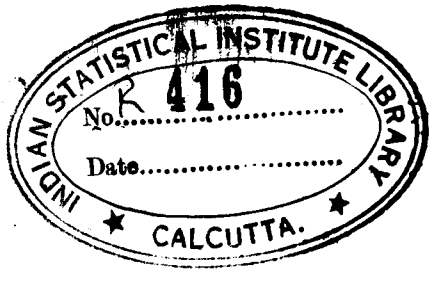
APPLICATION OF STATISTICAL THEORY
IN
TESTING HYPOTHESES

by

WALTER A. SHEWHART



Introductory Comment.

Statistical Method in Research

1. Testing Hypotheses.

2. Estimation

3. Design of Experiment.

INTRODUCTION

We shall start with the assumption that the object of the research worker is to discover the uniformities in observable phenomena so that he will be able to make valid predictions. The scientific procedure involves the making and the testing of hypotheses. The object of the present discussion is to indicate in a general way some of the ways in which statistical theory may be made to contribute in the process of taking and analyzing data in testing hypotheses.

Two of the important kinds of uniformity that the scientist seeks to discover are:

a. The constants of nature.

b. Physical and chemical laws.

For example, tables are available giving the values of literally hundreds of physical and chemical constants and scientific treatises provide us with numerous "laws" of nature.

Of course, these constants and laws must be inferred from measurements. For example, we conceive of the charge on an electron or the velocity of light as being fundamental constants. Likewise we conceive of the standard meter bar as having a constant length. Let X' represent the magnitude of some such constant. It is significant for our discussion that we can never know X' with certainty. Starting with any accepted method of measurement of such a constant the only thing we can observe is a sequence of observations

$$X_1, X_2, \ldots, X_i, \ldots, X_n, X_{n+1}, \ldots, X_{n+j}, \ldots \qquad (1)$$

corresponding to a ~~corresponding~~ sequence of repetitive operations or measurements. Since there is in general no limit to the number of times an operation of measurement can be repeated, the sequence (1) is infinite. In practice the research worker takes a finite number n of observations and from this, forms an estimate of the unknown true value X'.

At this point the statistician may be called upon to answer the question,

A. Given a finite number n of observed values $X_1, X_2, \ldots, X_i, \ldots, X_n$ what is the best estimate $\hat{X}$ of the unknown true value X'.

Let us next consider the simplest problem involved in the search for a law of nature, namely that giving the relationship between two physical variables X' and Y', such as the pressure and volume of a gas at constant temperature. Let us represent the conceived functional relationship f by the equation*

$$Y' = f(X', \lambda'_1, \lambda'_2, \ldots, \lambda'_i, \ldots, \lambda'_m). \qquad (2)$$

Let us represent by X and Y measurements of X' and Y' respectively. Given a method of measuring X' and one for

---

*Note in this case that X' and Y' stand for variables each of which may take on more than one value.

measuring Y' that may be repeated indefinitely, all that is observable about the relationship (2) is an infinite sequence of pairs of values,

$$X_1Y_1,\ X_2Y_2,\ \ldots,\ X_iY_i,\ \ldots,\ X_nY_n,\ X_{n+1},\ Y_{n+1},\ \ldots,\ X_{n+j}\ Y_{n+j}\ldots \quad (3)$$

Again the statistician may be called upon, this time to answer the question

A_1. Given a finite number n of observed pairs of measurements XY, what is the best choice $\hat{f}$ of the functional relationship f and the best estimates $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_i, \ldots, \hat{\lambda}_m$ of the parameters in this relationship?

In certain instances, of course, the functional relationship may be assumed to be known and in this case the problem reduces to that of making the best estimates of the parameters.

The question now arises, what is meant by "best" in both A and A_1? We might frame the answer in terms of some measure of the closeness of our estimate to objective true physical constants or laws as the case may be. If we do frame our answers in this way, there is no way of checking our conclusions in an operationally definite way for the obvious and simple reason that we can never know for sure what the true constant or the true law is in a specific case. What approach is there open to us under such conditions? To get at an answer let us confine our attention to the simpler of the two problems

A and A_1, namely to that of estimating the true value X' from a finite number n of observed values taken by some one method of measurement such as that corresponding to the sequence (1).

So far as I see, any finite set of n observed values after they are taken become, as it were, past history. These are useful in so far as scientific prediction is concerned only in so far as they enable us to make valid predictions about one or more of the characteristics of some of the as yet unexamined parts of the infinite sequence (1). As a scientific basis for any prediction, we must assume: a) that the infinite sequence (1) exhibits some specific kind of scientific uniformity and b) that there exists some rule of prediction making use of the n previously observed values that will in the long run of similar cases give the maximum number of valid predictions, provided the sequence exhibits the kind of uniformity assumed. Schematically we have the situation shown in (4).

$$X_1, X_2, \ldots, X_i, \ldots, X_n \;\Big|\; X_{n+1}, \ldots, X_{n+j}, \ldots \qquad (4)$$

Past | Present | Future

Starting with the n observed values to the left of the line marking the present what kinds of valid predictions can be made about the numbers that may be expected to occur on the

right and how shall we make these predictions? As already noted our answer to such a question depends upon the assumption or hypothesis that we adopt about the uniformity of the infinite sequence. From this viewpoint the sample of n may be used as a test of the hypothesis chosen.

The hypothesis chosen involves two choices:

a) The assumed distribution function

$$dy = f(X_1 \lambda_1', \lambda_2', \ldots, \lambda_i', \ldots, \lambda_m') \, dX \qquad (5)$$

where dy is the frequency of occurrence in the interval $X \pm 1/2\, dx$ and f is the functional relationship here assumed to be continuous.

b) The order in the sequence shows a particular kind of uniformity called random.

The sample of n may then be used to test any hypothesis of this character.

Likewise we may start with an assumption of an assumed relationship such as (2) and ask whether or not a given set of n pairs of observed values of X and Y may reasonably be expected to have arisen if the errors in respect to each parameter happen at random. This problem is beyond the scope of the present discussion although the methods herein considered may be extended to cover this more complicated problem.

TESTING STATISTICAL HYPOTHESES - SIMPLEST CASE

General

Given Hypothesis H_o $\begin{cases} 1. \text{ Sample from universe (5) with assumed form f and assumed values of parameters } \lambda_1', \lambda_2', \ldots, \lambda_i', \ldots, \lambda_m'. \\ 2. \text{ Sample drawn at random.} \end{cases}$

and (6)

Sample $X_1, X_2, \ldots, X_i, \ldots, X_n.$

Problem: The problem is to determine whether the observed sample is likely to have arisen in the process of drawing samples of n from the assumed universe.

The method of attack is to define one or more statistics $\theta_1, \theta_2, \ldots, \theta_j, \ldots, \theta_r$ of the sample where by definition

$$\theta_j = \varphi_j (X_1, X_2, \ldots, X, \ldots, X_n) \tag{7}$$

and then determine the distribution of this statistic

$$dy_{\theta_j} = f_{\theta_j} (\theta_j)\, d\theta_j. \tag{8}$$

Let us assume as the simplest case that the test is to be based on only one statistic θ, and that equation (8) may be integrated between any desired limits. Then equation (8) provides a means of computing the chance P of getting as large or larger value of θ than that given by the observed sample.

We then adopt a rule of action by which we agree to reject the hypothesis H_o.

Example 1

Given H_o { 1) $dy = \frac{1}{\sigma'\sqrt{2\pi}} e^{-\frac{(X-\bar{X}')^2}{2\sigma'}} dx$ where $\bar{X}' = 0$ and $\sigma' = 1$

2) Sample of n drawn at random.

and Sample 1.6, 2.0, 1.3, 1.9 $\quad \bar{X} = 1.7 \quad \bar{X} > 1.16$

Rule If the probability P of getting as large or larger average than that observed $P \leq P_1 = .01$, reject the hypothesis.

Solution Determine the distribution function for averages of $dy_{\bar{X}} = f_{\bar{X}}(\bar{X})\, d\bar{X}$ samples of size n drawn from the assumed universe. The solid line in Fig. 1 gives the theoretical distribution for samples of n = 4. The solid points show how closely the distribution of an observed set of 1000 averages of 4 follow this curve and provide empirical confirmation of the theory

Fig. 1 - Distribution of ...

Now the given sample has an average of 1.7 and computation shows that the area of the solid curve to the right of this value is less than .01. Hence upon the basis of the adopted rule, the hypothesis H_o is to be rejected.

Example 2

Hypothesis H_o same as in Example 1.

Given the Sample of 4 $\begin{matrix} -2.9 \\ 1.1 \\ -.2 \\ 2.3 \end{matrix}$ $\sigma = 1.932 > 1.8$

Rule If the probability P of getting as large or larger root mean square or standard deviation σ than that observed is $P \leq P_1 = .01$, reject the hypothesis.

Solution Determine the distribution function

$$dy_\sigma = f_\sigma(\sigma)\, d\sigma \qquad (8.2)$$

for standard deviations of samples of 4 drawn from the assumed universe. The solid line in Fig. 2 gives the theoretical distribution for

Fig. 2 - Distribution of σ's.

samples of 4 and the empirical check with theory as revealed by the distribution of standard deviations in 1000 samples of four. It will be observed that the observed $\sigma = 1.932$ lies well out on the distribution curve and well beyond the point corresponding $P \leq P_1 = .01$. Hence by the rule indicated above, the hypothesis H_o is to be rejected.

Example 3

H_o (Given an hypothesis H_o that is the same as that for the previous two examples except that σ is no longer assumed to be known.

Given the Sample of 4 .6, 1.4, 1.4, .8

$\bar{X} = 1.05$

$\sigma = .357$

$Z = \frac{\bar{X}}{\sigma} = 2.950 > 2.62$

Rule If the probability P of getting as large or larger value of Z than that observed is $P \leq P_1 = .01$, reject the hypothesis.

Solution Determine the distribution function

$$dy_z = f_z(Z)\, dZ \qquad (8.3)$$

for the ratio $Z = \frac{\bar{X}}{\sigma}$ for samples of size n drawn from the assumed universe. Fig. 3 gives the theoretical curve and the empirical check for the distribution of 1000 values for as

Fig. 3 - Distribution of z

many samples of 4. It also shows for purposes of contrast the normal error curve that we get under Example 4 when σ' is assumed to be known.

Since the observed Z = 2.950 is out on the error curve beyond the point $P \leq P_1 = .01$, we must upon the basis of our adopted rule reject the hypothesis H_o.

Some General Remarks

It will be noted that the development of tests of hypotheses depends upon the mathematical development of corresponding distribution functions for statistics of samples of size n. For example, the tests in the three examples above date respectively from the time of Gauss and La Place for the 1st, to 1890 for the second and 1908 for the 3rd.

It should also be noted that samples of any size n may be used in the above cases for testing the specified hypotheses. In fact a test by means of large samples is no better in this sense than the test by a small sample. This fact is illustrated in Fig. 4. Making use of the Z test used

Fig. 4

in example 3 given above, we may plot the range about the observed average within which the true value X' may be expected to fall for any given probability P of repetitive trials and for any size n of sample. In Fig. 4 the 1st hundred ranges are for samples of 4; the second forty are for samples of 40,

Operation of Control

Hypotheses against the acceptance of which we may wish to protect ourselves may not be simply random ones.

Important Dates

Normal Law - Gauss LaPlace

χ^2 - 1900 Large Sample.

z - Student 1908 - Beers.

{Consumer & Producer Risk 1924
{Control.

Neyman Pearson 1928

and the remaining four are for samples of 1000. Theoretically 50% of each group of ranges should cut the X = 0 line. The observed proportions are .51, .45, and .50.

Errors of the First Kind

Now we are in a position to examine more closely the significance of the probability P_1 in each of the three examples shown above. The choice of $P_1 = .01$ is of course arbitrary and any other value between 0 and 1 could have been chosen. We should note that P_1 is the probability of rejecting the hypothesis when true and we seldom want to do that in a large percentage of cases or as we say P_1 is the probability of committing an error of the 1st kind. Consequently P_1 is often arbitrarily chosen as either .01 or .05.

It follows from what has been said above that one may use small samples just as well as large to test the hypotheses introduced above in the case of Examples 1, 2, and 3. Resulting from this situation many modern theoretical statisticians have emphasized the importance of small samples, and many applied statisticians have gone gayly on using small samples and computing what have been termed errors of the first kind. Some have taken this result as the long sought for answer to the question, how large a sample shall one take? In fact, $P_1 = .01$ and $P_1 = .05$ are usually called the 1% and 5% levels of significance.

TESTING SIGNIFICANCE OF OBSERVED DIFFERENCES (Up to 1908)

Again and again in scientific research, we have two or more samples and the question is raised as to whether they are "significantly" different. We shall confine our attention to the case of two samples of n_1 and n_2 observed values of some variable X,

$$\begin{array}{ll} X_{11}, X_{12}, \cdots, X_{1i}, \cdots, X_{1n_1} & \text{Sample 1} \\ X_{21}, X_{22}, \cdots, X_{22}, \cdots, X_{2n_2} & \text{Sample 2} \end{array} \qquad (9)$$

For example the n_1 measurements may have been taken on one day and the set of n_2, on another day; both sets might have been taken on the same day but by different observers or different measuring equipments; or the two sets may represent the effects of two kinds of treatment on the quality characteristic X of a given kind of material. The list of examples might be extended indefinitely.

The problem here involved is fundamentally one of testing a specified hypothesis Ho, as, for example, the hypothesis that the two samples came from the same specified universe of chance causes or the hypothesis that they came from different specified universes.

The customary method of attack is to assume that both samples came from a normal universe and then to find the distribution function of $\delta = \bar{X}_1 - \bar{X}_2$ for samples of size n_1 and n_2. For convenience, let us assume that $n_1 = n_2$. Then we may write

$$dy_\delta = f_\delta(\delta)\, d\delta \,. \qquad (8.4)$$

Two Notes on Table 1.

In testing difference between samples 1 and 2 of Table 1.

Effect of Design Difference not significant on $P_1 = 0$ level if samples are taken as independent

Kinds of Tests

a) $z = \frac{1}{2} \log_e \frac{\sigma_1^2}{\sigma_2^2}$

This assumes $\sigma_1' = \sigma_2' = \sigma'$

z is not significantly different.

b) L Test.

This is independent of whether or not $\sigma_1' = \sigma_2' = \sigma'$.

Observed L is not significant.

This is a distribution function of the type (8) and subject to the same limitations as were imposed on (8) we may use (8.4) in computing the error P_1 of the first kind for certain rules of rejecting specified hypotheses.

Of course, the distribution function f_δ depends upon the hypothesis made about the universes. For example, if we assume that the universes are normal with standard deviations σ_1' and σ_2' respectively, it has been known since the time of LaPlace and Gauss that f_δ was normal with a standard deviation $\sigma' = \sqrt{\frac{(\sigma_1')^2}{n_1} + \frac{(\sigma_2')^2}{n_2}}$. In 1908 "Student" working in the research laboratory of the Guinness Brewery succeeded in finding f_δ for the case where the standard deviations σ_1' and σ_2' of the two normal universes are assumed to be equal but unknown.

Table 1 presents the classic example used by "Student" in his article of 1908. Two treatments were given each of 10 individuals. The question discussed in the literature is: Does treatment #2 give an effect that is significantly greater than treatment #1?

Table 1

Trial Number	Treatment #1	Treatment #2	Difference (#2 - #1)
1	0.7	1.9	1.2
2	-1.6	0.8	2.4
3	-1.2	1.1	1.3
4	-1.2	0.1	1.3
5	-.1	-0.1	0.0
6	3.4	4.4	1.0
7	3.7	5.5	1.8
8	0.8	1.6	0.8
9	0.0	4.6	4.6
10	2.0	3.4	1.4
Average X	$\bar{X}_1 = 0.75$	$\bar{X}_2 = 2.33$	$\bar{X}_2 - \bar{X}_1 = 1.58$
St nd rd	$\sigma_1 = 1.70$	$\sigma_2 = 1.90$	$\sigma_\delta = 1.17$

Calling $z = \frac{\bar{X}_2 - \bar{X}_1}{\sigma_\delta}$ $\frac{1.58}{1.17} =$, we can find the probability P of getting as large or larger value of z than that observed by the same method as used in example 3 of the previous section. If we adopt the rule of rejecting the hypothesis that the two samples came from the same universe with unknown standard deviation whenever $P \leq P_1 = .01$, then we would reject the hypothesis in this case.

Importance of Sample Size

It is of interest to see what the conclusion would have been if instead of using the whole 10 individuals, a sample of only the first 4, 5, 6, 7, 8 and 9 had been used. The last column of Table 2 shows that the answer would have been the same for samples of 7, 8, 9 and 10, because in each case $P < P_1 = .01$.

Table 2

Sample Size	$\bar{X}$	P
4	1.550	.013
5	1.240	.035
6	1.200	.013
7	1.286	<.01
8	1.225	<.01
9	1.600	<.01
10	1.580	<.01

Here as in the previous section one gets the impression that a small sample is just as good as a large sample to detect a significant difference as it is often called. There is, however, a fly in the ointment as has already been hinted. Let us therefore examine the meaning of significance.

Significant Differences

Broadly speaking, there are five kinds of significant differences between universes that may be of scientific and engineering interest. These are for any given parameter λ_i' of the universe:

1. Any difference $|\lambda_{i1}' - \lambda_{i2}'| = \Delta' > 0$.
2. Any difference Δ' large enough to be discovered at reasonable cost.
3. Any difference Δ' such as to indicate the presence of an assignable or findable cause.
4. Any difference Δ' large enough to be sensed.
5. Any difference Δ' large enough to be of economic significance in that it makes one universe more valuable than another.

Just so soon as we think of significant differences in this way it is obvious that we must be careful in choosing the test for rejecting an hypothesis to see that it gives us what we want in a specific case. This was realized as early as 1924 by members of the Laboratories staff in their development of inspection sampling plans and the operation of controlling the quality of product.

ERRORS OF THE SECOND KIND IN TESTING HYPOTHESES 1926 - 1939

In using any rule for rejecting a hypothesis Ho, we may make two kinds of errors:

1. We may reject the hypothesis H_o when in fact it is true.
2. We may accept the hypothesis H_o when in fact it is false.

These were called, respectively, errors of the 1st and 2nd kind by Neyman and Pearson in their first fundamental paper on testing hypotheses in 1928. In our own work beginning with about 1924 they had been thought of as consumer and producer risks. An extensive literature of many hundred pages has grown up since 1928 all of which is as yet contained only in original memoirs. All that we can do here is to indicate how the consideration of these two kinds of error influence our choice of the number of observations to be made in testing an hypothesis and also influences our choice of statistical hypothesis to be tested.

Sample Size Often Determined by Errors of 2nd Kind

Let us start with the two treatment samples of columns 2 and 3 of Table 1. Let us test the following hypothesis H_o;

H_o (a) Both samples drawn from the same normal universe with a standard deviation $\sigma' = 1.8$.

(b) Both samples are drawn at random.

Under these conditions the distribution of the difference $\bar{X}_1 - \bar{X}_2 = \delta$ for samples of size n is normal with standard deviation

$$\sigma'_\delta = \sqrt{2}\,\frac{\sigma'}{\sqrt{n}} = \frac{(1.4142)\ 1.8}{n} =$$

Let us assume that we wish to make the probability P_1 of errors of the 1st kind equal to the probability P_2 of errors of the second kind. Also let us assume that the economically significant error is 2. How large a sample must we take in applying our rule of rejecting so as not to make P_2 greater than let us say .01?

Under the conditions here assumed, it may be shown that in order for $P_1 = P_2 = .01$ we must have

$$\Delta' = \bar{X}'_1 - \bar{X}'_2 = 4.65\ \sigma'_\delta = \frac{11.83}{\sqrt{n}}$$

In this case Δ' is to be 2. Hence we have

$$2 = \frac{11.83}{\sqrt{n}} \quad \text{or} \quad n = 34.9$$

That is to say, there is a certain size sample necessary to reduce to not over .01 the probability of making an error of the second kind greater than the difference Δ' assumed to be economically significant.

We may, of course, use the relation $\Delta' = \frac{11.83}{\sqrt{n}}$ to see how Δ' depends on n. A few values are given in Table 3 below.

n	Δ'
1	11.83
9	3.90
16	2.96
25	2.36
36	1.97
49	1.69

In Fig. 5 is pictured a case corresponding to probabilities $P_1 = .01$ and $P_2 = .25$ of errors of the first and second kinds.

Fig. 5

Papers Presented - 1936-37

1. Some Comments on the Practical Significance of Tests for Significance - Cowles Commission July, 1936.

2. Use of Laws of Chance in Industrial Development Colorado Springs, Colorado College, July, 1936.

3. Collection, Compilation, and Publication of Statistics with Particular Reference to International Use - Discussion before World Power Conference, September, 1936. (Filed in Correspondence Folder)

4. Discussion of Roop's paper before Society of Naval Architects and Marine Engineers - November, 1936. (Filed in Correspondence Folder)

5. The Application of Statistical Methods to Manufacturing Problems - Franklin Institute, February, 1937. (I.E.B. 6).

6. Accuracy and Precision - A.S.T.M. Round Table Discussion, June, 1937. Filed in Correspondence Folder.