

WORLD METEOROLOGICAL ORGANIZATION

WMO
199
TP 103
C-2

TECHNICAL NOTE No. 81

**SOME METHODS OF
CLIMATOLOGICAL ANALYSIS**

by
H. C. S. Thom

Price: Sw. fr. 6.—

WMO - No. 199. TP. 103

Secretariat of the World Meteorological Organization - Geneva - Switzerland
1966

50-108-103-1966
WMO-199

English version only

Corrigendum

Page 21, line 15, amend the formula to read: $\hat{\beta} = \frac{\bar{x}}{\hat{\gamma}}$ ✓

Page 23, Table, third column, second line from bottom, replace 0.25549 by: 0.22549 ✓

Page 25, last line, amend the formula to read: $\beta^* = \frac{km}{n} \sum_{j=1}^6 b_{.j}^S / 5 + \frac{m'}{n} \sum_{j=1}^3 b_{.j}^S$ ✓

Page 26, Example 3, line 3, amend the formula to read:

$$F(x) = \exp \left[- \left(\frac{x}{\beta_2} \right)^{-\gamma} \right]$$

Page 27, line 5, sixth column, replace 3.9475 by: 3.8475 ✓

Page 27, fourth line from bottom, replace 3.2622 by: 3.7860 ✓

Page 27, second line from bottom, replace β_i^* by: β_1^* ✓

Page 28, line 8, amend the formula to read: $\ln v(F) = \ln \beta_2^* - \frac{\ln \ln \left(\frac{1}{F} \right)}{\gamma^*}$ ✓

Page 29, line 15, amend the formula to read: $P(p_L < \Phi < p_U) = 1 - 2\alpha$ ✓

Page 29, lines 21 and 22, amend the sentence to read:

These replace the formulas and tables for obtaining p_L and p_U . ✓

Page 29, ninth line from bottom, replace $F = P_U$ by: $F = p_U$

Page 31, line 9, amend the formula to read: $\log g_c = x \log \bar{x} - \log x! - 0.434.29 \bar{x}$ ✓

Page 31, lines 11 and 12, amend the sentence to read:

P_c is given in column 8 and is the estimate for each x .

Page 31, lines 13-15, amend the sentence to read:

In column 11 the frequencies g_c and g_o are compared by the χ^2 test whose total is given in the footing of the table.

Page 32, line 2, the division line in the coefficient should be extended over $\Gamma(k)$ to read:

$$f(x) = \frac{\Gamma(x+k)}{\Gamma(x+1) \Gamma(k)} \frac{p^x}{(1+p)^{k+x}}$$

Page 32, line 13, amend to read: $(1 + \frac{1}{p^*}) (k^* + 2) < 20$

Page 33, line 4, amend the formula to read: $g_c(x) = K \frac{p^{*x}}{(1 + p^*)^{k^* + x}}$

Page 33, line 5, amend the formula to read: $K = \frac{\Gamma(k^* + x)}{\Gamma(x+1) \Gamma(k^*)}$

Page 33, Table, column 6, line 1, replace $x \log \left[\frac{p}{p+1} \right]$ by: $x \log \left[\frac{p^*}{p^*+1} \right]$

Page 36, line 1, amend the formula to read: $y = \sum_{i=1}^m x_i$

Page 36, line 3, amend left hand side of the formula to read: $v \left(\sum_{i=1}^m k_i x_i \right)$

Page 37, paragraph 3.2.2, lines 1 and 2, amend the sentence to read:

The regression is a functional relationship between a dependent variable and one or more independent variables.

Page 45, line 6, amend the formula to read: $b_2 = c_{22} Q_{12} + c_{23} Q_{13}$

Page 45, line 8, amend the formula to read: $b_3 = c_{23} Q_{12} + c_{33} Q_{13}$

Page 46, line 5, amend the formula to read: $F(2, n-3) = \frac{(Q_{11} - Q_{1.23})/2}{Q_{1.23}/(n-3)}$

Page 49, fifth line from bottom, amend the formula to read:

$$s(x_{1c}) = \left\{ \frac{Q_{12 \dots k}}{n-k} \left[\frac{1}{n} + \sum_{i=2}^k \frac{1}{n} \sum_{j=2}^k c_{ij} (x_i - \bar{x}_i) (x_j - \bar{x}_j) \right] \right\}^{\frac{1}{2}}$$

Page 49, third line from bottom, amend the formula to read:

$$s(x_{1- \bar{x}}) = \left\{ \frac{Q_{12 \dots k}}{n-k} \left[1 + \frac{1}{n} + \sum_{i=2}^k \frac{1}{n} \sum_{j=2}^k c_{ij} (x_i - \bar{x}_i) (x_j - \bar{x}_j) \right] \right\}^{\frac{1}{2}}$$

Page 51, fourth line from bottom, amend the formula to read: $r^2 = \frac{Q_R}{Q_T} = \frac{15\ 030\ 556}{16\ 365\ 567} = 0.9\ 184$

Page 52, last line, amend the formula to read: $P(2\ 235 < Y < 3\ 805) = 0.90$

CONTENTS

| | <u>Page</u> |
|-----------------------------------------------------------|-------------|
| Foreword | V |
| Summaries (English, French, Russian, Spanish) | VII |
| <u>Chapter One - Climatological series</u> | |
| 1.1 The frequency distribution | 1 |
| 1.2 The cumulative distribution | 3 |
| 1.3 Homogeneity of data series | 4 |
| 1.4 Adjustment of climatological means | 7 |
| 1.4.1 The difference method | 8 |
| 1.4.2 The ratio method | 9 |
| <u>Chapter Two - Estimation of statistical parameters</u> | |
| 2.1 Statistics in general | 12 |
| 2.2 Common statistics of climatological variables | 13 |
| 2.3 Sampling variability of climatological means | 15 |
| <u>Chapter Three - General statistical methods</u> | |
| 3.1 Frequency distributions | 17 |
| 3.1.1 The normal distribution | 17 |
| Example 1 - Normal distribution | 19 |
| 3.1.2 The gamma distribution | 20 |
| Example 2 - Gamma distribution | 21 |
| 3.1.3 The extreme value distributions..... | 22 |
| Example 3 - Extreme value distributions..... | 26 |
| 3.1.4 The binomial distribution | 28 |
| 3.1.5 The Poisson distribution | 30 |
| Example 4 - Poisson distribution | 30 |
| 3.1.6 The negative binomial distribution | 31 |
| Example 5 - Negative binomial distribution | 32 |
| 3.2 Correlation and regression analysis | 34 |
| 3.2.1 Correlation analysis | 34 |
| 3.2.2 Regression analysis | 37 |
| Example 6 - Single regression | 50 |
| References | 53 |

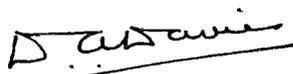
FOREWORD

At its third session (London, December 1960) the Commission for Climatology (CCL) established a Working Group on Statistical Methods in Climatology. The group was requested to review and expand the material on statistical methods in climatology already prepared by a previous working group of the CCL, and to advise on the application of statistical methods to specific climatological problems.

The membership of the group was as follows: H. C. S. Thom (U.S.A), chairman; M. I. Drozdov (U.S.S.R.); G. R. Kendall (Canada); G. O'Mahony (Australia); R. Sneyers (Belgium).

Mr. Thom, the chairman of the group, prepared a draft paper on methods of climatological analysis which, together with the final report of the group, was submitted to the fourth session of CCL in Stockholm (August 1965). The Commission expressed its satisfaction with the report and recommended that the paper be published in the series of WMO Technical Notes.

I am glad to have this opportunity of thanking the members of the working group, as well as others who have contributed to its work, for the time and effort they have devoted to the preparation of this Technical Note.



(D. A. Davies)
Secretary-General

SUMMARY

Modern statistical analysis is the mathematics of climatological analysis, the objective of which is climatological prediction. This Technical Note gives an introduction to the basic principles for the making of such predictions. The methods of analysis presented are applied to a series of illustrative examples.

After defining a climatological series to lay the basis for valid statistical analysis, the frequency distribution (the basic tool of climatological analysis) is discussed. From this the cumulative distribution for obtaining probabilities, which are the climatological predictions, follows naturally.

Since some meteorological records do not form climatological series because of heterogeneities, simple tests for homogeneity are given next. The difference and ratio methods for adjusting, averaging and totalizing variables are discussed, together with applications to actual climatological records. Limitations are also presented on their use and interpretation.

The fundamental problem of estimating statistical parameters is covered as it applies generally, and the ordinary statistical parameters are discussed critically. The approach to normality of several common statistics is also treated, and confidence limits are defined and given for the mean.

Several fundamental frequency distributions are treated, including the normal, gamma, extreme value, binomial, Poisson, and negative binomial distributions. Best estimates for the parameters are given together with complete instructions for fitting them to data. Examples of their application to climatological series are worked out. The application of the binomial distribution in order to obtain confidence limits for the probability estimates obtained from any distribution is also given.

Correlation and regression analysis are discussed in general. The propagation of variance in climatological series is treated, including the effect of covariance. The correct correlation is carefully differentiated from the autocorrelation which can only be used in this application with stationary data sequences. The formulas for propagation of variance are applied to an equation for cooling load in an air-conditioning system.

Regression analysis is discussed in detail, including linear regression forced through the origin. The analysis variance is employed for testing the significance of a relationship as well as to the regression itself. The test for linearity is also given. The standard errors of the regression, as well as the all-important standard error of a forecast, are presented. A complete example of application to a single independent variable is discussed. The simple regression and multiple correlation methods are extended to two independent variables and finally to many independent variables. Finally, there is a list of references to statistical textbooks and papers.

RESUME

L'analyse statistique moderne est l'aspect mathématique de l'analyse climatologique, qui a pour objectif la prévision climatologique. La présente Note technique est une introduction aux principes fondamentaux sur lesquels repose l'élaboration des prévisions climatologiques. Les méthodes d'analyse présentées sont illustrées par une série d'exemples.

Après avoir défini ce qu'est une série climatologique pour jeter les bases d'une analyse statistique valable, l'auteur traite de la distribution de fréquences (élément capital de l'analyse climatologique). De là, il passe naturellement à la distribution de fréquences cumulées pour obtenir des probabilités qui représentent les prévisions climatologiques.

Etant donné que certains relevés météorologiques ne constituent pas des séries climatologiques, en raison de leur hétérogénéité, l'auteur présente ensuite quelques tests simples permettant de déterminer l'homogénéité. Il expose les méthodes des différences et des quotients utilisées pour ajuster les variables, ainsi que pour calculer les moyennes et les sommes de ces variables; il mentionne leurs applications aux relevés climatologiques proprement dits. Il indique également les limites de leur emploi et de leur interprétation.

Après avoir abordé, d'une manière générale, le problème fondamental des paramètres statistiques, l'auteur passe au crible les divers paramètres statistiques ordinaires. Il montre également comment on arrive à déterminer le degré de normalité de divers paramètres statistiques courants; il définit et précise les seuils de confiance de la moyenne.

La Note passe en revue plusieurs distributions de fréquences fondamentales : gamma, valeurs extrêmes, binomiale, Poisson et binomiale négative. L'auteur fournit les meilleures estimations des paramètres et donne des instructions détaillées pour les adapter aux données. Il présente des exemples d'application de ces paramètres à des séries climatologiques. Il montre également comment on applique la distribution binomiale pour obtenir les seuils de confiance des estimations de probabilité à partir d'une distribution quelconque.

Les grandes lignes de l'analyse de corrélation et de régression sont esquissées. L'auteur traite de la propagation de la variance dans les séries climatologiques, notamment de l'effet de co-variance. La corrélation est soigneusement différenciée de l'autocorrélation qui ne peut être utilisée dans cette application qu'avec des séries de données stationnaires. Les formules se rapportant à la propagation de la variance sont appliquées à une équation permettant de déterminer le régime de refroidissement dans un système de climatisation.

L'auteur étudie en détail l'analyse de régression, notamment le cas du passage forcé des droites de régression par l'origine. La variance est utilisée pour vérifier la signification d'une relation et appliquée à la régression proprement dite. Le test de linéarité est également décrit. Les erreurs de la régression et l'erreur type - très importante - d'une prévision

sont exposées. La Note contient un exemple complet d'application de la méthode à une seule variable indépendante. Les méthodes de régression simple et de corrélation multiple sont étendues à deux variables indépendantes et, finalement, à de nombreuses variables indépendantes. La publication se termine par une bibliographie renvoyant à des manuels et à des études statistiques.

Резюме

Современные статистические анализы являются математическим аспектом климатологических анализов, целью которых является климатологический прогноз. В этой технической записке дано введение к основным принципам для разработки таких прогнозов. Методы представленных здесь анализов, иллюстрируются рядом примеров.

После определения климатологических серий, которые положены в основу статистического анализа, автор рассматривает распределение частот (основного элемента климатологического анализа). Из этого естественно вытекает обобщенное распределение полученных вероятностей, которое и есть ни что иное, как климатологический прогноз.

Учитывая, что некоторые метеорологические данные наблюдений не подготавливают климатологических серий в соответствующей форме в связи с разнородностью климатологических данных, автор предлагает несколько простых способов для приведения их к однородности. Он предлагает различные и рациональные методы для согласования, осреднения и суммирования различных изменений, используя фактические климатологические данные. Он указывает также на ограничения в их использовании и толковании.

После общей оценки основных проблем статистических параметров, автор высказывает критическое суждение, касающееся основных проблем расчета статистических параметров. Он создал также метод подхода к некоторым статистическим обобщениям; он определил и уточнил пределы для определения средних значений.

Были достигнуты некоторые основные повторяемости распределений: нормального, гамма, экстремального, биномиального, Пуассона и отрицательные биномиальные.

Автор находит наилучшие способы расчета параметров и дает детальные пояснения для использования при обработке данных наблюдений. Он дает примеры применения этих параметров в отношении климатологических серий. Также дается применение биномиального распределения для получения пределов оценки вероятностей, исходя из любых распределений.

Корреляционный анализ и анализ уравнений регрессии рассматривается в общем плане. Автор приводит разновидности вариаций в климатологических сериях, в частности эффект соизменений. Автор различает корреляцию от автокорреляции, которая может быть использована только с сериями стационарных данных. Формулы, относящиеся к разновидности вариаций применяются к уравнению, определяющему режим охлаждения в кондиционных системах.

Автор дает детальный анализ уравнению регрессии, включая линейное уравнение регрессии. Варианты анализа используются для проверки значения связей, также как и для самой регрессии. Дается также исследование линейности и приводятся весьма существенные ошибки регрессии и типичные ошибки в прогнозах.

Записка содержит пример использования метода по пределу независимой разновидности. Методы простой регрессии и сложной корреляции распространены на две независимые вариации и, в конечном итоге на многочисленные независимые вариации. Публикация заканчивается библиографией, которая ссылается на учебные пособия и статистические исследования.

RESUMEN

El análisis estadístico moderno constituye el aspecto matemático del análisis climatológico, cuyo objetivo es la predicción climatológica. Esta Nota Técnica es una introducción a los principios básicos necesarios para la elaboración de tales predicciones. Los métodos de análisis que se detallan se ilustran por medio de una serie de ejemplos.

Después de haber definido lo que es una serie climatológica, con el fin de establecer la base de un análisis estadístico valedero, el autor estudia la distribución de frecuencias que es un elemento fundamental del análisis climatológico. Siguiendo un orden lógico, se estudia a continuación la distribución acumulativa para la obtención de las probabilidades, que representan las predicciones climatológicas.

Como algunos registros de datos meteorológicos no forman series climatológicas debido a su heterogeneidad, el autor expone seguidamente algunos métodos sencillos que permiten verificar dicha homogeneidad. Explica los métodos de las diferencias y los cocientes utilizados para ajustar, promediar y totalizar las variables e indica sus aplicaciones a los registros climatológicos propiamente dichos. Se exponen también las limitaciones referentes a la utilización e interpretación de estos métodos.

Después de estudiar de un modo general el fundamental problema de la estimación de los parámetros estadísticos, el autor hace un examen crítico de los parámetros estadísticos ordinarios. Muestra asimismo el método que se sigue para determinar el grado de normalidad de varios parámetros estadísticos comunes y define y precisa los límites de confianza de la media.

La Nota estudia varias distribuciones fundamentales de frecuencia: distribución normal, distribución gamma, distribución de los valores extremos, distribución binomial, distribución de Poisson y distribución binomial negativa. El autor da cuenta de las estimaciones más aproximadas de los parámetros e incluye instrucciones detalladas para su adaptación a los datos. Se exponen ejemplos de su aplicación a las series climatológicas y se estudia también la aplicación de la distribución Binomial con el fin de obtener límites de confianza para las probabilidades estimadas que resultan a partir de una distribución cualquiera.

Se trata de una manera general del análisis de correlación y del de regresión y se estudia la propagación de la variancia en las series climatológicas, incluyendo el efecto de la covariancia. Seguidamente se explica con detalle la diferencia que existe entre la correlación y la autocorrelación, la cual sólo puede ser utilizada en esta aplicación si va acompañada de una serie de datos estacionales. Las fórmulas de propagación de la variancia se aplican a una ecuación que expresa el régimen de enfriamiento en un sistema de acondicionamiento de aire.

El autor estudia detalladamente el análisis de regresión, incluyendo la regresión lineal forzada a través del origen. El análisis de la variancia se utiliza para verificar la significación de una relación y se aplica también a la regresión misma. Se describe el método de verificación de la linealidad y los errores tipo de regresión, así como también el error tipo de una predicción, que es de la mayor importancia. La Nota contiene un ejemplo completo de aplicación del método a una sola variable independiente. Los métodos de regresión simple y de correlación múltiple se aplican a dos variables independientes y, finalmente, a numerosas variables independientes.

La publicación termina con una bibliografía en la que se hace referencia a diversos textos y documentos estadísticos.

CHAPTER ONE

CLIMATOLOGICAL SERIES

The methods of statistical analysis apply to climatological data because, to a large extent, if the data are properly taken, sequences of them behave like random variables. Since statistical analysis only applies to samples from populations of data, the sequences of climatological data must be defined so as to be samples from populations. To accomplish this we define a climatological series as a sample series of data consisting of one climatological value for each year of the record being considered. Thus the 30 January average temperatures for a 30-year record form a climatological series. The 30 daily precipitation amounts for 1 January form a climatological series. The 90 February, March, and April monthly precipitation amounts do not form a climatological series but are samples through different populations and are therefore different climatological series; hence they must be dealt with as three separate series. The series of 3 720 hourly temperatures for a five-year record during March does not form a climatological series because there are 24 x 31 different populations, so that really 744 different climatological series are involved. Under certain circumstances such populations can be mixed together, as were the February, March, and April series above, but the individual climatological series and populations must first be defined so that the exact meaning of the mixture of populations is defined in advance of statistical analysis.

Climatological series variables may be either discrete or continuous. Discrete series variables are usually counted values such as the number of days with precipitation greater than 1.0 mm for each of 30 Junes or the number of times the visibility is less than 1 km during each of 30 Julys. Continuous series variables are usually measured values such as temperature and precipitation, for example the series of 30 totals of spring precipitation (each the total of March, April, and May).

A climatological series is never more than a sample from a single population assumed to behave as if it were infinite in extent and having climatic properties such that the observed climatological series is a random sample from that infinite population, that is to say a sample drawn in a manner independent of the individual magnitudes of the members of the infinite population.

1.1 The frequency distribution

The frequency distribution is the basic tool for describing and analysing the population. This is accomplished by estimating the characteristics of the population frequency distribution from the sample or climatological series. To accomplish this the data of the climatological series are tallied in class intervals which are divisions of the range of the climatological variable. The number of class intervals is best taken to be between 10 and 20.

This divides the difference between the largest and smallest value or range of the climatological series into from 10 to 20 equal divisions. The procedure for division into class intervals is best illustrated by the following example for August precipitation amounts (in mm) for Geneva, Switzerland. The 30-year record for 1927-1956 given in the following table is used.

TABLE 1

August precipitation (mm), Geneva, Switzerland

| Year | p | Year | p | Year | p |
|------|-----|------|-----|------|-----|
| 1927 | 250 | 1938 | 79 | 1949 | 49 |
| 28 | 147 | 39 | 85 | 50 | 110 |
| 29 | 83 | 40 | 18 | 51 | 100 |
| 30 | 108 | 41 | 105 | 52 | 125 |
| 31 | 171 | 42 | 48 | 53 | 57 |
| 32 | 62 | 43 | 41 | 54 | 206 |
| 33 | 67 | 44 | 44 | 55 | 107 |
| 34 | 119 | 45 | 133 | 56 | 144 |
| 35 | 157 | 46 | 158 | | |
| 36 | 23 | 47 | 54 | | |
| 37 | 78 | 48 | 72 | | |

To find a class interval for this climatological series we follow our rule: the highest value is 250 mm and the lowest 18 mm. This gives a range of 232 mm. Since 20 mm is a convenient division and gives 13 divisions, this is a suitable class interval. Tallying these by classes we obtain the following table of precipitation p and frequency f :

TABLE 2

Frequency distribution of August precipitation, Geneva

| p | f | p | f |
|---------|---|---------|---|
| 0-19 | 1 | 140-159 | 4 |
| 20-39 | 1 | 160-179 | 1 |
| 40-59 | 6 | 180-199 | 0 |
| 60-79 | 5 | 200-219 | 1 |
| 80-99 | 2 | 220-239 | 0 |
| 100-119 | 6 | 240-259 | 1 |
| 120-139 | 2 | | |

If these frequencies are plotted as blocks proportional to f on the scale of precipitation, the histogram of precipitation for Geneva is obtained. The values of f may be divided by 30, the number of years in the climatological series, to obtain the relative frequencies in each class interval. These sample values are estimates of the probabilities in the population of precipitation amounts in the various class intervals.

1.2 The cumulative distribution

Usually the climatologist is more interested in estimates of probabilities over several class intervals which are more conveniently obtained from the cumulative distribution. The latter also provides better estimates of the probabilities, since the arbitrary division into class intervals, as in Table 2, tends to waste some of the information on the population given by the climatological series.

To obtain the cumulative distribution the data are first put in order as in the following table:

TABLE 3

Cumulative distribution, August precipitation

| m | p | F | m | p | F | m | p | F |
|----|----|-------|----|-----|-------|----|-----|-------|
| 1 | 18 | 0.032 | 11 | 72 | 0.355 | 21 | 119 | 0.677 |
| 2 | 23 | 0.065 | 12 | 78 | 0.387 | 22 | 125 | 0.710 |
| 3 | 41 | 0.097 | 13 | 79 | 0.419 | 23 | 133 | 0.742 |
| 4 | 44 | 0.129 | 14 | 83 | 0.452 | 24 | 144 | 0.774 |
| 5 | 48 | 0.161 | 15 | 85 | 0.484 | 25 | 147 | 0.806 |
| 6 | 49 | 0.194 | 16 | 100 | 0.516 | 26 | 157 | 0.839 |
| 7 | 54 | 0.226 | 17 | 105 | 0.548 | 27 | 158 | 0.871 |
| 8 | 57 | 0.258 | 18 | 107 | 0.581 | 28 | 171 | 0.903 |
| 9 | 62 | 0.290 | 19 | 108 | 0.613 | 29 | 206 | 0.935 |
| 10 | 67 | 0.323 | 20 | 110 | 0.645 | 30 | 250 | 0.968 |

The F values are the cumulative relative frequencies or estimates of the cumulative population probabilities, and are obtained by the formula $F = m/(n + 1)$ where m is the mth value in order of magnitude of the climatological series and n is the number of terms in the climatological series, in this case 30. The division by (n + 1) instead of n gives a better estimate of population probabilities especially at the ends of the distribution. It can be shown that $m/(n + 1)$ gives the best simple estimate of the probabilities.

The F values give the probabilities that precipitation is less than any value shown in the table. For example, the probability that p is less than 62 mm is 0.290, and that it is greater than 62 mm is $1 - F = 0.710$. Note that when probabilities are estimated for a continuous random variable, such as precipitation, it is a misunderstanding of sampling principles to use the wording "equalled or exceeded" or "less than or equal to", for the probability of any exact value occurring is zero. The probability that it is between 62 and 125 mm is $0.710 - 0.290 = 0.420$. Thus the cumulative distribution gives all the information available from histograms, and much in addition, since it uses every value of the climatological series individually to obtain the probability estimates. The sample cumulative distribution may also be put in graphical form by plotting F on the ordinate against p on the abscissa and connecting the points by straight lines. Climatological series with discrete variables may also be treated in a similar manner.

The average temperatures for August for Geneva shown in Table 4 may be analysed in a similar fashion as another example. The series has been arranged in order of magnitude in Table 4.

TABLE 4

Average temperature ($^{\circ}\text{C}$), August, Geneva

| m | t | F | m | t | F | m | t | F |
|----|------|-------|----|------|-------|----|------|-------|
| 1 | 16.9 | 0.032 | 11 | 18.6 | 0.355 | 21 | 19.8 | 0.677 |
| 2 | 17.4 | 0.065 | 12 | 18.7 | 0.387 | 22 | 19.9 | 0.710 |
| 3 | 17.5 | 0.097 | 13 | 18.7 | 0.419 | 23 | 20.3 | 0.742 |
| 4 | 17.8 | 0.129 | 14 | 18.9 | 0.452 | 24 | 20.4 | 0.774 |
| 5 | 17.9 | 0.161 | 15 | 18.9 | 0.484 | 25 | 20.7 | 0.806 |
| 6 | 17.9 | 0.194 | 16 | 19.2 | 0.516 | 26 | 20.8 | 0.839 |
| 7 | 18.1 | 0.226 | 17 | 19.3 | 0.548 | 27 | 20.9 | 0.871 |
| 8 | 18.3 | 0.258 | 18 | 19.5 | 0.581 | 28 | 20.9 | 0.903 |
| 9 | 18.5 | 0.290 | 19 | 19.5 | 0.613 | 29 | 22.0 | 0.935 |
| 10 | 18.6 | 0.323 | 20 | 19.7 | 0.645 | 30 | 22.9 | 0.968 |

Note that since the record length is the same, the F values are the same as in the previous table, and hence have the same interpretation as previously. The estimated probability that the average temperature for August at Geneva is less than 20.3°C is 0.742, and that it is greater than 20.3°C is $1 - 0.742 = 0.258$. The mean recurrence interval or return period (i.e. the average time between occurrences) for values exceeding any value t is $\frac{1}{(1 - F)}$.

Hence for temperatures exceeding 20.3°C the mean recurrence interval is $\frac{1}{0.258}$ or about four years.

1.3 Homogeneity of data series

A data series is said to be homogeneous if it is a sample from a single population. Hence by definition a climatological series is homogeneous and elementary probability analysis must be applied only to climatological series. The previous temperature and precipitation series were, of course, analysed on the assumption of homogeneity. If a series is not homogeneous, adjustments must be made so that statistical estimates will be valid estimates of the population parameters applying to the last terms in the series, or so that they are estimates obtained from a hypothetical homogeneous series including the latest data as elements.

In cases where instrument exposures have changed it is necessary to make a statistical test to ensure homogeneity. Many of the older methods of testing for homogeneity were incomplete in the sense that they provided inadequate criteria for accepting or rejecting the hypothesis of homogeneity. The valid test of homogeneity is a statistical test of hypothesis which provides a hypothesis of homogeneity (null hypothesis) and a rule for accepting or rejecting this hypothesis on the basis of probability of occurrence. Thus if the probability of the evidence for homogeneity is small, it is concluded that the series is heterogeneous; if it is large, the decision is for homogeneity. The rule specifies the probability limit (significance limit) beyond which the hypothesis of homogeneity would be rejected and some alternative to homogeneity accepted. In most instances distributions on the null hypothesis and the alternatives to homogeneity are difficult to specify; hence the so-called non-parametric tests must ordinarily be used.

The alternatives to homogeneity in a series of meteorological data are usually slippage of the mean, trend, or some form of oscillation. Since these alternatives, especially the latter, may be difficult to specify exactly, it is best to use a non-parametric test which does not require exact specification of these alternatives or the null distribution. A well-known non-parametric test which is sensitive to all of these alternatives is the run test. This test is made by counting the number of runs u above and below the median or middle value in a naturally ordered series, and testing this by means of a table of the distribution of u . The test is best illustrated by applying it to the August average temperatures for Geneva. These are given in their historical order in Table 5.

TABLE 5

Runs for observed Geneva temperature series

| | | | | | |
|------|------|---|------|------|---|
| 1927 | 17.4 | B | 1942 | 19.9 | A |
| 28 | 20.9 | A | 43 | 20.9 | A |
| 29 | 18.7 | B | 44 | 22.9 | A |
| 30 | 18.7 | B | 45 | 18.9 | B |
| 31 | 16.9 | B | 46 | 19.2 | A |
| 32 | 20.8 | A | | | |
| 33 | 20.4 | A | 1947 | 22.0 | A |
| 34 | 17.9 | B | 48 | 18.9 | B |
| 35 | 18.1 | B | 49 | 20.7 | A |
| 36 | 18.5 | B | 50 | 19.7 | A |
| | | | 51 | 19.5 | A |
| 1937 | 19.5 | A | 52 | 20.3 | A |
| 38 | 18.6 | B | 53 | 19.8 | A |
| 39 | 18.6 | B | 54 | 18.3 | B |
| 40 | 17.9 | B | 55 | 19.3 | A |
| 41 | 17.8 | B | 56 | 17.5 | B |

From Table 5 it can be seen that the median or middle value is between 18.9 and 19.2. It may be taken as half-way between these two values or 19.05. Using this value the entries in Table 5 may be marked with a B if they are below this value and with an A if above this value. The runs then are marked as sequences of A and B. The total number of runs is seen to be $u = 15$.

It is clear that too many runs would be an indication of oscillation, while too few runs would be an indication of a trend or a shift in the median during the sample record. Hence, if the probability of a u being exceeded were small, an oscillation would be suspected; whereas if the probability of being less than a sample u were small, a trend or shift in median would be suspected. If the probability of being either greater than or less than u is large, then neither oscillation nor trend is suspected and the series is said to be homogeneous or from a single population. To make this test a distribution table of u is required. This is given below. Since the median was chosen, the number of values above the median N_A will equal the number of values N_B below the median; hence the table is for $N_A = N_B$.

TABLE 6

Distribution table of number of runs u
 $N_A = N_B$

| P | | | P | | |
|-------|------|------|-------|------|------|
| N_A | 0.10 | 0.90 | N_A | 0.10 | 0.90 |
| 10 | 8 | 13 | 19 | 16 | 23 |
| 11 | 9 | 14 | 20 | 16 | 25 |
| 12 | 9 | 16 | 25 | 22 | 30 |
| 13 | 10 | 17 | 30 | 26 | 36 |
| 14 | 11 | 18 | 35 | 31 | 41 |
| 15 | 12 | 19 | 40 | 35 | 47 |
| 16 | 13 | 20 | 45 | 40 | 52 |
| 17 | 14 | 21 | 50 | 45 | 57 |
| 18 | 15 | 22 | | | |

Table 6 gives the lower and upper 0.10 significance limits, i.e. for probabilities P of 0.10 and 0.90. Significance limits of 0.10 are most satisfactory for many meteorological applications because, on account of frequent high variability, it is desirable to increase the significance limit probabilities since this in turn will increase the chances of accepting the alternative hypothesis. Since u is discrete, the u values shown in the tables are those corresponding to the probability closest to 0.10 and 0.90. The maximum divergence from exact probability values is ± 0.03 . If a sample u is below the lower limit, heterogeneity is due to trend or mean slippage; if above, to oscillation.

In order to illustrate further the application of the runs test the series in Table 5 has been deliberately made heterogeneous by subtracting 1°C from each of the first 12 years of record and subtracting 0.5°C from each of the next eight years. The heterogeneous series is shown in Table 7 (see following page).

It was seen in Table 5 that $u = 15$ for $N_A = N_B = 15$. The upper and lower limits from Table 6 for $N_A = 15$ are 12 and 19. $u = 15$ is within this range; hence this u is not significantly different from those expected from homogeneous series, and the series is concluded to be homogeneous.

TABLE 7

Runs for heterogeneous Geneva temperature series

| | | | | | |
|------|------|------|------|------|------|
| 1927 | 16.4 | 1934 | 16.9 | 1945 | 18.4 |
| 28 | 19.9 | 35 | 17.1 | 46 | 18.7 |
| 29 | 17.7 | 36 | 17.5 | 47 | 22.0 |
| 30 | 17.7 | 37 | 18.5 | 48 | 18.9 |
| 31 | 15.9 | 38 | 17.6 | 49 | 20.7 |
| 32 | 19.8 | 39 | 18.1 | 50 | 19.7 |
| 33 | 19.4 | 40 | 17.4 | 51 | 19.5 |
| | | 41 | 17.3 | 52 | 20.3 |
| | | 42 | 19.4 | 53 | 19.8 |
| | | 43 | 20.4 | 54 | 18.3 |
| | | 44 | 22.4 | 55 | 19.3 |
| | | | | 56 | 17.5 |

The number of runs is reduced to 11 by the two shifts of the mean which in effect produce a kind of trend. Table 6 at $N_A = 15$ shows that the probability of less than 12 runs is 0.10; and since Table 7 has only 11 runs the heterogeneity was found by the test. Of course it was already known that the heterogeneity was there because it was introduced deliberately. It will naturally be suspected from this example, and correctly so, that the ability of such tests to find heterogeneities when the exact alternatives to homogeneity are not known will not be very good. This brings out the very important point that the best way to determine heterogeneities is to determine their cause in the history of the record. If the history of a record shows changes which could cause heterogeneities and which can be described according to period and character, more powerful parametric tests such as Student's t-test may be employed to determine the significance of the heterogeneities. Such tests, however, may only be employed where the periods and character of the heterogeneities are known a priori.

1.4 Adjustment of climatological means

Heterogeneity in climatological data series is usually due to some disturbing factor such as change in station location or change in exposure. Although in the past attempts have been made to homogenize series having such disturbances, it must be made very clear that it is not possible to homogenize a series in the sense that a new series of individual values is derived with the same properties as a sample from the proper hypothetical population. In other terms, if the data from a particular station are unavailable for a particular period of record, it is impossible to reproduce the individual items of the series for that period. The reason for this is that any adjustment disturbs the variability of the series and hence changes the scale or dispersion of the frequency distribution. It is possible, however, to adjust certain statistics of the series so that these adjusted values are in effect like those estimated from samples taken from the proper hypothetical population. The most common application of such adjustments is to the means of data series for the purpose of obtaining normals. It is recommended that such adjustments be made if possible only on the basis of a priori known heterogeneities.

It may be shown by theoretical analysis that the classical difference and ratio methods are close to optimum for the adjustment of temperature and precipitation means. Such adjustments are often made to compensate for missing records and to remove heterogeneities. The difference method employs the difference between temperature means of two concurrent homogeneous series as an additive factor on the available series mean. The ratio method employs the ratio of precipitation totals or means of two concurrent homogeneous series as a multiplying factor on the available series total or mean. The adjustments are best illustrated by examples.

The method involves using a supplementary station with a concurrent homogeneous record. This station should be as close as possible to the station to be adjusted, as the effectiveness of the adjustment depends on the correlations between the two stations. Usually a station less than 80 km from the station to be adjusted, and in the same climatic régime, will serve the purpose. Several supplementary stations may be averaged and used as the supplementary record, but this usually does not increase the correlation greatly. If a supplementary station does not have a complete record the adjustment may have to proceed by stages, a different supplementary station being used for each period of record.

1.4.1 The difference method

In Table 7 deliberate heterogeneity was introduced into the average temperature record by subtracting 1.0°C from each of the first 12 years, 0.5°C from the next eight, and leaving the last ten unchanged. It is now assumed that during each of the first two periods the station was moved or the exposure of instruments changed, and that it is desired to adjust the 30-year mean to the exposure during the last 10 years. This is a typical adjustment problem. Other arrangements of the heterogeneities in a record are easily taken into account by a simple variation in the adjustment procedure.

To adjust the means of temperature and precipitation of the Geneva record, given the dates of heterogeneous periods and therefore also the dates of homogeneous ones, it has been found convenient to use Lausanne as the supplementary station. It is not presumed that Lausanne is the best supplementary station. It is only used because it serves the purpose of illustrating the adjustment of a known heterogeneity. The adjustment formula for temperature is

$$\bar{y} = a + \bar{x} \quad (1)$$

Here \bar{x} is the mean for the homogeneous period at the supplementary station corresponding to the heterogeneous period at the station whose record is being adjusted, and \bar{y} is the adjusted mean. The adjustment constant a is estimated by the equation

$$a = \bar{v} - \bar{u} \quad (2)$$

Here \bar{v} and \bar{u} are the means from concurrent periods of homogeneous record at the supplementary station and the station being adjusted respectively. The process of adjustment for temperature then consists of estimating a , using concurrent homogeneous records at the supplementary station and the station to be adjusted, and substituting this value in turn in Equation (1) to obtain the adjusted mean \bar{y} . The \bar{y} values for the various parts of the 30-year record are then weighted according to length of period in years and averaged to obtain the adjusted 30-year record.

The means for each period were obtained from Table 7 which is artificially heterogeneous. These are shown in Table 8:

TABLE 8

Mean temperature adjustment for Geneva

| Lausanne \bar{x} | Geneva-unadjusted means | Geneva \bar{y} |
|--------------------|-------------------------|------------------|
| 1927-38 17.9 | (17.9) | 19.3* |
| 1939-46 18.4 | (19.0) | 19.8* |
| 1947-56 18.2 | 19.6 | 19.6 |
| | Adjusted record mean | 19.5* |

Substituting the homogeneous values for \bar{u} and \bar{v} in Equation (2) gives an estimate of the adjustment factor $a = 19.6 - 18.2 = 1.4$. Inserting this in Equation (1) and substituting successively the homogeneous values 17.9 and 18.4 gives $\bar{y} = 17.9 + 1.4 = 19.3^*$ and $\bar{y} = 18.4 + 1.4 = 19.8^*$ the adjusted values. Next multiplying the values of \bar{y} by 12, 8, and 10, their respective lengths of record, summing these and dividing by 30 gives the weighted mean 19.5*. This is the estimated adjusted mean of August average temperature for Geneva. Note that this compares favourably to the actual value for the undisturbed record 19.3. The procedure provides the best estimate of the hypothetical mean for the 1927-1956 record at Geneva based on the homogeneous period 1947-1956.

1.4.2 The ratio method

In order to illustrate the application of the ratio method of adjustment which must be used for precipitation, the Geneva precipitation record for 1927-1956 was made heterogeneous by being subjected to a change of scale, the precipitation for each of the first 12 years being multiplied by 1.20 and each of the next eight by 0.90, the last 10 being left undisturbed. The resulting heterogeneous series is shown in Table 9.

TABLE 9

Heterogeneous precipitation series, Geneva, August

| | | | | | |
|------|-----|------|-----|------|-----|
| 1927 | 300 | 1936 | 28 | 1947 | 54 |
| 28 | 176 | 37 | 94 | 48 | 72 |
| 29 | 100 | 38 | 95 | 49 | 49 |
| 30 | 130 | 39 | 77 | 50 | 110 |
| 31 | 205 | 40 | 16 | 51 | 100 |
| 32 | 74 | 41 | 95 | 52 | 125 |
| 33 | 80 | 42 | 43 | 53 | 57 |
| 34 | 143 | 43 | 37 | 54 | 206 |
| 35 | 188 | 44 | 40 | 55 | 107 |
| | | 45 | 120 | 56 | 144 |
| | | 46 | 142 | | |

Before proceeding with the adjustment it is easy to test the homogeneity of the series to provide a further illustration of the use of the run test. Of course this test is really unnecessary, for the heterogeneities are known a priori. In this instance the median may be readily found by ordering the data to be 97.5 mm. The runs of values above and below the median may be marked as shown in Table 9. This is seen to give the number of runs $u = 9$. Since $N_A = N_B = 15$ as with Table 7, the upper and lower significance limits 12 and 19 are the same as previously. The value 9 lies outside this range; hence the series is not homogeneous. As would be expected, u has been made too small by slippage of the mean values for the periods 1927-1938 and 1939-1946.

Since heterogeneities in precipitation series are scale changes in the frequency distribution, it is proper to adjust for heterogeneities by scale adjustment, i.e. by using the ratio of homogeneous totals. This is seen to be equivalent to adjusting by the difference of homogeneous means.

By this principle, if y is the precipitation for one unit of the year on the station to be adjusted, and x is the corresponding value for the supplementary station, then

$$\Sigma y = b \Sigma x \quad (3)$$

where the summations are over a period heterogeneous at the station to be adjusted. Thus the estimated total precipitation on a unit of the year for a period of record is equal to the total for the same unit and period at the supplementary station times the adjustment constant b . The adjustment constant b is estimated by the equation:

$$b = \frac{\Sigma v}{\Sigma u} \quad (4)$$

where Σv is the sum of precipitation over the homogeneous period at the station to be adjusted and Σu is the sum for the corresponding period at the supplementary station. This, of course, should be the latest period of record for active stations since it is desired to adjust to a population from which observed values at the active station location will be obtained. The process of adjustment consists in estimating b for a homogeneous period by means of Equation (4) and applying Equation (3) with this statistic to the heterogeneous periods. The results are shown in Table 10.

TABLE 10

Mean precipitation adjustment for Geneva

| Lausanne | Σx | Geneva unadjusted totals | Geneva Σy |
|----------|------------|--------------------------|-------------------|
| 1927-38 | 1 602 | (1 613) | 1 295* |
| 1939-46 | 753 | (570) | 609* |
| 1947-56 | 1 267 | 1 024 | 1 024 |
| | | Adjusted record mean | 97.6* |

Substituting the values of Σx and Σy from Table 10 for the homogeneous period 1947-1956 for Σu and Σv in Equation (4) gives $b = \frac{1\ 024}{1\ 267} = 0.8\ 082$.

Inserting this value for b in Equation (3) and successively substituting the homogeneous totals 1 602 and 753 gives $\Sigma y = 0.8\ 082 \times 1\ 602 = 1\ 295^*$ and $\Sigma y = 0.8\ 082 \times 753 = 609^*$ the adjusted values. Finally averaging yields $\bar{y} = \frac{1\ 295 + 609 + 1\ 024}{30} = 97.6^*$ mm. This is a near optimum estimate of the

mean total precipitation for August at Geneva. The average value of the homogeneous 30-year series of August precipitation at Geneva is indeed 99.9 mm.

CHAPTER TWO

ESTIMATION OF STATISTICAL PARAMETERS

2.1 Statistics in general

A statistical parameter is a fixed value which is a function of all of the population values. Thus the mean for a population would be the average of all the values in that population. Since the entire population of values is never known in climatology, it is only feasible to estimate population parameters from samples or climatological series. Such an estimate of a population parameter is called a statistic. A statistic is a function of the sample or climatological series. Statistical parameters may be dealt with only in theory; in practice, statistics or estimates of the parameters must always be used.

Since every function of a random variable is also a random variable, statistics are random variables and are therefore subject to random variation similar to that in a climatological series. Every climatological statistic is therefore a random variable which forms a population for which there is a frequency distribution. The variability of this frequency distribution about the population parameter is called the dispersion of the statistic. There are always a number of functions of the sample or statistics which estimate the same population parameter. The best of these estimates will have the smallest dispersion. The estimate with the least dispersion will in general extract the most information from the sample on the value of the population parameter. The dispersion of a statistic decreases with increase in sample size, hence statistics for long climatological series have less dispersion than those for short climatological series. Since poor statistics have greater dispersion, their use in effect discards climatological record and thus is wasteful of usually scarce record length; it is therefore to be avoided if possible. An example is the use of the median to estimate the centre of a normal (Gaussian) distribution (for example a climatological series of temperature which has a distribution close to normal). Both the median and the mean are statistics for the centre of a normal distribution. The median, however, has a larger dispersion than the mean and in fact requires a climatological series about one third longer than the mean to obtain an equally good estimate of the centre of the distribution. A number of other inefficient statistics are used in climatology, for example the mean absolute deviation as an estimate of the standard deviation, and also certain short-cut estimates of the correlation coefficient. Statistics with the smallest dispersions are called efficient. It is naturally advantageous to employ either efficient statistics or those with high efficiencies in climatological analysis. If the distribution form is not known, little exact information can be inferred about the efficiency of a statistic.

While it is always desirable to use the most efficient statistic available, it is sometimes also desirable, but not necessarily essential in all problems, for it to be mean unbiased or what is commonly known as unbiased. A statistic is said to be (mean) unbiased if the mean of the statistic for m samples of size n approaches its parameter value as m increases without limit

or mn approaches the number of values in the whole population. Efficiency and lack of bias do not naturally occur together. In statistical analysis it is common practice to choose an efficient statistic and make it unbiased if the latter property is necessary, as in cases where statistics are to be added or averaged.

There are in general two kinds of statistics: (a) those which are direct estimates of the parameters of a frequency distribution, and (b) those which are estimates of other population properties. The mean and standard deviation are estimates of the population or distribution parameters of the normal distribution. The mean is also an estimate of the population mean or expected value independent of the distribution form.

2.2 Common statistics of climatological variables

The mode is defined as the value of the random variable where the density of probability is a maximum. If the analytical form of the frequency distribution is known, efficient estimates of the mode may be obtained by substituting efficient estimates of the distribution parameters and obtaining the maximum of the frequency curve by differentiation. If the analytical form of the frequency distribution is not known, there is no good method of estimating the mode. If the sample is large the centre of the class with the highest frequency may be taken as an estimate of the mode. In general the mode is not recommended for use in climatology.

There has been a good deal written about multimodal distributions in climatology. Most of the multimodality observed is caused by mixing small samples from several populations, which gives the false impression that large samples have been used. In these cases the multimodality is not real but only an effect resulting from improper statistical analysis.

The median of a population is defined as the value of the random variable below which the probability of occurrence is 0.50. If the frequency distribution is known, it may be obtained by integrating up to the value of the random variable where the probability reaches 0.50. If the distribution is not known the median is best obtained by reading the 0.50 value from cumulative distributions plotted from data such as those shown in Tables 3 and 4. Rough estimates of the median may be obtained by taking the middle value of an ordered series, or, if there are two middle values, they may be averaged to obtain the median. The median is one of a class of quantities called quantiles which are defined as X_F , where F is the probability of X being less than X_F . The median is then the 0.50 quantile. Quantiles should be estimated from fitted analytical distributions where possible, as those obtained either from the empirical cumulative distributions or from ordered series tend to be more variable.

The mean is the most used climatological parameter. In most cases it is best to obtain it by summing the climatological series and dividing by the number of years of record. It has two properties. First, it is an estimate of the well-known expected value or mathematical expectation, i.e. the mean of the population. This is important in applied climatology, for the mean of any linear function of the climatological series is a linear function of the mean of the series. Second, the mean is the centre of the normal distribution and is therefore the centre of the distribution for climatological series having this distribution. The mean, as computed above, is generally optimum for estimating the expected value for precipitation and optimum for both the expected value and the centre of the distribution for temperature.

The moments about the mean or central moments are also commonly employed in statistical-climatological work. These are defined for the population R by

$$\mu_r = \int_R (x - u)^r f(x) dx \quad (5)$$

Here μ_r is the r^{th} moment, u is the mean, $f(x)$ is the probability density function or frequency curve, and R is the population of interval or region over which $f(x)$ is defined. The unbiased estimate of the second moment or variance is

$$s^2 = \frac{\sum^n (x - \bar{x})^2}{n - 1} \quad (6)$$

The square root of this value is the standard deviation. The higher moments may be estimated by

$$m_r = \frac{\sum^n (x - \bar{x})^r}{n} \quad (7)$$

The third moment is often used to measure the skewness and the fourth moment the flatness of frequency distribution. For these purposes the statistics $g_1 = \frac{m_3}{s^3}$ and $g_2 = \frac{m_4}{s^4} - 3$ which are estimates of the parameters γ_1 and γ_2 may be employed. For the normal distribution $\gamma_1 = \gamma_2 = 0$. The statistic

$a = \frac{\sum^n |x - \bar{x}|}{ns}$ is often substituted for g_2 since it has a simpler distribution.

Moments higher than the 4th are ordinarily not recommended for climatological work since they are highly variable for the short climatological series usually available.

Again it should be stated that, if good estimates of the distribution parameters are available, Formula (5) should be used directly for estimating the moments. Another statistic occasionally used is the range. This statistic is not recommended except for very crude work, since it has a high variability. Related to the range are the extreme values of record. These are even more highly variable than the range and depend greatly on the length of record. The extreme values for each year may, of course, be fitted by appropriate frequency distributions. Statistics of these distributions give a much better appraisal of individual extremes. For example, quantiles from these distributions are independent of the length of record used; hence they give valid information about unusual values.

The coefficient of variation (or variability) or relative standard deviation has also been used in climatology. It is defined as the ratio of the standard deviation to the mean $\frac{s}{\bar{x}}$. The statistic in absolute value depends

on the interpretation which can be given the standard deviation. If the distribution is not normal the standard deviation has no simple meaning, and hence an individual relative standard deviation has little value. However, it is useful for comparison with other relative standard deviations from populations having the same analytical form of distribution. In this case the ordinary estimate may be an inefficient statistic. A better estimate could be obtained using the proper functions of the estimated parameters in Equation (5).

2.3 Sampling variability of climatological means

The sampling variability or accuracy of a statistic is often measured by its standard deviation, which, when applied to a statistic, is commonly called the standard error. In order for the standard deviation or standard error to have a valid interpretation, the distribution must be normal or near normal. Although the distributions of many climatological series are not normal, the distributions of their means for reasonably long records tend to normality. This is a result of the central limit theorem which states that the distribution of means tends to normal with increasing sample size, irrespective of the distributions of individual values, providing the second moments exist. Since the second moment exists for the distributions of every meteorological element, their means will be close to normally distributed for reasonably long records, such as 30 years.

The sample standard error of the mean of a climatological series is $s(\bar{x}) = \frac{s}{\sqrt{n}}$, where n is the number of years in the series and s is the standard deviation of the individuals in a climatological series. This is true regardless of the form of the distribution. In case the distribution is approximately normal and the sample size is 30 years or more, confidence limits may be established for the mean using normal tables. If s is obtained from $n < 30$, it is necessary to use the t distribution tables with $n-1$ degrees of freedom. Thus, for $n \geq 30$, the 0.90 confidence interval for the mean $\bar{x} - 1.64s(\bar{x}) < \mu < \bar{x} + 1.64s(\bar{x})$, where -1.64 and $+1.64$ are the 0.05 and 0.95 values obtained from a table of the normal distribution. This means that the probability is 0.90 that the true or population values of the mean will lie on this interval. Or, if such intervals were computed for successive periods of record of length used for \bar{x} , 9 out of 10 of these would contain μ . The confidence interval gives a good measure of the accuracy of \bar{x} . As in previous statistical tests, 0.90 probability, the complement of 0.10, has been used because most statistics in meteorology cannot be expected to attain an accuracy justifying any higher confidence that a parameter may be on an interval.

In order to determine how closely the distribution of means approaches normality, the skewness statistic $g_1(\bar{x}) = \frac{g_1}{\sqrt{n}}$, and the flatness or kurtosis statistic $g_2(\bar{x}) = \frac{g_2}{n}$, may be employed. With an extreme case of skewness and

flatness the above statistics could, for example, have the sample values $g_1 = 2$ and $g_2 = 6$ (a J-shaped distribution) in the original climatological series. According to the formulas given above, $g_1(\bar{x})$ is reduced to 0.365 and $g_2(\bar{x})$ to 0.2 for a 30-year mean. The small departure from normality shown by these statistics only increases the confidence interval probability from 0.900 to 0.901. The maximum effect at any single probability value will be less than 0.03. Thus, even with such extreme conditions of skewness and flatness in the climatological series, the distribution of 30-year means may be assumed normal without risk of serious bias in the probabilities.

CHAPTER THREE

GENERAL STATISTICAL METHODS

The basic problems of climatological analysis may be classified into three general types: (1) problems of specification which occur in the choice of the analytical form of the population, (2) problems of inference which arise in the estimation of population parameters and in testing hypotheses and establishing confidence intervals on the population parameters, and (3) problems of relationship which occur in relating several climatological variables and in relating climatological variables to non-climatological variables.

The problem of specification is solved by specifying the frequency distribution in the population of the climatological variable. This may be done either empirically or using theoretical reasoning. An empirical specification of the population usually consists of simply assuming the existence of a distribution of probability whose cumulative distribution has the characteristic ogive form. This was the approach followed previously in obtaining the distribution of August precipitation for Geneva. Occasionally on the basis of examination of numerous samples a mathematical form of distribution may be specified for convenience of computation. A theoretical specification of the population distribution is always expressed in mathematical form. This form is derived from a consideration of the bounds of the variable; scale, location, and shape behaviour, behaviour in convolution, etc. A theoretical specification of the normal distribution may result from an application of the central limit theorem.

The estimation part of the inference problem is solved by providing the most satisfactory statistics for estimating the population parameters. As was seen previously the most satisfactory statistics or estimators will be those having a small dispersion in their distributions. Usually maximum likelihood estimates will provide the best estimates of the parameters.

Confidence intervals for the parameter estimates should always be provided to give a measure of their accuracy. Tests of hypothesis may also be made to ascertain whether the population meets certain prescribed conditions or whether the parameters differ from other sets of parameters of similar character. Previously, for example, tests were made to examine the homogeneity of temperature and precipitation series. Confidence interval and test of hypothesis problems are similar in that they both involve distributions of the estimates or statistics.

The relationship problem may involve only climatological variables or it may involve climatological and other variables. The first problem arises when functions of climatological variables are needed to replace climatological variables which are not available, or to form a new variable which has some special properties. For example, statistics of daily temperatures may be impossible or too expensive to obtain directly, and it may be necessary to obtain

estimates of these from monthly statistics. The degree-day variable is a simple example of a function of temperature which had special useful properties not possessed by temperature. The second type of problem, where climatological variables are related to non-climatological variables, is encountered in every problem in applied climatology. The basic objective in such problems is to develop a relationship which will transform a frequency distribution on the climatological variable to one on the applied variable. A simple example would be a relationship between degree-days and heat consumption in a building which would give the distribution of heat consumption from the distribution of degree-days.

Since many of the inference problems of climatology are closely associated with specification problems, these will be discussed together. The test of hypothesis problem has already been introduced in connexion with tests of homogeneity, and space will not allow of further treatment. Further detail on the subject is readily available in statistical literature. The relationship problem will be treated separately.

3.1 Frequency distributions

An example of specification of the population has already been introduced at the beginning of Chapter One, where the empirical distribution was specified for August precipitation at Geneva. The only theory employed there was to assume the existence of a population and a random variable, and hence the set of cumulative probabilities. In many instances of climatological analysis the specification of an empirical distribution is all that is necessary or justified. It is only where the theory is strong, or where several distributions are to be fitted and comparison or smoothing of their statistics is required, that theoretical distributions are fitted. A mathematical fit adds little in other circumstances.

Frequency distributions are of two general types, discrete and continuous. In discrete distributions the probability density is a function of a discrete random variable, i.e. one that varies in steps. The most common discrete climatological variable is (absolute) frequency, for example the number of hail storms, days with rain, etc. In continuous distributions the probability density is a function of a continuous random variable. Temperature, pressure, precipitation, or any element measured on a continuous scale has a continuous random variable. Often for convenience a discrete random variable may be treated as continuous. Also for special application continuous random variables may be transformed to discrete random variables. Cloud height, for example, is a continuous variable which may be transformed into a discrete variable consisting of heights below and above an arbitrary height h .

While there has been a good deal of consideration given to fitting frequency distributions to meteorological data, much of this has been empirical in nature. Often also the fitting has been done to improperly defined populations such as mixtures of several climatological series which have led to quite anomalous interpretations. Because of lack of space only the most common distributions can be discussed.

3.1.1 The normal distribution

The most important continuous distribution in climatological analysis and, of course, in statistical analysis, is the normal or Gaussian distribution.

Its frequency or probability density function is

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

where μ is the population mean and σ is the population standard deviation. μ is best estimated by \bar{x} and σ by s . These are obtained from the sample values x by the relationships

$$\bar{x} = \frac{1}{n} \sum x$$

and

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

The normal distribution function cannot be expressed in terms of simple functions but must be evaluated by means of function expansions. Many tables of the normal distribution function and related functions have been prepared using the variable $u = \frac{(x - \mu)}{\sigma}$ as argument. u is called a standardized variable. Using this variable the distribution function becomes

$$F(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-\frac{1}{2} u^2} du$$

which can be converted to any desired normal distribution simply by varying μ and σ . Thus a single normal table with argument t , which is also a table of the distribution with mean zero and standard deviation unity, may be used to obtain the probabilities for any normal distribution. $F(t)$, of course, gives the probability that u is less than t , $1 - F(t)$ the probability that u is greater than t , and $F(t_2) - F(t_1)$ the probability that u is between t_1 and t_2 .

The importance of the normal distribution in climatology stems, to a considerable extent, from the central limit theorem. This causes means and sums of a sufficient number of climatological values to be normally distributed. For example, rainfall climatological series for short periods for which the mean rainfall is small would have very skewed distributions. As the period increases several shorter periods are added together and an increase in the mean occurs. Thus the size of the mean is some measure of how many periods have been added together; hence, as the mean value gets larger, the sum of the several component periods approaches a normal distribution. It may be shown that, under average conditions, periods with a mean rainfall of 500 mm or more will be close to normally distributed, the greatest discrepancy in probability being about 0.01 at the median. Even for 250 mm means under ordinary conditions the largest discrepancy in probability is only about 0.02.

The normal distribution also provides good fits in most instances to climatological variables which are unbounded above or below, such as temperature and pressure. The sample of data fitted must, of course, be a sample from

a homogeneous climatological series. It must not be a sample from mixed populations which in the past has led to erroneous conclusions such as frequency distributions having several modes, etc.

EXAMPLE 1 - NORMAL DISTRIBUTION

It is well known that monthly average temperature tends to be close to normally distributed. To fit the normal distribution it is necessary to estimate the mean and standard deviation. The estimation formulas are

$$\bar{x} = \frac{\sum x}{n}$$

and

$$s^2 = \frac{1}{n-1} \left[\sum x^2 - \frac{1}{n} (\sum x)^2 \right]$$

The necessary computations are shown below for average January temperature °C for Akureyri, Iceland.

| Year | Temperature | Year | Temperature |
|------|-------------|------|-------------|
| 1932 | -2.2 | 1947 | 3.2 |
| 1933 | 2.4 | 1948 | -0.5 |
| 1934 | -0.5 | 1949 | -3.3 |
| 1935 | 1.8 | 1950 | 1.7 |
| 1936 | -6.0 | 1951 | -3.5 |
| 1937 | 0.5 | 1952 | -2.9 |
| 1938 | -1.3 | 1953 | -0.4 |
| 1939 | -3.4 | 1954 | 1.6 |
| 1940 | -0.4 | 1955 | -3.5 |
| 1941 | -3.1 | 1956 | -3.8 |
| 1942 | 1.0 | 1957 | 0.8 |
| 1943 | -2.9 | 1958 | -3.6 |
| 1944 | -3.8 | 1959 | -5.7 |
| 1945 | -4.3 | 1960 | -0.6 |
| 1946 | 2.0 | 1961 | 0.0 |

The sum of the temperatures is -40.70; so $\bar{x} = \frac{-40.70}{30} = -1.36^\circ\text{C}$.
 $\sum x^2 = 240.49$ and $\frac{(\sum x)^2}{30} = \frac{1656.49}{30} = 55.22$

$$s^2 = \frac{1}{n-1} \left[\sum x^2 - \frac{1}{n} (\sum x)^2 \right] = \frac{1}{29} [240.49 - 55.22]$$

$$= \frac{185.27}{29} = 6.3986$$

$$s = \sqrt{6.3986} = 2.53^\circ\text{C}$$

The probability values are then obtained from the normal distribution function table by the equation

$$x(F) = \bar{x} + s t(F)$$

Here $x(F)$ is the x quantile for F and $t(F)$ is the standard normal quantile for F . To determine $x(0.10)$, i.e. the temperature below which x is expected to fall once in ten, the normal table gives $t(0.10) = -1.28$. Hence

$$\begin{aligned} x(0.10) &= -1.36 - 2.53 \times 1.28 \\ &= -4.6^{\circ}\text{C} \end{aligned}$$

3.1.2 The gamma distribution

Since there are a number of zero-bounded continuous variables in climatology, it is important to give a distribution which may be used for such variables. The gamma distribution which has a zero lower bound has been found to fit several such variables well. It is defined by its frequency or probability density function

$$g(x) = \frac{1}{\beta^{\gamma} \Gamma(\gamma)} x^{\gamma-1} e^{-\frac{x}{\beta}}$$

where β is a scale parameter, γ is a shape parameter, and $\Gamma(\gamma)$ is the ordinary gamma function of γ .

The moments in this instance give poor estimates of the parameters. Sufficient estimates are, however, available and these are closely approximated by

$$\hat{\gamma} = \frac{1}{4A} \left(1 + \sqrt{1 + \frac{4A}{3}} \right)$$

and

$$\hat{\beta} = \frac{\bar{x}}{\hat{\gamma}}$$

where A is given by

$$A = \ln \bar{x} - \frac{\sum \ln x}{n}$$

The distribution function, from which probabilities may be obtained, is

$$G(x) = \int_0^x g(t) dt$$

Pearson's "Tables of the Incomplete Γ -function" gives $G(u)$ where $u = \frac{x}{\beta\sqrt{\gamma}}$, $\gamma = p + 1$, and $u = \frac{x}{\sigma}$.

The gamma distribution has been found to give good fits to precipitation climatological series. In case these contain zeros the mixed distribution function of zeros and continuous precipitation amounts may be employed. This is given by

$$H(x) = q + p G(x)$$

where q is the probability of a zero and $p = 1 - q$. Thus when $x = 0$, $H(0) = q$ as it should be. If m is the number of zeros in a climatological series, q may be estimated by $\frac{m}{n}$.

EXAMPLE 2 - GAMMA DISTRIBUTION

The gamma distribution has been found to fit precipitation data closely. To fit this distribution it is necessary to estimate β and γ which are obtained from the maximum likelihood solutions.

$$\hat{\gamma} = \frac{1}{4A} \left(1 + \sqrt{1 + \frac{4A}{3}} \right)$$

and

$$\hat{\beta} = \frac{\bar{x}}{\hat{\gamma}}$$

where

$$A = \ln \bar{x} - \frac{\sum \ln x}{n}$$

The necessary computations for the November precipitation (mm) for Reykjavik, Iceland, are shown below.

| Year | Precipitation \bar{x} | $\ln x$ | Year | Precipitation \bar{x} | $\ln x$ |
|------|----------------------------|---------|------|----------------------------|---------|
| 1932 | 151.0 | 5.0173 | 1947 | 13.3 | 2.5877 |
| 1933 | 116.1 | 4.7545 | 1948 | 99.2 | 4.5972 |
| 1934 | 74.9 | 4.3162 | 1949 | 72.0 | 4.2767 |
| 1935 | 58.8 | 4.0742 | 1950 | 57.9 | 4.0587 |
| 1936 | 91.4 | 4.5153 | 1951 | 25.1 | 3.2229 |
| 1937 | 44.3 | 3.7910 | 1952 | 60.0 | 4.0943 |
| 1938 | 51.4 | 3.9397 | 1953 | 86.9 | 4.4648 |
| 1939 | 50.2 | 3.9160 | 1954 | 147.2 | 4.9918 |
| 1940 | 79.0 | 4.3694 | 1955 | 37.0 | 3.6109 |
| 1941 | 108.5 | 4.6868 | 1956 | 193.3 | 5.2642 |
| 1942 | 87.0 | 4.4659 | 1957 | 58.7 | 4.0725 |
| 1943 | 129.2 | 4.8614 | 1958 | 212.1 | 5.3571 |
| 1944 | 41.5 | 3.7257 | 1959 | 44.3 | 3.7910 |
| 1945 | 101.3 | 4.6181 | 1960 | 26.8 | 3.2884 |
| 1946 | 53.4 | 3.9778 | 1961 | 96.4 | 4.5685 |

From the table it is seen that $\bar{x} = \frac{\sum x}{n} = \frac{2468.2}{30} = 82.273$ and
 In $\bar{x} = 4.4100$. Averaging the logarithms gives $\frac{\sum \ln x}{n} = \frac{127.2760}{30} = 4.2425$.
 Hence $A = 4.4100 - 4.2425 = 0.1675$,

$$\hat{\gamma} = \frac{1 + \sqrt{1 + 4 \frac{0.1675}{3}}}{4 \times 0.1675}$$

$$= 3.14$$

and

$$\hat{\beta} = \frac{82.27}{3.14} = 26.20$$

To determine the probability that the precipitation is less than 50 mm it must be put in standard form $t(F) = \frac{x}{\hat{\beta}} = \frac{50}{26.20} = 1.91$. From tables of the gamma distribution it is seen that for $\hat{\gamma} = 3.14$ and $t = 1.91$, $F = 0.28$. Hence the probability of the precipitation being less than 50 mm is 0.28.

3.1.3 The extreme value distributions

Often in design problems the climatological variable of interest is the annual extreme, either upper or lower. This arises from the fact that if a designed structure can withstand the highest (lowest) value in a year it can also withstand all other values in the year. Hence a distribution of annual extreme values furnishes the proper climatological prediction. Up to the present the Fisher-Tippett Type I distribution has been of main interest. It has been widely applied by Gumbel. Its distribution function is given by

$$F(x) = \exp \left[-e^{\pm \frac{x - a}{\beta}} \right]$$

Here the negative of the double sign holds for maximum values and the positive sign applies for minimum values. The Type II distribution, which is an exponential transformation of the Type I distribution, has also been employed in climatology. It may be fitted by using the Type I distribution on $\ln z$ (see Example 3).

As with most other skewed distributions the moments give poor estimates of the parameters. Lieblein has provided a simple method of fitting the Type I distribution which gives estimates of the quantiles with minimum variance. This is a desirable property for climatological work, for our ultimate objective is always to obtain quantiles or probabilities.

The Lieblein fitting procedure involves carefully maintaining the original time order of the climatological series and dividing into suitable subgroups for the computations. The following table of weights is needed in the computations.

TABLE OF ORDER STATISTICS WEIGHTS

| m | | $x_{.1}$ | $x_{.2}$ | $x_{.3}$ | $x_{.4}$ | $x_{.5}$ | $x_{.6}$ |
|---|-----------------|----------|----------|----------|----------|----------|----------|
| 1 | | | | | | | |
| 2 | a. _j | 0.91637 | 0.08363 | | | | |
| | b. _j | -0.72135 | 0.72135 | | | | |
| 3 | a. _j | 0.65632 | 0.25571 | 0.08797 | | | |
| | b. _j | -0.63054 | 0.25582 | 0.37473 | | | |
| 4 | a. _j | 0.51100 | 0.26394 | 0.15368 | 0.07138 | | |
| | b. _j | -0.55862 | 0.08590 | 0.22392 | 0.24880 | | |
| 5 | a. _j | 0.41893 | 0.24628 | 0.16761 | 0.10882 | 0.05835 | |
| | b. _j | -0.50313 | 0.00653 | 0.13045 | 0.18166 | 0.18448 | |
| 6 | a. _j | 0.35545 | 0.22549 | 0.16562 | 0.12105 | 0.08352 | 0.04887 |
| | b. _j | -0.45928 | -0.03599 | 0.07319 | 0.12673 | 0.14953 | 0.14581 |

As previously, the sample climatological series is assumed to have n values. Retaining the original time order these n -values are to be divided into subgroups of size m . It will be noted that the table of weights allows m to be chosen from 2 to 6. It is best to choose m as large as possible. Thus, if the sample size is 30, $m = 6$ would be chosen rather than $m = 5$. If n is not divisible by $m = 4, 5, \text{ or } 6$, an additional weighting will be necessary. First consider that $n = 30$. The sample is maintained in original time order and divided into $k = 5$ subgroups of $m = 6$. The values within the subgroups are then arranged in order according to increasing magnitude. The i th subgroup would then appear as $x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6}$. All ordered subgroups are then arranged as in the table on page 24. The dot indicates no operation on the subscript it replaces.

Each column of x is first summed to obtain the $S_{.j}$. These are multiplied by the $a_{.j}$ and summed to obtain the row sum. Next the $S_{.j}$ are multiplied by the $b_{.j}$ and summed to obtain the second row sum.

In the Type I distribution function the exponent $\frac{(x-a)}{\beta}$ is a standardized variable, in other words it is a variable located at a and scaled in β .

| | | | | | | |
|----------------|----------------|----------------|----------------|----------------|----------------|-----------------------------|
| x_{11} | x_{12} | x_{13} | x_{14} | x_{15} | x_{16} | |
| x_{21} | x_{22} | x_{23} | x_{24} | x_{25} | x_{26} | |
| x_{31} | x_{32} | x_{33} | x_{34} | x_{35} | x_{36} | |
| x_{41} | x_{42} | x_{43} | x_{44} | x_{45} | x_{46} | |
| x_{51} | x_{52} | x_{53} | x_{54} | x_{55} | x_{56} | |
| $S_{.1}$ | $S_{.2}$ | $S_{.3}$ | $S_{.4}$ | $S_{.5}$ | $S_{.6}$ | |
| $a_{.1}$ | $a_{.2}$ | $a_{.3}$ | $a_{.4}$ | $a_{.5}$ | $a_{.6}$ | |
| $a_{.1}S_{.1}$ | $a_{.2}S_{.2}$ | $a_{.3}S_{.3}$ | $a_{.4}S_{.4}$ | $a_{.5}S_{.5}$ | $a_{.6}S_{.6}$ | $\sum_{j=1}^6 a_{.j}S_{.j}$ |
| $b_{.1}$ | $b_{.2}$ | $b_{.3}$ | $b_{.4}$ | $b_{.5}$ | $b_{.6}$ | |
| $b_{.1}S_{.1}$ | $b_{.2}S_{.2}$ | $b_{.3}S_{.3}$ | $b_{.4}S_{.4}$ | $b_{.5}S_{.5}$ | $b_{.6}S_{.6}$ | $\sum_{j=1}^6 b_{.j}S_{.j}$ |

If x_p is a quantile in x (a value of x corresponding to $F = p$), then

$$y_p = \frac{x_p - a}{\beta}$$

and

$$x_p = a + \beta y_p$$

Lieblein showed that a minimum variance estimate for a given y_p is given by

$$x_p^* = \frac{\sum_{j=1}^m a_{.j}S_{.j}/k}{\sum_{j=1}^m S_{.j}/k} + \left(\frac{\sum_{j=1}^m b_{.j}S_{.j}/k}{\sum_{j=1}^m S_{.j}/k} \right) y_p$$

Thus the minimum variance estimates for a and β are

$$a^* = \frac{\sum_{j=1}^m a_{.j}S_{.j}/k}{\sum_{j=1}^m S_{.j}/k}$$

and

$$\beta^* = \frac{\sum_{j=1}^m b_{.j}S_{.j}/k}{\sum_{j=1}^m S_{.j}/k}$$

For the sample of 30 under consideration they are

$$\alpha^* = \frac{6}{\sum a_{.j} S_{.j}} / 5$$

and

$$\beta^* = \frac{6}{\sum b_{.j} S_{.j}} / 5$$

When these values are substituted in the Type I distribution function, estimated probabilities are obtained.

In case $m = 5$ or 6 is not an even multiple of the sample size n , a further simple computation is necessary. Suppose that $n = 33$ instead of 30 . The last three values of the sample climatological series then form an additional subgroup $m' = 3$. These values are also arranged in order of increasing magnitude, giving x_{61} , x_{62} , and x_{63} . A similar table is formed with the weights for $m' = 3$ as follows:

| | | | |
|----------------------------------------------|----------|----------|------------------------|
| x_{61} | x_{62} | x_{63} | |
| $a_{.1}$ | $a_{.2}$ | $a_{.3}$ | |
| $a_{.1}x_{61}$ $a_{.2}x_{62}$ $a_{.3}x_{63}$ | | | $\sum^3 a_{.j} x_{6j}$ |
| $b_{.1}$ | $b_{.2}$ | $b_{.3}$ | |
| $b_{.1}x_{61}$ $b_{.2}x_{62}$ $b_{.3}x_{63}$ | | | $\sum^3 b_{.j} x_{6j}$ |

The estimator for this sample is then as before

$$u_p^* = \sum^3 a_{.j} x_{6j} + \left(\sum^3 b_{.j} x_{6j} \right) y_p$$

Lieblein has shown that the estimator for v_p the quantile for the variable in the sample $n = 33$ is

$$v_p^* = \frac{km}{n} x_p^* + \frac{m'}{n} u_p^*$$

For the final estimates this gives

$$\alpha^* = \frac{km}{n} \frac{6}{\sum a_{.j} S_{.j}} / 5 + \frac{m'}{n} \frac{3}{\sum a_{.j} S_{.j}}$$

and

$$\beta^* = \frac{km}{n} \frac{6}{\sum b_{.j} S_{.j}} / 5 + \frac{m'}{n} \frac{3}{\sum b_{.j} S_{.j}}$$

The fitting of any sample size is a simple variation of the above procedures. For minimum values or lower extremes the magnitude order arrangement in the rows of the computation tables is reversed, i.e. instead of going from low to high values they should go from high to low values. All other parts of the tables remain the same.

EXAMPLE 3 - EXTREME VALUE DISTRIBUTIONS

The commonly used distribution is the Type I. The Type II distribution has, however, been found useful in fitting extreme winds. Its distribution function is $F(x) = \exp\left[-\left(\frac{x}{\beta}\right)^\gamma\right]$. Since the Type I distribution on the logarithms is a Type II distribution, the fitting of both distributions may be illustrated by fitting the Type I distribution to the logarithms of annual extreme winds, in this instance the fastest mile of wind in miles per hour.

The computations are carried out for data of the airport at Birmingham, Alabama. Since engineers want design winds at a standard level, in this instance 30 feet (10 m), and anemometer heights vary, it is necessary to reduce all wind speeds to 30 feet. This is done by the power law where the exponent for this airport is assumed to be $\frac{1}{7}$. The formula in terms of logarithms is

$$\ln v(30) = \ln v(z) + \frac{\ln 30 - \ln z}{7}$$

$v(z)$ is the speed at the anemometer height z and $v(30)$ is the speed reduced to 30 feet. The computations are carried out in the table.

| Year | $v(z)$ | $\ln v(z)$ | z | $\ln z$ | $\frac{\ln 30 - \ln z}{7}$ | $\ln v(30)$ |
|------|--------|------------|-----|---------|----------------------------|-------------|
| 1944 | 52 | 3.9512 | 62 | 4.1271 | -0.1037 | 3.8475 |
| 1945 | 54 | 3.9890 | " | " | " | 3.8853 |
| 1946 | 49 | 3.8918 | " | " | " | 3.7881 |
| 1947 | 48 | 3.8712 | 63 | 4.1431 | -0.1060 | 3.7652 |
| 1948 | 47 | 3.8501 | " | " | " | 3.7441 |
| 1949 | 49 | 3.8918 | " | " | " | 3.7858 |
| 1950 | 47 | 3.8501 | " | " | " | 3.7441 |
| 1951 | 65 | 4.1744 | " | " | " | 4.0684 |
| 1952 | 60 | 4.0943 | " | " | " | 3.9883 |
| 1953 | 47 | 3.8501 | " | " | " | 3.7441 |
| 1954 | 48 | 3.8712 | " | " | " | 3.7652 |
| 1955 | 65 | 4.1744 | " | " | " | 4.0684 |
| 1956 | 56 | 4.0254 | " | " | " | 3.9194 |
| 1957 | 56 | 4.0254 | " | " | " | 3.9194 |
| 1958 | 45 | 3.8067 | " | " | " | 3.7007 |
| 1959 | 52 | 3.9512 | " | 4.1431 | -0.1060 | 3.8452 |
| 1960 | 59 | 4.0775 | " | " | " | 3.9715 |
| 1961 | 54 | 3.9890 | " | " | " | 3.8830 |
| 1962 | 43 | 3.7612 | " | " | " | 3.6552 |
| 1963 | 47 | 3.8501 | " | " | " | 3.7441 |
| 1964 | 43 | 3.7612 | " | " | " | 3.6552 |

| j | 1 | 2 | 3 | 4 | 5 | 6 |
|-----|----------|----------|----------|----------|----------|----------|
| a.j | 0.35545 | 0.22549 | 0.16562 | 0.12105 | 0.08352 | 0.04887 |
| b.j | -0.45928 | -0.03599 | 0.07319 | 0.12673 | 0.14953 | 0.14581 |
| i | x_{i1} | x_{i2} | x_{i3} | x_{i4} | x_{i5} | x_{i6} |
| 1 | 3.7441 | 3.7652 | 3.7858 | 3.7881 | 3.8475 | 3.8853 |
| 2 | 3.7441 | 3.7441 | 3.7652 | 3.9883 | 4.0684 | 4.0684 |
| 3 | 3.7007 | 3.8452 | 3.8830 | 3.9194 | 3.9194 | 3.9715 |
| S.j | 11.1889 | 11.3545 | 11.4340 | 11.6958 | 11.8353 | 11.9252 |

| j | 1 | 2 | 3 |
|----------|----------|---------|---------|
| a.j | 0.65632 | 0.25571 | 0.08797 |
| b.j | -0.63054 | 0.25582 | 0.37473 |
| x_{1j} | 3.65520 | 3.66520 | 3.74410 |

For convenience in using the computational notation let $x = \ln v(30)$. The best division into subsamples is three of size six and one of size three. For the three groups of six the a and b tables of subsample size six are employed, giving

$$\frac{\sum_{j=1}^6 a.j S.j}{3} = \frac{11.4183}{3} = 3.8061$$

and

$$\frac{\sum_{j=1}^6 b.j S.j}{3} = \frac{0.2800}{3} = 0.0933$$

as seen under the wind speed table. Similarly for the single group of three the a and b tables of subsample size three are used, giving

$$\sum_{j=1}^3 a.j x_{1j} = 3.6656$$

and

$$\sum_{j=1}^3 b.j x_{1j} = 0.0359$$

Since the six groups contain 18 of the 21 values their weight is $\frac{18}{21} = 0.8571$. For the single three group it is $\frac{3}{21} = 0.1429$ Hence

$$\begin{aligned} a^* &= 0.8571 \times 3.8061 + 0.1429 \times 3.6656 \\ &= 3.7860, \end{aligned}$$

and

$$\begin{aligned} b_1^* &= 0.8571 \times 0.0933 + 0.1429 \times 0.0359 \\ &= 0.0852 \end{aligned}$$

These are the parameters for the Type I distribution of $\ln v$. The parameters for the Type II distribution are given by $\beta_2 = \exp a$ and $\gamma = \frac{1}{\beta_1}$. Hence

$$\beta_2^* = \exp 3.7860 = 44.08$$

and

$$\gamma^* = \frac{1}{0.0852} = 11.74$$

Quantiles are readily obtained by taking logarithms twice of the distribution function giving

$$\ln v(F) = \ln \beta_2^* - \frac{\ln \ln \left(\frac{1}{F}\right)}{\gamma^*}$$

Hence for $F = 0.98$, a common design value,

$$v(0.98) = \exp \left[3.783 - \frac{-3.9021}{11.74} \right]$$

$$= 61.3 \text{ miles/hour}$$

3.1.4 The binomial distribution

This distribution does not in general fit climatological data well because of correlations which occur when the probabilities of occurrence are high enough to meet one of its requirements for application. It is important, however, because it is related to the Poisson and negative binomial distributions, which apply respectively for small probabilities (rare events, often uncorrelated) and for correlated events. Because of this relation it has occasionally been used to give simple rough probability estimates to replace the more crude observed extreme relative frequencies. The most important aspect of the binomial distribution in climatological analysis is that it is the distribution of the estimated probabilities obtained from any distribution function, empirical or theoretical. This makes it possible to obtain confidence limits for estimated probabilities and quantiles.

The binomial probability function is given by

$$f(x) = \binom{m}{x} p^x (1-p)^{m-x}$$

where p is the probability of an event occurring, $1-p$ is the probability of the event not occurring, x is the frequency of occurrence, and x can take the values $0, 1, \dots, m$. The distribution function is given by

$$F(x) = \sum_{t=0}^x \binom{m}{t} p^t (1-p)^{m-t}, \quad t=0, 1, \dots, m$$

This, of course, gives the probability that the frequency is x or less. p is usually estimated by $\frac{\sum x}{n}$ where n is the total number of occurrences and non-occurrences of the event. The climatological events which might be considered in this category are widely varied, for example days when it hailed and days when it did not; days when it rained and days when it did not; days with rainfall less than an amount u and those with rainfall greater than u ; observations with visibility less than V and those with visibility greater than V , etc. Most of these variables have the limitation that they are correlated and that the binomial distribution can therefore only be used for rough biased estimates

of probabilities where only summarized data are available or results are needed quickly.

The important application of the binomial distribution in climatological analysis is to obtain confidence bands for estimated probabilities. It may be seen that when an estimate $F(h)$ of the probability that $x < h$ is obtained from any distribution function, theoretical or empirical, the probabilities in random sampling are divided into those less than h and those greater than h . These form a binomial distribution. If the sample size is m and c values are below h , $F(h) = \frac{c}{m}$. Then the true value of F , ϕ , lies on the interval $p_L < \phi < p_U$ with probability $1 - 2\alpha$ where p_L and p_U are given by

$$\alpha = \sum_{x=c}^m \binom{m}{x} p_L^x (1-p_L)^{m-x}$$

and

$$1-\alpha = \sum_{x=c+1}^m \binom{m}{x} p_U^x (1-p_U)^{m-x}$$

Thus the probability

$$P(p_L < \phi < p_U) = 1 - 2\alpha$$

defines the confidence interval for ϕ with confidence coefficient $1 - 2\alpha$. The formulas have been arranged for use with the "Tables of the Binomial Probability Distribution" (U.S. National Bureau of Standards) where interpolation must be made on α and $1 - \alpha$, the function of the tables in this application. Dixon and Massey's book "An introduction to statistical analysis" gives convenient graphs of confidence limits for $1 - 2\alpha = 0.80, 0.90, 0.95,$ and 0.99 . These replace the formulas and tables for obtaining p_L and p_U . 0.90 is the largest confidence coefficient which should ordinarily be used in climatological analysis.

If the inverted function notation $h = F^{-1}(\frac{c}{m})$ is employed, the confidence interval for η the true value of the quantile h_F may be expressed as the probability relationship

$$P \left[F^{-1}(p_L) < \eta < F^{-1}(p_U) \right] = 1 - 2\alpha$$

This is obtained by simply finding the x values corresponding to $F = p_L$ and $F = p_U$ of the confidence interval.

It should be noted that both confidence intervals are independent of the functional form of F which in a sense makes them non-parametric. If the functional form of F is known, parametric confidence intervals may be available which will be shorter than those above. However, some authors simply assume that p and the corresponding quantiles are normally distributed. This can only give a good approximation at values near the middle of $F(x)$. For values of $F(x)$ near 0 or 1, it is better to use the binomial confidence intervals. They are slightly too broad but they reflect the right shape for the distribution of $F(h)$.

To test for the adequacy of the Poisson distribution

$$\begin{aligned}\chi^2(69) &= \frac{(70 \times 1199)}{261} - 261 \\ &= 60.6\end{aligned}$$

For this value from χ^2 tables

$$P[\chi^2(69) > 60.6] \approx 0.70$$

therefore it is not significant, and the Poisson distribution model is to be preferred.

The formula for the Poisson density function expressed in logarithms is

$$\log g_c = x \log \bar{x} - \log x! - 0.43429 \bar{x}$$

In the table all that is necessary is to subtract column 6 and $0.43429 \bar{x} = 0.43429 \times 3.73 = 1.61990$ from column 5 giving $\log P_c$. P_c is given in column 8 and is the estimate for each x . Multiplying the total frequency 70 by P_c gives column 9 the estimated frequency of occurrence. In column 10 the frequencies g_c and g_0 are compared by the χ^2 test whose total is given in the footing of the table. For this test it is necessary to consolidate observed frequencies less than 5 as shown in column 2. This leaves 7 x-cells from which one degree of freedom is lost for the total and one for having estimated the mean leaving 5. From the χ^2 table

$$P[\chi^2(5) > 5.7981] \approx 0.30$$

Thus the fit of the Poisson distribution is good.

3.1.6 The negative binomial distribution

The negative binomial distribution is useful in fitting discrete dichotomous random variables in which the individual events tend to be correlated. Thus, when too many events are packed on the average into an annual time interval, this distribution tends to fit better than the Poisson distribution. For example, annual hail days and annual frequency of typhoons tend to be fitted better by the negative binomial distribution when the mean annual occurrence is high. Continuous data should in general not be fitted with theoretical discontinuous distributions unless a simple transformation to a discrete variable is first made, for example to a dichotomous variable. There are a number of examples of such misfitting in meteorological literature. On the other hand the fitting of continuous distributions to discontinuous data is often useful.

A test of hypothesis is available to test the adequacy of the Poisson distribution. Thus, if the expression

$$\chi_{n-1}^2 = n \frac{\sum x^2}{\sum x} - \sum x$$

where n is the number years of record, is not greater than the 0.05 value in a chi-squared table with $n-1$ degrees of freedom, the Poisson distribution is adequate. If it exceeds the 0.05 value, the negative binomial distribution should be fitted.

The negative binomial probability function is

$$f(x) = \frac{\Gamma(x+k)}{\Gamma(x+1)\Gamma(k)} \frac{p^x}{(1+p)^{k+x}}$$

The distribution function is given by

$$F(x) = \sum_{t=0}^x f(t)$$

The moment estimates of p and k are

$$p^* = \frac{1}{\bar{x}} (s^2 - \bar{x})$$

and

$$k^* = \frac{\bar{x}^2}{s^2 - \bar{x}}$$

where \bar{x} is the sample arithmetic mean and s^2 is the sample variance.

The moment estimates are not always efficient enough. Fisher has given a criterion which suggests the use of a better-fitting procedure if the efficiency falls below 90%. Thus if

$$\left(1 + \frac{1}{p^*}\right) (k^* + 2) > 20$$

the method of maximum likelihood should be used. This method of fitting is too complex to consider here. For details of the method see Thom, 1957.

EXAMPLE 5 - NEGATIVE BINOMIAL DISTRIBUTION

The Poisson distribution has the population mean equal to the variance. When there is a packing of frequency in individual years, for example, the variance is increased above the mean. The distribution then becomes a negative binomial. This is the case with the number of days with hail (or hail frequency) at Abilene, Texas, for the record 1886-1950. Here the variance $s^2 = 5.25$ while the mean $\bar{x} = 3.58$. The necessary data for the fitting computations are shown in the table on page 33, where x is the number of days with hail, g_o is the observed frequency of hail days, g_c is the calculated frequency, and k and p are parameters of the negative binomial distribution. All logarithms are to base 10.

To test for adequacy of the Poisson distribution

$$\chi^2(64) = 65 \times 1.171/233 - 233 = 93.7$$

For this value from the χ^2 tables

$$P(\chi^2(64) > 93.7) < 0.02$$

Hence the departure from the Poisson distribution is significant and the negative binomial should be fitted.

The formula for the negative binomial density function is

$$g_c(x) = k \frac{p^{*x}}{(1 + p^*)^{k^* + x}}$$

where p^* and k^* are statistics and $k = \frac{\Gamma(k^* + x)}{\Gamma(x+1)\Gamma(k^*)}$. The problem is to find p^* and k^* so that g_c can be calculated.

| x | g_o | $g_o x$ | $g_o x^2$ | log K | $x \log \left[\frac{p}{p+1} \right]$ | log P_c | P_c | g_c |
|-----|-------|---------|-----------|---------|---------------------------------------|-----------|--------|-------|
| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| 0 | 3 | 0 | 0 | 0.0 | 0.0 | -1.27820 | 0.0527 | 3.43 |
| 1 | 11 | 11 | 11 | 0.88593 | -0.49812 | -0.89039 | 0.1287 | 8.37 |
| 2 | 11 | 22 | 44 | 1.52394 | -0.99624 | -0.75050 | 0.1776 | 11.54 |
| 3 | 8 | 24 | 72 | 2.03313 | -1.49436 | -0.73943 | 0.1822 | 11.84 |
| 4 | 10 | 40 | 160 | 2.46006 | -1.99248 | -0.81062 | 0.1547 | 10.06 |
| 5 | 9 | 45 | 225 | 2.82891 | -2.49060 | -0.93989 | 0.1148 | 7.46 |
| 6 | 7 | 42 | 252 | 3.15420 | -2.98872 | -1.11272 | 0.0771 | 5.01 |
| 7 | 2 | 14 | 98 | 3.44553 | -3.48684 | -1.31951 | 0.0479 | 3.11 |
| 8 | 2 | 16 | 128 | 3.70946 | -3.98496 | -1.55370 | 0.0279 | 1.81 |
| 9 | 1 | 9 | 81 | 3.95084 | -4.48308 | -1.81044 | 0.0155 | 1.01 |
| 10 | 1 | 10 | 100 | 4.17329 | -4.98120 | -2.08611 | 0.0082 | 0.53 |
| | 65 | 233 | 1 171 | | | | | |

The mean is $\bar{x} = \frac{233}{65} = 3.58$. The variance is $s^2 = \frac{1171 - (233)^2/65}{64} =$

5.2466. The moment estimates of k and p are then

$$k^* = \frac{\frac{\bar{x}^2}{s^2}}{s^2 - \bar{x}} = 7.6901$$

and

$$p^* = \frac{s^2 - \bar{x}}{\bar{x}} = 0.4655$$

Log K is found from gamma function tables using the value of k^* . The last term in the density function is divided into two factors of which only one involves x. The logarithm of this is given in column 6. The second factor is $k^* \log \left[\frac{1}{1+p^*} \right] = -1.27820$. Log P_c is then obtained by adding log K, column 6, and -1.27820. g_c is finally obtained from 65 P_c . The fit of the negative binomial is to be judged by comparing g_c and g_o .

This comparison may also be done by χ^2 . For this purpose the first two and last four frequencies must be consolidated. This leaves seven degrees of freedom, from which two must be subtracted for fitting k and p , leaving five. Hence

$$\chi^2(5) = \sum \frac{(g_c - g_o)^2}{g_c} = 2.83$$

From tables

$$P(\chi^2(5) > 2.83) > 0.70$$

Thus the fit of the negative binomial is good.

Fisher has shown that the moment estimates are not always efficient for the negative binomial. To test whether they are adequate he used the criterion that

$$C = \left(1 + \frac{1}{p^*}\right) (k^* + 2) > 20$$

For the present data

$$C = 30.51$$

which is greater than 20, so that the efficiency of the moment estimates is adequate. If $C \leq 20$ in an example, maximum likelihood estimation should be employed (see Thom, 1957).

3.2 Correlation and regression analysis

The most important use of correlation analysis in climatological analysis is in connexion with the correlation between climatological series caused by the natural persistence of the meteorological variable within the year. Correlation problems also occur in connexion with compound variables, that is where two or more variables are combined into a single variable, and also in connexion with the propagation of variability in relationships of theoretical or applied problems. Most other applications of correlation are supplementary to regression analysis.

Regression analysis is applied whenever the objective is to estimate a functional relationship for predicting the values of a variable from one or more others. Its main uses are in relating one or more meteorological variables so that one may be substituted for one or more others, and in relating applied variables to meteorological variables. There is also some application to the study of systematic variation of climatological variables in time, but as this is largely of specialized interest it will not be considered here. In any case the regression analysis in this instance is only a variation of that considered here, except that the independent variable is time and the regression terms may be harmonic functions or of some other form.

3.2.1 Correlation analysis

In a strict sense correlation analysis in climatology consists largely of accounting for the effect of correlation between climatological series. For example, if the climatological series for the average temperature

series for 1 and 2 May have sample variances s_1^2 and s_2^2 then the series for the average of 1 and 2 May has a variance which is affected by the correlation between the 1 May and 2 May series. Similarly the variance of the average of the 1, 2, ..., m May series will be affected by the correlations among the m climatological series. Clearly the climatological series could also be for weeks, months or any other portion of the year.

Just as it is necessary always to work with climatological series, so it is necessary to work with the proper correlation coefficients in the present aspect of climatological analysis. The only correlation coefficients useful in the type of analysis considered here are those computed between the two series in any pair of climatological series. If the two series are for the same element, they will be displaced in time within the year; hence it will be possible to have a whole sequence of such correlations. The pairs of climatological series may be separated by different units of time and so there will be a time lag between them. Because of the time-sequential nature of these correlation coefficients, and to differentiate them from autocorrelation coefficients, they will be called sequence correlation coefficients. The sequence correlation coefficient between the i th and j th climatological series is defined as

$$\rho(x_i, x_j) = \frac{E(x_i - \mu_i)(x_j - \mu_j)}{\sigma_i \sigma_j}$$

The numerator is the expected value of the product of the departures of the x_i and x_j from their respective population means, and is called the covariance. The denominator is the product of the population standard deviations of x_i and x_j . The sample estimate of the sequence correlation coefficient is given by

$$r(x_i, x_j) = \frac{\sum_{k=1}^m (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{n s_i s_j}$$

Here x_{ik} is the k th term (year) in the i th climatological series, and x_{jk} is the k th term (year) in the j th climatological series, and \bar{x}_i , s_i , and \bar{x}_j , s_j are their respective means and standard deviations.

The sequence correlation coefficient should be carefully differentiated from the autocorrelation coefficient (sometimes called serial correlation coefficient). The sequence correlation coefficient is really a single correlation coefficient with a time displacement so that the effect of variation in the mean and standard deviation through the year is removed. The autocorrelation coefficient, on the other hand, includes the variation in the mean and standard deviations. In the methods discussed here it is always wrong to use an autocorrelation coefficient.

For the 1, 2, ..., m May climatological series considered above there are $m(m-1)$ possible pairs of series. Since $\rho(x_i, x_j) = \rho(x_j, x_i)$, there are only $\frac{m(m-1)}{2}$ different sequence correlations. All of these must be considered in

obtaining the variance of the sum and average series formed by summing or averaging for each year. If i and j both run over the same sequence of series, the sample variance of the sum may be expressed by

$$v\left(\sum_{i=1}^m x_i\right) = \sum_{i=1}^m s_i^2 + 2 \sum_{i=1}^m \sum_{j>i}^m s_i s_j r(x_i, x_j)$$

This is the variance of the linear function $y = \sum_{i=1}^m k_i x_i$. If the x_i have different weights k_i so that the linear function is $y = \sum_{i=1}^m k_i x_i$, the variance becomes

$$v\left(\sum_{i=1}^m k_i x_i\right) = \sum_{i=1}^m k_i^2 s_i^2 + 2 \sum_{i=1}^m \sum_{j>i}^m k_i k_j s_i s_j r(x_i, x_j)$$

It may be noted that when the $r(x_i, x_j) = 0$, the relationship reduces to the simple variance formula

$$v\left(\sum_{i=1}^m k_i x_i\right) = \sum_{i=1}^m k_i^2 s_i^2$$

If $m = 2$ and k_2 has a negative sign the formula gives

$$v(k_1 x_1 - k_2 x_2) = k_1^2 s_1^2 + k_2^2 s_2^2 - 2k_1 k_2 s_1 s_2 r(x_1, x_2)$$

If $k_1 = 1$ and $k_2 = -1$,

$$v(x_1 - x_2) = s_1^2 + s_2^2 - 2s_1 s_2 r(x_1, x_2)$$

For $r(x_1, x_2) = 0$, $k_1 = 1$, $k_2 = -1$

$$v(x_1 - x_2) = s_1^2 + s_2^2$$

If $k = \frac{1}{m}$, so that the linear function is a simple average $\frac{\sum_{i=1}^m x_i}{m}$, the variance becomes

$$v\left(\frac{\sum_{i=1}^m x_i}{m}\right) = \frac{1}{m^2} \left[\sum_{i=1}^m s_i^2 + 2 \sum_{i=1}^m \sum_{j>i}^m s_i s_j r(x_i, x_j) \right]$$

Thus the average temperature for June has a variance formed from the daily variances and sequence correlations given by

$$v\left(\frac{\sum_{i=1}^{30} x_i}{30}\right) = \frac{1}{30^2} \left[\sum_{i=1}^{30} s_i^2 + 2 \sum_{i=1}^{30} \sum_{j>i}^{30} s_i s_j r(x_i, x_j) \right]$$

The variance of the total precipitation for June based on the individual daily variances and sequence correlation is

$$v\left(\sum_{i=1}^{30} x_i\right) = \sum_{i=1}^{30} s_i^2 + 2 \sum_{i=1}^{30} \sum_{j>i}^{30} s_i s_j r(x_i, x_j)$$

Since monthly total precipitation is not very near to normally distributed, there would be more interest in the variance of the mean or normal for n years $\frac{1}{n} \sum_{i=1}^{30} x_i$. This is $\frac{1}{n^2} v\left(\sum_{i=1}^{30} x_i\right)$.

All of the formulas also apply where the x_i are the variables of different elements which are observed simultaneously or otherwise. This makes them useful in applied problems where the relationship with the applied variable is linear. For example, the outside air cooling load for an air conditioning system may be closely approximated by the linear relationship

$$q = -k_1 t + k_2 t' + k_3$$

where t is dry-bulb temperature, t' is wet-bulb temperature, and the k results from purely physical considerations. Since t and t' are nearly normally distributed around ordinary design levels, the variance of q is important. By means of the formulas given above

$$v(q) = k_1^2 v(t) + k_2^2 v(t') - 2k_1 k_2 r(t, t')$$

The standard deviation of q is therefore $s(q) = \sqrt{v(q)}$ and the mean of q is given by

$$\bar{q} = -k_1 \bar{t} + k_2 \bar{t}' + k_3$$

Thus the normal distribution function $N[q; \bar{q}, s(q)]$ gives the probabilities for climatological predictions based on the distributions of t and t' .

Correlation analysis enters in other ways into climatological analysis, but most of these analyses are closely connected with regression analysis. In fact, wherever relationships are desired between random variables, regression analysis is the proper tool to employ.

3.2.2 Regression analysis

A regression is a functional relationship between an independent random variable and one or more dependent random variables. For a given set of values of the independent variables the regression gives a mean value of the dependent variable. Regression analysis is used in climatology to estimate the constants in functional relationships where these are not given directly as physical quantities. It is used for the establishment of relationships both between climatological series and between climatological series and applied variables. The latter may often be accomplished without climatological series by employing sets of values of the independent variables which are simply uncorrelated within each set and which vary over a range of values equal to that in the climatological series. Thus the relationship between an applied variable and climatological variables can often be established with a short simultaneous record of the two sets of variables.

The first problem in regression analysis is to estimate the constants. This is commonly done by the least squares method applied to the residuals about the regression function obtained when the values of the independent variables have been substituted. The minimization of the residuals of the dependent variable alone requires that the values of the independent variables be fixed or be measured essentially without error. If this condition is not met biases will be introduced in the regression constants. As mentioned above, the values of each variable must also be mutually independent. The least squares estimates have certain optimum properties which make the method a desirable one fitting regressions.

The least square principle is very general and may be applied to almost any type of function. If the regression function is of the form

$$y = R(x_1, \dots, x_k; \beta_0, \beta_1, \dots, \beta_k)$$

the sum of the square residuals may be expressed as

$$\begin{aligned} \sum_{j=1}^n e_j^2 &= \sum_{j=1}^n [y_j - R(x_{1j}, \dots, x_{kj}; \beta_0, \beta_1, \dots, \beta_k)]^2 \\ &= \sum_{j=1}^n (y_j - R_j)^2 \end{aligned}$$

where j runs over the sample values from 1 to n . The "least square" is obtained by minimizing the sums of squared residuals through differentiating and setting to zero. This gives the so-called normal equations

$$\frac{\partial}{\partial \beta_0} \sum_{j=1}^n (y_j - R_j)^2 = 0$$

$$\frac{\partial}{\partial \beta_1} \sum_{j=1}^n (y_j - R_j)^2 = 0$$

$$\begin{array}{ccc} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{array}$$

$$\frac{\partial}{\partial \beta_k} \sum_{j=1}^n (y_j - R_j)^2 = 0$$

The simultaneous solution of the normal equations gives the least squares estimates of $\beta_0, \beta_1, \dots, \beta_k$.

The regression function R can, of course, take an infinite variety of forms. As usual, the linear forms are the most used. Linear regressions for one and two independent variables are considered here. More complicated functions may be analysed by finding the proper normal equations by the process given above.

The linear regression equation in one independent variable is best written as

$$Y = a + \beta (x - \mu)$$

since measuring x from the mean μ makes the least squares estimate of a independent of that of β . The least squares estimates of a and β based on a sample of n pairs of (x, y) are

$$a = \bar{y}$$

and

$$b = \frac{\sum_{i=1}^n y(x-\bar{x})}{\sum_{i=1}^n (x-\bar{x})^2}$$

where the summation is over the sample values.

The regression equation may then be written as

$$y_c = a + b(x-\bar{x})$$

Frequently it is known by physical means that $a = 0$. In this case the regression equation becomes

$$y_c = bx$$

There is now only one normal equation which gives the least squares estimate

$$b = \frac{\sum xy}{\sum x^2}$$

It is often necessary to test the fitted regression for reality and for linearity. This is best done by the analysis of variance which is a technique devised by R. A. Fisher to analyse the mean squares due to several components of the variation. For the linear regression given above it may be observed that there is a total variability of the y which is divided into a variability accounted for by the regression, and a variability unaccounted for by the regression or residual variability. This may be expressed conveniently by an analysis of variance table:

ANALYSIS OF VARIANCE

| | Sum of squares | Degrees of freedom | Mean square |
|------------------------------------------|----------------------------------------|--------------------|-------------------|
| Accounted for by regression | $\sum_{i=1}^n (y_c - \bar{y})^2 = Q_R$ | 1 | $Q_R/1$ |
| Unaccounted for by regression (residual) | $\sum_{i=1}^n (y - y_c)^2 = Q_T - Q_R$ | $n-2$ | $(Q_T - Q_R)/n-2$ |
| Accounted for by mean (total) | $\sum_{i=1}^n (y - \bar{y})^2 = Q_T$ | $n-1$ | |

"Degrees of freedom" is a term used by R. A. Fisher to express the whole number by which the sum of squares is to be divided to give the mean square. When the mean has been estimated, and therefore fixed, only $n-1$ of the observations may vary, since once the mean is fixed and $n-1$ of the observations are chosen, the n th value is automatically fixed by the fact that the n values must average to the mean. One degree of freedom is therefore taken up by fitting the mean, or $n-1$ degrees of freedom remain for estimating the

total mean square which involves the mean. It will not be needed and so is not computed. A further degree of freedom is lost in estimating b ; hence there are $n-2$ degrees of freedom left for estimating the residual mean square. It is seen that the degrees of freedom of the components of variation in an analysis of variance table add to the total degrees of freedom. The sum of squares Q_T and Q_R are obtained from

$$Q_T = \sum y^2 - \frac{1}{n} (\sum y)^2$$

and

$$Q_R = \frac{[\sum y(x-\bar{x})]^2}{\sum (x-\bar{x})^2}$$

The squared correlation coefficient is given by

$$r^2 = Q_R/Q_T$$

From this it is seen that r^2 gives the proportion of the sum of squares or variability explained by the regression. Thus, in using the correlation coefficient as a measure of the goodness of relationship, it is best to square it in order to obtain a realistic estimate of the amount of variability which the linear relationship explains. This will, of course, always be less than r .

The analysis of variance table also provides a test of significance of the linear regression. The statistic F is given by

$$F(1, n-2) = \frac{Q_R/1}{(Q_T - Q_R)/(n-2)}$$

This is to be compared to an F or variance ratio table with 1 and $n-2$ degrees of freedom at the 0.10 or 0.05 significance level to determine whether a linear relationship really exists; or, in other terms, whether the mean square explained by the linear regression is large enough, in comparison to the residual mean square, to decide that the regression is due to a real effect rather than to random sampling.

There has been some tendency to attribute too much importance to tests of significance or tests of hypotheses. Thus it might be concluded that if a regression is significant no more is required: this, however, is far from true, for there are two kinds of significance, practical and statistical. If a regression is not practically significant it is of little use to test its statistical significance. If, however, it is practically significant, then the test of hypothesis must be made in order to test for reality. In the case of the linear relation, practical significance is measured by the squared correlation coefficient, that is by whatever proportion of the total variability is explained by the regression. It may be observed that if $r < 0.50$, i.e., $r^2 < 0.25$, the regression is of very doubtful practical use.

If the sample values of the independent variable x can be divided into, say, four or more classes or columns with at least two y values in each class, a second analysis of variance table may be prepared which will lead to a test of linearity. Such a test will tell whether it might be worth while to fit additional terms of higher degree.

With the data arranged into classes or columns with n_j in the j th column, the total variability may be divided into variability between column means arranged according to increasing x and variation within columns or residual. This leads to a second analysis of variance table:

ANALYSIS OF VARIANCE

| | Sum of squares | Degrees of freedom | Mean square |
|--------------|--------------------------------------------------------------------------------------------------------------------------------|--------------------|---------------------|
| Column means | $\sum_{j=1}^k \sum_{i=1}^{n \cdot j} n \cdot j (\bar{y}_{.j} - \bar{y})^2 = Q_M$ | $k-1$ | $Q_M/(k-1)$ |
| Residual | $\sum_{j=1}^k \sum_{i=1}^{n \cdot j} (y_{ij} - \bar{y}_{.j})^2 = Q_T - Q_M$ | $n-k$ | $(Q_T - Q_M)/(n-k)$ |
| Total | $\sum_{j=1}^k \sum_{i=1}^{n \cdot j} y_{ij}^2 - \frac{1}{n} \left(\sum_{j=1}^k \sum_{i=1}^{n \cdot j} y_{ij} \right)^2 = Q_T$ | $n-1$ | |

An F- test may be made on this table by computing

$$F(k-1, n-k) = \frac{Q_M/k-1}{(Q_T - Q_M)/n-k}$$

If this F is not significant, then there is no relation between y and x linear or otherwise. Had there been doubt about both linearity and whether there were a relationship at all, this test could have been made first.

It will be seen from the first analysis of variance table that the fitting of the linear regression leaves $Q_T - Q_R$ of the variability expressed as a sum of squares unexplained by the regression. If a more complicated function is to provide an improved fit, the improvement must come by removing or reducing this residual variability. Hence, this residual sum of squares may become the total for a third analysis of variance table. Since by the least squares principle a maximum amount of variability will be explained by fitting the column means, the residual from this fitting will be the smallest possible. If this residual is subtracted from the residual left by linear regression, the remainder is the amount explained by the column means over what was explained by the linear regression. The analysis of variance is as follows:

ANALYSIS OF VARIANCE

| | Sum of squares | Degrees of freedom | Mean square |
|----------------------------------|----------------|--------------------|-------------------|
| Column means about regression | $Q_M - Q_R$ | k-2 | $(Q_M - Q_R)/k-2$ |
| Column mean residual | $Q_T - Q_M$ | n-k | $(Q_T - Q_M)/n-k$ |
| Linear regression residual | $Q_T - Q_R$ | n-2 | |

The test for linearity is now made by comparing

$$F(k-2, n-k) = \frac{(Q_M - Q_R)/k-2}{(Q_T - Q_M)/n-k}$$

to the value corresponding to k-2 and n-k degrees of freedom of an F- table. If this is significant the linear regression does not explain all of the variability and it may be desirable to fit higher degree terms.

Once the regression line has been found significant in both the practical and statistical senses, the next interest will be in what error is committed in its use. This may be obtained from the confidence interval for Y, the true value of y_c and the prediction interval for $(y - Y)$, the departure from the true regression. These are found by taking the variance of y_c and $(y - y_c)$ using the regression equation. The square roots of these variances give the required standard deviations. The standard deviation of y_c at x is given by

$$s[y_c(x)] = \left\{ \frac{Q_T - Q_R}{n-2} \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2} \right] \right\}^{\frac{1}{2}}$$

The 0.90 confidence interval for y_c at a given value of x, $y(x)$, is given by

$$P\{y_c(x) - t_{.05}(n-2) s[y_c(x)] < Y(x) < y_c(x) + t_{0.95}(n-2) s[y_c(x)]\} = 0.90$$

where $Y(x)$ is the true value of $y_c(x)$ at x and $t_{.05}(n-2)$ is the value at 0.05 probability from a table of Student's t. It should be remembered that $y_c(x)$ is a conditional mean value, not a future y value, so the confidence interval is for this mean value; it is not the confidence interval for a particular predicted value. This must be obtained from the standard deviation of the observations of the y with respect to the true regression line. This will include the variation in the points about the sample regression line plus the variation in y_c or the sample regression.

This standard deviation at a given x is

$$s[\bar{y}-Y(x)] = \left\{ \frac{Q_T - Q_R}{n-2} \left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x-\bar{x})^2} \right] \right\}^{\frac{1}{2}}$$

and is sometimes called the standard error of a forecast in statistical language. The 0.90 prediction interval for a predicted value of y at the given x is then

$$P\{y_c(x) - t_{.05}(n-2) s[\bar{y}-Y(x)] < [y-Y(x)] < y_c(x) + t_{.95}(n-2) s[\bar{y}-Y(x)]\} = 0.90$$

where t is the same as in the confidence interval for Y .*

As in the case of the air conditioning design cooling load there may be two meteorological variables involved, but the equation connecting them with the design variable may not have its constants determined physically. In that case the problem is one of regression with two independent variables. Or, on the other hand, the simple linear regression may not account for all the variability and a quadratic might need to be added. This regression can be fitted in the same manner as the two independent variable linear regression.

The three-dimensional estimated linear regression may be conveniently expressed by

$$x_{1c} = b_1 + b_2(x_2 - \bar{x}_2) + b_3(x_3 - \bar{x}_3)$$

in which case the estimate $b_1 = \bar{x}_1$. The b are called the regression coefficients and are estimated from the normal equations which are the two independent variable case of the general normal equations given earlier. If the following general notation is used,

$$Q_{ij} = \sum_{i=1}^n (x_i - \bar{x}_i)(x_j - \bar{x}_j)$$

then

$$Q_{11} = \sum_{i=1}^n (x_1 - \bar{x}_1)^2$$

and

$$Q_{12} = \sum_{i=1}^n (x_1 - \bar{x}_1)(x_2 - \bar{x}_2)$$

etc. The normal equations for two independent variables may then be expressed as

$$Q_{22}b_2 + Q_{23}b_3 = Q_{12}$$

$$Q_{23}b_2 + Q_{33}b_3 = Q_{13}$$

* See Example 6 at the end of this chapter.

or in matrix notation

$$[Q] \begin{bmatrix} b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} Q_{12} \\ Q_{13} \end{bmatrix}$$

In this form it will be readily seen how the normal equations can be expanded for regressions of any number of dimensions. The b , other than b_1 , may be found directly from a simultaneous solution of the normal equations, but it will be found convenient to obtain the solution in terms of the Gaussian multipliers since they will be useful in extending the solutions to any number of independent variables. In terms of the Gaussian multipliers c_{ij} and in matrix notation the first equation is

$$\begin{bmatrix} c_{22} & c_{23} \\ c_{23} & c_{33} \end{bmatrix} \begin{bmatrix} Q_{22} & Q_{23} \\ Q_{23} & Q_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 1$$

or in general for k variables

$$[C][Q] = 1$$

where the subscript one does not appear because the x are taken about their means. Inverting gives

$$[C] = [Q]^{-1}$$

Thus the matrix of the c is the reciprocal matrix of the Q . The b are then found from the equation

$$\begin{bmatrix} b_2 \\ b_3 \\ \cdot \\ \cdot \\ \cdot \\ b_{2k} \end{bmatrix} = C \begin{bmatrix} Q_{12} \\ Q_{13} \\ \cdot \\ \cdot \\ \cdot \\ Q_{1k} \end{bmatrix}$$

If

$$D = Q_{23}^2 - Q_{22}Q_{33}$$

the c are given by

$$c_{22} = Q_{33}/D$$

$$c_{23} = Q_{23}/D$$

and

$$c_{33} = Q_{22}/D$$

The b are then given by the equations

$$b_1 = \bar{x}_1,$$

$$b_2 = c_{22}Q_{12} + c_{12}Q_{13}$$

and

$$b_3 = c_{12}Q_{12} + c_{33}Q_{13}$$

The solutions for k variables depend on the calculation of the reciprocal matrix of the Q. This is easily done by the method of pivotal condensation. While the method is simple to apply, space does not allow it to be discussed here. For details see Rao or Snedecor.

The tests of hypothesis on the regression are again facilitated by the analysis of variance. For this purpose two additional Q forms are needed

$$Q_{1.2\dots k} = Q_{11} - b_2Q_{12} - b_3Q_{13} - \dots - b_kQ_{1k}$$

and

$$Q_{p.q} = Q_{pp} - b_{p.q}Q_{pq}$$

where

$$b_{p.q} = Q_{pq}/Q_{qq}$$

$b_{p.q}$ is the sample regression coefficient between x_p and x_q .

The multiple regression analysis of variance is then

Multiple regression A/V

| | S/S | D/F |
|--------------------------------|---------------------|-----|
| Explained by x_2 and x_3 | $Q_{11} - Q_{1.23}$ | 2 |
| Unexplained by x_2 and x_3 | $Q_{1.23}$ | n-3 |
| Total | Q_{11} | n-1 |

where S/S is sum of squares and D/F is degrees of freedom. The multiple regression coefficient is

$$r_{1.23}^2 = (Q_{11} - Q_{1.23})/Q_{11}$$

and the significance test for the multiple regression is given by testing

$$F(2, n-3) = \frac{(Q_{11} - Q_{1.23})/2}{Q_{1.23}/(n-3)}$$

The three simple analyses are as follows:

A/V of x_1 on x_2

| | S/S | D/F |
|----------------------|--------------------|-------|
| Explained by x_2 | $Q_{11} - Q_{1.2}$ | 1 |
| Unexplained by x_2 | $Q_{1.2}$ | $n-2$ |
| Total | Q_{11} | $n-1$ |

The simple correlation coefficient between x_1 and x_2 is then given by

$$r_{1.2}^2 = (Q_{11} - Q_{1.2})/Q_{11}$$

and the F- test by

$$F(1, n-2) = \frac{(Q_{11} - Q_{1.2})}{Q_{1.2}/(n-2)}$$

A/V of x_1 on x_3

| | S/S | D/F |
|----------------------|--------------------|-------|
| Explained by x_3 | $Q_{11} - Q_{1.3}$ | 1 |
| Unexplained by x_3 | $Q_{1.3}$ | $n-2$ |
| Total | Q_{11} | $n-1$ |

$$r_{1.3}^2 = (Q_{11} - Q_{1.3}) / Q_{11}$$

$$F(1, n-2) = \frac{(Q_{11} - Q_{1.3})}{Q_{1.3} / (n-2)}$$

A/V of x_2 on x_3

| | S/S | D/F |
|----------------------|--------------------|-------|
| Explained by x_3 | $Q_{22} - Q_{2.3}$ | 1 |
| Unexplained by x_3 | $Q_{2.3}$ | $n-2$ |
| Total | Q_{22} | $n-1$ |

$$r_{2.3}^2 = (Q_{22} - Q_{2.3}) / Q_{22}$$

$$F(1, n-2) = \frac{(Q_{22} - Q_{2.3})}{Q_{2.3} / (n-2)}$$

From quantities already available in the above tables analyses may be made of the partial regression coefficients $r_{12.3}$ and $r_{13.2}$. These are,

respectively, the correlation between x_1 and x_2 after the influence of x_3 has been eliminated, and the correlation between x_1 and x_3 after the influence of x_2 has been eliminated. The analyses of variance tables again conveniently provide the term for the correlation coefficients and their tests of significance. They also provide tests of whether the fitting of x_3 significantly reduces the residual after x_1 on x_2 has been fitted and whether x_2 significantly reduces the residual after x_1 on x_3 has been fitted. This is important in determining the significance of an added variable, and as will be seen below, an added power.

Partial A/V of x_1 on x_3

| | S/S | D/F |
|--------------------------------|----------------------|-------|
| Increase due to x_3 | $Q_{1.2} - Q_{1.23}$ | 1 |
| Unexplained by x_2 and x_3 | $Q_{1.23}$ | $n-3$ |
| Unexplained by x_2 | $Q_{1.2}$ | $n-2$ |

$$r_{13.2}^2 = (Q_{1.2} - Q_{1.23}) / Q_{1.2}$$

The test of this partial correlation coefficient and whether x_3 adds significantly after x_1 on x_2 has been fitted is given by testing

$$F(1, n-3) = \frac{Q_{1.2} - Q_{1.23}}{Q_{1.23} / (n-3)}$$

Partial A/V of x_1 on x_2

| | S/S | D/F |
|--------------------------------|----------------------|-------|
| Increase due to x_2 | $Q_{1.3} - Q_{1.23}$ | 1 |
| Unexplained by x_2 and x_3 | $Q_{1.23}$ | $n-3$ |
| Unexplained by x_3 | $Q_{1.3}$ | $n-2$ |

$$r_{12.3}^2 = (Q_{1.3} - Q_{1.23}) / Q_{1.3}$$

The test of this coefficient and of x_2 after x_1 on x_3 has been fitted is given by

$$F(1, n-3) = \frac{Q_{1.3} - Q_{1.23}}{Q_{1.23}/(n-3)}$$

By observing the scheme of formation of the analysis of variance tables for two independent variables, the analyses may be extended to any number of variables. The analysis for the second degree equation

$$x_{1c} = b_1 + b_2x + b_3x^2$$

can be accomplished using the above methods by simply substituting the squares of the x values for x_3 and similarly substituting higher powers for further linear terms. The only difference is that b_1 will now be obtained from

$$b_1 = \bar{x}_1 - b_2\bar{x}_2 - \frac{1}{n} b_3 \Sigma x^2$$

The Gaussian multipliers will now be found to be a great convenience in obtaining the standard deviations of x_{1c} and $(x_1 - \Xi)$ from which the confidence bands may be obtained. Ξ is the true value of x_{1c} for a pair (x_2, x_3) . The standard deviations for the three-dimensional regression are then given by

$$s(x_{1c}) = \left\{ \frac{Q_{1.23}}{n-3} \left[\frac{1}{n} + c_{22}(x_2 - \bar{x}_2)^2 + c_{33}(x_3 - \bar{x}_3)^2 + 2c_{23}(x_2 - \bar{x}_2)(x_3 - \bar{x}_3) \right] \right\}^{\frac{1}{2}}$$

for particular pairs of (x_2, x_3) . The standard error of a prediction is given by

$$s(x_1 - \Xi) = \left\{ \frac{Q_{1.23}}{n-3} \left[1 + \frac{1}{n} + c_{22}(x_2 - \bar{x}_2)^2 + c_{33}(x_3 - \bar{x}_3)^2 + 2c_{23}(x_2 - \bar{x}_2)(x_3 - \bar{x}_3) \right] \right\}^{\frac{1}{2}}$$

As in the simple two-dimensional case the confidence band may be determined by employing Student's t with $(n-3)$ degrees of freedom. For the case of k independent variables the standard deviations become

$$s(x_{1c}) = \left\{ \frac{Q_{1.2\dots k}}{n-k} \left[\frac{1}{n} + \sum_{i=2}^k \sum_{j=2}^k c_{ij}(x_i - \bar{x}_i)(x_j - \bar{x}_j) \right] \right\}^{\frac{1}{2}}$$

where the summation is such that the cross-product terms occur twice and

$$s(x_1 - \Xi) = \left\{ \frac{Q_{1.2\dots k}}{n-k} \left[1 + \frac{1}{n} + \sum_{i=2}^k \sum_{j=2}^k c_{ij}(x_i - \bar{x}_i)(x_j - \bar{x}_j) \right] \right\}^{\frac{1}{2}}$$

Student's t for determining the confidence interval will in this case have $(n-k)$ degrees of freedom.

EXAMPLE 6 - SINGLE REGRESSION

The problem is to find the linear relation between a hotel's daily electrical consumption in kwh and degree-days above 65°F. The data given are as follows:

| x(DD) | y(kwh) |
|----------|-----------|
| 5 | 878 |
| 8 | 1 081 |
| 10 | 1 160 |
| 16 | 2 948 |
| 14 | 3 094 |
| 14 | 3 002 |
| 19 | 3 275 |
| 8 | 1 200 |
| 10 | 1 357 |
| 17 | 3 354 |
| 18 | 3 254 |
| 9 | 1 355 |
| <u>0</u> | <u>11</u> |
| 148 | 25 969 |

From these data $\Sigma x = 148$, $\Sigma y = 25\ 969$, $\Sigma xy = 370\ 330$, $\Sigma x^2 = 2\ 056$, and $\Sigma y^2 = 68\ 241\ 641$. Hence $\bar{y} = \frac{25\ 969}{13} = 1\ 997.62$, $n = 13$, $\frac{(\Sigma x)^2}{n} = 1\ 684.92$, $\frac{(\Sigma y)^2}{n} = 51\ 876\ 073.8$, and $\frac{\Sigma x \Sigma y}{n} = 295\ 647.08$. Reduced sums-of squares are immediately available:

$$Q_{yx} = \Sigma xy - \frac{1}{n} \Sigma x \Sigma y = 74\ 682.92$$

$$\Sigma(x-\bar{x})^2 = 2\ 056 - 1\ 684.92 = 371.08$$

and

$$Q_T = \Sigma(y-\bar{y})^2 = 68\ 241\ 641 - 51\ 876\ 074 = 16\ 365\ 567$$

Then

$$b = \frac{Q_{yx}}{\Sigma(x-\bar{x})^2} = \frac{74\ 682.92}{371.08} = 201.26$$

$$\bar{x} = \frac{148}{13} = 11.38$$

$$a = \bar{y} = \frac{25\ 969}{13} = 1\ 997.62$$

Hence

$$y_c = 1\ 997.6 + 201.3 (x - 11.4)$$

or removing \bar{x}

$$y_c = -297.2 + 201.3 x$$

The regression shows a small negative intercept of 297.2 kwh. This indicates that air conditioning is used only when the temperature is somewhere above 65°F. If it is bothersome, the regression can be forced through zero. The regression coefficient is then

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{370\ 330}{2\ 056} = 180.1$$

and

$$y'_c = 180.1 x$$

The regression may be tested and the correlation found through a simple analysis of variance:

| Variability | Sum of squares | Degrees of freedom | Mean square |
|------------------------------------------|---------------------------|--------------------|--------------|
| Accounted for by regression | $Q_R = 15\ 030\ 556$ | 1 | 15\ 030\ 556 |
| Unaccounted for by regression (residual) | $Q_T - Q_R = 1\ 335\ 011$ | 11 | 121\ 365 |
| Accounted for by the mean | $Q_T = 16\ 365\ 567$ | 12 | |

The ratio of the mean squares is distributed as F with 1 and 11 degrees of freedom so

$$F(1, 11) = \frac{15\ 030\ 556}{121\ 365} = 123.8$$

Referring to a table of F it is seen

$$P(F(1,11) > 4.84) < < 0.05$$

hence the regression is significant.

The square of the correlation coefficient between y and x which is proportional to the amount of information is given by

$$r^2 = \frac{Q_R}{Q_T - Q_R} = \frac{15\ 030\ 556}{16\ 365\ 567} = 0.9\ 184$$

This indicates that a very high proportion of the relationship between y and x is explained by the linear regression. The squared correlation coefficient is always to be preferred for an honest measure of a relationship.

To measure the error of a prediction one must compute the standard error of $y - Y$ where Y is the population value of y_c . This is given by

$$\begin{aligned} s(y - Y) &= \left\{ \frac{Q_T - Q_R}{n - 2} \left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\Sigma(x - \bar{x})^2} \right] \right\}^{\frac{1}{2}} \\ &= \left\{ 121\ 365 \left[1.077 + \frac{(x - \bar{x})^2}{371.08} \right] \right\}^{\frac{1}{2}} \end{aligned}$$

For $x = 15$, $(x - \bar{x})^2 = 184.96$; hence

$$\begin{aligned} s(y - Y) &= \left\{ 121\ 365 (1.077 + 0.498) \right\}^{\frac{1}{2}} \\ &= \sqrt{191\ 150} = 437.2 \end{aligned}$$

$(y - Y)$ is distributed as t with $n-2$ degrees of freedom. Referring to a t -table at $n-2 = 11$ degrees of freedom, the 0.05 and 0.95 points are ± 1.796 . $1.796 \times 437.2 = 785.2$. Therefore the true value of y_c or Y at $x = 15$ is covered by the random interval $(3\ 020 \pm 785.2)$ with probability 0.90 or

$$P(2\ 234 < Y < 3\ 805) = 0.90$$

REFERENCES

The first five books are especially recommended as general references.

- Aitken, A. C., -- Statistical mathematics (1939). Oliver and Boyd, London.
- Snedecor, G. W., -- Statistical methods (1956). Iowa State Press, Ames, Iowa.
- Dixon, W. J., and Massey, F. J., -- Introduction to statistical analysis, 1st and 2nd editions, (1951) and (1957). McGraw-Hill, New York.
- Hald, A., -- Statistical theory with engineering applications (1952). John Wiley, New York.
- Mood, A. M., -- Introduction to the theory of statistics (1950). McGraw-Hill, New York.
- Fisher, R. A., -- Statistical methods for research workers (1938), 7th edition or later. Oliver and Boyd, London.
- Gumbel, E. J., -- Statistics of extremes (1958). Columbia University Press, New York.
- Lieblein, J., -- A new method of analyzing extreme-value data (1954). U. S. Nat. Adv. Comm. For Aero., Technical Note No. 3053.
- Pearson, K., et al., -- Tables of the incomplete Γ -function (re-issue 1951). Cambridge University Press, London.
- Pearson, E. S., and Hartley, H. O., -- Biometrika tables for statisticians, Vol. 1, (1956). Cambridge University Press, London.
- Quenouille, M. H., -- Associated measurements (1952). Butterworths, London.
- Rao, C. R., -- Advanced statistics in biometric research (1952). John Wiley, New York.
- Thom, H. C. S., -- The frequency of hail occurrence, Archiv für Meteorologie, Geophysik und Bioklimatologie, Serie B, Band 8, 2 Heft (1957). pp. 185-194.
- U.S. National Bureau of Standards -- Tables of the binomial probability distribution (1950). Applied Mathematics Series 6.
- Weatherburn, C. E., -- A first course in mathematical statistics (1949). Cambridge University Press, London.
-

WMO TECHNICAL NOTES

| | <i>Price</i> |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| * No. 1 Artificial inducement of precipitation | <i>Sw. fr.</i> 1.— |
| * No. 2 Methods of observation at sea | |
| Part I: Sea surface temperature | <i>Sw. fr.</i> 1.— |
| Part II: Air temperature and humidity, atmospheric pressure, cloud height, wind, rainfall and visibility | <i>Sw. fr.</i> 1.— |
| * No. 3 Meteorological aspects of aircraft icing | <i>Sw. fr.</i> 1.— |
| * No. 4 Energy from the wind | <i>Sw. fr.</i> 10.— |
| * No. 5 Diverses expériences de comparaison de radiosondes. Dr. L. M. Malet | } <i>Sw. fr.</i> 1.— |
| * No. 6 Diagrammes aérologiques. Dr. P. Defrise | |
| * No. 7 Reduction of atmospheric pressure (Preliminary report on problems involved) | <i>Sw. fr.</i> 3.— |
| * No. 8 Atmospheric radiation (Current investigations and problems). Dr. W. L. Godson | } <i>Sw. fr.</i> 1.— |
| * No. 9 Tropical circulation patterns. Dr. H. Flohn | |
| * No. 10 The forecasting from weather data of potato blight and other plant diseases and pests. P. M. Austin Bourke | } <i>Sw. fr.</i> 2.— |
| * No. 11 The standardization of the measurement of evaporation as a climatic factor. G. W. Robertson | |
| * No. 12 Atmospheric techniques | <i>Sw. fr.</i> 3.— |
| * No. 13 Artificial control of clouds and hydrometeors. L. Dufour - Ferguson Hall - F. H. Ludlam - E. J. Smith | <i>Sw. fr.</i> 3.— |
| * No. 14 Homogénéité du réseau européen de radiosondages. J. Lugeon - P. Ackermann | } <i>Sw. fr.</i> 4.— |
| * No. 15 The relative accuracy of rawins and contour-measured winds in relation to performance criteria. W. L. Godson | |
| * No. 16 Superadiabatic lapse rate in the upper air. W. L. Godson | } <i>Sw. fr.</i> 3.— |
| * No. 17 Notes on the problems of cargo ventilation. W. F. McDonald | |
| * No. 18 Aviation aspects of mountain waves. M. A. Alaka | <i>Sw. fr.</i> 7.— |
| * No. 19 Observational characteristics of the jet stream (A survey of the literature). R. Berggren - W. J. Gibbs - C. W. Newton | <i>Sw. fr.</i> 9.— |
| * No. 20 The climatological investigation of soil temperature. Milton L. Blanc | } <i>Sw. fr.</i> 5.— |
| * No. 21 Measurement of evaporation, humidity in the biosphere and soil moisture. N. E. Rider | |
| * No. 22 Preparing climatic data for the user. H. E. Landsberg | <i>Sw. fr.</i> 4.— |
| * No. 23 Meteorology as applied to the navigation of ships. C. E. N. Frankcom - M. Rodewald - J. J. Schule - N. A. Lieurance | <i>Sw. fr.</i> 4.— |
| No. 24 Turbulent diffusion in the atmosphere. C. H. B. Priestley - R. A. McCormick - F. Pasquill | <i>Sw. fr.</i> 7.— |
| No. 25 Design of hydrological networks. Max A. Kohler | } <i>Sw. fr.</i> 4.— |
| No. 26 Techniques for surveying surface-water resources. Ray K. Linsley | |
| No. 27 Use of ground-based radar in meteorology (Excluding upper-wind measurements). J. P. Henderson - R. Lhermitte - A. Perlat - V. D. Rockney - N. P. Sellick - R. F. Jones | <i>Sw. fr.</i> 9.— |
| No. 28 Seasonal peculiarities of the temperature and atmospheric circulation regimes in the Arctic and Antarctic. Professor H. P. Pogosjan | <i>Sw. fr.</i> 3.— |
| No. 29 Upper-air network requirements for numerical weather prediction. A. Eliassen - J. S. Sawyer - J. Smagorinsky | } <i>Sw. fr.</i> 14.— |
| No. 30 Rapport préliminaire du Groupe de travail de la Commission de météorologie synoptique sur les réseaux. J. Bessemoulin, président - H. M. De Jong - W. J. A. Kuipers - O. Lönnqvist - A. Megenine - R. Pône - P. D. Thompson - J. D. Torrance | |

* Out of print

| | | |
|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------|
| No. 31 | Représentations graphiques en météorologie. P. Defrise - H. Flohn - W. L. Godson - R. Pône | Sw. fr. 3.— |
| No. 32 | Meteorological service for aircraft employed in agriculture and forestry. P. M. Austin Bourke - H. T. Ashton - M. A. Huberman - O. B. Lean - W. J. Maan - A. H. Nagle | Sw. fr. 3.— |
| No. 33 | Meteorological aspects of the peaceful uses of atomic energy. Part I - Meteorological aspects of the safety and location of reactor plants. P. J. Meade | Sw. fr. 5.— |
| No. 34 | The airflow over mountains. P. Quency - G. A. Corby - N. Gerbier - H. Koschmieder - J. Zierep | Sw. fr. 22.— |
| * No. 35 | Techniques for high-level analysis and forecasting of wind- and temperature fields (English edition) | Sw. fr. 8.— |
| No. 35 | Techniques d'analyse et de prévision des champs de vent et de température à haute altitude (édition française) | Sw. fr. 8.— |
| No. 36 | Ozone observations and their meteorological applications. H. Taba | Sw. fr. 5.— |
| No. 37 | Aviation hail problem. Donald S. Foster | } Sw. fr. 8.— |
| No. 38 | Turbulence in clear air and in cloud. Joseph Clodman | |
| No. 39 | Ice formation on aircraft. R. F. Jones | |
| No. 40 | Occurrence and forecasting of Cirrostratus clouds. Herbert S. Appleman | |
| No. 41 | Climatic aspects of the possible establishment of the Japanese beetle in Europe. P. Austin Bourke | } Sw. fr. 6.— |
| No. 42 | Forecasting for forest fire services. J. A. Turner - J. W. Lillywhite - Z. Pieślak | |
| No. 43 | Meteorological factors influencing the transport and removal of radioactive debris. Edited by Dr. W. Bleeker | Sw. fr. 8.— |
| No. 44 | Numerical methods of weather analysis and forecasting. B. Bolin - E. M. Dobrishman - K. Hinkelmann - K. Knighting - P. D. Thompson | Sw. fr. 4.— |
| No. 45 | Performance requirements of aerological instruments. J. S. Sawyer | Sw. fr. 4.— |
| No. 46 | Methods of forecasting the state of sea on the basis of meteorological data. J. J. Schule - K. Terada - H. Walden - G. Verploegh | } Sw. fr. 6.— |
| No. 47 | Precipitation measurements at sea. Review of the present state of the problem prepared by a working group of the Commission for Maritime Meteorology | |
| No. 48 | The present status of long-range forecasting in the world. J. M. Craddock - H. Flohn - J. Namias | Sw. fr. 4.— |
| No. 49 | Reduction and use of data obtained by TIROS meteorological satellites. (Prepared by the National Weather Satellite Center of the U.S. Weather Bureau) | Sw. fr. 6.— |
| No. 50 | The problem of the professional training of meteorological personnel of all grades in the less-developed countries. J. Van Mieghem | Sw. fr. 4.— |
| No. 50 | Le problème de la formation professionnelle du personnel météorologique de tous grades dans les pays insuffisamment développés. J. Van Mieghem | Sw. fr. 4.— |
| No. 51 | Protection against frost damage. M. L. Blanc - H. Geslin - I. A. Holzberg - B. Mason | Sw. fr. 6.— |
| No. 52 | Automatic weather stations. H. Treussart - C. A. Kettering - M. Sanuki - S. P. Venkiteshwaran - A. Mani | Sw. fr. 3.— |
| No. 52 | Stations météorologiques automatiques. H. Treussart - C. A. Kettering - M. Sanuki - S. P. Venkiteshwaran - A. Mani | Sw. fr. 3.— |
| No. 53 | The effect of weather and climate upon the keeping quality of fruit | Sw. fr. 8.— |
| No. 54 | Meteorology and the migration of Desert Locusts. R. C. Rainey | Sw. fr. 25.— |
| No. 55 | The influence of weather conditions on the occurrence of apple scab. J. J. Post - C. C. Allison - H. Burckhardt - T. F. Preece | Sw. fr. 5.— |
| No. 56 | A study of agroclimatology in semi-arid and arid zones of the Near East. G. Perrin de Brichambaut and C. C. Wallén | Sw. fr. 6.— |
| No. 56 | Une étude d'agroclimatologie dans les zones arides et semi-arides du Proche-Orient. G. Perrin de Brichambaut et C. C. Wallén | Sw. fr. 6.— |
| No. 57 | Utilization of aircraft meteorological reports. P. K. Rohan - H. M. de Jong - S. N. Sen - S. Simplicio | Sw. fr. 4.— |

* Out of print