Abstract: The paper sets out the basic statistical theory of the association between two or more measured quantities. The distinction is made between a mathematical variable and a random variable or variate, and by a simple example the concept of the "power" of a significance test is explained.

The analysis of association is necessarily carried out by the use of models and the theory involved in each of four such models is explained. The models are:

(i) The Regression model which contains only one random variable $(y_i)$ whose mean is linearly dependent on a mathematical variable $(x_i)$, and the power of the associated significance test is increased by examining the means of the $y_i$'s at widely separated values of $x$.

(ii) The Bivariate Normal Distribution model which contains two interdependent variates. Distinction is made between the regression line considered as a property of the bivariate distribution and alternatively as part of the definition of the regression model.

(iii) The Errors in Variables model in which two random variables are measurements involving random errors, and the model is used to arrive at an estimate of the underlying linear relationship between the unobserved structural variables which may be random or not.

(iv) Berkson's model in which analysis is made of the relationship between two quantities one of which is an observed random variable whose means are dependent linearly on unobserved random variables whose means, in turn, are known specified constants.
1. INTRODUCTION

The purpose of this expository note is to attempt to describe some of the basic ideas involved in the modern view of the statistical problem of dealing with the association or relationships between two or more kinds of measured quantity. This is a subject about which statisticians have been almost as confused as non-statisticians and it is only in recent times that, so it would seem, it has become possible to write a logical account of the subject.

2. DISCUSSION

In order to clarify the subject it is first necessary to consider the nature of statistical argument itself. A great deal of statistical work, especially that concerned with official statistics, is purely descriptive - that is to say, masses of numerical data are collected and presented in tabular form. For this purpose summarising quantities such as averages are used. The interpretation of such data is usually more a matter of insight into the economic or other factors lying behind the data than a problem of applying mathematical techniques. The work of the mathematical statistician begins when the mathematical theory of probability is invoked to draw inferences from the data. The manner in which this is done requires careful consideration if we are to understand correctly more complicated problems. Suppose that it is stated that the annual rainfall at a station A is larger than that at a station B. To examine this statement we make observations $x_1, \ldots, x_n$ of annual rainfall at Station A for n years and similarly $y_1, \ldots, y_m$ for m years at station B. We then compare the mean values

$$\bar{x} = \frac{1}{n} \sum x_i, \quad \bar{y} = \frac{1}{m} \sum y_i$$

If $\bar{x} > \bar{y}$ it is true that the mean rainfall at A, for the years considered, is greater than the mean rainfall at B. But it may be asked if this inequality will "continue for ever in the future" (ignoring all questions of climatic change). Without entering into the question of what is meant by a probability distribution in nature we may rephrase the question by asking if the mean of the underlying probability distribution of rainfall at A is greater than that at B, and thus if the underlying distributions are such that the "true" means $\mu_A$ and $\mu_B$ at A is greater than the true mean $\mu_B$ at B. To do this we have to use the theory of probability. We assume that the $x_i$ and $y_i$ have continuous probability distributions with means $\mu_A$ and $\mu_B$ and we consider the probability distribution of $\bar{x} - \bar{y}$. Unless $n$ and $m$ are small it is known that, under very wide assumptions, this will be approximately normally distributed with mean $\mu_A - \mu_B$ and a standard deviation which can be estimated
from \((x_1, ..., x_n)\) and \((y_1, ..., y_m)\). We then see if \(\bar{x} - \bar{y}\) is so large compared with its estimated standard error as to throw doubt on the hypothesis that \(\mu_1 = \mu_2\) (or \(\mu_1 \leq \mu_2\)).

In order to draw such inferences we have had to construct a mathematical model in which there is a conceptual universe of possibilities (all possible values of the \(x_i\) and \(y_i\)) on which is defined a probability distribution. The quantities \(x_i\) and \(y_i\) are no longer regarded as fixed numbers but as "random variables" or as is sometimes said "variates". It is an odd aspect of mathematical terminology that in statistics the word "variable" without the adjective "random" is, for the most part, used to mean a quantity which is kept fixed.

In the above problem we have been concerned with a "test of significance" i.e. we have set up a hypothesis (\(\mu_1 \leq \mu_2\) or \(\mu_1 = \mu_2\)) and we have asked whether this hypothesis is rendered unplausible by the observations. The other most common problem in statistics is to estimate something. For example we may wish to estimate the difference between the means \(\mu_1\) and \(\mu_2\). To do this we use the same probability model as before and calculate from the observations a quantity which is used as an "estimator". In the above problem we would use \(\bar{x} - \bar{y}\) as an estimator of \(\mu_1 - \mu_2\) and we would, of course, also calculate some measure of its accuracy such as its standard deviation.

Thus to apply mathematical statistics in practice we have to construct a model in which some or all of the observations are variates which have probability distributions specified by certain parameters (such as \(\mu_1\), and \(\mu_2\) above). In most cases it is possible to estimate these parameters from the sample and as the sample size grows larger the estimation becomes more and more accurate. If the model is such that a parameter can be estimated the parameter is said to be identifiable. This does not always happen. The model must be constructed from a knowledge of the process producing the observations and it sometimes happens that this results in some parameters in the model being in principle unidentifiable.

A somewhat trivial example is the following. Suppose that in a practical situation we know that an observed quantity \(Z\) is the sum of two unobservable quantities, \(X\) and \(Y\), which are independently normally distributed with means \(m_1\) and \(m_2\) and standard deviations \(\sigma_1\) and \(\sigma_2\). Then the observed values \(y_1, ..., y_n\) of \(Z\) are normally distributed with mean \(m_1 + m_2\) and standard deviation \(\sqrt{\sigma_1^2 + \sigma_2^2}\). It is possible to estimate these two quantities but not \(m_1\), \(m_2\), \(\sigma_1\), \(\sigma_2\) separately. The latter are therefore "unidentifiable". We shall come across this situation in a more important case later.

One aspect of tests of hypotheses also requires mention and that is the concept of "power". Suppose that in the above example of rainfall means we wish to test the hypothesis that \(\mu_1 = \mu_2\) and that
the observations $x_1$ and $y_1$ are known to be normally distributed (not a likely occurrence in practice with rainfall unless a transformation is used). Then we could use $\bar{x} - \bar{y}$ as a test criterion by comparing it with its estimated standard deviation. However another equally valid test would be obtained by testing the difference between the medians of the two samples when compared with its standard deviation (which is larger). This test is valid in the sense that the distribution of the test criterion can be calculated exactly if $\mu_1$ and $\mu_2$ are known, but is not as "powerful" as the test based on the difference of means because if the same significance level is used in both cases, the test using means is more likely to show a significant deviation if $\mu_1$, really is different from $\mu_2$.

In meteorological problems data are often scanty and it is important to use the most powerful tests available and also the most efficient methods of estimation. Sometimes this is not possible because we cannot be certain about the exact form of the distributions involved and in such cases we may have to use ranking, or other non-parametric methods which are, in general, less powerful.

We now apply these considerations to the problem of analysing statistically data in which each sampling unit consists of two or more measurements. We confine ourselves to the case of two measurements for simplicity as the more general situation involves the same problems. We suppose, then, that we have a series of $n$ pairs of observations $(x_1, y_1), \ldots, (x_n, y_n)$ and that we wish to draw inferences from this data by setting up probability models. An unmitigated nuisance in many meteorological problems is that such data is often only treatable under the assumption that the probability distributions associated with different pairs are not independent e.g. there may exist serial dependence. This will certainly happen if the pairs refer to meteorological observations at the same place on successive days. We shall ignore such considerations and assume also that the distributions associated with each pair are the same (e.g. no "trend"). We shall also usually assume normality in any distributions used in the models. This will often be inexact but is necessary for simplicity.

Under these restrictions four essentially different models have been proposed and differ markedly in their assumptions and the type of inference which can be drawn from them.

The first of these models is the Regression Model. Here we suppose that the $y_i$ are random variables which are distributed normally with the same standard deviations and means which are linearly dependent on the $x_i$ which are taken as fixed quantities and not as random variables. Thus we suppose that the expectation of $y_i$, $E(y_i)$, is equal to $\alpha + \beta x_i$. This is known as linear regression and the idea and the corresponding theory can be generalised to non-linear cases which do not concern us here.
The $y_i$ all have the same standard deviation, an assumption not always fulfilled, and the regression is therefore said to "homo-scedastic" (a word which I commend to meteorologists who wish to impress their colleagues. "Orographic" is sometimes used by statisticians for the same purpose.)

The essential feature of this model is that the $x_i$ are constants and not random variables. Thus the probability universe of reference consists of all possible values of the $y_i$ with their corresponding probabilities, the $x_1, \ldots, x_n$ being held fixed and equal to their observed values. The constants $\alpha$ and $\beta \cdot \gamma$ can be estimated by fitting a sample regression line, $Y = \alpha + \beta \cdot \gamma x_i$, by least squares i.e. by minimising the sum of squares of the deviations of the $y_i$ from $\alpha + \beta \cdot \gamma x_i$. This sum is

$$\sum_{i=1}^{n} (y_i - \alpha - \beta \cdot \gamma x_i)^2$$

Under the assumptions made, this provides the best possible estimators of $\alpha$ and $\beta$. Tests of significance can also be made. If we wish to test the hypothesis $\beta = 0$ (or more generally $\beta = \beta_0$) we calculate

$$t = \frac{(b - \beta_0) \sqrt{(n-2)}}{\sqrt{\sum (y_i - \alpha - \beta \cdot \gamma x_i)^2}}$$

where $b$ is the sample estimator of $\beta$ (the regression coefficient) given by

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

The quantity $t$ is then distributed in Student's $t$-distribution with $n - 2$ degrees of freedom. The important feature of this model is that the power of the test that $\beta = \beta_0$ (and in particular $\beta = 0$) is dependent on the values of the $x_i$. If we can choose the values of the $x_i$ before making the observations on the $y_i$ the best thing to do is to have half of them take one value, $X_0$ say, and the other half $X_1$, where $X_0$ and $X_1$ are as far apart as possible. The test of $\beta$ then reduces to a $t$ test for the difference of the means of the $y_i$ values at $X_0$ and $X_1$.

The above model has only a single random variable, $y$, dependent on $x$ and of which we have a sample for a number of different fixed values of $x_i$. The second type of model is the bivariate normal distribution model in which we have two dependent variates $x$ and $y$ and the set of pairs $(x_1, y_1), \ldots, (x_n, y_n)$ is regarded as a sample of $n$ paired variates from a bivariate distribution whose probability density is equal to
\[
\frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1 - \rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left(\frac{(x-m_1)^2}{\sigma_1^2} - 2\rho \frac{(x-m_1)(y-m_2)}{\sigma_1 \sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2}\right)\right).
\]

Here \(m_1\) and \(m_2\) are the means of \(x\) and \(y\), \(\sigma_1\) and \(\sigma_2\) their standard deviations, and \(\rho\) their correlation. The probability universe of reference now consists of all possible pairs of values \((x_i, y_i)\) with the probability density given above. The estimation of \(m_1, m_2, \sigma_1\) and \(\sigma_2\) is easy and to estimate the correlation coefficient we calculate the sample correlation coefficient

\[
r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum (x_i - \bar{x})^2\right) \left(\sum (y_i - \bar{y})^2\right)}}.
\]

If \(\rho = 0\) the two variates are independent and we may test this by calculating \(r\) and using its known distribution. When \(\rho = 0\) this is equivalent to calculating the quantity

\[
t = \frac{r}{\sqrt{(n-2)/(1-r^2)}}
\]

which, when \(\rho = 0\), is distributed in Student's \(t\)-distribution with \(n - 2\) degrees of freedom. In fact considered as a function of the sample values \((x_i, y_i)\) this is exactly the same quantity as we calculated previously in the regression model and moreover has the same distribution. However the tests of significance in the two cases are quite different because they refer to different probability universes. In the first cases we are considering the distribution of \(t\) over all possible values of the \(y_i\), keeping the \(x_i\) fixed, whilst in the second case we let both the \(x_i\) and the \(y_i\) vary. If the null hypothesis \(\rho = 0\) (or \(\beta = 0\)) is true the tests give the same answer but their powers are quite different as we can see from the fact that we can vary the power of the first test by choosing the \(x_i\) in different ways.

There are two properties of the distribution of \(r\) in a sample from a bivariate distribution which deserve mention here. If the joint distribution is such that \(x\) and \(y\) are independent, and one of them, \(y\) say, is normally distributed, the above quantity \(t\) still has Student's \(t\)-distribution, whatever the distribution of the other quantity \(x\). This is easily seen by observing that \(t\) has this distribution for all values of the \(y\) conditional on the \(x_i\)'s having the values actually observed. A further averaging over the values of the \(x_i\) does not then affect the distribution. We may even go further. If the \(x\) and \(y\) are independently distributed in any distribution whatever, \(r\) will have a distribution with zero mean and standard deviation \((n - 1)^{-\frac{1}{2}}\) exactly. Its exact distribution, apart from the first two moments, will depend on the distributions of \(x\) and \(y\) but is not very sensitive to them. This fact is occasionally useful.
When \( \rho \neq 0 \) and the distribution is bivariate normal the distribution of \( r \) is very complicated but has been studied in great detail and tabulated. \( r \) may be then regarded as an estimator of \( \rho \) which is a measure of the interdependence of the two variables. It is possible to define two "regression lines" associated with a bivariate normal distribution. From the formula for the joint distribution it is easy to see that the conditional distribution of \( y \) for a fixed value of \( x \) is a normal distribution with standard deviation \( \sigma_2 \sqrt{1 - \rho^2} \) and a mean \( m_2 + \rho \sigma_2 \sigma_1^{-1} (x - m_1) \), i.e., linearly dependent on \( x \). This is the regression of \( y \) on \( x \). The latter may be regarded as the best predictor of the value of \( y \) which will be associated with any given observed value of \( x \). Similarly we can define a regression line of \( x \) on \( y \). These two lines are to be regarded as properties of a bivariate distribution and it was in this sense that Galton originally introduced the idea of a regression line. It is only in relatively recent times that statisticians have realised the essential difference between a regression line considered as a property of a bivariate distribution and on the other hand as part of the definition of a regression model. If \( y \) is the variate in the latter, \( x \) is not, and it is meaningless to speak of the regression of \( x \) on \( y \). For example in considering problems of climatic trend we may consider the regression of rainfall on time but no meaning could be ascribed to the idea of the regression of time on rainfall. Time is not a variate and the model used is necessarily a regression model and not a correlation model.

Another model which is also of importance in meteorology and has been widely studied in economics may be called the "Errors in Variables" model. Suppose that \( X \) and \( Y \) are observed random variables which result from a model defined by the following relationships:

\[
\begin{align*}
X &= \lambda + u + \varepsilon \\
Y &= \mu + \beta u + \eta
\end{align*}
\]

Here \( \lambda, \mu, \beta \) and \( \varepsilon, \eta \) are independent random variables which are normally distributed with zero means and standard deviations \( \sigma_1 \) and \( \sigma_2 \). \( u \) may be either a variable or a variate. Then \( X \) and \( Y \) can be regarded as measurements, made with random errors \( \varepsilon \) and \( \eta \), of two underlying unobserved quantities

\[
\begin{align*}
v &= \lambda + u \\
w &= \mu + \beta u
\end{align*}
\]

which satisfy the equation \( (w - \mu) = \beta (v - \lambda) \)

This is the underlying linear relationship between the unobserved structural variables \( v \) and \( w \). Our aim is to estimate this relationship from a set of observations \((x_1, y_1), \ldots, (x_n, y_n)\) of \( X \) and \( Y \). If \( u \) is regarded as a quantity which varies from observation to observation, taking the (unknown) values \( u_1, \ldots, u_n \) we see that the unknown
constants in our model are \( u_1, \ldots, u_n, \lambda, \mu, \beta, \sigma \) and \( \sigma_2 \) and it will be found that \( \lambda, \mu \) and \( \beta \) can not be estimated without further knowledge. A more interesting model is obtained if we assume further that \( u \) is a random variable which is normally distributed with zero mean and standard deviation \( \sigma \) (any other mean can be taken care of by varying \( \lambda \) and \( \mu \)). We then have six underlying parameters in the model, \( \lambda, \mu, \beta, \sigma, \sigma_1 \) and \( \sigma_2 \). It is also easy to see that \( X \) and \( Y \) are then random variables which are distributed in a bivariate normal distribution such that

\[
\begin{align*}
E(X) &= \lambda \\
E(Y) &= \mu \\
\text{Var}(X) &= \sigma^2 + \sigma_1^2 \\
\text{Var}(Y) &= \beta^2 \sigma^2 + \sigma_2^2 \\
\text{Cov}(X, Y) &= \beta \sigma_2
\end{align*}
\]

Such a bivariate normal distribution is completely determined by the five quantities on the left of these equations. These five quantities are all identifiable and can be estimated simply in the usual way and no other quantities connected with the model and not expressible in terms of these five can be estimated. From the equations we see that these quantities determine \( \lambda \) and \( \mu \) but not \( \beta, \sigma_1, \sigma_1 \sigma_2, \sigma \) and \( \sigma_2 \). To see this suppose that \( k = \frac{\sigma_1^2}{\sigma_2^2} \). Then if \( k \) is known (0 \( \leq k \leq \infty \)) \( \beta, \sigma, \sigma_1^2 \) and \( \sigma_2^2 \) can be determined uniquely in terms of the quantities on the left and \( k \). We therefore see that \( \beta, \sigma, \sigma_1^2 \) and \( \sigma_2^2 \) are unidentifiable parameters in the model and unless we have some further knowledge (such as the value of \( k \)) they cannot be found. Thus in this example we cannot "fit a straight line when both variables are subject to error".

We could, of course, calculate the two regression lines. These could be used for prediction in the sense that given an observed value of \( X \), the regression of \( Y \) on \( X \) would give the best prediction of the associated observed value of \( Y \) (and similarly for \( X \) on \( Y \)) but unless \( \sigma_1^2 = 0 \) or \( \sigma_2^2 = 0 \) neither of these lines is the underlying structural relation.

There has been a great deal of confusion in scientific literature on this point. The distinction is essential for we may often be mainly interested in \( \beta \). If \( X \) is rainfall over a catchment area and \( Y \) is runoff (this may not fit the model very well in practice as a result of non-normality and lack of homoscedasticity) then \( \beta \) may be of great interest in determining how much increase of runoff one would get for a given increase of rainfall by artificial stimulation.

In some circumstances further knowledge such as the value of \( \sigma_1^2, \sigma_2^2 \), or \( \sigma_1^2 \sigma_2^{-2} \), may be available and it may then be possible to estimate the structural relationship.
However if we cannot estimate the line \( w - \mu = \mathbf{p} (v - \lambda) \) we can determine some things about it. It must pass through the point \((\mu, \lambda)\) which can be estimated by the means of the \(x_i\) and \(y_i\) and furthermore it must have a slope lying between the slopes of the two regression lines. Thus

\[
\frac{\text{cov}(X, Y)}{\text{var}(X)} \leq \beta \leq \frac{\text{var}(Y)}{\text{cov}(X, Y)}
\]

and both sides of this inequality can be estimated. Thus we can certainly put restrictions on the line even if we cannot estimate it. These restrictions themselves have to be estimated and with small samples may not be very secure.

There is a fourth model due to Berkson (1950) which has considerable interest to statisticians but is unlikely to occur in meteorology. This model is an attempt to make logical the analysis of the relationship between two quantities \( x \) and \( y \) which are linearly related. We suppose that we are doing an experiment in which we aim to control the variable \( x \) to certain values \( x_1, \ldots, x_n \) but that in doing this we make errors \( \xi_1, \ldots, \xi_n \) so that \( x_i = X_i + \xi_i \), and these errors are independently and normally distributed independently of the values \( X_i \) which are previously prescribed.

\( y_i = \alpha + b x_i \) and measurements, \( Y_i \), are made of \( y \) in such a way that \( Y_i = y_i + \xi_i \) where the \( \xi_i \) are independently and normally distributed. Then the regression of \( Y \) on \( X \) is a normal regression and the observed regression coefficient is an unbiased estimator of \( b \). The structure of this model is superficially similar to that of the previous one but the essential difference is that the \( \xi_i \) are now independent of the "observed" quantities \( X_i \) instead of being independent of the \( x_i \).

The first three models are the ones which are relevant to the meteorologist and the choice between them will depend mainly on what is known of the actual system being studied. Choice of a wrong model may lead to a quite incorrect understanding of the problem concerned.

References:

Kendall, M.G. 1951 Biometrika 38, pp. 11-25
Lindley, D.V. 1953 Biometrika 40, pp. 47-49