3).

The result is that if by means of equation (5) the values $F_1 < F_2$ of F corresponding to two values of $\Phi_m(x)$: $\Phi_1 < \Phi_2$, given in advance, are calculated, the interval (F_1, F_2) contains, with probability $(\Phi_2 - \Phi_1)$, the value of F corresponding to the order statistic x'_m. In other words, it can be stated that if several series of n observations are carried out, the value of $F(x)$ associated with the observation x'_m, i.e. $F_m = F(x'_m)$, is contained between F_1 and F_2 with a probability $(\Phi_2 - \Phi_1)$, that is to say, on average in $100(\Phi_2 - \Phi_1)\%$ of the cases. The interval (F_1, F_2) is a *confidence interval* of F_m at the $(\Phi_2 - \Phi_1)$ *level*.

It is clear that, depending on the requirements, we can make either $\Phi_1 = 1 - \Phi_2 = \alpha/2$ or $\Phi_1 = 0$ and $1 - \Phi_2 = \alpha$ or $\Phi_1 = \alpha$ and $\Phi_2 = 1$ in order to obtain an interval (F_1, F_2) at the level $(1 - \alpha)$.

In the first case, a *two-sided* confidence interval is constructed and in the two others, a *one-sided* confidence interval. In addition, F_1 and F_2 are the *confidence limits* at the level $(1 - \alpha)$.

Inversely, the determination of the confidence intervals (F_1, F_2) for the level $(1 - \alpha)$ for each order statistic $x'_1, x'_2, \ldots, x'_n$ allows a confidence interval to be constructed for value x_0 of x corresponding to a given value F_0 of F.

In reality, let x'_0 and x''_0 be the order statistics for which the confidence intervals are respectievely : (F_0, F'_2) and (F''_1, F_0). As the confidence limits F_1 and F_2 of $F(x'_m)$ increase with x'_m, we obtain $F_0 > F_1$ as soon as $x'_0 > x'_m$ and $F_2 > F_0$ as soon as $x'_m > x''_0$. From this it follows that for F_0 we have

$$F_2 > F_0 > F_1 \tag{6}$$

as soon as, for the corresponding value x_0 of x, we have

$$x'_0 > x_0 > x''_0 \tag{7}$$

It is deduced from this, that for all the values of x_0 for interval (7), the given value F_0 lies within the corresponding confidence interval (F_1, F_2).

The interval (x'_0, x''_0) can thus be considered as a (two-sided) confidence interval of x_0 for the level $(1 - \alpha)$.

In addition, if F_0 is equal to none of the values of the limit F_1, but is found to be contained between two consecutive values, by virtue of the continuity of $F(x)$, x'_0 can be estimated by linear interpolation. The same remark applies for x''_0.

2.1.2 Central values, variance and covariances of the variables F_m. Estimation of $F(x'_m)$.

In practical applications, the use of a confidence interval can be shown to be of little use. This is why it is generally desirable to give F_m a specific value, as appropriate as possible, which can then be used in the calculations. Preference should naturally be given to a central value so that the difference in absolute terms compared with the real value is likely to be very small. The choice can, therefore be made between the mean, the median or the mode.

The mean, in the case of the largest value, is immediately found by writting Φ_n for $\Phi_n(x)$:

$$E(F_n) = \int_0^1 F d\Phi_n = n/(n+1) \tag{1}$$

and generally, for the m^{th} value :

$$E(F_m) = m/(n+1) \tag{2}$$

Similarly, by seeking the value of F_m which makes $d\Phi_m$ a maximum, we find the mode $\tilde{F}_m$, with $m \neq 1$:

$$\tilde{F}_m = (m-1)/(n-1) \tag{3}$$

Finally, for the median $\check{F}_m$, it is known that the approximate value is

$$\check{F}_m = (m - 0{,}3)/(n + 0{,}4) \tag{4}$$

It can thus be stated that for m close to $n/2$, the three central values differ little amongst themselves. On the other hand, their values diverge fairly rapidly for the extreme values of m. In particular, if we call the probability

$(1 - F)$ the *exceedance probability*, it is immediately seen that for the largest value, the mean, the mode and the median of F_n assume exceedance probabilities equal to $1/(n+1)$, 0 and $0{,}7/(n+0{,}4)$.

We can see that as soon as $n > 1$, from the three central values, it is the mean which gives the probability of exceedance closest to the observed value $1/n$, from which we conclude that it is the mean which is best suited to be the estimate of F_m whatever the value of m. We can take as the empirical estimate of $F(x'_m)$

$$\hat{F}_m = \hat{F}(x'_m) = m/(n+1) \tag{5}$$

By associating with the estimates of F_m given by (5) the corresponding two-sided confidence intervals (F_1, F_2) calculated for the same level $(1 - \alpha)$, we obtain the best information regarding the accuracy of the estimates.

Another method for characterizing the accuracy of the estimate $\hat{F}_m$ consists of determining the variance of the variable F_m, by means of the equation

$$\sigma^2_{F_m} = \text{var } F_m = E[(F_m - \hat{F}_m)^2] = \hat{F}_m(1 - \hat{F}_m)/(n+2) \tag{6}$$

knowing that, for values of m close to $n/2$, the distribution of F_m is asymptotically normal.

The calculation of this variance can also be used to check the calculation of the confidence interval (F_1, F_2). If $u_0 > 0$ is the quantile of the reduced normal variate u, such that

$$1 - \alpha = P(|u| < u_0) \tag{7}$$

and if (F_1, F_2) is the two-sided confidence interval at the $(1 - \alpha)$ level of F_m, we have with good approximation

$$2u_0\sigma_{F_m} = F_2 - F_1 \tag{8}$$

and this is the case even for extreme values of m.

In addition, for order statistics of rank m and m', we note that

$$\text{cov}(F_m, F_{m'}) = E[(F_m - \hat{F}_m)(F_{m'} - \hat{F}_{m'})] = \hat{F}_m(1 - \hat{F}_{m'})/(n+2) \tag{9}$$

from which we derive that any variable of the form $(p_1 F_m + p_2 F_{m'})$ has the variance

$$\text{var}(p_1 F_m + p_2 F_{m'}) =$$

$$= [p_1^2 \hat{F}_m(1 - \hat{F}_m) + p_2^2 \hat{F}_{m'}(1 - \hat{F}_{m'}) + 2p_1 p_2 \hat{F}_m(1 - \hat{F}_{m'})]/(n+2) \tag{10}$$

In particular, in the case of estimation by linear interpolation between two order statistics with ranks m and $(m+1)$, with

$$x'_m < x_0 < x'_{m+1}, \quad p = (x_0 - x'_m)/(x'_{m+1} - x'_m), \quad q = 1 - p \tag{11}$$

we have

$$\hat{F}(x_0) = q\hat{F}_m + p\hat{F}_{m+1} \tag{12}$$

and with good approximation

$$\sigma[F(x_0)] = q\sigma_{F_m} + p\sigma_{F_{m+1}} \tag{13}$$

or again

$$\sigma[F(x_0)] = \sqrt{\hat{F}(x_0)[1 - \hat{F}(x_0)]/(n+2)} \tag{14}$$

2.1.3 The empirical distribution function. The Kolmogorov test

By definition, the empirical distribution function $F_e(x)$ is the function which gives the frequency of the observations less than x, that is to say we have

$$\begin{aligned} F_e(x) &= 0 && \text{for} && x \leqslant x'_1 \\ F_e(x) &= (m-1)/n && \text{for} && x'_{m-1} < x \leqslant x'_m,\ m = 2, 3, \ldots, n \\ F_e(x) &= 1 && \text{for} && x'_n < x \end{aligned}$$

As the true distribution function $F(x)$ is a monotonic increasing function, the difference $F(x) - F_e(x)$ over the interval $x \leqslant x'_1$ takes its largest value for $x = x'_1$. In the same way, for the interval $x'_{m-1} < x \leqslant x'_m$, $m = 2, 3, \ldots, n$ the difference $F(x) - F_e(x)$ takes its largest value for $x = x'_m$. If these various maxima are denoted by D_m this then becomes

$$D_m = F(x'_m) - F_e(x'_m) = F(x'_m) - (m-1)/n \tag{12}$$

for $m = 1, 2, \ldots, n$.

We know that for the probability $\frac{\alpha}{2}$ the constant $D_\alpha > 0$ can be determined such that

$$1 - \frac{\alpha}{2} = P(D_m < D_\alpha), \qquad \text{whatever the } m, \tag{3}$$

from which we can deduce that with probability $(1 - \alpha/2)$, all the inequalities

$$\begin{aligned} F(x) &< D_\alpha && \text{for} && x \leqslant x'_1 \\ F(x) - (m-1)/n &< D_\alpha && \text{for} && x'_{m-1} < x \leqslant x'_m, \quad m = 2, 3, \ldots, n \end{aligned} \tag{4}$$

are *simultaneously* satisfied.

In addition, when applied to the distribution for observations taking opposite signs, that is, to the order statistics

$$-x'_n \leqslant -x'_{n-1} \leqslant \ldots \leqslant -x'_1 \tag{5}$$

the same principle with the same probability $(1 - \alpha/2)$ gives the simultaneous inequalities :

$$\begin{aligned} 1 - F(x) &< D_\alpha && \text{for} && x \geqslant x'_n \\ \frac{m}{n} - F(x) &< D_\alpha && \text{for} && x'_{m+1} > x \geqslant x'_m, \quad m = 1, 2, \ldots, (n-1) \end{aligned} \tag{6}$$

By making $x = x'_m$, $m = 1, 2, \ldots, n$ in equations (4) and (6) we then obtain with the probability $(1 - \alpha)$, the simultaneous equations

$$\frac{m}{n} - D_\alpha < F(x'_m) < D_\alpha + \frac{m-1}{n} \tag{7}$$

used in the *Kolmogorov test* to decide whether a given distribution function $F(x)$ can be considered as a probability distribution for the observations or not.

It is clear that the confidence limits calculated in this way are larger than those given directly by the distribution functions of the order statistics. This is why Kolmogorov's method should be used only if we wish to determine a domain in which the distribution function $F(x)$ is completely specified with the probability $(1 - \alpha)$. On the other hand, if we are interested in a *particular value* of the observed variable, which is generally the case, it is the method based on the distribution functions for the order statistics which is used.

2.1.4 Empirical mean and variance

We know that the mean μ and the variance var x of a variate x are respectively the first-order moment with respect to zero and the second-order central moment of its distribution, i.e.

$$\mu = E(x) = \int_0^1 x dF \quad \text{and} \quad \text{var } x = E(x - \mu)^2 = \int_0^1 (x - \mu)^2 dF \tag{1}$$

In addition, the *shape* of the distribution for the variate x can be characterized in many cases by the value taken by the central moments of order $p > 2$:

$$\mu_p = \int_0^1 (x - \mu)^p dF \tag{2}$$

It follows that the estimation of the mean and the variance does not generally allow the distribution to be determined completely without preliminary hypothesis regarding its shape, i.e. the equations which link the higher-order moments to the mean and to the variance.

On the other hand, an *empirical estimation* of the mean and the variance can be carried out without making assumptions about the distributional shape as far as we can assume that the expressions defined in (1) exist and have a finite value (that is to say that the integrals do not diverge).

In particular, if we put

$$\hat{\mu} = \Sigma x_i / n \quad \text{and} \quad \hat{\text{var}}\, x = \Sigma (x - \hat{\mu})^2 / (n - 1) \tag{3}$$

we determine the empirical *mean* and *variance* of the element x respectively, which have the property of being *unbiased* estimates of the true mean and variance, that is to say,

$$E(\hat{\mu}) = \mu \quad \text{and} \quad E(\hat{\text{var}}\, x) = \text{var } x \tag{4}$$

whatever the shape of the distribution.

In addition, by applying the central limit theorem, a confidence interval for the true mean μ can be determined using the calculation of $\hat{\mu}$ and of $\hat{\text{var}}\, x$. In effect, because of this theorem, we can state that as soon as n is large enough and if several series of n independent observations belonging to the same population are available, the various values of $\hat{\mu}$ are themselves distributed asymptotically as normal with mean μ and variance $(\text{var } x)/n$. From this it follows that the ratio

$$(\hat{\mu} - \mu)\sqrt{n}/s \quad \text{with} \quad s^2 = \hat{\text{var}}\, x \tag{5}$$

is approximately distributed as a Student's t variate with $(n-1)$ degrees of freedom. From this, if $t_0 > 0$ denotes the quantile of this variate for which we have

$$P(|t| < t_0) = 1 - \alpha \tag{6}$$

the interval defined by the relation

$$|\hat{\mu} - \mu| < t_0 s/\sqrt{n} \tag{7}$$

is a confidence interval of the mean μ at the level $(1-\alpha)$.

2.1.5 Case of several samples. Kruskal-Wallis homogeneity test

In the previous section, it appears that the estimates considered become more accurate as the sample size n is increased. This is why, when several independent series of observations of the same element are available, it can be useful to consider the hypothesis that these series are homogeneous, that is to say that they all have the same distribution. In this case, all the observations can be grouped together as a single sample whose sample size is then larger than each of the sample sizes of the original series.

The overall sample will then give better estimates than those deduced from each of the separate series.

This being the case, in order to verify the homogeneity of the series, we could limit ourselves to verifying the homogeneity of the empirical means whilst ensuring the compatibility of their empirical variance with the empirical variance of the total sample by testing the ratio of the variances. However, such a test will be only approximate and, moreover, it leaves the question of the homogeneity of the variances unanswered.

It is for this reason that it is better to use a very efficient non-parametric test which can at the same time be used to verify the homogeneity of both the means and the variances.

In the proposed test (the Kruskal-Wallis test) the observations of each series are replaced by the rank that these observations occupy in the total ordered sample. Furthermore, if k is the number of independent series, n_j, $j = 1, 2, \ldots, k$ is the sample size of the series j and if r_{ij} is the rank occupied in the total ordered sample by the observation i of the series j, we put

$$R_j = \sum_i r_{ij}, \qquad n = \Sigma n_j \tag{1}$$

and calculate the statistic

$$X = \frac{12}{n(n+1)} \sum_j \frac{R_j^2}{n_j} - 3(n+1) \tag{2}$$

the large values of which lead to the rejection of the hypothesis of homogeneity of the series.

With the null hypothesis of homogeneous means, the statistic X has approximately a χ^2 distribution with $(k-1)$ degrees of freedom. The acceptance or rejection of this hypothesis can then be linked to the probability of the value found for X in this distribution.

The homogeneity of the variances is satisfied as in 1.4.3 by applying the same test to the series obtained, this time replacing each observation (or the rank r_{ij}) by the absolute value to the deviation from the general mean.

As both applications of this test can be considered as being independent, under the same conditions as those given in 1.4.3, the test for the whole is carried out using the methods described in 1.3.

Note further that, in the case of indeterminate rank because of the equality of certain values, to each value, as in 1.4.2.7, will be attributed the mean of the ranks that one or other value would have if they succeeded each other directly in the ordered sample.

In this case, the test statistic X is calculated as described, but it must then be corrected by dividing by

$$1 - \frac{\Sigma T}{n^3 - n} \tag{3}$$

where $T = t^3 - t$, t being the number of equal values in one group of equal values and the sum ΣT being extended to all the groups of equal values.

Finally, a check of the values obtained for R_j can be carried out using the equation

$$\Sigma R_j = n(n + 1)/2 \tag{4}$$

2.1.6 Graphical representation of a series of observations

The graphical representation of a series of observations generally aims to give an overall picture of the distribution of the observations. If, in addition, we proceed in such a way that, at each observation, the empirical estimate of the value of the distribution function and the confidence interval at the chosen level are made to correspond, an overall view of the results obtained and a method of making new estimates very rapidly by linear interpolation are simultaneously available.

The most general use consists of using a functional scale for the probabilities in such a way that the theoretical distribution function is represented on the graph by a line. In the absence of any indication regarding the shape of the distribution function, however, a linear scale can be used both for the observations and for the probabilities. The latter case, in fact, has the advantage that the estimate $\hat{F}_m$ of the probabilities associated with the order statistics are equidistant on the co-ordinate axis of the probabilities, and so an arbitrary linear scale can be used.

For all methods, while it is true that using a functional scale is a simple method of obtaining indications about the shape of the sample, it is however only by means of an adequate goodness-of-fit test that it can be decided whether the shape of the sample is representative of the shape of the theoretical distribution or whether it differs from it significantly.

2.1.7 **Example 4.** *The monthly rainfall totals collected at Brussels-Uccle in % of the average for odd months*

The example proposed relates to the monthly totals of precipitation collected at Brussels-Uccle during ten consecutive years. However, in order to eliminate the differences due either to the seasonal variation of the mean rainfall, or to the unequal length of the various months of the year, these quantities of rainfall have been expressed in percentages of the average over the ten years. In addition, in order to reduce the volume of the sample thus obtained, we have considered only the amounts collected during the months of January, March, May, July, September and November (I, III, V, VII, IX, XI).

The total ordered sample of 60 elements obtained in this way is given in Table 4, where the amounts of precipitation collected x are shown opposite the rank m assigned to them as a result of their arrangement in increasing order.

Table 5 restores the chronological order of these values during the ten consecutive years for each of the six months considered, whilst the ranks r opposite the observed values x are those given by the total ordered series, with the exception of equal values where r is the average of the values of m given to these observations.

TABLE 4. – Precipitation at Brussels-Uccle during ten consecutive years. Ordered sample of the monthly totals x in % of the average over the ten years and the associated ranks m

m	x	m	x	m	x
1	15	21	81	41	114
2	22	22	83	42	117
3	24	23	84	43	124
4	32	24	85	44	131
5	33	25	91	45	132
6	37	26	92	46	135
7	41	27	92	47	136
8	42	28	95	48	140
9	45	29	97	49	142
10	51	30	97	50	143
11	55	31	99	51	145
12	55	32	99	52	146
13	56	33	104	53	156
14	58	34	104	54	162
15	62	35	109	55	172
16	62	36	109	56	173
17	67	37	113	57	187
18	69	38	113	58	190
19	75	39	113	59	197
20	77	40	114	60	204

TABLE 5. – Precipitation at Brussels-Uccle during ten consecutive years. Monthly totals x in % of the average over the ten years for January (I), March (III), May (V), July (VII), September (IX) and November (XI). Ranks r, modified for equal values, of the observations in the total ordered series

Year	I		III		V		VII		IX		XI	
	x	r	x	r	x	r	x	r	x	r	x	r
1	114	40,5	99	31,5	32	4	15	1	37	6	83	22
2	140	48	131	44	22	2	135	46	146	52	85	24
3	77	20	109	35,5	197	59	97	29,5	92	26,5	58	14
4	62	15,5	104	33,5	132	45	95	28	109	35,5	51	10
5	187	57	99	31,5	113	38	113	38	143	50	136	47
6	92	26,5	97	29,5	113	38	162	54	75	19	56	13
7	104	33,5	142	49	55	11,5	81	21	156	53	67	17
8	114	40,5	91	25	124	43	45	9	33	5	173	56
9	41	7	42	8	24	3	55	11,5	62	15,5	117	42
10	69	18	84	23	190	58	204	60	145	51	172	55
R_j		306,5		310,5		301,5		298		313,5		300

2.1.7.1 Homogeneity of the means of the six series of observations

As shown in 2.1.5, the homogeneity test can be used first of all to verify the homogeneity of the means.

In this case, the values of R_j obtained from the ranks r of table 5 are calculated by means of formula 2.1.5(1). As the checking equation 2.1.5(4) which, for $n = 60$, here becomes

$$\Sigma R_j = 306{,}5 + 310{,}5 + 301{,}5 + 298 + 313{,}5 + 300 = 1830 = 60 \times 61/2 \qquad (1)$$

is statisfied, we can conclude that this is a correct calculation of the R_j.

Equation 2.1.5(2) then gives, with $n_j = 10$ and $n = 60$,

$$X = 0{,}063 \tag{2}$$

If we observe that taking out groups of equal values leads to the following table (r_e = ranks of equal values, t_e = number of equal values of the same group) :

r_e	11,5	15,5	26,5	29,5	31,5	33,5	35,5	38	40,5
t_e	2	2	2	2	2	2	2	3	2

in 2.1.5(3), the quantity ΣT becomes

$$\Sigma T = (2^3 - 2) \times 8 + (3^3 - 3) = 72 \tag{3}$$

that is to say, with $n = 60$,

$$1 - \Sigma T / n(n+1)(n-1) = 0{,}999667 \tag{4}$$

a value which is very close to 1. This is why the correction for equal values by means of division by the quantity (4) scarcely modifies the value (2) found for the statistic X.

As the number of series is equal to 6, the critical value with which the final value of X must be compared is that of the χ^2 variate with $6 - 1 = 5$ degrees of freedom. As, at the 0,05 level, this value is 11,07, we find that the value is very much smaller than the critical value, which allows us to accept the homogeneity of the means.

It must be stressed, however, that the agreement is too good and the exceedance probability, greater than 0,995, is too high for the value obtained to be attributable to chance.

This too good agreement must be ascribed to the fact that all the quantities of rainfall were expressed as percentages of the ten year average, which arbitrarily assigns the value 100 to all the averages.

2.1.7.2 Homogeneity of the variances of the six series of observations

To verify the homogeneity of the variances, Table 6 has been set up from the absolute values of the differences from the mean, in the same way as Table 5. In other words, for each value of x, the absolute value $d = |x - 100|$ has been calculated and the ranks r of Table 6 have been deduced from the total ordered series of values obtained for d.

After verifying that here also we have

$$\Sigma R_j = 1830 \tag{1}$$

TABLE 6. — Precipitation at Brussels-Uccle during ten consecutive years. Absolute value d of the difference of the monthly total from the mean for ten years. Ranks r, modified for the equal values, of the differences in the total ordered series

Year	I		III		V		VII		IX		XI	
	d	*r*	*d*	*r*	*d*	*r*	*d*	*r*	*d*	*r*	*d*	*r*
1	14	16,5	1	1,5	68	51	85	56	63	49	17	20,5
2	40	34	31	26,5	78	55	35	30	46	42	15	18
3	23	23	9	11	97	59	3	3,5	8	8,5	42	35,5
4	38	32,5	4	5,5	32	28	5	7	9	11	49	43
5	87	57	1	1,5	13	14	13	14	43	37	36	31
6	8	8,5	3	3,5	13	14	62	48	25	25	44	38
7	4	5,5	42	35,5	45	40	19	22	56	45	33	29
8	14	16,5	9	11	24	24	55	44	67	50	73	53
9	59	47	58	46	76	54	45	40	38	32,5	17	20,5
10	31	26,5	16	19	90	58	104	60	45	40	72	52
R_j		267		161		397		324,5		340		340,5

with $n_j = 10$, $n = 60$, the relation 2.1.5(2) gives this time

$$X = 10{,}987 \tag{2}$$

Furthermore, knowing that the table of equal values is as follows :

r_e	1,5	3,5	5,5	8,5	11	14	16,5	20,5	26,5	32,5	35,5	40
t_e	2	2	2	2	3	3	2	2	2	2	2	3

it follows that for the correction divider 2.1.5(3)

$$\Sigma T = (2^3 - 2) \times 9 + (3^3 - 3) \times 3 = 126 \tag{3}$$

which gives

$$1 - \Sigma T/n(n+1)(n-1) = 0{,}999417 \tag{4}$$

The corrected value of X thus becomes

$$X = 10{,}993 \tag{5}$$

which is a little less than the critical value 11,07 at the 0,05 level of the χ^2 variate with five degrees of freedom. Hence the hypothesis of homogeneity of variance can also be accepted.

Although the combination of the two tests has no meaning in this case since the first hypothesis, the homogeneity of the means, is a certainty because of the definition of the variable x, we should nevertheless note that with $\alpha_1 = 0{,}9995$, $\alpha_2 = 0{,}05$, the Fisher test (see 1.3.1) gives

$$-2 \log \alpha_1 - 2 \log \alpha_2 = 0{,}001 + 5{,}991 = 5{,}992 \tag{6}$$

a value which is less than 9,49, the critical value at the 0,05 level of the χ^2 variate with four degrees of freedom.

In the same way, the test based on the binomial (see 1.3.2) for $k = 2$ at the 0,05 level gives 0,025 and 0,22 as the lower limits for the smallest and largest of the two values α_1 and α_2, limits which are not exceeded.

Therefore both tests allow us to conclude that the null hypotheses may be accepted.

2.1.7.3 Calculation of the confidence intervals of the probabilities associated with the order statistics

The limits F_1 and F_2 of the two-sided confidence intervals of the probabilities associated with the order statistics are given in Table 7.

As shown in 1.3.2, their calculation has been carried out using the equation

$$F = \nu_1 v^2/(\nu_2 + \nu_1 v^2) \tag{1}$$

where $\nu_1 = 2m$, $\nu_2 = 2(61 - m)$ and where v^2 is the quantile of the variance ratio with ν_1 and ν_2 degrees of freedom, of order $P = 0{,}025$ for F_1 and order $P = 0{,}975$ for F_2. In this connection, note that $v^2(\nu_1, \nu_2)$ for $P = 0{,}025$ is the inverse $1/v^2(\nu_2, \nu_1)$ of $v^2(\nu_2, \nu_1)$ for $P = 0{,}975$. Furthermore, the calculation has been done with equation (1) for $m(1,30)$ whereas for $m(31,60)$ account has been taken of the equalities

$$F_{1,m} = 1 - F_{2,(61-m)} \quad \text{and} \quad F_{2,m} = 1 - F_{1,(61-m)} \tag{2}$$

In addition, by using 2.1.2(6), the standard deviation has been calculated :

$$\sigma_{F_m} = [\hat{F}_m(1 - \hat{F}_m)/62]^{1/2} \quad \text{with} \quad \hat{F}_m = m/61 \tag{3}$$

from which the amplitude of the two-sided confidence interval at the 0,95 level has been derived, approximated by the normal distribution

$$d_0 = 2 \times 1{,}96\, \sigma_{F_m} \tag{4}$$

where 1,96 is the quantile of the standard normal variate u for which we have :

$$P(|u| < 1{,}96) = 0{,}95 \tag{5}$$

TABLE 7. – Limits $F_1 = F_{0,025}$ and $F_2 = F_{0,975}$ of the two-sided confidence interval at the 0,95 level of the probabilities associated with the order statistics of rank m for a sample size $n = 60$. Quantiles of the variance ratio v^2 for P = 0,975 with the degrees of freedom ν_1 and ν_2 or ν_2 and ν_1 used in the calculation of these limits. Amplitude d_0 of the two-sided confidence interval approximated by the normal distribution

m	ν_1	ν_2	$1/v^2(\nu_2,\nu_1)$	$F_1 = F_{0,025}$	$v^2(\nu_1,\nu_2)$	$F_2 = F_{0,975}$	d_0	$F_2 = F_{0,975}$	$F_1 = F_{0,025}$	m
1	2	120	1/39,5	0,0004	3,81	0,060	0,063	0,9996	0,940	60
2	4	118	8,31	0,0041	2,90	0,090	0,089	0,9959	0,910	59
3	6	116	4,91	0,010	2,52	0,115	0,108	0,990	0,885	58
4	8	114	3,73	0,018	2,31	0,139	0,123	0,982	0,861	57
5	10	112	3,15	0,028	2,16	0,162	0,137	0,972	0,838	56
6	12	110	2,80	0,038	2,07	0,184	0,148	0,962	0,816	55
7	14	108	2,56	0,048	1,99	0,205	0,159	0,952	0,795	54
8	16	106	2,40	0,059	1,93	0,226	0,168	0,941	0,774	53
9	18	104	2,27	0,071	1,88	0,246	0,177	0,929	0,754	52
10	20	102	2,17	0,083	1,85	0,266	0,184	0,917	0,734	51
11	22	100	2,09	0,095	1,81	0,285	0,191	0,905	0,715	50
12	24	98	2,02	0,108	1,78	0,304	0,198	0,892	0,696	49
13	26	96	1,97	0,121	1,76	0,323	0,204	0,879	0,677	48
14	28	94	1,93	0,134	1,75	0,343	0,209	0,866	0,657	47
15	30	92	1,89	0,147	1,73	0,361	0,214	0,853	0,639	46
16	32	90	1,86	0,160	1,72	0,379	0,219	0,840	0,621	45
17	34	88	1,83	0,174	1,71	0,398	0,223	0,826	0,602	44
18	36	86	1,80	0,189	1,70	0,416	0,227	0,811	0,584	43
19	38	84	1,78	0,203	1,69	0,433	0,231	0,797	0,567	42
20	40	82	1,76	0,217	1,68	0,450	0,234	0,783	0,550	41
21	42	80	1,74	0,232	1,67	0,467	0,237	0,768	0,533	40
22	44	78	1,73	0,246	1,67	0,485	0,239	0,754	0,515	39
23	46	76	1,72	0,260	1,66	0,501	0,241	0,740	0,499	38
24	48	74	1,70	0,276	1,66	0,518	0,243	0,724	0,482	37
25	50	72	1,70	0,290	1,65	0,534	0,245	0,710	0,466	36
26	52	70	1,69	0,305	1,65	0,551	0,246	0,695	0,449	35
27	54	68	1,68	0,321	1,65	0,567	0,247	0,679	0,433	34
28	56	66	1,68	0,336	1,66	0,585	0,248	0,664	0,415	33
29	58	64	1,67	0,352	1,66	0,601	0,249	0,649	0,398	32
30	60	62	1,67	0,367	1,66	0,616	0,249	0,633	0,384	31

Thus it is observed that the equation

$$d_0 = F_2 - F_1 \tag{6}$$

is relatively well statisfied even for the extreme values of m. Moreover, it is seen that d_0 decreases in a ratio of about 1/4 from the central values of m to the extreme values.

Otherwise, it is clear that the values of F_1 give the lower limits of the one-sided confidence intervals at the 0,975 level and that the values of F_2 give the upper limits of the one-sided confidence interval at the same level.

2.1.7.4 Estimation of the value of x for a given probability F_0. Estimation of the probability F for a given value x_0. Associated two-sided confidence intervals

By way of an example the probability 1/11 has been chosen. This is the empirical estimation of the probability F associated with the smallest value in each of the monthly series of ten observations. From the equation

$$F_0 = \frac{m_0}{61} = \frac{1}{11} = 0{,}091 \tag{1}$$

we get

$$m_0 = \frac{61}{11} = 5,55 \tag{2}$$

Therefore, for applying the linear interpolation method, we have

$$p = 0,55 \quad \text{and} \quad q = 0,45 \tag{3}$$

which, with 2.1.2(11) and the data in Table 4, leads to

$$\hat{x}(F_0) = x'_{5,55} = qx'_5 + px'_6 = 0,45 \times 33 + 0,55 \times 37 = 35,2 \tag{4}$$

To determine the two-sided confidence interval for $x(F_0)$, one should determine by means of Table 7 the value x'_0 of x'_m where $F_1 = F_0$ as well as the value x''_0 of x'_m for which $F_2 = F_0$. As in both cases F_0 is placed between two consecutive values, one should again proceed by linear interpolation.

In this way it is found that

$$x'_0 = qx'_{10} + px'_{11} \quad \text{with} \quad p = 0,67 \quad \text{and} \quad p = 0,33 \tag{5}$$

that is to say (Table 4), with $x'_{10} = 51$ and $x'_{11} = 55$: $x'_0 = 53,7$, whereas for x''_0 we obtain

$$x''_0 = qx'_2 + px'_3 \quad \text{with} \quad p = 0,04 \quad \text{and} \quad q = 0,96 \tag{6}$$

from which, with $x'_2 = 22$ and $x'_3 = 24$, $x''_0 = 22,1$.

It follows that the two-sided confidence interval at the 0,95 level for $x(F_0)$ is the interval

$$(22,1 \ , \ 53,7) \tag{7}$$

It is clear that, for the inverse problem, the same example can be taken. In this case, the value $x_0 = 35,2$ is considered as given and, as the values of the coefficients p and q in (3) remain the same, this then leads to

$$\hat{F}(x_0) = q\hat{F}_5 + p\hat{F}_6 = 0,091 \tag{8}$$

In addition, applying equation (8) to the limits of the confidence intervals obtained for x'_5 and x'_6 gives

$$F_1(x_0) = 0,034 \quad \text{and} \quad F_2(x_0) = 0,174 \tag{9}$$

2.1.7.5 Estimation of the probabilities in the case of equal values

The existence in Table 4 of equal values of x for different values of m can be a source of ambiguity in estimating the probabilities when the given value x_0 is identical with one of the repeated values of this table.

However, if we consider this as a simple case of linear interpolation, it is clear that the probabilities which should be associated with such a value are, for an empirical estimation, the one which comes from the average of the corresponding ranks in Table 4, and for the limits of the confidence interval, the averages of the limits attributed to the same ranks in Table 7.

In particular, for $x = 55$, in Table 4, we have $m = 11$ and 12, from which

$$\hat{F}(55) = 11,5/61 = 0,189 \tag{1}$$

whilst from Table 7 we obtain

$$F_1(55) = (0,095 + 0,108)/2 = 0,102 \quad \text{and} \quad F_2 = (0,285 + 0,304)/2 = 0,294 \tag{2}$$

2.1.7.6 Two-sided Kolmogorov limits of the true distribution function

Table 8 gives the extreme limits at the 0,05 level between which the value of the distribution function is found for each order statistic.

These limits have been calculated knowing that for the 0,95 level and $n=60$, the quantity D_α is equal to $1{,}36/\sqrt{60}=0{,}176$. If we then put

$$F_{1K}(m)=\frac{m}{60}-0{,}176 \quad \text{and} \quad F_{2K}(m)=0{,}176+\frac{m-1}{60} \tag{2}$$

for $m=1$ to 30, the values shown in Table 8 are found. Moreover, for $m=31$ to 60, the following relations were used

$$F_{2K}(m)=1-F_{1K}(61-m) \quad \text{and} \quad F_{1K}(m)=1-F_{2K}(61-m) \tag{3}$$

It follows from relations 2.1.3(4) and (6) that the maximum region inside which the true distribution function is found at the 0,95 level is defined by the set of inequalities :

$$\begin{aligned}
&F(x) < F_{2K}(1) && \text{for} && x \leqslant x'_1 \\
&F(x) < F_{2K}(m) && \text{for} && x'_{m-1} < x \leqslant x'_m && m=2,3,\cdots,60 \\
&F(x) > F_{1K}(m) && \text{for} && x'_{m+1} > x \geqslant x'_m && m=1,2,\cdots,59 \\
&F(x) > F_{1K}(60) && \text{for} && x \geqslant x'_{60}
\end{aligned} \tag{4}$$

TABLE 8. – Two-sided Kolmogorov limits F_{1K} and F_{2K} at the 0,95 level

m	F_{1K}	F_{2K}	F_{2K}	F_{1K}	m
1	−0,159	0,176	1,159	0,824	60
2	−0,143	0,193	1,143	0,807	59
3	−0,126	0,209	1,126	0,791	58
4	−0,109	0,226	1,109	0,774	57
5	−0,093	0,243	1,093	0,757	56
6	−0,076	0,259	1,076	0,741	55
7	−0,059	0,276	1,059	0,724	54
8	−0,043	0,293	1,043	0,707	53
9	−0,026	0,309	1,026	0,691	52
10	−0,009	0,326	1,009	0,674	51
11	0,007	0,343	0,993	0,657	50
12	0,024	0,359	0,976	0,641	49
13	0,041	0,376	0,959	0,624	48
14	0,057	0,393	0,943	0,607	47
15	0,074	0,409	0,926	0,591	46
16	0,091	0,426	0,909	0,574	45
17	0,107	0,443	0,893	0,557	44
18	0,124	0,459	0,876	0,541	43
19	0,141	0,476	0,859	0,524	42
20	0,157	0,493	0,843	0,507	41
21	0,174	0,509	0,826	0,491	40
22	0,191	0,526	0,809	0,474	39
23	0,207	0,543	0,793	0,457	38
24	0,224	0,559	0,776	0,441	37
25	0,241	0,576	0,759	0,424	36
26	0,257	0,593	0,743	0,407	35
27	0,274	0,609	0,726	0,391	34
28	0,291	0,626	0,709	0,374	33
29	0,307	0,643	0,693	0,357	32
30	0,324	0,659	0,676	0,341	31

Note moreover, that the amplitude of the Kolmogorov intervals is constant for all the values of m and that here it is equal to

$$F_{2K} - F_{1K} = 2D_\alpha - \frac{1}{60} = 0{,}335 \tag{5}$$

Obviously, in Table 8, the values of $F_{1K} < 0$ should be replaced by zero and the values of $F_{2K} > 1$ should be set to 1.

2.1.7.7 Calculation of the empirical mean and variance. Confidence interval of the mean

Applying formulae 2.1.4(3) to the data in Table 4 gives :

$$\hat{\mu} = \Sigma x_i/n = 5998/60 = 99{,}97 \tag{1}$$

$$\hat{\text{var}}\, x = \Sigma(x - \hat{\mu})^2/(n-1) = \left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right]/(n-1)$$

$$= \left[728470 - \frac{(5998)^2}{60}\right]/59 = 2184{,}24 \tag{2}$$

which are the empirical mean and variance respectively of the total sample.

From these we deduce

$$s = \sqrt{\hat{\text{var}}\, x} = \sqrt{2184{,}24} = 46{,}74 \tag{3}$$

Knowing, then, that for $\nu = 59$ degrees of freedom, 2,001 is the quantile of the Student t variate, for which we have

$$P(|t| < 2{,}001) = 0{,}95 \tag{4}$$

the two-sided confidence interval for the true mean μ at the 0,95 level with 2.1.4(7) becomes

$$|\hat{\mu} - \mu| = |99{,}97 - \mu| < 2{,}001\, s/\sqrt{n} = 2{,}001 \times 46{,}74/\sqrt{60} \tag{5}$$

in other words

$$|99{,}97 - \mu| < 12{,}07 \tag{6}$$

or even

$$87{,}90 < \mu < 112{,}04 \tag{7}$$

which sets the accuracy of the empirical estimation of the mean.

The same method can be used, for example, to make an estimation of the mean of the ten year minimum from the data in Table 5, that is, the mean of the order statistic x' corresponding to $\hat{F} = 1/11$ in the series of ten observations.

Since the minimum values of x for each of the six months are respectively

$$x'_1 = 41, 42, 22, 15, 33 \quad \text{and} \quad 51 \tag{8}$$

we obtain from then :

$$\hat{\text{var}}\, x'_1 = \Sigma[x'_1 - \hat{E}(x'_1)]^2/5 = \left[7844 - \frac{(204)^2}{6}\right]/5 = 181{,}6 \tag{9}$$

which in turn gives

$$s = \sqrt{\hat{\text{var}}\, x'_1} = 13{,}5 \tag{10}$$

Furthermore, seeing that, for $\nu = 5$ degrees of freedom, 2,571 is the quantile of the Student t variate for which we have

$$P(|t| < 2{,}571) = 0{,}95 \tag{11}$$

the two-sided confidence interval of the mean $E(x'_1)$ of the ten year minima at the 0,95 level is written

$$|34 - E(x'_1)| < 2{,}571 \times 13{,}5/\sqrt{6} = 14{,}2 \tag{12}$$

from which we get

$$19{,}8 < E(x'_1) < 48{,}2 \tag{13}$$

Although $E(x'_1)$ cannot be made identical with $x(F)$ for $F = 1/11$, however, both may be considered as central values of the probability distribution for the statistic x'_{m_0} of the total series with $m_0 = 5{,}55$, that is $\hat{F} = 1/11$.

It is consequently interesting to compare the interval (13) with the interval 2.1.7.4(7) to show that the two estimates have comparable accuracies.

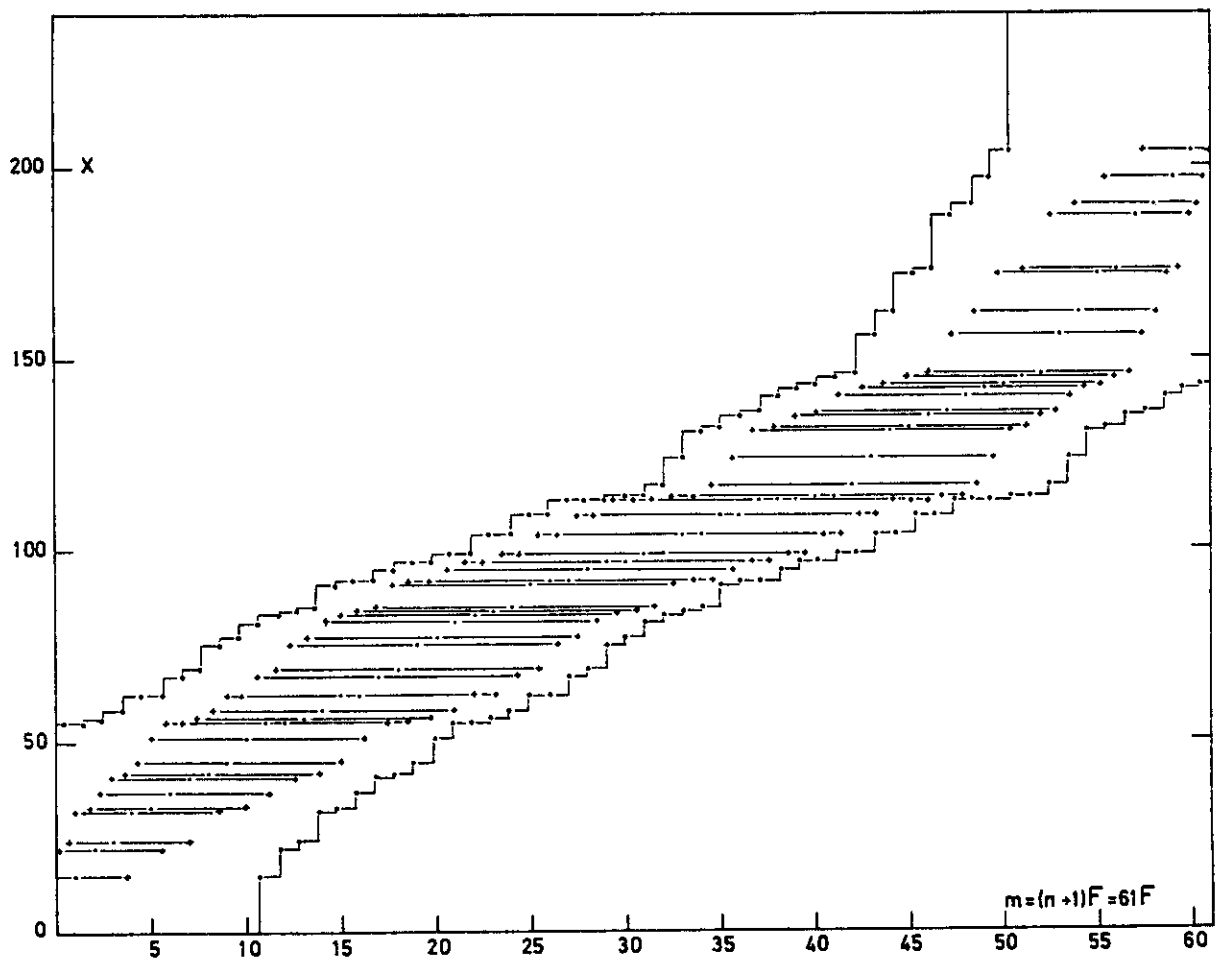

FIG. 3. — Precipitation at Brussels-Uccle during ten consecutive years. Ordered series of the monthly totals for the odd months (January to November) in % of the average over ten years for each month. Confidence intervals at 0,95 of the probabilities associated with the order statistics. Kolmogorov limits at 0,95 of the true distribution function

2.1.7.8 Graphical representation of the results

The results obtained have been given in Figure 3. For this purpose, two coordinate axes with linear scales have been used, the horizontal axis for the probabilities, the vertical axis for the values of x. In addition, to aid the construction of the graph, the scale for the probabilities has been graduated as a function of m. The value F of the distribution function is therefore related to m by the relationship

$$m = (n + 1)F \tag{1}$$

The graphical representation includes :

(1) The representation of the values x'_m in terms of m from the data in Table 4;
(2) The representation of the confidence intervals for F_m at the 0,95 level given in Table 7;
(3) The representation of the Kolmogorov limits, taken from Table 8 and the equations 2.1.7.6(4).

In the case of equal values in the ordered series, the corresponding confidence intervals are placed on the same horizontal line. Because of the foregoing, the interval adopted for each of these values takes the extremities as the mean position of the extremities of the various corresponding intervals.

For the Kolmogorov limits, on the other hand, there is no need to introduce any correction for equal values.

Thus it is confirmed that for extreme values, the Kolmogorov limits are farther away from the extremities of the confidence intervals than for the central values.

2.1.8 Case of discrete distributions

When, because of their definition, the observations are arranged in two or more distinct groups $g_1, g_2, \ldots, g_k$, such that the number of cases in each group is respectively $n_1, n_2, \ldots, n_k$ with $\Sigma n_i = n$, the distribution of the observations is a *discrete distribution*.

The estimation problem to be solved is then that of determining the best values to give to the probabilities $p_1, p_2, \ldots, p_k$ which govern the frequencies in each group and for which we have :

$$\Sigma p_i = 1 \tag{1}$$

It is clear that for the statistical treatment of these observations, there is no drawback in giving the values 1, 2, ..., k to the observations according to whether they belong to the group $g_1, g_2, \ldots$ or g_k. It follows that if these observations appear as a time series, this operation will allow the methods described in Chapter 1 to be applied in order to verify the simple random character of the series.

Under these conditions, the discrete distribution considered is a *multinomial distribution* for which the estimation of any one of the probabilities p_i or any of the sums $p_{i_1} + p_{i_2} + \ldots$ of these probabilities is reduced to estimating the probability in the case of the binomial distribution.

2.1.9 The binomial distribution

In this case, $k = 2$ and by putting $p_1 = p$ and $p_2 = q$, we have $p + q = 1$. Furthermore, we also have

$$\mathrm{E}(n_1) = np \quad \text{and} \quad \mathrm{var}\, n_1 = npq \tag{1}$$

from which

$$\mathrm{E}\left(\frac{n_1}{n}\right) = p \quad \text{and} \quad \mathrm{var}\, \frac{n_1}{n} = \frac{pq}{n} \tag{2}$$

If we adopt as the estimate for p

$$\hat{p} = \frac{n_1}{n} \tag{3}$$

we get, from (2)

$$E(\hat{p}) = p \tag{4}$$

that is to say, the estimation in *unbiased*.

In addition, by making $F = p$ in the expression ϕ_m of 2.1.1(5), we have

$$P(n_1 \geqslant m) = \phi_m(p) \tag{5}$$

and, in particular

$$P(n_1 \geqslant 0) = \phi_0(p) = (p+q)^n = 1 \tag{6}$$

whilst for n sufficiently large and n_1 sufficiently close to $n/2$, n_1 has an asymptotically normal distribution.

So knowing p, one or other of these two properties can be used to construct a confidence interval for the frequency n_1.

Conversely, given n_1, the confidence interval for p will group together all the values of p leading to a confidence interval of the frequency containing the observed frequency n_1.

2.1.9.1 Confidence interval for a frequency

From 2.1.9(5), for a given p, $m_1 < m_2$ means that $\phi_{m_1} > \phi_{m_2}$, whereas we have

$$P(m_1 < n_1 < m_2) = \phi_{m_{1+1}} - \phi_{m_2} \tag{1}$$

It follows that if m_1 and m_2 are chosen in such a way that :

$$\Phi_{m_1+1} \geqslant 1 - \frac{\alpha}{2} > \Phi_{m_1+2}$$

$$\Phi_{m_2-1} > \frac{\alpha}{2} \geqslant \Phi_{m_2} \tag{2}$$

from (1), the relationship

$$m_1 < n_1 < m_2 \tag{3}$$

will be satisfied with a probability $\geqslant 1 - \alpha$.

The interval (3) is therefore a two-sided confidence interval of frequency n_1 with a level $\geqslant 1 - \alpha$ In addition, as the conditions (2) ensure the uniqueness of the values m_1 and m_2, we agreee to define it as a *two-sided confidence interval of the frequency n_1 at the level* $(1 - \alpha)$.

To calculate the limits m_1 and m_2, we note that for $\alpha_0 = \phi_m(p)$, from 1.3.2(2) and (3), we get

$$v^2_{\alpha_0} = \frac{p\nu_2}{q\nu_1} \tag{4}$$

where $\nu_1 = 2m$ and $\nu_2 = 2(n - m + 1)$.

The value of m_1 is then that for which $m = m_1 + 1$ leads to $v^2_{\alpha_0} = v^2_{(1-\alpha/2)}$ or to the smallest value of $v^2_{\alpha_0} > v^2_{(1-\alpha/2)}$. Similarly, the value of m_2 is that for which $m = m_2$ gives $v^2_{\alpha_0} = v^2_{\alpha/2}$ or the largest value of $v^2_{\alpha_0} < v^2_{\alpha/2}$.

For the latter calculation, however, the equation used is

$$v^2_{\alpha_0}(\nu_1, \nu_2) = 1/v^2_{1-\alpha_0}(\nu_2, \nu_1) \tag{5}$$

The number of trials used in this calculation can moreover be reduced by starting out from the values of m_1 and m_2 given by the normal approximation. In this case, if $u_0 > 0$ is the value of the standard normal variate u for which we have

$$P(|u| < u_0) = 1 - \alpha \tag{6}$$

because of 2.1.9(1), the approximate confidence limits are

$$m'_1 = np - u_0\sqrt{npq} \quad \text{and} \quad m'_2 = np + u_0\sqrt{npq} \tag{7}$$

By rounding the values of m'_1 and m'_2 to the nearest integer, the best starting values for the exact determination of m_1 and m_2 are found.

2.1.9.2 **Example 5.** *Calculation of the two-sided confidence interval at the 0,95 level of the number of cases in 100 years of an event whose probability is 1/10*

In this case $n = 100$, $np = 10$, $\sqrt{npq} = 3$, $\alpha/2 = 0{,}025$ and $u_0 = 1{,}96$. From these values, using 2.1.9.1(7) we get :

$$m'_1 = 10 - 1{,}96 \times 3 = 10 - 5{,}88 = 4{,}12 \qquad m'_2 = 15{,}88 \tag{1}$$

The starting values for determining m_1 and m_2 are then 4 and 16.

The calculation of $v^2_{\alpha_0}$ and of $1/v^2_{\alpha_0}$ by means of equation 2.1.9.1(4) gives :

m_1	$m_1 + 1$	ν_1	ν_2	$v^2_{\alpha_0}$	$v^2_{0,975}(\nu_1,\nu_2)$
4	5	10	192	2,13	2,11
5	6	12	190	1,76	2,01

m_2	ν_1	ν_2	$v^2_{\alpha_0}$	$1/v^2_{\alpha_0}$	$v^2_{0,975}(\nu_2,\nu_1)$
16	32	170	0,59	1,69	1,82
17	34	168	0,55	1,82	1,79

From comparing the values found for $v^2_{\alpha_0}$, it seems that the largest value of m_1 for which $\alpha_0 = \Phi_{m_1+1}$ gives $v^2_{\alpha_0} > v^2_{0,975}$ is $m_1 = 4$.

Similarly $m_2 = 17$ is the smallest value for which $\alpha_0 = \Phi_{m_2}$ gives $v^2_{\alpha_0} < v^2_{0,025}(\nu_1, \nu_2) = 1/v^2_{0,975}(\nu_2, \nu_1)$.

The confidence interval sought is then

$$4 < n_1 < 17 \tag{2}$$

2.1.9.3 Confidence interval of a probability

The two-sided confidence interval at the level $(1-\alpha)$ of the probability p associated with the observed frequency n_1 is the set of values of p leading at the level $(1-\alpha)$ to a two-sided confidence interval of the frequency including the observed frequency n_1.

To determine this, we see that if p_1 is taken from the equality

$$\Phi_{n_1+1}(p_1)=1-\frac{\alpha}{2} \tag{1}$$

for any value of $p<p_1$, we will have

$$\Phi_{n_1+1}(p)<1-\frac{\alpha}{2} \tag{2}$$

So using relations 2.1.9.1(2), n_1 will be greater than the limit m_1 of the confidence interval of the frequency.

Similarly, by taking p_2 from the equation

$$\Phi_{n_1}(p_2)=\frac{\alpha}{2} \tag{3}$$

for any value of $p>p_2$, we will have

$$\Phi_{n_1+1}(p)>\frac{\alpha}{2} \tag{4}$$

which means that, using 2.1.9.1(2), n_1 will be lower than the upper limit m_2 of the confidence interval.

The interval

$$p_1>p>p_2 \tag{5}$$

is therefore clearly the two-sided confidence interval we are looking for.

Tables exist which give the values of p_1 and p_2 directly for given values of n_1 and n. However, if we wish to calculate p_1 and p_2, the form of the equation (1) and (3) allows the method indicated in 1.3.2 to be used.

This calculation can then be simplified by basing it on the normal approximation to the distribution of the frequency. In this case, equations (1) and (3) are replaced by equations

$$np_1'-u_0\sqrt{np_1'(1-p_1')}=n_1 \qquad \text{and} \qquad np_2'+u_0\sqrt{np_2'(1-p_2')}=n_1 \tag{6}$$

where u_0 is the quantile of the normal distribution defined in 2.1.9.1(6), equations whose solutions are the roots of the equation

$$(np-n_1)^2=u_0^2\,np(1-p) \tag{7}$$

that is to say that

$$p_1', p_2'=\frac{2n_1+u_0^2\pm u_0\sqrt{u_0^2+4n_1(n-n_1)/n}}{2(n+u_0^2)} \tag{8}$$

However, it is advisable to stress here that as long as the normal approximation is sufficiently accurate, the probability associated with the confidence interval (p_1', p_2') is exactly $(1-\alpha)$. The result is that the confidence intervals

of the frequencies that are made to correspond to the values of p in this interval no longer contain n_1 with a probability $\geqslant 1-\alpha$, but only with an average probability equal to $(1-\alpha)$.

This happens, for example, if m_1 or m_1+1 are adopted as the lower limit of the interval depending on whether Φ_{m_1+1} or Φ_{m_1+2} is best approximated by $1-\alpha/2$, and whether the same is done with the upper limit. It also follows that the interval (p'_1, p'_2) is narrower than the interval (p_1, p_2).

Equivalently, the limits p''_1 and p''_2 of p obtained from equations

$$\frac{1}{2}[\Phi_{n_1}(p''_1)+\Phi_{n_1+1}(p''_1)]=1-\frac{\alpha}{2} \quad \text{and} \quad \frac{1}{2}[\Phi_{n_1}(p''_2)+\Phi_{n_1+1}(p''_2)]=\frac{\alpha}{2} \tag{9}$$

give the limits of a confidence interval for which the associated probability is *exactly* equal to the level $(1-\alpha)$.

These are obtained to a good approximation by taking the average of the solutions of the equations, firstly

$$\Phi_{n_1}(p)=1-\frac{\alpha}{2} \quad , \quad \Phi_{n_1+1}(p)=1-\frac{\alpha}{2}$$

and secondly

$$\Phi_{n_1}(p)=\frac{\alpha}{2} \quad , \quad \Phi_{n_1+1}(p)=\frac{\alpha}{2}$$

2.1.9.4 **Example 6.** *Calculation of the confidence interval of the probability associated with the frequency $n_1=10$ in a total number of cases $n=60$*

The choice of $n=60$ permits the results given in Table 7 to be used. With 2.1.9(3), this becomes

$$\hat{p}=10/60=0{,}167 \tag{1}$$

whereas the limits of the confidence interval at the 0,95 level are deduced from equations 2.1.9.3(1) and (3), that is to say

$$\Phi_{11}(p_1)=0{,}975 \quad \text{and} \quad \Phi_{10}(p_2)=0{,}025 \tag{2}$$

from which we obtain, using the results of Table 7 given for $m=11$ and 10,

$$p_1=0{,}285 \quad \text{and} \quad p_2=0{,}083 \tag{3}$$

Furthermore, the normal approximation leads, with 2.1.9.3(8) to

$$p'_1=0{,}280 \quad \text{and} \quad p'_2=0{,}093 \tag{4}$$

whereas the averages for $m=10$ and 11 of the values derived for p_1 and p_2 in Table 7 give

$$p''_1=(0{,}266+0{,}285)/2=0{,}276$$

and

$$p''_2=(0{,}083+0{,}095)/2=0{,}089 \tag{5}$$

2.1.9.5 The mean recurrence interval

To express the probability associated with a particular value of the variable x, it is fairly common to use the concept of the mean recurrence interval T, linked to the distribution function $F(x)$ by the relation

$$T=\frac{1}{F(x)} \quad \text{or} \quad T=\frac{1}{1-F(x)} \tag{1}$$

depending on whether $F(x)<0{,}5$ or $F(x)>0{,}5$.

In order to justify this idea, the case of a binomial distribution with probability p is considered again. In this case, if a series of consecutive independent trials is carried out, the probability f_i that an event associated with the probability p will occur only at trial i, that is to say, for its recurrence interval to be equal to i, is given by

$$f_i=q^{i-1}p \quad \text{with} \quad q=1-p \tag{2}$$

Since on the other hand we have

$$\sum_1^\infty f_i=p\sum_1^\infty q^{i-1}=p\,.\,\frac{1}{1-q}=1 \tag{3}$$

we can conclude from this that the probabilities f_i completely define the distribution of the recurrence intervals of the event considered.

In partitular, it follows that the mean of the interval i is deduced from the equation

$$T=E(i)=\sum_1^\infty if_i=p\sum_1^\infty iq^{i-1}=p\,.\,\frac{1}{(1-q)^2}=\frac{1}{p} \tag{4}$$

a relation which justifies relation (1) for $F(x)<0{,}5$.

Moreover, the choice of $T=1/[1-F(x)]$ for $F(x)>0{,}5$ is logical since it gives the same mean recurrence interval to equally rare values of x, because $F(x)$ is either close to 0 or close to 1.

2.1.10 The asymptotic form of the multinomial distribution. The Pearson test

It is known that the term $\binom{n}{n_1}p^{n_1}q^{n-n_1}$ of the expansion of the binomial $(p+q)^n$ with $p+q=1$, gives the probability that an event of probability p occurs n_1 times in a series of n independent trials. Similarly, the term

$$f(n_1, n_2, \ldots, n_k)=\frac{n!}{n_1!\,n_2!\ldots n_k!}\,p_1^{n_1}p_2^{n_2}\ldots p_k^{n_k} \tag{1}$$

with

$$\sum_1^k n_i=n,$$

of the expansion of the polynomial $(p_1+p_2\ldots+p_k)^n$, gives the probability that, in a series of n independent trials, k mutually exclusive events of probabilities $p_1, p_2, \ldots, p_k$ with

$$\sum_1^k p_i=1,$$

occur $n_1, n_2, \ldots, n_k$ times respectively.

As n_k is fixed once the $(k-1)$ other values of n_i are determined, the probability that we have simultaneously $n_1 \geqslant n'_1, n_2 \geqslant n'_2, \ldots, n_{k-1} \geqslant n'_{k-1}$ is given by the sum

$$\Phi_{n'_1, n'_2, \ldots, n_{k-1}} = \sum_{n_1 \geqslant n'_1} \sum_{n_2 \geqslant n'_2} \ldots \sum_{n_{k-1} \geqslant n'_{k-1}} f(n_1, n_2, \ldots, n_k) \tag{2}$$

or again that the probability P that the frequencies $n_1, n_2, \ldots, n_{k-1}$ are simultaneously within the intervals $n'_1 < n''_1$, $n'_2 < n''_2, \ldots, n'_{k-1} < n''_{k-1}$ is caculated from the relation :

$$P = \Phi_{n'_1+1, n'_2+1, \ldots, n'_{k-1}+1} - \Phi_{n''_1, n''_2, \ldots, n''_{k-1}} \tag{3}$$

a relation which generalizes equation 2.1.9.1(1).

It is clear that relation (3) can be used either, when the probabilities p_i are known, to determine a confidence domain within which the frequencies n_i are found simultaneously with a given probability $P = 1 - \alpha$, or, when the frequencies n_i are given, to determine the whole series of values of probabilities p_i each leading to a confidence domain for the frequencies containing the frequencies n_i.

As manipulation of equation (3) is fairly difficult, it is preferable to solve these problems by using the asymptotic form of the multinomial distribution.

Returning to the case of the binomial, since the frequency n_1 has an asymptotically normal (Gaussian) distribution with mean np and variance npq, as a result, the expression

$$X = \frac{(n_1 - np)^2}{npq} \tag{4}$$

which can be put in the form

$$X = \frac{(n_1 - np)^2}{np} + \frac{(n_2 - nq)^2}{nq}, \quad \text{with} \quad n_1 + n_2 = n \tag{5}$$

has a χ^2 distribution with one degree of freedom.

Similarly, it can be shown that, in the case of a multinomial law, the frequencies $n_1, n_2, \ldots, n_k$ have a joint asympotically normal distribution from which it follows that the statistic

$$X = \sum_1^k \frac{(n_i - np_i)^2}{np_i} \tag{6}$$

called the *K. Pearson statistic*, has an asymptotic χ^2 distribution with $(k-1)$ degrees of freedom. This statistic can be used in place of expression (2) with a constraint similar to that in 2.1.9.3 regarding the intervals (p_1, p_2) and (p'_1, p'_2).

In particular, if we denote by χ^2_0 the quantile of the χ^2 variate with $(k-1)$ degrees of freedom for which we have

$$P(\chi^2 < \chi^2_0) = 1 - \alpha \tag{7}$$

by writing

$$\sum_1^k \frac{(n_i - np_i)^2}{np_i} = \chi^2_0 \tag{8}$$

when the probabilities p_i are known, a confidence region will be defined for the frequencies n_i at the $(1 - \alpha)$ level, whilst if the probabilities p_i are unknown, the introduction in (8) of a series of observed frequencies n_i will determine all of the values of the probabilities p_i which are acceptable for the same level.

In addition, in order to verify the compatibility of a series of observed frequencies n_i with a given series of values of probabilities p_i, the statistic (6) can be used to accept or reject the null hypothesis of compatibility depending on whether the statistic X is lower than or greater than the critical value χ_0^2.

It should be said that because of the asymptotic character of the distribution for this statistic X, it is generally advisable to use it only on condition that the quantities np_i are not too small. In reality, it can be shown that the behaviour of X for high values ($\alpha = 0{,}05$ or less) remains excellent when the quantities np_i fall to 3 or 2 or even, when there are sufficiently many groups, when some of these drop to one.

From this it is concluded that the range of application of the test remains very wide.

2.1.10.1 **Example 7.** *Probabilities of the sequences of consecutive dry years at Brussels-Uccle*

The rainfall totals collected at Brussels-Uccle from 1833 to 1960 have been expressed as percentages of the general average and a year is considered as a dry year when this quantity is less than 100%. The other years are rainy years. If it is assumed that the dry and wet years follow each other by chance and that the probability p of a wet year is equal to 0,5 , the probability p_i that it is followed by a sequence of i consecutive dry years is then :

$$p_i = p^i q = (0{,}5)^i \,.\, (0{,}5) = (0{,}5)^{i+1} \tag{1}$$

During the period mentioned, 68 wet years have been counted and if we denote by n_i the number of wet years which have been followed by a sequence of $i = 0, 1, 2, \ldots$, dry years, the 68 wet years are distributed as shown in Table 9. The mean values np_i of the frequencies n_i have also been given in this table; these values are calculated from (1) with $n = 68$. Finally, for $i = 5$, we have given the value of np_i for $i \geqslant 5$.

TABLE 9. – Distribution of sequences of dry years at Brussels-Uccle

i	n_i	p_i	np_i	$(n_i - np_i)^2/np_i$
0	37	0,50	34	0,265
1	17	0,25	17	0
2	7	0,125	8,5	0,265
3	4	0,0625	4,25	0,015
4	1	0,0312	2,125	0,596
5	1	0,0156	2,125	0,007
6	0	0,0078	—	—
$\geqslant 7$	1	0,0078	—	—

Calculation of the statistic X then leads to

$$X = 0{,}265 + 0 + 0{,}265 + 0{,}015 + 0{,}596 + 0{,}007 = 1{,}148 \tag{2}$$

With $(6 - 1) = 5$ degrees of freedom, the critical value of the variable χ^2 at the 0,05 level is equal to 11,07 , so there is excellent agreement between the observed frequencies n_i and the probabilities p_i.

Incidentally, it would be wrong to infer too good an agreement here between the series of observations and the theoretical distribution. Following the remark made in the previous section, since the value of X is relatively small, the distribution of the χ^2 variate does not give the exact value of the corresponding probability.

In other words, we actually have

$$P(X < 1{,}148) > P(\chi^2 < 1{,}148) \tag{3}$$

2.1.11 Reference works consulted

The empirical estimation based on the distribution of the order statistics has been used in [15]. It is inspired by concepts of intervals and of control curves introduced by Gumbel in [4]. The properties of the distributions of the order statistics and the graphical representation deriving from them have also been taken from [4], whereas the presentation of the Kolmogorov test comes from reference [11] and that of the Kruskal-Wallis test from [10].

The empirical estimation of a probability from a frequency is the transposition of the empirical estimation in the case of a continuous variable. It is based on the identity in relation 2.1.1(5), which gives the distribution of the order statistic x'_m, and relation 1.3.2 (2), which gives the probability associated with the observed frequency for a specified generating probability.

The definition of the mean recurrence interval is the symmetrical version of the one given by Gumbel in [4]. The comments on the appoximation of the multinomial distribution by χ^2 appear in [11].

For the remainder, reference is made to what has already been said in section 1.5.

2.2 PARAMETRIC ESTIMATION

The method of parametric estimation is founded on the hypothesis that the observations considered are distributed according to a probability distribution whose analytical form is known and completely specified by the value of one or more parameters.

If x denotes any observation whatsoever and x_0 a particular value, this amounts to assuming we can write

$$P(x < x_0) = F(x_0, \theta) \tag{1}$$

where the distribution function $F(x, \theta)$ is fully determined as soon as the value of the parameter θ (or of the vector θ with components $\theta_1, \theta_2, \ldots, \theta_k$) is fixed.

In the case of a single parameter θ, if θ_0 is its true value, the problem of estimating the probability associated with the observation x_0 boils down to choosing a suitable value $\hat{\theta}_0$ enabling us to put

$$\hat{F}(x_0, \theta_0) = F(x_0, \hat{\theta}_0) \tag{2}$$

As this choice is generally made by using a function $\hat{\theta}(x_1, x_2, \ldots, x_n)$ determined from the sample formed by the observations $x_1, x_2, \ldots, x_n$, a function called the *estimator*, we then have

$$\hat{\theta}_0 = \hat{\theta}(x_1, x_2, \ldots, x_n) \tag{3}$$

which, because of the random character of the sample of observed values of x, allows us to consider the estimator $\hat{\theta}(x_1, x_2, \ldots, x_n)$ as a random variable also.

If, furthermore, the true value θ_0 occupies a central position in the distribution of the estimator $\hat{\theta}(x_1, x_2, \ldots, x_n)$, from the differential, for $\theta = \theta_0$,

$$dF(x_0, \theta_0) = \left.\frac{\partial F(x_0, \theta)}{\partial \theta}\right|_{\theta=\theta_0} \cdot d\theta \tag{4}$$

we deduce that the variance can be expressed to a good approximation as

$$\operatorname{var} \hat{F}(x_0, \theta_0) = \left|\frac{\partial F(x_0, \theta)}{\partial \theta}\right|^2_{\theta=\theta_0} \cdot \operatorname{var} \hat{\theta} \tag{5}$$

which characterizes the error associated with the estimate $\hat{F}(x_0, \theta_0)$ obtained by (2) as a result of (3).

Inversely, if the quantile of order P of the variable x is required, the relevant equation is

$$F(x_P, \hat{\theta}_0) = P \tag{6}$$

and the variance of the error associated with the estimate $\hat{x}_P$ obtained in this way is determined, noting that since

$$dx_P = \left(\frac{\partial F}{\partial x}\right)^{-1} dF \tag{7}$$

we obtain under the same conditions

$$\operatorname{var} \hat{x}_P = \left(\frac{\partial F}{\partial x}\right)^{-2} . \operatorname{var} \hat{F}_0 \tag{8}$$

where $\operatorname{var} \hat{F}_0$ is given by (5) and where the derivatives are calculated at $x = x_P$ and $\theta = \theta_0$.

An important case is that where the parameters specifying the distribution are a location parameter μ and a scale parameter $\sigma > 0$ where the variable x observed is related to the reduced variable u by means of the linear relationship

$$x = \mu + \sigma u \tag{9}$$

If $F(u)$ is the distribution function of the reduced variable u, equation (1) then becomes

$$P(x < x_0) = F(u_0) \quad \text{with} \quad u_0 = (x_0 - \mu)/\sigma \tag{10}$$

In this case, if $\hat{\mu}$ and $\hat{\sigma}$ are estimates of the parameters μ and σ, functions of x can be estimated using the reduced variable u by means of (9). In particular, (2) becomes

$$\hat{P}(x < x_0) = F\left(\frac{x_0 - \hat{\mu}}{\hat{\sigma}}\right) = F(\hat{u}_0) \tag{11}$$

with

$$\operatorname{var} F(\hat{u}_0) = \left(\frac{\partial F}{\partial u}\right)^2 \operatorname{var} \hat{u}_0 \tag{12}$$

where

$$\operatorname{var} \hat{u}_0 = \frac{\operatorname{var} \hat{\mu} + 2u_0 \operatorname{cov}(\hat{\mu}, \hat{\sigma}) + u_0^2 \operatorname{var} \hat{\sigma}}{\sigma^2} \tag{13}$$

whilst the solution of equation (6) is directly given by the relation

$$\hat{x}_P = \hat{\mu} + \hat{\sigma} u_P \tag{14}$$

where u_P is derived from equation $F(u_P) = P$ and for which we have

$$\operatorname{var} \hat{x}_P = \operatorname{var} \hat{\mu} + 2u_P \operatorname{cov}(\hat{\mu}, \hat{\sigma}) + u_P^2 \operatorname{var} \hat{\sigma} \tag{15}$$

The foregoing concerns the case where the variable x is continuous. Its extension to discrete variables does not present any difficulties; however, it will be examined later.

Several remarks can be made about the parametric estimation that has just been derived.

Indeed, whereas the domain of the empirical estimation for the observed values remains limited to the interval extending between the extreme values x'_1 and x'_n and, for the associated probabilities, to the interval defined by the probabilities $1/(n+1)$ and $n/(n+1)$ given to these extreme values, parametric estimation makes estimation possible over the whole area of definition of the variable x and for any value between 0 and 1 of the distribution function of this variable.

In addition, in the domains where both types of estimation are applicable, as will be seen in the examples, parametric estimation generally leads to estimates at least as accurate. Because of relations (5) and (8), it is however obvious that the quality of the parametric estimation depends directly on the choice of the estimator of the parameter which, as has been stated, must be well centred (low bias) and have a small variance.

Parametric estimation thus appears as a more powerful method than empirical estimation. It must never, however, be forgotten that its use remains dependent on the *a priori* knowledge of the shape of the distribution function. In particular, when, in the absence of any information on this subject, the function fitted to the observations is adopted arbitrarily, it is advisable to ensure the compatibility of the fitted distribution with the observations. In any case, such verification is always useful.

Moreover, when choosing the function to be fitted, it should be remembered that choosing functions containing the largest possible number of parameters with a view to obtaining the best fit to the observed values is usually counterproductive.

In fact, from the generalization of equations (5) and (8), the introduction of any new parameter involves an increase in the variance of the error of estimation, and, as a result, the mediocrity of fit will increase proportionately with the number of parameters specifying the distribution.

2.2.1 Estimation by the method of moments. Fitting the normal distribution

When the one or more parameters specifying the distribution are moments of this same distribution, estimation by the method of moments consists of adopting as the fitted function that obtained by using the empirical values derived from the series of observations as parameter estimates.

In particular, if the distribution function is determined by the values of the mean and the variance, that is, if these parameters are the moments

$$\mu_1 = \int_0^1 x\,dF \quad \text{and} \quad \mu_2 = \int_0^1 (x-\mu_1)^2\,dF \tag{1}$$

the method amounts to using the estimators

$$\hat{\mu}_1 = \Sigma x_i/n \quad \text{and} \quad \hat{\mu}_2 = \Sigma (x_i - \hat{\mu}_1)^2/n \tag{2}$$

for which we have, when the observations are independent,

$$\text{var}\ \hat{\mu}_1 = \mu_2/n \tag{3}$$

and, to an approximation of the order of $1/n$,

$$\text{var}\hat{\mu}_2 = (\mu_4 - \mu_2^2)/n \quad \text{and} \quad \text{cov}(\hat{\mu}_1 , \hat{\mu}_2) = \mu_3/n \tag{4}$$

since the moments μ_p , $p = 3,4$ have been defined by relation 2.1.4(2).

Furthermore, if, instead of the variance μ_2, the standard deviation $\sigma = \sqrt{\mu_2}$ is considered, the estimator $\hat{\mu}_2$ is replaced by the estimator $\hat{\sigma} = \sqrt{\hat{\mu}_2}$, while from

$$d\hat{\sigma} = \frac{d\hat{\mu}_2}{2\sqrt{\mu_2}} \tag{5}$$

we obtain, with (4)

$$\text{var}\,\hat{\sigma} = \frac{\text{var}\,\hat{\mu}_2}{4\mu_2} = \frac{\mu_4 - \mu_2^2}{4n\mu_2} \tag{6}$$

and

$$\text{cov}(\hat{\mu}_1 , \hat{\sigma}) = \frac{1}{2\sqrt{\mu_2}} \,.\, \text{cov}(\hat{\mu}_1 , \hat{\mu}_2) = \frac{\mu_3}{2n\sigma} \tag{7}$$

In the case of the normal distribution of mean μ_1 and standard deviation σ , we have

$$\mu_3 = 0 \quad \text{and} \quad \mu_4 = 3\sigma^4$$

from which, with $\mu_2 = \sigma^2$, we derive

$$\text{var}\,\hat{\mu}_1 = \frac{\sigma^2}{n} , \quad \text{var}\,\hat{\sigma} = \frac{\sigma^2}{2n} \quad \text{and} \quad \text{cov}(\hat{\mu}_1 , \hat{\sigma}) = 0 \tag{9}$$

The distribution of the variable x can then be defined by means of a relation of the type 2.2(9), since for the reduced variable u, we have

$$dF(u) = f(u)du = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)du \tag{10}$$

The result is that the estimate of the probability associated with a value x_0 of the observed variable becomes with 2.2(11)

$$\hat{P}(x < x_0) = F(\hat{u}_0) \quad \text{where} \quad \hat{u}_0 = \frac{x_0 - \hat{\mu}}{\hat{\sigma}} \tag{11}$$

for which, with 2.2(12) and (13), we have

$$\text{var}\,F(\hat{u}_0) = \frac{1}{n}[f(u_0)]^2 . \left[1 + \frac{u_0^2}{2}\right] \tag{12}$$

Inversely, calculation of the quantile x_P corresponding to the probability $P = F(u_P)$ is derived as shown by equation 2.2(14), for which 2.2(15) here becomes

$$\text{var}\,\hat{x}_P = \frac{\sigma^2}{n} . \left[1 + \frac{u_P^2}{2}\right] \tag{13}$$

In practice, it is preferable to use as the estimator for the empirical variance $s^2 = \hat{\text{var}}\,x$ defined in 2.1.4(3), because s^2 gives an unbiased estimate for μ_2. Note, however, that $s = \sqrt{\text{var}\,x}$ has a systematic error with respect to σ

that, in the case of the normal, can be made negligible by taking

$$\hat{\sigma}=s\sqrt{\frac{n-1}{n-1,5}} \tag{14}$$

Let us add that introducing the estimates $\hat{u}_0$ and $\hat{\sigma}$ in (12) and (13) instead of the true unknown values u_0 and σ introduces only a negligible error of order $1/n$.

2.2.1.1 **Example 8.** *Estimation by fitting a normal distribution*

We consider again example 4 given in 2.1.7 of the monthly rainfall totals collected at Brussels-Uccle given in percentages of the normal total for which we shall assume that the distribution is normal. In reality, the actual distribution differs slightly from the normal and we shall show later how this hypothesis can be verified.

From the results obtained in 2.1.7.7, we obtain, with $n=60$,

$$\hat{\mu}_1=99,97 \quad \text{and} \quad \hat{\sigma}=s\sqrt{\frac{n-1}{n-1,5}}=46,74 \times 1,0043=46,94 \tag{1}$$

In order to allow comparison with the results given by empirical estimation, the probabilities associated with the values $x_0=15$, 51 and 97 corresponding respectively to $m=1$, 10 and 30 are calculated.

With 2.2.1(11) the value of $\hat{u}_0$ corresponding to $x_0=15$ is

$$\hat{u}_0=\frac{15-99,97}{46,94}=-1,810 \tag{2}$$

Similarly, for $x_0=51$ and 97, we obtain $\hat{u}_0=-1,043$ and $-0,063$.

Since the tables of $F(u)$ and $f(u)$ give the corresponding values for the normal,

$$\begin{aligned} F(\hat{u}_0)&=0,0352, \quad 0,148 \quad \text{and} \quad 0,475 \\ f(\hat{u}_0)&=0,0775, \quad 0,232 \quad \text{and} \quad 0,398 \end{aligned} \tag{3}$$

we deduce from these, with 2.2.1(12),

$$\sigma[F(\hat{u}_0)]=\sqrt{\operatorname{var} F(\hat{u}_0)}=f(u_0)\sqrt{(1+u_0^2/2)/n}$$

$$=0,016, \quad 0,037 \quad \text{and} \quad 0,051 \tag{4}$$

The empirical estimation gives

$$\hat{F}_0=0,0164, \quad 0,164 \quad \text{and} \quad 0,492 \tag{5}$$

with for $m=1$, 10 and 30, due to 2.1.2(6),

$$\sigma(F_0)=0,016, \quad 0,047 \quad \text{and} \quad 0,063 \tag{6}$$

It thus seems that parametric estimation gives a better accuracy here only in the vicinity of the mean. Done in this way, the comparison is not however correct, because the values of F_0 used in calculating the variances are not the same in both cases. If the values of F_0 given in (3) are used for this comparison, assuming that these values were obtained by the empirical method, using 2.1.2(14), we should have

$$\sigma(F_0)=0,023, \quad 0,045 \quad \text{and} \quad 0,063 \tag{7}$$

which shows, on the contrary a relatively better precision at the extreme values than in the vicinity of the mean.

Inversely, if we want the value of the observed variable corresponding to the probability $P = 1/11 = 0{,}0909$, the equation $F(u_p) = 0{,}0909$ gives, using the table for $F(u)$, $u_p = -1{,}335$, from which

$$\hat{x}_p = 99{,}97 - 46{,}94 \times 1{,}335 = 37{,}3 \tag{8}$$

Moreover, with 2.2.1(13), we obtain

$$\sigma(\hat{x}_p) = \sigma\sqrt{(1 + u_p^2/2)/n}$$

$$= 46{,}94 \times 0{,}1775 = 8{,}3 \tag{9}$$

As the distribution of $\hat{x}_p$ is approximately normal, the values $37{,}3 \pm 1{,}96 \times 8{,}3$ (21,0 and 53,6) define a two-sided confidence interval of x_0 at the 0,95 level. The empirical estimation in 2.1.7.4 gave the interval (22,1 ; 53,7), which is almost identical. This agreement between the two results can however be considered as fortuitous; the statement made on the subject of the values obtained in (4) and (7) allows us to predict, in fact, that the confidence interval given by the parametric estimation will be on average narrower than that obtained by the empirical estimation.

The empirical mean, however, is an exception and in this connection, it will be noted that the calculation of a confidence interval for the mean is carried out in the parametric case as given in 2.1.4. It is clear, however, that the interval obtained in this way is exact only if the distribution is normal. In other cases, it is only a good approximation obtained by virtue of the Central Limit Theorem.

2.2.2 Sufficient statistics. Estimates with minimum variance

From the foregoing, it follows that the accuracy of the parametric estimation is directly linked to the choice of the estimator of the parameter. In particular, the results from example 8 show that the choice of estimators with variance larger than those which were used can lead, as a rule, to less good estimates than empirical estimation. Seeking estimators with minimum variance is therefore the prime task. An important case where such estimators exist is that where the distribution possesses *sufficient statistics*. In order to define these, let

$$dF(x\ ,\ \theta) = f(x\ ,\ \theta)dx \tag{1}$$

be the differential equation which determines the distribution of the observed values $x_1, x_2, \ldots, x_n$ (assumed to be independent of each other) and let us consider the function

$$g(x\ ,\ \theta) = f(x_1\ ,\ \theta)f(x_2\ ,\ \theta) \ldots f(x_n\ ,\ \theta) \tag{2}$$

which is called the *likelihood function*, for which, by taking the logarithm, we put

$$L(\theta) = \log g(x\ ,\ \theta) = \Sigma \log f(x_i\ ,\ \theta) \tag{3}$$

If, in addition, T denotes a function $T(x_1, x_2, \ldots, x_n)$ of the observed values, the statistic T is called *sufficient* if the function $L(\theta)$ can be put in the form

$$L(\theta) = h(T\ ,\ \theta) + k(x_1, x_2, \ldots, x_n) \tag{4}$$

where $h(T\ ,\ \theta)$ depends only on T and θ and where $k(x_1, x_2, \ldots, x_n)$ depends only on the observed values $x_1, x_2, \ldots, x_n$.

The sufficient statistic as defined by (4) is unique, except that if T is sufficient, any one-to-one function of T will also be sufficient. Moreover, in this case, estimates of θ with minimum variance exist and are expressed as a

function of the statistic T only. To obtain such an estimate, we can proceed as follows.

Calculation of the mean E(T) of T by means of the integral

$$E(T) = \int T\, dF_1\, dF_2 \cdots dF_n \tag{5}$$

where dF_i stands for $f(x_i, \theta)dx_i$, leads to a relation of the form

$$E(T) = T(\theta) \tag{6}$$

By solving this equation with respect to θ, by substituting T for E(T), we find the estimator

$$\hat{\theta} = \theta(T) \tag{7}$$

wich may be corrected for bias by calculating the difference : $b(\theta) = \theta - E(\hat{\theta})$.

Similarly, the variance of T is obtained by means of the integral

$$\text{var } T = \int [T - E(T)]^2\, dF_1\, dF_2 \cdots dF_n \tag{8}$$

from which we derive the variance

$$\text{var } \hat{\theta} = \left[\frac{\partial \theta(T)}{\partial T}\right]^2 \cdot \text{var } T \tag{9}$$

which possibly can be corrected for bias of $\hat{\theta}$.

The preceding calculations are simplified if the function $b(T, \theta)$ in (4) can be put in the form

$$b(T, \theta) = A(\theta)T + B(\theta) \tag{10}$$

In this case, the sufficient estimator is the solution of the equation

$$\frac{\partial L(\theta)}{\partial \theta} = 0 \tag{11}$$

and if the estimator $\hat{\theta}$ obtained in this way is unbiased, we have

$$\text{var } \hat{\theta} = 1/I(\theta) \tag{12}$$

with

$$I(\theta) = E\left[\frac{\partial L(\theta)}{\partial \theta}\right]^2 = -E\left[\frac{\partial^2 L(\theta)}{\partial \theta^2}\right]$$

In the case where $\hat{\theta}$ is not without bias, relation (12) is only approximate. The exact relation is then

$$\text{var } \hat{\theta} = [1 + b'(\theta)]^2 / I(\theta) \tag{13}$$

in which we have

$$b(\theta) = \theta - E(\hat{\theta}) \quad \text{and} \quad b'(\theta) = \frac{db(\theta)}{d\theta} \tag{14}$$

Finally, when the parameter θ is a vector with several components $\theta_1, \theta_2, \ldots, \theta_k$, the generalization of the preceding ideas follows easily. In particular, equation (11) is replaced by the system of equations

$$\frac{\partial L(\theta_1, \theta_2, \ldots, \theta_k)}{\partial \theta_i} = 0, \quad i = 1, 2, \ldots, k \tag{15}$$

and the variance (12) becomes the covariance matrix

$$\| \operatorname{cov}(\hat{\theta}_i, \hat{\theta}_j) \| = \left\| - E\left[\frac{\partial^2 L(\theta_1, \theta_2, \cdots, \theta_k)}{\partial \theta_i \partial \theta_j} \right] \right\|^{-1} \tag{16}$$

The determination of the estimators by means of equations (11) or (15) and their variances by equations (12) or (16) is called estimation by the *maximum likelihood method*. This name is justified by the fact that if the estimator $\hat{\theta}$ derived from (11) is unbiased, it can be shown that the likelihood function $L(\theta)$ passes through a maximum for $\theta = \hat{\theta}$.

2.2.2.1 Case of the normal distribution

Since for this distribution, we have

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{u^2}{2}\right)$$

with

$$u = \frac{x - \mu}{\sigma}, \quad -\infty < x < +\infty \quad \text{and} \quad \sigma > 0 \tag{1}$$

by putting

$$T_1 = \Sigma x_i / n \quad \text{and} \quad T_2 = \Sigma (x - T_1)^2 \tag{2}$$

the likelihood function expressed in natural logarithms leads to

$$L = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{T_2 + n(T_1 - \mu)^2}{2\sigma^2} \tag{3}$$

from which it is evident that the statistics T_1 and T_2 are jointly sufficient.

As, on the other hand, we have

$$E(T_1) = \mu \quad \text{and} \quad E(T_2) = (n-1)\sigma^2 \tag{4}$$

we obtain the estimators already known

$$\hat{\mu} = T_1 \quad \text{and} \quad \hat{\sigma}^2 = T_2/(n-1) \tag{5}$$

which, by what has gone before, are estimators without bias at minimum variance.

In addition, applied to the function L, the maximum likelihood method gives the equations

$$\frac{\partial L}{\partial \mu} = \frac{2n(T_1 - \mu)}{2\sigma^2} = 0, \quad \frac{\partial L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{T_2 + n(T_1 - \mu)^2}{2\sigma^4} = 0 \tag{6}$$

from which we obtain

$$\hat{\mu} = T_1 \quad \text{and} \quad \hat{\sigma}^2 = T_2/n \tag{7}$$

Furthermore, from the relations

$$\frac{\partial^2 L}{\partial \mu^2} = -\frac{n}{\sigma^2}; \quad \frac{\partial^2 L}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{2[T_2 + n(T_1 - \mu)^2]}{2\sigma^6}; \quad \frac{\partial^2 L}{\partial \mu \partial \sigma^2} = -\frac{n(T_1 - \mu)}{\sigma^4} \tag{8}$$

we obtain

$$\| \operatorname{cov}(\hat{\mu}, \hat{\sigma}^2) \| = \left\| -E\left[\frac{\partial^2 L}{\partial \mu, \partial \sigma^2}\right] \right\|^{-1} = \begin{Vmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{Vmatrix} \tag{9}$$

Since, on the other hand, T_1 is a normal variate and T_2/σ^2 a χ^2 variate with $(n-1)$ degrees of freedom, we have

$$\operatorname{var} T_1 = \frac{\operatorname{var} x}{n} = \frac{\sigma^2}{n} \quad \text{and} \quad \operatorname{var} T_2 = 2(n-1)\sigma^4 \tag{10}$$

This shows that for the estimators (5) the exact relations are

$$\operatorname{var} \hat{\mu} = \frac{\sigma^2}{n} \quad \text{and} \quad \operatorname{var} \hat{\sigma}^2 = \frac{2\sigma^4}{n-1} \tag{11}$$

Thus we see that, in the case of the joint estimation of $\hat{\mu}$ and $\hat{\sigma}^2$, the solution by the maximum likelihood method is correct for $\hat{\mu}$, but is only approximate for $\hat{\sigma}^2$.

2.2.2.1.1 **Example 9.** *Fitting of a log-normal distribution*

When the variable studied is non-negative, the log-normal distribution can often give a good fit. To analyse this, it is sufficient to replace the values of the variable x by those of $\log x$ in the preceding equations. The daily averages of the wind speed in Table 10 give an example where such an adjustement is applicable.

In this case, by using natural logarithms, we obtain

$$\hat{\mu}(\log x) = 3{,}398\,, \quad \hat{\sigma}(\log x) = s\sqrt{\frac{n-1}{n-1{,}5}} = 0{,}3439 \times 1{,}0043 = 0{,}3454 \tag{1}$$

In particular, for $x_0 = 14$, with $\log x_0 = 2{,}639$, we have

$$\hat{u}_0 = \frac{2{,}639 - 3{,}398}{0{,}3456} = -2{,}20 \tag{2}$$

which gives

$$\hat{F}_0 = F(\hat{u}_0) = 0{,}0139, \quad f(\hat{u}_0) = 0{,}0355$$

and, as in 2.2.1.1(4)

$$\sigma(\hat{F}_0) = 0{,}0085 \tag{3}$$

Inversely, if we had taken $F(u_0) = 0{,}0139$, $u_0 = -2{,}20$ would be obtained, i.e. $\log \hat{x}_0 = 2{,}639$, and, as in 2.2.1.1(9), $\sigma(\log \hat{x}_0) = 0{,}082$.

Since, on the other hand, we have

$$d \log x = \frac{dx}{x}, \text{ that is to say }, \; dx = x \,.\, d\log x \tag{4}$$

we derive from this, to a good approximation,

$$\sigma(\hat{x}_0) = \hat{x}_0 \sigma(\log \hat{x}_0) = 14 \times 0{,}082 = 1{,}15 \tag{5}$$

TABLE 10. — Daily averages of the wind speed at Uccle in July. Ordered simple random sample of 60 values (in 0,1m/s).

m	x	m	x	m	x
1	14	21	26	41	34
2	15	22	26	42	36
3	17	23	26	43	37
4	18	24	27	44	38
5	19	25	28	45	38
6	19	26	28	46	39
7	20	27	28	47	41
8	21	28	28	48	43
9	21	29	29	49	44
10	22	30	29	40	44
11	22	31	29	51	45
12	22	32	30	52	45
13	22	33	30	53	45
14	23	34	30	54	46
15	23	35	30	55	47
16	23	36	31	56	48
17	24	37	31	57	49
18	24	38	32	58	51
19	24	39	32	59	62
20	26	40	34	60	68

2.2.2.2 The gamma distribution

In this case, we have

$$dF(x,\alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} dx \tag{1}$$

with

$$x \geqslant 0, \quad \alpha > 0 \quad \text{and} \quad \beta > 0$$

By using natural logarithms and by putting

$$T_1 = \Sigma \log x_i \quad \text{and} \quad T_2 = \Sigma x_i \tag{2}$$

we get

$$L = -n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha - 1)T_1 - T_2/\beta \tag{3}$$

from which it is found that T_1 and T_2 are jointly sufficient statistics.

On the other hand, as we have for the gamma distribution

$$E(T_1) = n\,E(\log x) = n \log \beta + n\psi(\alpha) \quad \text{with} \quad \psi(\alpha) = \frac{d}{d\alpha} \log \Gamma(\alpha) \tag{4}$$

and

$$E(T_2) = nE(x) = n\alpha\beta \tag{5}$$

by substituting in the previous relations T_1 for $E(T_1)$ and T_2 for $E(T_2)$, equations are obtained in α and β whose solutions are the sufficient estimates of these parameters. In particular, by eliminating β between equations (4) and (5) the equation in α is obtained :

$$\log \alpha - \psi(\alpha) = A \quad \text{with} \quad A = \log \frac{T_2}{n} - \frac{T_1}{n} \tag{6}$$

If we adopt the following approximation

$$\psi(\alpha)=\log\alpha-\frac{1}{2}\alpha^{-1}-\frac{1}{12}\alpha^{-2} \tag{7}$$

equation (6) becomes

$$12A\alpha^2-6\alpha-1=0 \tag{8}$$

which gives

$$\hat{\alpha}=\frac{1+\sqrt{1+4A/3}}{4A} \tag{9}$$

From (5), we then obtain

$$\hat{\beta}=T_2/(n\hat{\alpha}) \tag{10}$$

The accuracy of this estimate is adequate for all practical purposes. Note, however, that as the estimators which follow from this are not linear functions of T_1 and T_2, they cannot be expected to be unbiased.

Furthermore, the determination of the covariance matrix could be made directly by calculating the appropriate integrals. In order to simplify the calculation, we can, however, be satisfied with the approximate values given by the maximum likelihood method.

We thus find

$$\operatorname{var}\hat{\alpha}=\frac{\alpha}{n[\alpha\psi'(\alpha)-1]},\quad \operatorname{var}\hat{\beta}=\frac{\beta^2\psi'(\alpha)}{n[\alpha\psi'(\alpha)-1]},\quad \operatorname{cov}(\hat{\alpha},\hat{\beta})=\frac{-\beta}{n[\alpha\psi'(\alpha)-1]} \tag{11}$$

in which we can make in an approximate way

$$\psi'(\alpha)=\alpha^{-1}+\frac{1}{2}\alpha^{-2}+\frac{1}{6}\alpha^{-3}-\frac{1}{30}\alpha^{-5}+\frac{1}{84}\alpha^{-7}$$

and

$$\alpha\psi'(\alpha)-1=\frac{1}{2}\alpha^{-1}+\frac{1}{6}\alpha^{-2}-\frac{1}{30}\alpha^{-4}+\frac{1}{84}\alpha^{-6} \tag{12}$$

Moreover, for the calculation of $F(x,\alpha,\beta)$, we have

$$F(x,\alpha,\beta)=\frac{1}{\beta^\alpha\Gamma(\alpha)}\int_0^x x^{\alpha-1}e^{-x/\beta}dx \tag{13}$$

$$=\frac{1}{\Gamma(\alpha)}\int_0^t t^{\alpha-1}e^{-t}\,dt \tag{14}$$

$$=\frac{1}{\Gamma(\alpha)}\int_0^{u\sqrt{p+1}} t^{\alpha-1}e^{-t}dt \tag{15}$$

with

$$t=\frac{x}{\beta}\quad,\quad u=\frac{t}{\sqrt{p+1}}\quad\text{and}\quad p=\alpha-1$$

which reduces the calculation of $F(x,\alpha,\beta)$ to the tables of the incomplete gamma function.

As regards the calculation of var $\hat{F}_0$, for a given value of x, we see that from (15), we obtain

$$dF = \frac{\partial F}{\partial \alpha} \cdot d\alpha + \frac{\partial F}{\partial u} \cdot du \tag{16}$$

where, with $u = x/(\beta\sqrt{\alpha})$, we have

$$du = -\frac{u}{\beta}\, d\beta - \frac{u}{2\alpha}\, d\alpha \tag{17}$$

From this we obtain

$$\begin{aligned} \operatorname{var} \hat{F}_0 = &\left(\frac{\partial F}{\partial \alpha} - \frac{u}{2\alpha} \cdot \frac{\partial F}{\partial u}\right)^2 \cdot \operatorname{var} \hat{\alpha} + \left(\frac{u}{\beta} \cdot \frac{\partial F}{\partial u}\right)^2 \cdot \operatorname{var} \hat{\beta} \\ &- 2\frac{u}{\beta} \cdot \frac{\partial F}{\partial u}\left(\frac{\partial F}{\partial \alpha} - \frac{u}{2\alpha} \cdot \frac{\partial F}{\partial u}\right) \cdot \operatorname{cov}(\hat{\alpha}, \hat{\beta}) \end{aligned} \tag{18}$$

knowing that for the calculation of $\dfrac{\partial F}{\partial \alpha}$ we have

$$\left.\frac{\partial F}{\partial \alpha}\right|_{\alpha = \alpha_0} = \frac{F(\alpha_0 + \Delta\alpha) - F(\alpha_0 - \Delta\alpha)}{2\Delta\alpha} \tag{19}$$

with an approximation which improves as $\Delta\alpha$ decreases, and that for $\dfrac{\partial F}{\partial u}$ a similar relation can be written.

Inversely, for the calculation of var $\hat{x}_0$, when the value of F is given, the relation

$$\frac{\partial F}{\partial x} = \frac{\partial F}{\partial u} \cdot \frac{du}{dx} = \frac{1}{\beta\sqrt{\alpha}} \cdot \frac{\partial F}{\partial u} \tag{20}$$

gives, with 2.2(8)

$$\sigma(\hat{x}) = \beta\sqrt{\alpha}\left(\frac{\partial F}{\partial u}\right)^{-1} \sigma(\hat{F}_0) \tag{21}$$

where $\sigma(\hat{F}_0) = \sqrt{\operatorname{var} \hat{F}_0}$ is given by (18).

2.2.2.2.1 **Example 10.** *Fitting of a gamma distribution*

This type of fit also suits the case of a non-negative variable. Although often considered as an alternative solution for use in a log-normal distribution, the two solutions cannot be considered as equivalent.

Considering data in Table 10 again, with 2.2.2.2(2) and (6), we obtain

$$\frac{T_1}{n} = 3{,}398, \quad \frac{T_2}{n} = 31{,}72 \quad A = 3{,}457 - 3{,}398 = 0{,}059 \tag{1}$$

and thus using 2.2.2.2(9) and (10), gives us

$$\hat{\alpha} = 8{,}6 \quad \text{and} \quad \hat{\beta} = 3{,}69 \tag{2}$$

In addition, 2.2.2.2(12) gives

$$\psi'(\alpha) = 0{,}123, \quad \alpha\psi'(\alpha) - 1 = 0{,}0604 \tag{3}$$

which with 2.2.2.2(11) leads to

$$\text{var}\,\hat{\alpha}=2{,}37, \quad \text{var}\,\hat{\beta}=0{,}462, \quad \text{cov}(\hat{\alpha},\hat{\beta})=-1{,}018 \tag{4}$$

In particular, for $x_0=14$, we find

$$\hat{u}_0=x_0/(\hat{\beta}\sqrt{\hat{\alpha}})=14/(3{,}69\times\sqrt{8{,}6})=1{,}3 \tag{5}$$

With $p=8{,}6-1=7{,}6$, $u_0=1{,}3$, the tables for the gamma distribution give

$$\hat{F}_0=F(\hat{u}_0)=0{,}0238 \tag{6}$$

For the calculation of var $\hat{F}_0$ with $u=1{,}3$, we find, using the same tables,

$$\frac{\partial F}{\partial \alpha}=[F(p=7{,}8)-F(p=7{,}4)]/0{,}4=(0{,}0210061-0{,}0268726)/0{,}4=-0{,}0147 \tag{7}$$

whereas for $p=7{,}6$, we obtain

$$\frac{\partial F}{\partial u}=[F(u=1{,}4)-F(u=1{,}2)]/0{,}2=(0{,}0350552-0{,}0153433)/0{,}2=0{,}09856 \tag{8}$$

From this, with $u/2\alpha=0{,}0756$, $u/\beta=0{,}3523$ and using 2.2.2.2(18), we obtain

$$\begin{aligned}\text{var}\,\hat{F}_0 &= (-0{,}0147-0{,}0756\times 0{,}09856)^2.2{,}37+(0{,}3523\times 0{,}09856)^2\times 0{,}462\\ &\quad -2(0{,}3523)(0{,}09856)(-0{,}0147-0{,}0756\times 0{,}09856)(-1{,}018)\\ &= 0{,}000154\end{aligned} \tag{9}$$

that is to say

$$\sigma(\hat{F}_0)=0{,}0124 \tag{10}$$

For the same probability, the estimation by the log-normal law and the empirical estimation give respectively

$$\sigma(\hat{F}_0)=0{,}0125 \quad \text{and} \quad \sigma(F_0)=0{,}0194 \tag{11}$$

Inversely, if we had taken $F(u_0)=0{,}0238$, we should find $\hat{x}_0=14$. From 2.2.2.2(21), we then get

$$\sigma(\hat{x}_0)=3{,}69\times\sqrt{8{,}6}\times 0{,}0124/0{,}09856=1{,}36 \tag{12}$$

2.2.3 Estimation by the maximum likelihood method. Efficiency of an estimator

When no sufficient statistic exists because of the form of the distribution function, the problem of seeking good estimators can be solved by using the maximum likelihood method defined by equations 2.2.2(11) and (15). Under certain regularity conditions and notably when the number of parameters is small with respect to the size of the sample, the estimators obtained in this way are *asymptotically* :

1. unbiased;
2. distributed jointly as normal;
3. with minimum variance.

In addition, the variance or the covariance matrix of these estimators takes as an asymptotic value that found from relations 2.2.2(12) or (16).

Furthermore, if v denotes the variance of any estimator t and v_0 the variance of the corresponding estimator t_0 by the method of maximum likelihood, the limiting value, for large samples, of the ratio

$$Ef = \lim \frac{v_0}{v} \tag{1}$$

is called the *efficiency* of the estimator under consideration.

In the assumption where v_0 and v are functions of $1/n$, the ratio Ef gives the proportion by which the sample size must be changed for the estimator t_0 to be as accurate as the estimator t.

If the estimator t_0 is itself any estimator whatsoever taken as reference, the ratio Ef is called the *relative efficiency* of the estimator t with respect to the estimator t_0.

In example 10, for $\hat{F}_0 = 0{,}0238$, we found for the gamma distribution

$$\sigma(\hat{F}_0) = 0{,}0124 \tag{2}$$

whereas for the log-normal distribution and for the empirical estimate, we had

$$\sigma(\hat{F}_0) = 0{,}0125 \quad \text{and} \quad 0{,}194 \tag{3}$$

The efficiency of the two last methods for $\hat{F}_0 = 0{,}0238$ compared with the first method is thus respectively

$$\left(\frac{0{,}0124}{0{,}0125}\right)^2 \cong 1 \qquad \left(\frac{0{,}0124}{0{,}0194}\right)^2 = 0{,}41 \tag{4}$$

From this it is concluded that, for the value of F considered, the estimates given by the gamma and by the log-normal distributions have equivalent efficiencies, whereas the efficiency of the empirical estimate shows that, for a sample of 60 observations the accuracy obtained is the same as that obtained by the gamma distribution from a sample of 25 observations. The efficiency thus supplies a measure of the quantity of information which is neglected by adopting one estimator rather than another.

Another example is given by fitting the gamma distribution by means of the method of moments. In this case, the estimates of the parameters α and β are derived from the equations

$$\beta\alpha = \hat{\mu} \quad , \quad \beta^2\alpha = s^2 \tag{5}$$

where $\hat{\mu}$ and s^2 are the empirical mean and variance of the sample.

It has been shown that, for $\alpha = 1$, the efficiency of the estimates thus obtained does not exceed 0,5 , whereas if $\alpha = 10$, it is close to 0,9. From this it is deduced that we can use the simplifications into the calculations given by relations (5) only if $\alpha > 10$.

2.2.3.1 The double exponential distribution (Fisher-Tippett Type I)

For this distribution, we have

$$F(u) = \exp[-e^{-u}] \quad \text{with} \quad u = \frac{x - \mu}{\sigma} \tag{1}$$

thas is to say, the observed variable x is linked to the reduced variable u by a location parameter μ and a scale parameter

$\sigma > 0$. This is often of use in the case of monthly or annual extremes. From this we get

$$dF(x) = \exp\left[-e^{-u}\right] . e^{-u} \frac{dx}{\sigma} \tag{2}$$

The result is that if we denote by u_i , $i = 1, 2, \ldots, n$, the values of u corresponding to the observations x_i and given by relation (1), the likelihood function, using natural logarithms, leads to the expression

$$L = -\Sigma e^{-u_i} - \Sigma u_i - n \log \sigma \tag{3}$$

The estimates by the maximum likelihood method thus satisfy the equations

$$\frac{\partial L}{\partial \mu} = -\frac{1}{\sigma}(\Sigma e^{-u_i} - n) = 0, \quad \frac{\partial L}{\partial \sigma} = -\frac{1}{\sigma}(\Sigma e^{-u_i} u_i - \Sigma u_i + n) = 0 \tag{4}$$

whilst from the relations

$$E\left(\frac{\partial^2 L}{\partial \mu^2}\right) = -\frac{n}{\sigma^2}, \quad E\left(\frac{\partial^2 L}{\partial \sigma^2}\right) = -\frac{n}{\sigma^2}\left[\frac{\pi^2}{6} + (\gamma - 1)^2\right], \quad E\left(\frac{\partial^2 L}{\partial \mu \partial \sigma}\right) = -\frac{n}{\sigma^2}(\gamma - 1) \tag{5}$$

be obtained after some calculations, using $\gamma = E(u) = 0{,}577216 \ldots$ (the Euler number) and $\pi^2/6 = \text{var } u$, we get the asymptotic values :

$$\text{var}_0 \hat{\mu} = \frac{\sigma^2}{n}\left[\frac{\pi^2}{6} + (\gamma - 1)^2\right] \Big/ \frac{\pi^2}{6} = 1{,}10866 \frac{\sigma^2}{n}$$

$$\text{var}_0 \hat{\sigma} = \frac{\sigma^2}{n}\left(1 \Big/ \frac{\pi^2}{6}\right) = 0{,}60793 \frac{\sigma^2}{n}$$

$$\text{cov}_0(\hat{\mu}, \hat{\sigma}) = \frac{\sigma^2}{n}(1 - \gamma) \Big/ \frac{\pi^2}{6} = 0{,}25702 \frac{\sigma^2}{n} \tag{6}$$

As equations (4) are fairly complicated to solve and as, in addition, they lead to estimates which are biased, preference is generally given to other solutions a little less efficient, but unbiased and easier to calculate. One of these solutions is that given by Lieblein in which for $n > 6$ the efficiency of $\hat{\mu}$ is close to 1 whereas that of $\hat{\sigma}$ is maintained between 0,7 and 0,8. Another acceptable solution is that proposed by Kimball. It consists of noting that, because of the first of the two equations (4), the second gives

$$n\sigma = \Sigma x_i - \Sigma e^{-u_i} x_i \tag{7}$$

In addition, if the x_i are arranged in increasing order $x'_1 \leqslant x'_2 \leqslant \ldots \leqslant x'_n$, with (1), relation (7) becomes

$$n\sigma = \Sigma x'_m (1 + \log F_m) \tag{8}$$

Kimball's solution is derived from this by replacing $\log F_m$ by its expectation

$$E(\log F_m) = -\left(\frac{1}{m} + \frac{1}{m+1} + \ldots + \frac{1}{n}\right) \tag{9}$$

The estimate obtained in this way is also biased. The bias is eliminated by multiplying the estimate of σ by the coefficient b_n taken either from a table or, if $n > 11$, from the relation

$$\log(b_n - 1) = -0{,}975652 \log(n-1) + 1{,}043532 - 1{,}950309/\log(n-1) + 3{,}574231/\log^2(n-1) \tag{10}$$

The corresponding estimate for μ is then

$$\hat{\mu} = \bar{x} - \gamma\hat{\sigma} \quad \text{with} \quad \bar{x} = \Sigma x_i/n \tag{11}$$

The efficiency of these estimators can be considered as excellent. For $n > 10$, $Ef(\hat{\mu})$ is already close to 1, whereas $Ef(\hat{\sigma})$ exceeds 0,80. Compared with the Lieblein estimators, they have in addition the advantage that their efficiency increases regularly with n.

If we note also that their covariance rapidly approaches the value given in (6), we can see that, to a good approximation, the values of $\text{var}_0\,\hat{\mu}$, $\text{var}_0\,\hat{\sigma}$ and $\text{cov}_0(\hat{\mu}, \hat{\sigma})$ given in (6) can be used during the calculation of the errors of estimation.

Finally the method which has just been described can be used for distributions derived from the double exponential distribution, notably in the case where $x = \pm \log(\pm t)$ where, depending on the case, t is either non-negative or non-positive. The functions obtained in this way are simplified forms of the second and third type Fisher-Tippett asymptotic distribution.

2.2.3.1.1 **Example 11.** *Fitting a double exponential distribution*

The annual maximum of the daily rainfall at Uccle (Table 11) gives an example where the simplified form of the Fisher-Tippett Type II distribution can be used. In other words, it can be assumed that the logarithm of the variable considered has a double exponential distribution.

In this case, we get :

$$1 + E(\log F_1) = 1 - \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{60}\right) = -3{,}67987$$

$$1 + E(\log F_2) = 1 - \left(\frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{60}\right) = -2{,}67987$$

$$1 + E(\log F_3) = 1 - \left(\frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{60}\right) = -2{,}17987$$

$$\ldots$$

$$1 + E(\log F_{60}) = 1 - \frac{1}{60} = 0{,}98333 \tag{1}$$

Using natural logarithms then gives : $\log x'_1 = \log 187 = 5{,}2311$, $\log x'_2 = \log 198 = 5{,}2883$ etc., whilst from 2.2.3.1(10) we obtain $b_{60} = 1{,}04084$.

Using 2.2.3.1(8), we find

$$\hat{\sigma} = b_{60} \sum_{1}^{60} \log x'_m \,[1 + E(\log F_m)]/60 = 0{,}2421 \tag{2}$$

TABLE 11. – Annual maximum of the daily rainfall at Uccle. Ordered random sample of 60 values (in 0,1mm).

m	x	m	x	m	x
1	187	21	277	41	343
2	198	22	279	42	354
3	198	23	282	43	356
4	203	24	284	44	368
5	204	25	285	45	372
6	214	26	290	46	375
7	217	27	298	47	378
8	222	28	301	48	392
9	232	29	306	49	398
10	238	30	325	50	405
11	240	31	329	51	406
12	243	32	333	52	406
13	249	33	334	53	412
14	255	34	336	54	415
15	265	35	338	55	468
16	265	36	338	56	480
17	265	37	340	57	507
18	266	38	340	58	511
19	271	39	342	59	600
20	272	40	343	60	723

whilst with

$$\overline{\log x} = \Sigma(\log x)/60 = 5{,}752 \tag{3}$$

2.2.3.1(11) gives

$$\hat{\mu} = 5{,}752 - 0{,}577216 \times 0{,}2421 = 5{,}612 \tag{4}$$

In particular, for $x_0 = 723$, $\log x_0 = 6{,}5834$:

$$\hat{u}_0 = \frac{6{,}5834 - 5{,}612}{0{,}2421} = 4{,}012$$

$$\hat{F}_0 = F(\hat{u}_0) = 0{,}9821 \quad \text{and} \quad f(\hat{u}_0) = 0{,}01777 \tag{5}$$

Moreover, with the asymptotic values given in 2.2.3.1(6), 2.2(13) leads to

$$\text{var}\ \hat{u}_0 = \frac{1}{n}\,(1{,}10866 + 2u_0\ 0{,}25702 + u_0^2\ 0{,}60793) \tag{6}$$

With $u_0 = 4{,}012$, we obtain

$$\text{var}\ \hat{u}_0 = 0{,}2159 \tag{7}$$

and therefore

$$\sigma(\hat{u}_0) = 0{,}4647 \tag{8}$$

which, with 2.2(12), gives

$$\sigma(\hat{F}_0) = 0{,}01777 \times 0{,}4647 = 0{,}00826 \tag{9}$$

For the same value of $\hat{F}_0$, the empirical estimation gives

$$\sigma(F_0) = 0{,}0168 \tag{10}$$

Inversely, for F = 0,9821 , we should have obtained $\log \hat{x}_0 = 6{,}5834$, $\hat{x}_0 = 723$, whilst 2.2(13) and (15) here lead to

$$\sigma(\log \hat{x}_0) = \hat{\sigma} \times \sigma(\hat{u}_0) = 0{,}2421 \times 0{,}4647 = 0{,}1125 \tag{11}$$

in other words, as in 2.2.2.1.1(5)

$$\sigma(\hat{x}_0) = \hat{x}_0 \, \sigma(\log \hat{x}_0) = 723 \times 0{,}1125 = 81{,}3 \tag{12}$$

2.2.4 Estimation by least squares

Consider the linear form

$$x = u_1\theta_1 + u_2\theta_2 + \ldots + u_k\theta_k + e \tag{1}$$

where θ_1, θ_2, ..., θ_k are constants, x, u_1, u_2, ..., u_k are observations which are all not necessarily random, but amongst which u_1, u_2, ..., u_k are linearly independent, and where e is a random quantity distributed independently, with zero mean and variance σ^2 which may be unknown.

In this case the function

$$x = \Sigma u_j \theta_j \tag{2}$$

is called the linear regression of x in $u_1, u_2, \ldots, u_k$. Furthermore, if x_i, u_{ij} and e_i, $i = 1, 2, \ldots, n$, $(n > k)$ are n corresponding values of the observation x, the observations u_1, u_2, ..., u_k and the variable e, the determination of the constants θ_j by the method of least squares consists of minimizing the quadratic form

$$Q = \Sigma e_i^2 = \sum_i (x_i - \sum_j u_{ij}\theta_j)^2 \tag{3}$$

which is obtained by resolving the system of equations

$$\frac{\partial Q}{\partial \theta_j} = -2\sum_i u_{ij}(x_i - \sum_j u_{ij}\theta_j) = 0 \tag{4}$$

The expression for the solution of this system of equations can be made particularly simple by introducing matrix notation. In this case, if x, θ and e denote the component vectors x_i, θ_j and e_i and if u is the matrix $\| u_{ij} \|$, equations (1), (3) and (4) become respectively

$$x = u\theta + e \tag{5}$$

$$Q = e'e = (x - u\theta)'(x - u\theta) \tag{6}$$

and

$$\frac{\partial Q}{\partial \theta} = -2u'(x - u\theta) = 0 \tag{7}$$

that is

$$u'x = u'u\theta \tag{8}$$

the solution of which is written

$$\hat{\theta} = (u'u)^{-1}\, u'x \tag{9}$$

where for a given matrix a, the matrix a^{-1} denotes the inverse of the matrix a and the matrix a', the transpose (inversion of the lines and columns) of the matrix a.

A first generalization of the method consists of assuming that the components e_i of the random error e have different known variances σ_i^2. In this case, the quadratic form (3) becomes :

$$Q = \sum \frac{e_i^2}{\sigma_i^2} = \sum_i \left(\frac{x_i - \Sigma u_{ij}\theta_j}{\sigma_i} \right)^2 \tag{10}$$

If we note that the covariance matrix of variables e_i is none other than the diagonal matrix

$$v = \begin{Vmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & & \\ \vdots & & & \\ 0 & \cdots & & \sigma_n^2 \end{Vmatrix} \tag{11}$$

the quadratic form Q can be written

$$Q = e'v^{-1}e \tag{12}$$

This new expression makes a last generalization possible by assuming that the random variables e_i are no longer independent but have any, but a known, covariance matrix, of order n :

$$v = \| v_{rs} \|, \quad r, s = 1, 2, \ldots, n \tag{13}$$

where $v_{rs} = \text{cov}(e_r, e_s)$. The quadratic form (12) then keeps the same expression

$$Q = e'v^{-1}e = (x - u\theta)'v^{-1}\,(x - u\theta) \tag{14}$$

and using the method leads successively to

$$\frac{dQ}{d\theta} = -2u'v^{-1}(x - u\theta) = 0 \tag{15}$$

that is to say

$$u'v^{-1}x = u'v^{-1}\,u\theta \tag{16}$$

from which we obtain

$$\hat{\theta} = (u'v^{-1}u)^{-1}u'v^{-1}x \tag{17}$$

It is clear that the estimates obtained in this way for the components of the vector θ are linear forms of the components of the vector x. In addition, it can be shown that we have $E(\hat{\theta}) = \theta$ and that var $\hat{\theta}$ is a minimum, which means that, amongst all the linear estimators of θ, the estimator $\hat{\theta}$ is unbiased and possesses the smallest generalized variance.

Furthermore, this variance (which is in fact the covariance matrix of the components of θ) can be calculated directly as follows. From (17), with (5), we obtain :

$$\begin{aligned}\hat{\theta} &= (u'v^{-1}u)^{-1}u'v^{-1}(u\theta + e)\\ &= (u'v^{-1}u)^{-1}u'v^{-1}u\theta + (u'v^{-1}u)^{-1}u'v^{-1}e\\ &= \theta + (u'v^{-1}u)^{-1}u'v^{-1}e\end{aligned} \tag{18}$$

from which we find

$$\hat{\theta} - \theta = (u'v^{-1}u)^{-1}u'v^{-1}e \tag{19}$$

and, as a result

$$\operatorname{var}\hat{\theta} = \mathrm{E}[(\hat{\theta} - \theta)(\hat{\theta} - \theta)']$$

$$= \mathrm{E}[(u'v^{-1}u)^{-1}u'v^{-1}ee'v^{-1}u(u'v^{-1}u)^{-1}]$$

With $\mathrm{E}(ee') = v$, after simplification we obtain

$$\operatorname{var}\hat{\theta} = (u'v^{-1}u)^{-1} \tag{20}$$

Considering our knowledge of the covariance matrix v, it should be stressed that, because of (17), the multiplication of all of its elements by the same factor does not change the solution $\hat{\theta}$. Then again, it can be shown that using the hypothesis where all the components of the vector e possess a joint normal distribution, the quadratic form Q given by (14) has a χ^2 distribution with $(n - k)$ degrees of freedom when $\theta = \hat{\theta}$.

Therefore if, in the most simple case, the variance σ^2 is not known, the value of Q given by (3) could be used to construct a confidence interval for σ^2.

On the other hand, if the matrix v is known, the calculation of Q by means of (14) will enable us to verify the compatibility of the series of observations with the hypothesis posed by relation (1).

It is clear that during the application of a χ^2 test to the statistic Q, large values of this statistic implies that the linear form (2) is not sufficient to fully explain the variation of x as a function of u. Inversely, too small values are evidence of either an overestimation of the value of the elements of the matrix v or the overabundance of the components of θ used in the regression (2). The test can thus be used in its single-sided or two-sided form depending on the alternative hypotheses considered.

For calculation of Q, it can be useful to know that from (14) we obtain

$$\hat{\mathrm{Q}} = x'v^{-1}x - \hat{\theta}'u'v^{-1}x - x'v^{-1}u\hat{\theta} + \hat{\theta}'u'v^{-1}u\hat{\theta} \tag{21}$$

Furthermore, from (17) we get simultaneously

$$u'v^{-1}x = u'v^{-1}u\hat{\theta} \quad \text{and} \quad x'v^{-1}u = \hat{\theta}'u'v^{-1}u \tag{22}$$

As a result, after simplification, (21) is written

$$\hat{\mathrm{Q}} = x'v^{-1}x - \hat{\theta}'u'v^{-1}u\hat{\theta} \tag{23}$$

or again, by putting $\beta = u'v^{-1}x$ and using (22)

$$\hat{\mathrm{Q}} = x'v^{-1}x - \hat{\theta}'\beta \tag{24}$$

The linear estimators obtained by the least squares method also have a normal asymptotic distribution as soon as the joint distribution of the variables e_i fulfils certain very general conditions for regularity. As a result, the calculation of var $\hat{\theta}$ by means of (20) can also be used to advantage in significance tests on $\hat{\theta}$.

More specifically, if we wish to establish that the true value of θ differs from a particular given value θ_0, the conclusions will be based on the value taken by the quadratic form

$$X=(\hat{\theta}-\theta_0)'(\text{var }\hat{\theta})^{-1}(\hat{\theta}-\theta_0) \tag{25}$$

in other words, with (20)

$$X=(\hat{\theta}-\theta_0)'(u'v^{-1}u)(\hat{\theta}-\theta_0) \tag{26}$$

whose distribution under the null hypothesis $\theta=\theta_0$ is approximately a χ^2 variate with a number of degrees of freedom equal to the number of components of the vector θ.

When the joint distribution of the variables e_i is normal (Gaussian), it is clear that the distribution of the preceding statistic is exactly that of a χ^2 variate.

2.2.4.1 Examples of estimation by the least squares

2.2.4.1.1 **Example 12.** *Calculation of the normal of the number of frost days in Belgium in January from the normal of the daily mean minimum air temperature*

a) *Estimation of the parameters*

The normals shown in Table 12 of the number of frost days and of the mean daily minimum of the corresponding air temperature were drawn up on the basis of 30 year records.

If, for the relation between the number of frost days g and the mean daily minimum air temperature t for the month of January we adopt a linear model of the form

$$g=\theta_1+\theta_2 t+e \tag{1}$$

which is valid for all the stations in the Belgian climatological network, where θ_1 and θ_2 denote constants and e is an independent random quantity with zero mean, it can be predicted that between the normals $\overline{g}$ and $\overline{t}$ a relation of the form

$$\overline{g}=\theta_1+\theta_2\overline{t}+\overline{e} \tag{2}$$

also exists, where θ_1 and θ_2 are the same constants as in (1), where $\overline{g}$ and $\overline{t}$ are the normals for a given station and where $\overline{e}$ is a random quantity with zero mean, whose values are independent from one station to another. Moreover, since the quantity $\overline{e}$ is itself an average, its distribution is, at least asymptotically, normal.

Thus the estimation of the constants θ_1 and θ_2 can be made using the leastsquares method in its most simple form. Note however that by virtue of (2), we have

$$\text{var }\overline{g}=\theta_2^2\text{ var }\overline{t}+\text{var }\overline{e} \tag{3}$$

from which we find that var $\overline{e}$ can be estimated only after calculation of θ_2 and by knowing var $\overline{g}$ and var $\overline{t}$.

This being so, using the method in its matrix version, let us recall that if a, b, u, ... denote the matrices $\|a_{ij}\|$, $\|b_{ij}\|$, $\|u_{ij}\|$, ..., the product of the matrices a and b is the matrix

$$ab=\|\sum_j a_{ij}b_{jl}\| \tag{4}$$

from which it transpires that

$$u'u=\|\sum_i u_{ij}u_{il}\| \tag{5}$$

and that if the matrix b is the inverse a^{-1} of the matrix a, we have

$$ab = \| \sum_j a_{ij} b_{jl} \| = \| \delta_{il} \| \tag{6}$$

with $\delta_{ij}=1$ for $i=l$ and $\delta_{ij}=0$ for $i \neq l$.

Under these conditions, taking up again the notation of 2.2.4, the use of relation (2) leads to

$$x_i = \bar{g}_i \ , \ u_{i1} = 1 \ , \ u_{i2} = \bar{t}_i \ , \ v_{ii} = \text{var } \bar{e} \ , \ v_{ii'} = 0 \tag{7}$$

for $i \neq i' = 1, 2, \ldots, 15$.

Moreover, putting $a = u'u$, with (5), we obtain :

$$a_{11} = \Sigma u_{i1} u_{i1} = 15, \quad a_{12} = a_{21} = \Sigma u_{i1} u_{i2} = \Sigma u_{i2} u_{i1} = \Sigma \bar{t}_i, \tag{8}$$

$$a_{22} = \Sigma u_{i2} u_{i2} = \Sigma \bar{t}_i^2$$

which gives us

$$a^{-1} = (u'u)^{-1} = \frac{1}{a_{11}a_{22} - a_{12}^2} \begin{Vmatrix} a_{22} & -a_{12} \\ -a_{12} & a_{11} \end{Vmatrix} \tag{9}$$

On the other hand, we also have

$$u'x = \begin{Vmatrix} \Sigma u_{i1} x_i \\ \Sigma u_{i2} x_i \end{Vmatrix} = \begin{Vmatrix} \Sigma \bar{g}_i \\ \Sigma \bar{t}_i \bar{g}_i \end{Vmatrix} \tag{10}$$

that is to say that, by putting $\beta = u'x$, we make :

$$\beta_1 = \Sigma \bar{g}_i \ ; \ \beta_2 = \Sigma \bar{t}_i \bar{g}_i \tag{11}$$

The solution 2.2.4(9) is thus written with (9)

$$\hat{\theta} = (u'u)^{-1} u'x = a^{-1}\beta = \frac{1}{a_{11}a_{22} - a_{12}^2} \begin{Vmatrix} a_{22} & -a_{12} \\ -a_{12} & a_{11} \end{Vmatrix} \cdot \begin{Vmatrix} \beta_1 \\ \beta_2 \end{Vmatrix} \tag{12}$$

in other words,

$$\begin{aligned} \hat{\theta}_1 &= (a_{22}\beta_1 - a_{12}\beta_2)/(a_{11}a_{22} - a_{12}^2) \\ \hat{\theta}_2 &= (-a_{12}\beta_1 + a_{11}\beta_2)/(a_{11}a_{22} - a_{12}^2) \end{aligned} \tag{13}$$

As, in addition to $a_{11}=15$, we extract from Table 12 :

$$\begin{aligned} a_{12} &= \Sigma \bar{t}_i = (-30) + (-11) + \cdots + (-03) = -122 \\ a_{22} &= \Sigma \bar{t}_i^2 = (-30)^2 + (-11)^2 + \cdots + (-03)^2 = 3080 \\ \beta_1 &= \Sigma \bar{g}_i = (220) + (173) + \cdots + (132) = 2338 \\ \beta_2 &= \Sigma \bar{t}_i \bar{g}_i = (-30)(220) + (-11)(173) + \cdots + (-03)(132) = -24660, \end{aligned} \tag{14}$$

with (13), we obtain :

$$\begin{aligned} \hat{\theta}_1 &= [3080 \times 2338 - (-122) \times (-24660)]/[15 \times 3080 - (-122)^2] = 4192520/31316 = 133{,}9 \\ \hat{\theta}_2 &= [-(-122) \times 2338 + 15 \times (-24660)]/31316 = -2{,}704 \end{aligned} \tag{15}$$

If then we adopt the values 169 and 13,3 for var $\bar{g}$ and var $\bar{t}$ respectively, obtained from the series of observations for g_i and t_i, with (3), we obtain

$$\text{var } \bar{e} = 169 - (-2{,}704)^2 \times 13{,}3 = 71{,}8 \tag{16}$$

From this we deduce, with 2.2.4(20),

$$\text{var } \hat{\theta} = 71{,}8(u'u)^{-1} = 71{,}8a^{-1} \tag{17}$$

that is to say, on means of (9) and (14) :

$$\begin{aligned} \text{var } \hat{\theta}_1 &= 71{,}8 \times 3080/31316 = 7{,}06 \\ \text{var } \hat{\theta}_2 &= 71{,}8 \times 15/31316 = 0{,}0344 \\ \text{cov } (\hat{\theta}_1, \hat{\theta}_2) &= 71{,}8 \times 122/31316 = 0{,}280 \end{aligned} \tag{18}$$

b) *Significance tests*

The calculation of $\hat{Q}$ given in 2.2.4(24) here becomes, with (14) and (15) and the values of $\bar{g}_i$ in Table 12,

$$\begin{aligned} \hat{Q} &= (\Sigma \bar{g}_i^2 - \hat{\theta}_1\beta_1 - \hat{\theta}_2\beta_2)/71{,}8 \\ &= (380378 - 133{,}9 \times 2338 - 2{,}704 \times 24660)/71{,}8 \\ &= 639{,}16/71{,}8 = 8{,}90 \end{aligned} \tag{19}$$

Knowing that, because of the normality of the quantity $\bar{e}$, the variable $\hat{Q}$ has a χ^2 distribution with $15 - 2 = 13$ degrees of freedom, and since in this distribution we have

$$P(\chi^2 > 8{,}90) \cong 0{,}78 \tag{20}$$

we conclude that, at the 0,05 level, all of the hypotheses posed are acceptable.

In order to establish the significant character of the coefficients of the linear regression obtained, the quadratic form 2.2.4(25) with $\theta_0 = 0$ can be calculated. In this case, this quadratic form becomes

$$X = \theta'(u'u)\hat{\theta}/71{,}8 \tag{21}$$

in other words, with (14) and (15),

$$\begin{aligned} X &= (a_{11}\hat{\theta}_1^2 + 2a_{12}\hat{\theta}_1\hat{\theta}_2 + a_{22}\hat{\theta}_2^2)/71{,}8 \\ &= [15(133{,}9)^2 + 2(-122)(133{,}9)(-2{,}704) + (3080)(-2{,}704)^2]/71{,}8 \\ &= 379802/71{,}8 = 5290 \end{aligned} \tag{22}$$

a value which for a χ^2 variate with two degrees of freedom is very significant .

TABLE 12. – Normals of frost days in January $\bar{g}$ (in tenths of a day) and of mean minimum daily corresponding air temperature $\bar{t}$ (in 0.1°C) for fifteen stations in the Belgian climatological network

Station i	$\bar{g}_i$	$\bar{t}_i$
1	220	−30
2	173	−11
3	110	09
4	158	−11
5	213	−26
6	142	−03
7	131	−01
8	174	−17
9	109	10
10	132	02
11	177	−18
12	128	03
13	177	−20
14	162	−06
15	132	−03

Since, moreover, with (15) and (18), the ratios :

$$\hat{\theta}_1^2/\text{var}\,\hat{\theta}_1 = (133{,}9)^2/7{,}06 = 2540$$

$$\hat{\theta}_2^2/\text{var}\,\hat{\theta}_2 = (-2{,}704)^2/0{,}0344 = 213 \qquad (23)$$

also lead to very significant values for the χ^2 variate with one degree of freedom, the hypotheses $\theta_1 = 0$ and $\theta_2 = 0$ can both be rejected.

c) *Estimates by means of (2) and (15)*

If we take the estimator

$$\hat{\bar{g}} = \hat{\theta}_1 + \hat{\theta}_2 \bar{t} \qquad (24)$$

we shall have

$$\text{var}\,\hat{\bar{g}} = \text{var}\,\hat{\theta}_1 + 2\bar{t}\,\text{cov}\,(\hat{\theta}_1, \hat{\theta}_2) + \bar{t}^2\,\text{var}\,\hat{\theta}_2 \qquad (25)$$

that is, with (18),

$$\text{var}\,\hat{\bar{g}} = 7{,}06 + 2 \times 0{,}280\bar{t} + 0{,}0344\bar{t}^2 \qquad (26)$$

In particular, in the case of station 1 (Table 12) for $\bar{t} = -30$, it becomes

$$\hat{\bar{g}} = 133{,}9 - 2{,}704(-30) = 215$$

and

$$\text{var}\,\hat{\bar{g}} = 7{,}06 + 2 \times 0{,}280(-30) + 0{,}0344(-30)^2 = 21{,}2 \qquad (27)$$

which by comparison with var $\bar{e} = 71{,}8$ shows that the estimate $\hat{\bar{g}} = 215$ can be considered as an improved estimate of the observed $\bar{g} = 220$.

Furthermore, $\hat{\bar{g}}$ can also be considered as an estimate of the true mean m_g of the distribution of the number of days of frost. In this case, the value of var $\hat{\bar{g}}$ must, however, be increased from the variance of the error of t with respect to its true mean m_t, which is achieved by adding the term $\hat{\theta}_2^2$ var $\bar{t}$ to var $\hat{\bar{g}}$ which is the value already utilized in (16) :

$$(-2{,}704)^2 \times 13{,}3 = 97{,}2 \qquad (28)$$

We thus obtain

$$\text{var}\,\hat{\bar{g}} = 21{,}2 + 97{,}2 = 118{,}4 \qquad (29)$$

which, compared with var $\bar{g} = 169$, shows that $\hat{\bar{g}}$ is also an estimation of the true mean m_g better than $\bar{g}$.

We should note here that if changes of the origin in $\bar{g}$ and $\bar{t}$ are carried out, only the values of $\hat{\theta}_1$, var $\hat{\theta}_1$ and cov $(\hat{\theta}_1, \hat{\theta}_2)$ are modified. Furthermore, these changes of origin do not affect the estimates $\hat{\bar{g}}$ nor the value of var $\hat{\bar{g}}$ given by relations (24) and (25), calculated in the new reference system. In particular, if we take as the origin of the temperatures $\bar{t}$ the mean of the values of $\bar{t}_i$ from Table 12, in the new reference system we shall have $\sum \bar{t}_i = 0$ and, because of (9) and (17), cov $(\hat{\theta}_1, \hat{\theta}_2) = 0$, from which it results that var $\hat{\bar{g}}$ will have the form

$$\text{var}\,\hat{\bar{g}} = \text{var}\,\hat{\theta}_1 + \bar{t}^2\,\text{var}\,\hat{\theta}_2 \qquad (30)$$

We come to the conclusion that the error of estimation which affects the values of $\hat{\bar{g}}$ increases when the corresponding value of $\bar{t}$ is well away from the average of $\bar{t}$ values considered.

This is all the more useful as, in the example above, the value $t = -30$ is the one which in Table 12 differs most from the average -08. It can thus be seen that the least squares estimate is better here than the initial estimate over the whole range of the values observed. It will be noted that this property is generally satisfied only when the number of estimated parameters is small with respect to the number of observations n.

One point must still be made regarding the use of relation (1) by means of (13) to calculate the number of frost days where that value is missing.

If $\hat{g}$ denotes such an estimate, we can write

$$g = \hat{g} + \hat{e} \tag{31}$$

from which, because of the independence of $\hat{g}$ and $\hat{e}$, we get

$$\text{var } g = \text{var } \hat{g} + \text{var } \hat{e} \tag{32}$$

where var $\hat{g}$ characterizes the scatter of the estimates $\hat{g}$ around the common mean of g and $\hat{g}$.

From this we deduce

$$\text{var } \hat{g} < \text{var } g \tag{33}$$

from which we find that the introduction of the values of $\hat{g}$ in a series of values of g is a source of heterogeneity since the scattering of the estimate $\hat{g}$ is smaller than that of the observed variable. A solution which obeys the distribution of the initial variable g will be examined later.

2.2.4.1.2 **Example 13.** *Normals of the monthly rainfall totals in January in Belgium. Calculation as a function of the altitude of the station*

a) *Estimation of the parameters*

Table 13 gives the mean January rainfall R_i figures found over a period of 30 years as well as the altitude b_i of the station and the variance var e_i of the error of the mean with respect to the true mean estimated from the observations made at each of the nine stations considered.

The linear model adopted is of the form

$$R_i = \theta_1 + \theta_2 b_i + e_i \quad , \quad i = 1, 2, \ldots, 9 \tag{1}$$

where θ_1 and θ_2 are constants and where the error e_i from the normal is a random quantity of zero mean, whose variance differs from one station to another, but whose values from one station to another are assumed to be independent and distributed following the normal distribution.

Refering to the notations in 2.2.4, we therefore have here

$$x_i = R_i \ , \ u_{i1} = 1 \ , \ u_{i2} = b_i \ , \ v_{ii} = \text{var } e_i \ , \ v_{ii'} = 0 \tag{2}$$

$$\text{for } i \neq i' = 1, 2, \ldots, 9,$$

and

$$v^{-1} = \begin{Vmatrix} v_{11}^{-1} & 0 \cdots & 0 \\ 0 & v_{22}^{-1} & \\ \vdots & & \\ 0 & 0 \cdots & v_{99}^{-1} \end{Vmatrix} \tag{3}$$

which, for the matrix $a = u'v^{-1}u$, leads to :

$$\begin{aligned} a_{11} &= \Sigma u_{i1} v_{ii}^{-1} u_{i1} = \Sigma v_{ii}^{-1} \\ a_{12} = a_{21} &= \Sigma u_{i1} v_{ii}^{-1} u_{i2} = \Sigma v_{ii}^{-1} h_i \\ a_{22} &= \Sigma u_{i2} v_{ii}^{-1} u_{i2} = \Sigma v_{ii}^{-1} h_i^2 \end{aligned} \tag{4}$$

and

$$a^{-1} = (u'v^{-1}u)^{-1} = \frac{1}{a_{11}a_{22} - a_{12}^2} \begin{Vmatrix} a_{22} & -a_{12} \\ -a_{12} & a_{11} \end{Vmatrix} \tag{5}$$

TABLE 13 — Normals of rainfall totals in January R_i for nine Belgian stations (in mm). Corresponding altitudes h_i of the stations (in m), values of the variances var e_i, of the estimations $\hat{R}_i$ and of the difference $R_i - \hat{R}_i$

Stations i	R_i	h_i	var e_i	$\hat{R}_i$	$R_i - \hat{R}_i$
1	125	672	98	130	−5
2	68	50	29	66	2
3	73	120	34	73	0
4	136	370	117	99	37
5	70	236	30	85	−15
6	65	10	27	62	3
7	61	4	23	61	0
8	97	240	59	86	11
9	67	100	28	71	−4

As in addition

$$u'v^{-1}x = \begin{Vmatrix} \Sigma u_{i1} v_{ii}^{-1} R_i \\ \Sigma u_{i2} v_{ii}^{-1} R_i \end{Vmatrix} = \begin{Vmatrix} \Sigma v_{ii}^{-1} R_i \\ \Sigma v_{ii}^{-1} h_i R_i \end{Vmatrix} \tag{6}$$

by putting $\beta = u'v^{-1}x$, which is

$$\beta_1 = \Sigma v_{ii}^{-1} R_i \quad \text{and} \quad \beta_2 = \Sigma v_{ii}^{-1} h_i R_i \tag{7}$$

we obtain the solution

$$\hat{\theta} = a^{-1}\beta = \frac{1}{a_{11}a_{22} - a_{12}^2} \begin{Vmatrix} a_{22}\beta_1 - a_{12}\beta_2 \\ -a_{12}\beta_1 + a_{11}\beta_2 \end{Vmatrix} \tag{8}$$

with

$$\text{var}\,\hat{\theta} = a^{-1} \tag{9}$$

From Table 13, with (4) and (7) we obtain :

$$\begin{aligned} a_{11} &= 1/98 + 1/29 + \cdots + 1/28 = 0{,}24915 \\ a_{12} &= 672/98 + 50/29 + \cdots + 100/28 = 31{,}32 \\ a_{22} &= (672)^2/98 + (50)^2/29 + \cdots + (100)^2/28 = 9482 \\ \beta_1 &= 125/98 + 68/29 + \cdots + 67/28 = 18{,}36 \\ \beta_2 &= (125 \times 672)/98 + (68 \times 50)/29 + \cdots + (67 \times 100)/28 = 2881 \end{aligned} \tag{10}$$

From this we get, with (5), (7), (8) and (9),

$$\begin{aligned} \hat{\theta}_1 &= \frac{a_{22}\beta_1 - a_{12}\beta_2}{a_{11}a_{22} - a_{12}^2} = \frac{9482 \times 18{,}36 - 31{,}32 \times 2881}{0{,}24916 \times 9482 - (31{,}32)^2} = \frac{83856{,}6}{1381{,}5} = 60{,}7 \\ \hat{\theta}_2 &= \frac{-a_{12}\beta_1 + a_{11}\beta_2}{a_{11}a_{22} - a_{12}^2} = \frac{-31{,}32 \times 18{,}36 + 0{,}24915 \times 2881}{1381{,}5} = 0{,}1033 \\ \text{var}\,\hat{\theta}_1 &= \frac{a_{22}}{a_{11}a_{22} - a_{12}^2} = \frac{9482}{1381{,}5} = 6{,}86 \\ \text{var}\,\hat{\theta}_2 &= \frac{0{,}24915}{1381{,}5} = 0{,}0001803 \text{ et } \text{cov}(\hat{\theta}_1, \hat{\theta}_2) = \frac{-31{,}32}{1381{,}5} = -0{,}02267 \end{aligned} \tag{11}$$

b) *Significance test*

The quadratic form 2.2.4(24)

$$\hat{Q}=x'v^{-1}x-\hat{\theta}'\beta \tag{12}$$

here becomes

$$\hat{Q}=\Sigma v^{-1}_{ii}\,R_i^2-\hat{\theta}_1\beta_1-\hat{\theta}_2\beta_2 \tag{13}$$

where

$$\Sigma v^{-1}_{ii}\,R_i^2=(125)^2/98+(68)^2/29+\ldots+(67)^2/28=1435{,}1 \tag{14}$$

Using (10) and (11), we derive from this

$$\hat{Q}=1435{,}1-60{,}7\times 18{,}36-0{,}1033\times 2881=23{,}0 \tag{15}$$

a value which for a χ^2 variate with $9-2=7$ degrees of freedom is significant for a level less than 0,01.

In addition, if the estimates $\hat{R}_i$ of the true mean are calculated by means of the relation

$$\hat{R}_i=60{,}7+0{,}1033\,b_i \tag{16}$$

the values of $R_i-\hat{R}_i$ which are deduced from this in Table 13 do not allow us to attribute this disagreement to the fact that the relation of R_i with b_i will not be linear. It is necessary therefore to assume that local sources of heterogeneity exist whose values of var e_i are not accounted for.

c) *Estimates by means of (1) using (11)*

Assuming that we have been able to accept the hypotheses posed in (1), by estimating the true mean by means of relation

$$\hat{R}_i=\hat{\theta}_1+\hat{\theta}_2 b_i=60{,}7+0{,}1033\,b_i \tag{17}$$

we should have

$$\text{var}\,\hat{R}_i=\text{var}\,\hat{\theta}_1+2\,\text{cov}\,(\hat{\theta}_1\,,\,\hat{\theta}_2)b_i+\text{var}\,\hat{\theta}_2 b_i^2 \tag{18}$$

that is to say, with (11)

$$\text{var}\,\hat{R}_i=6{,}86-0{,}04534\,b_i+0{,}0001803\,b_i^2 \tag{19}$$

This formula shows that the estimation $\hat{R}_i$ is again better in all cases than the initial values R_i and that var $\hat{R}_i$ reaches its minimum value 4,0 for $b=126$.

d) *Note*

It should be noted that, in the preceding example, the independence of the random quantities e_i from one station to another is a simplifying hypothesis. In fact, because of the dimension of the meteorological phenomena which affect a region such as Belgium, it is to be expected that a positive correlation will generally link the monthly totals of water collected, a correlation which will be greater, the shorter the distance between the stations. It follows that the correct solution of the problem here would be to adopt a matrix v where the covariances are not all zero. As the solution to this problem is possible only by means of inversion of the matrix v and this inversion can generally be obtained in certain conditions only by means of computer-programmed methods, such a solution has not been developed here.

The fact that these correlations have been neglected in our example, however, does not imply that the solution obtained is not acceptable. The only reservation which must be made in such cases is that by neglecting the correlation which links the accidental errors, the estimation error given by relation (18) is underestimated as is the value of the quadratic form $\hat{Q}$.

To understand this, it is sufficient to consider the simple case of correlation coefficients which are all equal to 1 and where the variances var e_i are all equal amongst themselves.

In this case, the values of R_i verify without error the relationship

$$R_i = (\theta_1 + e) + \theta_2 b_i \tag{20}$$

which signifies that θ_2 is determined without error, unlike θ_1, which remains affected by the same error as the initial values R_i. It is clear that by neglecting the correlations here, we arrive erroneously at var $\hat{\theta} = \hat{Q} = 0$.

2.2.4.2 Application to the distributions defined by a position parameter μ and a scale parameter σ. Linear estimators of the parameters

In this case, the random variable x is linked to a reduced variate u by a relation of the form

$$x = \mu + u\sigma \tag{1}$$

as the distribution of the reduced variate u is fully known.

If, furthermore, $x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n$ is a simple ordered random series of observations of this variable, we have for the order statistics the relations

$$x_i = \mu + u_i \sigma \ , \quad i = 1, 2, \ldots, n \tag{2}$$

where the u_i denote the order statistics corresponding to the reduced variate u. By taking the average, we obtain

$$E(x_i) = \mu + E(u_i) \cdot \sigma \tag{3}$$

and from this

$$x_i - E(x_i) = [u_i - E(u_i)]\sigma \tag{4}$$

The result is that if we denote by v the covariance matrix of $e_i = x_i - E(x_i)$, that is to say, the order statistics x_i, and by w, the covariance matrix of $d_i = u_i - E(u_i)$, that is to say, of the order statistics u_i, we have

$$v = \sigma^2 \cdot w \tag{5}$$

Returning then to matrix notation, it follows that if the vectors $(x_1, x_2, \ldots, x_n)$, (μ, σ) and $(e_1, e_2, \ldots, e_n)$ are denoted by x, θ and e and the matrix $\| 1 \ E(u_i) \|$ by u, we have again the equation

$$x = u\theta + e \tag{6}$$

whose solution by the least squares method is written

$$\hat{\theta} = (u' v^{-1} u)^{-1} u' v^{-1} x \tag{7}$$

thus, by means of (5), after simplification by σ^2,

$$\hat{\theta} = (u' w^{-1} u)^{-1} u' w^{-1} x \tag{8}$$

It follows that as soon as the matrix u and the matrix w are known, i.e. the mathematical expectation and the covariance matrix of the order statistics of the reduced variate, we can calculate the linear estimators of the parameters μ and σ.

As, in addition 2.2.4(20) here becomes

$$\text{var}\,\hat{\theta} = (u'v^{-1}u)^{-1} = \sigma^2(u'w^{-1}u)^{-1} \tag{9}$$

the efficiency of the estimators obtained can be directly calculated.

In the case of the normal distribution, μ denotes the mean and σ, the standard deviation, and the linear estimator obtained for σ has enabled Shapiro and Wilk to develop a test for normality for series of observations of sample size $n = 3(1)50$. This test will be examined later.

In the case of the double exponential distribution, Lieblein has calculated the matrices u and w for the sample size $n = 2, 3, 4, 5$ and 6. Applications of these results have been given by Thom in Technical Note N° 81. As has been already mentioned in 2.2.3.1, we should note, however, that the efficiency of the estimator of σ obtained in this way does not exceed 77%.

2.2.4.3 Estimators with polynomial coefficients

The generalized application of the method of linear estimators as we have just indicated comes up against the fact that as soon as the sample size of the series of observations increases, the corresponding matrix w becomes practically incalculable.

Linear estimators of a more simple type have been introduced by Downton taking into consideration the variables

$$R_s = \Sigma i^s x_i, \quad s = 0, 1, 2, \ldots, N \tag{1}$$

where x_i are still the order statistics of the variable observed.

With 2.2.4.2(2), we obtain

$$R_s = \Sigma i^s \mu + \Sigma i^s u_i \sigma \tag{2}$$

and similarly

$$E(R_s) = \Sigma i^s \mu + \Sigma i^s E(u_i)\sigma \tag{3}$$

from which we get

$$R_s - E(R_s) = \Sigma i^s \left[u_i - E(u_i)\right]\sigma = b_s\sigma \tag{4}$$

If the vectors $(R_0, R_1, \ldots, R_N)$,(μ,σ),$(b_0, b_1, \ldots, b_N)$ are then denoted by R, θ and b, the matrix $\|\Sigma i \Sigma i^s E(u_i)\|$ and the matrix $\|E(b_r b_s)\|$ by u and w respectively, we have the equation :

$$R = u\theta + b\sigma \tag{5}$$

whose solution can again be written

$$\hat{\theta} = (u'w^{-1}u)^{-1}u'w^{-1}R \tag{6}$$

and for which we also have

$$\text{var}\,\hat{\theta} = \sigma^2(u'w^{-1}u)^{-1} \tag{7}$$

which enables us once more to calculate the efficiencies of the estimators.

The advantage of the method rests in the fact that the existence of certain recurrent relations between the mathematical expectations and the covariances of the order statistics makes the calculation of the matrices u and w relatively easy. In addition, the efficiency of the estimators obtained for μ and σ rapidly increases with N, so that we can limit ourselves to a small number of components of the vector R.

In particular, in the case of the double exponential distribution, by making N = 2, Downton has obtained estimators whose efficiency rapidly becomes satisfactory for increasing sample sizes. For the application of his method the reader should refer to his publication which contains an example. Note, however, that the distribution function of the reduced variate considered by Downton is $G(u) = 1 - F(u)$. It follows that, in the formulae, the order statistics of the double exponential distribution are arranged in decreasing order, that is to say we have : $x_1 \geqslant x_2 \geqslant \ldots \geqslant x_n$.

In the case of the normal distribution, Downton has given a linear estimator of σ with N = 1, which has been used by D'Agostino to construct a test for normality applicable to series of sample size $n \geqslant 50$. This test will also be examined later.

2.2.4.4 Reference works consulted

The application of differential calculus to the calculation of the variance of functions depending on random variables is justified in reference [8]. We should recall in this connection that this application is valid only if, in the total differential of the function, the derivatives of order greater than the first are negligible with respect to those of the first order. In particular, this is achieved asymptotically in the cases which are considered here, as soon as the first derivative is not zero and the other derivatives are finite, because of the fact that the variances of the estimators of the parameters tend towards zero when the sample size of the series of observations increases indefinitely.

Moreover, the problem of estimating the standard deviation of the normal distribution without bias is dealt with in references [8] and [16]. Our presentation of the theory of sufficient statistics and the method of maximum likelihood is principally taken from [8] and [11], the case of the gamma distribution and the double exponential distribution from [17] and [18] respectively and the method of estimation by least squares from [8].

The linear estimators of the parameters of the double exponential distribution have been introduced by Lieblein in [19] and the results obtained by Downton in this area for the normal and the double exponential distribution are published in [20] and [21].

The data for examples 9, 10 and 11 are taken from the Uccle archives, those for example 13 from [22] and example 12 is given in [23].

As regards the notation of the matrix equations which appear in the least squares method, note that we have not considered it useful to use special notation, the introduction of matrices being made, as in [1], without ambiguity.

2.2.5. On the choice of the distribution to be fitted

As has been stated in the preceding examples, fitting distribution functions to series of observations generally leads to more accurate results than those derived from the non-parametric properties of the order statistics. Therefore parametric estimation should be preferred to non-parametric estimation whenever this is possible. It is, on the other hand, very clear that the best accuracy obtained by parametric estimation is real only if the shape of the fitted distribution closely matches that which governs the distribution of the observations.

In practice, it must be recognized that the actual shape of the distribution function of the observations is not generally known and experience shows that the fact that a climatological datum is an average or an extreme value does not necessarily mean that its distribution is represented by a normal or by a double exponential distribution. It often happens, as a matter of fact, that the populations from which we take the elements whose averages or extreme values are being determined are in reality mixtures of populations composed in such a way that the final distribution of the averages and extreme values are outside the very general framework in which the normal and the double exponential distribution are used.

It will therefore be seen that parametric estimation is possible only after the function which agrees best with the series of observations has been selected.

To make such a choice, we could work in two ways :

(a) Fit a relatively general function to several parameters and by means of significance tests retain only a minimum number of estimated parameters so as to ensure a satisfactory accuracy of the estimates;

(b) Carry out a goodness-of-fit test with a relatively simple function such as the normal, and in the case of non-goodness-of-fit, make another choice based on the values obtained for the third-and fourth-order empirical moments and on the result given by a graphical representation of the series of observations on probability paper with appropriate scales.

The first method is theoretically the more correct; however, it implies that the fitted function is sufficiently general to meet all the cases which can occur in practice for the element considered.

The second method, on the other hand, is an easier application, as long as an efficient goodness of fit test is available. However, in the case of non-goodness-of-fit, the choice of a probability function based on a simple graphical representation, even completed by the indications taken from the higher-order empirical moments, will always retain an arbitrary aspect.

This is why, whatever the method used, it is always as well to ensure a good agreement a posteriori between the fitted distribution with the series of observations using the properties of non-parametric estimation. In addition, it is desirable to subject the solutions selected to another statistical check when sufficient secondary information is available.

Moreover, we shall see when examining the tests for normality that, apart from use which is made of graphical representation in the second case, the two methods actually follow a common principle.

2.2.5.1 The exponential distribution as a particular case of the gamma distribution

For the exponential distribution we have

$$dF(x\ ,\beta) = dF(u) = e^{-u}du \quad \text{with} \quad u = x/\beta \tag{1}$$

β being a scale parameter, which gives the distribution function

$$F(x\ ,\beta) = F(u) = 1 - e^{-u} = 1 - \frac{dF(u)}{du} \tag{2}$$

Comparison with 2.2.2.2(1) shows that the exponential is a gamma distribution for which $\alpha = 1$. It follows that the likelihood function 2.2.2.2(3) is simplified according to

$$L = -\,n \log \beta - T_2/\beta \tag{3}$$

from which results the sufficient statistic for β,

$$\hat{\beta} = T_2/n \tag{4}$$

with

$$\operatorname{var}\hat{\beta} = \left[-E\left(\frac{\partial^2 L}{\partial \beta^2}\right)\right]^{-1} = \beta^2/n \tag{5}$$

To verify whether a gamma can be identical to an exponential, we note that with $\alpha = 1$, 2.2.2.2(11) and (12) give

$$\operatorname{var}\hat{\alpha} = 1{,}5498/n\ , \quad \operatorname{var}\hat{\beta} = 2{,}5498\beta^2/n \quad \text{and} \quad \operatorname{cov}(\hat{\alpha}\ ,\hat{\beta}) = -\,1{,}5498\beta/n \tag{6}$$

So, the corresponding elements of the inverse matrix for the covariance matrix are

$$[(\operatorname{var}\hat{\alpha})^{-1}\ ,\quad (\operatorname{var}\hat{\beta})^{-1}\ ,\quad (\operatorname{cov}(\hat{\alpha}\ ,\hat{\beta})^{-1}] \equiv n[\,1{,}64524\ ,\ 1/\beta^2\ ,\ 1/\beta] \tag{7}$$

Under these conditions, to verify the identity of a gamma distribution with an exponential distribution of parameter β_0 we can calculate the statistic

$$X = n\,[\,1{,}64524(\Delta\hat{\alpha})^2 + (\Delta\hat{\beta}/\beta)^2 + 2(\Delta\hat{\alpha})(\Delta\hat{\beta}/\beta)] \tag{8}$$

where $\Delta\hat{\alpha} = \hat{\alpha} - 1$ and $\Delta\hat{\beta} = \hat{\beta} - \beta_0$, using the null hypothesis that this statistic has a χ^2 distribution with two degrees of freedom.

An alternative test which may be used here, the *likelihood ratio test*, is based on the fact that if $g(x\ ,\ \hat{\theta})$ is the maximum of the likelihood function of the general distribution, estimated from the series of observations and if $g(x\ ,\ \hat{\theta}_0)$ is the corresponding maximum for the general distribution when one or more parameters have given values, for large samples, the statistic $X = -\ 2 \log l$ with $l = g(x\ ,\ \hat{\theta}_0)/g(x\ ,\ \hat{\theta})$ has approximately a χ^2 distribution with a number of degrees of freedom equal to the difference between the numbers of parameters estimated with each of the likelihood functions, when $g(x\ ,\ \theta_0)$ is the true likelihood function. Large values are significant of the disagreement with the null hypothesis.

Finally, note that, in the case of the exponential, 2.2(5) and (8) become

$$\operatorname{var} \hat{F}_0 = (u_0 \frac{\partial F}{\partial u})^2/n = (u_0\, e^{-u_0})^2/n$$

and

$$\operatorname{var} \hat{x}_0 = {x_0}^2/n \tag{9}$$

2.2.5.1.1 **Example 14.** *The duration of glaze at Brussels*

The analysis of synoptic reports has allowed us to establish the frequency of the duration of glaze given in Table 14. With 2.2.2.2(2) we obtain

$$T_1 = 18{,}6278 \quad \text{and} \quad T_2 = 125 \tag{1}$$

from which, with $n = 60$, 2.2.2.2(6) gives

$$A = \log \frac{125}{60} - \frac{18{,}6278}{60} = 0{,}73397 - 0{,}31045 = 0{,}42352 \tag{2}$$

From this, using 2.2.2.2(9) and (10), we derive

$$\hat{\alpha} = 1{,}329\ , \quad \hat{\beta} = 1{,}5676 \tag{3}$$

Furthermore, from 2.2.5.1(4) we obtain

$$\hat{\beta}_0 = 125/60 = 2{,}0833 \tag{4}$$

Thus, to verify the null hypothesis for an exponential distribution with parameter β_0, we have

$$\Delta\hat{\alpha} = 1{,}329 - 1 = 0{,}329\ , \quad \Delta\hat{\beta}/\hat{\beta}_0 = (1{,}5676 - 2{,}0833)/2{,}0833 = -\ 0{,}2475$$

which leads to the value of the statistic 2.2.5.1(8)

$$X = 60\,[\,1{,}64524(0{,}329)^2 + (0{,}2475)^2 + 2(0{,}329)(-0{,}2475)\,] = 4{,}59 \tag{5}$$

As this value is less than 5,99 , the critical value of the χ^2 variate with two degrees of freedom at the 0,05 level, fitting an exponential distribution with parameter (4) can be accepted.

For the likelihood ratio test, with (1), (4) and $n = 60$, 2.2.5.1(3) gives : $\hat{L}_1 = -\ 104{,}038$ whereas with (1), (3), $n = 60$ and $\Gamma(\hat{\alpha}) = 0{,}89350$, 2.2.2.2(3) gives : $\hat{L}_2 = -\ 102{,}701$. The value of the statistic is thus : $-\ 2(\hat{L}_1 - \hat{L}_2) = 2{,}67$ with $2 - 1 = 1$ degree of freedom. At the 0,05 level, the critical value being 3,84 , the conclusion remains unchanged.

TABLE 14 — Duration of glaze at Brussels (in hours)

Duration	Average duration	Number of cases
0- 1	0,5	22
0- 2	1	8
1- 3	2	12
2- 4	3	8
3- 5	4	4
4- 6	5	4
5- 7	6	—
6- 8	7	—
7- 9	8	—
8-10	9	—
9-11	10	1
10-12	11	—
11-13	12	1

The accuracy of the chosen fit is, with 2.2.5.1(5), characterized by

$$\operatorname{var} \hat{\beta} = 0{,}0723 \tag{6}$$

In particular, for $F_0 = 0{,}95$, the tables of the gamma distribution from H. C. S. Thom give, for $\alpha = 1$

$$\hat{u}_0 = 2{,}9957 \quad \text{and} \quad \hat{x}_0 = \hat{u}_0 \hat{\beta} = 2{,}9957 \times 2{,}0833 = 6{,}24 \tag{7}$$

whilst from 2.2.5.1(9) we deduce

$$\sigma(\hat{x}_0) = 6{,}24/\sqrt{60} = 0{,}81 \tag{8}$$

Inversely, for $x_0 = 6{,}24$, we have $F_0 = 0{,}95$ and, as the same tables give

$$\frac{\partial F}{\partial u} = 0{,}0500 \quad \text{and} \quad u_0 = 2{,}9957$$

from 2.2.5.1(9) we obtain

$$\sigma(\hat{F}_0) = 2{,}9957 \times 0{,}05/\sqrt{60} = 0{,}019 \tag{9}$$

For the same probability, the empirical estimate leads to

$$\sigma(F_0) = 0{,}028 \tag{10}$$

2.2.5.2 Tests for normality

In order to allow useful comparisons with existing tests, the most general way in which one can design a normality test will first be examined.

As the moments of the normal distribution exist for any order, it can be considered that this distribution belongs to a family of probability functions with an infinity of functionally independent parameters and that, in this family, the normal function is itself defined by particular values for these parameters.

In this situation, if we wish to fit a normal distribution to a series of n independent observations we can, by using the first method given in 2.2.5, proceed with the estimation, which can the assumed to be without bias, of n amongst the infinity of parameters which define the functions of the family and follow with a compatibility test of the estimates obtained with the exact values of these n parameters, under the null hypo-

thesis considered. Note that, in theory, the choice of the n parameters does not matter much since n observations lead at most to n independent estimates.

If $\| v_{ij} \|$ then denotes the covariance matrix of the estimates of the parameters under the null hypothesis and if β_i , $\hat{\beta}_i$, $i = 1, 2, ..., n$ are respectively the exact values of these parameters and their estimates, the goodness of fit of the normal distribution can then be verified by calculating the statistic

$$X_n = \Sigma (\hat{\beta}_i - \beta_i) v_{ij}^{(-1)} (\hat{\beta}_j - \beta_j), \quad i, j = 1, 2, ..., n \tag{1}$$

where $\| v_{ij}^{(-1)} \|$ is the inverse matrix of the matrix $\| v_{ij} \|$. Since we know that under certain conditions, this statistic is approximately distributed as a χ^2 variate with n degrees of freedom, large values of this statistic show the disagreement between the series of observations and the null hypothesis and, as a result, form the critical region of the test.

If the value of the statistic X_n is not significant at the chosen level, the null hypothesis for a distribution following the normal distribution with parameters β_i , $i = 1, 2, ..., n$ can be accepted and we note that in this case we have found a fit without making any *error of estimation* since an overall test has been carried out on a series of values which have all been fixed in advance.

If, on the contrary, the value of the statistic X_n is significant at the chosen level, it is necessary to estimate at least one of the n parameters, starting, for example, with the one for which the ratio $(\hat{\beta}_{i_1} - \beta_{i_1})^2 / v_{i_1 i_1}$ is highest. This first estimate leads to a new statistic X_{n-1} which, as before, has approximately a χ^2 distribution with $(n - 1)$ degrees of freedom and, depending on whether its value is significant or not, the estimate of a second parameter is necessary or the first estimate suffices.

Moreover, if as a preliminary, we have estimated the mean or the variance of the distribution, it is obvious that the number of degrees of freedom of the initial statistic (1) comes down to $(n - 2)$.

Thus we see that by carrying out the test in the way outlined, not only can we ensure the possible goodness of fit of a normal distribution, but in addition, if we reject this, the alternative can be determined which agrees best with the series of observations.

This being the case, the tests for normality available at the present time can be classed into three categories depending upon whether they are based on the calculation of empirical central moments greater than the second order and more precisely on the reduced central moments (moments divided by the corresponding power of the empirical standard deviation); are based on the order statistics deduced from the series of observations; or whether they introduce, for these order statistics, values corresponding to the distribution of the normal distribution.

2.2.5.2.1 Tests for normality based on calculation of the empirical moments

Applied to reduced central moments, the method which has just been described leads to a statistic of the form :

$$X_{n-2} = \Sigma \hat{\beta}_i v_{ij}^{(-1)} \hat{\beta}_j, \quad i, j = 3, 4, ..., n \tag{1}$$

where it can be assumed that the $\hat{\beta}_i$ are the normal transforms of mean zero of these moments, and where the matrix $\| v_{ij} \|$ is the covariance matrix of these transforms.

On the other hand, because of the independence between each empirical central moment of odd order and all the even ordered empirical central moments, and although division of the empirical moments by powers of the same

empirical standard deviation leads to a non-zero correlation between the quotients obtained, we can consider that the correlation between the variables $\hat{\beta}_i$ from the odd-order moments and those linked to the even-order moments is slight. We deduce from this that the statistic X_{n-2} is split up into the sum of two practically independent statistics $X_{1,n-2}$ and $X_{2,n-2}$, one dependent on the odd-order moments and the other on the even-order moments.

It follows from this that the test for normality can be carried out by proceeding separately to a test of the symmetry of the distribution of the observations by means of the statistic $X_{1,\,n-2}$ and to a test of the peakedness of the dispersion of these observations using the statistic $X_{2,\,n-2}$, and by then determining the probability associated with the global test by means of the tests considered in 1.3.

A simplified version of such a test is available in the form of tests in $\sqrt{b_1}$ and b_2 by K. Pearson. In particular, if we put

$$\sqrt{b_1} = m_3/(m_2)^{3/2} \quad \text{and} \quad b_2 = m_4/(m_2)^2 \tag{2}$$

with

$$m_j = \Sigma(x_i - \bar{x})^j/n, \quad j = 2, 3, 4 \tag{3}$$

where $\bar{x} = \Sigma x_i/n$ and x_i, $i = 1, 2, \ldots, n$ are the n observations, following tables published by E. S. Pearson and H. O. Hartley, the work by D'Agostino on the normal transform of $\sqrt{b_1}$ and that of D'Agostino and Tietjen on the quantiles of b_2, the preceding test can be applied whenever $n \geqslant 7$ or 8.

We might think that the use of the simplified form of the test means that part of the information contained in the empirical central moments greater then the fourth order has apparently been neglected. However, because of the correlation linking the third-order empirical moment to the empirical moments of odd-order greater than three and that linking the fourth order empirical moment to the empirical moments of even order higher than fourth, it is only a fraction of this information which is neglected in actual fact.

As regards the general use of the joint test based on the statistics $\sqrt{b_1}$ and b_2, the tables drawn up by Bowman and Shenton for sample sizes extending from 20 to 1000, show that the actual distribution of the Fisher statistic calculated from the probabilities associated with the statistics $\sqrt{b_1}$ and b_2 is more spread out than that for the χ^2 variate with four degrees of freedom. Moreover, if an agreement already exists at the 0,05 level when $n > 50$, such an agreement appears at the 0,01 level only for $n > 1000$. Thus, when using these tables for the χ^2 variate in applying the joint test, we must take into account the fact that the probabilities (significance levels) which they give as corresponding to large values of the Fisher statistic are smaller than the real values.

2.2.5.2.1.1 **Example 15.** *Air temperature at 225mb at Uccle in January. Distribution of daily averages. Test of normality based on the central empirical moments.*

From Table 15 we get

$$\bar{x} = (7{,}0 + 8{,}0 + \cdots + 34{,}0)/45 = 14{,}6$$

$$m_2 = [(7{,}0 - 14{,}6)^2 + (8{,}0 - 14{,}6)^2 + \cdots + (34{,}0 - 14{,}6)^2]/45 = 25{,}4731$$

$$\sqrt{m_2} = 5{,}0471$$

$$m_3 = [(7{,}0 - 14{,}6)^3 + (8{,}0 - 14{,}6)^3 + \cdots + (34{,}0 - 14{,}6)^3]/45 = 167{,}3842$$

$$m_4 = [(7{,}0 - 14{,}6)^4 + (8{,}0 - 14{,}6)^4 + \cdots + (34{,}0 - 14{,}6)^4]/45 = 3832{,}54$$

TABLE 15 — Air temperature at 225mb at Uccle in January. Ordered sample of 45 independent daily averages. Coefficients $a_m(W)$ of the Shapiro and Wilk test and $[(n+1)/2 - m]$ of the D'Agostino test.

m	Daily averages : 200 °K+		m	$(x_{n+1-m} - x_m)$	$a_m(W)$	$(n+1)/2 - m$
1	7,0	34,0	45	27,0	0,3850	22
2	8,0	23,2	44	15,2	0,2651	21
3	8,0	23,1	43	15,1	0,2313	20
4	8,4	22,0	42	13,6	0,2065	19
5	8,4	21,2	41	12,8	0,1865	18
6	9,0	20,0	40	11,0	0,1695	17
7	9,8	18,8	39	9,0	0,1545	16
8	10,0	18,5	38	8,5	0,1410	15
9	10,2	18,4	37	8,2	0,1286	14
10	11,0	18,0	36	7,0	0,1170	13
11	11,0	18,0	35	7,0	0,1062	12
12	11,0	17,9	34	6,9	0,0959	11
13	11,6	17,2	33	5,6	0,0860	10
14	11,8	16,4	32	4,6	0,0765	9
15	12,0	16,4	31	4,4	0,0673	8
16	12,6	15,2	30	2,6	0,0584	7
17	12,8	15,2	29	2,4	0,0497	6
18	12,9	14,9	28	2,0	0,0412	5
19	12,9	14,8	27	1,9	0,0328	4
20	13,0	14,5	26	1,5	0,0245	3
21	13,0	14,5	25	1,5	0,0163	2
22	13,2	14,0	24	0,8	0,0081	1
23	13,5					

and hence

$$\sqrt{b_1} = m_3/(\sqrt{m_2})^3 = 167,3842/(5,0471)^3 = 1,3019$$

$$b_2 = m_4/(m_2)^2 = 3832,54/(25,4731)^2 = 5,9063$$

To obtain the normal transform of $\sqrt{b_1}$, we make

$$Y = \sqrt{b_1}\left\{\frac{(n+1)(n+3)}{6(n-2)}\right\}^{1/2} = 1,3019\left\{\frac{46 \times 48}{6 \times 43}\right\}^{1/2} = 3,8086$$

$$\gamma^2 = -1 + \left\{2\left(\frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)} - 1\right)\right\}^{1/2}$$

$$= -1 + \{4,9562\}^{1/2} = 1,2263$$

$$\delta = 1/\sqrt{\log \gamma} = 3,1311$$

$$\varepsilon = \{2/(\gamma^2 - 1)\}^{1/2} = \{2/0,2263\}^{1/2} = 2,9728$$

$$u_0 = \delta \log [Y/\varepsilon + \{(Y/\varepsilon)^2 + 1\}^{1/2}]$$

$$= 3,1311 \log [1,2811 + \{1,6412 + 1\}^{1/2}]$$

$$= 3,34$$

Using the table for the reduced normal variate, we deduce that, under the null hypothesis, we have

$$P(|u|>3,34)=0,00084<0,001$$

that is to say, the value of $\sqrt{b_1}$ is significant at the 0,001 level in a two-sided test. For the value of b_2, the table of percentiles from D'Agostino and Tietjen (Table A in the annex) shows that, with the null hypothesis, we have

$$P(b_2>5,90)<0,01$$

We can therefore say that the value obtained for b_2 is significant at the 0,02 level in a two-sided test.
To apply the Fisher test (cf. 1.3.1), we can thus make

$$\alpha_1=0,001 \quad \alpha_2=0,02$$

which gives us

$$X=-2\log 0,001-2\log 0,02=(+6,91+3,91)\times 2=21,64$$

Table B of the annex shows that the value obtained for the statistic X is greater than the critical value at the 0,01 level, that is, under the null hypothesis, we have

$$P(X>21,64)<0,01$$

From this we conclude that the hypothesis of normality can be rejected at a significance level greater than 0,01.

2.2.5.2.2 Tests for normality based on the order statistics or on the corresponding values of the normal distribution function

If we consider the order statistics x_m, $m=1, 2, \ldots, n$ derived from the ordered series of observations, the test for the normality of the distribution of the observations must be based on the compatibility of these statistics with their mathematical expectation. If μ and σ are the mean and the standard deviation of the normal function considered, it is obvious that under the null hypothesis the joint distribution of the variables $u_m=(x_m-\mu)/\sigma$ will be that of the order statistics of the standard normal distribution. If, moreover, $\| v_{mm'} \|$ denotes the covariance matrix of these statistics and $\bar{u}_m=E(u_m)$, their mathematical expectations, the test statistic then becomes

$$X=\Sigma(u_m-\bar{u}_m)\, v^{(-1)}_{mm'}\,(u_{m'}-\bar{u}_{m'})$$

$$=\Sigma(x_m-\mu-\sigma\bar{u}_m)\, v^{(-1)}_{mm'}\,(x_{m'}-\mu-\sigma\bar{u}_{m'})/\sigma^2 \qquad (1)$$

where $\| v^{(-1)}_{mm'} \|$ denotes the inverse matrix of the matrix $\| v_{mm'} \|$.

Here again large values of the statistic make up the critical region of the test.

In practice, such a test of normality does not seem to have been developed up to date, probably because the matrix v is known only for small values of n and particularly because of the complexity of calculations which the inversion of the matrix v involves.

On the other hand, a series of tests of normality exists which can be considered as a simplified form of the previous test since in this case the statistic is of the form

$$X=\Sigma a_m x_m / s \qquad (2)$$

in other words, the ratio of a linear function of the order statistics to the empirical standard deviation.

Because of the symmetry of the normal distribution, such a statistic must be transformable to

$$X = \Sigma a_m (x_{n+1-m} - x_m)/s , \quad m = 1, 2, ..., (n-1)/2 \quad \text{or} \quad n/2 \tag{3}$$

which means that it must be sensitive at the same time to asymetric alternative hypotheses and to alternatives where the kurtosis does not conform with the kurtosis of the normal distribution. It thus appears that the statistic (3) can fulfil a similar role to that of statistic (1) considered in 2.2.5.2.1.

To choose the best values to give to the constants a_m , we can see that when these are fixed, the mathematical expectation of the numerator of statistic (3) is, to within a constant factor, equal to the true value of the standard deviation, σ. On the other hand, the best values are those which make var X a minimum so the choice imposed consists of placing the linear estimator of σ with least variance (cf. 2.2.4.2) in the numerator of statistic (3).

With $s^2 = \Sigma (x_m - \bar{x})^2/n$ and by taking $W = X^2$, the statistic of the Shapiro and Wilk test is obtained where small values make up the critical region of the test and for which tables of quantiles have been set out for $n = 3,4,...,50$.

For $n \geqslant 50$, D'Agostino has extended the range of application of the test by taking the Downton linear estimator (to within a constant factor) for the numerator of X in (3), (cf 2.2.4.3), that is to say, by taking

$$a_m = (n+1)/2 - m \tag{4}$$

and by transforming the statistic X in such a way as to obtain a variable of zero mean and unit variance.

Furthermore, by making $a_m = 1/n$, the statistic (3) becomes the statistic of the Geary test for even n; for odd n, this test statistic, with $a_m = 1/n$, takes the form

$$X + |x_{m_0} - \bar{x}|/ns \tag{5}$$

where $m_0 = (n-1)/2 + 1$, that is to say, x_{m_0} is the median of the series of observations.

Finally, with $a_1 = 1$, $a_m = 0$ for $m \neq 1$ and $s^2 = \Sigma (x_m - \bar{x})^2/(n-1)$, the statistic (3) is identical to that of David-Hartley-Pearson.

In the same category, several tests similar to the Kolmogorov test exist; we shall mention two of them.

In the first, the Lilliefors test, for each order statistic, the difference between the value of the fitted normal distribution function $\hat{F}(x_m)$ and the fraction m/n is considered, and the test statistic D is the maximum of the absolute value of the differences obtained in this way :

$$D = \text{Max} \left| \hat{F}(x_m) - m/n \right| \tag{6}$$

In the second, the Kuiper-Louter-Koerts test, the statistic V is the width of the sample of differences considered in the previous test :

$$V = \text{Max} \left[\frac{m}{n} - \hat{F}(x_m) \right] - \min \left[\frac{m}{n} - \hat{F}(x_m) \right] \tag{7}$$

2.2.5:2.2.1 **Example 15** (continued). *Application of tests for normality based on the order statistics*

(a) Using the Shapiro and Wilk coefficients (Table C in the annex), we obtain from Table 15 :

$$\Sigma a_m(W)(x_{n+1-m} - x_m) = 0{,}3850 \times 27{,}0 + 0{,}2651 \times 15{,}2 + \ldots + 0{,}0081 \times 0{,}8 = 32{,}4580$$

In example 15, we found $m_2 = 25{,}4731$, so the Shapiro and Wilk statistic takes the value

$$W = \frac{[\Sigma a_m(W)(x_{n+1-m} - x_m)]^2}{n m_2} = \frac{(32{,}4580)^2}{45 \times 25{,}4731} = 0{,}919$$

From the table of quantiles of W (Table D in the annex) for $n = 45$, we derive

$$P(W < 0{,}919) < 0{,}01$$

that is to say that, in a one-sided test, the value obtained is significant at the 0,01 level.

(b) For the D'Agostino statistic, Table 15 leads to

$$\Sigma [\, (n+1)/2 - m \,] (x_{n+1-m} - x_m) = 22 \times 27{,}0 + 21 \times 15{,}2 + \ldots + 1 \times 0{,}8 = 2719{,}3$$

from which, using the values obtained in example 15, we obtain

$$X = \frac{\Sigma[(n+1)/2 - m](x_{n+1-m} - x_m)}{n^2 \sqrt{m_2}} = \frac{2719{,}3}{(45)^2 \times 5{,}0471} = 0{,}266066$$

The D'Agostino transform then gives the value

$$Y = \frac{\sqrt{n}(X - 0{,}28209479)}{0{,}02998598} = \frac{6{,}7082 \times (-0{,}016029)}{0{,}02998598} = -3{,}59$$

for which the table of quantiles of Y (Table E in the annex) shows that we have

$$\text{Prob}\,(Y < -3{,}59) < 0{,}01$$

in other words, in a two-sided test, the value obtained for Y is significant at the 0,02 level.

(c) Using Table 15, the Geary statistic here becomes

$$X = \frac{\Sigma(x_{n+1-m} - x_m) + |x_{23} - \bar{x}|}{n\sqrt{m_2}} = \frac{27{,}0 + 15{,}2 + \cdots + 0{,}8 + |13{,}5 - 14{,}6|}{45 \times 5{,}0471}$$

$$= \frac{168{,}6 + 1{,}1}{227{,}12} = 0.7472$$

The table of quantiles of the Geary statistic shows that we have

$$\text{Prob}\,(X < 0{,}747) \cong 0{,}05$$

From this we deduce that in a two-sided test, the value obtained for the statistic is significant at the 0,10 level.

(d) With the same data, the David statistic gives

$$X = \frac{x_n - x_1}{\sqrt{m_2}} \times \sqrt{\frac{n-1}{n}} = \frac{27{,}0}{5{,}0471} \times \sqrt{\frac{44}{45}}$$

$$= 5{,}3496 \times 0{,}9888 = 5{,}29$$

The table of quantiles for the David statistic enables us to write

$$\text{Prob}\,(X > 5{,}29) \cong 0{,}95$$

From this we conclude that in a two-sided test, the value obtained for the statistic is also significant at the 0,10 level.

(e) For the Lilliefors and Kuiper tests, we shall limit ourselves to noting that for the first $D = 0{,}119$ and for the second $V = 0{,}173$, values which are significant for a level of about 0,10.

2.2.5.2.3 On the choice of the test for normality

The diversity of the answers given by using each of the tests performed on the same example immediately poses the question of which is the best test.

The results obtained show indeed that, if the usual significance level of 0,05 is adopted, the null hypothesis is rejected by the test based jointly on the central moments of the third and fourth order as well as by the Shapiro and Wilk and D'Agostino tests, but that it is accepted by the others.

As we have seen for each of the two classes of test, none of these is capable of taking full advantage of the information contained in the series of observations. On the other hand, in the second cartegory, the fact that the statistics used by Shapiro-Wilk and by D'Agostino are constructed with linear estimators of the standard deviation, with least variance or asymptotically with least variance, should in general give them a greater power than the simplified versions of Geary and David.

This conlusion extends to the tests derived from the Kolmogorov test, since in this last case, consideration of only the extreme value in a series of values is also, in general, abandonment of part of the information.

However, this does not prevent one of these last tests being more powerful than the first for a particular alternative. Specifically, in the case of the David test when the alternative is the uniform law ($F(x) = x/a$ for $0 \leqslant x \leqslant a$), the difference $(x_n - x_1)$ is a sufficient statistic and the David statistic becomes, as a result, the most powerful test.

In the general case of a non-specified alternative, the choice therefore lies between the Shapiro-Wilk and D'Agostino tests and the test based on the central moments of the third and fourth order.

Although the study of the power of these various tests relative to a series of alternatives seems to show a slight advantage for the first two tests, the simultaneous use of the first and the last remains desirable since, if the hypothesis of normality is rejected, the choice of the alternative can be based on the values found for the coefficients $\sqrt{b_1}$ and b_2.

2.2.5.3 On the choice of the alternative when the hypothesis of normality is rejected

So as to allow an alternative to be adopted when the hypothesis of normality is rejected, we give below the reduced central moments of the third and fourth orders : $\sqrt{\beta_1} = \mu_3/\mu_2^{3/2}$ and $\beta_2 = \mu_4/\mu_2^2$ for distributions which have been considered previously. The observed variable has been denoted by x, the reduced variate by u.

Reduced moments $\sqrt{\beta_1}$ and β_2 *of some general distributions*

Distributions	$\sqrt{\beta_1}$	β_2
a. *Distributions linked with the standard normal :*		
$dF(u) = (2\pi)^{-1/2} \exp(-u^2/2)du$		
– Normal distributions : $x = \mu + \sigma u$	0	3
– Log-normal distributions :		
$\log x = \alpha + \beta u,\ \log \omega = \beta^2/2$	$(\omega^2 + 2)\sqrt{\omega^2 - 1}$	$3 + (\omega^2-1)(\omega^6+3\omega^4 + 6\omega^2+6)$
$\beta = 0{,}14 \quad \omega = 1{,}01$	0,43	3,33
$= 0{,}44 \quad = 1{,}10$	1,47	7,08
$= 0{,}90 \quad = 1{,}50$	4,75	60,6
$= 1{,}18 \quad = 2$	10,39	429
b. *Gamma distributions :*		
$dF(u) = \frac{u^{\alpha-1}}{\Gamma(\alpha)} e^{-u} du,\ x = \beta u$	$2/\sqrt{\alpha}$	$3(1 + 2/\alpha)$
$\alpha = 0{,}1$	6,32	63
$= 0{,}5$	2,83	15
(exponential) $= 1$	2	9
$= 2$	1,41	6
$= 10$	0,63	3,6
c. *Distributions connected with the double exponential :*		
$F(u) = \exp[-e^{-u}]$		
Type I Fisher-Tippett asymptotic distribution :		
$x = \mu + \sigma u$	1,396	5,4
Weibull's distribution : $F(u) = 1 - \exp[-e^{-u}]$,		
$\log x = \alpha - \beta u$		
$\beta = 0{,}1$	−0,64	3,57
$= 0{,}5$	0,63	3,25
(exponential) $= 1$	2	9
$= 2$	6,62	87,72

For Weibull's distribution, note that we have

$$\sqrt{\beta_1} = [\Gamma(1 + 3\beta) - 3\Gamma(1 + 2\beta)\Gamma(1 + \beta) + 2\Gamma^3(1 + \beta)]B^3(\beta)$$

$$\beta_2 = [\Gamma(1 + 4\beta) - 4\Gamma(1 + 3\beta)\Gamma(1 + \beta) + 6\Gamma(1 + 2\beta)\Gamma^2(1 + \beta) - 3\Gamma^4(1 + \beta)]B^4(\beta)$$

with

$$B(\beta) = [\Gamma(1 + 2\beta) - \Gamma^2(1 + \beta)]^{-1/2}$$

So the exponential distribution appears as a particular case of both the gamma and the Weibull distributions. Similarly, in the log-normal, gamma and Weibull distributions, we see that a distribution exists where the values of $\sqrt{\beta_1}$ and β_2 are very close to those of the first Fisher-Tippett asymptotic distribution.

The result of this is that, beyond conditions imposed by the definition space and by certain special properties of the observed variable, it will often be ease of calculation which determines the final choice of the fitted distribution.

It is appropriate to recall a characteristic property of the family gamma distributions when the parameter β is fixed. In this case, they have the well known additive property of the χ^2 laws, that is to say that the distribution of the sum of two independent gamma variables with parameters α_1 and α_2 is a gamma distribution with parameter $\alpha = \alpha_1 + \alpha_2$. It follows that, in this respect, the gamma distribution, for non-negative variables, is similar to the normal distribution for non-limited variables.

2.2.5.4 On graphical representation on probability paper. Final goodness-of-fit of a fitted distribution

In the case where the relationship between the variable of the fitted distribution x, and the reduced variate u is a linear relation of the form

$$x = \mu + \sigma u \tag{1}$$

graphical representation of this relationship in a co-ordinate axis system with linear scales in x and in u is a line. If, in addition, the corresponding values $F(u)$ of the distribution function of the reduced variate u are marked on the co-ordinate axis u, a graphical representation with a linear scale in x and a probability scale in $F(u)$ is obained with the relation $x = x(F)$ represented by a line. When we then proceed to the graphical representation of the points $[x_m, F(u_{0m})]$ where u_{0m} is a central value of the probability distribution of the order statistic u_m, we get a simultaneous representation of the series of observations and the distribution to be fitted. In particular, if we take for u_{0m} the mean $\bar{u}_m = E(u_m)$ of the order statistic u_m, the test agreement corresponding to the case where the test statistic 2.2.5.2(1) is zero will be characterized by the fact that all the points $[x_m, F(\bar{u}_m)]$ are aligned on the line of equation (1).

So the joint representation of the distribution which has been fitted and the series of observations has a double usefulness. In fact, it permits :

(a) when the fitted distribution is rejected by a goodness-of-fit test, the determination of whether the reason for the rejection results from the form of the fit, or from the presence of aberrant observations, or from both of these;

(b) when the fitted distribution has been accepted, the easy reading of the probability F associated with any particular value of the observed variable x.

It is sometimes customary to mark the scale of the probabilities in terms of the mean recurrence interval (cf. 2.1.9.5), in other words, to indicate the value of T given by either of the relationships

$$T = 1/F \quad \text{for} \quad F \leqslant 0{,}5 \quad \text{and} \quad T = 1/(1-F) \quad \text{for } F \geqslant 0{,}5 \tag{2}$$

In practice, use of the average $\overline{u}_m$ to represent the observations of the series can be shown to be inconvenient because it is necessary to use a table of mathematical expectations for the order statistics.

When we wish to avoid this disadvantage, we can for $F(u_{0m})$ take the value

$$E[F(u_m)] = m/(n+1) \tag{3}$$

which we know is the mathematical expectation of the probability F associated with the order statistic u_m (cf. 2.1.2). Such a solution has the advantage that, in the vicinity of the median, it is practically equivalent to the first method and that, at the extremes, the difference between the values given by each of the methods is small relative to the variability of the extreme values.

In short, in order to judge the final goodness-of-fit of the distribution which has been fitted to the series of observations, it is desirable to calculate the probabilities attributed to the median and to the extreme values by the fitted function and to verify whether the probabilities obtained are within the corresponding confidence intervals calculated as shown in 2.1.1. The same method can possibly be used to underline the aberrant character of either of the extremes.

Finally, note that because of the asymptotic independence of the distribution of the extremes and the median, we can also carry out a global test using the tests considered in 1.3.

Be that as it may, it is very clear that in the case of rejection of an extreme value, the fit can be accepted only if the main reason for eliminating this value is that it is really doubtful or it is actually linked to a very rare event.

2.2.5.4.1 **Example 15***(continued). Final fitting and graphical representation*

In Figure 4 the vertical axis has been marked in 4K intervals from 204K to 240K. The horizontal axis has been marked in values of 100 F = 1, 5, 10, 20, 50, 80, 90, 95, 98, 99, 99,8, for which we have T = 100, 20, 10, 5, 2, 5, 10, 20, 50, 100 and 500 which for the standard normal correspond respectively to the values of $u = -2{,}33$, $-1{,}64$, $-1{,}28$, $-0{,}84$, 0 , 0,84, 1,28 , 1,64 , 2,05 , 2, 33 and 2, 88.

This being so, by giving to the observations 207,0K, 208,0K, ..., 234,0K of Table 15 the probabilities $F = 1/46 = 0{,}022$, $2/46 = 0{,}043$, ..., $45/46 = 0{,}978$ respectively, the graphical representation shown is obtained (on normal probability paper).

On the other hand, with $\overline{x} = 214{,}6$K and $\hat{\sigma}^2 = nm_2/(n-1) = 45 \times 25{,}4731/44 = 26{,}0520$, the fitted normal function can, because of 2.2.1(2) and 2.2.1(14), be represented by equation (1) :

$$\hat{x} = 214{,}6 + 5{,}13\,u \tag{1}$$

which gives for F = 0,02 and F = 0,98 ($u = \pm 2{,}05$)

$$\hat{x} = 214{,}6 \pm 5{,}13 \times 2{,}05 = 204{,}1 \quad \text{and} \quad 225{,}1 \tag{2}$$

By joining the two points [204,1 , F = 0,02] and [225,1 , F = 0,98], the line is found which represents the fitted normal distribution.

It thus appears that rejection of the normal distribution is explained by a systematic deviation from the graph of the series of observations with respect to that of the fit and perhaps also by a divergent extreme observation.

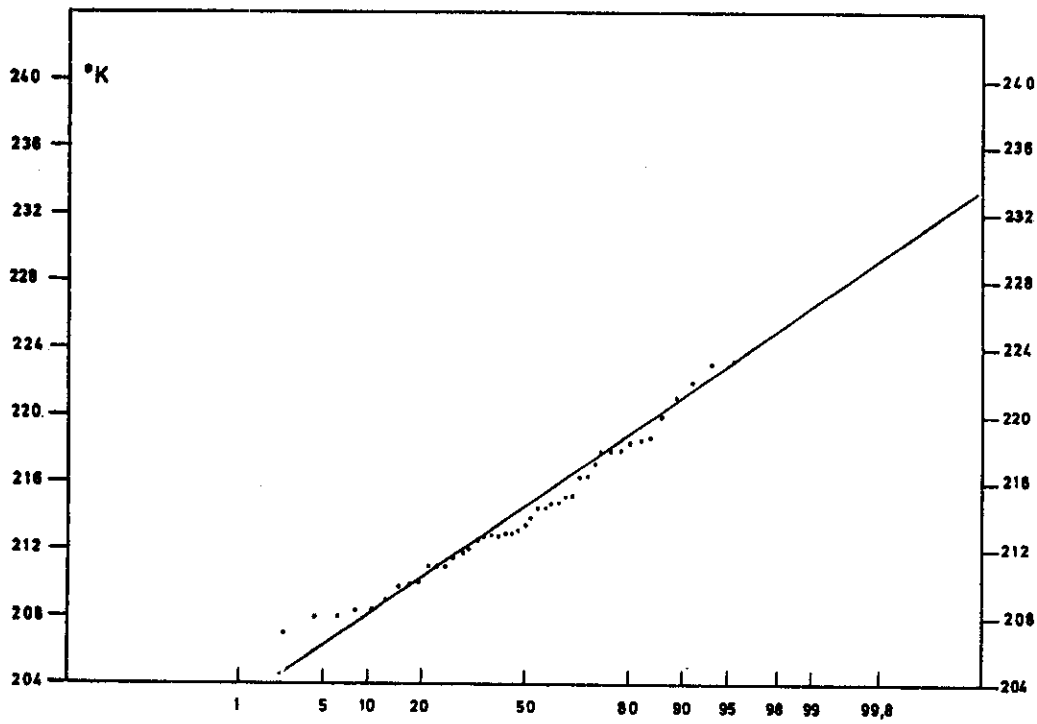

FIG.4 — Air temperature at 225 mb at Uccle, in January. Graph of the fitted normal distribution and the series of observations

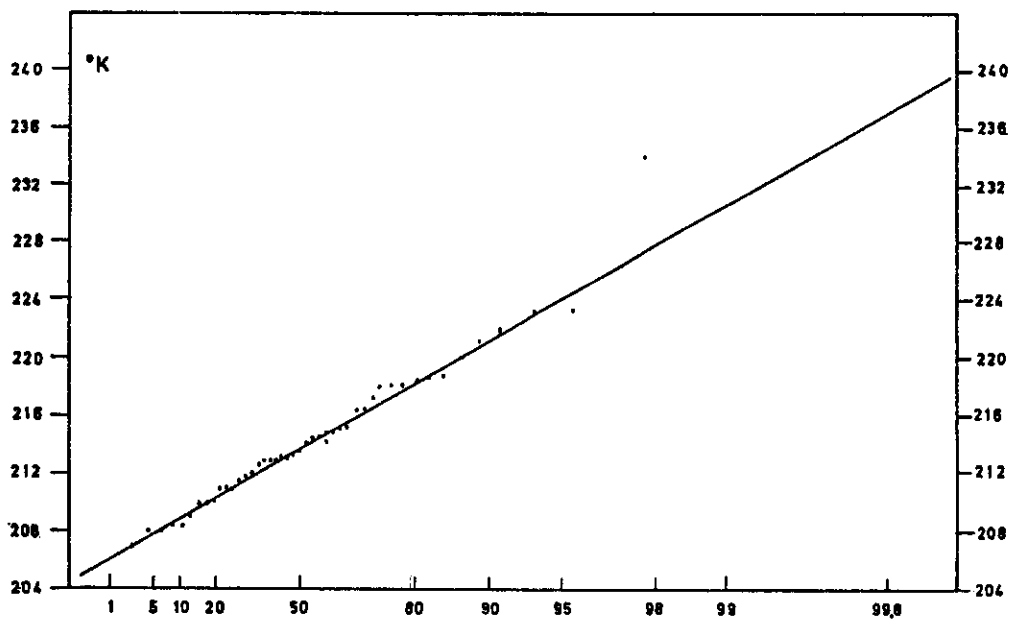

FIG. 5 — Air temperature at 225mb at Uccle in January. Graph of the fitted double exponential distribution and of the series of observations

To obtain a better fit, we note that the values obtained for $\sqrt{b_1}$ and b_2 , i.e. 1,30 and 5,91 , are very close to the values of $\sqrt{\beta_1}$ and β_2 of the double exponential distribution, that is to say 1,396 and 5,4.

The fitting of such a function to the series of observations by means of the method described in 2.2.3.1 leads to

$$\hat{\mu}=212,3 \quad \text{and} \quad \hat{\sigma}=3,9 \tag{3}$$

The graph of the new distribution which has been fitted and the series of observations on double exponential probability paper (Figure 5) shows an excellent agreement this time with, however, the same peculiarity for the last value.

To make sure that this value is allowable, we note that we have

$$\hat{x}=212,3 \ +3,9\,u \tag{4}$$

where u denotes the reduced variate of the double exponential distribution.

The result is that for $x=234,0$, we have $u=5,564$, which corresponds to $F(u)=0,99617$. In a series of 45 observations, the distribution function of the last value is F^{45}; the probability associated with $x_{45}=234,0$ is therefore : $(0,99617)^{45}=0,842$, a value which is within the confidence interval (0,025 , 0,975) at the 0,05 level. There is therefore no reason to suspect the last value and the fit (4) can be accepted.

We should also point out that the choice of the double exponential distribution is preferable to that of one of the other possible forms, because of the fact that, for the first function, the space defining the reduced variate is not limited.

In order to apply the goodness-of-fit test based on the two extreme values and the median, we have in the case of the fitted normal function (1) successively :

(*a*) for the first value :

$$x_1=207,0\ , \quad u_1=(7,0-14,6)/5,13=-1,48\ , \quad F(u_1)=0,0694;$$

the probability associated with the first value is thus

$$(1-F)^{45}=0,0393 \quad \text{and for a two sided test} \quad \alpha_1=0,0393 \times 2=0,079$$

(*b*) for the median :

$$x_{23}=213,5\ , \quad u_{23}=(13,5-14,6)/5,13=-0,21\ , \quad F(u_{23})=0,4168$$

As in the case of the probability associated with the median, we have also $\sigma^2(F)=0,5 \times 0,5/47$, that is to say $\sigma(F)=0,0729$, the normal approximation for the distribution of $F(u_{23})$ make the value of the standard normal variate correspond to 0,4168 :

$$(0,4168-0,5)/0,0729=-1,14$$

We thus have, using tables of the normal;

$$P[F(u_{23})<0,4168]=0,1271,$$

that is, in a two-sided test

$$\alpha_2=0,1271 \times 2=0,254$$

(*c*) for the last value ;

$$x_{45}=234,0\ , \quad u_{45}=(34,0-14,6)/5,13=3,78\ , \quad F(u_{45})=0,9999216;$$

the probability associated with the last value is thus

$$(F)^{45}=(0,9999216)^{45}=0,9965$$

which is, in a two-sided test,

$$\alpha_3=(1-0,9965) \times 2=0,007.$$

For the global test, the Fisher statistic here becomes

$$X = -2 \Sigma \log \alpha_i = 2(2{,}53831 + 1{,}37042 + 4{,}96185) = 17{,}74$$

for which the table of the χ^2 distribution with six degrees of freedom shows that we have

$$P(\chi^2 > 17{,}74) < 0{,}01$$

In the case of the fitted double exponential function (4), we find in the same way

$$\alpha_1 = 0{,}795 \ , \alpha_2 = 0{,}787 \quad \text{and} \quad \alpha_3 = 0{,}318$$

The Fisher statistic is then

$$X = -2 \Sigma \log \alpha_i = 2(0{,}22941 + 0{,}23953 + 1{,}14570) = 3{,}23$$

for which the table for the χ^2 distribution with six degrees of freedom shows that we have :

$$P(\chi^2 > 3{,}23) > 0{,}7$$

The rejection of the normal and the goodness-of-fit of the double exponential distribution are thus confirmed.

2.2.5.5 Reference works consulted

Theory and applications of the maximum likelihood ratio test may be found in [8] and [9].

The account of the normality tests is taken from [24] where an attempt to synthetize tests given in [25] has been made. Disregarding the importance of the problem, this description also seems useful because of the recent developments which have been achieved in this area and because of the illustration it gives concerning the general problem of applying tests of hypothesis. We should add that a very full document on this subject has been prepared by Subcommittee 2 of the Committee ISO/TC 69 and the methods described here have benefited both from work already accomplished and from contacts with Professor E.S. Pearson of this committee.

Moreover, the normal transform of $\sqrt{b_1}$ is given in [26], Table A of the annex comes from [T6] and from [27], Table B of the annex has been drawn up from the tables in [28], the Shapiro-Wilk test and Tables C and D of the annex are taken from [29], the D'Agostino test and Table E have been published in [30] and [31] and the respective powers of the tests have been examined in [32]. Note that the tables in [33] enable us to calculate the probabilities associated with the values of b_2 when n is between 20 and 200 and the probability between 0,999 and 0,001.

Otherwise, the preceding sections can be considered as direct applications of the theory and principles given in earlier sections.

As regards the examples, example 14 is taken from [34] and example 15 from [35].

The tables which appear in the annex have been reproduced with the kind permission of their authors.

2.2.6 The parametric estimation in the case of distributions of discrete variables

When the random variable is a frequency, the values that it takes are whole numbers 0, 1, 2, ..., k, $k + 1$, ... ; in which case, it is said to be a discrete variable. The distribution function of such a variable is a discontinuous function since it progresses from 0 to 1 in steps while the variable successively takes the values 0, 1, ..., k, $k + 1$, ...

In climatology, chance phenomena such as rain, snow, hail, storm, glaze, fog, etc. are examples of events which can be counted, generally in the form of days of rain, snow, hail, ... and the series of numbers of days thus obtained pose the problem of seeking the probability function which governs the statistical distribution of these frequencies.

In spite of the very wide use which is made of fitting discrete distributions to the series of numbers of days of particular chance phenomena, two criticisms must always precede the use of such distributions.

The first concerns the very justification of their use. As a matter of fact, the discrete distributions are mainly intended to give a probability to the various numbers or times that a particular phenomenon can occur during a series of observations and the fact of having distributed the observed values of a meteorological variable into classes of the same interval does not imply that the distribution obtained can be represented adequately by a discrete distribution function. For example, when the variable observed is a period expressed in whole numbers of time units, this does not in any way alter the fact that it is a continuous variable. Furthermore, it seems justified to ask whether this remark does not apply to the statistics of consecutive days with special meteorological characteristics, since the appearance of such a day is linked mainly to the existence of a definite type of weather whose life-time often exceeds that of a day.

On the other hand, in order to simplify the calculations, we can obviously still use approximate values given by certain continuous asymptotic forms of discrete distribution functions, such as the normal or the gamma distribution.

The second problem which is often encountered in the fitting of discrete distributions concerns the actual quality of the series. When considering chance phenomena observed during relatively short periods, such as storm, hail, glaze, etc., the majority of the series of these observations are affected by the inevitable irregularity with which they are observed by voluntary observers. In such cases, it is therefore always wise to ensure that the possible deficiencies of an observer are compensated by a sufficiently general definition of the phenomenon, as for example a definition at the scale of all the stations of one region, rather than at the purely local level of one special station.

2.2.6.1 Notation and generalities. The principle of minimum χ^2.

For the sake of convenience, in the following, we shall discriminate between the positive observations, that is to say those during which the phenomenon considered does occur, and the negative observations, that is to say, those where this phenomenon does not occur.

If x then denotes the frequency of the positive observations and if $k = 0, 1, 2, \ldots, n$ are the various possible values of this frequency, n, the total number of observations possibly being arbitrarily large, we put :

$$p_k = P(x = k) \tag{1}$$

$$P_k = P(x \geqslant k) = \sum_k^n p_j \quad ; \quad Q_k = P(x \leqslant k) = \sum_0^k p_j \tag{2}$$

$$S_k = \sum_k^n j p_j \tag{3}$$

from which we obtain

$$P_0 = Q_n = \sum_0^n p_j = 1 \quad \text{and} \quad Q_k = 1 - P_{k+1} \tag{4}$$

and also

$$E(x) = \sum_0^n j p_j / \sum_0^n p_j = S_0/P_0 = S_0 \tag{5}$$

It follows that the Q_k denote the values of the distribution function of the discrete variable x for the various values of k.

From this we also get

$$\text{var } x = E(x^2) - [E(x)]^2 = \sum_0^n j^2 p_j - S_0^2 \tag{6}$$

More generally, we also have

$$E(x \geqslant k) = S_k/P_k \tag{7}$$

and from this we deduce that the expression

$$l_k = S_k/P_k - (k-1) \tag{8}$$

which generalizes the *Borel mean length*, gives a mean of $[x - (k-1)]$ for $x \geqslant k$. In particular, if x denotes the sequences of consecutive positive observations (consecutive days of storm, hail, ...), l_k gives the average length of the sequences greater than or equal to k, when their length is counted from k.

For $k = 0$, (8) becomes

$$l_0 = S_0 + 1 = E(x) + 1 \tag{9}$$

For the observed frequencies, by denoting by f_k the number of cases when $x = k$, we put in the same way for the number of cases with $x \geqslant k$ or $x \leqslant k$

$$F_k = \sum_k^n f_j \quad \text{or} \quad G_k = \sum_0^k f_j \tag{10}$$

and for the sum of the $x \geqslant k$

$$N_k = \sum_k^n j f_j \tag{11}$$

As a result, the total number of the observed values of the variable x is F_0 and its empirical mean $\bar{x}$ is given by

$$\bar{x} = N_0/F_0 \tag{12}$$

Similarly, the empirical mean lengths are defined by means of the relationship

$$l_k^* = N_k/F_k - (k-1) \tag{13}$$

from which we get

$$l_0^* = N_0/F_0 + 1 = \bar{x} + 1 \tag{14}$$

In the following, the binomial distribution and its most common derivatives are considered and, as regards their fitting, it will be noted that the principles of mathematical statistics which have been used in the case of continuous distributions are completely applicable to the discrete probability distributions.

In particular, estimates which come from the fit of a distribution are valid only on condition that the observed frequencies follow one another independently. Before any fitting, it is therefore necessary to ensure that this hypothesis is completely satisfied by means of the tests described in Chapter 1.

Under these conditions, we use the form taken by the likelihood function expressed in natural logarithms

$$L = \sum_{0}^{n} f_k \log p_k \tag{15}$$

from which, as for continuous distributions, we decide whether the distribution has sufficient statistics or not.

In particular, if, in the case of a distribution with a single parameter α, it appears that the sum N_0 is a sufficient statistic for this parameter, and if we have

$$E(x) = \phi(\alpha) \tag{16}$$

with (12) we obtain

$$E(N_0) = F_0 E(x) = F_0 \phi(\alpha) \tag{17}$$

from which it results that the sufficient estimate of α satisfies the equation

$$N_0 = F_0 \phi(\alpha) \tag{18}$$

As we also get from this

$$\text{var } N_0 = F_0^2 \left[\frac{d\phi(\alpha)}{d\alpha} \right]^2 \text{var } \hat{\alpha} \tag{19}$$

and because of the independence of the F_0 observed values of x, we have further

$$\text{var } N_0 = F_0 \text{ var } x \tag{20}$$

we finally obtain

$$\text{var } \hat{\alpha} = \left[\frac{d\phi(\alpha)}{d\alpha} \right]^{-2} \frac{\text{var } x}{F_0} \tag{21}$$

To verify the goodness-of-fit of a given discrete probability distribution for a series of observations, the K. Pearson statistic is used, which here becomes

$$X = \sum_{0}^{n} \frac{(f_k - F_0 p_k)^2}{F_0 p_k} \tag{22}$$

and for which the number of degrees of freedom is $(n + 1) - 1 = n$ if the probabilities p_k have been fixed in advance or $(n - r)$ degrees of freedom if these probabilities have been calculated from the estimates of r parameters. When interpreting the result, we shall take account, however, of the recommendations given in this connection in 2.1.10.

It should be noted that when no sufficient statistic exists, an alternative method of estimation to the method of maximum likelihood is that known as the minimum χ^2 method, which consists of seeking the values of the parameters which make the statistic (22) a minimum.

The advantage of this last method is that, in addition to good estimates of the parameters, it provides the means of directly ensuring the goodness-of-fit of the fitted distribution. However, as its solutions are no more precise and calculations involved are less convenient, solutions obtained using the maximum likelihood function are to be preferred.

2.2.6.2 The binomial distribution

When p is the probability of a positive observation and in a series of n independent observations, the total number x of positive observations is counted, the binomial distribution gives the probabilities associated with the various possible values of x, that is : 0, 1, 2, ..., n.

With $q = 1 - p$, we have in this case

$$p_k = \frac{n!}{k!(n-k)!} p^k q^{n-k} \tag{1}$$

which are the successive terms of the expansion of the expression $(p+q)^n$. Furthermore, we also have :

$$p_0 = q^n \; ; \; p_k = p_{k-1} \times \frac{p}{q} \times \frac{n-k+1}{k} \; , \; k = 1, 2, ..., n \tag{2}$$

From this, and using the recurrence relation (2), we get

$$\mathrm{E}(x) = \sum_0^n k p_k = np, \quad \mathrm{E}(x^2) = \sum_0^n k^2 p_k = n^2 p^2 + npq$$

$$\mathrm{var}\, x = \mathrm{E}(x^2) - [\mathrm{E}(x)]^2 = npq \tag{3}$$

It also follows from this, using (2), 2.2.6.1(10) and (11), that if F_0 series of n independent observations are carried out, the likelihood function 2.2.6.1(15) becomes

$$\mathrm{L} = n\mathrm{F}_0 \log q + \mathrm{N}_0 \log \frac{p}{q} + \sum_1^n \mathrm{F}_k \log \frac{n-k+1}{k} \tag{4}$$

The result is that N_0 is a sufficient statistic for the parameter p.

With (3) and using 2.2.6.1(18), the sufficient estimate of p is derived from the equation : $N_0 = \mathrm{F}_0 \times np$, that is

$$\hat{p} = \mathrm{N}_0/(n\mathrm{F}_0) \tag{5}$$

Furthermore, as we have

$$\frac{d\mathrm{E}(x)}{dp} = n$$

with 2.2.6.1(21), we obtain

$$\mathrm{var}\, \hat{p} = \frac{1}{n^2} \cdot \frac{\mathrm{var}\, x}{\mathrm{F}_0} = \frac{pq}{n\mathrm{F}_0} \tag{6}$$

In the case where $\mathrm{F}_0 = 1$, where only one series of n observations is carried out and if $x_0 = \mathrm{N}_0$ is the number of positive observations, the sufficient estimate of the probability associated with a positive observation is

$$\hat{p} = x_0/n \tag{7}$$

for which we have

$$\mathrm{var}\, \hat{p} = pq/n \tag{8}$$

As the probability p is linked with no other parameter in this case, it can be considered as an estimate of the probability in the non-parametric case.

In the general case, the error associated with the estimation of the values $Q_k = P(x \leqslant k)$ of the distribution function can be calculated as follows.

From (1), using natural logarithms, we get

$$d \log p_k = (n-k) d \log q + k d \log p \tag{9}$$

that is,

$$dp_k = \left[-\frac{n-k}{q} p_k + \frac{k}{p} p_k \right] dp \tag{10}$$

which gives, with the recurrence relation (2),

$$dp_k = \left[-\frac{n-k}{q} p_k + \frac{n-k+1}{q} p_{k-1} \right] dp \tag{11}$$

It follows that we have

$$dQ_k = d(p_0 + p_1 + \cdots p_k) = -\frac{n-k}{q} p_k dp \tag{12}$$

from which

$$\operatorname{var} \hat{Q}_k = \frac{(n-k)^2}{q^2} p_k^2 \operatorname{var} \hat{p} \tag{13}$$

2.2.6.2.1 **Example 16.** *The frequency per year of months with low rainfall at Uccle*

Table 16 gives for a period of 128 years the number of years f_k during which k months of low rainfall (total rainfall less than normal) have been observed. As the monthly rainfall levels at Uccle can be considered as independent of each other, we can predict that the number per year of low rainfall months has a binomial distribution.

With

$$F_0 = \sum_0^{12} f_k = 128, \qquad N_0 = \sum_0^{12} k f_k = 825 \quad \text{and} \quad n = 12,$$

we have, using 2.2.6.2(5) and (6),

$$\hat{p} = 825/(12 \times 128) = 0{,}53711$$

and

$$\sigma(\hat{p}) = \sqrt{0{,}53711 \times 0{,}46289/(12 \times 128)} = 0{,}0127 \tag{1}$$

The calculation of $\sigma(\hat{p})$ from the estimate $\hat{p}$ involves only a negligible error.

By using relations 2.2.6.2(2), the values of $\hat{p}_k$ shown in the same table are found.

To verify the goodness-of-fit of the sample distribution, calculation of $F_0 \hat{p}_k$ and of the statistic X in 2.2.6.1(22) leads to

$$X = 4{,}38 \tag{2}$$

As this statistic has a χ^2 distribution with $\nu = 10 - 1 - 1 = 8$ degrees of freedom (bearing in mind the groupings in Table 16) for which the critical value at the 0,05 level is 15,5, we conclude that the agreement between the fit and the series of observations is excellent.

TABLE 16. – Distribution of the number k of low-rainfall months per year at Uccle (1833-1960). Values of $\hat{p}_k$ given by the fitted binomial distribution.

k	f_k	F_k	N_k	l_k^*	$\hat{p}_k$	$F_0\hat{p}_k$	$\hat{Q}_k$	$\hat{T}_k$
0	—	128	825	7,45	0,00010	—	0,00010	10.000
1	1	128	825	6,45	0,00135	—	0,00145	690
2	—	127	824	5,49	0,00860	1,29	0,01005	99,5
3	7	127	824	4,49	0,03226	4,26	0,04331	23,1
4	13	120	803	3,39	0,08683	11,11	0,13014	7,68
5	19	107	751	3,02	0,16120	20,63	0,29134	3,43
6	25	88	656	2,45	0,21823	27,93	0,50957	—
7	23	63	506	2,03	0,21704	27,78	0,72661	2,04
8	22	40	345	1,62	0,15740	20,15	0,88401	3,66
9	12	18	169	1,39	0,08117	10,39	0,96518	8,62
10	5	6	61	1,17	0,02826	3,62	0 99344	28,7
11	1	1	11	1,00	0,00586	0,84	0,99940	152
12	—	—	—	—	0,00058	—	0,99998	1.667

On the other hand, as we also have

$$X = (\hat{p} - 0{,}50)^2 / \text{var}\,\hat{p} = (0{,}03711)^2/(0{,}0127)^2 = 8{,}54 \tag{4}$$

a value which is very much greater than 3,84 , the critical value at the 0,05 level, in a χ^2 distribution with one degree of freedom, we deduce that the probability associated with low-rainfall months is greater than that of high-rainfall months. This is in agreement with the fact that the distribution of monthly rainfall levels is not symmetrical.

Finally, from the values obtained for $\hat{p}_k$ we also get, for example, the value of $\hat{Q}_2$:

$$\hat{Q}_2 = \hat{p}_0 + \hat{p}_1 + \hat{p}_2 = 0{,}01005 \tag{5}$$

for which we have, from 2.2.6.2(13) :

$$\sigma(\hat{Q}_2) = \frac{12-2}{0{,}46289} \times 0{,}00860 \times 0{,}0127 = 0{,}0024 \tag{6}$$

For $p = 0{,}01005$, non-parametric estimation, with 2.2.6.2(8), leads to :

$$\sigma(\hat{Q}_2) = \sqrt{0{,}01005 \times 0{,}98995/128} = 0{,}0088 \tag{7}$$

The mean recurrence interval $\hat{T}_2 = 1/0{,}01005 = 99{,}5$ corresponds to $\hat{Q}_2 = P(x \leqslant 2) = 0{,}01005$; furthermore the two-sided confidence interval at the 0,05 level that (6) associates with Q_2, i.e. $0{,}01005 \pm 1{,}96 \times 0{,}0024$, for T_2 leads to the interval

$$67{,}8 \leqslant T_2 \leqslant 187 \tag{8}$$

Similarly, for high values of k, we have

$$T_k = 1/(1 - Q_{k-1}) \tag{9}$$

and notably, for $x \geqslant 7$

$$\hat{T}_7 = 1/(1 - 0{,}50957) = 2{,}04 \tag{10}$$

2.2.6.3 The Poisson distribution

When, in a binomial distribution, the number of observations n is very large and the probability p is very small so that the mean $\alpha = np$ remains finite, relations (1) and (2) in 2.2.6.2 take the following asymptotic forms which define Poisson's distribution

$$p_k = e^{-\alpha} \cdot \frac{\alpha^k}{k!}, \quad k = 0, 1, 2, \ldots \tag{1}$$

from which we obtain

$$p_0 = e^{-\alpha}, \; p_k = p_{k-1} \cdot \frac{\alpha}{k}, \; k = 1, 2, \ldots \tag{2}$$

From this we obtain, via the recurrence (2) :

$$E(x) = \sum_0^\infty k p_k = \alpha, \quad E(x^2) = \sum_0^\infty k^2 p_k = \alpha^2 + \alpha$$

$$\text{var}\, x = E(x^2) - [E(x)]^2 = \alpha^2 + \alpha - \alpha^2 = \alpha \tag{3}$$

It then follows that in the case of a series of observations governed by a Poisson distribution, in which r denotes the highest observed value of x , the likelihood function 2.2.6.1(15) becomes

$$L = -\alpha F_0 + N_0 \log \alpha - \sum_1^r f_k \log k! \tag{4}$$

which shows that N_0 is again a sufficient statistic for the parameter α . We thus deduce, with 2.2.6.1(18), that the estimate of α is given by the equation $N_0 = F_0\, \alpha$, that is to say,

$$\hat{\alpha} = N_0 / F_0 \tag{5}$$

for which, with

$$\frac{dE(x)}{d\alpha} = 1,$$

we obtain, using 2.2.6.1(21)

$$\text{var}\, \hat{\alpha} = \frac{\text{var}\, x}{F_0} = \frac{\alpha}{F_0} \tag{6}$$

Similarly, for the calculation of the error associated with the estimate of $Q_k = P(x \leqslant k)$, we note that, by expressing relation (1) in natural logarithms, we obtain

$$d \log p_k = -d\alpha + \frac{k}{\alpha} d\alpha \tag{7}$$

which, with the recurrence (2), gives

$$dp_k = (-p_k + p_{k-1})\, d\alpha \tag{8}$$

and consequently

$$\sigma(\hat{Q}_k) = p_k \sigma(\hat{\alpha}) \tag{9}$$

2.2.6.3.1. **Example 17.** *The frequency of storms in Belgium in January*

Table 17 gives for a period of 27 years the number f_k of years for which a number k of stormy days was counted in January in Belgium. As storms are very rare in January, we can presume that their frequency follows a Poisson distribution.

With $F_0 = \Sigma f_k = 27$, $N_0 = \Sigma k f_k = 33$, 2.2.6.3(5) and (6) give

$$\hat{\alpha} = 33/27 = 1{,}2222$$

$$\text{var}\,\hat{\alpha} = 1{,}2222/27 = 0{,}04527 \tag{1}$$

or

$$\sigma(\hat{\alpha}) = 0{,}2128 \tag{2}$$

knowing that the calculation of var $\hat{\alpha}$ from the estimate $\hat{\alpha}$ involves only a negligible error.

The relations 2.2.6.3(2) next lead to the estimates $\hat{p}_k$ indicated in the same table, from which the values of $F_0\hat{p}_k$ required for the calculation of the statistic X defined in 2.2.6.1(22) are deduced.

Since we find

$$X = 0{,}58 \tag{3}$$

and since in a χ^2 distribution with $\nu = 5 - 1 - 1 = 3$ degrees of freedom, the critical value at the 0,05 level is 7,8 , we can conclude that there is good agreement between the fit and the series of observations.

From this we obtain, in particular, the estimate

$$\hat{Q}_3 = \hat{P}(x \leqslant 3) = 0{,}9642 \tag{4}$$

for which, with 2.2.6.3(9), we have

$$\sigma(\hat{Q}_3) = p_3 \times \sigma(\hat{\alpha}) = 0{,}0896 \times 0{,}2128 = 0{,}0191 \tag{5}$$

against

$$\sigma(\hat{Q}_3) = \sqrt{0{,}9642(1-0{,}9642)/27} = 0{,}0358 \tag{6}$$

for the non-parametric estimate.

It follows that for Q_3 we have, at the 0,95 level, the one-sided confidence interval

$$Q_3 \geqslant 0{,}9642 - 1{,}645 \times 0{,}0191 = 0{,}9328 \tag{7}$$

or for T_4 , at the same level, the one-sided interval

$$T_4 \geqslant 1/(1 - 0{,}9328) = 14{,}9 \tag{8}$$

TABLE 17. — Number of stormy days in Belgium, in January (1946-1973). Values of $\hat{p}_k$ given by the fitted Poisson function.

k	f_k	F_k	N_k	l_k^*	$\hat{p}_k$	$F_0\hat{p}_k$	$\hat{Q}_k$	$\hat{T}_k$
0	9	27	33	2,22	0,2946	7,95	0,2946	3,39
1	8	18	33	1,83	0,3600	9,72	0,6546	—
2	6	10	25	1,50	0,2200	5,94	0,8746	2,90
3	3	4	13	1,25	0,0896	2,42	0,9642	8,00
4	1	1	4	1,00	0,0274	0,97	0,9916	27,9
$\geqslant 5$	—	—	—	—	0,0084	—	1,0000	119

2.2.6.4. The negative binomial distribution

Because of the continuous nature of the evolution of the weather, any succession of days containing chance phenomena does not take place in an independent way. It follows that the binomial distribution and its asymptotic form, the Poisson distribution, do not generally provide good models to represent the distribution of the number of days observed over a fixed period of the year, such as a month, a season or the year itself. Indeed, it is observed that, when one or other of these distributions is fitted, then the theroretical variance is too low relative to the empirical variance; this precludes these forms being used to obtain good estimates of the probabilities associated with the observed frequencies.

Since it is defined in terms of two parameters, the negative binomial distribution allows a fit to be made to the series of observations which takes account both of the empirical mean value and of the empirical variance. This distribution has another advantage arising from the possibility of its definition as the resultant distribution obtained by considering an infinity of Poisson distributions whose mean values would be distributed in accordance with a gamma distribution.

Each chance phenomenon – such as a hail storm, for example – is charateristic of particular conditions of the state of the atmosphere which are themselves related to types of weather which can differ from each other to a varying extent. Thus, through its second definition, the negative binomial distribution also offers a means of characterizing the degree of heterogeneity of the causes which produce the chance phenomena considered.

So, with $0<\alpha<1$ and $\beta>0$, we have for this distribution

$$p_0=(1-\alpha)^{\beta} \quad \text{and} \quad p_k=p_0\,\alpha^k\,\frac{\beta(\beta+1)\ldots(\beta+k-1)}{k!},\ k=1,2,\ldots \tag{1}$$

which implies the recurrence

$$p_k=p_{k-1}\,\alpha\,\frac{\beta+k-1}{k},\ k=1,2,\ldots \tag{2}$$

From this we have

$$\mathrm{E}(x)=\sum_0^{\infty} kp_k=\sum_0^{\infty}\alpha p_{k-1}(\beta+k-1)=\alpha\beta+\alpha\,\mathrm{E}(x)$$

which gives

$$\mathrm{E}(x)=\alpha\beta/(1-\alpha) \tag{3}$$

Similarly, we find

$$\mathrm{E}(x^2)=\alpha\beta(1+\alpha\beta)/(1-\alpha)^2 \tag{4}$$

and consequently

$$\operatorname{var} x=\mathrm{E}(x_2)-[\mathrm{E}(x)]^2=\alpha\beta/(1-\alpha)^2 \tag{5}$$

In anticipation of subsequent theoretical conclusions, we note that relationships exist between the two parameters α and β of the negative binomial distribution and the location parameter a and the scale parameter b of the gamma distribution describing the distribution of the mean values m of the Poisson distributions

$$d\mathrm{F}(m)=\frac{m^{a-1}e^{-m/b}}{b^a\Gamma(a)}dm \tag{6}$$

These relations are given by

$$a=\beta \quad \text{and} \quad b=\alpha/(1-\alpha) \tag{7}$$

which allow us to calculate the variance $ab^2 = \alpha^2\beta/(1-\alpha)^2$ of the gamma distribution and thus to obtain a mesure of the degree of heterogeneity of the family of Poisson distributions which govern the phenomenon being studied.

If β becomes very large and α very small, such that $\alpha\beta$ remains constant, we have, at the limit, $ab^2 = 0$ and the negative binomial converges to the Poisson distribution.

In fitting a negative binomial distribution to a series of observations, we note that the likelihood function 2.2.6.1(15) becomes

$$L = F_0\beta\log(1-\alpha) + N_0\log\alpha + F_1\log\beta + F_2\log(\beta+1) + \ldots + F_r\log(\beta+r-1) - \sum_1^r F_k\log k \tag{8}$$

where r denotes the highest observed value of x.

We can thus see that for the negative binomial distribution, N_0 is sufficient only if β is given. Thus estimates do not exist which are jointly sufficient for α and β.

By eliminating α between the maximum likelihood equations

$$\frac{\partial L}{\partial \alpha} = 0 \quad \text{and} \quad \frac{\partial L}{\partial \beta} = 0,$$

we obtain the following equation in β:

$$-F_0\log\left(1+\frac{\bar{x}}{\beta}\right) + \frac{F_1}{\beta} + \frac{F_2}{\beta+1} + \cdots + \frac{F_r}{\beta+r-1} = 0 \tag{9}$$

the solution of which then allows us to calculate α by means of the equation

$$\alpha = \bar{x}/(\beta+\bar{x}) \tag{10}$$

Similarly, using

$$a_{11} = \frac{\beta}{\alpha(1-\alpha)^2}, \quad a_{12} = \frac{1}{1-\alpha} \quad \text{and} \quad a_{22} = \sum_1^\infty \frac{P_j}{(\beta+j-1)^2} = \sum_1^\infty \frac{1-Q_{j-1}}{(\beta+j-1)^2} \tag{11}$$

where

$$P_j = \sum_j^\infty p_k,$$

calculation of

$$-E\left(\frac{\partial^2 L}{\partial\alpha^2}\right), \quad -E\left(\frac{\partial^2 L}{\partial\alpha\partial\beta}\right) \text{ and of } -E\left(\frac{\partial^2 L}{\partial\beta^2}\right)$$

gives the asymptotic values

$$\operatorname{var}\hat{\alpha} = a_{22}/(F_0\, a), \quad \operatorname{var}\hat{\beta} = a_{11}/(F_0\, a) \quad \text{and} \quad \operatorname{cov}(\hat{\alpha},\hat{\beta}) = -a_{12}/(F_0\, a) \tag{12}$$

where $a = a_{11}\, a_{22} - (a_{12})^2$.

For the solution of equation (9), we note that the method of moments leads to the estimates

$$\alpha^* = (s^2 - \bar{x})/s^2 \quad \text{and} \quad \beta^* = (\bar{x})^2/(s^2 - \bar{x}) \tag{13}$$

where $\bar{x}$ and s^2 denote the empirical mean value and variance; this allows us to take β^* as the initial value in obtaining the solution of equation (9) by successive approximation.

We note (in common with Fisher) that if we have $(\beta^* + 2)/\alpha^* > 20$, then the estimates (13) are efficient enough to be used directly.

It only remains to point out that for calculation of var $\hat{Q}_k$, we again have

$$dQ_k = d\sum_0^k p_j = A d\alpha + B d\beta \tag{14}$$

with

$$A = \sum_0^k j p_j/\alpha - \beta \sum_0^k p_j/(1-\alpha)$$

and

$$B = \sum_0^k p_j \log(1-\alpha) + \sum_1^k p_j/\beta + \sum_2^k p_j/(\beta+1) + \ldots + p_k/(\beta+k-1) \tag{15}$$

from which it follows that

$$\text{var } \hat{Q}_k = A^2 \text{ var } \hat{\alpha} + B^2 \text{ var } \hat{\beta} + 2AB \text{ cov } (\hat{\alpha}, \hat{\beta}) \tag{16}$$

2.2.6.4.1. **Example 18.** *The frequency of storms in Belgium in March*

Table 18 gives, for a period of 27 years, the number f_k of years for which a number k of stormy days was counted in March in Belgium. To fit a negative binomial distribution, we extract from this table :

$$\bar{x} = N_0/F_0 = 93/27 = 3{,}4444 \tag{1}$$

$$s^2 = \Sigma f_k (k - \bar{x})^2/(F_0 - 1)$$

$$= [3(-3{,}4444)^2 + 4(-2{,}4444)^2 + \ldots + 2(4{,}5556)^2]/26$$

$$= 5{,}7949 \tag{2}$$

The estimates of α and β deduced from these values by the method of moments are, using 2.2.6.4(13) :

$$\alpha^* = (5{,}7949 - 3{,}4444)/5{,}7949 = 0{,}406$$

$$\beta^* = (3{,}4444)^2/(5{,}7949 - 3{,}4444) = 5{,}047 \tag{3}$$

these being estimates for which Fisher's relation $(\beta^* + 2)/\alpha^* > 20$ is not satisfied.

In this case equation 2.2.6.4(9) becomes

$$\log\left(1 + \frac{3{,}4444}{\beta}\right) = \left[\frac{24}{\beta} + \frac{20}{\beta+1} + \frac{16}{\beta+2} + \frac{12}{\beta+3} + \frac{11}{\beta+4} + \frac{5}{\beta+5} + \frac{3}{\beta+6} + \frac{2}{\beta+7}\right] \Big/ 27 \tag{4}$$

We obtain from this (see below in 2.2.6.4.1.1) $\hat{\beta} = 4{,}16$, so that, using 2.2.6.4(10), we have

$$\hat{\alpha} = 3{,}4444/(4{,}16 + 3{,}4444) = 0{,}453 \tag{5}$$

We then deduce, by virtue of 2.2.6.4(1) and (2),

$$\hat{p}_0 = (1 - 0{,}453)^{4{,}16} = 0{,}081290 \tag{6}$$

together with the values of $\hat{p}_k$, for $k = 1, 2, \ldots$, indicated in Table 18.

TABLE 18. — Number of stormy days in Belgium, in March (27 years). Values of $\hat{p}_k$ given by the fitted negative binomial distribution.

k	f_k	F_k	N_k	l_k^*	$\hat{p}_k$	$F_0\hat{p}_k$	$\hat{Q}_k$
0	3	27	93	4,44	0,081288	2,19	0,081288
1	4	24	93	3,88	0,153186	4,14	0,234474
2	4	20	89	3,45	0,179034	4,83	0,413507
3	4	16	81	3,06	0,166531	4,50	0,580038
4	1	12	69	2,75	0,135035	3,65	0,715073
5	6	11	65	1,91	0,099831	2,70	0,814904
6	2	5	35	2,00	0,069041	1,86	0,883945
7	1	3	23	1,67	0,045394	1,23	0,929339
8	2	2	16	1,00	0,028686	1,90	0,958025
9	—	—	—	—	0,017557	—	0,975582
10	—	—	—	—	0,010467	—	0,986049
11	—	—	—	—	0,006104	—	0,992153
12	—	—	—	—	0,003493	—	0,995646
13	—	—	—	—	0,001967	—	0,997613
14	—	—	—	—	0,001092	—	0,998705
15	—	—	—	—	0,000599	—	0,999304
16	—	—	—	—	0,000325	—	0,999629
17	—	—	—	—	0,000175	—	0,999804
18	—	—	—	—	0,000093	—	0,999897

The resulting values of $F_0\hat{p}_k$ then give for the statistic 2.2.6.1(22)

$$X = 6{,}52 \tag{7}$$

This value is less than 12,6 , the critical value at the 0,05 level in a χ^2 distribution with $\nu = 9-1-2 = 6$ degrees of freedom. We can thus conclude that there is good agreement between the fit and the series of observations. Consequently, we can also characterize the heterogeneity of the family of March storms by means of the variance of the associated gamma distribution; that is, using 2.2.6.4(7),

$$ab^2 = 4{,}16\left(\frac{0{,}453}{1-0{,}453}\right)^2 = 2{,}85 \tag{8}$$

and this value can serve as a basis of comparison for the other months of the year.

Furthermore, for the calculation of the errors of estimation, 2.2.6.4(11) gives :

$$a_{11} = \frac{4{,}16}{0{,}453 \times (0{,}547)^2} = 30{,}69, \quad a_{12} = 1/0{,}547 = 1{,}8282$$

$$a_{22} = \frac{1-0{,}081288}{(4{,}16)^2} + \frac{1-0{,}234474}{(5{,}16)^2} + \cdots = 0{,}114225 \tag{9}$$

and hence

$$a = 30{,}69 \times 0{,}114225 - (1{,}8282)^2 = 0{,}16325 \tag{10}$$

With 2.2.6.4(12), we then deduce :

$$\operatorname{var}\hat{\alpha} = 0{,}114225/(27 \times 0{,}16325) = 0{,}025915$$

$$\operatorname{var}\hat{\beta} = 30{,}69/(27 \times 0{,}16325) = 6{,}9627$$

$$\operatorname{cov}(\hat{\alpha}, \hat{\beta}) = -1{,}8282/(27 \times 0{,}16325) = -0{,}41477 \tag{11}$$

In particular, for $\hat{Q}_4 = \hat{P}(x \leqslant 4) = 0{,}715073$, we have, by virtue of 2.2.6.4(15),

$$A = (0{,}153186 \times 1 + 0{,}179034 \times 2 + \ldots + 0{,}135035 \times 4)/0{,}453 - 4{,}16 \times 0{,}715073/0{,}547 = -2{,}014$$

$$B = 0{,}715073 \log 0{,}547 + (0{,}153186 + 0{,}179034 + \ldots + 0{,}135035)/4{,}16$$

$$+ (0{,}179034 + \ldots + 0{,}135035)/5{,}16 + \ldots + 0{,}135035/7{,}16 = -0{,}1182 \qquad (12)$$

With 2.2.6.4(16), we thus have

$$\text{var } \hat{Q}_4 = (-2{,}014)^2 \times 0{,}025915 + (-0{,}1182)^2 \times 6{,}9627 + 2(-2{,}014)(-0{,}1181)(-0{,}41477)$$

$$= 0{,}00508 \qquad (13)$$

whence

$$\sigma(\hat{Q}_4) = 0{,}071 \qquad (14)$$

The corresponding non-parametric estimate gives

$$\sigma(\hat{Q}_4) = \sqrt{0{,}715073(1 - 0{,}715073)/27} = 0{,}087 \qquad (15)$$

2.2.6.4.1.1 Calculation of the coefficient β of the negative binomial distribution by successive approximation

Equation 2.2.6.4.1(14) is used as an example to illustrate the method. We therefore have (in natural logarithms)

$$\log\left(1 + \frac{3{,}4444}{\beta}\right) = \left[\frac{24}{\beta} + \frac{20}{\beta+1} + \frac{16}{\beta+2} + \frac{12}{\beta+3} + \frac{11}{\beta+4} + \frac{5}{\beta+5} + \frac{3}{\beta+6} + \frac{2}{\beta+7}\right] \Big/ 27 \qquad (1)$$

We take $\beta_1 = 5$, the nearest integer value to β^*, and we insert this into the right-hand side of (1). We obtain

$$\log\left(1 + \frac{3{,}4444}{\beta}\right) = 0{,}521506 \qquad (2)$$

that is,

$$1 + \frac{3{,}4444}{\beta} = 1{,}68456 \qquad (3)$$

hence

$$\beta'_1 = 5{,}0316 \quad \text{and} \quad \beta_1 - \beta'_1 = -0{,}0316 \qquad (4)$$

Similarly, for $\beta_2 = 5{,}1$ we find $\beta'_2 = 5{,}1348$ and $\beta_2 - \beta'_2 = -0{,}0348$, or an increase in the difference between β and β' of $(0{,}0348 - 0{,}0316) = 0{,}0032$ for an increase of β of 0,1. We can thus approximate a zero difference by using

$$\beta_3 = 5 - 0{,}1 \times \frac{0{,}0316}{0{,}0032} = 5 - 1{,}0 = 4{,}0 \qquad (5)$$

We thus obtain

$$\beta'_3 = 3{,}99303 \quad \text{and} \quad \beta_3 - \beta'_3 = 0{,}00697 \qquad (6)$$

Furthermore, with $\beta_4 = 4{,}1$, we have

$$\beta'_4 = 4{,}09738 \quad \text{and} \quad \beta_4 - \beta'_4 = 0{,}00262 \qquad (7)$$

Consequently the corrected new value of β becomes

$$\beta_5 = 4{,}1 + 0{,}1 \times \frac{0{,}00262}{0{,}00697 - 0{,}00262} = 4{,}1 + 0{,}06 = 4{,}16 \tag{8}$$

for which we have

$$\beta'_5 = 4{,}1599 \quad \text{and} \quad \beta_5 - \beta'_5 = 0{,}0001 \tag{9}$$

With $\beta = 4{,}16$ we thus have a sufficiently accurate value of the solution of equation (1).

2.2.6.5 The geometric distribution

The geometric distribution is a negative binomial distribution for which the parameter β is unity. For such a distribution we thus have, using 2.2.6.4(1) and (2),

$$p_0 = 1 - \alpha \quad \text{and} \quad p_k = (1 - \alpha)\alpha^k, \quad k = 1, 2, \ldots \tag{1}$$

together with the recurrence

$$p_k = p_{k-1}\alpha \tag{2}$$

When the probability of a positive observation is α and when successive observations are independent, this distribution function gives the probabilities p_k that a negative observation will be followed by a sequence of k consecutive positive observations. Consequently, the geometric distribution appears as the complement of the ordinary binomial distribution in that it gives the probabilities associated with the various sequences of positive observations into which the total number of positive observations may be decomposed.

Putting $\beta = 1$ in 2.2.6.4(3), (4) and (5), we also obtain

$$\mathrm{E}(x) = \frac{\alpha}{1 - \alpha}, \quad \mathrm{E}(x^2) = \frac{\alpha(1 + \alpha)}{(1 - \alpha)^2} \tag{3}$$

and

$$\operatorname{var} x = \frac{\alpha}{(1 - \alpha)^2} \tag{4}$$

Since the likelihood function 2.2.6.4(8) reduces to

$$\mathrm{L} = \mathrm{F}_0 \log(1 - \alpha) + \mathrm{N}_0 \log \alpha \tag{5}$$

the statistic N_0 can be used for calculating a sufficient estimate of α.

With (3) and 2.2.6.1(18) this estimate can be obtained from the equation

$$\mathrm{N}_0 = \mathrm{F}_0 \cdot \frac{\alpha}{1 - \alpha} \tag{6}$$

that is to say,

$$\hat{\alpha} = \mathrm{N}_0/(\mathrm{N}_0 + \mathrm{F}_0) \tag{7}$$

for which with (4), 2.2.6.1(21) and

$$\frac{d\mathrm{E}(x)}{d\alpha} = (1 - \alpha)^{-2}$$

we have

$$\operatorname{var}\hat{\alpha} = (1-\alpha)^4 \cdot \frac{\alpha}{(1-\alpha)^2} \cdot \frac{1}{F_0} = \frac{\alpha(1-\alpha)^2}{F_0} \tag{8}$$

Finally, from the relation

$$Q_k = P(x \leqslant k) = \sum_0^k p_j = (1-\alpha) + (\alpha - \alpha^2) + \ldots (\alpha^k - \alpha^{k+1}) = 1 - \alpha^{k+1} \tag{9}$$

we obtain

$$\operatorname{var}\hat{Q}_k = (k+1)^2 \alpha^{2k} \operatorname{var}\hat{\alpha} \tag{10}$$

As F_0 is the total number of negative observations and N_0 that of the positive observations, estimate (7) becomes the same as that obtained in 2.2.6.2(5) for the ordinary binomial distribution, since $N_0 + F_0$ is the total number of observations. Furthermore, the expression (8) for the variance of this estimate does not differ essentially from that given in 2.2.6.2(6) in the case of the ordinary binomial distribution, since we have

$$E(F_0) = (1-\alpha)N \quad \text{with} \quad N = F_0 + N_0 \tag{11}$$

On the other hand, if in the geometric distribution we neglect $p_0, p_1, \ldots, p_r$, the truncated distribution thereby obtained is still geometric, and this allows other estimates of α to be found using the same formulae.

However, since for these new estimates the value of F_0 in the new geometric distribution is the value of F_{r+1} in the first, it is immediately apparent from the relation

$$F_{r+1} < F_0 \tag{12}$$

and from (8) that these new estimates will be less accurate than the first.

2.2.6.5.1 Example 19. *Sequences of consecutive months with below-normal rainfall at Uccle*

Table 19 gives the number f_k of months with above-normal rainfall which were followed by k months with below-normal rainfall, over a total of 128 years (1536 months). In particular, over this period 179 above-normal months which were followed by a single below-normal month were counted.

We have seen in 2.2.6.2.1 that the distribution over the year of the below-normal months follows a binomial distribution, so fitting a geometric distribution would appear to be justified.

We therefore deduce, with 2.2.6.5(7) and (8),

$$\hat{\alpha} = 825/(711+825) = 0{,}537109 \tag{1}$$

and

$$\operatorname{var}\hat{\alpha} = 0{,}537109(1-0{,}537109)^2/711 = 0{,}000162 \tag{2}$$

whence, $\sigma(\hat{\alpha}) = 0{,}0127$, these values coinciding with those obtained in 2.2.6.2.1, as we expected.

From the resulting values of $\hat{p}_k$ and $F_0\hat{p}_k$, we then deduce for the statistic defined in 2.2.6.1(22) the value

$$X = 7{,}45 \tag{3}$$

this value being less than 16,9 , the critical value at the 0,05 level of the χ^2 variate with $\nu = 11 - 1 - 1 = 9$ degrees of freedom. We can thus conclude that there is good agreement between the fitted distribution and the series of observations.

For $x \leqslant 4$ we have, in particular

$$\hat{Q}_4 = \hat{P}(x \leqslant 4) = 0{,}955300 \tag{4}$$

TABLE 19. – Number of months, at Uccle, with above-normal rainfall followed by a sequence of k consecutive months with below-normal rainfall (1833-1960). Values of $\hat{p}_k$ given by the fitted geometric distribution

k	f_k	F_k	N_k	l_k^*	$\hat{p}_k$	$F_0\hat{p}_k$	$\hat{Q}_k$	$\hat{T}_k$
0	333	711	825	2,16	0,462891	329,12	0,462891	2,2
1	179	378	825	2,18	0,248623	176,77	0,711514	—
2	99	199	646	2,25	0,133538	94,95	0,845052	3,5
3	41	100	448	2,48	0,071724	51,00	0,916776	6,5
4	23	59	325	2,51	0,038524	27,39	0,955300	12,0
5	13	36	233	2,47	0,020691	14,71	0,975991	22,4
6	9	23	168	2,30	0,011114	7,90	0,987105	41,7
7	8	14	114	2,14	0,005969	4,24	0,993074	77,5
8	3	6	58	2,66	0,003206	2,28	0,996280	144
9	2	3	34	3,33	0,001722	1,22	0,998002	269
10	—	—	—	—	0,000925	1,42	0,998927	501
11	—	—	—	—	0,000497	—	0,999424	—
12	—	—	—	—	0,000267	—	0,999691	—
13	—	—	—	—	0,000143	—	0,999834	—
14	—	—	—	—	0,000077	—	0,999911	—
15	—	—	—	—	0,000041	—	0,999952	—
16	1	1	16	1,00	0,000022	—	0,999974	—

for which we obtain, using 2.2.6.5(10)

$$\sigma(\hat{Q}_4) = 5 \times (0{,}537109)^4 \times 0{,}0127 = 0{,}0053 \tag{5}$$

For the corresponding non-parametric estimate, we would have

$$\sigma(\hat{Q}_4) = \sqrt{0{,}955300(1 - 0{,}955300)/711} = 0{,}0077 \tag{6}$$

We deduce from this that the mean recurrence interval for at least five consecutive below-normal months can be estimated as

$$\hat{T}_5 = 1/(1 - \hat{Q}_4) = 1/(1 - 0{,}955300) = 22{,}4 \tag{7}$$

it being understood that this mean interval is expressed in terms of above-normal months. To express this in years, we note that the mean number of above-normal months per year is equal to $12 \times (1 - \alpha) = 12 \times 0{,}462891 = 5{,}55$. It therefore follows that, expressed in years, the mean interval $\hat{T}_5$ becomes

$$\hat{T}_5 = 22{,}4/5{,}55 = 4{,}0 \tag{8}$$

2.2.6.6 *The logarithmic distribution*

When in a negative binomial distribution the parameter β tends towards 0, the value of p_0 tends towards unity and asymptotically we have

$$p_k/\beta = p_0\alpha^k/k \tag{1}$$

Furthermore, with $0 < \alpha < 1$, the logarithmic distribution is defined by means of the relations

$$p_0 = -\alpha \log^{-1}(1 - \alpha) \quad \text{and} \quad p_k = p_0\alpha^k/(k+1) \tag{2}$$

which gives the recurrence

$$p_k = p_{k-1}\,\alpha\,\frac{k}{k+1} \tag{3}$$

which implies, in particular, that

$$p_1 < p_0/2 \tag{4}$$

By comparing (1) and (2), we see that the logarithmic distribution can be considered as the asymptotic form of a negative binomial distribution with $\beta = 0$, truncated for $k = 0$. It follows that in the logarithmic distribution, the value $k = 0$ corresponds to $k = 1$ in the corresponding negative binomial distribution.

Logarithmic distributions often give a good fit in the case of sequences of positive observations of certain persistent events.

Such being the case, we obtain from (2) and from the recurrence (3)

$$\mathrm{E}(x) = \frac{p_0}{1-\alpha} - 1, \quad \mathrm{E}(x^2) = \frac{\alpha p_0}{(1-\alpha)^2} - \frac{p_0}{1-\alpha} + 1 \tag{5}$$

and hence

$$\operatorname{var} x = \mathrm{E}(x^2) - [\mathrm{E}(x)]^2 = \frac{p_0(1-p_0)}{(1-\alpha)^2} \tag{6}$$

Furthermore, since the likelihood function 2.2.6.1(15) becomes

$$\mathrm{L} = \mathrm{F}_0 \log p_0 + \mathrm{N}_0 \log \alpha - \Sigma f_k \log (k+1) \tag{7}$$

we deduce that the statistic N_0 is sufficient.

Consequently, with (5) and 2.2.6.1(18), the sufficient estimate of α is found from the equation

$$\mathrm{N}_0 = \mathrm{F}_0 \left(\frac{p_0}{1-\alpha} - 1 \right) \tag{8}$$

that is, with $\bar{x} = \mathrm{N}_0/\mathrm{F}_0$, (2) and 2.2.6.1(13)

$$\frac{p_0}{1-\alpha} = \bar{x} + 1 = l_0^*, \text{ or : } \log(1-\alpha) = -\frac{\alpha}{(1-\alpha) l_0^*} \tag{9}$$

Knowing that

$$\frac{d}{d\alpha}\left(\frac{p_0}{1-\alpha}\right) = \frac{p_0(1-p_0)}{\alpha(1-\alpha)^2} = \frac{\operatorname{var} x}{\alpha} \tag{10}$$

we then deduce, using 2.2.6.1(12)

$$\sigma(\hat{\alpha}) = \alpha / \sqrt{\mathrm{F}_0 \operatorname{var} x} \tag{11}$$

Furthermore, we also find for

$$\hat{\mathrm{Q}}_k = \hat{\mathrm{P}}(x \leqslant k) = \sum_0^k p_j :$$

$$\sigma(\hat{\mathrm{Q}}_k) = \left| \sum_0^k j p_j - \mathrm{Q}_k \, \mathrm{E}(x) \right| \frac{\sigma(\hat{\alpha})}{\alpha} \tag{12}$$

2.2.6.6.1 **Example 20.** *Sequences at Uccle of consecutive days with mean temperature* $\leqslant -5^\circ$ *(70 years)*

Table 20 gives the numbers f_k of sequences of $(k+1)$ consecutive days with mean temperature $\leqslant -5^\circ$ observed at Uccle from 1901 to 1970. The number k thus indicates the number of consecutive days with mean temperature $\leqslant -5^\circ$ which followed an initial day with a mean temperature which was as low as this.

As in similar cases the logarithmic distribution gives a good fit, this kind of distribution was also fitted here.

From the data in Table 20 we obtain

$$l^*_0 = (205/118) + 1 = 2{,}73729 \tag{1}$$

which, with 2.2.6.6(9), leads to the equation

$$\log(1-\alpha) = -\frac{\alpha}{(1-\alpha) \times 2{,}73729} \tag{2}$$

Using a method analogous to that described in 2.2.6.4.1.1 and taking as the initial value $\alpha = 0{,}6$, the nearest to the estimate of the parameter of the geometric distribution $\hat{\alpha} = N_0/(N_0 + F_0) = 0{,}635$, we find

$$\hat{\alpha} = 0{,}828 \tag{3}$$

which is very close to that obtained ($\hat{\alpha} = 0{,}833$) for the logarithmic distribution of the sequences of dry days at Uccle.

From this we deduce, using 2.2.6.6(2),

$$p_0 = -0{,}828 \log^{-1} 0{,}172 = 0{,}828/1{,}76026 = 0{,}470385 \tag{4}$$

so that, using 2.2.6.6(6), we have

$$\operatorname{var} x = 0{,}470385(1 - 0{,}470385)/(0{,}172)^2 = 8{,}4209 \tag{5}$$

and with 2.2.6.6(11)

$$\sigma(\hat{\alpha}) = 0{,}828/\sqrt{118 \times 8{,}4209} = 0{,}0263 \tag{6}$$

TABLE 20. — Sequences of length $(k+1)$ at Uccle of consecutive days with mean temperature $\leqslant -5^\circ$ (1901-1970). Values of $\hat{p}_k$ given by the fitted logarithmic distribution

k	f_k	F_k	N_k	l^*_k	$\hat{p}_k$	$F_0\hat{p}_k$	$\hat{Q}_k$	$\hat{T}_k$
0	52	118	205	2,74	0,47038	55,5	0,47038	2,1
1	20	66	205	3,11	0,19474	23,0	0,66512	—
2	21	46	185	3,02	0,10750	12,7	0,77262	3,0
3	7	25	143	3,72	0,06676	7,9	0,83938	4,4
4	5	18	122	3,78	0,04422	5,22	0,88360	6,2
5	3	13	102	3,85	0,03051	3,60	0,91411	8,6
6	4	10	87	3,70	0,02165	2,55	0,93576	11,7
7	—	—	—	—	0,01569	3,21	0,95145	15,6
8	3	6	63	3,50	0,01155	—	0,96300	20,6
9	—	—	—	—	0,00860	1,78	0,97160	27,0
10	1	3	39	4,00	0,00648	—	0,97808	35,2
11	—	—	—	—	0,00492	1,02	0,98300	45,6
12	1	2	29	3,50	0,00376	—	0,98676	58,8
13	—	—	—	—	0,00289	1,52	0,98965	75,5
14	—	—	—	—	0,00223	—	0,99188	96,6
15	—	—	—	—	0,00173	—	0,99361	123
16	—	—	—	—	0,00135	—	0,99496	156
17	1	1	17	1,00	0,00106	—	0,99602	198

The values of $\hat{p}_k$ obtained by means of 2.2.6.6(2) and (3) and those of $F_0\hat{p}_k$ which result from these values give for the statistic 2.2.6.1(22) the value

$$X = 7,6 \tag{7}$$

Since this value is less than 16,9 , the critical value at the 0,05 level of the χ^2 variate with $\nu = 11 - 1 - 1 = 9$ degrees of freedom, we can conclude that there is good agreement between the series of observations and the fitted distribution.

Furthermore, we also have, for example, for $x \leqslant 4$,

$$\hat{Q}_4 = 0,88360 \tag{8}$$

with, by virtue of 2.2.6.6(12)

$$\sigma(\hat{Q}_4) = |0,19474 + 2 \times 0,10750 + \cdots 4 \times 0,04422 - 0,88360 \times 1,73729| \times \frac{0,0263}{0,828}$$
$$= 0,024 \tag{9}$$

whereas the non-parametric estimate of the same probability leads to

$$\sigma(\hat{Q}_4) = \sqrt{0,88360\,(1 - 0,88360)/118} = 0,030 \tag{10}$$

It should be pointed out that the mean recurrence intervals $\hat{T}_k$, are expressed in terms of sequences of consecutive days with mean temperature $\leqslant -5^\circ$. To express these in years, we note that in 70 years 118 were counted, giving a mean value of 1,69 sequences per year.

So in the case of $\hat{T}_5$, for example, the mean interval, expressed in years, becomes

$$\hat{T}'_5 = 8,6/1,69 = 5,1 \tag{11}$$

2.2.6.7 Choice of the discrete distribution to be fitted

As has already been pointed out, most meteorological phenomena are somewhat persistent in nature since their development extends over periods of several days.

It follows that a good fit of the binomial or its asymptotic form, Poisson's distribution, to the distribution of the frequencies of days with particular chance phenomena will be obtained only exceptionally, and that, on the contrary, it will more usually be the negative binomial and its derivatives, the geometric distribution or the logarithmic distribution which will actually provide the best fit.

To help make the necessary choice, we can invoke principles analogous to those which were used for continuous distributions. More precisely, we can initially apply a test for the goodness-of-fit of the binomial distribution and if this test is not adequately fulfilled, we can make some choice based on the behaviour of the empirical lengths or proceed to a significance test on the estimate of the coefficient β in a fit of the negative binomial distribution, comparing this estimate, for example, with the value $\beta = 1$ corresponding to the geometric distribution.

As regards the goodness-of-fit of the binomial distribution, we know that if p is the probability of a positive observation, then the mean of the number x of positive observations in a series of n observations is np. In this case, the number of negative observations is $(n - x)$, its mean is $n(1 - p)$ and the statistic 2.2.6.1(22) becomes, for a particular series of n independent observations,

$$X = \frac{(x - np)^2}{np} + \frac{[(n - x) - n(1 - p)]^2}{n(1 - p)} = \frac{(x - np)^2}{np(1 - p)} \tag{1}$$

the number of degrees of freedom being $\nu = 2 - 1 = 1$. We thus recover the statistic 2.1.10(4).

If, moreover, we perform F_0 independent series of n observations which are also independent, then by adding the statistics X obtained for each series we are led to a new statistic

$$X = \Sigma \frac{(x - np)^2}{np(1 - p)} \tag{2}$$

for which this time, by the principle of additivity of the χ^2 variates, the number of degrees of freedom is equal to $\nu = F_0$.

Finally, if we replace the true mean value np by the empirical mean value : $\bar{x} = n\hat{p}$, the statistic (2) becomes

$$X = \Sigma \frac{(x - \bar{x})^2}{(1 - \hat{p})\bar{x}} \tag{3}$$

for which, under the assumption of a binomial distribution of the frequency x, the distribution is χ^2 with $\nu = F_0 - 1$ degrees of freedom.

The test based on the statistic (3) can thus be used as a goodness-of-fit test for the binomial distribution. It then also follows that the statistic

$$X = \Sigma \frac{(x - \bar{x})^2}{\bar{x}} \tag{4}$$

obtained by making p tend towards zero, possesses the same distribution under the assumption of a Poisson distribution of the frequencies x and consequently can be used as a statistic for a goodness-of-fit test of the latter distribution.

If this test is rejected, the behaviour of the mean lengths l_k as a function of k gives useful indications for the possible choice of the distribution which should provide a reasonable fit. We know, in fact, that the mean lengths l_k decrease with k for the binomial distribution, for Poisson's distribution and for the negative binomial distribution with $\beta > 1$, that these lengths are all equal to each other for the geometric distribution and that they are increasing with k for the negative binomial with $\beta < 1$ and for the logarithmic distribution.

So if the values of the empirical lengths for the initial values of k reveal a distinct trend or, on the other hand, if they appear to fluctuate about a mean value, then this directly points towards the use of a negative binomial distribution with $\beta > 1$ in the case of a decreasing trend, towards a negative binomial distribution with $\beta < 1$ or a logarithmic distribution in the case of an increasing trend and towards the use of a geometric distribution in the case of a reasonably small fluctuation. In the above examples 16 to 20 (2.2.6.2.1 to 2.2.6.6.1), calculation of the empirical mean lengths l^*_k confirms that the trend of successive values of these lengths each time constituted an indicator for the type of distribution which should be fitted.

As a last resort, it is still possible to estimate the coefficients α and β of the negative binomial distribution together with the coefficient α under the null hypothesis $\beta = \beta_0$. If $\hat{\alpha}$ and $\hat{\beta}$ are the estimates of the parameters of the negative binomial and if $\hat{\alpha}_0$ is the estimate of α when $\beta = \beta_0$, the statistic of the test of the null hypothesis $\beta = \beta_0$ becomes, by virtue of 2.2.6.4(11) and (12),

$$X = F_0[a_{11}(\Delta\hat{\alpha})^2 + a_{22}(\Delta\hat{\beta})^2 + 2a_{12}\Delta\hat{\alpha}.\Delta\hat{\beta}] \tag{5}$$

with $\Delta\hat{\alpha} = \hat{\alpha} - \hat{\alpha}_0$ and $\Delta\hat{\beta} = \hat{\beta} - \beta_0$.

In particular, for the calculation of a_{22} in the case of a negative binomial distribution where $\beta_0 = 1$ (geometric) we note that we have

$$p_k = \alpha^k \tag{6}$$

and that, consequently, the last of the relations 2.2.6.4(11) becomes, by virtue of 2.2.6.5(9)

$$a_{22} = \sum_{1}^{\infty} \alpha^j / j^2 \tag{7}$$

Note that the likelihood ratio test may also be used as an alternative test (see 2.2.5.1).

It remains to point out that, whatever final fit one arrives at, it is still useful to check the goodness-of-fit of the sample distribution with the series of observations, by means of the statistic X defined in 2.2.6.1(22).

2.2.6.7.1 **Example 21.** *Number of sequences per year at Uccle of consecutive days with mean temperature* $\leqslant -5^\circ$ *(70 years)*

Table 21 gives, for a total of 70 years, the number of years at Uccle for which k sequences of consecutive days with mean temperature $\leqslant -5^\circ$ were observed.

With

$$\bar{x} = N_0/F_0 = 1{,}68571 \tag{1}$$

the statistic 2.2.6.7(4) of the goodness-of-fit test for the Poisson distribution gives

$$X = [(0 - 1{,}68571)^2 \times 24 + (1 - 1{,}68571)^2 \times 16 + \ldots + (8 - 1{,}68571)^2 \times 2)]/1{,}68571 = 146{,}6 \tag{2}$$

Since for the χ^2 variate with $\nu = 70 - 1 = 69$ degrees of freedom the critical value at the 0,05 level is 89,4 , the goodness-of-fit of the Poisson distribution must be deemed unsatisfactory, and must be rejected. Furthermore, the fit of the binomial is even less statisfactory, since whatever total number n of observations the particular sequences considered are referred to, the presence of the divisor $(1 - \hat{p})$ in the corresponding statistic X would make its value even larger than that found in (2).

Moreover, the values of the empirical lengths l^*_k in Table 21 reveal the absence of any distinct systematic variation as a function of k, this being compatible with the assumption of a geometric distribution. To verify this, we can proceed to estimate the parameters α and β in a negative binomial distribution and proceed to test the hypothesis $\beta = 1$.

To this end, we note that with 2.2.6.5(7) we have for the geometric distribution :

$$\hat{\alpha}_0 = 118/(70 + 118) = 0{,}627660 \tag{3}$$

whereas for the negative binomial distribution with $\beta = 1$, equations 2.2.6.4(11) and 2.2.6.7(7) give :

$$a_{11} = \frac{1}{0{,}627660(1 - 0{,}627660)^2} = 11{,}4920, \quad a_{12} = \frac{1}{1 - 0{,}627660} = 2{,}68572$$

$$a_{22} = \sum_{1}^{\infty} \frac{(0{,}627660)^k}{k^2} = 0{,}770465 \tag{4}$$

Furthermore, equation 2.2.6.4(9) here becomes :

$$\log\left(1 + \frac{1{,}68571}{\beta}\right) = \left(\frac{46}{\beta} + \frac{30}{\beta+1} + \frac{19}{\beta+2} + \frac{10}{\beta+3} + \frac{5}{\beta+4} + \frac{4}{\beta+5} + \frac{2}{\beta+6} + \frac{2}{\beta+7}\right) \Big/ 70 \tag{5}$$

From this we obtain $\beta = 1{,}38$, and hence with (1) and 2.2.6.4(10) :

$$\hat{\alpha} = 1{,}68571/(1{,}38 + 1{,}68571) = 0{,}54986 \tag{6}$$

With $\Delta\hat{\alpha} = 0{,}54986 - 0{,}62766 = -0{,}0778$ and $\Delta\hat{\beta} = 1{,}38 - 1 = 0{,}38$, the statistic 2.2.6.7(5) becomes :

$$X = 70[11{,}492(-0{,}0778)^2 + 0{,}770465(0{,}38)^2 + 2 \times 2{,}68572(-0{,}0778)(0{,}38)] = 1{,}54$$

Since this value is less than 5,99 , the critical value at the 0,05 level of the χ^2 variate with two degrees of freedom, the hypothesis of a geometric distribution can be accepted.

For the likelihood ratio test, with (3), $F_0 = 70$ and $N_0 = 118$ (Table 21), 2.2.6.5(5) gives : $\hat{L}_1 = -124{,}116$, whereas with $\hat{\alpha} = 0{,}54986$, $\hat{\beta} = 1{,}38$ and the values of F_k and N_k in Table 21, 2.2.6.4(8) gives $\hat{L}_2 = -123{,}662$. Thus : $X = -2(\hat{L}_1 - \hat{L}_2) = 0{,}91$ with $2 - 1 = 1$ degree of freedom. At the 0,05 level, the critical value being 3,84 , the conclusion remains unchanged.

TABLE 21. — Number of sequences per year at Uccle of consecutive days with mean temperature $\leqslant -5^{\circ}$ (70 years). Values of $\hat{p}_k$ given by the fitted geometric distribution

k	f_k	F_k	N_k	l_k^*	$\hat{p}_k$	$F_0\hat{p}_k$
0	24	70	118	2,69	0,372340	26,06
1	16	46	118	2,57	0,233703	16,36
2	11	30	102	2,40	0,146686	10,27
3	9	19	80	2,21	0,092069	6,44
4	5	10	53	2,30	0,057788	4,05
5	1	5	33	2,60	0,036271	2,54
6	2	4	28	2,00	0,022766	1,59
7	—	2	—	—	0,014289	2,69
8	2	2	16	1,00	0,008969	—

The calculation of $\hat{p}_k$ and of $F_0\hat{p}_k$ next gives for the statistic 2.2.6.1(22)

$$X = 2{,}68 \tag{7}$$

that is, a value less than 12,6 , the critical value at the 0,05 level of the χ^2 variate with $\nu = 8 - 1 - 1 = 6$ degrees of freedom. The goodness-of-fit of the distribution fitted is thus confirmed.

Otherwise, we still have, by virtue of 2.2.6.5(8),

$$\operatorname{var} \hat{\alpha} = 0{,}627660(1 - 0{,}627660)^2/70 = 0{,}001243 \tag{8}$$

2.2.6.8 Distribution of maximal sequences

The knowledge, for a period of the year (month, season or the year itself), of the distribution of the sequences of positive observations of various lengths, together with that of the number of sequences during this period, allows us to calculate the probability associated with the maximal sequences for the period considered.

Indeed, if for the periods during which non zero sequences occur we have for the length x of these sequences

$$P(x \leqslant k) = Q_k \tag{1}$$

we know, from what we have seen in 2.1.1, that the probability of having, over n independent sequences, a maximal sequence of length $L \leqslant k$ is

$$P(L \leqslant k) = [Q_k]^n \tag{2}$$

If, in addition the number of sequences is variable and if p_n denotes the probability of having n sequences during this period, the total probability for $n \geqslant 1$ becomes

$$P(L \leqslant k) = E[(Q_k)^n] \cong [Q_k]^{n_0} \tag{3}$$

with

$$n_0 = \sum_1^{\infty} np_n \Big/ \sum_1^{\infty} p_n = S_1/P_1 = l_1 \tag{4}$$

from which it follows that n_0 is the mean length for $k = 1$ of the distribution of the number of sequences.

Finally, since p_0 denotes the probability of a year without any sequences, under the hypothesis of the independence of the succession of sequences we have :

$$Q_{M,0} = P(L=0) = p_0 \tag{5}$$

and

$$Q_{M,k} = P(L \leqslant k) = p_0 + (1-p_0)[Q_k]^{n_0} \tag{6}$$

For the calculation of the error in the estimates $\hat{Q}_{M,k}$, we have

$$\operatorname{var} \hat{Q}_{M,0} = \operatorname{var} \hat{p}_0 \tag{7}$$

while for $\hat{Q}_{M,k}$ with $k \geqslant 1$, we note that with the use of natural logarithms we obtain from (6)

$$\log[Q_{M,k} - p_0] = \log(1-p_0) + n_0 \log Q_k \tag{8}$$

which implies

$$dQ_{M,k} = \left[1 - \frac{Q_{M,k} - p_0}{1-p_0}\right] dp_0 + (Q_{M,k} - p_0) \log Q_k \, dn_0 + \frac{n_0 (Q_{M,k} - p_0)}{Q_k} dQ_k \tag{9}$$

where dp_0, dn_0 and dQ_k must be expressed explicitly as a function of the parameters of the two distributions.

In particular, if the distribution of the number of sequences n depends on only a single parameter α, then relation (9) can be written in the form

$$dQ_{M,k} = A\,d\alpha + B\,dQ_k$$

with

$$A = \left[1 - \frac{Q_{M,k} - p_0}{1-p_0}\right] \frac{dp_0}{d\alpha} + (Q_{M,k} - p_0) \log Q_k \frac{dn_0}{d\alpha}$$

$$B = \frac{n_0 (Q_{M,k} - p_0)}{Q_k} \tag{10}$$

from which it follows, by the assumed independence of the two distributions, that

$$\operatorname{var} \hat{Q}_{M,k} = A^2 \operatorname{var} \hat{\alpha} + B^2 \operatorname{var} \hat{Q}_k \tag{11}$$

2.2.6.8.1 **Exemple 22.** *The annual maximal sequence at Uccle of consecutive days with mean temperature* $\leqslant -5°$ (70 years)

Table 22 gives the number of years f_k for which the maximal sequence of consecutive days with mean temperature $\leqslant -5°$ at Uccle comprised $k = 0, 1, 2, \ldots$ days, over a total of 70 years.

Since the number of sequences per year at Uccle of the days considered follows a geometric distribution (2.2.6.7.1, Example 21), we immediately have

$$\hat{Q}_{M,0} = \hat{p}_0 = 1 - \hat{\alpha} = 1 - 0{,}6277 = 0{,}3723 \tag{1}$$

just as, by 2.2.6.8(4), 2.2.6.1(9) and 2.2.6.5(3),

$$\hat{n}_0 = \hat{l}_1 = \hat{l}_0 = \hat{S}_0 + 1 = E(\hat{x}) + 1 = 2{,}6857 \tag{2}$$

since, in a geometric distribution, all the mean lengths are equal to each other.

TABLE 22. – Yearly maximal sequences at Uccle of consecutive days with mean temperature $\leqslant -5^{\circ}$ (70 years). Values of $\hat{p}_{M,k}$ given by the fitted distribution of maximal sequences

k	f_k	F_k	$\hat{Q}_{M,k}$	$\hat{p}_{M,k}$	$F_0\hat{p}_{M,k}$	$\hat{T}_k$
0	24	70	0,3723	0,3723	26,06	2,7
1	7	46	0,4551	0,0828	5,80	2,2
2	7	39	0,5822	0,1271	8,90	—
3	12	32	0,6863	0,1041	7,29	2,4
4	4	20	0,7645	0,0782	5,47	3,2
5	4	16	0,8225	0,0580	4,06	4,2
6	2	12	0,8655	0,0430	3,01	5,6
7	4	10	0,8975	0,0320	2,25	7,4
8	—	—	0,9215	0,0240	—	9,8
9	3	6	0,9396	0,0181	2,95	12,7
10	—	—	0,9533	0,0137	—	16,6
11	1	3	0,9637	0,0104	1,69	21,4
12	—	—	0,9718	0,0081	—	27,5
13	1	2	0,9779	0,0061	0,99	35,5
14	—	—	0,9827	0,0048	1,53	45,2
15	—	—	0,9864	0,0037	—	57,8
16	—	—	0,9893	0,0029	—	73,5
17	—	—	0,9915	0,0022	—	93,5
18	1	1	0,9933	0,0018	—	118

The values of $\hat{Q}_k$ of the distribution at Uccle of the sequences of various lengths (2.2.6.6.1, Example 20 - Table 20), taking account of the fact that in this distribution the values of k here correspond to $(k+1)$, next give with 2.2.6.8(6), for $k=1$

$$\hat{Q}_{M,1} = 0{,}3723 + (1 - 0{,}3723)(0{,}4704)^{2{,}6857} = 0{,}4551 \tag{3}$$

We find in similar fashion the other values of $\hat{Q}_{M,k}$ indicated in Table 22. Then we deduce for $\hat{p}_{M,k}$ and for $F_0\hat{p}_{M,k}$ values which lead, for the statistic 2.2.6.1(22), to

$$X = 6{,}42 \tag{4}$$

for which the number of degrees of freedom is $\nu = 12 - 1 - 2 = 9$, since each of the distributions used (lengths of the sequences and number of sequences) was calculated on the basis of the estimate of one parameter. Since the critical value at the 0,05 level is here 16,9 , we can conclude that there is a good agreement between the law fitted and the series of observations.

For calculating the error, taking again the case of a maximal length $k=5$, we have

$$\hat{Q}_{M,5} = 0{,}8225 \tag{5}$$

Moreover, from the relations

$$p_0 = 1 - \alpha \quad \text{and} \quad n_0 = l_0 = E(x) + 1 = \frac{1}{1-\alpha} \tag{6}$$

which are satisfied in the case of a geometric distribution [2.2.6.5(1) and (3)], we obtain

$$\frac{dp_0}{d\alpha} = -1 \text{ et } \frac{dn_0}{d\alpha} = \frac{1}{(1-\alpha)^2} = \frac{1}{p_0^2} \tag{7}$$

It follows that 2.2.6.8(10) becomes for A

$$A = \left(1 - \frac{Q_{M,k} - p_0}{1 - p_0}\right)(-1) + \frac{(Q_{M,k} - p_0)\log Q_k}{p_0^2} \tag{8}$$

With $\hat{Q}_5 = 0{,}8836$ [2.2.6.6.1(8)] we have, in particular, for $k = 5$

$$A = \left[1 - \frac{0{,}8225 - 0{,}3723}{1 - 0{,}3723}\right](-1) + \frac{(0{,}8225 - 0{,}3723)\log 0{,}8836}{(0{,}3723)^2}$$

$$= -0{,}28278 - 0{,}40195 = -0{,}68473 \tag{9}$$

while for B, 2.2.6.8(10) gives

$$B = \frac{2{,}6857(0{,}8225 - 0{,}3723)}{0{,}8836} = 1{,}3684 \tag{10}$$

With var $\hat{\alpha} = 0{,}001243$ and var $\hat{Q}_5 = (0{,}024)^2 = 0{,}000576$ calculated respectively in 2.2.6.7.1(11) and 2.2.6.6.1(9), we finally obtain

$$\text{var } \hat{Q}_{M,5} = (-0{,}68473)^2 \times 0{,}001243 + (1{,}3684)^2 \times 0{,}000576 = 0{,}001661$$

or

$$\sigma(\hat{Q}_{M,5}) = 0{,}041 \tag{11}$$

The corresponding non-parametric estimate gives

$$\sigma(\hat{Q}_{M,5}) = \sqrt{0{,}8225 \times (1 - 0{,}8225)/70} = 0{,}046 \tag{12}$$

2.2.6.9 The approximate continuous forms of discrete distributions

When the variance of a discrete distribution is sufficiently large for the probabilities associated with the various values of k all to be relatively small, the discrete distribution can be approximated by a continuous form. Furthermore, for the discrete distributions which have been considered above, the approximate form is, depending on the case considered, either a gamma or a normal distribution.

If these conditions are satisfied, the estimate of the various probabilities can be made through the intermediary of the formulae relating the parameters of the discrete distribution to those of the corresponding approximate continuous distribution function.

Nonetheless, it is often simpler to fit a continuous function directly to the observed frequencies, treating the discrete variable as a continuous variable; this also applies to the choice of the distribution to be fitted.

For the reduction of the continuous distribution obtained in this manner to the appropriate discrete form, we need only note that the probability interval of the continuous function, which corresponds to a particular value of the frequency k, must be centred on this value; that is to say, we should put

$$p_k = \int_{k-\frac{1}{2}}^{k+\frac{1}{2}} p(x)dx \tag{1}$$

where $p(x)$ denotes the probability density of the continuous function fitted.

2.2.6.9.1. **Example 23.** *Number of days with measurable precipitation at Uccle in July (70 years)*

Table 23 gives the number f_k of months of July for which k days of measurable precipitation (at least 0,1mm) were recorded over a period of 70 years.

We obtain for the mean value

$$\overline{x} = N_0/F_0 = 1174/70 = 16,771 \tag{1}$$

Furthermore, for the calculation of the statistic 2.2.6.7(3), we have

$$\hat{p} = \overline{x}/n = 16,771/31 = 0,54101 \tag{2}$$

which, for this statistic, leads to the value

$$X = \Sigma \frac{(k - 16,771)^2 f_k}{(1 - 0,54101)16,771} = 273,4 \tag{3}$$

Since for $\nu = 70 - 1 = 69$ degrees of freedom the critical value at the 0,05 level of the χ^2 variate is 89,4 , we must conclude that a binomial distribution must be rejected.

TABLE 23. — Number of days of measurable precipitation at Uccle in July (1901-1970). Values of $\hat{p}_k$ given by the sample fitted normal distribution

k	f_k	F_k	N_k	l_k^*	a_k	$\hat{P}_k$	$\hat{p}_k$	$F_0\hat{p}_k$	$\hat{T}_k$
0	—	70	1174	17,8	3,12	(0,9991)	0,0016	—	625
1	—	—	—	—	2,94	0,9984	0,0013	—	345
2	—	—	—	—	2,76	0,9971	0,0022	—	196
3	—	—	—	—	2,57	0,9949	0,0033	—	119
4	—	—	—	—	2,39	0,9916	0,0052	—	74
5	1	70	1174	12,8	2,21	0,9864	0,0076	—	47,2
6	—	—	—	—	2,03	0,9788	0,0110	—	31,1
7	3	69	1169	10,9	1,85	0,9678	0,0153	3,33	21,1
8	2	66	1148	10,4	1,67	0,9525	0,0206	1,44	14,7
9	2	64	1132	9,7	1,49	0,9319	0,0270	1,89	10,5
10	2	62	1114	9,0	1,31	0,9049	0,0341	2,39	7,7
11	2	60	1094	8,2	1,13	0,8708	0,0419	2,93	5,8
12	7	58	1072	7,5	0,95	0,8289	0,0495	3,46	4,5
13	2	51	988	7,4	0,77	0,7794	0,0570	3,99	3,6
14	4	49	962	6,6	0,59	0,7224	0,0633	4,43	2,9
15	2	45	906	6,1	0,41	0,6591	0,0681	4,77	2,4
16	6	43	876	5,4	0,23	0,5910	0,0711	4,98	2,1
17	4	37	780	5,1	0,05	0,5199	0,0716	5,01	—
18	4	33	712	4,6	−0,13	0,4483	0,0700	4,90	2,2
19	4	29	640	4,1	−0,31	0,3783	0,0662	4,63	2,6
20	8	25	564	3,6	−0,49	0,3121	0,0607	4,25	3,2
21	2	17	404	3,8	−0,67	0,2514	0,0537	3,76	4,0
22	3	15	362	3,1	−0,85	0,1977	0,0462	3,23	5,1
23	4	12	296	2,7	−1,03	0,1515	0,0384	2,69	6,6
24	3	8	204	2,5	−1,21	0,1131	0,0308	2,16	8,8
25	3	5	132	2,4	−1,39	0,0823	0,0241	1,69	12,2
26	—	—	—	—	−1,57	0,0582	0,0190	4,07	17,2
27	—	—	—	—	−1,76	0,0392	0,0130	—	25,5
28	1	2	57	1,5	−1,94	0,0262	0,0092	—	38,2
29	1	1	29	1,0	−2,12	0,0170	0,0063	—	59
30	—	—	—	—	−2,30	0,0107	0,0041	—	94
31	—	—	—	—	−2,48	0,0066	0,0066	—	152

The use of the D'Agostino statistic [2.2.5.2.2.1(b)] for checking the goodness of fit of the normal distribution next gives, by arranging the 70 observed values of k in non-decreasing order

$$\sum_{1}^{35} [71/2 - m](x_{71-m} - x_m) = 7669 \tag{4}$$

With $m_2 = \Sigma(k - 16{,}771)^2 f_k/70 = 30{,}06204$, that is $\sqrt{m_2} = 5{,}4829$, we next obtain

$$X = \frac{7669}{(70)^2 \times 5{,}4829} = 0{,}2854515 \tag{5}$$

which gives for the D'Agostino transform

$$Y = \frac{\sqrt{70}(0{,}2854515 - 0{,}28209479)}{0{,}02998598} = 0{,}937 \tag{6}$$

Since this value lies within the confidence interval at the 0,95 level (−2,15 , 1,04) for this statistic, the assumption of a normal distribution can be accepted. Under these conditions, the unbiased estimate of the standard deviation, by virtue of 2.2.1(14), is given by

$$\hat{\sigma} = \sqrt{m_2}\sqrt{\frac{F_0}{F_0 - 1}} \cdot \sqrt{\frac{F_0 - 1}{F_0 - 1{,}5}} = 5{,}4829\sqrt{\frac{70}{68{,}5}} = 5{,}5426 \tag{7}$$

from which it follows that the quantiles of the standard normal distribution

$$\hat{u}_k = \left[\bar{x} - \left(k - \frac{1}{2}\right)\right] \Big/ \hat{\sigma} \tag{8}$$

allow us to calculate the values of $\hat{P}_k$ of the discrete function by means of the relation

$$\hat{P}_k = F(\hat{u}_k) \tag{9}$$

where $F(\hat{u}_k)$ is the value of the distribution function of the standard normal corresponding to $\hat{u}_k$. In particular, for $k = 1$ we obtain

$$\hat{u}_1 = \frac{16{,}771 - 0{,}5}{5{,}5426} = 2{,}94$$

so that, using a table of the standard normal distribution function, we have

$$\hat{P}_1 = F(\hat{u}_1) = F(2{,}94) = 0{,}998359 \tag{10}$$

The values thereby deduced for $\hat{p}_k$ and $F_0\hat{p}_k$ then lead for the statistic 2.2.6.1(22) to

$$X = 14{,}8 \tag{11}$$

for $\nu = 20 - 1 - 2 = 17$ degrees of freedom.

Since the critical value at the 0,05 level of the corresponding χ^2 variate is 27,6 , we can conclude that there is good fit with the series of observations.

2.2.6.10 Reference works consulted

Parametric estimation in the case of distributions of discrete variables does not differ essentially from that in the case where the random variable is continuous. For this reason, the solutions presented comprise direct applications of the methods described above, with the exception of the minimal χ^2 method, the mention of which is justified here because it concerns more particularly the case of

a discrete variable. For details relating to the latter method we can refer to [11].

The theoretical framework which we have used as the basis for the discrete distributions is taken from [36], whereas the estimation theory in the case of the negative binomial distribution is taken from [37] and [38]; Fisher's condition for the application, in this case, of the method of moments was introduced in [39]. Furthermore, we can consider the determination of the distribution of maximal sequences as being a direct application of the theory of extreme values. We should also note that the use of truncated distributions can also prove useful as has been pointed out in [40].

Examples 17, 18 and 23 relate to series of observations which have been used previously in [38] and in [15], in some cases completed by recent data obtained during the last few years. The data in the other examples were drawn from the climatological archives for Uccle.

2.3 PARAMETRIC ESTIMATION IN THE CASE OF CORRELATED SERIES

The problem presented most frequently to the climatologist is that which results from the presence in the climatological networks of central stations, which may have been in existence for a hundred years or even longer, alongside local stations which have been in operation generally for no more than 10 or 20 years and only rarely for 30 years. Furthermore, in the long-term stations the situation is often complicated by the fact that, largely due to changes in the methods of observation as time has progressed, the series obtained are not homogeneous.

We know, on the other hand, that when series of observations are used for estimating climatological parameters, the methods used lead to values which become progressively less accurate as the series become shorter.

It follows that as long as it is possible to correct the long series for any possible heterogeneity, the components of the climate of a region are generally determined with high accuracy for the long-term stations whereas, in contrast, for the short-term stations, these same components can be obtained with only mediocre accuracy.

This situation, however, is not always as bad as it may appear at first sight. Indeed, when the local station is not too far away from the central station, it can be expected that, in the course of their evolution, the variations in the meteorological elements in the two stations will remain more or less in agreement, i.e. the behaviour of the series of observations will be more or less correlated. It would thus appear that taking advantage of this correlation must lead to a better determination than when any such correlation is absent.

In the following we shall show that, by using this correlation, it is actually possible to provide more precise estimates of the climatological parameters of the local stations and that, in addition, when the correlation is high, the improvement in accuracy can be very significant.

2.3.1 Generalities and fundamental problems

Let x and y be the random variables representing the values taken respectively at the long-term station and the short-term station by the same climatological element. Suppose, furthermore, that the statistical distribution of these two variables is determined for each of them by location parameters μ and μ_1 and scale parameters σ and σ_1 through the intermediary of a single standardized variable t.

It follows that the quantiles of same order of the variables x, y, and t are then connected by the relations

$$x = \sigma t + \mu \quad \text{and} \quad y = \sigma_1 t + \mu_1 \tag{1}$$

and that for the variables x and y we also have

$$y = a_1 x + b_1 \tag{2}$$

with

$$a_1 = \sigma_1/\sigma \quad \text{and} \quad b_1 = \mu_1 - a_1\mu = \mu_1 - r\sigma_1 \tag{3}$$

where $r = \mu/\sigma$, which also gives

$$a_1^2 = \text{var } y/\text{var } x \tag{4}$$

If we also assume that the two series are in correlation, this time denoting by x and y a pair of corresponding values from these series and by ρ_1 the coefficient of correlation between x and y, then in the case of a linear regression, we have

$$y - \mu_1 = \rho_1 a_1 (x - \mu) + e \tag{5}$$

where e represents a random variable with zero mean value, independent of x.

It then follows that the expression $(y - \rho_1 a_1 x)$ is also a random variable independent of x, whose mean value is $(\mu_1 - \rho_1 a_1 \mu)$ and whose variance is $(1 - \rho_1^2)$ var y.

In view of this, the problems considered here are as follows :

(a) The optimal estimation of the parameters of the distribution of the variable y, taking account of the correlation which relates it to the variable x;

(b) The calculation of missing values in the series of values of the variable y, taking the series of values of x as a reference series.

It is clear that the solution of the first problem provides that of the fitting of the *normal* (or climatological average) of the variable y, since the mathematical expectation of this normal is the same as that of the location parameter μ_1 or has with this parameter some relationship which is determined by the type of probability distribution of the variable.

For the second problem, we note that the method of calculating the missing values by using the series of values of x as a reference series comes down to deriving the value of the standardized variable t from the value of x corresponding to the missing value of y by means of the first relation (1), that is

$$t = (x - \mu)/\sigma \tag{6}$$

and introducing this value of t into the second relation (1), which, in fact, consists of applying relation (2).

This method also allows us to extend, in homogeneous fashion, the series of values of y when, this being of recent origin, we wish to reconstitute it, with the aid of the variable x, for earlier periods of time.

We should emphasize straight away that the quantile method used for calculating the missing values must be preferred to that which consists of applying the relation

$$y' = \rho_1 a_1 (x - \mu) + \mu_1 \tag{7}$$

deduced from (5) since, with (4), it implies the relation

$$\text{var } y' = \rho_1^2 a_1^2 \text{ var } x = \rho_1^2 \text{ var } y < \text{var } y \tag{8}$$

which shows that, in the latter method, the corrected variable y' has a smaller variance than the original variable.

Furthermore, if in order to avoid this difficulty we adopt

$$y_1 = y' + e_1 \tag{9}$$

where e_1 is a quantity chosen arbitrarily from a population with zero mean value and with variance $(1 - \rho_1^2)$ var y, we should have

$$\text{var}(y - y_1) = 2(1 - \rho_1^2)\text{var } y \tag{10}$$

whereas for the estimate via the quantile method we have

$$\text{var}(y - a_1 x - b_1) = 2(1 - \rho)\text{var } y \tag{11}$$

We therefore conclude that the quantile method is the more accurate.

2.3.2 Optimal parametric estimation

Let x_i , y_i, $i = 1, 2, \ldots, n_1$ be a simple random sample of pairs of corresponding observed values of the random variables x and y considered in 2.3.1, for which we assume that the parameters μ and σ of the variable x are known. The problem of estimating the parameters μ_1 and σ_1 of the variable y thus reduces to an estimation for a bivariate distribution for which a marginal distribution is known.

The optimal estimation of σ_1 and μ_1 can be considered in two cases :

(a) The bivariate distribution is normal :

(b) The location and scale parameters of the distributions of x and y possess linear estimators.

2.3.2.1 Case of a bivariate normal distribution

When the distribution of the variables x and y is normal and when the parameters μ and σ of the marginal distribution of x are known, the equations deduced by the maximum likelihood method lead for μ_1 to the estimator

$$\mu_1^* = \hat{\mu}_1 - \hat{\rho}_1(\hat{\sigma}_1/\hat{\sigma})(\hat{\mu} - \mu) \tag{1}$$

with

$$\hat{\mu} = \Sigma x_i/n_1, \quad \hat{\mu}_1 = \Sigma y_i/n_1, \quad \hat{\sigma}^2 = \Sigma(x_i - \hat{\mu})^2/n_1,$$
$$\hat{\sigma}_1^2 = \Sigma(y_i - \hat{\mu}_1)^2/n_1, \quad \hat{\rho}_1 = \Sigma(x_i - \hat{\mu})(y_i - \hat{\mu}_1)/(n_1\ \hat{\sigma}\ \hat{\sigma}_1) \tag{2}$$

whereas for σ_1 and ρ_1 it can be shown that the estimators

$$\sigma_1^* = \hat{\sigma}_1(\sigma/\hat{\sigma})^{\hat{\rho}_1^2} \quad \text{and} \quad \rho_1^* = \hat{\rho}_1(\sigma/\hat{\sigma})^{(1-\hat{\rho}_1^2)} \tag{3}$$

possess optimal variances.

Furthermore, for these estimators, we have the following variances :

$$\operatorname{var}\mu_1^* = (1-\rho_1^2)\operatorname{var}\hat{\mu}_1, \quad \operatorname{var}\sigma_1^* = (1-\rho_1^4)\operatorname{var}\hat{\sigma}_1,$$

$$\operatorname{var}\rho_1^* = \frac{2-\rho_1^2}{2}\operatorname{var}\hat{\rho}_1$$

$$\operatorname{cov}(\mu_1^*, \sigma_1^*) = 0, \quad \operatorname{cov}(\mu_1^*, \rho_1^*) = 0$$

$$\text{and} \quad \operatorname{cov}(\sigma_1^*, \rho_1^*) = (1-\rho_1^2)\operatorname{cov}(\hat{\sigma}_1, \hat{\rho}_1) \tag{4}$$

where we have written

$$\operatorname{var}\hat{\mu}_1 = \sigma_1^2/n_1, \quad \operatorname{var}\hat{\sigma}_1 = \sigma_1^2/(2n_1), \quad \operatorname{var}\hat{\rho}_1 = (1-\rho_1^2)^2/n_1$$

$$\text{and} \quad \operatorname{cov}(\hat{\sigma}_1, \hat{\rho}_1) = \rho_1\sigma_1(1-\rho_1^2)/(2n_1) \tag{5}$$

Since the variances indicated in (5) are also the variances of the estimators $\hat{\mu}_1$, $\hat{\sigma}_1$ and $\hat{\rho}_1$ which can be applied to the sample of size n_1 in the absence of any other information on the parameters μ and σ , it is immediately apparent that the estimators μ_1^* , σ_1^* and ρ_1^* are more precise.

In practice, the values of μ and σ are not given, but are themselves estimates $\hat{\hat{\mu}}$ and $\hat{\hat{\sigma}}$ deduced from the complete series of values of x, which we assume to be of sample size n. In this case, relations (1) and (3) remain applicable, and we note, for the calculation of the variance of the error of estimation, that switching to the differentials we obtain

$$d\mu_1^* = d\hat{\mu}_1' + \rho_1 a_1 d\hat{\hat{\mu}} \quad \text{and} \quad d\sigma_1^* = d\hat{\sigma}_1' + \rho_1^2 a_1 d\hat{\hat{\sigma}} \tag{6}$$

where $\hat{\mu}_1'$ and $\hat{\sigma}_1'$ are random variables which are independent of $\hat{\hat{\mu}}$ and $\hat{\hat{\sigma}}$, of which the variances and the covariance are the same as in (4). We thus immediately deduce that

$$\operatorname{var}\mu_1^* = (1-\rho_1^2)\operatorname{var}\hat{\mu}_1 + \rho_1^2 a_1^2 \operatorname{var}\hat{\hat{\mu}}, \quad \operatorname{var}\sigma_1^* = (1-\rho_1^4)\operatorname{var}\hat{\sigma}_1 + \rho_1^4 a_1^2 \operatorname{var}\hat{\hat{\sigma}} \tag{7}$$

or again

$$\operatorname{var}\mu_1^* = (1-\rho_1^2)\operatorname{var}\hat{\mu}_1 + \rho_1^2 \operatorname{var}\hat{\hat{\mu}}_1, \quad \operatorname{var}\sigma_1^* = (1-\rho_1^4)\operatorname{var}\hat{\sigma}_1 + \rho_1^4 \operatorname{var}\hat{\hat{\sigma}}_1 \tag{8}$$

with $\operatorname{var}\hat{\hat{\mu}}_1 = a_1^2 \operatorname{var}\hat{\hat{\mu}}$ and $\operatorname{var}\hat{\hat{\sigma}}_1 = a_1^2 \operatorname{var}\hat{\hat{\sigma}}$, i.e. that $\operatorname{var}\hat{\hat{\mu}}_1$ and $\operatorname{var}\hat{\hat{\sigma}}_1$ are the values taken by $\operatorname{var}\hat{\hat{\mu}}$ and $\operatorname{var}\hat{\hat{\sigma}}$ when we put $\sigma = \sigma_1$.

We also conclude from this that the estimates μ_1^* and σ_1^* become more accurate than the estimates $\hat{\mu}_1$ and $\hat{\sigma}_1$ as soon as $n > n_1$.

2.3.2.1.1 **Example 24.** *The annual rainfall on the drainage basin of the Senne at Vilvoorde (Belgium)*

Table 24 gives the annual precipitation on the drainage of the Senne at Vilvorde calculated from the accumulated total precipitation recorded in the climatological stations of this basin between 1951 and 1970. This table also gives the total annual precipitation recorded at Uccle for the same period. Since the station at Uccle is one of these stations and since the basin is relatively small in area (1.006 km^2), we can expect that a good correlation will exist between the total precipitation at Uccle and the precipitation in the basin considered.

Since we have, for Uccle, a series of rainfall observations corrected for heterogeneities between 1833 and 1970, we can, by applying the above methods, determine the parameters of the distribution of the annual rainfall with greater accuracy than that of the estimates derived solely from the sample over twenty years.

Since we know that the assumption of nomality is acceptable for the two variables, the application of relations 2.3.2.1(2) to the data in Table 24 leads, with the correction for the bias of the standard deviation [see 2.2.1(14)], to

$$\hat{\mu} = (879 + 922 + \cdots + 727)/20 = 802$$

$$\hat{\mu}_1 = (901 + 839 + \cdots + 742)/20 = 800$$

$$\hat{\sigma}^2 = [(879 - 802)^2 + (922 - 802)^2 + \cdots + (727 - 802)^2]/(20 - 1{,}5) = 20606{,}05$$

$$\hat{\sigma} = 143{,}55$$

$$\hat{\sigma}_1^2 = [(901 - 800)^2 + (939 - 800)^2 + \cdots + (742 - 800)^2]/18{,}5 = 20334{,}11$$

$$\hat{\sigma}_1 = 142{,}60$$

$$\hat{\rho}_1 = [(879 - 802)(901 - 800) + (922 - 802)(939 - 800) + \cdots$$
$$+ (727 - 802)(742 - 800)]/(18{,}5 \times 143{,}55 \times 142{,}60) = 0{,}9746 \qquad (1)$$

Furthermore, the total rainfall series at Uccle between 1833 and 1970 leads to the estimates

$$\mu = \hat{\mu} = 781 \quad \text{and} \quad \sigma = \hat{\sigma} = 117{,}9 \qquad (2)$$

It then follows, using 2.3.2.1(1) and (3), that we have

$$\mu_1^* = 800 - 0{,}9746(142{,}60/143{,}55)(802 - 781) = 779{,}7$$

$$\sigma_1^* = 142{,}60(117{,}9/143{,}55)^{(0{,}9746)^2} = 118{,}3$$

TABLE 24. – Total annual precipitation recorded at Uccle and annual precipitation on the drainage basin of the Senne at Vilvorde, in mm, between 1951 and 1970.

Year	Uccle	The Senne at Vilvoorde
1951	879	901
1952	922	939
1953	556	557
1954	740	774
1955	616	656
1956	794	786
1957	786	843
1958	834	819
1959	561	556
1960	963	918
1961	904	909
1962	862	800
1963	713	677
1964	786	772
1965	1074	1088
1966	1056	1049
1967	707	722
1968	777	776
1969	777	713
1970	727	742

Since, moreover, we have

$$\text{var } \hat{\mu}_1 = (118{,}3)^2/20 = 699{,}7, \quad \text{var } \hat{\sigma}_1 = (\text{var } \hat{\mu}_1)/2 = 349{,}9$$

$$\text{var } \hat{\hat{\mu}}_1 = (118{,}3)^2/138 = 101{,}4, \quad \text{var } \hat{\hat{\sigma}}_1 = (\text{var } \hat{\hat{\mu}}_1)/2 = 50{,}7$$

we can deduce, using 2.3.2.1(8)

$$\text{var } \mu_1^* = [1 - (0{,}9746)^2]\,699{,}7 + (0{,}9746)^2\,101{,}4 = 131{,}4$$

$$\text{var } \sigma_1^* = [1 - (0{,}9746)^4]\,349{,}9 + (0{,}9746)^4\,50{,}7 = 80{,}0$$

Since var $\hat{\mu}_1/\text{var } \mu_1^* = 699{,}7/131{,}4 = 5{,}3$ and var $\hat{\sigma}_1/\text{var } \sigma_1^* = 349{,}9/80{,}0 = 4{,}4$ we can assume that the normal μ_1 is known with the same accuracy as if it had been deduced from a series over $20 \times 5{,}3 = 106$ years, whereas for the standard deviation σ_1 the accuracy is equivalent to that given by a series over $20 \times 4{,}4 = 88$ years. We thus see that the increase in accuracy obtained is considerable.

2.3.2.2 Case in which the estimators of the location and scale parameters are linear

When the parameters of the marginal distributions of x and y are estimated by means of linear estimators, if x_j and y_j, $j = 1, 2, \ldots, n_1$, are the values of x and of y of the sample of size n_1 arranged in order of magnitude, the estimators of the location and scale parameters are of the form

$$\hat{\sigma} = \Sigma \lambda_j x_j, \quad \hat{\sigma}_1 = \Sigma \lambda_j y_j, \quad \hat{\mu} = \Sigma \nu_j x_j, \quad \hat{\mu}_1 = \Sigma \nu_j y_j \tag{1}$$

where the coefficients λ_j and ν_j have well-defined values for each value of n_1. Under these conditions, if σ and μ are known and if $\hat{\rho}_1$ denotes the usual estimate of the coefficient of correlation, as defined in 2.3.2.1(2), more accurate estimates than those given in (1) are obtained by means of the relations

$$\sigma_1^* = \hat{\sigma}_1 (\sigma/\hat{\sigma})^{\hat{\rho}_1}, \quad \mu_1^* = \hat{\mu}_1 - \hat{\rho}_1 a_1^* (\hat{\mu} - \mu) \tag{2}$$

with $a_1^* = \sigma_1^*/\sigma$, estimates for which we have :

$$\text{var } \sigma_1^* = (1 - \rho_1^2)\,\text{var } \hat{\sigma}_1, \quad \text{var } \mu_1^* = (1 - \rho_1^2)\,\text{var } \hat{\mu}_1$$
$$\text{cov}(\sigma_1^*, \mu_1^*) = (1 - \rho_1^2)\,\text{cov}(\hat{\sigma}_1, \hat{\mu}_1) \tag{3}$$

Furthermore, if the values σ and μ are themselves estimates $\hat{\hat{\sigma}}$ and $\hat{\hat{\mu}}$ deduced from a series of observations of x of size n, then the relations (2) remain applicable. In this case switching to differentials leads to

$$d\sigma_1^* = d\hat{\sigma}_1' + \rho_1 a_1 d\hat{\hat{\sigma}}, \quad d\mu_1^* = d\hat{\mu}_1' + \rho_1 a_1 d\hat{\hat{\mu}}$$

where $\hat{\sigma}_1'$ and $\hat{\mu}_1'$ denote random variables which are independent of $\hat{\hat{\mu}}$ and of $\hat{\hat{\sigma}}$, of which the variances and the covaraiance are the same as in (3), and we have :

$$\text{var } \sigma_1^* = (1 - \rho_1^2)\,\text{var } \hat{\sigma}_1 + \rho_1^2 a_1^2\,\text{var } \hat{\hat{\sigma}},$$
$$\text{var } \mu_1^* = (1 - \rho_1^2)\,\text{var } \hat{\mu}_1 + \rho_1^2 a_1^2\,\text{var } \hat{\hat{\mu}},$$
$$\text{cov}(\sigma_1^*, \mu_1^*) = (1 - \rho_1^2)\,\text{cov}(\hat{\sigma}_1, \hat{\mu}_1) + \rho_1^2 a_1^2\,\text{cov}(\hat{\hat{\sigma}}, \hat{\hat{\mu}}) \tag{5}$$

or again

$$\operatorname{var}\sigma_1^* = (1-\rho_1^2)\operatorname{var}\hat{\sigma}_1 + \rho_1^2 \operatorname{var}\hat{\hat{\sigma}}_1,$$

$$\operatorname{var}\mu_1^* = (1-\rho_1^2)\operatorname{var}\hat{\mu}_1 + \rho_1^2 \operatorname{var}\hat{\hat{\mu}}_1,$$

$$\operatorname{cov}(\sigma_1^*, \mu_1^*) = (1-\rho_1^2)\operatorname{cov}(\hat{\sigma}_1, \hat{\mu}_1) + \rho_1^2 \operatorname{cov}(\hat{\hat{\sigma}}_1, \hat{\hat{\mu}}_1) \tag{6}$$

where var $\hat{\hat{\sigma}}_1$, var $\hat{\hat{\mu}}_1$ and cov$(\hat{\hat{\sigma}}_1$, $\hat{\hat{\mu}}_1)$ are the values taken by var $\hat{\hat{\sigma}}$, var $\hat{\hat{\mu}}$ and cov $(\hat{\hat{\sigma}}, \hat{\hat{\mu}})$ when we put $\sigma = \sigma_1$.

It then follows that the estimates (2) are more accurate than the estimates (1) when $n > n_1$, as long as the variances and the covariance of the estimates (1) are decreasing functions of the sample size n_1.

Furthermore, if we put

$$y^* = \sigma_1^* t + \mu_1^* \,,\ \hat{y} = \hat{\sigma}_1 t + \hat{\mu}_1 \quad \text{and} \quad \hat{\hat{y}} = \hat{\hat{\sigma}}_1 t + \hat{\hat{\mu}}_1, \tag{7}$$

then (6) allows us to write

$$\operatorname{var} y^* = (1-\rho_1^2)\operatorname{var}\hat{y} + \rho_1^2 \operatorname{var}\hat{\hat{y}}_1 \tag{8}$$

In particular, when the variances and the covariance of the linear estimators are inversely proportional to the sample size of the series of observations, we have :

$$n_1 \operatorname{var}\hat{y} = n \operatorname{var}\hat{\hat{y}} \tag{9}$$

and relation (8) becomes

$$\operatorname{var} y^* = \left[1 - \rho_1^2\left(1 - \frac{n_1}{n}\right)\right] \operatorname{var}\hat{y} \tag{10}$$

It follows that, in this case, the accuracy of the estimate of the quantile y^* defined in (7) is equivalent to that which would be obtained directly from the estimates given by a series of observations of y of length kn_1, with

$$k = 1/\left[1 - \rho_1^2\left(1 - \frac{n_1}{n}\right)\right] \tag{11}$$

2.3.2.2.1 **Example 25.** *The annual maximum of the daily rainfall in the stations neighbouring the station at Uccle. Estimation of the parameters of the distribution.*

Table 25 gives the values of the annual maximum rainfall at Uccle (Fs 1) and at the station Fs 19 located 2,5 km from the station at Uccle, for the years 1951 to 1970, as well as the values of this maximum for station Fs 9 located 12 km from Uccle, for the years 1961 to 1969.

Using the series available, the problem of estimating the parameters of the distribution for the variable at station Fs 9 can be considered in two ways : either we can estimate these parameters directly as a function of those at Uccle, or they can be estimated from the preliminary estimates of those at station Fs 19.

As we have seen in example 11 in 2.2.3.1.1. that the natural logarithm of this variable has a double-exponential distribution, we find that we here have a case in which the estimators of the location and scale parameters are linear and in which relations 2.3.2.2(10) and (11) are applicable.

If then we denote by x, y_1 , and y_2 the logarithm of the element considered for the station at Uccle, for station Fs 19 and for station Fs 9, respectively, and if ρ_1 and ρ_2 are the coefficients of correlation of stations Fs 19 and Fs 9 with the station at Uccle, then the values k_1 and k_2 of k in 2.3.2.2(11) are, for each of these stations, given respectively by

$$k_i = 1/\left[1 - \rho_i^2\left(1 - \frac{n_i}{n}\right)\right], \quad i = 1; 2 \tag{1}$$

with $n_1 = 20$, $n_2 = 9$ and $n = 60$. We find that the estimate based on the relations 2.3.2.2(2), and using the correlation with the station at Uccle, leads to an estimation accuracy equivalent to that given by series with sample sizes

$$n_1' = k_1 n_1 \quad \text{and} \quad n_2' = k_2 n_2 \tag{2}$$

On the other hand, if we proceed to estimate the parameters of station Fs 9 by means of relations 2.3.2.2(2) through the intermediary of the improved estimate of the parameters of station Fs 19, then the variances and the covariance of these estimates are again given to a good approximation by relations analogous to relations 2.3.2.2(6); i.e., if ρ_{12} is the coefficient of correlation between the stations considered, then the value k_{12} of k

$$k_{12} = 1/[\,1 - \rho_{12}^2 (1 - \frac{n_2}{n_1'})\,] \tag{3}$$

will lead for these estimates to the equivalent sample size

$$n_2'' = k_{12} n_2 \tag{4}$$

Consequently, the most accurate estimate is the one corresponding to the larger of the values k_2 and k_{12}. Under these conditions, we find for Fs 1 and Fs 19, using the natural logarithms of the maximum rainfall between 1951 and 1970,

$$(\log 282 + \log 511 + \cdots + \log 291)/20 = 5{,}8326$$

$$(\log 304 + \log 597 + \cdots + \log 504)/20 = 5{,}9315$$

$$[(\log 282 - 5{,}8326)^2 + (\log 511 - 5{,}8326)^2 + \cdots + (\log 291 - 5{,}8326)^2]/20 = 0{,}136011$$

$$[(\log 304 - 5{,}9315)^2 + (\log 597 - 5{,}9315)^2 + \cdots + (\log 504 - 5{,}9315)^2]/20 = 0{,}128710$$

$$[(\log 282 - 5{,}8326)(\log 304 - 5{,}9315) + (\log 511 - 5{,}8326)(\log 597 - 5{,}9315) + \cdots$$

$$+ (\log 291 - 5{,}8326)(\log 504 - 5{,}9315)]/20 = 0{,}114477$$

TABLE 25. – Annual maximum of the daily rainfall (in 0,1mm) at stations Fs 1, Fs 9 and Fs 19

Year	Fs 1	Fs 19	Fs 9
1951	282	304	—
1952	511	597	—
1953	375	562	—
1954	343	402	—
1955	222	204	—
1956	356	329	—
1957	342	318	—
1958	243	282	—
1959	203	262	—
1960	480	518	—
1961	324	330	335
1962	596	495	343
1063	604	605	700
1064	270	373	651
1965	458	483	394
1966	398	405	380
1967	216	200	168
1968	197	252	183
1969	544	540	539
1970	291	504	—

From this we obtain

$$\hat{\rho}_1 = 0{,}114477/\sqrt{0{,}136011 \times 0{,}128710} = 0{,}865$$

Similarly, using the data between 1961 and 1969, we have

$$\hat{\rho}_2 = 0{,}664 \quad \text{and} \quad \hat{\rho}_{12} = 0{,}836$$

The values of k_1 and k_2 given by relation (1) are thus :

$$k_1 = 1\Big/\left[1 - (0{,}865)^2\left(1 - \frac{20}{60}\right)\right] = 1/0{,}501 = 1{,}995$$

$$k_2 = 1\Big/\left[1 - (0{,}664)^2\left(1 - \frac{9}{60}\right)\right] = 1/0{,}625 = 1{,}599 \tag{5}$$

so that we have

$$n_1' = 1{,}995 \times 20 = 39{,}9 \quad \text{and} \quad n_2' = 1{,}599 \times 9 = 14{,}4 \tag{6}$$

Similarly, (3) gives

$$k_{12} = 1/\left[1 - (0{,}836)^2\left(1 - \frac{9}{39{,}9}\right)\right] = 1/0{,}459 = 2{,}18 \tag{7}$$

so that we have

$$n_2'' = 2{,}18 \times 9 = 19{,}6 \tag{8}$$

By comparing relations (6) and (8), we find that the second method of estimation is the more accurate. Furthermore, relation (8) shows that the accuracy obtained is equivalent to that given directly by a sample whose size would be approximately double that of the initial sample size.

Furthermore, using the estimators given in 2.2.3.1, we deduce from the observations between 1951 and 1970, for the variables x and y_1 ,

$$\hat{\sigma} = 0{,}3178 \quad \hat{\mu} = 5{,}649 \quad \hat{\sigma}_1 = 0{,}3233 \quad \hat{\mu}_1 = 5{,}745 \tag{9}$$

On the basis of the values obtained in 2.2.3.1.1

$$\sigma = 0{,}2421 \quad \text{and} \quad \mu = 5{,}612 \tag{10}$$

we obtain, using relations 2.3.2.2(2)

$$\begin{aligned} \sigma_1^* &= 0{,}3233(0{,}2421/0{,}3178)^{0{,}865} \\ &= 0{,}2555 \end{aligned}$$

$$\begin{aligned} \mu_1^* &= 5{,}745 - 0{,}865(0{,}2555/0{,}2421)(5{,}649 - 5{,}612) \\ &= 5{,}711 \end{aligned} \tag{11}$$

Similarly, the series of observations between 1961 and 1969 gives for the variables y_1 and y_2

$$\hat{\sigma}_1 = 0{,}3534 \ , \quad \hat{\mu}_1 = 5{,}755 \ , \quad \hat{\sigma}_2 = 0{,}4645 \ , \quad \hat{\mu}_2 = 5{,}645 \tag{12}$$

from which, using the same formulae 2.3.2.2(2), we have

$$\begin{aligned} \sigma_2^{**} &= 0{,}4645(0{,}2555/0{,}3534)^{0{,}836} \\ &= 0{,}3542 \end{aligned}$$

$$\begin{aligned} \mu_2^{**} &= 5{,}645 - 0{,}836(0{,}3542/0{,}2555)(5{,}755 - 5{,}711) \\ &= 5{,}594 \end{aligned} \tag{13}$$

2.3.2.3 Estimation of missing values

The problem of estimating the missing values can take a variety of forms depending on whether we have two series in correlation or a larger number of series, and on the manner in which the common periods of observation occur.

To examine this, we shall however confine ourselves to considering the two following cases which are typical :

(a) We have two series of observations x and y_1, the series x being long-term and the series y_1 being of short duration and recent. The parameters of the series y_1 are estimated by using the correlation with the series x and the missing values of y_1 are calculated as a function of the corresponding values of x.

(b) We have three series of observations x , y_1 and y_2, the series x being long-term, the series y_1 and y_2 both being short, the first being an old series whilst the second is recent. The parameters of the series y_1 and y_2 are estimated by using the correlation with the series x, and the missing values of y_2 are calculated as a function of the corresponding values of y_1.

As was indicated in 2.3.1, in the absence of any error of estimation on the parameters σ, μ, σ_1 and μ_1, the quantile method leads to an error whose variance is given in 2.3.1(11). On the contrary, when these parameters are estimated, this error is increased by that contained in the estimates.

In case (a), denoting by (y_1) the estimate of the missing value y_1 and differentiating the relations

$$(y_1)=\sigma_1 t+\mu_1 \quad \text{and} \quad t=(x-\mu)/\sigma \tag{1}$$

we obtain for the estimation error $d(y_1)$

$$d(y_1)=dy_1-a_1\,dx \tag{2}$$

where dy_1 and dx are the errors associated with the quantiles y_1 and x for given t which result from the estimation of the parameters.

For normally distributed variables, we obtain

$$\operatorname{var}(y_1)=(1-\rho_1^2)\operatorname{var}\hat{\mu}_1+(1-\rho_1)^2\operatorname{var}\hat{\hat{\mu}}_1+t^2[(1-\rho_1^4)\operatorname{var}\hat{\sigma}_1+(1-\rho_1^2)^2\operatorname{var}\hat{\hat{\sigma}}_1] \tag{3}$$

whereas with linear estimators we find

$$\operatorname{var}(y_1)=(1-\rho_1^2)\operatorname{var}\hat{y}_1+(1-\rho_1)^2\operatorname{var}\hat{\hat{y}}_1 \tag{4}$$

knowing that var $\hat{\hat{\mu}}_1$, var $\hat{\hat{\sigma}}_1$ and var $\hat{\hat{y}}_1$ are the values taken by var $\hat{\hat{\mu}}$, var $\hat{\hat{\sigma}}$ and var $\hat{\hat{y}}$, when we put $\sigma=\sigma_1$.

In case (b), making the indices 1 and 2 correspond to the variables y_1 and y_2 respectively, we find similarly

$$d(y_2)=dy_2-a_{21}\,dy_1 \tag{5}$$

where $a_{21}=\sigma_2/\sigma_1$.

We then obtain, for normally distributed variables,

$$\begin{aligned}\operatorname{var}(y_2)=&\left\{1-\rho_2^2-2\frac{n'}{n_1}(\rho_{12}-\rho_1\rho_2)+\frac{t^2}{2}\left[1-\rho_2^4-2\frac{n'}{n_1}(\rho_{12}^2-\rho_1^2\rho_2^2)\right]\right\}\operatorname{var}\hat{\mu}_2\\&+\left[(1-\rho_1^2)+\frac{t^2}{2}(1-\rho_1^4)\right]\operatorname{var}\hat{\mu}_{2(1)}\\&+\left[(\rho_2-\rho_1)^2+\frac{t^2}{2}(\rho_2^2-\rho_1^2)^2\right]\operatorname{var}\hat{\hat{\mu}}_2\end{aligned} \tag{6}$$

where n' is the length of the observation period common to the series y_1 and y_2, and where var $\hat{\mu}_{2(1)}$ and var $\hat{\hat{\mu}}_2$ are the values taken by var $\hat{\mu}_1$ and var $\hat{\hat{\mu}}$ when we put respectively $\sigma_1 = \sigma_2$ and $\sigma = \sigma_2$.

Similarly, for linear estimators, we can write

$$\text{var}(y_2) = \left[1 - \rho_2^2 - 2\frac{n'}{n_1}(\rho_{12} - \rho_1\rho_2)\right]\text{var}\,\hat{y}_2 + (1 - \rho_1^2)\,\text{var}\,\hat{y}_{2(1)} + (\rho_2 - \rho_1)^2\,\text{var}\,\hat{\hat{y}}_2 \tag{7}$$

These relations are deduced directly from the fact that for normal variables we have

$$\text{cov}(\hat{\mu}_1, \hat{\mu}_2) = \frac{n'}{n_1 n_2}\rho_{12}\sigma_1\sigma_2 \qquad \text{and} \qquad \text{cov}(\hat{\sigma}_1, \hat{\sigma}_2) = \frac{1}{2}\cdot\frac{n'}{n_1 n_2}\rho_{12}^2\sigma_1\sigma_2 \tag{8}$$

whereas in the case of linear estimators, we obtain an analogous relation for cov$(\hat{y}_1, \hat{y}_2)$.

Furthermore, we note that relations (3) and (4) can be deduced from relations (6) and (7) as a consequence of case (*a*) being a special case of case (*b*) in which $x \equiv y_1$, i.e. with $n = n_1$, $n' = n_2$, $\rho_1 = 1$ and $\rho_{12} = \rho_2$.

2.3.2.3.1 **Example 25** (continued). *The annual maximum of the daily rainfall in the stations neighbouring the station at Uccle. Estimation of missing values*

The estimates of the parameters of the distribution functions for the variables x, y_1 and y_2 obtained in 2.3.2.2.1 allow us to solve the problem of calculating the missing values of y_2 in three different ways, depending on whether we write

$$(y_2)_1 = y_2^*(t) \quad \text{with} \quad t = (x - \mu)/\sigma \tag{1}$$

$$(y_2)_2 = y_2^*(t) \quad \text{with} \quad t = (y_1 - \mu_1^*)/\sigma_1^* \tag{2}$$

or

$$(y_2)_3 = y_2^{**}(t) \quad \text{with} \quad t = (y_1 - \mu_1^*)/\sigma_1^* \tag{3}$$

it being understood that the variance of the error of estimation for the quantile, var (y_2), is to be added to the quantity $2(1 - \rho_2)$ var y_2 or $2(1 - \rho_{12})$ var y_2 respectively for case (*a*) or (*b*), which is the variance of the error due to use of the quantile method. So the method does not give any information on the missing values unless we have

$$\text{var}(y_2) + 2(1 - \rho)\,\text{var}\,y_2 < \text{var}\,y_2 \tag{4}$$

where var y_2 denotes the variance of the distribution of y_2, and where ρ is written for ρ_2 or ρ_{12} depending on the case, i.e. if we have

$$\text{var}(y_2) < (2\rho - 1)\,\text{var}\,y_2 \tag{5}$$

We recognize that relations (1) and (3) correspond in 2.3.2.3 to case (*a*), and relation (2) to case (*b*). So using 2.3.2.3(4) we obtain for the estimate (1)

$$\text{var}(y_2)_1 = (1 - \rho_2^2)\,\text{var}\,\hat{y}_2 + (1 - \rho_2)^2\,\text{var}\,\hat{\hat{y}}_2 \tag{6}$$

or with

$$\text{var}\,\hat{\hat{y}}_2 = (9/60)\,\text{var}\,\hat{y}_2,$$

$$\text{var}(y_2)_1 = \left[1 - \rho_2^2 + \frac{9}{60}(1 - \rho_2)^2\right]\text{var}\,\hat{y}_2, \tag{7}$$

while for the estimate (3), we have

$$\text{var}\,(y_2)_3 = (1 - \rho_{12}^2)\text{var}\,\hat{y}_2 + (1 - \rho_{12})^2\,\text{var}\,\hat{\hat{y}}_2 \tag{8}$$

where, by virtue of 2.3.2.2.1(6), we have

$$\operatorname{var} \hat{\hat{y}}_2 = \frac{9}{39{,}9} \operatorname{var} \hat{y}_2 = 0{,}226 \operatorname{var} \hat{y}_2 \tag{9}$$

so that

$$\operatorname{var}(y_2)_3 = [(1 - \rho_{12}^2) + 0{,}226\,(1 - \rho_{12})^2] \operatorname{var} \hat{y}_2 \tag{10}$$

Finally, for the estimate (2), using

$$\operatorname{var} \hat{y}_{2(1)} = \frac{9}{20} \operatorname{var} \hat{y}_2 \quad \text{and} \quad \operatorname{var} \hat{\hat{y}}_2 = \frac{9}{60} \operatorname{var} \hat{y}_2,$$

2.3.2.3(7) becomes

$$\operatorname{var}(y_2)_2 = \left\{1 - \rho_2^2 - 2\frac{9}{20}(\rho_{12} - \rho_1\rho_2) + \frac{9}{20}(1 - \rho_1^2) + \frac{9}{60}(\rho_2 - \rho_1)^2\right\} \operatorname{var} \hat{y}_2 \tag{11}$$

With the values $\rho_1 = 0{,}865$, $\rho_2 = 0{,}664$ and $\rho_{12} = 0{,}836$ found in 2.3.2.2.1, we finally obtain

$$\operatorname{var}(y_2)_1 = \left[1 - (0{,}664)^2 + \frac{9}{60}(1 - 0{,}664)^2\right] \operatorname{var} \hat{y}_2 = 0{,}576 \operatorname{var} \hat{y}_2$$

$$\operatorname{var}(y_2)_3 = [0{,}301 + 0{,}006] \operatorname{var} \hat{y}_2 = 0{,}307 \operatorname{var} \hat{y}_2$$

$$\operatorname{var}(y_2)_2 = [0{,}324 + 0{,}113 + 0{,}006] \operatorname{var} \hat{y}_2 = 0{,}443 \operatorname{var} \hat{y}_2 \tag{12}$$

We conclude that method (3) for calculating the missing values leads to the most accurate values. Since for the distribution considered $\operatorname{var} y_2 = 1{,}645\,\sigma_2^2$, then using 2.2.3.1(6), condition (5) becomes

$$\operatorname{var}(y_2)_3 = 0{,}307 \operatorname{var} \hat{y}_2 =$$

$$= 0{,}307(1{,}10866 + 0{,}60793t^2 + 2 \times 0{,}25702t)\frac{\sigma_2^2}{9} < (2 \times 0{,}836 - 1)1{,}645\,\sigma_2^2 \tag{13}$$

that is

$$0{,}0207\,t^2 + 0{,}0175\,t - 1{,}068 < 0 \tag{14}$$

this relation being satisfied in the interval

$$-7{,}62 < t < 6{,}77 \tag{15}$$

that is to say, over practically the entire interval of the observed values.

Under these conditions, the relation giving the missing values of y_2 becomes

$$(y_2) = a_{21}y_1 + b_{21} \qquad \text{where} \qquad a_{21} = \sigma_2^{**}/\sigma_1^*$$

$$\text{and} \quad b_{21} = \mu_2^{**} - a_{21}\mu_1^*$$

or, using 2.3.2.2.1(11) and (13),

$$(y_2) = 1{,}39\,y_1 - 2{,}32 \tag{16}$$

In particular, for the missing value in 1970 at station Fs 9, we have

$$(y_2) = 1{,}39 \log (504) - 2{,}32 = 6{,}329$$

so that this missing value can be estimated as

$$\exp(6{,}329) = 561 \tag{17}$$

For the calculation of the total error of estimation, we have with $t = (6{,}239 - 5{,}594)/0{,}3542 = 2{,}075$ and using (13)

$$\text{var}(y_2) = 1{,}471\ (0{,}3542)^2/9 = 0{,}02051 \tag{18}$$

Since we also have

$$2(1 - \rho_{12})\ \text{var}\ y_2 = 0{,}328 \times 1{,}645 \times (0{,}3542)^2 = 0{,}06769 \tag{19}$$

we obtain a total of 0,08820 , i.e. a standard deviation equal to 0,30. The standard deviation of the total error of estimation associated with the value found in (17) is thus of the order of 30%.

2.3.3. The method of differences

When the mean of the distribution of the variable x is known and when we have a short series of corresponding values of the variable x and of a variable y : x_i , y_i , $i = 1, 2, \ldots, n$, a simple method for estimating the mean value of the distribution of the variable y consists of calculating the mean value $(\hat{\mu}_1 - \hat{\mu})$ of the differences between the corresponding values $(y_i - x_i)$ and adding to it the known mean value of x.

If μ_1 and μ are respectively the mean values of the distributions of y and of x , this comes down to making

$$\mu_1^* = \mu + (\hat{\mu}_1 - \hat{\mu}) = \hat{\mu}_1 - (\hat{\mu} - \mu) \tag{1}$$

whereas the optimal estimation is given by the relation

$$\mu_1^* = \hat{\mu}_1 - a_1 \rho\, (\hat{\mu} - \mu) \tag{2}$$

The comparison of the two relations shows that if $a_1 = 1$, the two relations differ by progressively smaller amounts as ρ becomes larger.

Furthermore, since, with $a_1 = 1$, we have for relation (1)

$$\text{var}\ \mu_1^* = 2(1 - \rho)\ \text{var}\ \hat{\mu}_1 \tag{3}$$

and for relation (2)

$$\text{var}\ \mu_1^* = (1 - \rho^2)\ \text{var}\ \hat{\mu}_1 \tag{4}$$

we thereby deduce that the ratio $2/(1 + \rho)$ of the variances of the estimation errors of the two methods varies by a factor of 2 when ρ varies from 1 to 0.

We conclude that the method of differences is to be preferred in view of its simplicity, as long as the correlations are high or the variance of the differences is small.

2.3.4 Reference works consulted

The methods of estimation when series are correlated have been taken from [41]. In the case of example 24 we have used hydrological data given in [42] and the rainfall series for Uccle obtained after applying the corrections calculated in [43]. The data for example 25 were extracted from the climatological archives of Uccle.

3. ESTIMATION IN THE CASE OF RECURRENT SERIES

A time series x_i, $i = 1, 2, ..., n$, is random recurrent (also autoregressive) when the terms of this series satisfy a linear relation of the form

$$x_{i+k} = a_0 + a_1 x_i + a_2 x_{i+1} + ... + a_k x_{i+k-1} + d_i, \quad i = 1, 2, ..., n-k \tag{1}$$

where the elements $a_0, a_1, ..., a_k$ are assumed constant and where the d_i are random quantities with zero mean which possess a well-defined joint distribution. If $a_1 \neq 0$, equation (1) is said to be of order k; it is homogeneous if a_0 and d_i are identically zero.

The most simple case is that of the series $x_i = x_0$, where x_0 is a constant. This series satisfies the homogeneous recurrence relation of first order with constant coefficients

$$x_{i+1} = x_i \tag{2}$$

Similarly, the series $x_i = e_i$, which is simple random whenever the quantities e_i follow each other independently and come from the same population, is another simple example of a recurrent series. This satisfies the non-homogeneous relation of first order with constant coefficients

$$x_{i+1} = x_i + d_i \tag{3}$$

where the d_i are random quantities for which, denoting by var e the variance of the quantities e_i, we have

$$\mathrm{E}(d_i) = 0, \ \mathrm{var}\ d_i = 2\ \mathrm{var}\ e, \ \mathrm{cov}(d_i, d_{i+1}) = -\ \mathrm{var}\ e$$

$$\mathrm{cov}(d_i, d_{i+j}) = 0 \ \text{ for } j \geqslant 2 \tag{4}$$

We thus see that the scheme offered by the class of recurrent series is more general than that of the simple random series considered hitherto. Note, however that the range of application of the former is at least as large as that of the latter since it covers the case of all the series of hourly, multi-hourly or daily observations, this being due both to the existence of daily and seasonal variations and to the persistency governing, at various scales, the evolution of the observed meteorological phenomena.

For the applications, we recall that the general solution of the inhomogeneous equation of order k is given by the sum of the general solution of the corresponding homogeneous equation and a particular solution of the non-homogeneous equation. More precisely, if $x_i^{(1)}, x_i^{(2)}, ..., x_i^{(k)}$ are k linearly-independent solutions of the homogeneous equation

$$x_{i+k} = a_1 x_i + a_2 x_{i+1} + ... + a_k x_{i+k-1} \tag{5}$$

and if e_i is a random series which satisfies equation (1), then any solution of x_i of equation (1) can be expressed in the form

$$x_i = c_1 x_i^{(1)} + c_2 x_i^{(2)} + \ldots + c_k x_i^{(k)} + e_i \tag{6}$$

by attributing suitable values to the constants $c_1, c_2, \ldots, c_k$; that is, x_i is the sum of a deterministic component which satisfies the corresponding homogeneous equation and a random part which satisfies the non-homogeneous equation.

Consequently, for a given series x_i, two problems can be considered, depending on whether we wish to find the recurrence relation (1) or alternatively to express the series x_i in the form (6).

The first problem is that of determining an arbitrary term of the series from the k preceding elements, which is made possible to within an error d_i by the knowledge of equation (1).

The second problem concerns the separation of the deteministic component of the series x_i from its random part.

In each case, we have to estimate either the constants $a_0, a_1, \ldots, a_k$ or the constants $c_1, c_2, \ldots, c_k$ the most appropriate method for doing so being the method of least squares as described in 2.2.4.

For the first problem, if we denote by x, θ, d and u respectively the vectors and the matrix

$$x = \begin{Vmatrix} x_{k+1} \\ x_{k+2} \\ \vdots \\ x_n \end{Vmatrix} \qquad \theta = \begin{Vmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{Vmatrix} \qquad d = \begin{Vmatrix} d_1 \\ d_2 \\ \vdots \\ d_{n-k} \end{Vmatrix}$$

$$u = \begin{Vmatrix} 1 & x_1 & x_2 \ldots & x_k \\ 1 & x_2 & x_3 & x_{k+1} \\ \vdots & & & \\ 1 & x_{n-k} & x_{n-k+1} & x_{n-1} \end{Vmatrix} \tag{7}$$

then equation (1) can be written

$$x = u\theta + d \tag{8}$$

and that takes the form 2.2.4(5).

We find that if v is the covariance matrix for the elements of the vector d, then estimation of θ via the method of least squares is, as in 2.2.4(17), given by the relation

$$\hat{\theta} = (u'v^{-1}u)^{-1}u'v^{-1}x \tag{9}$$

the order of the matrix v here being $(n-k)$, and we have

$$\operatorname{var}\hat{\theta} = (u'v^{-1}u)^{-1} \tag{10}$$

The solution of the problem is then given by the estimator

$$\hat{x} = u\hat{\theta} \tag{11}$$

for which we also have

$$\operatorname{var}\hat{x} = u(\operatorname{var}\hat{\theta})u' \tag{12}$$

Furthermore, the significance tests are based on the fact that with $\hat{d}=x-u\hat{\theta}$, we have

$$x'v^{-1}x=\hat{\theta}'u'v^{-1}u\hat{\theta}+\hat{d}'v^{-1}\hat{d} \tag{13}$$

and the elements of the vector $\hat{\theta}$ possess a joint normal distribution, at least asymptotically.

It follows in particular that, under the null hypothesis $\theta=0$, the quadratic form $\hat{\theta}'u'v^{-1}u\hat{\theta}$ possesses, at least asymptotically, a χ^2 distribution with $(k+1)$ degrees of freedom and, if the joint distribution of the elements of the vector d is normal, then the quadratic form $\hat{d}'v^{-1}\hat{d}$ possesses a χ^2 distribution with $(n-2k-1)$ degrees of freedom independent of that of $\hat{\theta}'u'v^{-1}u\hat{\theta}$.

Similarly, for the second problem, introducing the matrix notations

$$x=\begin{Vmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{Vmatrix} \qquad u=\begin{Vmatrix} x_1^{(1)}x_1^{(2)}\cdots & x_1^{(k)} \\ x_2^{(1)}x_2^{(2)} & x_2^{(k)} \\ \vdots & \\ x_n^{(1)}x_n^{(2)} & x_n^{(k)} \end{Vmatrix}$$

$$\theta=\begin{Vmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{Vmatrix} \qquad e=\begin{Vmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{Vmatrix} \tag{14}$$

relation (6) can be expressed in the form

$$x=u\theta+e \tag{15}$$

from which we obtain relations analogous to relations (9), (10), (11), (12) and (13) and for which analogous remarks can be made, taking into account however that the quadratic forms of relation (13), $\hat{\theta}'u'v^{-1}u\hat{\theta}$ and $\hat{e}'v^{-1}\hat{e}$ possess numbers of degrees of freedom equal to k and $(n-k)$ respectively.

3.1 PERIODIC SERIES

When the recurrence relation 3(1) reduces to the equation

$$x_{i+k}=x_i+d_i \tag{1}$$

where $\mathrm{E}(d_i)=0$, the series x_i is said to be periodic. The deterministic or systematic component of such a series has a period k. Futhermore, the general solution 3(6) becomes

$$x_i=a_0+\sum_1^{k'-1}(a_j\sin j\alpha i+b_j\cos j\alpha i)+a_k(-1)^i+e_i \tag{2}$$

where $\alpha=2\pi/k$, $k'=k/2$ or $k'=(k+1)/2$ depending on whether k is even or odd, where a_k, is a constant which is identically zero if k is odd and in which $\mathrm{E}(e_i)=0$. In this expression, it is also useful to designate as *harmonic* j the sum $(a_j\sin j\alpha i+b_j\cos j\alpha i)$.

The two types of problem considered above reduce either, when the period k is known, to the estimation of the constants a_0, a_j , b_j and a_k , of the general solution, or, when the period k is unknown, to the actual determination of this period. Moreover, when the period is known, we can again discern two cases depending on whether the matrix v is known or not. In the latter case, we assume that the quantities e_i in (2) are independent and identically distributed, i.e. that we can write $v = I\sigma^2$ where I is the unit matrix and σ^2 an unknown constant.

For the estimation of the deterministic component, we then have the following possibilities :

1° $n = k$ with $v = I\sigma^2$, σ^2 being an unknown constant;

2° $n = n'k (n' > 1)$ with $v = I\sigma^2$, σ^2 being an unknown constant;

3° $n = k$, v being known.

As examples of application, we can cite the estimation of the true mean values of a meteorological element from :

1° The monthly normals of a meteorological element for the 12 months of the year ($k = 12$), the errors of the normals in relation to the true mean value being assumed to be independent and distributed with the same unknown variance;

2° A series of monthly averages of a meteorological element for the 12 months ($k = 12$) of n' years, the distributions of the monthly averages for each month being assumed independent and with the same unknown variance;

3° The normals of a meteorological element for each of the 12 months of the year ($k = 12$) or for each of the 24 hours ($k = 24$) of a given day of the year, the covariance matrix of the errors of the normals with respect to the true mean values being known.

As we shall see, estimation of the deterministic components of the periodic series allows us to find estimates of the true mean values which are more accurate than the initial values when $n = k$, or than those given by the *normals* deduced from the period of n' years.

The problem of seeking the deterministic component of a periodic series when the period is arbitrary and is either known or unknown will be examined later.

3.1.1 Estimation of the deterministic component of a periodic series when the period is known and the covariance matrix of the random component is unknown

As was indicated above, we assume that for the covariance matrix of the random component we have

$$v = I\sigma^2 \tag{1}$$

where I denotes the unit matrix and σ^2 is an unknown constant.

Consequently, relations 3(9), (10) and (13) simplify to

$$\hat{\theta} = (u'u)^{-1}u'x \ , \quad \text{var}\,\hat{\theta} = (u'u)^{-1}\sigma^2 \tag{2}$$

and

$$x'x = \hat{\theta}'u'u\hat{\theta} + \hat{e}'\hat{e} \tag{3}$$

Furthermore, if the length n of the series is equal to k or to a multiple of k, and if $\alpha=2\pi/k$, when k is even, we have both :

$$u=\begin{Vmatrix} 1 & \sin\alpha & \cos\alpha & \sin 2\alpha & \cos 2\alpha & \cdots & \cos(k'-1)\alpha & (-1) \\ 1 & \sin 2\alpha & \cos 2\alpha & \sin 4\alpha & \cos 4\alpha & \cdots & \cos 2(k'-1)\alpha & (-1)^2 \\ \vdots & & & & & & & \\ 1 & \sin n\alpha & \cos n\alpha & \sin 2n\alpha & \cos 2n\alpha & \cdots & \cos n(k'-1)\alpha & (-1)^n \end{Vmatrix}$$

$$u'u=\begin{Vmatrix} n & 0 & \cdots & & 0 \\ 0 & \frac{n}{2} & & & \cdot \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & \frac{n}{2} & 0 \\ 0 & & & 0 & n \end{Vmatrix}$$

and

$$(u'u)^{-1}=\begin{Vmatrix} \frac{1}{n} & 0 & \cdots & & 0 \\ 0 & \frac{2}{n} & & & \cdot \\ \cdot & & & \frac{2}{n} & 0 \\ 0 & \cdots & & 0 & \frac{1}{n} \end{Vmatrix} \tag{4}$$

whereas if k is odd, the matrix u loses its last column and the two other matrices lose their last row and last column.

If follows that the elements of the vector $\hat{\theta}$ are given by the relations :

$$\hat{a}_0=\Sigma x_i/n,$$

$$\hat{a}_j=2(\Sigma x_i \sin j\alpha i)/n, \quad \hat{b}_j=2(\Sigma x_i \cos j\alpha i)/n, \quad j=1, 2, \ldots, k'-1$$

$$\hat{a}_{k'}=\Sigma x_i(-1)^i/n \tag{5}$$

and that we also have :

$$\operatorname{var}\hat{a}_0=\operatorname{var}\hat{a}_{k'}=\sigma^2/n$$

$$\operatorname{var}\hat{a}_j=\operatorname{var}\hat{b}_j=2\sigma^2/n, \quad j=1, 2, \ldots, k'-1, \tag{6}$$

whereas the covariances of the elements $\hat{a}_0, \hat{a}_{k'}, \hat{a}_j$ and $\hat{b}_j$ taken in pairs are all equal to zero.

Finally, if we put

$$\sqrt{(u'u)} = \begin{Vmatrix} \sqrt{n} & 0 \cdots & 0 & 0 \\ 0 & \sqrt{n/2} & & \\ \cdot & & & \\ \cdot & & & \\ & & \sqrt{n/2} & 0 \\ 0 & & 0 & \sqrt{n} \end{Vmatrix} \tag{7}$$

as well as

$$s = \sqrt{(u'u)}\hat{\theta} \tag{8}$$

we have

$$\text{var } s = I\,\sigma^2 \tag{9}$$

and (3) becomes

$$x'x = s's + \hat{e}'\hat{e} \tag{10}$$

We deduce from this that the elements of the vector s : $\hat{a}_0\sqrt{n}$, $\hat{a}_j\sqrt{n/2}$, $\hat{b}_j\sqrt{n/2}$ and $\hat{a}_k\sqrt{n}$, are distributed independently with the same variance σ^2 and that if we denote these elements by s_l, $l = 1, 2, \ldots, k$, then relation (10) can be written

$$\sum_1^n x_i^2 = \sum_1^k s_l^2 + \sum_1^n \hat{e}_i^2 \tag{11}$$

Since, under the assumption of a joint normal distribution of the elements of the vector e, the quantity $\hat{e}'\hat{e}/\sigma^2 = \Sigma\,\hat{e}_j^2/\sigma^2$ possesses a χ^2 distribution with $(n-k)$ degrees of freedom, then in order to test the null hypothesis $\theta = 0$ or, what comes down to the same thing, $E(s) = 0$, we shall have to distinguish the case in which $n = k$ from that in which $n = n'k (n' > 1)$. In the former case, $\hat{e}'\hat{e}$ will in fact be identically zero, whereas in the second case $\hat{e}'\hat{e}/(n-k)$ will give an unbiased estimate of σ^2.

(1) $n = k$

Under the hypothesis $E(s) = 0$, the quantities s_l^2 form k independent estimates with one degree of freedom of the same variance σ^2. Moreover, as the quantities s_l are distributed normally or approximately normally - we can test the hypothesis $E(s) = 0$ by means of a parametric test of the homogeneity of the variances and use for this the likelihood ratio statistic derived by Bartlett. This statistic, expressed in natural logarithms, here becomes

$$M = k \log (\Sigma s_l^2/k) - \Sigma \log s_l^2 \tag{12}$$

the large values of which are significant and with which we associate the constant

$$c_1 = k - \frac{1}{k} \tag{13}$$

when using the tables of critical values given by Pearson and Hartley.

If the value of the statistic M is not significant, we conclude that the deterministic component of the series x_i is identically zero and that this series is indistinguishable from the random part e_i.

On the other hand, when the statistic M leads to the null hypotheses being rejected, we select significant elements by taking from the set of the k values of s_l^2 the highest values, and re-applying the homogeneity test to the

$(k-1)$ remaining values. If this new test again leads to the rejection of the null hypothesis, we once again take the highest value from amongst the $(k-1)$ values of s_l^2 considered, and repeat the operation until a non-significant value of the statistic is found. Note that when selecting one of the constants a_j or b_j, the selection should take the pair $(\hat{a}_j, \hat{b}_j)$ giving the highest value of the sum $(\hat{a}_j^2 + \hat{b}_j^2)$ owing to the fact that this pair completely defines the j^{th} harmonic of the deterministic component of the series. Moreover, if we put $s'^2_1 = s_1^2$, $s'^2_{k'+1} = s_k^2$, $s'^2_{j+1} = (s_{2j}^2 + s_{2j+1}^2)/2$; $j = 1, 2, \ldots,$ $(k'-1)$, with $s'^2_{k'+1} = 0$ if k is odd, we get a more powerful test by replacing in the statistic (12) $\Sigma \log s_l^2$ by $\Sigma \nu_j \log s'^2_j$, with $\nu_1 = \nu_{k'+1} = 1$ and $\nu_{j+1} = 2$, $j = 1, 2, \ldots, (k'-1)$, the constant (13) then being : $c_1 = \sum_{t=1}^{k'+1} \frac{1}{\nu_t} - \frac{1}{k}$.

The quantities s_l^2 which have not been selected give an estimate of the variance σ^2. In particular, if k_1 is their number and $\Sigma' s_l^2$ their sum, this estimate is

$$\hat{\sigma}^2 = \Sigma' s_l^2 / k_1 \tag{14}$$

(2) $n = n'k (n' > 1)$

We proceed in analogous fashion, given that under the null hypothesis $E(s) = 0$, the statistic

$$X = \frac{(\Sigma s_l^2)/k}{(\Sigma \hat{e}_i^2)/(n-k)} \tag{15}$$

has an F distribution with $\nu_1 = k$ and $\nu_2 = (n-k)$ degrees of freedom, if we can assume that the elements of the vector e are normally distributed.

Note that for the calculation of this statistic, relation (11) gives

$$\Sigma \hat{e}_i^2 = \Sigma x_i^2 - \Sigma s_l^2 \tag{16}$$

and that, for the final estimate of the variance σ^2 , we can use the relation

$$\hat{\sigma}^2 = (\Sigma x_i^2 - \Sigma'' s_l^2)/(n - k + k_1) \tag{17}$$

in which $\Sigma'' s_l^2$ denotes the sum of the quantities s_l^2 selected and k_1 the number of quantities s_l^2 which are not significant.

So, as far as the estimate of the deterministic component is concerned, we have in both cases

$$\hat{x} = u\hat{\theta} \tag{18}$$

where we have retained only the significant constants of the general solution. For the errors of estimation, in addition to relations (6), using 3(12), we also find

$$\operatorname{cov}(\hat{x}_i, \hat{x}_{i'}) = \frac{\sigma^2}{n} [\delta + \delta' \cos \pi(i' - i) + 2 \Sigma'' \cos j\alpha(i' - i)] \tag{19}$$

where $\delta = 0$ or 1 depending on whether $a_0 = 0$ or $a_0 \neq 0$, where $\delta' = 0$ or 1 depending on whether $a_{k'} = 0$ or $a_{k'} \neq 0$ and where the sum Σ'' is extended to the values of j corresponding to the harmonics selected.

It follows that if, owing to the selection of the harmonics, the error of estimation for the elements of the vector $\hat{x}$ has a variance less than that of the random part of the elements of the vector x, on the other hand, the values of this error no longer form a simple random series.

For the actual selection, reference is made to the instructions given in 1.3.3. In particular, when $n = k$, since the statistic (12) has approximately a χ^2 distribution with $(k-1)$ degrees of freedom, for the level $\alpha_0 = 0{,}05$, the value after each selection must be compared to an appropriate critical level. These levels can be obtained by subtracting from the initial critical value M_c the ratio M_c/χ_c^2 where χ_c^2 is the corresponding critical value of the variable χ^2 with $(k-1)$ degrees of freedom. This is done a number of times equal to the number of components s_l^2 selected. If s_0^2 is the last component selected or the mean value of the last pair selected, then the statistic

$$X = s_0^2 / \hat{\sigma}^2 \tag{20}$$

where $\hat{\sigma}^2$ is the estimate (14) of the final estimate of the variance, establishes the significance of all the components selected once it exceeds the critical value of the corresponding F statistic for the level α obtained from the equation

$$1 - (1-\alpha)^k = \alpha_0 \quad \text{or} \quad 1 - (1-\alpha)^{k/2} = \alpha_0 \tag{21}$$

depending on the case, α_0 being the initial level of the multiple test.

In the case where $n = n'k (n' > 1)$, the calculation of the critical values of the statistic (15) can be performed by noting that the statistic kX has approximately a χ^2 distribution with k degrees of freedom. It follows that if F_c and χ_c^2 are the critical values at the $\alpha_0 = 0{,}05$ level, then the critical value of the statistic X after k_1 selections becomes

$$X_c = (kF_c - k_1 kF_c/\chi_c^2)/(k - k_1) \tag{22}$$

The significance of all the components selected is then tested in the same way, using the statistic (20) where in this case $\hat{\sigma}^2$ denotes the estimate (17).

3.1.1.1 Example 26. *The mean daily value of rainfall at Uccle, for days with measurable precipitation*

The mean value of rainfall at Uccle has been determined by dividing the normal of the rainfall total for each month by the corresponding normal of the number of days with measurable precipitation (at least 0,1 mm), these normals being established on the basis of observations over a period of 137 years (1833-1969) for rainfall amounts and over 128 years (1833-1960) for the number of days.

The values of the mean level x_i thus obtained are shown in Table 26 together with the values of the circular functions $\sin \alpha i$ and $\cos \alpha i$ for $\alpha = 2\pi/12$. Since we have neglected the calculation of the error of estimation which affects these values and since the rainfall data from one month to another are practically independent, we can assume that the problem falls into the first category considered in 3.1.1 where the length of the series x_i is equal to the period, i.e. that $n = k$.

Application of the relations 3.1.1(5) gives successively :

$$\hat{a}_0 = (343 + 327 + \ldots + 369)/12 = 391{,}08$$

$$\hat{a}_1 = 2(343 \times 0{,}5 + 327 \times 0{,}86603 + \ldots + 369 \times 0)/12 = -55{,}41 \tag{1}$$

together with the other values of the elements of the vector $\hat{\theta}$ shown in Table 26.

For this calculation, values of the functions $\sin 2\alpha i$ and $\cos 2\alpha i$ are obtained by taking in the series of values of $\sin \alpha i$ and of $\cos \alpha i$ those values corresponding to the even values of i. Similarly, those corresponding to values of i which are multiples of j give the values of $\sin j\alpha i$ and $\cos j\alpha i$.

Next, by multiplying $\hat{a}_0$ and $\hat{a}_6$ by $\sqrt{12}$ and $\hat{a}_j$, $\hat{b}_j$, $j = 1, 2, 3, 4, 5$, by $\sqrt{6}$, we find by 3.1.1(8) the elements of the vector s, the absolute values and their natural logarithms of which are also shown in Table 26.

TABLE 26. Mean daily value of rainfall, in 0,01 mm, for days with measurable precipitation at Uccle. Selective harmonic analysis of the monthly values x_i

i	x_i	sin ωi	cos ωi	$\hat{\theta}$	$\lvert s \rvert$	$\log \lvert s \rvert$	$\hat{x}_i$	$\hat{\sigma}(\hat{x}_i)$
1	343	0,5	0,86603	$\hat{a}_0 = 391,08$	1354,74	7,21136	329,7	7,3
2	327	0,86603	0,5	$\hat{a}_1 = -55,41$	135,73	4,91067	323,7	
3	324	1,0	0	$\hat{b}_1 = -38,83$	95,11	4,55503	335,7	
4	351	0,86603	−0,5	$\hat{a}_2 = 12,12$	29,69	3,39081	362,5	
5	390	0,5	−0,86603	$\hat{b}_2 = 11,50$	28,17	3,33826	397,0	
6	437	0	−1,0	$\hat{a}_3 = -2,17$	5,32	1,67147	429,9	
7	472	−0,5	−0,86603	$\hat{b}_3 = 2,67$	6,54	1,87794	452,4	
8	468	−0,86603	−0,5	$\hat{a}_4 = 1,73$	4,24	1,44456	458,5	
9	427	−1,0	0	$\hat{b}_4 = -1,83$	4,48	1,49962	446,5	
10	408	−0,86603	0,5	$\hat{a}_5 = 1,75$	4,29	1,45629	419,7	
11	377	−0,5	0,86603	$\hat{b}_5 = 2,16$	5,29	1,66582	385,2	
12	369	0	1,0	$\hat{a}_6 = 2,25$	7,79	2,05284	352,2	

For calculation of the statistic 3.1.1(12), we then get, for $k = 12$,

$$\Sigma s_i^2 = \Sigma x_i^2 = (343)^2 + (327)^2 + \ldots + (369)^2 = 1864715 \tag{2}$$

from which, using the values of $\log \lvert s \rvert$ from Table 26, we obtain

$$M_{12} = 12 \log (1864715/12) - 2(7,21136 + 4,91067 + \ldots + 2,05284)$$

$$= 12 \times 11,953712 - 2 \times 35,07467 = 73,30 \tag{3}$$

For $c_1 = 12 - (1/12) \cong 12$, the tables of Pearson and Hartley (Table 32, line (a)) give for the 0,05 level the critical value 22,56. We can therefore conclude that there exists a nonzero deterministic component.

From the values of the elements of the vector s, the first element which we should extract from the group is a_0. For the new test, we then have

$$\Sigma s_i^2 = 1864715 - (1354,74)^2 = 29394,53 \tag{4}$$

from which the value of the statistic M, for $k = 11$, is given by

$$M_{11} = 11 \log (29394,53/11) - 2(35,07467 - 7,21136) = 31,07 \tag{5}$$

The critical value of the χ^2 variate with 11 degrees of freedom is 19,7 at the 0,05 level, so that the critical value of M_{11} is : 22,56 − (22,56/19,7) = 22,56 − 1,15 = 21,41, this value being less than that found in (5). Selection therefore has to be continued. As a_1 and b_1 give the component of the sum of squares $(\hat{a}_j^2 + \hat{b}_j^2)$ with the highest value, we have for the following test

$$\Sigma s_i^2 = 29394,53 - (135,73)^2 - (95,11)^2 = 1925,99 \tag{6}$$

which gives

$$M_9 = 9 \log (1925,99/9) - 2(35,07467 - 7,21136 - 4,91067 - 4,55503) = 11,5 \tag{7}$$

As the critical value is this time 22,56 − 3 x 1,15 = 19,11, we can assume that the remaining elements of the vector θ are all zero.

We also have the estimate

$$\hat{\sigma}^2 = 1925,99/9 = 214,0 \tag{8}$$

so that, for the pair formed by the last two components selected, the value of the statistic 3.1.1(20) gives

$$[(135,73)^2 + (95,11)^2]/(2 \times 214,0) = 64,2$$

with which we can associate the level α obtained from the relation

$$1 - (1 - \alpha)^{12/2} = 0,05$$

that is, $\alpha = 1 - 0,9915$.

Since the critical value of the F variate with $\nu_1 = 2$ and $\nu_2 = 9$ degrees of freedom is, for this level, 9,2, the significance of the selected components is established.

The estimator of the deterministic component of the series x_i is given by the relation

$$\hat{x}_i = 391{,}08 - 55{,}41 \sin \alpha i - 38{,}83 \cos \alpha i \tag{9}$$

for which we have, using 3.1.1(19)

$$\operatorname{var} \hat{x}_i = \frac{\sigma^2}{12}(1+2) = \sigma_2/4 \tag{10}$$

We conclude that the estimates $\hat{x}_i$ of the true mean values thus obtained are as accurate as normals which would be given by a series of observations four times longer than those which have been used.

By means of (8), the residual error associated with the estimates $\hat{x}_i$ given in Table 26 can be characterized by the estimate of the standard deviation

$$\hat{\sigma}(\hat{x}_i) = \sqrt{214{,}0/4} = 7{,}3 \tag{11}$$

For the computation of the modified statistic 3.1.1(12), we have : $|s'_1| = 1354{,}73$, $|s'_2| = 117{,}19$, $|s'_3| = 28{,}94$, $|s'_4| = 5{,}96$, $|s'_5| = 4{,}36$, $|s'_6| = 4{,}82$ and $|s'_7| = 7{,}79$, which gives :

$$M_7 = 12 \times 11{,}953712 - 2(7{,}21136 + 2 \times 4{,}76380 + \ldots + 2 \times 1{,}57277 + 2{,}05284) = 73{,}08$$

with a critical value of 14,83. Selecting s'_1 gives : $M_6 = 30{,}85$ with a critical value of 13,65. Selecting then s'_2, we have finally : $M_5 = 11{,}41$ with a critical value of 12,47. If follows that if the selection remains the same, this time M_5 is nearly significant. Moreover, if we select s'_3, the statistic 3.1.1(20) gives : $X = 27{,}12$ with $\nu_1 = 2$ and $\nu_2 = 7$ degrees of freedom, value which is significant at the level $\alpha = 0{,}001$ which makes, with 3.1.1(20) $\alpha_0 < 0{,}05$.

3.1.1.2 **Example 27.** *Estimation of the true mean values of the monthly averages of the air temperature at Uccle, from a series of monthly averages over two years (1971 and 1972)*

The 24 successive monthly averages of the air temperature at Uccle (in 0.1°C) for 1971 and 1972 are given in Table 27 ($i = 1, 2, \ldots, 24$). If we assume that the distributions of these monthly averages about the true mean values are, for each month, normal distributions whose unknown variances are all equal, then the problem of estimating the true mean values reduces to the second case considered in 3.1.1, i.e. that in which the length of the series x_i is a multiple of the period : $n = n'k$, with $n' = 2$ and $k = 12$.

The application of relations 3.1.1(5) then gives, using the values of $\sin \alpha i$ and $\cos \alpha i$ given in Table 26,

$$\hat{a}_0 = (36 + 40 + \ldots + 47)/24 = 98{,}00$$

$$\hat{a}_1 = 2(36 \times 0{,}5 + 40 \times 0{,}86603 + \ldots + 47 \times 0)/24 = -38{,}87 \tag{1}$$

together with the other values of the elements of the vector $\hat{\theta}$ given in Table 27.

As in the preceding example, by multiplying $\hat{a}_0$ and $\hat{a}_6$ by $\sqrt{24}$ and a_j, b_j, $j = 1, 2, 3, 4, 5$, by $\sqrt{12}$, we also find, using 3.1.1(8), the elements of the vector s, absolute values of which are shown in Table 27.

For the calculation of the statistic 3.1.1(15), we have

$$\Sigma x_i^2 = (36)^2 + (40)^2 + \ldots + (47)^2 = 294072$$

and

$$\Sigma s_i^2 = (480{,}10)^2 + (134{,}65)^2 + \ldots + (1{,}22)^2 = 291880{,}2648 \tag{2}$$

From this we obtain, using 3.1.1(16),

$$\Sigma \hat{e}_i^2 = 294072 - 291880{,}2648 = 2191{,}7352 \tag{3}$$

The statistic 3.1.1(15) thus becomes

$$X_{12} = \frac{291880{,}2648/12}{2191{,}7352/12} = 133{,}2 \tag{4}$$

TABLE 27. Monthly averages x_i of the air temperature at Uccle in 1971 and 1972 (in 0,1°C). Selective harmonic analysis of the monthly averages x_i

i	x_i	i	x_i	$\hat{\theta}$	$\lvert s \rvert$	$\hat{x}_i$	1901-30
1	36	13	20	$a_0 =$ 98,00	480,10	27,7	27
2	40	14	46	$a_1 = -38,87$	134,65	35,0	31
3	33	15	78	$\hat{b}_1 = -58,76$	203,55	59,1	55
4	91	16	82	$a_2 =$ 6,50	22,52	93,7	82
5	151	17	119	$\hat{b}_2 =$ 0,75	2,60	129,5	128
6	144	18	141	$a_3 =$ 0,75	2,60	158,8	149
7	189	19	176	$\hat{b}_3 =$ 2,75	9,53	168,3	168
8	173	20	161	$a_4 =$ 0,00	0,00	161,0	164
9	143	21	122	$\hat{b}_4 = -3,00$	10,39	136,9	140
10	107	22	95	$a_5 =$ 1,12	3,88	102,3	100
11	50	23	56	$\hat{b}_5 =$ 9,51	32,94	66,5	52
12	52	24	47	$a_6 =$ 0,25	1,22	39,2	33

Since the critical value at the 0,05 level of the F variate with $\nu_1 = 12$ and $\nu_2 = 12$ degrees of freedom is 2,69, we can reject the hypothesis $\theta = 0$ or $E(s) = 0$.

From the values of the elements of the vector s, the first element which we should extract from the group is a_0. For the new test we have

$$\Sigma s_l^2 = 291880{,}2648 - (480{,}10)^2 = 61384{,}2548 \tag{5}$$

from which we obtain the statistic

$$X_{11} = \frac{61384{,}2548/11}{2191{,}7352/12} = 30{,}6 \tag{6}$$

The critical value at the 0,05 level of the χ^2 variate with 12 degrees of freedom is 21,0, so for X_{11}, using 3,1,1(22),

$$[2{,}69 \times 12 - (2{,}69 \times 12/21{,}0)]/11 = (32{,}28 - 1{,}54)/11 = 2{,}79$$

Consequently, the hypothesis $E(s) = 0$ for the remaining elements of the vector s can again be rejected. Since the elements a_1 and b_1 give the highest value of the sum $(\hat{a}_j^2 + \hat{b}_j^2)$, for the following test, we obtain

$$\Sigma s_l^2 = 61384{,}2548 - (134{,}65)^2 - (203{,}55)^2 = 1821{,}0298 \tag{7}$$

which gives the statistic

$$X_9 = \frac{1821{,}0298/9}{2191{,}7352/12} = 1{,}11 \tag{8}$$

As the critical value of X_9 is this time $(32{,}28 - 3 \times 1{,}54)/9 = 3{,}07$, we can assume that the remaining elements of the vector θ are all zero. It follows that we also have

$$\hat{\sigma}^2 = (2191{,}74 + 1821{,}03)/(12 + 9) = 191{,}08 \tag{9}$$

and that for the last component selected, the statistic 3.1.1(20) becomes

$$[(134{,}65)^2 + (203{,}55)^2]/(2 \times 191{,}08) = 155{,}9$$

with which we can associate, as in 3.1.1.1, the level $\alpha = 1 - 0{,}9915$. As the critical value of the F variate with $\nu_1 = 2$ and $\nu_2 = 21$ degrees of freedom for this level is 6,1, the significance of the selected components is established.

The estimator of the deterministic component of the series x_i is given by the relation

$$\hat{x}_i = 98{,}0 - 38{,}87 \sin \alpha i - 58{,}76 \cos \alpha i \tag{10}$$

for which we have, using 3,1,1(19),

$$\operatorname{var} \hat{x}_i = \frac{\sigma^2}{24}(1+2) = \sigma^2/8 \tag{11}$$

We conclude that the estimates $\hat{x}_i$ of the true mean values thus obtained are as accurate as the normals which would be given by a series of observations over eight years.

To allow comparison, we have given in Table 27, beside our estimates, normals calculated over the period 1901-1930. It is apparent that the agreement is satisfactory in the majority of cases, except in April, June and November, for which the disagreement reaches or exceeds $10(1{,}0^{\circ})$, which may appear quite large in comparison with the estimate of the standard deviation : $\sqrt{191{,}08/8} \cong 5$.

The explanation for this slight disagreement lies in the fact both that the value given for $\hat{\sigma}^2$ is somewhat underestimated and that the hypothesis of a constant variance for the 12 months of the year is not satisfied.

3.1.2 Estimation of the deterministic component of a periodic series when the period of the deterministic component and the covariance matrix of the random part are known.

As indicated in 3.1, we assume that the length n of the series is equal to the period k and that the matrix v is known. If we then note that relation 3(9) can be written

$$u' v^{-1} u\hat{\theta} = u' v^{-1} x \tag{1}$$

then putting $\beta = u' v^{-1} u\hat{\theta} = u' v^{-1} x$ and $\hat{e} = x - u\hat{\theta}$, relation 3(13) becomes

$$x' v^{-1} x = \hat{\theta}'\beta + \hat{e}' v^{-1} \hat{e} \tag{2}$$

which can be used in the selection of the non-zero elements of the vector θ.

This selection can be performed in two different ways depending on whether we consider progressively more complex models until we obtain a statistic

$$X = \hat{e}' v^{-1} \hat{e} = x' v^{-1} x - \hat{\theta}'\beta \tag{3}$$

which is not significant, or whether we first estimate the k elements of the vector θ using equation (1) obtained from a complete matrix u, the selection than being performed on the basis of relation (2) which becomes

$$x' v^{-1} x = \hat{\theta}'\beta \tag{4}$$

In the first method, the simplest model which we shall start by examining corresponds to the null hypothesis $\theta = 0$. Under this hypothesis the statistic

$$X_k = x' v^{-1} x = \hat{e}' v^{-1} \hat{e} \tag{5}$$

has a χ^2 distribution with k degrees of freedom when the joint distribution of the elements of the vector e is normal.

If this assumption is rejected, we next examine the series of k models in which the vector θ has a single non-zero element and from these choose the model which leads to the smallest value of the statistic

$$X_{k-1} = \hat{e}' v^{-1} \hat{e} = X_k - \hat{\theta}'\beta \tag{6}$$

To this end, we calculate the solutions $\hat{\theta}_l$, $l = 1, 2, \ldots, k$ of each of the k equations

$$\sum_{ij} u_{ii} v_{ij}^{-1} u_{jl} \hat{\theta}_l = \beta_l, \quad l = 1, 2, \ldots, k \tag{7}$$

where

$$\beta_l = \sum_{ij} u_{ii} v_{ij}^{-1} x_j$$

is the l^{th} element of the vector β and where v_{ij}^{-1} is the element (ij) of the inverse matrix of the matrix v. We next put $\hat{\theta}'\beta = \hat{\theta}_l \beta_l$ in relation (6).

If $\hat{\theta}_{l_0}$ is the solution which gives, the smallest value of the statistic (6), there is a difference depending on whether θ_{l_0} is one of the constants a_0 and a_k , or whether θ_{l_0} is one of the constants a_j and b_j, $j = 1,2,\ldots,(k'-1)$.

In the first eventuality, the model employed is, depending on the case,

$$x_i = a_0 + e_i \quad \text{or} \quad x_i = a_k (-1)^i + e_i \tag{8}$$

whereas in the second, we use from the $(k' - 1)$ models

$$x_j = a_j \sin j\alpha i + b_j \cos j\alpha i + e_i, \quad j = 1, 2, \ldots, (k'-1) \tag{9}$$

the one which leads to the lowest value of the statistic

$$X_{k-2} = \hat{e}' v^{-1} \hat{e} = X_k - \hat{\theta}'\beta \tag{10}$$

and hence requires the estimation of the coefficients a_j and b_j by means of the systems of two equations with two unknowns which are obtained from relation (1).

If moreover, the value of the statistic X_{k-1} or X_{k-2} is less than the critical value of the corresponding χ^2 variate, at the level calculated using equation 1.3.3(3), this model is chosen.

Otherwise, we make an analogous examination of the models obtained by adding to the model any of the other terms of the general solution 3.1(2), and the process is continued until a non-significant value of the statistic (3) is found. We must remember that if one of the pair a_j or b_j is chosen, the other should be chosen too.

The significance of the components selected is then established by ensuring that the decrease X_k brought about by the last selection is greater than the critical value of the χ^2 variate with one or two degrees of freedom at the level α obtained from the relation

$$1 - (1 - \alpha)^k = \alpha_0 \quad \text{or} \quad 1 - (1 - \alpha)^{k/2} = \alpha_0$$

respectively, α_0 being the initial level adopted.

The final estimate of the deterministic component of the vector x is then given by the relation

$$\hat{x} = u\hat{\theta} \tag{11}$$

in which the matrix u has been reduced to only those columns corresponding to the non-zero elements of the vector θ and we have for the vectors $\hat{\theta}$ and $\hat{x}$ the covariance matrices

$$\text{var}\,\hat{\theta} = (u'v^{-1}u)^{-1} \quad \text{and} \quad \text{var}\,\hat{x} = u(\text{var}\,\hat{\theta})u' \tag{12}$$

In the second method, we consider relation (1) in the case of a vector θ reduced to only some of its elements. Then the matrix $u'v^{-1}u$ of the coefficients of these elements in the system of equations (1) is found from the complete matrix $w = u'v^{-1}u$ by suppressing in matrix w the rows and columns corresponding to the elements which are absent in the reduced vector θ.

Since in all cases we have

$$(\text{var}\,\hat{\theta})^{-1} = u'v^{-1}u \tag{13}$$

the hypothesis $\theta = 0$ is satisfied for any reduced vector θ for which an estimate $\hat{\theta}$ leads to a value of the corresponding statistic

$$\hat{\theta}'u'v^{-1}u\hat{\theta} = \hat{\theta}'w\hat{\theta} \tag{14}$$

which is not significant.

Since relation (1) can also be written

$$u'v^{-1}(u\hat{\theta} - x) = 0 \tag{15}$$

and since, in the case of a complete vector θ, this relation implies the relation

$$u\hat{\theta} = x \tag{16}$$

then we deduce that the estimate $\hat{\theta}$ of the complete vector θ is the same as that given by relations 3.1.1(5).

Under these conditions, if the solution of equation (15) leads to a significant value of the statistic

$$\chi_k = \hat{\theta}'w\hat{\theta} = \hat{\theta}'\beta \tag{17}$$

then the method consists of successively setting to zero in this statistic each of the elements of the vector $\hat{\theta}$, in order of decreasing size, until a non-significant value is obtained for the statistic (14), the critical values being calculated as in 1.3.3. In particular, for the level $\alpha_0 = 0{,}05$, these values can be obtained by subtracting from the initial critical value of the χ^2 variate with k degrees of freedom the number of elements of the vector θ which have been selected. The significance of the selected components can be established on the basis of the decrease brought about by the last selection in the value of the statistic χ_k.

The solution adopted is that represented by the elements of the vector $\hat{\theta}$ which have had to be set to zero, and the final estimate of the reduced vector θ thus defined is calculated on the basis of the system of equation represented by (1).

For the final estimate of the deterministic components of the vector x as well as for the calculation of the covariance matrices of the vectors $\hat{\theta}$ and $\hat{x}$, we use, as in the first method, relations (11) and (12).

In choosing the elements of the complete vector $\hat{\theta}$ which should be successively set to zero, by putting $\hat{\theta}_i = 0$ in $\hat{\theta}'w\hat{\theta}$, we reduce the value of the statistic by the quantity

$$\Delta_{1,i} = 2\hat{\theta}_i\beta_i - \hat{\theta}_i^2 w_{ii} \tag{18}$$

If we simultaneously put $\hat{\theta}_i = 0$ and $\hat{\theta}_l = 0$, this quantity becomes

$$\Delta_{2,il} = 2(\hat{\theta}_i \beta_i + \hat{\theta}_l \beta_l) - (\hat{\theta}_i^2 w_{ii} + \hat{\theta}_l^2 w_{ll} + 2\hat{\theta}_i \hat{\theta}_l w_{il}) \tag{19}$$

and, generally speaking, if $\hat{\theta}_0$ is the vector formed by the p elements of the vector $\hat{\theta}$ which have been set to zero in $\hat{\theta}' w \hat{\theta}$, the decrease becomes

$$\Delta_p = 2\hat{\theta}'_0 \beta - \hat{\theta}'_0 w \hat{\theta}_0 \tag{20}$$

These relations can be used either to determine the size of the contribution in the statistic (17) of the value of each element of the vector $\hat{\theta}$ (depending on whether one of the elements $\hat{a}_0$ or $\hat{a}_k$, or one of the pairs $\hat{a}_j$ and $\hat{b}_j$ is used), or to calculate the final value of the statistic (14).

3.1.2.1 **Example 28.** *Daily averages of the air temperature at Uccle. Estimation of the true mean values on the basis of 12 equally spaced daily normals (period 1901-1970)*

As the length of year is 365,25 days and that of the month of February is 28,25 days, the normal of the daily mean of the air temperature was calculated for each of the 12 days of the year for which the ordinal number best approximates the numbers $-15{,}22 + 30{,}42i$, $i = 1, 2, \ldots, 12$, that is respectively : 15 January, 15 February, 17 March, 16 April, 17 May, 16 June, 17 July, 16 August, 15 September, 16 October, 15 November and 16 December. The normals thus obtained are the values of x_i given in Table 28 expressed in 0,1 °C. The variance v_i of the sampling error associated with these normals are also shown alongside the values of x_i. These were obtained by dividing by 70 the mean value of the empirical variances of the temperatures of the ten days which bracket each of the dates chosen.

So we are dealing with a periodic series for which the length is equal to the period : $n = k = 12$. Since the dates considered are separated by an interval of some 30 days, we can assume that the temperatures on these various dates are practically independent of each other. The covariance matrix of the error associated with the 12 normals is therefore a diagonal matrix whose elements are the variances v_i. The inverse matrix v^{-1} is then also diagonal, and its elements are given by $1/v_i$.

To test the validity of the hypothesis $\theta = 0$ we first calculate the value of the statistic 3.1.2(5), which gives

$$X_{12} = (23)^2/26{,}4 + (24)^2/26{,}4 + \ldots (29)^2/25{,}2 = 10551{,}5 \tag{1}$$

Since this value exceeds the critical value at the 0,05 level of the χ^2 variate with 12 degrees of freedom, 21,0, we can conclude that at least one element of the vector θ is not zero.

(a) *Selection of the non-zero elements of the vector θ by the first method*

To determine the element of the vector θ which should be taken as the first non-zero element, we solve equation 3.1.2(1) in each of the cases in which the matrix

$$u = \| 1 \quad \sin \alpha i \quad \cos \alpha i \quad \sin 2\alpha i \quad \cos 2\alpha i \cdots \sin 5\alpha i \quad \cos 5\alpha i \quad (-1)^i \|$$
$$i = 1, 2, \cdots, 12 \tag{2}$$

reduces respectively to the first, second ..., 12th column, where $\alpha = 2\pi/12$.

With $u = \| 1 \|$, we obtain successively

$$u' v^{-1} u = w_{11} = \Sigma(1/v_i) = 1/26{,}4 + 1/26{,}4 + \cdots + 1/25{,}2 = 0{,}734$$

$$u' v^{-1} x = \beta_1 = \Sigma(x_i/v_i) = 23/26{,}4 + 24/26{,}4 + \cdots + 29/25{,}2 = 79{,}21$$

From this we obtain

$$\hat{\theta}_1 = \beta_1 / w_{11} = 79{,}21/0{,}734 = 107{,}92 \tag{3}$$

With $u = \| \sin \alpha i \|$, we find in the same way, using the values of $\sin \alpha i$ given in Table 26 :

$$u'v^{-1}u = w_{22} = \Sigma(\sin^2 \alpha i)/v_i = (0,5)^2/26,4 + (0,86603)^2/26,4 + \cdots + (0)^2/25,2 = 0,395$$

$$u'v^{-1}x = \beta_2 = \Sigma(\sin \alpha i)x_i/v_i = 0,5 \times 23/26,4 + 0,86603 \times 24/26,4 + \cdots + 0 \times 29/25,2 = -26,98$$

From this we deduce

$$\hat{\theta}_2 = \beta_2/w_{22} = -26,98/0,395 = -68,30 \tag{4}$$

Similarly we find the other values of w_{ii}, β_i and $\hat{\theta}_i$ given in Table 28.
It follows that for $X_{11,1}$ we have the value

$$X_{11,1} = X_{12} - \hat{\theta}_1 \beta_1 = 10551,5 - 107,92 \times 79,21 = 2003 \tag{5}$$

as well as the eleven other values of $X_{11,i}$ for $i = 2,3,\ldots, 12$ also given in Table 28.
It turns out in this way that the smallest value of the statistic X_{11} is that of θ_1. The model to be used is thus

$$x_i = \theta_1 + e_i = a_0 + e_i \tag{6}$$

Since furthermore the critical value with which $X_{11,1}$ should be compared is X_{12} diminished by 1, that is, $21,0 - 1 = 20,0$, i.e. a value less that $X_{11,1}$, it is necessary to consider a model in which, in addition to a_0, at least one other element of the vector θ is not zero. As the value of $X_{11,12}$ differs very little from X_{12}, i.e. the estimate of a_6 contributes little to the reduction of the value of X_{12}, we can eliminate this element from the comparison and examine the models with three non-zero constants

$$x_i = a_0 + a_j \sin j\alpha i + b_j \cos j\alpha i + e_i \; ; \; j = 1, 2, \ldots, 5 \tag{7}$$

for which the matrix u reduces to the form

$$u = \| 1 \; \sin j\alpha i \; \cos j\alpha i \| \; ; \; j = 1, 2, \ldots, 5 \; ; \; i = 1, 2, \ldots, 12 \tag{8}$$

In addition to the values of w_{ii} already calculated, we also have, in view of the diagonal form of the matrix v,

$$\begin{aligned} w_{1\,2j} &= \Sigma u_{i1} v_i^{-1} u_{i\,2j} = \Sigma(\sin j\alpha i)/v_i \\ w_{1\,2j+1} &= \Sigma u_{i1} v_i^{-1} u_{i\,2j+1} = \Sigma(\cos j\alpha i)/v_i \\ w_{2j\,2j+1} &= \Sigma u_{i\,2j} v_i^{-1} u_{i\,2j+1} = \Sigma(\sin j\alpha i)(\cos j\alpha i)/v_i \end{aligned} \tag{9}$$

TABLE 28. – Daily averages of the air temperature at Uccle (in 0,1°C). Selective harmonic analysis of 12 equally spaced daily normals (period 1901-1970).

i	j	x_i	v_i	w_{ii}	β_i	$\hat{\theta}_i$	$X_{11,i}$	$X_{9,j}$	$X_{7,j}$	$\hat{x}_i$	var $\hat{x}_i$	$\hat{\theta}$	Δ_1	Δ_2	$X_{12} - \Delta_p$
1		23	26,4	0,734	79,21	107,92	2003			18,1	10,1	95,17	8428,8		2122,7
2	1	24	26,4	0,395	−26,98	−68,3	8709	23		32,4	9,4	−40,90	1546,2	3830,7	
3		64	17,5	0,339	−28,63	−84,5	8132			60,0	8,0	−64,60	2284,3		24,0
4	2	92	17,5	0,367	6,76	18,4	10427	1873	11	93,1	7,4	2,454	31,0	96,7	
5		125	18,5	0,367	−7,60	−20,7	10394			126,2	7,2	−4,917	65,9		11,3
6	3	155	15,6	0,366	0,351	0,96	10551	1982	22	154,4	6,6	1,000	0,336	5,4	
7		173	13,7	0,368	3,89	10,6	10510			170,7	5,7	0,667	5,03		
8	4	164	10,4	0,367	−0,124	−0,34	10551	2002	15	167,4	5,1	4,186	−7,47	−8,0	
9		143	12,0	0,367	0,388	1,06	10551			141,2	5,2	2,583	−0,444		
10	5	102	13,7	0,395	2,55	6,46	10535	2001	21	98,7	6,0	2,401	9,97	6,7	
11		48	16,5	0,339	−1,61	−4,74	10544			54,7	7,2	0,931	−3,29		
12		29	25,2	0,734	1,73	2,36	10547			24,9	9,0	−0,833	−3,93		

the system of equations 3.1.2(1) which gives the estimates $\hat{a}_0$, $\hat{a}_j$ and $\hat{b}_j$ and can be written :

$$\begin{aligned} w_{11}\hat{a}_0 + w_{1\,2j}\hat{a}_j + w_{1\,2j+1}\hat{b}_j &= \beta_1 \\ w_{1\,2j}\hat{a}_0 + w_{2j\,2j}\hat{a}_j + w_{2j\,2j+1}\hat{b}_j &= \beta_{2j} \\ w_{1\,2j+1}\hat{a}_0 + w_{2j\,2j+1}\hat{a}_j + w_{2j+1\,2j+1}\hat{b}_j &= \beta_{2j+1} \end{aligned} \tag{10}$$

where β_1 , β_{2j} and β_{2j+1} are the elements (1), (2j) and (2j +1) of the vector β .

In particular, for j =1, in addition to the values of w_{11}, w_{22} and w_{33} given in Table 28, we have, using the values of sin αi and cos αi given in Table 26 and those of v_i given in Table 28 :

$$\begin{aligned} w_{12} &= (0,5)/26,4 + (0,86603)/26,4 + \cdots + (0)/25,2 = -0,1112 \\ w_{13} &= (0,86603)/26,4 + (0,5)/26,4 + \cdots + (1)/25,2 = -0,0704 \\ w_{23} &= (0,5 \times 0,86603)/26,4 + (0,86603 \times 0,5)/26,4 + \cdots + (0 \times 1)/25,2 = 0,000047 \end{aligned} \tag{11}$$

Using the values of β_1 , β_2 and β_3 given in Table 28, the system of equations (10) leads to the solution

$$\hat{a}_0 = 95,44 \ , \ \hat{a}_1 = -41,42 \ , \ \hat{b}_1 = -64,66 \tag{12}$$

and to the value of the statistic

$$\begin{aligned} X_{9,1} = X_{12} - \hat{\theta}'\beta &= 10551,5 - 95,44 \times 79,21 - (-41,42) \times (-26,98) \\ &\quad -(-64,66) \times (-28,63) = 23,0 \end{aligned} \tag{13}$$

Proceeding in similar fashion for j =2, 3, 4, and 5, we find the values of the statistics $X_{9,j}$ from Table 28. We see that the model to be used is that with j =1. Since the critical value of X_9 is here 21,0 − 3 =18,0 , i.e. a value less that $X_{9,1}$, it is necessary to continue the selection. Discarding, as above, the element c_6 from the selection, we are thus led to consider the models

$$x_i = a_0 + a_1 \sin \alpha i + b_1 \cos \alpha i + a_j \sin j\alpha i + b_j \cos j\alpha i + e_i \quad j = 2, 3, 4, 5 \tag{14}$$

The solution of the four systems of five equations with five unknowns gives in similar fashion the values of the statistics $X_{7,j}$ of Table 28, from which we find the model to be used is j =2. As the value of $X_{7,2}$ is now less than the critical value 21,0 − 5 =16,0 of X_7 , the model can be adopted. Moreover, the decrease in the statistic X_k corresponding to the last selection, i.e. 23 − 11 =12, is greater than 9,62 , the critical value of the χ^2 variate with two degrees of freedom at the level α, which here is still (1 − 0,9915); the significance of the selected components is therefore established.

With, for this model,

$$\hat{a}_0 = 95,14 \ , \ \hat{a}_1 = -40,58 \ , \ \hat{b}_1 = -64,71 \ , \ \hat{a}_2 = 2,30 \ , \ \hat{b}_2 = -5,48 \tag{15}$$

we obtain the values $\hat{x}_i$ of the deterministic component given in Table 28.

In the calculation of the error associated with these estimates, we note that, using 3.1.2(12), the covariance matrix of the vector $\hat{\theta}$ is the inverse matrix of the matrix of the coefficients of the unknowns in the system of equations 3.1.2(1), i.e. in this case, the inverse of the matrix

$$u'v^{-1}u = \| w_{1l}\, w_{2l}\, w_{3l}\, w_{4l}\, w_{5l} \| \quad l = 1, 2, \ldots, 5 \tag{16}$$

For the vector $\hat{x}$, the covariance matrix is given by the matrix product u(var $\hat{\theta}$)u'.

The inversion of the matrix (16) was carried out by means of a computer subroutine and the quantities var $\hat{\theta}_l$ and cov $(\hat{\theta}_l , \hat{\theta}_{l'})$ allow the elements var $\hat{x}_t$ and cov $(\hat{x}_t , \hat{x}_{t'})$ to be calculated. It was in particular the relations

$$\text{var } \hat{x}_i = \sum_l \sum_{l'} u_{il} \text{ cov } (\hat{\theta}_l , \hat{\theta}_{l'}) u_{il'}$$

and

$$|| u_{il} || = || 1 \sin \alpha i \cos \alpha i \sin 2\alpha i \cos 2\alpha i ||$$

which supplied the values given in Table 28.

The mean value of the ratio $v_i/(\text{var } \hat{x}_i)$ deduced from this, 2,45 , shows that estimation of the deterministic component via selective harmonic analysis has led to estimates of the true mean values which are as accurate as if these had been obtained by means of normals calculated for a period 2,45 times longer, i.e. a period of 70 x 2,45 =172 years.

(b) *Selection of the non-zero elements of the vector θ by the second method*

The application of relations 3.1.1(5) gives the values of the elements of the vector $\hat{\theta}$ shown in column $\hat{\theta}$ of Table 28 and application of relation 3.1.2(18) then leads, using the values of β_i and w_{ii} appearing in Table 28, for $i=1$ to :

$$\Delta_1 = 2 \text{ x } 95,17 \text{ x } 79,21 - (95,17)^2 \text{ x } 0,734 = 8428,8 \tag{17}$$

together with the values of Δ_1 given in Table 28 for the other values of i.

Similarly, application of relation 3.1.2(19) gives for each of the pairs $\hat{a}_j$ and $\hat{b}_j$, $j=1, 2, ..., 5$, the values of Δ_2 given in Table 28.

We see that Δ_1 has its highest value for $i=1$, whereas that of Δ_1 for $i=12$ is exceeded by at least one of the two values of Δ_1 corresponding to the same harmonic. We conclude that $\hat{\theta}_1$ and $\hat{\theta}_{12}$ must be placed respectively in first and last position in the order of selection. Furthermore, the values of Δ_2 show the order to be adopted for the elements relating to each of the five harmonics as $j = 1, 2, 5, 3, 4$.

In this way, successive values of $X_{12} - \Delta_p$, for $p=1, 3, 5$ lead, for $p=5$, to a value of the statistic which is less than the critical value $21,0 - 5 = 16,0$. Moreover, since the decrease from $p=3$ to $p=5$, i.e. $24,0 - 11,3 = 12,7$, is greater than 9,62 , the critical value of the χ^2 variate with two degrees of freedom at the level $\alpha = 1 - 0,9915$ given by the equation $1 - (1-\alpha)^{12/2} = 0,05$, all the components selected can be considered to be significant.

Since this corresponds to the selection $(a_0 , a_1 , b_1 , a_2 , b_2)$, we obtain once again the model found by the first method. The final solution is therefore that represented by the estimates given in (15).

We should point out that when the matrix v is diagonal as is the case in the preceding example, a simple approximate solution can be obtained by replacing the elements v_i by their mean value.

If this mean value is v_0 , the selection is performed by successively subtracting from the quantity $x'x/v_0$ the quantities s_i^2/v_0, where the quantities s_i are the elements of the vector s defined in 3.1.1. As above, the subtraction is performed in the order of decreasing size and by coupling each a_j to each b_j.

In the last example, we have

$$v_0 = (26,4 + 26,4 + ... + 25,2)/2 = 17,78 \tag{18}$$

and

$$X_{12} = [(23)^2 + (24)^2 + ... + (29)^2]/17,78 = 8106,7 \tag{19}$$

with

$$s_1 = 95,17 \text{ x } \sqrt{12} = 329,68 \quad , \quad s_2 = -40,90 \text{ x } \sqrt{6} = -100,18,$$

$$s_3 = -158,24 \quad , \quad s_4 = 6,01 \quad \text{and} \quad s_5 = 12,04$$

we obtain successively :

$$X_{11} = 8106,7 - s_1^2/v_0 = 1993,7$$

$$X_9 = 1993,7 - (s_2^2 + s_3^2)/v_0 = 20,9$$

$$X_7 = 20,9 - (s_4^2 + s_5^2)/v_0 = 10,8$$

which gives the same selection.

It is verified moreover that, if we adopt the solution formed by the first five elements of the vector $\hat{\theta}$ of the complete solution, we obtain a solution which differs very little from the exact solution. Finally, the reduction of the variance of the estimated values of x, associated with the simplified solution, which is 5/12 of the initial variance, leads to an estimation accuracy equivalent to that given by mean values found from series $2/4$ times greater in length. The agreement between the two solutions is thus excellent.

3.1.3 Reference works consulted

The method of selection adopted for the harmonic analysis of a series of observations when the period is known is a direct application of the method of least squares. The description is reproduced from that given in [44].

When we have k independent estimates and when we wish to proceed to a significance test for each of these estimates, we actually have a multiple test which should be treated as indicated in 1.3. This is why the considerations described in 1.3.3 can be applied here.

We should point out here that the selection methods usually employed are applications of the test based on the binomial distribution as are presented by Fisher in [45] or more recently by Tiago de Oliveira in [46]. We have opted to use as first choice a method of selection based on the statistic of a multiple test of the type used by Fisher, however, because this is more generally applicable since in such a test the covariance matrix of the joint distribution of the estimates on which the multiple test is performed may not be diagonal.

The data used in example 26 were drawn up on the basis of the rainfall series at Brussels-Uccle corrected in accordance with [43] and the mean number of days with measurable precipitation is that found in [15]. For example 27, we have referred to the monthly averages of the air temperature at Uccle published in [47], and for example 28 we have used the data published in [48].

3.1.4 Estimation of the deterministic component of a periodic series when the period of the deterministic component is known but has an arbitrary value, the covariance matrix of the random part being unknown

Here we assume that the time series is of the form

$$x_i = a_0 + a_1 \sin \alpha i + b_1 \cos \alpha i + e_i \quad , \quad i = 1, 2, \ldots, n \tag{1}$$

where $\alpha = 2\pi/T$ is known, without T necessarily being an integer, where a_0 , a_1 and b_1 are unknown constants and where e_i are random quantities with zero mean value but with unknown variance σ^2, distributed independently with the same distribution.

If we denote by x , e and u respectively the vectors and the matrix,

$$x = \| x_i \| \ , \ e = \| e_i \| \ , \ u = \| 1 \ \sin \alpha i \ \cos \alpha i \ \|$$

where $i = 1, 2, \ldots, n$, and by θ the vector of unknown elements a_0 , a_1 and b_1 , equation (1) is written

$$x = u\theta + e \tag{2}$$

The solution $\hat{\theta}$ by the method of least squares is given, as in 3.1.1, by the relation

$$u'u\hat{\theta} = u'x \tag{3}$$

More precisely, the estimates $\hat{a}_0$, $\hat{a}_1$ and $\hat{b}_1$ form the solution of the system of three linear equations with three unknowns for which the matrix of the coefficients of the unknowns is the matrix

$$u'u = \begin{Vmatrix} n & \Sigma \sin \alpha i & \Sigma \cos \alpha i \\ \Sigma \sin \alpha i & \Sigma \sin^2 \alpha i & \Sigma \sin \alpha i \cos \alpha i \\ \Sigma \cos \alpha i & \Sigma \sin \alpha i \cos \alpha i & \Sigma \cos^2 \alpha i \end{Vmatrix} \tag{4}$$

and for which the vector $\beta = u'x$ of the independent terms is formed by the elements

$$\beta_1 = \Sigma x_i \ , \ \beta_1 = \Sigma \ x_i \sin \alpha i \quad , \quad \beta_3 = \Sigma x_i \cos \alpha i \tag{5}$$

where in (4) and (5) we have written Σ to denote $\sum_{i=1}^{n}$.

If σ^2 is the common variance of the random quantities e_i , we also have

$$\operatorname{var} \hat{\theta} = (u'u)^{-1} \sigma^2 \tag{6}$$

whereas the relation

$$\hat{e} = x - u\hat{\theta} \tag{7}$$

leads, as in 3.1.2, to

$$\hat{e}'\hat{e} = x'x - \hat{\theta}'\beta \tag{8}$$

which gives the unbiased estimate

$$\hat{\sigma}^2 = \hat{e}'\hat{e}/(n-3) = (x'x - \hat{\theta}'\beta)/(n-3) \tag{9}$$

To establish the significance of the periodic part of the deterministic component, if the matrix

$$w = \begin{Vmatrix} w_{11} & w_{12} \\ w_{12} & w_{22} \end{Vmatrix} \tag{10}$$

is the inverse matrix of the covariance matrix of the estimates $\hat{a}_1$ and $\hat{b}_1$, under the null hypothesis $a_1 = b_1 = 0$, the statistic

$$X = \hat{a}_1^2 w_{11} + 2\hat{a}_1 \hat{b}_1 w_{12} + \hat{b}_1^2 w_{22} \tag{11}$$

has a χ^2 distribution with two degrees of freedom which may be exact or approximate.

Moreover, we can put

$$a_1 = c_1 \cos \varphi \quad \text{and} \quad b_1 = c_1 \sin \varphi \tag{12}$$

with $c_1 > 0$, which implies

$$c_1 = \sqrt{a_1^2 + b_1^2} \quad \text{and} \quad \operatorname{tg} \varphi = b_1/a_1 \tag{13}$$

and allows us to express relation (1) in the form

$$x_1 = a_0 + c_1 \sin (\alpha i + \varphi) + e_i \tag{14}$$

If we estimate c_1 and φ by (13), by differentiation we obtain from (12) :

$$\begin{aligned} da_1 &= \cos\varphi dc_1 - c_1 \sin\varphi d\varphi \\ db_1 &= \sin\varphi dc_1 + c_1 \cos\varphi d\varphi \end{aligned} \tag{15}$$

from which we deduce :

$$\begin{aligned} dc_1 &= \cos\varphi da_1 + \sin\varphi db_1 \\ d\varphi &= (\cos\varphi db_1 - \sin\varphi da_1)/c_1 \end{aligned} \tag{16}$$

If the errors of estimation are sufficiently small, we have ;

$$\operatorname{var}\hat{c}_1 = \cos^2\varphi \operatorname{var}\hat{a}_1 + \sin^2\varphi \operatorname{var}\hat{b}_1 + \sin 2\varphi \operatorname{cov}(\hat{a}_1, \hat{b}_1)$$

$$\operatorname{cov}(\hat{c}_1, \hat{\varphi}) = [\sin\varphi\cos\varphi(\operatorname{var}\hat{b}_1 - \operatorname{var}\hat{a}_1) + \cos 2\varphi \operatorname{cov}(\hat{a}_1, \hat{b}_1)]/c_1$$

and

$$\operatorname{var}\hat{\varphi} = [\cos^2\varphi \operatorname{var}\hat{b}_1 + \sin^2\varphi \operatorname{var}\hat{a}_1 - \sin 2\varphi \operatorname{cov}(\hat{a}_1, \hat{b}_1)]/c_1^2 \tag{17}$$

When n is sufficiently large, the above relations can be simplified because the terms $\Sigma\sin\alpha i$, $\Sigma\cos\alpha i$ and $\Sigma\sin\alpha i\cos\alpha i$ of the matrix (4) become negligible relative to n and since, to a good approximation, we have

$$\Sigma\sin^2\alpha i = \Sigma\cos^2\alpha i = n/2 \tag{18}$$

We obtain, in this case

$$\operatorname{var}\hat{a}_1 = \operatorname{var}\hat{b}_1 = 2\sigma^2/n \quad \text{and} \quad \operatorname{cov}(\hat{a}_1, \hat{b}_1) = 0, \tag{19}$$

statistic (11) simplifies to

$$X = n\hat{c}_1^2/(2\sigma^2) \tag{20}$$

and relations (17) become respectively :

$$\operatorname{var}\hat{c}_1 = 2\sigma^2/n \ , \quad \operatorname{var}\hat{\varphi} = 2\sigma^2/(nc_1^2)$$

$$\operatorname{cov}(\hat{c}_1, \hat{\varphi}) = 0 \tag{21}$$

For completeness, we should add that under the same conditions we also have

$$\operatorname{var}\hat{a}_0 = \sigma^2/n \ , \quad \operatorname{cov}(\hat{a}_0, \hat{a}_1) = \operatorname{cov}(\hat{a}_0, \hat{b}_1) = \operatorname{cov}(\hat{a}_0, \hat{c}_1) = \operatorname{cov}(\hat{a}_0, \hat{\varphi}) = 0 \tag{22}$$

It is clear that if the errors of estimation are not sufficiently small, there is little interest in attempting to find the correct values for $\operatorname{var}\hat{c}_1$ and $\operatorname{var}\hat{\varphi}$ since in this case the estimate $\hat{c}_1$ will not generally be significant.

We shall see later, however, that it can be useful to know the mean value and the variance of $\hat{c}_1$ with good accuracy when c_1 is zero or close to zero.

To this end, let us put

$$\hat{a}_1 = a_1 + e_1, \quad \hat{b}_1 = b_1 + e_2 \text{ et } \hat{c}_1 = c'_1 + t \tag{23}$$

where $c'_1 = \mathrm{E}(\hat{c}_1)$ which implies that $\mathrm{E}(t) = 0$ and $\operatorname{var} t = \operatorname{var}\hat{c}_1$, since $\hat{c}_1$ is deduced from $\hat{a}_1$ and $\hat{b}_1$ by means of the first of relations (13). It follows that we have

$$\hat{c}_1^2 = c_1'^2 + 2c'_1 t + t^2 = c_1^2 + 2a_1e_1 + 2b_1e_2 + e_1^2 + e_2^2 \tag{24}$$

Putting var $\hat{a}_1 = \text{var}\ \hat{b}_1 = v$, we then deduce

$$\mathrm{E}(\hat{c}_1^2) = c'^2_1 + \text{var}\ t = c_1^2 + 2v, \tag{25}$$

whilst by assuming n to be sufficiently large and adopting the approximation of a normal distribution for the distributions of e_1, e_2 and t, we have ; var $e_1^2 = \text{var}\ e_2^2 = 2v^2$, var $t^2 = 2\text{var}^2 t$ and we obtain

$$\text{var}\ \hat{c}_1^2 = 4c'^2_1\ \text{var}\ t + 2\ \text{var}^2 t = 4c_1^2 v + 4v^2 \tag{26}$$

The solution of equations (25) and (26) with respect to var t and c'^2_1 next gives

$$\text{var}\ t = \text{var}\ \hat{c}_1 = c_1^2 + 2v - \sqrt{c_1^4 + 2c_1^2 v + 2v^2} \tag{27}$$

and

$$c_1'^2 = \sqrt{c_1^4 + 2c_1^2 v + 2v^2} \tag{28}$$

which, for $c_1 = 0$, leads to

$$c'^2_1 = v\sqrt{2} \quad \text{and} \quad \text{var}\ \hat{c}_1 = v(2 - \sqrt{2}) \tag{29}$$

With regard to the suitable level at which the test on the statistic (11) should be applied, because of the asymptotic independance of the estimates of the coefficients a_1 and b_1 and of the estimate of a_0, there is no need to correct the level in the manner of 1.3.3 when we wish only to establish the significance of the periodic part of the deterministic component.

3.1.4.1 **Example 29.** *Estimation of a simulated periodic component introduced in advance into the total annual rainfall series for Uccle*

The annual rainfall totals for Brussels-Uccle between 1833 and 1969, in mm, were increased by the quantity 50 sin αi, with $\alpha = 2\pi/9{,}5$, i being the chronological ordinal number of the rainfall total in the series. The first ten and the last ten terms of the series thereby obtained are given in Table 29.

The amplitude of the periodic component introduced into the series is of the order of the standard deviation of this series and it appeared interesting to compare the true deterministic component with its estimate given by the method of least squares.

For the calculation of the matrix 3.1.4(4) besides $n = 137$, we obtain, expressing αi in radians :

$$\Sigma \sin \alpha i = \sin 0{,}661388 + \sin (2 \times 0{,}661388) + \cdots + \sin (137 \times 0{,}661388) = 2{,}975$$

$$\Sigma \cos \alpha i = -0{,}2465, \quad \Sigma \sin^2 \alpha i = 68{,}88, \quad \Sigma \sin \alpha i \cos \alpha i = -0{,}06377$$

$$\Sigma \cos^2 \alpha i = 68{,}12 \tag{1}$$

Calculation of the elements of the vector β by means of the relations 3.1.4(5) gives successively :

$$\beta_1 = \Sigma x_i = 841{,}5 + 630{,}7 + \cdots + 800{,}3 = 107143{,}9$$

$$\beta_2 = \Sigma x_i \sin \alpha i = 841{,}5 \times \sin 0{,}661388 + 630{,}7 \times \sin (2 \times 0{,}661388) + \cdots$$

$$+ 800{,}3 \times \sin (137 \times 0{,}661388) = 4788{,}7$$

$$\beta_3 = \Sigma x_i \cos \alpha i = 999{,}5 \tag{2}$$

The system of equations represented by relation 3.1.4(3) and satisfied by the elements of the vector $\hat{\theta}$ thus becomes here :

$$\begin{cases} 137\hat{a}_0 + 2{,}975\hat{a}_1 - 0{,}2465\hat{b}_1 = 107143{,}9 \\ 2{,}975\hat{a}_0 + 68{,}88\hat{a}_1 - 0{,}06377\hat{b}_1 = 4788{,}7 \\ -0{,}2465\hat{a}_0 - 0{,}06377\hat{a}_1 + 68{,}12\hat{b}_1 = 999{,}5 \end{cases} \tag{3}$$

From this we obtain

$$\hat{a}_0 = 781{,}3 \ , \ \hat{a}_1 = 35{,}79 \ , \ \hat{b}_1 = 17{,}54 \tag{4}$$

Furthermore, with

$$\Sigma x_i^2 = (841{,}5)^2 + (630{,}7)^2 + \ldots + (800{,}3)^2 = 85751323{,}99 \tag{5}$$

and

$$\hat{\theta}'\beta = 781{,}3 \times 107143{,}9 + 35{,}79 \times 4788{,}7 + 17{,}54 \times 999{,}5 = 83900447{,}87 \tag{6}$$

we obtain

$$\hat{e}'\hat{e} = x'x - \hat{\theta}'\beta = 85751323{,}99 - 83900447{,}87 = 1850876{,}12 \tag{7}$$

which gives the estimate of the variance of the random part as

$$\hat{\sigma}^2 = 1850876{,}12/(137 - 3) = 13812{,}51 \tag{8}$$

For the calculation of the covariance matrix of the elements of the vector $\hat{\theta}$, we note first that 3.1.4(6) can be expressed as

$$\text{var}\,\hat{\theta} = (u'u)^{-1}\sigma^2 = [n(u'u)^{-1}]\,\frac{\sigma^2}{n} = [\,(\frac{u'u}{n})^{-1}]\,\frac{\sigma^2}{n} \tag{9}$$

and this allows var $\hat{\theta}$ to be expressed as a function of σ^2/n.

TABLE 29. – Total annual rainfall x_i for Uccle, in mm, increased by the quantity 50 sin αi ($\alpha = 2\pi/9{,}5$). Values of the empirical autocovariances $\bar{v}_i$ and of the statistics X_i.

Year	i	x_i	$\bar{v}_i$	X_i
1833	1	841,5	1697	1,88
1834	2	630,7	370	1,97
1835	3	734,7	−1589	3,59
1836	4	913,6	872	4,07
1837	5	801,5	−830	4,51
1838	6	631,2	1985	6,98
1839	7	793,2	−380	7,07
1840	8	671,5	1215	7,98
1841	9	828,5	754	8,33
1842	10	719,1	1263	9,30
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
1960	128	971,0	−3539	120,4
1961	129	879,8	11894	125,2
1962	130	816,1	4291	125,8
1963	131	664,2	−10333	128,3
1964	132	755,0	−5818	129,0
1965	133	1073,9	7732	129,8
1966	134	1086,3	−3511	130,0
1967	135	755,9	2329	130,0
1968	136	822,4	—	
1969	137	800,3	—	

Dividing the elements of the matrix of the unknowns of the system (3) by 137 and calculating the inverse of the matrix thereby obtained, we find

$$\left(\frac{u'u}{n}\right)^{-1}_{11} = 1{,}0009454, \quad \left(\frac{u'u}{n}\right)^{-1}_{12} = -0{,}0432326, \quad \left(\frac{u'u}{n}\right)^{-1}_{13} = 0{,}00358235$$
$$\left(\frac{u'u}{n}\right)^{-1}_{22} = 1{,}990798, \quad \left(\frac{u'u}{n}\right)^{-1}_{23} = 0{,}00170731, \quad \left(\frac{u'u}{n}\right)^{-1}_{33} = 2{,}011255 \tag{10}$$

Multiplying by $\sigma^2/n = 13812{,}51/137 = 100{,}82$, we finally have

$$\text{var}\,\hat{a}_0 = 100{,}9 \ , \ \text{var}\,\hat{a}_1 = 200{,}7 \ , \ \text{var}\,\hat{b} = 202{,}8$$
$$\text{cov}\,(\hat{b}_0\,,\hat{a}_1) = -4{,}36 \ , \ \text{cov}\,(\hat{a}_0\,,\hat{b}_1) = 0{,}361 \ , \ \text{cov}\,(\hat{a}_1\,,\hat{b}_1) = 0{,}172 \tag{11}$$

It is thus confirmed that the approximations given by relations 3.1.4(19) are indeed justified.

To verify the significance of the periodic deterministic component, the calculation of the matrix 3.1.4(10) gives :

$$w_{11} = 202{,}8/[200{,}7 \times 202{,}8 - (0{,}172)^2] = 202{,}8/40701{,}9 = 0{,}00498$$
$$w_{22} = 200{,}7/40701{,}9 = 0{,}00493$$
$$w_{12} = -0{,}172/40701{,}9 = -0{,}00000423 \tag{12}$$

which gives the value of the statistic 3.1.4(11) as

$$X = (35{,}79)^2 \times 0{,}00498 - 2(35{,}79)(17{,}54) \times 0{,}00000423 + (17{,}54)^2 \times 0{,}00493 = 7{,}89 \tag{13}$$

Since this value exceeds 5,99, the critical value at the 0,05 level of the variate χ^2 with two degrees of freedom, the hypothesis $a_1 = 0$ and $b_1 = 0$ can be rejected.

If we use the approximate value 3.1.4(20) of the statistic X with $\hat{c}_1^2 = \hat{a}_1^2 + \hat{b}_1^2$, we obtain

$$X = 137[(35{,}79)^2 + (17{,}54)^2]/(2 \times 13812{,}51) = 7{,}88 \tag{14}$$

i.e., practically the same value as in (13).

In addition, we also have

$$\hat{c}_1 = \sqrt{(35{,}79)^2 + (17{,}54)^2} = 39{,}86 \ , \ \text{tg}\,\hat{\varphi} = 17{,}54/35{,}79 = 0{,}49008 \tag{15}$$

which gives us, in radians,

$$\hat{\varphi} = 0{,}45568 \tag{16}$$

these being estimates for which, using 3.1.4(21), we have

$$\text{var}\,\hat{c}_1 = 2 \times 13812{,}51/137 = 201{,}64 \quad \text{and} \quad \text{var}\,\hat{\varphi} = 201{,}64/(39{,}86)^2 = 0{,}12691 \tag{17}$$

Since the statistic

$$X = \frac{(50 - 39{,}86)^2}{201{,}64} + \frac{(0{,}45568)^2}{0{,}12691} = 2{,}15 \tag{18}$$

is less than 5,99 , the critical value at the 0,05 level of the χ^2 variate with two degrees of freedom, we can conclude, as might be expected, that c_1 and φ do not differ significantly from 50 and zero, respectively.

The significance of c_1 can also be established by means of relations 3.1.4(29).

With $v = 201{,}64$, assuming $c_1 = 0$, we in fact have

$$c_1' = \sqrt{201{,}64 \times \sqrt{2}} = 16{,}89 \ , \ \text{var}\,\hat{c}_1 = 201{,}64(2 - \sqrt{2}) = 118{,}12 \tag{19}$$

which gives the value of the statistic as

$$X=(\hat{c}_1 - c_1')^2/\text{var}\,\hat{c}_1 = (39{,}86 - 16{,}89)^2/118{,}12 = 4{,}47 \tag{20}$$

which is greater than 3,84, the critical value at the 0,05 level of the χ^2 variate with one degree of freedom.

The conclusions deduced from the values of statistics (13) and (14) are thus confirmed.

We should also point out that if we adopt

$$\hat{x}_i = 781{,}3 + 35{,}79 \sin \alpha i + 17{,}54 \cos \alpha i \tag{21}$$

where $\alpha = 2\pi/9{,}5$ as an estimate of the deterministic component of the series, we have, using (8) and 3.1.4(19) and (22),

$$\text{var}\,\hat{x}_i = \sigma^2/n + 2\sigma^2/n = 3\sigma^2/n = 3(13812{,}51)/137 = 302{,}46 \tag{22}$$

or

$$\sigma(\hat{x}_i) = 17{,}39 \tag{23}$$

which shows that the accuracy of the periodic part of the deterministic component is relatively mediocre.

3.1.5 Estimation of the deterministic component of a periodic series when the period of the component is known only approximately or is unknown, the variance of the random component also being unknown.

This estimation problem is considered under the same conditions as in 3.1.4, the only difference being that in this case the argument α also has to be estimated.

Two cases can be distinguished, depending on whether the existence of the deterministic component is well established *a priori* or whether this existence is itself in question.

In both cases, the method of determining α or the period $T = 2\pi/\alpha$ consists of calculating a_0, a_1 and b_1 for a set of values of T using the method described in 3.1.4 so as to retain only the value leading to the lowest estimate 3.1.4(9) of $\hat{\sigma}^2$ or, what comes down to the same thing, to the largest estimate of c_1. If the values of T cover a defined interval, the value of T adopted can consequently be considered as the estimate of T (or at least its approximate value) by means of the method of least squares.

The application of a significance test to the estimation of the deterministic component has different meanings for each case, so that the test itself cannot be applied in the same way. In the first case, the test aims simply at verifying that the estimate of the expected component possesses a sufficiently large amplitude to enable it to be considered as non-zero. In the second case, the test must establish whether, among the estimates obtained for c_1, the maximal value differs in significant fashion from the maximum which would be obtained under the assumption of a simple random series. Since the second test is, in principle, more severe than the former, we must anticipate that, in certain cases, for a given time series, we may end up concluding either that there is a significant periodic component, or that a periodic component is absent, depending on whether we are dealing with the first case or the second case. We should therefore always make clear, and justify, which case is being considered.

Since the estimation of α is an example of application of the method of non-linear least squares, we have, in the following, started by examining the problem of calculating the errors in the latter estimation case.

We have supplemented this calculation with a few considerations about the use of empirical covariances in the search for periodic components.

We should also note that generalizing the calculations to the case of models containing several periodic components with different periods follows directly.

3.1.5.1 Least squares estimation in the non-linear case.

Let x denote the vector representing the time series with sample size n, and let us consider the model

$$x = f(\theta) + e \quad \text{or} \quad f(\theta) = x - e \tag{1}$$

where $f(\theta)$ denotes another vector dependent upon a vector θ with n_1 elements, and e a vector whose components are independent and identically distributed with zero mean value.

The estimate $\hat{\theta}$ of the vector θ via the method of least squares is then the solution of the system of n_1 equations in the n_1 elements of this vector

$$\frac{d'f}{d\theta} \cdot f(\theta) = \frac{d'f}{d\theta} x \tag{2}$$

where $\frac{d'f}{d\theta}$ denotes the matrix transpose of the matrix $\frac{df}{d\theta}$ of the partial derivatives of $f(\theta)$ with respect to the elements of θ. Multiplying relation (1) by $\frac{d'f}{d\theta}$ and subtracting the two sides of the resulting equation from the two sides of the equality (2), respectively, with $\theta = \hat{\theta}$ in the latter equation, with

$$f(\hat{\theta}) - f(\theta) = \frac{df}{d\tilde{\theta}}(\hat{\theta} - \theta) \tag{3}$$

we obtain

$$\frac{d'f}{d\tilde{\theta}} \cdot \frac{df}{d\tilde{\theta}}(\hat{\theta} - \theta) = \frac{d'f}{d\tilde{\theta}} e \tag{4}$$

which leads to

$$\hat{\theta} - \theta = \left(\frac{d'f}{d\tilde{\theta}} \cdot \frac{df}{d\tilde{\theta}}\right)^{-1} \frac{d'f}{d\tilde{\theta}} e \tag{5}$$

and consequently to

$$(\hat{\theta} - \theta)(\hat{\theta} - \theta)' = \left(\frac{d'f}{d\tilde{\theta}} \cdot \frac{df}{d\tilde{\theta}}\right)^{-1} \frac{d'f}{d\tilde{\theta}} ee' \frac{df}{d\tilde{\theta}} \left(\frac{d'f}{d\tilde{\theta}} \cdot \frac{df}{d\tilde{\theta}}\right)^{-1} \tag{6}$$

Taking the mean value, we thus have

$$\operatorname{var} \hat{\theta} = \left(\frac{d'f}{d\theta} \cdot \frac{df}{d\theta}\right)^{-1} \sigma^2 \tag{7}$$

where σ^2 denotes the common variance of the elements of the vector e.

In the case of relation 3.1.4(1), the matrix $\frac{df}{d\theta}$ is formed by the four vectors with n elements.

$$\frac{\partial f}{\partial a_0} = \| 1 \|, \quad \frac{\partial f}{\partial a_1} = \| \sin \alpha i \|, \quad \frac{\partial f}{\partial b_1} = \| \cos \alpha i \|, \quad \frac{\partial f}{\partial \alpha} = \| c_1 i \cos (\alpha i + \varphi) \| \tag{8}$$

where $i = 1, 2, \ldots, n$, with $c_1 = \sqrt{a_1^2 + b_1^2}$, $\cos \varphi = a_1/c_1$ and $\sin \varphi = b_1/c_1$.

Thus we have :

$$(\operatorname{var} \hat{\theta})^{-1} = \frac{1}{\sigma^2} \begin{Vmatrix} n & \Sigma \sin \alpha i & \Sigma \cos \alpha i & c_1 \Sigma i \cos (\alpha i + \varphi) \\ \Sigma \sin \alpha i & \Sigma \sin^2 \alpha i & \Sigma \sin \alpha i \cos \alpha i & c_1 \Sigma i \sin \alpha i \cos (\alpha i + \varphi) \\ \Sigma \cos \alpha i & \Sigma \sin \alpha i \cos \alpha i & \Sigma \cos^2 \alpha i & c_1 \Sigma i \cos \alpha i \cos (\alpha i + \varphi) \\ c_1 \Sigma i \cos (\alpha i + \varphi) & c_1 \Sigma i \sin \alpha i \cos (\alpha i + \varphi) & c_1 \Sigma i \cos \alpha i \cos (\alpha i + \varphi) & c_1^2 \Sigma i^2 \cos^2 (\alpha i + \varphi) \end{Vmatrix} \tag{9}$$

from which, by inversion, we obtain the covariance matrix $\operatorname{var} \hat{\theta}$ of the errors of estimation for the elements of the vector θ.

3.1.5.2 *Empirical autocovariances of a time series with a periodic deterministic component*

The empirical autocovariances of order k, for a time series x_i, $i = 1, 2, \ldots, n$, can be defined by the relations

$$\bar{v}_k = \left[\sum_1^{n-k} x_i x_{i+j} - \left(\sum_1^{n-k} x_i \right) \left(\sum_1^{n-k} x_{i+k} \right) \Big/ (n-k) \right] \Big/ (n-k-1) \tag{1}$$

for $k = 1, 2, \ldots, n-2$, while the empirical variance $\bar{v}_0$ is still determined by the relation

$$\bar{v}_0 = \left[\sum_1^n x_i^2 - \left(\sum_1^n x_i \right)^2 \Big/ n \right] \Big/ (n-1) \tag{2}$$

For n sufficiently large, we have, to a first approximation,

$$\left(\sum_1^n \sin^2 \alpha i \right) \Big/ n = \left(\sum_1^n \cos^2 \alpha i \right) \Big/ n = \frac{1}{2}, \quad \left(\sum_1^n \sin \alpha i \right) \Big/ n = \left(\sum_1^n \cos \alpha i \right) \Big/ n = 0 \tag{3}$$

from which we deduce, in the case of a series of the form 3.1.4(1)

$$\mathrm{E}(\bar{v}_0) = \frac{c_1^2}{2} + \sigma^2 \quad \text{and} \quad \mathrm{E}(\bar{v}_k) = \frac{c_1^2}{2} \cos \alpha k \tag{4}$$

with $c_1^2 = a_1^2 + b_1^2$ and where σ^2 is the common variance of the quantities e_i.

For normal variables e_i we have $\operatorname{var} e_i^2 = 2\sigma^4$ and $\operatorname{cov} (e_{i+k} e_{i+l}, e_i^2) = 0$, with k and $l \neq 0$, neglecting the terms in σ^2/n, giving, to a first approximation,

$$\operatorname{var} \bar{v}_0 = 2\sigma^4/(n-1); \quad \operatorname{var} \bar{v}_k = \sigma^4/(n-k-1) \text{ et } \operatorname{cov} (\bar{v}_k, \bar{v}_l) = 0 \tag{5}$$

for $k \neq l \geqslant 1$.

We conclude that the series $\bar{v}_k$ possesses a deterministic component with the same period as the initial series x_i. Furthermore, if we put for the series x_i

$$r^2 = \sigma^2/c_1^2 \tag{6}$$

the corresponding ratio for the series $\bar{v}_k$ becomes

$$r^2(\bar{v}_k) = \frac{\sigma^4}{n-k-1} \cdot \frac{4}{c_1^4} = \frac{4r^2}{n-k-1} \cdot r^2 \tag{7}$$

The relative magnitude of the random component in the terms of the series of the autocovariances is less than that of the random component in the initial series, when we have

$$n - k - 1 > 4r^2 \tag{8}$$

Consequently, when the series is sufficiently long, calculation of the series of the autocovariances can be used to reveal the possible existence of a periodic deterministic component.

To do this, we can use the statistic

$$X_k = \left[\sum_1^k (n-i-1)\bar{v}_i^2\right] \Big/ \bar{v}_0^2 \tag{9}$$

since, using the preceding argument, under the null hypothesis for a simple, random, initial series x_i, this statistic has approximately a χ^2 distribution with k degrees of freedom when n and $(n-k)$ are sufficiently large.

It is obvious that, for $k = 1$, the test based on the statistic (9) is equivalent to the serial correlation test described in 1.4.1. Furthermore, if $\cos\alpha = 0$, i.e. if in the time series the period of the deterministic component is equal to 4, then the test on the statistic X_1 possesses zero power since in (4), we have $E(\bar{v}_1) = 0$. We can deduce that in general the calculation of the statistics X_k for $k > 1$ constitutes a useful extension of the test based on the serial correlation used in the examination of the simple random character of a series of observations.

3.1.5.3 Determination of the periodic deterministic component when the period is unknown

From the foregoing, it follows that, if n is sufficiently large, the series of the empirical autocovariances will bring out quite clearly the existence of the periodic deterministic component. In this case, $\bar{v}_1$ and $\bar{v}_2$ are good estimates of $E(\bar{v}_1)$ and $E(\bar{v}_2)$ and a first estimate of the argument α can be found by noting that relation 3.1.5.2(4) written for $k = 1$ and $k = 2$ gives

$$\frac{2}{c_1^2} = \frac{\cos\alpha}{E(\bar{v}_1)} = \frac{\cos 2\alpha}{E(\bar{v}_2)} \tag{1}$$

which, replacing $E(\bar{v}_1)$ and $E(\bar{v}_2)$ respectively by $\bar{v}_1$ and $\bar{v}_2$, gives the equation in $\cos\alpha$:

$$2\bar{v}_1 \cos^2\alpha - \bar{v}_2 \cos\alpha - \bar{v}_1 = 0 \tag{2}$$

of which only the root with the same sign as $\bar{v}_1$ should be retained.

Next, we deduce from (1) a first estimate of c_1^2, then, using the first of relations 3.1.5.2(4) and the calculation of $\bar{v}_0$, a first estimate of σ^2, which allows us to calculate the ratio r^2.

Differentiating relation (2) and replacing $\bar{v}_1$ and $\bar{v}_2$ by the values of $E(\bar{v}_1)$ and $E(\bar{v}_2)$ given by 3.1.5.2(4), we also obtain

$$-\frac{c_1^2}{2}\sin\alpha(2\cos^2\alpha+1)d\alpha=\cos\alpha d\bar{v}_2-(2\cos^2\alpha-1)d\bar{v}_1 \tag{3}$$

which, adopting σ^4/n as an approximate value of var $\bar{v}_1$ and of var $\bar{v}_2$, gives

$$\operatorname{var}\hat{\alpha}=\frac{4\cos^4\alpha-3\cos^2\alpha+1}{(2\cos^2\alpha+1)^2\sin^2\alpha}\cdot\frac{4r^4}{n} \tag{4}$$

It follows that the optimal value of α can be obtained by attributing to α a series of values covering a sufficiently large interval around the solution $\hat{\alpha}$ obtained from (2) and calculating for each of these values the corresponding estimates of a_0, a_1, b_1 and $c_1=\sqrt{a_1^2+b_1^2}$ using the method described in 3.1.4. The final estimate is that which leads to the highest value of c_1, and the covariance matrix associated with this estimation is deduced by inversion of the matrix 3.1.5.1(9)

When n is not large enough to enable equation (2) to give an adequate estimate of α, determination of the optimal value of α is still possible if we possess *a priori* information on the interval in which this value should be sought.

Otherwise, the choice of such an interval can be only arbitrary and the adoption of the value of α which would be obtained by this method would be justified only if the corresponding estimate of c_1 differs significantly from the maximal estimate of c_1 which would be given in this interval by a simple random series.

The latter question is examined later in applying the properties of the extreme values of the recurrent series.

To determine the level at which the significance test on the estimates of a_1 and b_1 should be applied, it is obviously necessary to take into account a joint estimate of the period. In particular, at the level $\alpha_0=0{,}05$, the critical value of the statistic 3.1.4(11) can be obtained by subtracting 1 from the critical value of the χ^2 variate with three degrees of freedom.

3.1.5.4 **Example 30.** *Estimate of a periodic deterministic component introduced in advance into the total annual rainfall series for Uccle, when the period itself is subject to estimation*

The series investigated is once again that of the total annual rainfall at Uccle, in mm, increased by the quantity 50 sin αi, with $\alpha=2\pi/9{,}5$ and whose values x_i are indicated in Table 29.

Using the results obtained earlier in 3.1.4.1(2) and (5), with 3.1.5.2(2), we have

$$\bar{v}_0=85751323{,}99/136-(107143{,}9)^2/(137\times136)=14390 \tag{1}$$

Similarly, with 3.1.5.2(1), we obtain

$$\begin{aligned}\bar{v}_1 &= (841{,}5\times630{,}7+630{,}7\times734{,}7+\cdots+822{,}4\times800{,}3)/135\\ &-(841{,}5+630{,}7+\cdots+822{,}4)(630{,}7+734{,}7+\cdots+800{,}3)/(136\times135)=1697\end{aligned} \tag{2}$$

together with the values of $\bar{v}_i$ for $i=2, 3, \ldots, 135$ given in Table 29.

We then deduce, with the aid of 3.1.5.2(9),

$$X_1 = (137 - 1 - 1)(1697)^2/(14390)^2 = 1,88$$

$$X_2 = [135(1697)^2 + 134(370)^2]/(14390)^2 = 1,97 \qquad (3)$$

together with the values of X_i for $i = 3, 4, \ldots, 135$ given in Table 29.

Thus we see that the values X_i are all very close to the mathematical expectation i of the χ^2 variate with i degrees of freedom, i.e., that the distribution of the empirical autocovariances does not differ from that given by a simple random series. We therefore anticipate that the first estimate of the argument α based on these autocovariances will have only a very mediocre accuracy.

With $\bar{v}_2/\bar{v}_1 = 0,21803$, equation 3.1.5.3(2) becomes

$$2\cos^2\alpha - 0,21803\cos\alpha - 1 = 0 \qquad (4)$$

of which the positive root (of the same sign as $\bar{v}_1$) is equal to

$$\cos\hat{\alpha} = 0,76371 \qquad (5)$$

from which we obtain the estimate in radians

$$\hat{\alpha} = 0,70175 \qquad (6)$$

and the estimate of the period

$$\hat{T} = 2\pi/\hat{\alpha} = 6,28318/0,70175 = 9,0 \qquad (7)$$

With 3.1.5.3(1), we also get

$$c_1^2 = 2 \times 1697/0,76371 = 4444 \qquad (8)$$

and, using 3.1.5.2(4),

$$\hat{\sigma}^2 = 14390 - (4444/2) = 12168 \qquad (9)$$

that is to say,

$$r^2 = 12168/4444 = 2,74 \qquad (10)$$

With $\cos^2\hat{\alpha} = 0,58325$ and $\sin^2\hat{\alpha} = 0,41675$, 3.1.5.3(4) finally gives

$$\operatorname{var}\hat{\alpha} = \frac{4(0,58325)^2 - 3 \times 0,58325 + 1}{[2(0,58325) + 1]^2 \times 0,41675} \cdot \frac{4(2,74)^2}{137} = 0,31234 \times 0,21920 = 0,06847 \qquad (11)$$

or

$$\sigma(\hat{\alpha}) = 0,26166 \qquad (12)$$

From this we deduce that the actual value of α lies within the 0,95 confidence interval (approximately) : $0,70175 \pm 1,96 \times 0,26166$, which corresponds to a period $T = 2\pi/\alpha$ lying between 5,17 and 33,3.

The estimates of a_0, a_1, b_1 and c_1 for $T = 5,0$ up to $T = 12,0$, with values increasing in stages of 0,1 , give values of c_1 which vary relatively continuously with maximal values for $T = 5,6$, 7,1 , 8,1 , 8,8 , 9,6 and 12,0 , amongst which the largest value is that of $T = 9,6$. These results have been extended by some new estimates calculated for $T = 9,50$ up to $T = 9,70$, using values increasing in stages of 0,01. These various estimates are given in Table 30. They show that the optimal value for T, to an accuracy of 0,01 , is $T^* = 9,60$.

For the calculation of the covariance matrix of the vector $\hat{\theta}$, we note that by writing the matrix 3.1.5.1(9) in the form

$$(var\,\hat{\theta})^{-1} = \frac{n}{\sigma^2}\left\| \frac{1}{n}\frac{d'f}{d\theta} \cdot \frac{df}{d\theta} \right\|$$

and taking the inverse, we find for $\alpha^* = 2\pi/9{,}60$ and the values of c_1 and of φ deduced from the corresponding values of a_1 and b_1 of Table 30 :

$$\left(\frac{1}{n}\frac{d'f}{d\theta}\frac{df}{d\theta}\right)^{-1}_{11} = 1{,}00157, \qquad \left(\frac{1}{n}\frac{d'f}{d\theta}\frac{df}{d\theta}\right)^{-1}_{12} = -0{,}093216$$

$$\left(\frac{1}{n}\frac{d'f}{d\theta}\frac{df}{d\theta}\right)^{-1}_{13} = 0{,}035713, \qquad \left(\frac{1}{n}\frac{d'f}{d\theta}\frac{df}{d\theta}\right)^{-1}_{14} = -2{,}7317.10^{-5}$$

$$\left(\frac{1}{n}\frac{d'f}{d\theta}\frac{df}{d\theta}\right)^{-1}_{22} = 5{,}8420, \qquad \left(\frac{1}{n}\frac{d'f}{d\theta}\frac{df}{d\theta}\right)^{-1}_{23} = -3{,}0413$$

$$\left(\frac{1}{n}\frac{d'f}{d\theta}\frac{df}{d\theta}\right)^{-1}_{24} = 0{,}0016990, \qquad \left(\frac{1}{n}\frac{d'f}{d\theta}\frac{df}{d\theta}\right)^{-1}_{33} = 4{,}3892$$

$$\left(\frac{1}{n}\frac{d'f}{d\theta}\frac{df}{d\theta}\right)^{-1}_{34} = -0{,}0013323, \qquad \left(\frac{1}{n}\frac{d'f}{d\theta}\frac{df}{d\theta}\right)^{-1}_{44} = 7{,}4833.10^{-7} \qquad (13)$$

Multiplying by $\hat{\sigma}^2/n = 13709/137 = 100{,}1$, we obtain the elements of the required covariance matrix, where the indices 1, 2, 3 and 4 correspond respectively to a_0, a_1, b_1 and α.

TABLE 30. – Modified total annual rainfall for Uccle. Estimates of a_0, a_1, b_1 and σ^2 for various values for the period $T = 2\pi/\alpha$

T	a_0	a_1	b_1	c_1	σ^2
5,6	781,9	22,8	22,3	31,8	14088
7,1	782,3	−20,3	33,2	38,9	13838
8,1	781,9	−8,6	−36,0	37,0	13910
8,8	781,5	27,3	−11,3	29,5	14163
9,50	781,3	35,8	17,5	39,9	13792
9,51	781,3	35,3	19,3	40,2	13777
9,52	781,3	34,6	21,0	40,5	13763
9,53	781,3	33,9	22,7	40,8	13751
9,54	781,3	33,0	24,4	41,0	13740
9,55	781,4	32,0	26,0	41,3	13732
9,56	781,4	31,0	27,5	41,4	13724
9,57	781,4	29,8	29,0	41,6	13718
9,58	781,4	28,6	30,3	41,66	13714
9,59	781,4	27,2	31,6	41,72	13711
9,60	781,5	25,8	32,8	41,75	13709
9,61	781,5	24,3	33,9	41,74	13709
9,62	781,5	22,8	34,9	41,70	13711
9,63	781,5	21,2	35,8	41,6	13714
9,64	781,6	19,5	36,6	41,5	13719
9,65	781,6	17,9	37,3	41,4	13724
9,66	781,6	16,2	37,9	41,2	13732
9,67	781,7	14,4	38,4	41,0	13740
9,68	781,7	12,7	38,8	40,8	13750
9,69	781,7	10,9	39,0	40,5	13761
9,70	781,7	9,2	39,2	40,3	13773
12,0	782,7	−21,8	24,2	32,5	14065

In particular, the statistic 3.1.4(11) has the value

$$X = \frac{(25,8)^2 \times 4,39 + (32,8)^2 \times 5,84 + 2(25,8)(32,8) \times 3,04}{[5,84 \times 4,39 - (3,04)^2] \times 100,1} = 8,74 \tag{14}$$

which is greater than 6,81 , the critical value at the 0,05 level of the χ^2 variate with three degrees of freedom, reduced by 1.

Furthermore, for the final estimate $\alpha^* = 2\pi/9,60 = 0,6545$ we also have

$$\text{var } \alpha^* = 7,4833.10^{-7} \times 100,1 = 7,491.10^{-5} \tag{15}$$

that is :

$$\sigma(\alpha^*) = 0,008655 \tag{16}$$

So the interval $0,6545 \pm 1,96 \times 0,008655$ forms an approximate confidence interval for α at the 0,95 level, i.e., for T, this interval extends from 9,36 to 9,86.

This result is perfectly compatible with the actual period : $T = 9,5$.

3.2 RECURRENT SERIES

When a time series is not a simple random series, the alternative assumption which can be considered is that of a recurrence relation between the terms of the series. As long as this assumption is satisfied, estimation of the coefficients of the recurrence by the method of least squares, and that of the variance of the random residual, offers a means of characterizing both the chronological progression of the terms of the series and their statistical distribution. We should note, however, that this estimation is single-valued only if the residual is simple-random.

As we have already emphasized, estimation in the case of such a series is an important problem since it can be applied to all series of observations whose successive terms are not independent, such as, for instance, series of hourly, bi-hourly and daily observations, etc.

Furthermore, the theory of recurrent series leads to new applications in the field of significance tests, in particular in seeking periodic deterministic components in a time series.

3.2.1 General properties of recurrent series

Using the same notation as in 3, but with d_i replaced by e_{i+k} , we consider the recurrence relation of order $k \geqslant 1$

$$x_{i+k} = a_0 + a_1 x_i + a_2 x_{i+1} + \cdots + a_k x_{i+k-1} + e_{i+k},$$
$$i = 1, 2, \cdots, n - k \tag{1}$$

for which we assume $a_1 \neq 0$ and where the quantities e_{i+k} are random variables distributed independently and identically with zero mean and with variance $\text{var } e = \sigma^2$. The case of a simple random series $x_i = a_0 + e_i$, $(a_j = 0, j \geqslant 1)$ therefore becomes that of a recurrence of order $k = 0$.

We immediately deduce

$$\text{cov}(x_i, e_j) = 0 \quad \text{for } i < j$$

and

$$\text{cov}(x_i, e_j) = \sigma^2 \quad \text{for } i = j \tag{2}$$

Furthermore, if we denote by x_i' the solution of the homogeneous recurrence relation associated with equation (1),

$$x_{i+k} = a_1 x_i + a_2 a_{i+1} + \ldots + a_k x_{i+k-1} \tag{3}$$

defined by the relations $x_i' = x_i$, for $i = 1, 2, \ldots, k$, and by x_i'', the solution of the same equation defined by the relations $x_i'' = 0$, for $i = 2, 3, \ldots, k$ and $x_{k+1}'' = 1$, then relation (1) implies

$$\begin{aligned} x_{k+1} &= x_{k+1}' + (e_{k+1} + a_0)x_{k+1}'' \\ x_{k+2} &= x_{k+2}' + (e_{k+1} + a_0)x_{k+2}'' + (e_{k+2} + a_0)x_{k+1}'' \end{aligned} \tag{4}$$

and, in general,

$$x_{k+j} = x_{k+j}' + (e_{k+1} + a_0)x_{k+j}'' + (e_{k+2} + a_0)x_{k+j-1}'' + \cdots + (e_{k+j} + a_0)x_{k+1}'' \tag{5}$$

which leads to the recurrent series x_i in the form of a sum of particular solutions of the homogeneous equation (3) directly related to the successive values of the random quantity e_{i+k}.

The general solution of the homogeneous equation (3) is the linear combination of k particular solutions of the form

$$x_i^{(j)} = b_r^i, \quad b_s^i \sin \alpha_s i \quad \text{ou} \quad b_s^i \cos \alpha_s i \tag{6}$$

where b_r, $b_s(\cos \alpha_s \pm \sqrt{-1}.\sin \alpha_s)$ are the real or complex roots of the characteristic equation

$$z^k - a_k z^{k-1} - a_{k-1} z^{k-2} - \ldots - a_1 = 0 \tag{7}$$

Consequently, if the series x_i is stationary (i.e. if $E(x_i)$ and var x_i are constants), then when we put it into the form (5) we must have

$$|b_r| < 1 \quad \text{and} \quad |b_s| < 1 \tag{8}$$

which, in particular, implies for the coefficient a_1 of equations (1) and (3)

$$|a_1| < 1 \tag{9}$$

The various components of a stationary recurrent time series x_i which appear in its form (5) are progressively damped as i increases, this damping occurring in oscillatory fashion when equation (7) possesses negative or complex roots.

This also shows that model (1) is unsuitable for representing a stationary time series possessing a periodic component since the independence of the quantities e_{i+k} in the relation

$$x_{i+k} = x_i + e_{i+k} \tag{10}$$

implies the relation

$$\text{var } x_{i+jk} = \text{var } x_i + j \text{ var } e \tag{11}$$

Taking the mean of relation (1), with $E(e_{i+k}) = 0$, we obtain for the mean value $m = E(x)$ of the series

$$m = a_0 + m(a_1 + a_2 + \ldots + a_k) \tag{12}$$

while by subtracting the latter relation from (1), multiplying the relation obtained by $(x_j - m)$ and putting

$$v_{r-s} = \operatorname{cov}(x_r, x_s) = \mathrm{E}(x_r - m, x_s - m) = v_{s-r} \tag{13}$$

we obtain in similar fashion for $j < i + k$, using the first of relations (2),

$$v_{i-j+k} = a_1 v_{i-j} + a_2 v_{i-j+1} + \cdots + a_k v_{i-j+k-1} \tag{14}$$

which shows that the series of the autocovariances of the series satisfies the homogeneous recurrence relation (3) associated with equation (1). Consequently, for l sufficiently large, v_l becomes negligible for all values of l.

In the case of $j = i + k$, the same operation leads, using the second relation in (2), to the equation

$$v_0 = a_1 v_k + a_2 v_{k-1} + \ldots + a_k v_1 + \sigma^2 \tag{15}$$

which in conjunction with equation (14) for $j = i, i+1, \ldots, i+k-1$, gives a system of $(k+1)$ linear equations from which we can obtain $v_0, v_1, \ldots, v_k$ as a function of $a_1, a_2, \ldots, a_k$ and of σ^2. In particular, we deduce that expressions for $v_0, v_1, \ldots, v_k$ thus obtained are of the form

$$v_l = \sigma^2 f_l(a_1, a_2, \cdots, a_k), \quad l = 0, 1, 2, \cdots, k \tag{16}$$

and that if $k = 1$, we have

$$v_0 = \sigma^2/(1 - a_1^2) \tag{17}$$

3.2.2 Estimation and determination of the recurrence

If we denote by x, e and u respectively the vectors and the matrix

$$x = \| x_{i+k} \|, \quad e = \| e_{i+k} \| \quad \text{and} \quad u = \| 1\ x_i\ x_{i+1} \ldots x_{i+k-1} \|$$

where $i = 1, 2, \ldots, n-k$ and by θ the vector of the unknown elements $a_0, a_1, \ldots, a_k$, then equation 3.2.1(1) is written

$$x = u\theta + e \tag{1}$$

and the estimate $\hat{\theta}$ given by the method of least squares is once again that given by the relation

$$u'u\hat{\theta} = u'x = \beta \tag{2}$$

where this time we have

$$u'u = \begin{Vmatrix} n-k & \Sigma x_i & \Sigma x_{i+1} \cdots & \Sigma x_{i+k-1} \\ \Sigma x_i & \Sigma x_i^2 & \Sigma x_i x_{i+1} & \Sigma x_i x_{i+k-1} \\ \vdots & & & \\ \Sigma x_{i+k-1} & \Sigma x_i x_{i+k-1} & & \Sigma x_{i+k-1}^2 \end{Vmatrix}, \quad \beta = u'x = \begin{Vmatrix} \Sigma x_{i+k} \\ \Sigma x_i x_{i+k} \\ \vdots \\ \Sigma x_{i+k-1} x_{i+k} \end{Vmatrix} \tag{3}$$

and where Σ indicates $\sum_{i=1}^{n-k}$.

The covariance matrix of the vector $\hat{\theta}$ is given by the relation

$$\text{var}\,\hat{\theta} = (u'u)^{-1}\,\sigma^2 \tag{4}$$

and an estimate of $\sigma^2 = \text{var}\,e$ is given by the relation

$$\hat{\sigma}^2 = \hat{e}'\hat{e}/(n-2k-1) \tag{5}$$

where we have put $\hat{e} = x - u\hat{\theta}$.

In particular, for $k=0$, we obtain

$$u'u = n\,, \quad \beta = \Sigma x_i\,, \quad \hat{a}_0 = \Sigma x_i/n \quad \text{and} \quad \text{var}\,\hat{a}_0 = \sigma^2/n \tag{6}$$

Since, moreover, putting $a_1 = 0$ in 3.2.1(17) gives $\sigma^2 = v_0$, we again find, using (4), the well known result

$$\text{var}\,\hat{a}_0 = v_0/n \tag{7}$$

Similarly, for $k=1$, putting

$$V_0 = (n-1)\Sigma x_i^2 - (\Sigma x_i)^2 \text{ et } V_1 = (n-1)\Sigma x_i x_{i+1} - (\Sigma x_i)(\Sigma x_{i+1}) \tag{8}$$

we obtain

$$\hat{a}_1 = V_1/V_0 \tag{9}$$

and

$$\text{var}\,\hat{a}_1 = (n-1)\sigma^2/V_0 \tag{10}$$

Since to a first approximation we have $E(V_0) = (n-1)^2 v_0$, with 3.2.1(17), we have

$$\text{var}\,\hat{a}_1 = (1-a_1^2)/(n-1) \tag{11}$$

So we can test the hypothesis $a_1 = 0$ by means of the statistic

$$X_1 = (n-1)V_1^2/V_0^2 \tag{12}$$

which under the null hypothesis $a_1 = 0$ has an approximate χ^2 distribution with one degree of freedom. Thus we recover the statistic 3.1.5.2(9) for $k=1$.

Generally, it is obvious that estimation of the recurrence relation of a time series can be reduced to seeking a residual $\hat{e}$ whose elements are subject to a recurrence of order zero. Consequently, we can fit to the series, in turn, recurrences of increasing order until a residual $\hat{e}$ is obtained for which the statistic (12) is no longer significant.

To calculate this statistic, using (5), we have

$$\bar{v}_0(e) = \Sigma \hat{e}_{i+k}^2/(n-2k-1) \tag{13}$$

and the corresponding empirical autocovariance of first order is given by

$$\bar{v}_1(e) = [\Sigma \hat{e}_{1+k}\hat{e}_{i+k+1} - (\Sigma \hat{e}_{i+k})(\Sigma \hat{e}_{i+k+1})/(n-k-1)]/(n-2k-2) \tag{14}$$

which leads to the statistic

$$X_1 = (n-2k-2)\bar{v}_1^2(e)/\bar{v}_0^2(e) \tag{15}$$

We can also check the order of the recurrence when starting to fit a recurrence of given order $k > 1$. In this case, the test of the hypothesis $a_1 = 0$ obtained by comparing the estimate of the coefficient a_1 with its variance var $\hat{a}_1$ under the hypothesis $a_1 = 0$ by means of the statistic $X = (\hat{a}_1)^2/(\text{var } \hat{a}_1)_{a_1=0}$ will allow us to decide either that the order of the recurrence is less than k, or that this order is at least equal to k.

3.2.3 Estimation of the mean value, the variance and the autocovariances of a recurrent series

Since equations 3.2.1(12) , (14) and (15) relate the mean value, the variance and the autocovariances of a recurrent series to the coefficients $a_0, a_1, \ldots, a_k$ of the recurrence and to the variance σ^2 of the residual, calculation of the first quantities by means of these relations, using estimates of the second quantities, gives appropriate estimates.

In calculating the variances and the covariances associated with these estimates, if we eliminate a_0 by subtracting both sides of relation 3.2.1(12) from both sides of relation 3.2.1(1), we obtain a relation of the form

$$x_{i+k} = f_i(\theta_1) + e_{i+k} \tag{1}$$

with

$$f_i(\theta_1) = m + a_1(x_i - m) + a_2(x_{i+1} - m) + \ldots + a_k(x_{i+k-1} - m) \tag{2}$$

where θ_1 this time denotes the vector formed by the elements $m, a_1, a_2, \ldots, a_k$.

If, additionally, we denote by $\frac{df}{d\theta_1}$ the matrix whose columns are the $(k+1)$ vectors with $(n-k)$ elements

$$\frac{\partial f}{\partial m} = \| 1 - \Sigma a_j \|, \quad \frac{\partial f}{\partial a_j} = \| x_{i+j-1} - m \|, \quad j = 1, 2, \ldots, k \tag{3}$$

then the covariance matrix of the estimates of the elements of the vector $\hat{\theta}_1$ is defined, as in 3.1.5.1, by the relation

$$(\text{var } \hat{\theta}_1)^{-1} = \frac{1}{\sigma^2}\frac{d'f}{d\theta_1}\frac{df}{d\theta_1} = \frac{1}{\sigma^2}\begin{Vmatrix} (n-k)(1-\Sigma a_j)^2 & (1-\Sigma a_j)\Sigma(x_i - m)\cdots & (1-\Sigma a_j)\Sigma(x_{i+k-1} - m) \\ (1-\Sigma a_j)\Sigma(x_i - m) & \Sigma(x_i - m)^2 & \Sigma(x_i - m)(x_{i+k-1} - m) \\ \vdots & & \\ (1-\Sigma a_j)\Sigma(x_{i+k-1} - m) & \Sigma(x_i - m)(x_{i+k-1} - m) & \Sigma(x_{i+k-1} - m)^2 \end{Vmatrix} \tag{4}$$

which, taking the mean, gives

$$(\text{var } \hat{\theta}_1)^{-1} = \frac{n-k}{\sigma^2}\begin{Vmatrix} (1-\Sigma a_j)^2 & 0 & 0\cdots & 0 \\ 0 & v_0 & v_1\cdots & v_{k-1} \\ \vdots & & & \\ 0 & v_{k-1}\cdots & & v_0 \end{Vmatrix} \tag{5}$$

From this we deduce

$$\operatorname{var} \hat{m} = \sigma^2/[(n-k)(1-\Sigma a_j)^2], \quad \operatorname{cov}(\hat{m}, \hat{a}_j) = 0$$

and

$$\operatorname{cov}(\hat{a}_j, \hat{a}_{j'}) = (\delta_{jj'}/\Delta_{k-1})\sigma^2/(n-k), \quad j, j' = 1, 2, \ldots, k \tag{6}$$

where we have denoted by Δ_{k-1} the value of the determinant formed by the matrix

$$\| A_{k-1} \| = \begin{Vmatrix} v_0 & v_1 \cdots & v_{k-1} \\ v_1 & v_0 \cdots & \\ \vdots & & \\ v_{k-1} & v_{k-2} \cdots & v_0 \end{Vmatrix} \tag{7}$$

and by $\delta_{jj'}$, the minor, in this determinant, of the element located on the row j and the column j'.

Since the estimate of the quantities $v_0, v_1, \ldots, v_k$ using relations 3.2.1(14) and (15) does not depend on that of m, it follows that the variances and covariances of these estimates are deduced uniquely from those of the estimates $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_k$.

Moreover, the form of the matrix (4) also shows that apart from the error of $\Sigma x_{i+j}/(n-k)$ on m, the estimates of $v_0, v_1, \ldots, v_k$ given through relations 3.2.1(14) and (15) are identical to the corresponding empirical variance and autocovariances of the series.

We conclude that, if we adopt the empirical variance $\overline{v}_0$ as an estimate of the variance of a recurrent series, then the variance of the associated error of estimation can be deduced from the expression for v_0 as a function of $a_1, a_2, \ldots, a_k$ and σ^2 obtained from relations 3.2.1(14) and (15).

In particular, for $k = 1$, we obtain

$$v_0 = \sigma^2/(1-a_1^2) \tag{8}$$

which, by differentiation, gives

$$\frac{d\hat{v}_0}{v_0} = \frac{d\hat{\sigma}^2}{\sigma^2} + \frac{2a_1 d\hat{a}_1}{1-a_1^2} \tag{9}$$

from which, assuming that the distribution of the quantities e_{i+k} is normal, i.e. that we have $\operatorname{var} \hat{\sigma}^2 = 2\sigma^4/(n-3)$ with 3.2.2(11), we obtain

$$\frac{\operatorname{var} \hat{v}_0}{v_0^2} \cong \frac{2}{n-1} \cdot \frac{1+a_1^2}{1-a_1^2} \tag{10}$$

Similarly, for $k = 2$, we have

$$v_0 = \frac{\sigma^2(1-a_1)}{(1+a_1)(1-a_1+a_2)(1-a_1-a_2)}, \quad v_1 = \frac{a_2 v_0}{1-a_1} \tag{11}$$

whereas (7) gives

$$\| A_1 \| = \begin{Vmatrix} v_0 & v_1 \\ v_1 & v_0 \end{Vmatrix}, \quad \| A_1 \|^{-1} = \frac{1}{v_0^2 - v_1^2} \begin{Vmatrix} v_0 & -v_1 \\ -v_1 & v_0 \end{Vmatrix}$$

From this we obtain

$$\operatorname{var} \hat{a}_1 = \operatorname{var} \hat{a}_2 = (1-a_1^2)/(n-2) \; ; \; \operatorname{cov}(\hat{a}_1, \hat{a}_2) = -a_2(1+a_1)/(n-2) \tag{13}$$

By differentiation, relation (11) next gives

$$\frac{d\hat{v}_0}{v_0} = \frac{d\hat{\sigma}^2}{\sigma^2} + A d\hat{a}_1 + B d\hat{a}_2 \tag{14}$$

with

$$A = -\frac{1}{1-a_1} - \frac{1}{1+a_1} + \frac{1}{1-a_1+a_2} + \frac{1}{1-a_1-a_2};$$

$$B = \frac{1}{1-a_1-a_2} - \frac{1}{1-a_1+a_2} \tag{15}$$

so that, with $\operatorname{var} \hat{\sigma}^2/\sigma^4 = 2/(n-4)$

$$\frac{\operatorname{var} \hat{v}_0}{v_0^2} = \frac{2}{n-4} + (A^2 + B^2) \operatorname{var} \hat{a}_1 + 2AB \operatorname{cov}(\hat{a}_1, \hat{a}_2) \tag{16}$$

The calculation of the covariance matrix associated with the estimate of the roots of the characteristic equation 3.2.1(7) is obtained in similar fashion. In particular, for $k = 2$ and when the roots are complex and of the form $b(\cos\alpha \pm \sqrt{-1} \cdot \sin\alpha)$, where $b > 0$, the relations

$$a_1 = -b^2 \quad \text{and} \quad a_2 = 2b\cos\alpha \tag{17}$$

imply :

$$\operatorname{var} \hat{b} = (\operatorname{var} \hat{a}_1)/4b^2, \quad \operatorname{var} \hat{\alpha} = \left[\left(1 + \frac{\cos^2\alpha}{b^2}\right) \operatorname{var} \hat{a}_1 + \frac{2\cos\alpha}{b} \operatorname{cov}(\hat{a}_1, \hat{a}_2)\right] \Big/ 4b^2 \sin^2\alpha$$

$$\operatorname{cov}(\hat{b}, \hat{\alpha}) = \left[\frac{\cos\alpha}{2b^2} \operatorname{var} \hat{a}_1 + \frac{\operatorname{cov}(\hat{a}_1, \hat{a}_2)}{2b}\right] \Big/ 2b \sin\alpha \tag{18}$$

3.2.3.1 **Example 31.** *Determination and estimation of the recurrence between the daily averages of the atmospheric pressure at Uccle (two years of observations)*

We have calculated the excess x_i over 7000 of the daily average of the atmospheric pressure at Uccle, in tenths of mm of standard mercury. The first 20 and the last 10 values of the series thus obtained are shown in Table 31.

From this series, we obtain the value of the empirical variance,

$$\bar{v}_0 = [(531)^2 + (576)^2 + \cdots + (305)^2 + (351)^2 - (531 + 576 + \cdots + 305 + 351)^2/730]/729$$
$$= 4891{,}4, \tag{1}$$

the value of the empirical autocovariance of order 1,

$$\bar{v}_1 = [531 \times 576 + 576 \times 610 + \cdots + 333 \times 305 + 305 \times 351$$
$$-(531 + 576 + \cdots + 333 + 305)(576 + 610 + \cdots + 305 + 351)/729]/728 = 3978 \tag{2}$$

as well as the other values of the autocovariances $\bar{v}_i$ shown in Table 31.

We find :

$$X_1 = (730 - 2) \times (3978/4891{,}4)^2 = 481$$

$$X_2 = X_1 + (730 - 3)(2650/4891{,}4)^2 = 695 \qquad (3)$$

together with the values of X_i which are also indicated in Table 31.

It appears that the hypothesis of a simple random series can be rejected, and that the alternative of a recurrent series can be considered.

For fitting a recurrence of order 1, we first have :

$$(u'u)_{11} = n - 1 = 730 - 1 = 729$$

$$(u'u)_{12} = \Sigma x_i = 531 + 576 + \cdots + 333 + 305 = 3{,}81442.10^5$$

$$(u'u)_{22} = \Sigma x_i^2 = (531)^2 + (576)^2 + \cdots + (333)^2 + (305)^2 = 2{,}0312198.10^8$$

$$\beta_1 = \Sigma x_{i+1} = 576 + 610 + \cdots + 305 + 351 = 3{,}81262.10^5$$

$$\beta_2 = \Sigma x_i x_{i+1} = 531 \times 576 + 576 \times 610 + \cdots + 333 \times 305 + 305 \times 351 = 2{,}023874.10^8 \qquad (4)$$

TABLE 31. – Daily averages ($7000 + x_i$) of the atmospheric pressure at Uccle in 0,1 mm of standard mercury (two years of observations). Empirical autocovariances $\bar{v}_i$ of order i and statistics X_i. Residuals $\hat{e}_i^*$ and $\hat{e}_i^{**}$ of the recurrences of order 1 and of order 2 fitted to the series

i	x_i	$\bar{v}_i$	X_i	$\hat{e}_i^*$	$\hat{e}_i^{**}$
1	531	3978	481	—	—
2	576	2650	695	46,7	—
3	610	1680	780	43,8	31,5
4	644	1024	812	50,0	43,8
5	632	517	820	10,1	6,3
6	578	195	821	−34,1	−22,6
7	557	76	822	−10,8	11,8
8	527	30	822	−23,6	−14,0
9	523	13	822	−3,1	7,7
10	539	16	822	16,2	17,5
11	561	−21	822	25,1	20,5
12	616	−151	822	62,0	56,9
13	665	−281	825	66,0	52,9
14	659	−383	829	19,9	12,2
15	604	−435	835	−30,2	−18,9
16	552	−464	841	−37,1	−12,4
17	571	−640	853	24,4	44,6
18	574	−767	871	11,9	8,4
19	505	−719	886	−59,6	−57,2
20	517	−644	898	8,9	31,9
.					
.					
.					
721	655	2923	3056	75,7	75,4
722	678	1377	3056	47,1	33,8
723	655	−1044	3057	5,3	7,6
724	593	−3639	3059	−37,9	−21,0
725	559	−4885	3063	−21,1	5,3
726	565	−2647	3064	12,7	27,0
727	474	288	3064	−83,2	−82,5
728	333	1035	3064	−149,7	−121,1
729	305	—		−62,2	−25,5
730	351	—		6,7	1,8

from which we deduce the estimates

$$\hat{a}_0 = 94{,}509 \quad \text{and} \quad \hat{a}_1 = 0{,}818906 \tag{5}$$

by solving the system of equations represented by the relation $u'u\hat{\theta} = \beta$.

With $x_1 = 531$ and $x_2 = 576$, we then obtain

$$\hat{e}_2^* = 576 - 94{,}509 - 0{,}818906 \times 531 = 46{,}7 \tag{6}$$

together with the other values of $\hat{e}_i^*$ given in Table 31 for which, with 3.2.2(13), we find the empirical variance

$$\begin{aligned} \bar{v}_0(e^*) &= [(46{,}7)^2 + (43{,}8)^2 + \cdots + (-62{,}2)^2 + (6{,}7)^2]/(730 - 2 \times 1 - 1) \\ &= 1645{,}15 \end{aligned} \tag{7}$$

and with 3.2.2(14), the empirical autocovariance of order 1

$$\begin{aligned} v_1(e^*) &= [46{,}7 \times 43{,}8 + 43{,}8 \times 50{,}0 + \cdots + (-62{,}2) \times 6{,}7 \\ &\quad -(46{,}7 + 43{,}8 + \cdots - 62{,}2)(43{,}8 + 50{,}0 + \cdots + 6{,}7)/(730 - 1 - 1)]/(730 - 2 \times 1 - 2) \\ &= 466{,}6 \end{aligned} \tag{8}$$

For the statistic 3.2.2(15) we get the value

$$X_1 = 726 \times [466{,}6/1645{,}15]^2 = 58{,}4 \tag{9}$$

which is much larger than 3,84, the critical value at the 0,05 level of the χ^2 variate with one degree of freedom. The hypothesis of a recurrence of order 1 must therefore be rejected.

In fitting a recurrence of order 2, the quantities

$$\begin{aligned} (u'u)_{11} &= n - 2 = 730 - 2 = 728 \\ (u'u)_{12} &= \Sigma x_i = 531 + 576 + \cdots + 474 + 333 = 3{,}81137.10^5 \\ (u'u)_{13} &= \Sigma x_{i+1} = 576 + 610 + \cdots + 333 + 305 = 3{,}80911.10^5 \\ (u'u)_{22} &= \Sigma x_i^2 = (531)^2 + (576)^2 + \cdots + (474)^2 + (333)^2 = 2{,}0302895.10^8 \\ (u'u)_{23} &= \Sigma x_i x_{i+1} = 531 \times 576 + 576 \times 610 + \cdots + 474 \times 333 + 333 \times 305 = 2{,}0228034.10^8 \\ (u'u)_{33} &= \Sigma x_{i+1}^2 = (576)^2 + (610)^2 + \cdots + (333)^2 + (305)^2 = 2{,}0284001.10^8 \\ \beta_1 &= \Sigma x_{i+2} = 610 + 644 + \cdots + 305 + 351 = 3{,}80686.10^5 \\ \beta_2 &= \Sigma x_i x_{i+2} = 531 \times 610 + 576 \times 644 + \cdots + 474 \times 305 + 333 \times 351 = 2{,}0123079.10^8 \\ \beta_3 &= \Sigma x_{i+1} x_{i+2} = 576 \times 610 + 610 \times 644 + \cdots + 333 \times 305 + 305 \times 351 = 2{,}0208154.10^8 \end{aligned} \tag{10}$$

lead to the estimates

$$\hat{a}_0 = 129{,}84027, \quad \hat{a}_1 = -0{,}351247, \quad \hat{a}_2 = 1{,}102713 \tag{11}$$

via the solution of the system of equations represented by the relation $u'u\hat{\theta} = \beta$.

With $x_1 = 531$, $x_2 = 576$ and $x_3 = 610$, we then obtain

$$\hat{e}_3^* = 610 - 129{,}84027 + 0{,}351247 \times 531 - 1{,}102713 \times 576 = 31{,}5 \tag{12}$$

together with the other values of $\hat{e}_i^{**}$ given in Table 31 for which, with 3.2.2(13), we find the empirical variance

$$\bar{v}_0(e^{**}) = [(31,5)^2 + (43,8)^2 + \ldots + (-25,5)^2 + (1,8)^2]/(730 - 2 \times 2 - 1) = 1445,9 \tag{13}$$

and, with 3.2.2(14), the empirical autocovariance of order 1

$$\begin{aligned}\bar{v}_1(e^{**}) &= [31,5 \times 43,8 + 43,8 \times 6,3 + \cdots + (-25,5) \times (1,8) \\ &-(31,5 + 43,8 + \cdots - 25,5)(43,8 + 6,3 + \cdots + 1,8)/(730 - 2 - 1)]/(730 - 2 \times 2 - 2) \\ &= 56,51\end{aligned} \tag{14}$$

The statistic 3.2.2(15) thus takes the value

$$X_1 = 724 \times (56,51/1445,9)^2 = 1,11 \tag{15}$$

which this time is less than 3,84 , the critical value at the 0,05 of the χ^2 variate with one degree of freedom.

We conclude that the series of daily means of the atmospheric pressure at Uccle can be represented by a recurrence of order 2.

We note, however, that this conclusion does not rule out the existence of a seasonal variation of the mean value of these pressures. In reality, statistical analysis of the monthly means shows that this variation exists, but that it does not affect the variance of the residual e^{**} to more than about 3%. At the scale of the daily averages, it can therefore be neglected.

Adoption of the model defined by the estimates (11) implies, on the basis of 3.2.1(12),

$$\hat{m} = 129,84027/(1 + 0,351247 - 1,102713) = 522,4 \tag{16}$$

and with $\hat{\sigma}^2 = \bar{v}_0(e^{**})$, using 3.2.3(6),

$$\text{var } \hat{m} = 1445,9/[(730 - 2) \times (1 + 0,351247 - 1,102713)^2] = 32,154 \tag{17}$$

Similarly, with

$$\begin{aligned}1 - \hat{a}_1 &= 1,351247 \quad , \quad 1 + \hat{a}_1 = 0,648753 \\ 1 - \hat{a}_1 + \hat{a}_2 &= 2,45396 \quad \text{and} \quad 1 - \hat{a}_1 - \hat{a}_2 = 0,248534\end{aligned} \tag{18}$$

relation 3.2.3(11) gives

$$\hat{v}_0 = 1445,9 \times 1,351247/[0,648753 \times 2,45396 \times 0,248534] = 4937,88 \tag{19}$$

whereas 3.2.3(13) and (15) give

$$\begin{aligned}\text{var } \hat{a}_1 &= \text{var } \hat{a}_2 = [1 - (0,351247)^2]/728 = 0,001204 \\ \text{cov } (\hat{a}_1 , \hat{a}_2) &= 1,102713 \times 0,648753/728 = -0,000983 \\ A &= 2,149623 \quad \text{and} \quad B = 3,61609\end{aligned} \tag{20}$$

so that, with 3.2.3(16), we have

$$\frac{\text{var } \hat{v}_0}{v_0^2} = \frac{2}{726} + [(2,149623)^2 + (3,61609)^2] \times 0,001204 - 2 \times 2,149623 \times 3,61609 \times 0,000983 = 0,008780 \tag{21}$$

that is

$$\sigma(\hat{v}_0) = \sqrt{0,008780} \times 4937,88 = 462,7 \tag{22}$$

It is apparent that the accuracy obtained for $\hat{v}_0$ is equivalent to that which would be given by a simple random series with the same variance of which the sample size would be 0,008780/(2/726) = 3,2 times smaller. Furthermore, for the estimate $\hat{m}$, this ratio is slightly smaller and has the value

$$32{,}154/(4937{,}88/729) = 4{,}75$$

Furthermore, using relations 3.2.3(17) and (18), we also have :

$$\hat{b} = \sqrt{0{,}351247} = 0{,}59266, \quad \cos\hat{\alpha} = 1{,}102713/(2 \times 0{,}59266) = 0{,}903307,$$

$$\hat{\alpha} = 0{,}375547 \text{ (en radians) et } \hat{T} = 2\pi/\hat{\alpha} = 16{,}7 \tag{23}$$

with

$$\text{var } \hat{b} = 0{,}000857\,, \quad \text{var } \hat{\alpha} = 0{,}005738 \quad \text{and} \quad \text{cov}(\hat{b}\,, \hat{\alpha}) = 0{,}001760, \tag{24}$$

which gives

$$\text{var } \hat{T} = (T/\alpha)^2 \text{ var } \hat{\alpha} = 11{,}3 \tag{25}$$

We conclude that the fitted recurrence behaves like a damped oscillatory system whose period can be estimated as 16,7 days with a standard error equal to $\sqrt{11{,}3} = 3{,}4$ days.

Since $\bar{v}_0(\epsilon^{**})/\hat{v}_0 = 0{,}29$, we see that the recurrence "explains" 71% of the variance of x.

3.2.4 Extreme value of a recurrent series. Bartels' equivalent number of repetitions

The determination of the distribution of the largest value of a recurrent series is, in principle, a complicated problem because the correlation relating the elements of the series to each other does not allow us to deduce this function directly from that of the elements of the series.

More precisely, if $F(x)$ is the distribution function for the elements x_i, $i = 1, 2, \ldots, n$ of a recurrent series of order equal at least to 1 and if $x_{(n)}$ is the largest value in the series, we have

$$\text{Prob}\,(x_{(n)} < x) = \text{Prob}\,(x_1 < x, x_2 < x, \ldots, x_n < x) \neq [F(x)]^n \tag{1}$$

The Bartels' equivalent number of repetitions defined by the relation

$$\omega_n = n \text{ var } \bar{x}_n / \text{var } x \tag{2}$$

where $\bar{x}_n$ is the arithmetic mean of the series, allows us to circumvent the difficulty.

First we note, in the context of this number, that by virtue of the discussion in 3.2.3 the arithmetic mean $\bar{x}_n$ and the estimate $\hat{m}$ possess, to a first approximation, the same variance. It therefore follows, by virtue of the value of var $\hat{m}$ given in 3.2.3(6) and of the form of v_0 given by 3.2.1(16) that the quantity ω_n depends only on the coefficients $a_1, a_2, \ldots, a_k$ of the recurrence.

In particular, for $k = 1$, with 3.2.3(8), we have

$$\omega = \frac{1 + a_1}{1 - a_1} \tag{3}$$

whereas for $k = 2$, with 3.2.3(11), we obtain

$$\omega = \frac{1 + a_1}{1 - a_1} \cdot \frac{1 - a_1 + a_2}{1 - a_1 - a_2} \tag{4}$$

From the definition (2), we obtain

$$\text{var } \bar{x}_n = (\omega \text{ var } x)/n \tag{5}$$

from which if follows formally that, by putting $n = \omega$, we have

$$\text{var } \bar{x}_\omega = \text{var } x \tag{6}$$

This implies that if, in a recurrent series with sample size n, we calculate the non-overlapping means of ω consecutive elements (assuming that ω is an integer), we obtain n/ω means whose variance is equal to that of the elements of the initial series and whose distribution is simple-random, because to a first approximation relation (5) is true for all values of n (var $\bar{x}_{l\omega} = (\text{var } x)/l$ implies successively $v_l = 0$ for $l = 1, 2, \ldots$, i.e. the linear independence of the terms of this series from each other, at least in the case of a joint normal distribution).

Consequently, if we designate as *equivalent sample size* the quantity

$$n' = n/\omega = \text{var } x/\text{var } \bar{x}_n \tag{7}$$

the distribution of the largest value in the series of values of $\bar{x}_\omega$ and, in approximate fashion, that of the largest value of the initial series, will be given by the relation

$$\text{Prob } (x_{(n)} < x) = [F(x)]^{n'} \tag{8}$$

Therefore we deduce that the asymptotic form of the distribution of the extreme values of a recurrent series is also the double exponential distribution. Moreover, we can show that in the vicinity of the median, the distribution of the extreme values can be derived with good accuracy from the initial distribution. In particular, we have

$$E(x_{(n)}) = m + \frac{\sigma(x)\sqrt{6}}{\pi} \cdot \log n' = m + 0{,}7797\,\sigma(x) \log n' \tag{9}$$

where $\sigma(x) = \sqrt{v_0}$, since, if we take the distribution function $F(x)$ to be a double exponential F' with the same mean and the same variance, then the scale parameter of the function F' will be $\sigma(x)\sqrt{6}/\pi$, and furthermore relation (8) gives

$$[\exp(-e^{-u})]^{n'} = \exp[-e^{\log n' - u}] \tag{10}$$

Conversely, when relation (9) is not satisfied, we may generally deduce from this that the recurrent model is not applicable to the series of observations in question.

3.2.4.1 **Example 32.** *The maximum values of the daily average of the atmospheric pressure at Uccle, in April*

We have determined, for 30 consecutive years, the maximum value of the daily average of the atmospheric pressure at Uccle, in 0,1 mm of standard mercury. The excess x_i' of these values over 7000, arranged in increasing order, is indicated in Table 32.

If we adopt for these pressures the recurrent model obtained in 3.2.3.1, we get, using 3.2.4(4),

$$\omega = \frac{1 - 0{,}351247}{1 + 0{,}351247} \cdot \frac{1 + 0{,}351247 + 1{,}102713}{1 + 0{,}351247 - 1{,}102713} = 4{,}74 \tag{1}$$

Moreover, the monthly averages of the atmospheric pressure in April at Uccle obtained over 140 years give

$$\text{var } \bar{x}_{30} = 907{,}8 \tag{2}$$

which, using 3.2.4(2), leads to

$$\text{var } x = 30 \times 907{,}8/4{,}74 = 5745{,}57 \tag{3}$$

that is, $\sigma(x) = 75{,}8$.

With the value $m = 522$ obtained in 3.2.3.1, this allows us to associate with x'_1 the value of the reduced variable

$$(552 - 522)/75{,}8 = 0{,}40 \tag{4}$$

and consequently, the value $F(x'_1) = 0{,}655$ of the distribution function of the standard normal variate.

Moreover, the equivalent sample size n' for the recurrent series of the 30 daily values of a single month are here $30/4{,}74 = 6{,}33$ so that the value of the distribution function of the April maximum corresponding to x'_1 is $[F(x'_1)]^{6{,}33} = (0{,}655)^{6{,}33} = 0{,}07$. Finally, using the method indicated in 2.1.7.3, we find that the limits $F_{0{,}025}$ and $F_{0{,}975}$ of the two-sided confidence interval at the 0,95 level, for the probability associated with the maximum value x'_1 are respectively $F_{0{,}025} = 0{,}001$ and $F_{0{,}975} = 0{,}12$.

TABLE 32. — Maximum values of the daily average of the atmospheric pressure at Uccle in April, in 0,1 mm of standard mercury. Excess x'_i over 7000 for 30 consecutive years, arranged in increasing order. Corresponding values $F(x'_i)$ of the distribution of the pressures and values of $[F(x'_i)]^{6{,}33}$ for the distribution of the largest monthly value. Limits $F_{0{,}025}$ and $F_{0{,}975}$ of the two-sided confidence interval at the 0,95 level of the probabilities associated with the maximum values x'_i.

i	x'_i	$F(x'_i)$	$[F(x'_i)]^{6{,}33}$	$F_{0{,}025}$	$F_{0{,}975}$
1	552	0,655	0,07	0,001	0,12
2	564	709	11	01	17
3	571	742	15	02	22
4	583	788	22	04	27
5	601	851	36	06	31
6	603	858	38	08	35
7	605	862	39	10	39
8	608	871	42	12	42
9	608	871	42	15	46
10	615	891	48	17	49
11	618	898	51	20	53
12	618	898	51	23	56
13	619	900	51	25	59
14	619	900	51	28	63
15	627	915	57	31	66
16	627	915	57	34	69
17	632	927	62	37	72
18	632	927	62	41	75
19	635	932	64	44	77
20	636	933	65	47	80
21	638	937	66	51	83
22	640	941	68	54	85
23	651	955	75	58	88
24	652	956	75	62	90
25	654	959	77	65	92
26	654	959	77	69	94
27	658	963	79	74	96
28	665	971	83	78	98
29	689	986	92	83	99
30	703	992	948	88	999

Proceeding in this way for all the values x_i', the values obtained for $F(x_i')$, $[F(x_i')]^{6,33}$, $F_{0,025}$ and $F_{0,975}$ show (Table 32) that except for disagreement for $i = 5$, 6 and 7, all the probabilities found for x_i' lie within the corresponding confidence intervals. If we take into account that the number ω is in reality here only an estimate, this result can be considered as very satisfactory.

With respect to $E(x_i')$ of 3.2.4(9), we obtain

$$E(x'_i) = 522 + 0{,}7797 \times 75{,}8 \times \log 6{,}33 = 631 \tag{5}$$

Since the values of x_i' in Table 32 give

$$\hat{E}(x_i') = (552 + 564 + \ldots + 703)/30 = 625{,}9$$

and

$$\hat{\text{var}}\, x_i' = [(552 - 625{,}9)^2 + (564 - 625{,}9)^2 + \ldots + (703 - 625{,}9)^2]/29 = 1127 \tag{6}$$

we obtain for the statistic

$$X = (631 - 625{,}9)^2 \times 30/1127 = 0{,}69 \tag{7}$$

a value which is much less than 3,84, the critical value at the 0,05 level of the χ^2 variate with one degree of freedom, and therefore, *a fortiori*, lower than the corresponding critical value of the F variate with $v_1 = 1$ and $v_2 = 30$ degrees of freedom. The agreement is, this time, excellent.

3.2.5 Application to the determination of the periodic deterministic components of a time series

We saw in 3.1.4 that if, for a simple random series of elements x_i, $i = 1, 2, \ldots, n$ with variance σ^2, we calculate the estimate $\hat{c}_1$ of the semi-amplitude of a hypothetical periodic component with given period $T = 2\pi/\alpha$, then the quantity $\hat{c}_1^2/v$, with $v = 2\sigma^2/n$, possesses either exactly or approximately a χ^2 distribution with two degrees of freedom. Furthermore, the mean value and the variance of $\hat{c}_1$ are, to a first approximation,

$$E(\hat{c}_1) = \sqrt{v\sqrt{2}} \quad \text{and} \quad \text{var}\, \hat{c}_1 = v(2 - \sqrt{2}) \tag{1}$$

If we perform this calculation for a continuous series of values of T in the manner indicated in 3.1.5.3, then the corresponding series of values of $\hat{c}_1$ will also vary continuously. Consequently, the theory of simple random series is not able to provide any significance test which allows us to establish from this series of values of $\hat{c}_1$ the possible existence of a periodic deterministic component in the initial series.

Such a test, on the other hand, can be developed on the basis of the properties of recurrent series. In fact, as long as the values of T are regularly spaced and sufficiently close together, the series of values of $\hat{c}_1$ can be likened to a recurrent series and the significance test comes down to comparing the estimates of the mean and the variance of $\hat{c}_1$ deduced from the series with the theoretical values given in (1).

Moreover, we can use the properties of the extreme values of the recurrent series to establish the significance or otherwise of the maximum value of the series of values of $\hat{c}_1$. However, since the quantity $\hat{c}_1$ is a non-negative variable, the most accurate test is obtained by using as a test statistic the probability given by a χ^2 variate with two degrees of freedom to the value of $\hat{c}_1^2/v$ supplied by the largest value of $\hat{c}_1$.

3.2.5.1 **Example 33.** *Determination of any periodic deterministic component in the (unmodified) total annual rainfall series for Brussels-Uccle (137 years)*

The calculation was carried out in the same way for the unmodified series as for the modified series considered in 3.1.5.4. We thus find successively

$$\hat{v}_0 = 13845{,}52 \ , \ \hat{v}_1 = 1263{,}26 \ , \ \hat{v}_2 = 221{,}26 \tag{1}$$

which, with $\bar{v}_2/\bar{v}_1 = 0{,}17515$, gives the equation

$$2\cos^2\alpha - 0{,}17515\cos\alpha - 1 = 0 \tag{2}$$

of which the solution $\cos\hat{\alpha} = 0{,}752249$ gives, in radians,

$$\hat{\alpha} = 0{,}719328 \quad \text{and} \quad \hat{T} = 2\pi/\hat{\alpha} = 8{,}73 \tag{3}$$

Since the values of $\bar{v}_0$, $\bar{v}_1$, $\bar{v}_2$ and $\hat{T}$ are here of the same order of magnitude as those obtained in 3.1.5.4, we next estimated c_1, for the same values of T, that is T = 5,0 (0,1)12,0. In this way we find the series with sample size $n = 71$ given in Table 33.

Fitting a recurrence to this series and calculating the residual $\hat{e}$ next give, for $k = 1$, the statistic 3.2.2(15)

$$X_1 = (71 - 2 - 2) \times [\bar{v}_1(e^*)/\bar{v}_0(e^*)]^2 = 7{,}90 \tag{4}$$

while, for $k = 2$, this statistic becomes

$$X_1 = (71 - 4 - 2) \times [\bar{v}_1(e^{**})/\bar{v}_0(e^{**})]^2 = 0{,}26 \tag{5}$$

The critical value of the χ^2 variate with one degree of freedom is 3,84, so that the fit of a recurrence of order 2 can be accepted.

This fit gives, in particular,

$$\hat{a}_0 = 4{,}55, \quad \hat{a}_1 = -0{,}3686 \text{ et } \hat{a}_2 = 1{,}1313 \tag{6}$$

from which, in the same way as in 3.2.3.1, with $\bar{v}_0(e^{**}) = 23{,}63$, we obtain

$$\hat{m} = 19{,}2, \quad \text{var}\,\hat{m} = 6{,}08, \quad \hat{v}_0(\hat{c}_1) = 86{,}34, \quad \text{var}\,\hat{v}_0(\hat{c}_1) = 720{,}7 \tag{7}$$

In addition, under the hypothesis of a simple random series for the total rainfall the quantity v defined in 3.1.4(19) and (25) becomes, using the value of $\bar{v}_0$ given in (1),

$$v = 2 \times 13845{,}52/137 = 202{,}12 \tag{8}$$

from which we obtain, using 3.1.4(29),

$$E(\hat{c}_1) = c_1' = \sqrt{202{,}12\sqrt{2}} = 16{,}9 \ , \ \text{var}\,\hat{c}_2 = 202{,}12(2 - \sqrt{2}) = 118{,}4 \tag{9}$$

Comparison of the results (7) with the mathematical expectations (9) thus leads to the statistics :

$$X_1(\hat{m}) = (19{,}2 - 16{,}9)^2/6{,}08 = 0{,}87; \quad X_2[\hat{v}_0(\hat{c}_1)] = (118{,}4 - 86{,}3)^2/720{,}7 = 1{,}43 \tag{10}$$

whose values are both less than 3,84 , the critical value at the 0,05 level of the χ^2 variate with one degree of freedom. Moreover, their sum 0,87 + 1,43 = 2,30 is less than the critical value 5,99 at the same level of the χ^2 variate with two degrees of freedom.

The series of values of $\hat{c}_1$ in Table 33 is thus compatible with the hypothesis of a simple random series for the total rainfall and consequently with the absence of a periodic deterministic component with period lying within the interval considered.

TABLE 33. – (Unmodified) total annual rainfall series for Brussels-Uccle (137 years). Estimates $\hat{c}_1$ of the semi-amplitude of the periodic deterministic component for the values of the period T=5,0(0,1)12,0

T	$\hat{c}_1$	T	$\hat{c}_1$	T	$\hat{c}_1$
5,0	35,4	7,4	13,7	9,8	13,0
,1	19,4	,5	14,7	,9	9,3
,2	27,7	,6	14,6	10,0	6,7
,3	26,1	,7	7,9	,1	6,2
,4	8,4	,8	4,5	,2	7,2
,5	21,0	,9	16,8	,3	8,4
,6	31,7	8,0	26,4	,4	9,2
,7	29,4	,1	31,2	,5	9,4
,8	27,7	,2	29,9	,6	9,1
,9	21,4	,3	23,2	,7	9,0
6,0	9,5	,4	14,2	,8	9,3
,1	7,4	,5	7,3	,9	10,0
,2	7,2	,6	11,0	11,0	11,0
,3	6,8	,7	18,2	,1	11,9
,4	8,3	,8	23,3	,2	12,8
,5	10,7	,9	26,1	,3	13,8
,6	19,8	9,0	27,1	,4	15,1
,7	26,6	,1	27,3	,5	17,0
,8	27,9	,2	26,9	,6	19,7
,9	29,7	,3	25,9	,7	23,0
7,0	35,6	,4	24,4	,8	26,5
,1	38,7	,5	22,6	,9	29,9
,2	33,1	,6	20,1	12,0	32,8
,3	21,7	,7	16,9		

Also using 3.2.4(4), we have for the recurrent series

$$\omega = \frac{1 - 0{,}3686}{1 + 0{,}3686} \cdot \frac{1 + 0{,}3686 + 1{,}1313}{1 + 0{,}3686 - 1{,}1313} = 4{,}86 \tag{11}$$

i.e.,

$$n' = 71/4{,}86 = 14{,}61 \tag{12}$$

It follows that the probability F_c associated with the maximum value of the series is given, for the 0,05 level, by the relation

$$(F_c)^{14{,}61} = 0{,}95 \tag{13}$$

which gives $F_c = 0{,}9965$.

The value of the χ^2 variate with two degrees of freedom corresponding to the probability $F_c = 0{,}9965$ is 11,6 , so that the critical value of the maximum value of $\hat{c}_1$ reduces, using (8) in the relation

$$\hat{c}_1^2 = 11{,}6 \times v = 11{,}6 \times 202{,}12 = 2344{,}6 \tag{14}$$

i.e., $\hat{c}_1 = 48{,}4$.

As all the values in Table 33 are less than the latter value, we again arrive at the same conclusion as above.

We should also point out that the same analysis applied to the modified series of total rainfall leads to a series of values of $\hat{c}_1$ which is only very slightly different. Since the maximum value of $\hat{c}_1$, which in this case is 41,75 , is also less than the critical value 48,4 , we see that in the absence of any preliminary information regarding the existence of the periodic deterministic component, this existence cannot be proved if we adopt the 0,05 critical level.

Moreover, as the table of the χ^2 variate with two degrees of freedom associates with : $(41{,}75)^2/202{,}12 = 8{,}62$, the probability $P = 0{,}9852$, and since we have : $(0{,}9852)^{14{,}61} = 0{,}804$, we deduce that the maximum value of $\hat{c}_1$ could be declared significant only by adopting the level 0,196 $\cong$0,2. This would lead to rejecting the null hypothesis when this is true one time in five, and this is certainly too high an error rate for a Type I error.

3.3 PERSISTENT SERIES

If we can assume that the elements of a time series x_i, $i = 1, 2, \ldots, n$ possess the same distribution with variance $\operatorname{var} x = v_0$, then the persistence model constitutes a new alternative to be set up against the hypothesis of a simple random series.

This hypothesis demands that the condition

$$\operatorname{var} \sum_{j=1}^{l} x_{i+j} = l v_0 \tag{1}$$

be satisfied for all values of i and l, since it implies that $v_k = 0$ for all $k > 0$, and the new alternative hypothesis comes down to assuming that the relation

$$\operatorname{var} \sum_{j=1}^{l} x_{i+j} = l \omega_l v_0 \tag{2}$$

defines for $l > 1$ quantities ω_l which are well defined and independent of i, amongst which at least one differs from 1.

Since in the latter case we have between the quantities ω_l and the autocovariances $v_1, v_2, \ldots, v_{l-1}$, the relation

$$\omega_l = 1 + \frac{2}{l}[(l-1)v_1 + (l-2)v_2 + \ldots + v_{l-1}]/v_0 \tag{3}$$

we see that knowledge of the autocovariances v_1, v_2, v_{l-1} leads to the quantities ω_l and conversely, and also that the condition $\omega_l \neq 1$ for at least one value of $l > 1$ is equivalent to the condition $v_l \neq 0$ for at least one value of $l > 0$.

It can be recognized that the quantities ω_l fit in with the definition of the Bartels equivalent number of repetitions introduced in 3.2.4 in the case of a recurrent series. Note, however, that the number ω defined in 3.2.4 is in reality the limiting value towards which the ω_l tend. In particular, for $k = 1$, we have

$$\omega_l = \frac{1 + a_1}{1 - a_1} - \frac{2a_1(1 - a_1^l)}{l(1 - a_1)^2} \tag{4}$$

of which the asymptotic value for large values of l is $\omega = (1 + a_1)/(1 - a_1)$.

In practice, calculation of the empirical values of ω_l is made from the value of the autocovariances, by using relation (3); in this connection it is worth saying something about the method of calculating these autocovariances when the series of observations comprises the daily values for the same month over a series of years. So as to eliminate the effect due to a possible seasonal variation, we first calculate the empirical variances and covariances for the whole set formed by the various days of the month. In particular, if for a given month we represent the overall set of observations of the days j of the years i by the matrix $\| x_{ij} \|$ with $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, n'$, then the empirical variances and covariances are given by the relations

$$\tilde{v}_{jj'} = \left[\sum_{i=1}^{N} x_{ij} x_{ij'} - (\Sigma x_{ij})(\Sigma x_{ij'})/N \right] / (N - 1) \tag{5}$$

where $j, j' = 1, 2, \ldots, n'$. The empirical variance and autocovariances of the overall set are next obtained from the relations

$$\tilde{v}_0 = \Sigma \tilde{v}_{jj}/n', \quad \tilde{v}_l = (\Sigma \tilde{v}_{jj+l})/(n' - l), \quad l = 1, 2, \cdots, n' - 1 \tag{6}$$

Under these conditions, when fitting a recurrence of order k, the system of equations 3.2.2(2) becomes

$$\| \Delta_{k-1} \| . \theta = \beta \tag{7}$$

with

$$\| \Delta_{k-1} \| = \begin{Vmatrix} v'_0 & v'_1 & \cdots & v'_{k-1} \\ v'_1 & & & \\ \vdots & & & \\ v'_{k-1} & v'_{k-1} & \cdots & v'_0 \end{Vmatrix}, \quad \theta = \begin{Vmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{Vmatrix}, \quad \beta = \begin{Vmatrix} v''_k \\ v''_{k-1} \\ \\ v''_1 \end{Vmatrix} \tag{8}$$

where

$$v'_0 = \sum_1^{n'-k} \bar{v}_{jj}/(n'-k), \quad v'_l = \sum_1^{n'-k} \bar{v}_{jj+l}/(n'-k), \quad l = 1, 2, \cdots, k-1 \tag{9}$$

and

$$v''_l = \sum_1^{n'-k} \bar{v}_{j+k-l\, j+k}/(n'-k), \quad l = 1, 2, \cdots, k \tag{10}$$

The solution of the system (7) will be the mean value of the solutions given by the series of observations of each of the N months considered, after eliminating the seasonal variation of the mean.

3.3.1 **Example 34.** *The daily maxima of the air temperature recorded at 8 a.m. at Uccle in July.*

The 8 a.m.. daily maximum air temperature is the temperature read at 8 a.m. on a maximum thermometer, the thermometer being reset after each reading. Since this reading is taken at a time of the day when the air temperature is relatively close to its minimal value, the successive maxima at 8 a.m. belong in principle to distinct 24 hour periods. The sequence of the daily maxima at 8 a.m. is consequently related only to the evolution from one day to the next of the meteorological conditions, and is not perturbed by any repetition effect such as is the case, for example, for the daily maximum at midday.

This being so, we have considered the daily maxima of the air temperature at 8 a.m. at Uccle, recorded in July between 1921 and 1970 and the application of relation 3.3 (5) and (6) led to the empirical values of the autocovariances given in Table 34, where we have $n'=31$ and $N=50$.

Since, with $\bar{v}_0 = 1750$, the statistic 3.2.2(12) becomes

$$X_1 = N(n'-1)\bar{v}_1^2/\bar{v}_0^2 = 50 \times 30\ (1154/1750)^2 = 652{,}3 \tag{1}$$

i.e., a value much greater than 3,84 , the critical value at the 0,05 level of the χ^2 variate with one degree of freedom, then the hypothesis of a simple random series can be rejected. This result is confirmed by the behaviour (Table 34) of the empirical values ω_l^* deduced from the autocovariances by means of relation 3.3(3).

Therefore we can test the fit of a recurrence. Furthermore, negative values of the autocovariances appear for $l \geqslant 15$, so the trial was made straight away for $k=2$.

With

$$v'_0 = 1787{,}8 \ , \ v'_1 = 1165{,}7 \ , \ v''_1 = 1149{,}7 \quad \text{and} \quad v''_2 = \bar{v}_2 = 759{,}75 \tag{2}$$

the system of equations 3.3(7) reduces to

$$a_1 v'_0 + a_2 v'_1 = v''_2$$

$$a_1 v'_1 + a_2 v'_0 = v''_1$$

and gives the solution

$$\hat{a}_1 = 0{,}009838 \quad \text{and} \quad \hat{a}_2 = 0{,}6367 \tag{3}$$

Under the null hypothesis $a_1 = 0$, 3.2.3(13) becomes

$$\text{var } \hat{a}_1 = 1/N(n' - 2) = 1/50 \text{ x } 29 = 0{,}000690 \tag{4}$$

which leads to the statistic

$$\hat{a}_1^2/\text{var } \hat{a}_1 = (0{,}009838)^2/0{,}000690 = 0{,}14 \tag{5}$$

this value being much less than 3,84 , the critical value at the 0,05 level of the χ^2 variate with one degree of freedom. Consequently we can assume $a_1 = 0$ and adopt the hypothesis of a recurrence of first order.

In the latter case, with $v_0' = 1767{,}9$ and $v_1'' = \bar{v}_1 = 1153{,}9$, the system of equations 3.3(7) reduces to

$$a_1 v_0' = v_1'' \tag{6}$$

which gives

$$\hat{a}_1 = 1153{,}9/1767{,}9 = 0{,}6527 \tag{7}$$

and, with 3.2.2(11)

$$\text{var } \hat{a}_1 = [1 - (0{,}6527)^2]/50 \text{ x } 30 = 0{,}000383 \tag{8}$$

i.e., a standard deviation of $\sigma(\hat{a}_1) \cong 0{,}02$.

TABLE 34. – Daily maxima at 8 a.m. of the air temperature in July at Uccle. Values of the quantities $\bar{v}_l, \omega^*_{l+1}, \hat{\omega}_{l+1}, \Delta v_l$ and a^*_1.

l	$\bar{v}_l$	ω^*_{l+1}	$\hat{\omega}_{l+1}$	Δv_l	a^*_1
1	1154	1,66	1,65		0,659
2	760	2,17	2,15	−1	0,659
3	544	2,58	2,54	33	0,673
4	397	2,91	2,85	45	0,690
5	392	3,21	3,09	130	0,741
6	300	3,47	3,29	42	0,745
7	226	3,70	3,45	28	0,746
8	197	3,90	3,58	48	0,761
9	183	4,09	3,69	53	0,778
10	194	4,26	3,78	73	0,803
11	203	4,42	3,86	75	0,806
12	128	4,56	3,93	−6	0,804
13	57	4,70	3,99	−27	0,768
14	29	4,81	4,04	−9	0,746
15	−43	4,91	4,08	−62	
16	−86	4,99	4,12	−58	
17	−88	5,06	4,16	−31	
18	−129	5,11	4,19	−71	
19	−150	5,15	4,22	−65	
20	−172	5,17	4,24	−73	
21	−207	5,18	4,27	−94	
22	−153	5,19	4,29	−16	
23	−121	5,18	4,31	−20	
24	−33	5,18	4,33	47	
25	−43	5,17	4,34	−21	
26	−188	5,16	4,36	−160	
27	−416	5,13	4,37	−292	
28	−428	5,09	4,39	−154	
29	−442	5,03	4,40	−160	
30	−240	4,97	4,40	51	

Using 3.3(4), we find estimates $\hat{\omega}_{l+1}$ of Table 34, to which we have added the value of Δv_l calculated by means of the relation

$$\Delta v_l = \overline{v}_l - a_1 \overline{v}_{l-1} = \overline{v}_l - 0{,}6527\ \overline{v}_{l-1} \tag{9}$$

together with those of a_1^* deduced from the relation

$$(a_1^*)^l = \overline{v}_l / \overline{v}_0 \tag{10}$$

We see that the empirical values ω_{l+1}^* increase more rapidly than the estimated values $\hat{\omega}_{l+1}$, and we can reconcile this with the positive values of the residual Δv for $l = 3$ to 11 and the systematic increase, for the same values of l, of a_1^* up to a value lying far outside the confidence interval at the 0,95 level assigned by the value (8) of var $\hat{a}_1$ to the true value of a_1.

Although the latter observation does not constitute a rigorous significance test, a confirmation of the incompatibility of the model can also be found in the behaviour of the monthly absolute extreme values.

Indeed, with $\omega = (1 + 0{,}6527)/(1 - 0{,}6527) = 4{,}7587$, from which we obtain $n'' = 31/4{,}7587 = 6{,}51$, and $\sigma(x) = \sqrt{\overline{v}_0} = 41{,}8$, relation 3.2.4(9) gives

$$0{,}7797 \times \sigma(x) \times \log n'' = 0{,}7797 \times 41{,}8 \times 1{,}873 = 61{,}05 \tag{12}$$

which, with the general mean over 50 years $\hat{m} = 234$ leads to the mathematical expectations

$$E(TX) = 234 + 61 = 295 \quad \text{and} \quad E(TN) = 234 - 61 = 173 \tag{13}$$

for the absolute extreme values TX and TN of the month (largest and smallest maximum of the month).

Also, the 50 observed values of TX and TN give

$$\overline{TX} = \Sigma TX/50 = 317{,}9\ ,\ \text{vâr}\ \overline{TX} = 17{,}06\ ,\ \overline{TN} = \Sigma TN/50 = 169{,}3\ ,\ \text{vâr}\ \overline{TN} = 7{,}49 \tag{14}$$

and hence the statistics

$$X(\overline{TX}) = (317{,}9 - 295)^2/17{,}06 = 30{,}7 \quad \text{and} \quad X(\overline{TN}) = (169{,}3 - 173)^2/7{,}49 = 1{,}83 \tag{15}$$

whose values demonstrate a good agreement only for TN.

In reality, examination of the series shows that the disagreement for $\overline{TX}$ can be explained by a larger variability of the residual of the recurrence when it is associated with the largest values of the maxima than in the other cases.

If we add to this that the application of the Shapiro-Wilk normality test to the 50 values of TN and TX shows in both cases that there is excellent agreement with the hypothesis of a normal distribution of these values, this agreement excludes any distribution function of the form $[F(x)]^{n''}$, and we must conclude from this that the first-order recurrent model gives an imperfect representation of the time series considered, and that in reality we are dealing with a heterogeneous process.

3.4 MOVING AVERAGES. WEIGHTED AVERAGES. FILTERS

Given a time series x_i, $i = 1, 2, \ldots, n$ for which we assume $\Sigma x_i = 0$, we designate as *moving averages* series any series derived from a relation of the form :

$$y_{i+k-1} = a_1 x_i + a_2 x_{i+1} + \cdots + a_k x_{i+k-1}, \quad i = 1, 2, \cdots, n - k + 1 \tag{1}$$

with $k \geqslant 2$ and where the quantities $a_1, a_2, \ldots, a_k$ are given constants amongst which a_1 and a_k are not equal to zero.

If, furthermore, we have the relation

$$\Sigma a_l = 1 \tag{2}$$

then the moving averages are said to be *weighted*.

Denoting by y and by θ, respectively, the vectors formed by the elements y_{i+k-1} and the elements a_l, and by u, the matrix

$$u = \begin{Vmatrix} x_1 & x_2 & \cdots & x_k \\ x_2 & x_3 & \cdots & x_{k+1} \\ \vdots & & & \\ x_{n-k+1} & & \cdots & x_n \end{Vmatrix} \tag{3}$$

then relation (1) can also be written as

$$y = u\theta \tag{4}$$

and we say that the vector y is the result of applying the filter θ of order k to the series x_i. If we adopt as matrix u the matrix

$$u = \begin{Vmatrix} \frac{x_1}{n-1} & \frac{x_2}{n-1} & \cdots & & \frac{x_n}{n-1} \\ 0 & \frac{x_2}{n-2} & \cdots & & \frac{x_n}{n-2} \\ \vdots & & & & \\ 0 \cdots & & 0 & x_{n-1} & x_n \end{Vmatrix} \tag{5}$$

and as filter θ the vector formed by the elements x_i of the initial series, then the derived series $y = u\theta$ turns out to be the same as the series of the empirical autocovariances, which we have seen plays an important role in periodic or recurrent series.

In general, the derived series obtained by application of a filter has random properties which are different from those of the initial series. In fact, it follows from the relation

$$\operatorname{var} y = \mathrm{E}(yy') = \mathrm{E}(u\theta\theta' u') \tag{6}$$

that the autocovariances of the series y are well defined linear forms of the autocovariances of the initial series.

In particular, for a persistent series with autocovariances

$$v_{i'} = \operatorname{cov}(x_i, x_{i+i'}) \tag{7}$$

we have, for example, for $k = 2$,

$$\begin{aligned} v_{l'}(y) &= \operatorname{cov}(y_{l+1}, y_{l+l'+1}) = \mathrm{E}[(a_1 x_l + a_2 x_{l+1})(a_1 x_{l+l'} + a_2 x_{l+l'+1})] \\ &= (a_1^2 + a_2^2) v_{l'} + a_1 a_2 (v_{l'-1} + v_{l'+1}) \end{aligned} \tag{8}$$

It follows that if the initial series is simple random, then we have, for $k = 2$

$$v_0(y) = (a_1^2 + a_2^2) v_0, \quad v_1(y) = a_1 a_2 v_0 \quad \text{and} \quad v_{i'}(y) = 0, \ \text{for } i' \geqslant 2 \tag{9}$$

while the application of a filter of arbitrary order k gives in the same case

$$v_{i'}(y) = \Sigma a_i a_{l+i'} v_0, \quad i' = 0, 1, 2, \cdots, k-1, \text{and } v_{l'} = 0, \quad i' \geqslant k. \tag{10}$$

Any series of moving averages of a simple random series is, consequently, a persistent series.

Application of a filter to a time series is generally aimed at transforming the initial series in such a way as to retain the properties of the systematic component of the series and to reduce the variance of the random component. Therefore, among the filters of order k with the same mean $a = \Sigma a_l/k$, the one which leads to the smallest variance $v_0(y)$ with a simple random series is the one for which we have : $a_l = a$. In particular, if $\Sigma a_l = 1$, we then obtain

$$a_l = 1/k \tag{11}$$

and relations (10) become

$$v_0(y) = v_0/k \;\; ; \;\; v_{i'}(y) = (k - i')v_0/k^2 \;\; , \;\; i' = 1, 2, \ldots, k-1 \;\; , \;\; \text{and} \;\; v_{i'}(y) = 0, i' \geqslant k. \tag{12}$$

In practice, the use of filters can be justified when we wish to enhance phenomena whose scale differs significantly from the order of the filter. In this connection, in a periodic series the application of a filter of the type (11) brings about a reduction in the amplitude of the periodic component which becomes progressively larger as the order of the filter gets closer to the period. We can therefore also make use of this property to eliminate a particular component of the initial series and thus make the other components stand out better.

Although the use of filters constitutes a simple means of determining, in an initial analysis, the behaviour of a time series, any conclusions drawn on the basis of such an analysis must be considered valueless if they are not confirmed by appropriate tests.

Without such precautions, their use will inevitably lead only to disappointing results, and we can indeed attribute the failures experienced with certain prediction methods based empirically on the use of moving averages to the absence of adequate testing.

3.4.1 **Example 35.** *Cumulative sums of the rainfall at Uccle over 3, 6 and 12 consecutive months. Application to the case of the droughts of 1921, 1949 and 1864*

The cumulative sums considered are special cases of application of filters of order $k = 3$, 6 and 12, with $a_l = 1$. These provide a simple means of characterizing the evolution of the meteorological series in climatological reports and, in the case of rainfall series, of describing the structure of the succession of above-average and below-average cases. They therefore allow, to some extent, a better understanding of their influence on certain phenomena in agriculture, hydrology and certain other areas.

As an example, we have considered (Table 35) the cumulative sums of the rainfall at Uccle over 3, 6 and 12 consecutive months between 1920 and 1922 and, for comparison, we have also included the cumulative sums over 12 consecutive months from 1948 to 1950 and from 1863 to 1865. The cumulative deficit in the year 1921 is in fact the largest recorded since the start of meteorological observations at Brussels in 1833; the next largest deficit is that of 1949, while that of 1864 is the highest recorded in the last century (from 1833).

The normal for the annual amount of precipitation at Uccle is 780 mm, i.e. for the three months a mean of 195 mm and for the six months a mean of 390 mm; the data in Table 35 thus show that the drought of 1921 corresponded to a continuous deficit in the moving cumulative totals over three months from October 1920 to December 1921. For the moving totals over six months, the period of deficit extends from October 1920 to March 1922, whereas for those over 12 months, this starts only in December 1920 but extends until July 1922. We see that the deficit reached its highest point for the three months totals in July and August 1921, for the six months totals in October 1921, and for the twelve months totals in September 1921.

So, depending on the interval over which the cumulative totals are calculated, the least favourable periods range from May to August 1921, from May to October 1921, and from October 1920 to September 1921.

The evolution of the cumulative totals also shows that the drought of 1921 followed a period of heavy precipitation up to the middle of 1920 and that the return of such amounts of precipitation occurred from the beginning of 1922.

For the year 1949, the behaviour of the moving cumulative totals over 12 months was similar, with a minimum at the start of the autumn followed by a progressive recovery, returning in 1950 to significantly above-average rainfall observations. In contrast, unlike the above-average rainfall which was present over almost the whole of the year 1920, below-average values were already becoming apparent during the first half of 1948.

Finally, the behaviour of the moving cumulative totals over 12 months from 1863 to 1865 shows a quite different character. All the 12 months sums are below average, and the minimum is reached in December 1864. Although the deficit did not reach the proportions of that of 1921, it is nonetheless distinguished from the latter by a greater degree of persistence.

Thus we see, on the basis of the cumulative sums, that the three droughts had their own distinct characteristics which distinguish each one quite clearly from the others.

TABLE 35. – Total rainfall in mm at Uccle over one month (1) and accumulated over three (3), six (6) or twelve (12) consecutive months (the cumulative sums are attributed to the last month of the period)

Month	Year	(1)	(3)	(6)	(12)	Year	(12)	Year	(12)
1	1920	119	389	589	1021	1948	738	1863	700
2		31	309	585	998		760		684
3		28	178	534	953		689		672
4		103	162	551	970		684		638
5		76	207	516	1020		702		656
6		48	227	405	1020		756		644
7		125	249	411	1000		792		585
8		52	225	432	1017		832		609
9		50	227	454	988		836		668
10		15	117	366	917		858		618
11		29	94	319	835		822		621
12		61	105	332	737		791		631
1	1921	77	167	284	695	1949	694	1864	627
2		10	148	242	674		674		641
3		39	126	231	685		683		657
4		33	82	249	615		672		637
5		18	90	238	557		674		611
6		17	68	194	526		598		601
7		6	41	123	407		488		582
8		20	43	133	375		487		578
9		26	52	120	351		485		561
10		23	69	110	359		499		574
11		81	130	173	411		510		559
12		56	160	212	406		520		498
1	1922	93	230	299	422	1950	552	1865	520
2		51	200	330	463		635		564
3		57	201	361	481		609		557
4		86	194	424	534		641		558
5		12	155	355	528		680		585
6		120	218	419	631		724		552
7		113	243	437	736		812		670
8		93	324	479	809		834		686
9		96	300	518	879		890		623
10		50	239	482	906		860		684
11		80	226	550	905		928		673
12		60	190	490	909		952		674

3.5 CONCLUSIONS

To summarize, we can see on the basis of the few examples we have considered that analysis of a persistent series plays an overwhelmingly important role in climatology. We must, in fact, steer towards the use of such models if we wish to provide as complete a statistical representation as possible of the phenomena, i.e. a representation which gives a description of their behaviour on the smallest as well as on the largest time-scales.

The importance of persistence models had in fact already become apparent during the investigation of the distributions of discrete variables and we may conclude that the range of available models is sufficiently wide to represent the series of frequencies in climatology.

The situation is clearly less favourable in the case of series of observations of a continuous variable, and the example of the daily maximum of the air temperature at Uccle in July has shown that the problem of the statistical representation of a meteorological phenomenon can extend beyond the domain of recurrent series and even that of persistent series.

We can thus conclude that the answer to this question lies in the design of more general models in which the mean and the variance would not, for example, be stationary, even though the increase in the number of parameters which results inevitably leads to a lower precision in the estimation of the parameters.

Nonetheless, although it is desirable purely from the point of view of attempting to represent the phenomena by the most reliable model, it would appear reasonable in practice always to link the adoption of a particular model to the accuracy of the results necessary in the applications in question.

3.6 REFERENCE WORKS CONSULTED

The estimation methods presented both in the case of periodic random series and in that of recurrent random series are direct applications of the method of least squares.

Although consideration of recurrent random models goes back to Yule [49] and to Walker [50] and although it is possible to find in [8] a collection of properties of these series, a fresh description of the theory of these series was still necessary in order to solve the problems of estimation. We should also emphasize the importance taken by the series of the autocovariances because of the very varied behaviour which its study reveals, depending on whether it is calculated from a periodic random series or a recurrent random series.

Note that Box and Jenkins [55] have generalized the autoregressive series by allowing the residuals to be correlated by defining them as a finite moving average (see Kendall [56]). As may be expected, in this case, the problem of estimation has no longer a solution by the method of least squares, but is rather involved, using preliminary estimates for finding final results with the assistance of a computer routine. Moreover, when testing estimated residuals in the case of autoregressive series, is should be kept in mind that the variance of the distribution of the statistic 3.2.2(15) is modified owing to the fact that the error of estimation of the residual is a moving average of the elements of the autoregressive series. However, for large samples, the bias being small, the test remains valid in a first approximation.

We would like the fundamentals on which the use of the Bartels equivalent number (to determine the distribution of the extreme values of a recurrent series) is based to be more rigorous. Nonethless, its use fills a gap in a way which appears acceptable, and although progress still remains to be made in this area we believe that our method should prove useful.

The concept of persistent series generalizes that introduced by Bartels in [51], together with the concept of the equivalent number of repetitions, which was also taken up by vander Bijl in [52]; the concept of a filter conforms to the definition given, for example, in [8].

Also, [53] gives some indications on the efficiency of a simplified method for estimating the Bartels equivalent number in the case of a recurrent random series of first order (Markov series) and [54] mentions failures encountered when filters are used as a prediction method.

The source of the data used in examples 29, 30, 33 and 35 has been stated earlier; for examples 31, 32 and 34 we referred to the climatological archives for Uccle.

4. ACKNOWLEDGEMENTS

I should like to thank the members of the working group (Messrs. K. Cehak (Austria), E. Chryssaidos (Greece), J.M. Craddock (England), J. Dettwiller (France) and H.C.S. Thom (USA)) over whom I was honoured to preside, both for their valued collaboration on the final choice of the material chosen for inclusion in the present work, this choice having been discussed during the meeting of the working group at Brussels in May 1972, and for the suggestions and comments which they were kind enough to communicate after reading the texts which I sent to them.

Thanks are also extended to my colleagues Mesdames Hubrecht and Lietar, recpectively, for the production of the computer programs necessary for certain of the calculations, and for putting together the manuscript of this work.

ANNEX

TABLE A. — K. Pearson's test for normality. Quantiles of the statistic b_2.

n	P = 0,01	0,05	0,95	0,99
7	1,25	1,41	3,55	4,23
8	1,31	1,46	3,70	4,53
9	1,35	1,53	3,86	4,82
10	1,39	1,56	3,95	5,00
12	1,46	1,64	4,05	5,20
15	1,55	1,72	4,13	5,30
20	1,65	1,82	4,17	5,36
25	1,72	1,91	4,16	5,30
30	1,79	1,98	4,11	5,21
35	1,84	2,03	4,10	5,13
40	1,89	2,07	4,06	5,04
45	1,93	2,11	4,00	4,94
50	1,95	2,15	3,99	4,88
75	2,08	2,27	3,87	4,59
100	2,18	2,35	3,77	4,39
125	2,24	2,40	3,71	4,24
150	2,29	2,45	3,65	4,13
200	2,37	2,51	3,57	3,98
250	2,42	2,55	3,52	3,87
300	2,46	2,59	3,47	3,79
350	2,50	2,62	3,44	3,72
400	2,52	2,64	3,41	3,67
450	2,55	2,66	3,49	3,63
500	2,57	2,67	3,37	3,60
550	2,58	2,69	3,35	3,57
600	2,60	2,70	3,34	3,54
650	2,61	2,71	3,33	3,52
700	2,62	2,72	3,31	3,50
750	2,64	2,73	3,30	3,48
800	2,65	2,74	3,29	3,46
850	2,66	2,74	3,28	3,45
900	2,66	2,75	3,28	3,43
950	2,67	2,76	3,27	3,42
1000	2,68	2,76	3,26	3,41

TABLE B. — Joint test on the statistics $\sqrt{b_1}$ and b_2. Critical values of Fisher's statistic for the level α.

n	α = 0,05	α = 0,01
20	10,3	16,6
30	10,1	16,3
40	10,0	15,9
50	10,0	15,7
75	9,8	15,4
100	9,7	15,1
200	9,6	14,8
300	9,5	14,6
500	9,5	14,3
1000	9,5	13,8

TABLE C. — Shapiro - Wilk's test for normality. Coefficients a_m for the statistic W.

m	$n = 2$	3	4	5	6	7	8	9	10	
1	0,7071	0,7071	0,6872	0,6646	0,6431	0,6233	0,6052	0,5888	0,5739	
2	—	,0000	,1677	,2413	,2806	,3031	,3164	,3244	,3291	
3	—	—	—	,0000	,0875	,1401	,1743	,1976	,2141	
4	—	—	—	—	—	,0000	,0561	,0947	,1224	
5	—	—	—	—	—	—	—	,0000	,0399	

	11	12	13	14	15	16	17	18	19	20
1	0,5601	0,5475	0,5359	0,5251	0,5150	0,5056	0,4968	0,4886	0,4808	0,4734
2	,3315	,3325	,3325	,3318	,3306	,3290	,3273	,3253	,3232	,3211
3	,2260	,2347	,2412	,2460	,2495	,2521	,2540	,2553	,2561	,2565
4	,1429	,1586	,1707	,1802	,1878	,1939	,1988	,2027	,2059	,2085
5	,0695	,0922	,1099	,1240	,1353	,1447	,1524	,1587	,1641	,1686
6	0,0000	0,0303	0,0539	0,0727	0,0880	0,1005	0,1109	0,1197	0,1271	0,1334
7	—	—	,0000	,0240	,0433	,0593	,0725	,0837	,0932	,1013
8	—	—	—	—	,0000	,0196	,0359	,0496	,0612	,0711
9	—	—	—	—	—	—	,0000	,0163	,0303	,0422
10	—	—	—	—	—	—	—	—	,0000	,0140

	21	22	23	24	25	26	27	28	29	30
1	0,4643	0,4590	0,4542	0,4493	0,4450	0,4407	0,4366	0,4328	0,4291	0,4254
2	,3185	,3156	,3126	,3098	,3069	,3043	,3018	,2992	,2968	,2944
3	,2578	,2571	,2563	2554	,2543	2533	,2522	2510	,2499	2487
4	,2119	,2131	,2139	,2145	,2148	,2151	,2152	,2151	,2150	,2148
5	,1736	,1764	,1787	,1807	,1822	,1836	,1848	,1857	,1864	,1870
6	0,1399	0,1443	0,1480	0,1512	0,1539	0,1563	0,1584	0,1601	0,1616	0,1630
7	,1092	,1150	,1201	,1245	,1283	,1316	,1346	,1372	,1395	,1415
8	,0804	,0878	,0941	,0997	,1046	,1089	,1128	,1162	,1192	,1219
9	,0530	,0618	,0696	,0764	,0823	,0876	,0923	,0965	,1002	,1036
10	,0263	,0368	,0459	,0539	,0610	,0672	,0728	,0778	,0822	,0862
11	0,0000	0,0122	0,0228	0,0321	0,0403	0,0476	0,0540	0,0598	0,0650	0,0697
12	—	—	,0000	,0107	,0200	,0284	,0358	,0424	,0483	,0537
13	—	—	—	—	,0000	,0094	,0178	,0253	,0320	,0381
14	—	—	—	—	—	—	,0000	,0084	,0159	,0227
15	—	—	—	—	—	—	—	—	,0000	,0076

	31	32	33	34	35	36	37	38	39	40
1	0,4220	0,4188	0,4156	0,4127	0,4096	0,4068	0,4040	0,4015	0,3989	0,3964
2	,2921	,2898	,2876	,2854	,2834	,2813	,2794	,2774	,2755	,2737
3	,2475	,2463	,2451	,2439	,2427	2415	,2403	,2391	,2380	,2368
4	,2145	,2141	,2137	,2132	,2127	,2121	,2116	,2110	,2104	,2098
5	,1874	,1878	,1880	,1882	,1883	,1883	,1883	,1881	,1880	,1878
6	0,1641	0,1651	0,1660	0,1667	0,1673	0,1678	0,1683	0,1686	0,1689	0,1691
7	,1433	,1449	,1463	,1475	,1487	,1496	,1505	,1513	,1520	,1526
8	,1243	,1265	,1284	,1301	,1317	,1331	,1344	,1356	,1366	,1376
9	,1066	,1093	,1118	,1140	,1160	,1179	,1196	,1211	,1225	,1237
10	,0899	,0931	,0961	,0988	,1013	,1036	,1056	,1075	,1092	,1108

TABLE C. — (Continued)

m	$n = 31$	32	33	34	35	36	37	38	39	40
11	0,0739	0,0777	0,0812	0,0844	0,0873	0,0900	0,0924	0,0947	0,0967	0,0986
12	,0585	,0629	,0669	,0706	,0739	,0770	,0798	,0824	,0848	,0870
13	,0435	,0485	,0530	,0572	,0610	,0645	,0677	,0706	,0733	,0759
14	,0289	,0344	,0395	,0441	,0484	,0523	,0559	,0592	,0622	,0651
15	,0144	,0206	,0262	,0314	,0361	,0404	,0444	,0481	,0515	,0546
16	0,0000	0,0068	0,0131	0,0187	0,0239	0,0287	0,0331	0,0372	0,0409	0,0444
17	—	—	,0000	,0062	,0119	,0172	,0220	,0264	,0305	,0343
18	—	—	—	—	,0000	,0057	,0110	,0158	,0203	,0244
19	—	—	—	—	—	—	,0000	,0053	,0101	,0146
20	—	—	—	—	—	—	—	—	,0000	,0049

	41	42	43	44	45	46	47	48	49	50
1	0,3940	0,3917	0,3894	0,3872	0,3850	0,3830	0,3808	0,3789	0,3770	0,3751
2	,2719	,2701	,2684	,2667	,2651	,2635	,2620	,2604	,2589	,2574
3	,2357	,2345	,2334	,2323	,2313	,2302	,2291	,2281	,2271	,2260
4	,2091	,2085	,2078	,2072	,2065	,2058	,2052	,2045	,2038	,2032
5	,1876	,1874	,1871	,1868	,1865	,1862	,1859	,1855	,1851	,1847
6	0,1693	0,1694	0,1695	0,1695	0,1695	0,1695	0,1695	0,1693	0,1692	0,1691
7	,1531	,1535	,1539	,1542	,1545	,1548	,1550	,1551	,1553	,1554
8	,1384	,1392	,1398	,1405	,1410	,1415	,1420	,1423	,1427	,1430
8	,1249	,1259	,1269	,1278	,1286	,1293	,1300	,1306	,1312	,1317
10	,1123	,1136	,1149	,1160	,1170	,1180	,1189	,1197	,1205	,1212
11	0,1004	0,1020	0,1035	0,1049	0,1062	0,1073	0,1085	0,1095	0,1105	0,1113
12	,0891	,0909	,0927	,0943	,0959	,0972	,0986	,0998	,1010	,1020
13	,0782	,0804	,0824	,0842	,0860	,0876	,0892	,0906	,0919	,0932
14	,0677	,0701	,0724	,0745	,0765	,0783	,0801	,0817	,0832	,0846
15	,0575	,0602	,0628	,0651	,0673	,0694	,0713	,0731	,0748	,0764
16	0,0476	0,0506	0,0534	0,0560	0,0584	0,0607	0,0628	0,0648	0,0667	0,0685
17	,0379	,0411	,0442	,0471	,0497	,0522	,0546	,0568	,0588	,0608
18	,0283	,0318	,0352	,0383	,0412	,0439	,0465	,0489	,0511	,0532
19	,0188	,0227	,0263	,0296	,0328	,0357	,0385	,0411	,0436	,0459
20	,0094	,0136	,0175	,0211	,0245	,0277	0,307	,0335	,0361	,0386
21	0,0000	0,0045	0,0087	0,0126	0,0163	0,0197	0,0229	0,0259	0,0288	0,0314
22	—	—	,0000	,0042	,0081	,0118	,0153	,0185	,0215	,0244
23	—	—	—	—	,0000	,0039	,0076	,0111	,0143	,0174
24	—	—	—	—	—	—	,0000	,0037	,0071	,0104
25	—	—	—	—	—	—	—	—	,0000	,0035

TABLE D. — Shapiro-Wilk's test for normality. Quantiles of the statistic W.

n	P = 0,01	0,02	0,05	0,10	0,50	0,90	0,95	0,98	0,99
3	0,753	0,756	0,767	0,789	0,959	0,998	0,999	1,000	1,000
4	,687	,707	,748	,792	,935	,987	,992	,996	,997
5	,686	,715	,762	,806	,927	,979	,986	,991	,993
6	0,713	0,743	0,788	0,826	0,927	0,974	0,981	0,986	0,989
7	,730	,760	,803	,838	,928	,972	,979	,985	,988
8	,749	,778	,818	,851	,932	,972	,978	,984	,987
9	,764	,791	,829	,859	,935	,972	,978	,984	,986
10	,781	,806	,842	,869	,938	,972	,978	,983	,986
11	0,792	0,817	0,850	,876	0,940	0,973	0,979	0,984	0,986
12	,805	,828	,859	,883	,943	,973	,979	,984	,986
13	,814	,837	,866	,889	,945	,974	,979	,984	,986
14	,825	,846	,874	,895	,947	,975	,980	,984	,986
15	,835	,855	,881	,901	,950	,975	,980	,984	,987
16	0,844	0,863	0,887	0,906	0,952	0,976	0,981	0,985	0,987
17	,851	,869	,892	,910	,954	,977	,981	,985	,987
18	,858	,874	,897	,914	,956	,978	,982	,986	,988
19	,863	,879	,901	,917	,957	,978	,982	,986	,988
20	,868	,884	,905	,920	,959	,979	,983	,986	,988
21	0,873	0,888	0,908	0,923	0,960	0,980	0,983	0,987	0,989
22	,878	,892	,911	,926	,961	,980	,984	,987	,989
23	,881	,895	,914	,928	,962	,981	,984	,987	,989
24	,884	,898	,916	,930	,963	,981	,984	,987	,989
25	,888	,901	,918	,931	,964	,981	,985	,988	,989
26	0,891	0,904	0,920	0,933	0,965	0,982	0,985	0,988	0,989
27	,894	,906	,923	,935	,965	,982	,985	,988	,990
28	,896	,908	,924	,936	,966	,982	,985	,988	,990
29	,898	,910	,926	,937	,966	,982	,985	,988	,990
30	,900	,912	,927	,939	,967	,983	,985	,988	,990
31	0,902	0,914	,929	0,940	0,967	0,983	0,986	0,988	0,990
32	,904	,915	,930	,941	,968	,983	,986	,988	,990
33	,906	,917	,931	,942	,968	,983	,986	,989	,990
34	,908	,919	,933	,943	,969	,983	,986	,989	,990
35	,910	,920	,934	,944	,969	,984	,986	,989	,990
36	0,912	0,922	0,935	0,945	0,970	0,984	0,986	0,989	0,990
37	,914	,924	,936	,946	,970	,984	,987	,989	,990
38	,916	,925	,938	,947	,971	,984	,987	,989	,990
39	,917	,927	,939	,948	,971	,984	,987	,989	,991
40	,919	,928	,940	,949	,972	,985	,987	,989	,991
41	0,920	0,929	0,941	0,950	0,972	0,985	0,987	0,989	0,991
42	,922	,930	,942	,951	,972	,985	,987	,989	,991
43	,923	,932	,943	,951	,973	,985	,987	,990	,991
44	,924	,933	,944	,952	,973	,985	,987	,990	,991
45	,926	,934	,945	,953	,973	,985	,988	,990	,991
46	0,927	0,935	0,945	0,953	0,974	0,985	0,988	0,990	0,991
47	,928	,936	,946	,954	,974	,985	,988	,990	,991
48	,929	,937	,947	,954	,974	,985	,988	,990	,991
49	,929	,937	,947	,955	,974	,985	,988	,990	,991
50	,930	,938	,947	,955	,974	,985	,988	,990	,991

TABLE E. – D'Agostino's test for normality. Quantiles of the statistic Y.

n	P = 0,005	0,010	0,025	0,05	0,10	0,90	0,95	0,975	0,990	0,995
10	−4,66	−4,06	−3,25	−2,62	−1,99	0,149	0,235	0,299	0,356	0,385
12	−4,63	−4,02	−3,20	−2,58	−1,94	0,237	0,329	0,381	0,440	0,479
14	−4,57	−3,97	−3,16	−2,53	−1,90	0,308	0,399	0,460	0,515	0,555
16	−4,52	−3,92	−3,12	−2,50	−1,87	0,367	0,459	0,526	0,587	0,613
18	−4,47	−3,87	−3,08	−2,47	−1,85	0,417	0,515	0,574	0,636	0,667
20	−4,41	−3,83	−3,04	−2,44	−1,82	0,460	0,565	0.628	0,690	0,720
22	−4,36	−3,78	−3,01	−2,41	−1,81	0,497	0,609	0,677	0,744	0,775
24	−4,32	−3,75	−2,98	−2,39	−1,79	0,530	0,648	0,720	0,783	0,822
26	−4,27	−3,71	−2,96	−2,37	−1,77	0,559	0,682	0,760	0,827	0,867
28	−4,23	−3,68	−2,93	−2,35	−1,76	0,586	0,714	0,797	0,868	0,910
30	−4,19	−3,64	−2,91	−2,33	−1,75	0,610	0,743	0,830	0,906	0,941
32	−4,16	−3,61	−2,88	−2,32	−1,73	0,631	0,770	0,862	0,942	0,983
34	−4,12	−3,59	−2,86	−2,30	−1,72	0,651	0,794	0,891	0,975	1,02
36	−4,09	−3,56	−2,85	−2,29	−1,71	0,669	0,816	0,917	1,00	1,05
38	−4,06	−3,54	−2,83	−2,28	−1,70	0,686	0,837	0,941	1,03	1,08
40	−4,03	−3,51	−2,81	−2,26	−1,70	0,702	0,857	0,964	1,06	1,11
42	−4,00	−3,49	−2,80	−2,25	−1,69	0,716	0,875	0,986	1,09	1,14
44	−3,98	−3,47	−2,78	−2,24	−1,68	0,730	0,892	1,01	1,11	1,17
46	−3,95	−3,45	−2,77	−2,23	−1,67	0,742	0,908	1,02	1,13	1,19
48	−3,93	−3,43	−2,75	−2,22	−1,67	0,754	0,923	1,04	1,15	1,22
50	−3,91	−3,41	−2,74	−2,21	−1,66	0,765	0,937	1,06	1,18	1,24
60	−3,81	−3,34	−2,68	−2,17	−1,64	0,812	0,997	1,13	1,26	1,34
70	−3,73	−3,27	−2,64	−2,14	−1,61	0,849	1,05	1,19	1,33	1,42
80	−3,67	−3,22	−2,60	−2,11	−1,59	0,878	1,08	1,24	1,39	1,48
90	−3,61	−3,17	−2,57	−2,09	−1,58	0,902	1,12	1,28	1,44	1,54
100	−3,57	−3,14	−2,54	−2,07	−1,57	0,923	1,14	1,31	1,48	1,59
150	−3,409	−3,009	−2,452	−2,004	−1,520	0,990	1,233	1,423	1,623	1,746
200	−3,302	−2,922	−2,391	−1,960	−1,491	1,032	1,290	1,496	1,715	1,853
250	−3,227	−2,861	−2,348	−1,926	−1,471	1,060	1,328	1,545	1,779	1,927
300	−3,172	−2,816	−2,316	−1,906	−1,456	1,080	1,357	1,528	1,826	1,983
350	−3,129	−2,781	−2,291	−1,888	−1,444	1,096	1,379	1,610	1,863	2,026
400	−3,094	−2,753	−2,270	−1,873	−1,434	1,108	1,396	1,633	1,893	2,061
450	−3,064	−2,729	−2,253	−1,861	−1,426	1,119	1,411	1,652	1,918	2,090
500	−3,040	−2,709	−2,239	−1,850	−1,419	1,127	1,423	1,668	1,938	2,114
550	−3,019	−2,691	−2,226	−1,841	−1,413	1,135	1,434	1,682	1,957	2,136
600	−3,000	−2,676	−2,215	−1,833	−1,408	1,141	1,443	1,694	1,972	2,154
650	−2,984	−2,663	−2,206	−1,826	−1,403	1,147	1,451	1,704	1,986	2,171
700	−2,969	−2,651	−2,197	−1,820	−1,399	1,152	1,458	1,714	1,999	2,185
750	−2,956	−2,640	−2,189	−1,814	−1,395	1,157	1,465	1,722	2,010	2,199
800	−2,944	−2,630	−2,182	−1,809	−1,392	1,161	1,471	1,730	2,020	2,211
850	−2,933	−2,621	−2,176	−1,804	−1,389	1,165	1,476	1,737	2,029	2,221
900	−2,923	−2,613	−2,170	−1,800	−1,386	1,168	1,481	1,743	2,037	2,231
950	−2,914	−2,605	−2,164	−1,796	−1,383	1,171	1,485	1,749	2,045	2,241
1000	−2,906	−2,599	−2,159	−1,792	−1,381	1,174	1,489	1,754	2,052	2,249

BIBLIOGRAPHY

A. General works

[1] AITKEN, A. C. *Determinants and matrices*, Oliver and Boyd, Edinburgh and London, 1958.
[2] ANDERSON, T. W. *Introduction to multivariate statistical analysis*, Wiley, New York, 1958.
The statistical analysis of time series, Wiley, New York, 1971.
[3] CRAMER, M. *Mathematical methods of statistics*. Princeton University Press, Princeton, 1958.
[4] GUMBEL, E. J., *Statistics of extremes*. Columbia University Press, New York, 1958.
[5] FISHER, R. A., *Statistical methods for research workers*. Oliver and Boyd, Edinburgh and London, 1963.
[6] HALD, A. *Statistical theory with engineering applications*. Wiley, New York, 1952.
[7] HOEL, P. G. *Introduction to mathematical statistics*, Wiley, 1962.
[8] KENDALL, M. G. and STUART, A. *The advanced theory of statistics*, (in three volumes). Griffin, London, 1963.
[9] MOOD, A. M. *Introduction to the theory of statistics*. McGraw-Hill, New York, 1950.
[10] SIEGEL, S. *Nonparametric statistics for the behavioral sciences*, McGraw-Hill, New York, 1956.
[11] VAN DER WAERDEN, B. L. *Statistique mathématique*. Dunod, Paris, 1967.

For illustration of the applications of statistical methods in climatology and meteorology, we should also mention, in addition to WMO Technical Notes Nos. 71, 79, 81, 98 and 100 :

BROOKS, C. E. P., and CARRUTHERS, N. *Handbook of statistical methods in meteorology*. Met. Office, London, Her Majesty's Stationery Office, 1953.
GRISOLLET, H., GUILMET, B. et ARLÉRY, R. *Climatologie, Méthodes et Pratiques*, Gauthier-Villars, Paris, 1962.
VIALAR, J. *Calcul des probabilités et statistique*, 5 fascicules, Météorologie Nationale, Paris 1956.
PANOFSKY, H. A., and BRIER, G. W. *Some applications of statistics to meteorology, The Pennsylvania State University*, University Park, Pennsylvania, 1958.
ESSENWANGER, O.M. *Applied Statistics in Atmospheric Science, Part A : Frequencies and Curve Fitting*. Elsevier, Amsterdam, 1976.

Further developments in Statistical Climatology may be found in the proceedings of the International Meetings on Statistical Climatology:

Statistical Climatology, Proceedings of the First International Conference on Statistical Climatology, Hachioji, Tokyo (Japan), Elsevier, 1980.
II. International Meeting on Statistical Climatolgoy. Proceedings, Instituto Nacional de Meteorologia e Geofisica, Lisboa, Portugal, 1983.
Third International Conference on Statistical Climatology, June 23-27, 1986, Vienna, Austria. Österreichische Geselschaft für Meteorologie, Wien.
Fourth International Meeting on Statistical Climatology. 27-31 March, 1989 Rotorua, New Zealand, Proceedings, New Zealand Meteorological Service. Wellington, N.Z.

B. Particular references

[12] SNEYERS, R. *Sur la détermination de la stabilité des séries climatologiques. Recherches sur la zone aride, XX. Les changements de climat* (Actes du Colloque de Rome organisé par l'Unesco et l'O.M.M.) Unesco 1963, p. 38.
[13] SNEYERS, R. *Les pointes maximales du vent en Belgique*, Inst. R. Mét. de Belgique, Pub. A, 73, 1971.
[14] SNEYERS, R. *La détermination de la stabilité du climat par l'analyse statistique des séries d'observations; un exemple: la température de l'air et l'eau recueillie à Bruxelles-Uccle de 1833 à 1969*, Annalen der Meteorologie, Neue Folge, 5, 1971, pp. 205-208.
[15] SNEYERS, R. *Du test de validité d'un ajustement basé sur les fonctions de l'ordre des observations*, Inst. R. Mét. de Belgique, Pub. B, 39, 1963.
[16] GURLAND, J. et TRIPATHI, R. C. *A simple approximation for unbiased estimation of the standard deviation*, The American Statistician, 25, 4, oct. 71, pp. 30-32.
[17] THOM, H. C. S. *A note on the gamma distribution*, Month. Weather Rev., 86, 4, 1958, pp. 117-122.
[18] SNEYERS, R. et VAN ISACKER, J. *Sur l'ajustement de la loi de répartition de Fisher-Tippett du Type I au moyen des estimateurs de Kimball*, Rev. b. Stat. Inf. et Rech. Opér., 12, 1, 1972, pp. 36-43.
[19] LIEBLEIN, J. *A New Method of Analyzing Extreme Value Data*, National Advisory Committee for Aeronautics, Tech. Note 3053, Washington D.C., 1954.
[20] DOWNTON, F. *Linear Estimates of parameters in the extreme-value distribution*, Technometrics, 8, 1, 1966, pp. 3-17.
[21] DOWNTON, F. *Linear estimates with polynomial coefficients*. Biometrika, 53, 1 and 2, 1966, pp. 129-141.
[22] PONCELET, L. et MARTIN, H. *Esquisse climatologique de la Belgique*, Inst. R. Mét. de Belgique, Mém., 27, 1947.
[23] SNEYERS, R. *De l'estimation de la fréquence des gelées en Belgique à partir du minimum moyen journalier de la température de l'air*. Proceedings of the XIII International Meeting on Alpine Meteorology, Saint-Vincent (Aosta), 1974, Riv. It. Geof. Sc. Af., 1975.

[24] SNEYERS, R. *Sur les tests de normalité*, Rev. Stat. Appl., 22, 2, 1974, p. 29-36.
[25] MORICE, E. *Tests de normalité d'une distribution observée*, Rev. Stat. Appl., 20, 2, 1972, pp. 5-35.
[26] D'AGOSTINO, R. B. *Transformation to normality of the null distribution of g_1*, Biometrika, 57, 1970, pp. 679-681.
[27] D'AGOSTINO, R. B. et TIETJEN, G. L. *Simulation probability points of b_2 for small samples*, Biometrika, 58, 1971, pp. 669-672.
[28] BOWMAN, K. O. and SHENTON, L. R., « *Omnibus* » *test contours for departures from normality based on $\sqrt{b_1}$, b_2* (à paraître dans Biométrika).
[29] SHAPIRO, S. S. and WILK, M. B. *An analysis of variance test for normality (complete samples)*, Biometrika, 52, 1965, pp. 591-611.
[30] D'AGOSTINO, R. B. *An omnibus test for normality for moderate and large size samples*, Biometrika, 58, 1971, pp. 341-348.
[31] D'AGOSTINO, R. B. *Small sample probability points for the D-test of normality*, Biometrika, 59, 1972, pp. 219-221.
[32] SHAPIRO, S. S., WILK, M. B. and CHEN, H. J. *A comparative study of various tests for normality*, Journ. Am. Stat. Assoc., 63, 1968, pp. 1343-1372.
[33] D'AGOSTINO, R. B. and PEARSON, E. S. *Tests for departure from normality. Empirical results for the distributions of b_2 and $\sqrt{b_1}$*. Biometrika, 60, 3, 1973, pp. 613-622.
[34] SNEYERS, R. *La formation du verglas en Basse et Moyenne Belgique comparée à celle du verglas en Haute Belgique*. XII[e] Congrès international de Météorologie Alpine, Sarajevo, 1972, Zbornik met. i hidrol. radova, Br. 5, 1974, pp. 139-142.
[35] DEFRISE, P., SNEYERS, R., STRUYLAERT, W. et VAN ISACKER, J. *Climatologie aérologique, Station d'Uccle (06447)*. Inst. R. Mét. de Belgique, Pub. A, 77, 1972.
[36] SNEYERS, R. *Sur l'analyse des répartitions statistiques discrètes*, Arch. Met., Geoph. u. Biokl. B, 7, 1, 1955, pp. 117-132.
[37] THOM, H. C. S. *The frequency of hail occurence*, Arch. Met., Geoph. u. Biokl., B, 8, 2, 1957, pp. 185-194.
[38] SNEYERS, R. *La fréquence des orages en Belgique*, Inst. R. Mét. de Belgique, Pub. A, 22, 1961.
[39] FISHER, R. A. *The negative binomial distribution*, Annals Eugenics, 11, 1941, pp. 182-187.
[40] SNEYERS, R. *Un type de distribution discrète tronquée*, Arch. Met., Geoph. u. Biokl., B, 10, 3, 1960, pp. 404-411.
[41] SNEYERS, R. *Notes sur la notion d'indépendance climatologique*, Rev. Stat. Appl., 7, 4, 1969, pp. 45-53.
[42] Institut Royal Météorologique de Belgique, *Lames d'eau mensuelles, saisonnières et annuelles sur les bassins hydrographiques belges (période 1951-1970)*. Inst. R. Mét. de Belgique, Misc., B, 23, 1972.
[43] SNEYERS, R. *La statistique des précipitations à Bruxelles-Uccle*, Inst. R. Mét. de Belgique, Contrib. 94, 1964.
[44] SNEYERS, R. *Application of least squares to the search for periodicities*. Journ. Appl. Meteor., 15, 4, 1976, pp. 387-392.
[45] FISHER, R. A. *Tests of significance in Harmonic Analysis*, Proc. Roy. Soc. London, A, 125, 1929, pp. 54-59.
[46] TIAGO DE OLIVEIRA, J. *Harmonic time-series and least squares*, Rev. b. Stat., Inf. et Rech. Opér., 13, 3, 1973, pp. 3-15.
[47] Institut Royal Météorologique de Belgique, *Bulletin mensuel*, Observations Climatologiques.
[48] Institut Royal Météorologique de Belgique, *La température de l'air à Uccle aux heures paires (période 1931-1960). La moyenne et les extrêmes journaliers de la température de l'air à Uccle (période 1901-1970)*. Misc. B, 27, 1974.
[49] YULE, G. U., *On a method of investigating periodicities in disturbed series with special reference to Wolfer's sunspot numbers*, Phil. Trans. A. 226, 1927, p. 267.
[50] WALKER, G. T. *On periodicity in series of related terms*. Proc. Roy. Soc. A, 131, 1931, p. 518.
[51] BARTELS, J. *Gesetz und Zufall in der Geophysik*, Naturwiss, 31, 1943, pp. 421-435.
[52] VANDER BIJL, W. *Toepassing van statistische methoden in de klimatologie*, Kon. Ned. Met. Inst., Med. en Verhand., 58, 1952.
[53] SNEYERS, R. *Sur l'estimation du nombre équivalent de répétitions*, Rev. Stat. Appl., 19, 2, 1971, pp. 35-47.
[54] LELOUCHIER, P. *De l'efficacité des prévisions à courte et moyenne échéances de la pression atmosphérique basées sur des informations locales de la pression*, Arch. Met. Geoph. u. Biokl., A, 13, 1961, pp. 350-365.
[55] BOX, G.E.P. and JENKINS, G.M., *Time-Series Analysis, Forecasting and Control*. Holden Day, Revised Edition, 1976.
[56] KENDALL, M., *Time-Series*, Charles Griffin, London, 2nd Edition, 1976.

C. Numerical tables

[T1] HALD, A. *Statistical tables and formulas*, Wiley, New York, 1952.
[T2] C.R.C. *Standard mathematical tables*, Chemical Rubber Publishing Co., Cleveland, 1954.
[T3] PEARSON, K. *Tables of the incomplete gamma-function*, Cambridge University Press, Cambridge, 1957.
[T4] United States Department of Commerce, *Probability tables for the analysis of extreme-value data*, Nat. Bur. Stand., Appl, Math. Series, 22, Washington D.C., 1953.
[T5] THOM, H. C. S., *Direct and inverse tables of the gamma distribution*, U.S. Dep. Com., Essa Techn. Rep., EDS-2, Silver Spring, Md, 1968.
[T6] PEARSON, E. S. and HARTLEY, H. O. *Biometrika tables for statisticians*, Vol. I, Cambridge University Press, Cambridge, 1962.

Recent WMO Technical Notes

No. 140 Upper-air sounding studies. Volume I: Studies on radiosonde performance. By A. H. Hooper. Volume II: Manual computation of radiowinds. By R. E. Vockeroth.
No. 141 Utilization of aircraft meteorological reports. By S. Simplicio.
No. 144 Rice and weather. By G. W. Robertson.
No. 145 Economic benefits of climatological services. By R. Berggren.
No. 146 Cost and structure of meteorological services with special reference to the problem of developing countries. Part I, Part II. By E. A. Bernard.
No. 147 Review of present knowledge of plant injury by air pollution. By E. I. Mukammal.
No. 148 Controlled climate and plant research. By R. J. Downs and H. Hellmers.
No. 149 Urban climatology and its relevance to urban design. By T. J. Chandler.
No. 150 Application of building climatology to the problems of housing and building for human settlements. By J. K. Page.
No. 152 Radiation regime of inclined surfaces. By K. Ya. Kondratyev.
No. 154 The scientific planning and organization of precipitation enhancement experiments, with particular attention to agricultural needs. By J. Maybank.
No. 155 Forecasting techniques of clear-air turbulence including that associated with mountain waves. By R. H. Hopkins.
No. 157 Techniques of frost prediction and methods of frost and cold protection. By A. Bagdonas, J. C. Georg and J. F. Gerber.
No. 158 Handbook of meteorological forecasting for soaring flight. Prepared by the Organisation Scientifique et Technique Internationale du Vol à Voile (OSTIV) in collaboration with WMO.
No. 159 Weather and parasitic animal disease. By J. Armour, D. Branagan, J. Donelly, M. A. Gemmell, G. Gettinby, T. E. Gibson, J. N. R. Grainger, M. J. Hope Cawdery, F. Leimbacher, N. D. Levine, J. Nosek, J. G. Ross, M. W. Service, L. P. Smith, D. E. Sonenshine, D. W. Tarry, R. J. Thomas, M. H. Williamson and R. A. Wilson. Edited by T. E. Gibson.
No. 160 Soya bean and weather. By F. S. da Mota.
No. 161 Estudio agroclimatológico de la zona andina. Por M. Frère, J. Q. Rijks y J. Rea.
No. 162 The application of atmospheric electricity concepts and methods to other parts of meteorology. Edited by H. Dolezalek.
No. 164 The economic value of agrometeorological information and advice. By M. H. Omar.
No. 165 The planetary boundary layer. By G. A. McBean, K. Bernhardt, S. Bodin, Z. Lityńska, A. P. Van Ulden and J. C. Wyngaard with contributions from B. R. Kerman, J. D. Reid, Z. Serbjan and J. L. Wolmsley. Edited by G. A. McBean.
No. 166 Quantitative meteorological data from satellites. By Ch. M. Hayden, Lester F. Hubert, E. P. McClain and R. S. Seaman. Edited by J. S. Winston.
No. 167 Meteorological factors affecting the epidemiology of the cotton leaf worm and the pink bollworm. By M. H. Omar.
No. 168 The role of agrometeorology in agricultural development and investment projects. By G. W. Robertson *et al.*
No. 170 Meteorological and hydrological aspects of siting and operation of nuclear power plants. Volume I: Meteorological aspects. Volume II: Hydrological aspects.
No. 172 Meteorological aspects of the utilization of solar radiation as an energy source.
No. 174 The effect of meteorological factors on crop yields and methods of forecasting the yield. Based on a report by the CAgM Working Group on the Effect of Agrometeorological Factors on Crop Yields and Methods of Forecasting the Yield.
No. 175 Meteorological aspects of the utilization of wind as an energy source.
No. 176 Tropospheric chemistry and air pollution. By H. Rodhe, A. Eliassen, I. Isaksen, F. B. Smith and D. M. Whelpdale.
No. 177 Review of atmospheric diffusion models for regulatory applications. By S. R. Hanna.
No. 178 Meteorological aspects of certain processes affecting soil degradation – especially erosion. Report by the CAgM Working Group on Meteorological Factors associated with Certain Aspects of Soil Degradation and Erosion.
No. 179 A study of the agroclimatology of the humid tropics of south-east Asia. By L. R. Oldeman and M. Frère.
No. 180 Weather-based mathematical models for estimating development and ripening of crops. By G. W. Robertson.
No. 181 Use of radar in meteorology. By G. A. Clift, CIMO Rapporteur on Meteorological Radars.
No. 182 The analysis of data collected from international experiments on lucerne. Report of the CAgM Working Group on International Experiments for the Acquisition of Lucerne/Weather Data.
No. 184 Land use and agrosystem management under severe climatic conditions.
No. 185 Meteorological observations using navaid methods. By A. A. Lange.
No. 186 Land management in arid and semi-arid areas.
No. 187 Guidance material for the calculation of climatic parameters used for building purposes. *(In preparation).*
No. 188 Applications of meteorology to atmospheric pollution problems. By D. J. Szepesi, CCl Rapporteur on Atmospheric Pollution.
No. 189 The contribution of satellite data and services to WMO programmes in the next decade.
No. 190 Weather, climate and animal performance. By J. R. Starr.
No. 191 Animal health and production at extremes of weather. *(In preparation).*
No 192 Agrometeorological aspects of operational crop protection.
No 193 Agroclimatology of the sugar-cane crop.