SOME COMMENTS
ON THE
PRACTICAL SIGNIFICANCE OF TESTS FOR SIGNIFICANCE

BY

W. A. SHEWHART

INTRODUCTION

Can you imagine an experimental scientist who is not interested in the significance of his results? Can you imagine one who would not jump at the chance to make use of any generally accepted measure [illegible] of such significance? Witness for example the long history of the theory of errors and its application in many fields of science. This theory provided measures of accuracy and precision and criteria for the rejection of observations. These were applied to all kinds of data as "measures of significance".

But all scientists did not accept the theory and its applications. For example, no less a light than the late Lord Raliegh once said that the theory of error is something good to read and then forget. More recently Millikan[1] has expressed emphatically his belief that, in measuring the error of the charge on an electron, his graphical method of getting at an estimate of uncertainty ~~was~~ is better than that derived from the method of least squares. Other scientists of equal prominence have either expressed similar views or ignored the theory of errors. They have failed to attach much [illegible] significance to such tests for significance.

[1] Millikan, R.A., "Most Probable 1930 Values of the Electron and Related Constants", The Physical Review, Vol.35, 2nd series, #10, pp. 1231-1237.

Then, shortly after the beginning of this century, a new school of statisticians sprung up and what did they do? One of the first things was to criticize the early work on error theory. In many instances they threw out, as it were, the old and substituted new tests for significance. That is to say, they considered the significance of the early tests of significance and found them wanting as had some scientists, but for a different reason. Now, how have the new tests been accepted by natural scientists and engineers? In many cases these men have ignored the new tests and in others they have spoken out against their use as for example, Norman Campbell who says: "It is as certain as anything can be that the great majority of us will fail to use the weapons they (modern statistical methods) offer; it is equally certain that this great majority will include almost all of the great masters of the experimental art."[1] I know many successful physicists, chemists, and engineers who would ~~say~~ chant "Amen" to this statement.

How is it today in the field of statistics? Well, we have significance, statistical significance, levels of significance, measures of reliability, confidence limits, confidence coefficients, measures of ignorance, measures of degree of belief, fiducial limits, and probability limits. The statistician talks about probability, a priori

[1] "The Statistical Theory of Errors", Proc. of the Physical Society, Vol. 47, pp.800-809, 1935.

1. Viewpoint of Shop in about 1925

2. Particular characteristic of our work.

(1) Trying to do the same thing again and again. Hence any prediction is tested again and again.

(2) ~~Statements~~, about significance or prediction must be definite OPERATIONLY.

THE EFFECT OF STATISTICAL METHODS ON MOST OF US and a posteriori, mathematical probability, statistical probability and fiducial probability. Add to these the older concepts of accuracy and precision, and one gets a picture of the ~~Babel~~ babble of tongues that tends to confound the practical man. This ~~Babel~~ babble is none the less confusing when one finds that the theoretical statisticians seemingly do not all agree among themselves!

With this background in mind, I wish to make a few comments on the practical significance of tests for significance. The theoretical statistician as a rule offers his wares to the practical man with the remark that these are but tools to be used with common sense and judgment. What I have to say may indicate that the common sense and judgment required of the practical man is, at least in certain instances, of an uncommon variety.

It ~~is perhaps~~ should be needless for me to say that I am an enthusiastic believer in the usefulness of tests for significance as they are called. In fact there are [illegible] kinds [illegible] There is, however, quite a difference between "significance" in the sense that such tests furnish a measure or level of such significance and the practical significance of such test results. Those of us in the practical field must try to bridge this gap. My comments are directed to this end. My approach is from the viewpoint of the practical man.

ILLUSTRATIVE STATEMENTS ABOUT SIGNIFICANCE

Problem 1

To get started let us consider a practical problem. Let us take one that is now almost famous in the so-called theory of small samples - one first discussed by "Student" in 1908. An experiment on the effect of the optical isomers of hyoscyamine hydrobromide in producing sleep gave the results shown in Table 1. The question is: Is there a significant difference between the effects of the two drugs?

TABLE I

Additional hours of Sleep gained by the Use of Hyoscyamine Hydrobromine

Patient	1(Dextro)	2(Lacvo)	Difference(2-1)=X
1	+0.7	+1.9	+1.2
2	-1.6	+0.8	+2.4
3	-0.2	+1.1	+1.3
4	-1.2	+0.1	+1.3
5	-0.1	-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
Mean	+7.5	+2.33	+1.58

In discussing this set of data Fisher[3] says in effect: Calculate for the differences

[3] Fisher, R.A., Statistical Methods for Research Workers, 4th Edition Oliver and Boyd, London, pp. 112-113, 1932. Fisher treats this problem under the heading - Significance of mean of a small sample - emphasis being placed on small

$$\bar{X} = \frac{1}{n} \Sigma X$$

$$\frac{s^2}{n} = \frac{1}{n(n-1)} \Sigma (X-\bar{X})^2$$

$$t = \bar{X} \cancel{\sqrt{\frac{n}{s^2}}} \sqrt{\frac{n}{s^2}}$$

With t thus found enter a certain table of (n-1) " degrees of freedom " and P and if the P thus found is less than .01, then the difference is significant. Applying this test to the ten differences, he gets a value t = 4.06, and since for this value of t and for n-1 = 9, the corresponding value of P is less than .01, he concludes: "The difference between the results is clearly significant".

Now suppose the research man getting the original data had been inclined to be a little lazy - I say just suppose that there were such a man and that he had gotten tired and stopped after making only seven measurements. Suppose he had applied the same test for significance to his 7 measurements. Lo and behold, he would have found P < .01, and from the viewpoint of the test, could have concluded that the difference was clearly significant.

Table 2 shows the results of similar computations for samples of 4, 5, 6, 7 - 10.

TABLE 2

.0675
.3125
.015625
[illegible]
[illegible]
.9[illegible]5
[illegible]
.97[illegible]5625

Sample Size	$\overline{X}$	t' for P=.01	$\frac{p}{\sqrt{n}}$	$t'\frac{p}{\sqrt{n}}$	t	P
4	1.550	5.841	.2843	1.66	5.45	.013
5	1.240	4.604	.3803	1.75	3.26	.035
6	1.200	4.032	.3130	1.26	3.83	.013
7	1.286	3.707	.2781	1.03	4.62	<.01
8	1.225	3.499	.2484	.87	4.93	<.01
9	1.600	3.355	.4343	1.46	3.68	<.01
10	1.580	3.250	.3890	1.26	4.06	<.01

What happens when a practical man - busy as he can be with the details of his problem - is introduced to such results. He may ask himself, why go to 10 measurements when 7 will do, at least in such a case. Usually he is a little cautious and may examine the discussion of such problems in statistical texts. Having, however, been informed through such an investigation that the tests for small samples are just as rigorous under conditions where they apply as are similar tests for large samples, will he have the common sense to question further whether a test that is significant on the basis of a sample of a thousand or even a million is any more significant than one that is significant on the basis of a sample of four? I am afraid that such common sense is a little too uncommon. The trouble here is perhaps that the term "significance" does not mean the same in the mind of the practical man as it does as used in the test.

Problem 2

"Student" as we know, gave a table for the value of the probability that the means of a sample of n, drawn at random from a population following the normal law, will not exceed in the algebraic sense the mean of that population by more than z times the standard deviation of the sample. Knowing this, it follows that we can write down for any such sample a range

$$\overline{X} \pm \Delta X$$

for which the probability P is any previously specified value that the range thus written down will include the mean of the population. Suppose I take one sample at random from each of two normal populations and get ranges

123.350 ± .040 for 1st sample

and 123.0445 ± .0428 for 2nd sample

Under the assumptions, is the true average of the first universe any more or less likely to lie within the range 123.350 ± .040 than the true average of the second universe to lie within its range 123.0445 ± .0428? You answer no because the probability is the same in both cases. Is your answer changed if I tell you that in the first case the sample size was 4 and in the second case 1000? Again I believe you must answer no in order to be correct. Of all the facts that I have come to know about the theory of sampling, none has been more impressive than this one.

For example, this means that if I had before me a series of N bowls containing normal universes but with unknown parameters and if I could draw a random sample of size n from each, I could then set up a range for each bowl corresponding to any given probability P. Just think I could set up a range for each bowl with a sample of 4 from each with just as good reason to expect that PN of the averages of the universes within the bowls would lie within these ranges as if the sample size had been as large as one cared to make it.

Fig. 1 shows for example, 100 such ranges for samples of 4 drawn from the same bowl in which the average of the universe (distribution of numbers on the 998 chips) was zero. Please note that the zero line is cut 52 times out of a possible 100 - pretty close indeed to 50%. Now, what would happen if we had taken samples of 1000 instead of 4? Fig. 2 shows the results of 4 samples - I got tired and quit at 4000 drawings. Lady luck gave me just a 50/50 break! Obviously in this case, I didn't do a better job with a sample of 1000 than with one of 4. Of course, I do not offer these results to justify (they are too few in number to do that) but to illustrate the theory.

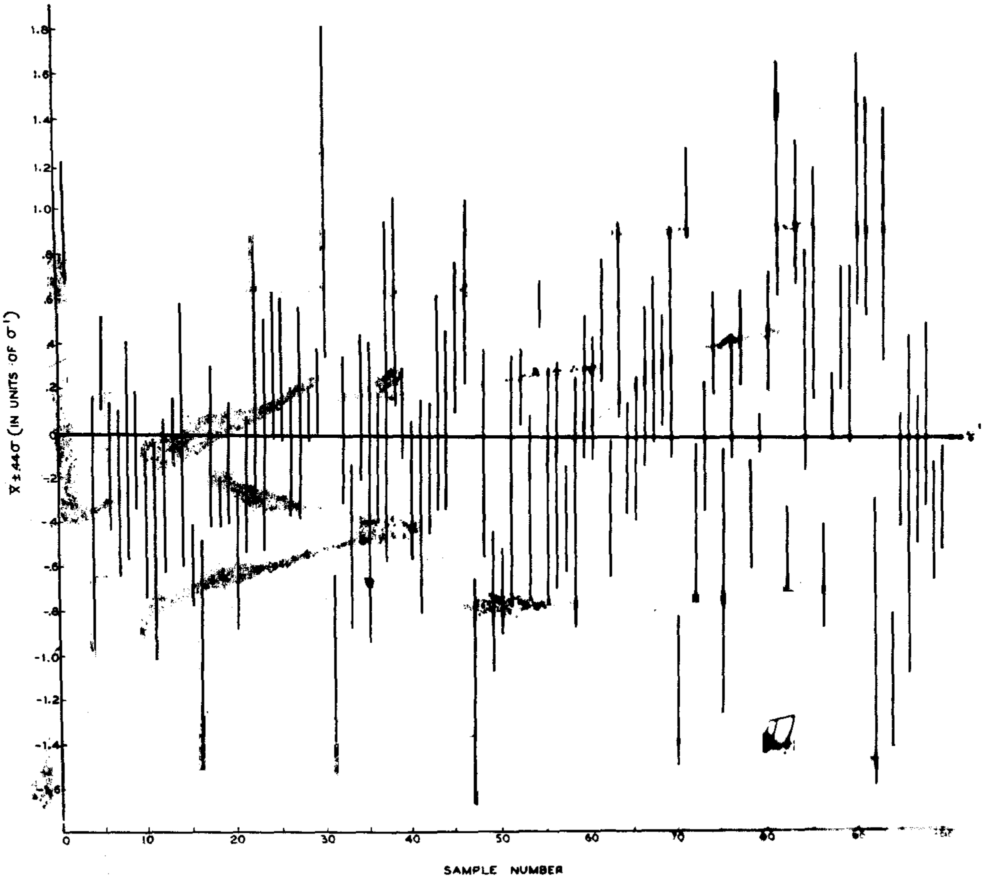

X̄ ± .44σ (IN UNITS OF σ')
1.8
1.6
1.4
1.2
1.0
.8
.6
.4
.2
0
-.2
-.4
-.6
-.8
-1.0
-1.2
-1.4
-1.6
0
10
20
30
40
50
60
70
80
90
SAMPLE NUMBER

Fig. 1

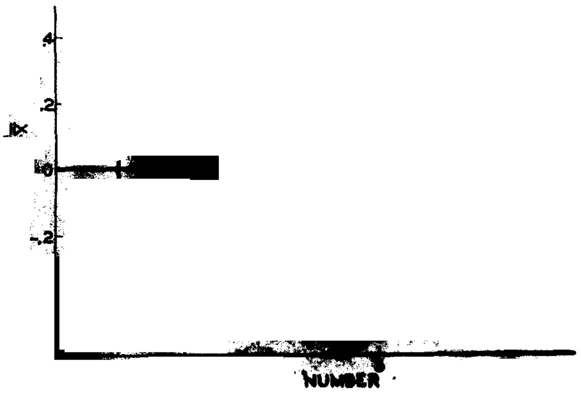

.4
.2
X̄
0
-.2
NUMBER

Fig. 2

Now, let me ask: Is the result obtained from the sample of 4 just as significant as that obtained from the sample of 1000? If it is, why can't we use small samples, so long as we use correct small sample theory? Think how important the answer to this question is from the viewpoint of determining how large a sample to take.

The answer to the question about significance obviously depends upon the meaning to be attached to the term. Let us therefore see how such results are sometimes used. In a recent paper by Eddington[4], he puts the question: Suppose I have occasion to use Planck's constant and I find in reference books two determinations

$$h \times 10^{27} = 6.551 \pm .013 = \bar{X}_1 \pm \Delta X_1$$

$$h \times 10^{27} = 6.547 \pm .008 = \bar{X}_2 \pm \Delta X_2$$

Assuming that these are to be taken at their face value which one should be chosen? He argues that the latter is the more useful to him because <u>it limits h to a narrower range</u>, and hence will lead to sharper conclusions. This, today, from a great astronomer is typical of how many have interpreted the interval

$$,\bar{X} \pm \Delta X$$

[4] Eddington, A.S. "Notes on the Methods of Least Squares", Proceedings of the Physical Society, Vol.45, Part 2, No.247, pp. 271-282, 1933.

Ever since the origin of the theory of errors, that is to say, they look at the <u>magnitude</u> of the ΔX, <u>without</u> raising any question as to the size n of the sample from which ΔX was determined, as fixing the range in which the unknown true value is to be assumed to lie. Now, of course, the small sample may in a single trial as in the case of the previously referred to drawings from a normal distribution give a smaller ΔX than a large sample from the same universe. Certainly, however, my "common sense" suggests that a ΔX from a small random sample is less significant in the long run than one from a large random sample when considered from the viewpoint of establishing the range to be used in practice.

The situation thus revealed is that if the world were filled with normal universes that could be sampled at random and all we wanted to do was to set up ranges for these such that the probability of the true averages lying within the ranges thus set up is any value P, this could be done with small samples just as well as with large. But if we are interested in the <u>precision</u> (width of ranges) with which we do this, then we must give attention to sample size. There is no lazy man's road to this goal. Tyng

Going back now to the argument that the second value of h is the one to be used because .008 is less than .013, we see that this attaches equal significance to the two ranges.

[illegible]

Even under ideal conditions of random sampling where the objective distributions of the errors for both determinations are the same, this would not be justified unless the sizes of the samples were the same. Such is the nature of the evidence that the ~~hand~~ brand of common sense often used in judging the significance of certain tests for significance needs to be modified.

Without more ado, let us try to approach in a fundamental way the problem of setting up ways of judging the practical significance of tests for significance.

THE MEANING OF SIGNIFICANCE

Have you ever sat on the side lines at political arguments over "personal liberty", "personal rights", "due process of law", "constitutional rights", "the common man or Johnny Q Public", or religious arguments about "God", "immortality" or ethical arguments about "the right", "the good", "the better", "the best"? All of us have had such experiences, but how many times have we had such arguments settled once and for all? Contrast with statements about such entities the following statements: The Bell Telephone Laboratories is located at the corner of Bethune and West Streets; iron is heavier than water; the density of water at 40°C is one gram per cubic centimeter. What is the difference between these

The answer is

two classes of statements? One can be verified, the other cannot. Likewise we can talk and argue about significance, for example, in the abstract, or we can talk about it in a verifiable way. By choosing the latter, we can keep our feet on the ground as is fitting for a practical man. The term ~~An example~~ significance is in much the same situation. ~~may help to get across what I have in mind~~. Tell a practical man that a test measures significance. He will-at least if he is cautious and if he has the common sense he is supposed to have in using statistical tools - ask:

1. Significant to whom?
2. Significant for what?

Tell a practical man that a test measures the confidence, ~~or~~ degree of rational belief, degree of ignorance and, if he has the prescribed common sense, he will ask:

1. Whose confidence, ~~or~~ whose rational belief? Whose ignorance
2. Confidence in, / ~~B~~elief in and ignorance of what?
3. How can I verify the significance operationally?

Test for Significance

What does it mean in an operational

From this viewpoint what constitutes a test for significance? Let us look at the Student-Fisher test for significant difference between the observed mean $\bar{X}$ and an assumed true value $\bar{X}'$ of this mean, given a sample of n observed values X_1, X_2, ..., X_i ...., X_n. We are told to

manner?

Bridgman 1928.
C I Lewis

calculate the statistics $\frac{s}{\sqrt{n}}$ and X and the ratio $\frac{(X' - X)}{s}\sqrt{n} = t$. Then using this value of t and n-1, enter a certain table and read a number P. If this number is less than .01 call the difference clearly significant. Obviously the test thus described consists of certain formal rules for operating on the observed data - simple operations too that any one can apply and all will get the same result.

You may wish to call my attention to the fact that in my description, I forgot to state that two provisos are included in the test - (a) the sample must be random and (b) the universe must be normal or at least approximately so. ¶ Yes, you are right. But if you make a mistake in applying the test to a sample of n data and get a certain value P when the sample is not random and the universe not normal you will get the same value of P as though you didn't make the mistake. The test itself without you is insensitive to such mistakes. It is in this sense a pure formal or mathematical rule of operating on a set of numbers that you choose to operate on in this way. ¶ Right here is where the practical man with the right brand of common sense gets a severe shock. The probability P or measure of significance as it is called is just the same for a set of n data when the one making the test doesn't have the common sense to choose to use the test under condition when it is supposed to work as when he does.

The practical man of common sense can't swallow that. Yet all brands of practical significance start with a consideration of the data from the viewpoint of how they were gotten and who got them. Should be thoroughly [illegible]

The practical man is often quite justly disturbed over the fact that the P given by the Student-Fisher test was called probability by Student and is now being called fiducial probability by some statisticians. It has also become a custom among many to refer to P as a level of significance or confidence. Thus we have

P is a probability

P is a fiducial probability

P is a level of significance

P is a measure of [illegible]

P is a measure of confidence

$P < .01$ is clearly significant.

But from the viewpoint of the rule of getting P from the set of data you choose, P is just a number - a number by any other name is a number just the same. The tool that the mathematical statistician gives the practical man in this case is a rule for getting a number P from any set of n numbers (observed values) chosen by the statistician. What is the practical significance of P? That is what must be supplied by common sense.

Practical Significance

Let us consider here the very simple case of testing for significant difference between two samples $X_{11}, X_{12}, \ldots. X_{1n_1}$ and $X_{21}, X_{22}, \ldots. X_{2n_2}$ such as the difference between the effects of two drugs in producing sleep, or the difference in the quality of a given kind of product produced by two machines, or the difference in the effects of two kinds of fertilizers. The Student-Fisher t test gives a test for the significant difference in such averages.

Assuming that the universes from which the samples were drawn have averages $\bar{X}'_1$ and $\bar{X}'_2$ respectively and let

$$\Delta \bar{X}' = \bar{X}'_1 - \bar{X}'_2$$

Now from the two samples we get an observed difference $\Delta\bar{X}$ in sample means:

$$\Delta \bar{X} = \bar{X}_1 - \bar{X}_2$$

Confining our attention first to the true but unknown difference $\Delta \bar{X}'$ there are four questions which a practical man may and usually does ask. They are:

1. Is there a difference?
2. Even if there is a difference is it economically feasible to discover this difference?
3. Even if there is a difference, will it be one that will be sensed in use?

4. Even though there is a true difference that is large enough to be sensed, is it large enough to be of any appreciable value?

For example, in the case of the two sleep producing drugs there is the question as to whether or not there is any real difference between the average effects produced. That difference might of course be in absolute value only a few seconds or even just a fraction of a second. In fact the test as applied by Student and Fisher in the way indicated above simply tests as it were the hypothesis that $\Delta X' = 0$. We might apply the same test to test the hypothesis that $\Delta \bar{X}$ [illegible]

Even if there is a real difference but of only a few seconds let us say, you can imagine what a job it would be to find it and to establish its magnitude with any great
2 degree of assurance. From this viewpoint the question as to whether or not there is a difference even though very small is of more or less academic interest. Certainly if the actual difference in the average effects is only a few
3 seconds or even a few minutes, it would not likely be sensed by those using the drugs. But even if it were sensed in such a case, the difference would not likely be worth very much to the users of the drug at least when compared with the cost of trying to discover the difference and to control it under production.

In considering the significance from the viewpoint of whether or not the difference $\Delta \bar{X}'$ can be sensed, the drug example is perhaps as simple a case as one can conceive. In the majority of cases, the difference is in terms of a physical or chemical measurement whereas the sensory experience is a more or less complicated function of such a measurement. For example, the minimum detectable sound intensity varies with sound frequency and other factors.

Enough has been said to indicate that before we can say anything about the practical significance of the test for significance, we must see what we can infer from it as to the existence of a difference, the feasibility of finding and controlling the difference in future experience, and the magnitude of the difference.

From a study of the results given in Table 1 the practical man might conclude:

1. There is (or is not) a difference in the sleep producing effects of the two drugs.
2. The difference is (or is not) findable and controllable.
3. The difference is (or is not) one that will be sensed by the users.
4. The difference is (or is not) one that will be valued by the user.

determining the answer to practical question

Any one of these is a judgment or probable inference J, based upon evidence Q, a part of which may be the results given by one or more tests for significance. In the inferences as stated, the difference considered as of importance is not simply the differences in the true or expected averages.

To get at the practical significance of tests for significance, we must consider their interpretation.

Interpretation of Test for Significance

what does it mean

First let us consider the difference between the test itself and its interpretation. For the sake of definiteness, let us consider a very simple example. The previous example is a little too complicated to begin with. I have before me two bowls containing circular pasteboard chips like the one I hold in my hand. On each chip there is a number. Now, I draw with replacement a sample of four from Bowl A and now a similar sample from Bowl B. The numbers thus obtained are:

TABLE 3

	Bowl A		Bowl B
	+1.0		-2.1
	- .8		- .5
	+ .3		-2.8
	- .2		-3.4
Mean $\bar{X}_1$ =	+0.75	$\bar{X}_2$ =	-2.2

The practical question to be considered is: Is the true average X_1' of the numbers on the chips in A different from the similar average for B?

Let us apply the Student-Fisher test to these two sets of n=4 data. This consists in calculating from the data in a prescribed way a number t which in this case turns out to be t = 2.650. Then entering a table of 't with the approximate "number of degrees of freedom" we get another number P = .08. This constitutes the formal part of the rule.

Now in applying the test we say that *if* the two samples are drawn at random, *if* the distribution of the numbers in each bowl is normal and *if* $P < .01$, the difference is significant. What does such a statement mean? How, in other words, can it be shown to be either true or false?

If, for example, you look at every chip in each bowl, write down the number found thereon, calculate the true averages $\bar{X}'_1$ and $\bar{X}'_2$ of the numbers in the two bowls and find that $\bar{X}'_1 - \bar{X}'_2 = 0$, would this constitute a verification of what we say when applying the test? Obviously it would verify the statement that there is no difference but I prefer a negative answer to the question as asked for the following reason.

¶ Here is where we try to attach meaning to the number P as a probability. We interpret P as indicating that, if the test is applied to samples taken under the conditions prescribed, then in a *series* of N trials in which samples were thus taken and the P calculated as indicated, in PN of these cases where the actual differences between the averages in the bowls is zero we would get a value of t as great as or greater than that observed. Obviously we couldn't verify this meaning by a single case!

The practical man asks a very simple question, Is $X'_{11} - X'_2 = 0$, and we as statisticians tell him well, if that is so, we may expect to observe a value of $t > 2.650$ under the conditions here assumed about .08 N times in every N trials. Stated in another way this means that if the practical man takes the bull by the horns and concludes under every such condition that the true difference is not zero when $t = 2.650$ he will make a mistake in 8% of the cases when the difference is zero. On the other hand if he concludes that the difference is zero he will run a chance of overlooking a real difference.
Here are two kinds of errors which a practical man must consider in making his [illegible]
We shall return to this later.

[illegible]

Next, however, let us consider the practical significance of the fact that as yet we do not know that the conditions imposed on the one applying the test have been fulfilled. Were the samples drawn at random and are the universes in the bowls at least approximately normal? We as theoretical statisticians may dismiss such question as being somewhat trite. What we say is the gospel truth if the conditions are met. But when the practical man applies the test to a set of observed data and interprets it in a way that is correct if the assumptions are justified, he will be making a mistake if the assumptions are not justified.

This fact is usually considered too trite to be emphasized in texts on statistical theory. We spend a lot of time showing how to calculate a level of significance out to two or three decimal places but we spend comparatively little (if any) time considering the chance that the user of the test will be correct in his judgment of the conditions. ¶For example, if one makes a mistake of this kind on the average of p% of the number N of times he uses a given test, the interpretation of P certainly holds strictly in just (1-p)N cases on the average. Only if p is very small may the practical man justly overlook this source of error. ¶In my limited experience in the field of applied physics and chemistry and engineering, I have found this source of error to be quite large unless used only after enough observations have first been taken and analyzed in a way to detect lack of randomness. In my own work under best conditions this means at least something like 150 observations - something like a 100 in the process of detecting and removing assignable causes and at least 50 which give no evidence of the presence of assignable causes. ¶Note that I didn't say anything about the significance of lack of normality.

E. S. Pearson and others have removed much of the need for worry over this source by showing that a small variation from normality does not affect materially the value of P in the Student-Fisher test. We shall return to this later.

Now, let us summarize what we have found out about the interpretation of a test for a significant difference in universe averages. Two factors are important, the meaning of the test and the knowledge provided by the test, it being noted that the test itself is a formal rule of operating on the observed numbers. (1) The meaning in order to be verifiable must be operational and in the case of the application of a statistical test, the operation is of the nature of a repetition of a simple operation in which an event E is to be observed to happen on the average PN times in N repetitions. (2) Now knowledge about the world in which we live is always of the nature of a probable inference. In so far as it is operationally verifiable, it consists in a certain degree of belief p, based upon specified evidence Q, that an event E will happen if a certain specified operation is carried out. For example, the results given by the test applied to the data in Table 3 constitute certain evidence Q. We must therefore later consider the

The first important thing to note is that we may use this evidence as a basis of predicting more than one kind of event. We have chosen to consider its significance, however, as a basis of predicting whether or not the true difference $\bar{X}_1 - \bar{X}_2$ is zero. We may interpret the test as indicating that in N repetitions of the experiment and test, we could expect to observe a value of t as large as or larger in absolute value than that here observed only .08 N times on the average provided two conditions are met in sampling. Subject to these two conditions, if we conclude that there is a difference whenever t = .08 we shall make a mistake in .08N cases out of N trials when there is no difference. It helps us to estimate a certain kind of error. But it didn't answer the question about ~~Referring to~~ the data in Table 3; ~~what shall we decide.~~ Is the difference $\bar{X}'_1 - \bar{X}'_2$ equal to zero? Now let us consider this question for a moment.

I think most practical men would answer no. If pushed to choose one or the other bowl upon the basis of the data of Table 3, it being advantageous to choose the one with the highest average, I believe they would choose the one which gave the sample with the highest average. They would do this irrespective of the results obtained by applying the test. Would a statistician do differently? I think not, if he is a good statistician, even though the observed difference is not clearly significant! Certainly, at least, the observed value is the most likely in this case.

In what way, therefore, is the test of practical significance? As far as I see it simply gives one a little side information as to how badly one may be fooled in acting as if there were a real difference equal to that observed i.e. +2.275 in at least 8% of the cases you would get a value of t as large as or larger in absolute value than that observed even though the true difference is zero.

Interpretation of Test for Significance - Further Comment

The problem just considered is very simple in character compared with that, for example, of testing the difference in the sleep producing effects of the two drugs. In particular, it is possible to find the true difference in the case of the bowl and also to check the distributions of numbers on the chips, assuming that we can look at all of the chips.[1] It is not feasible to do this in the case of the effects of the two drugs. In this case, if the sample difference $\Delta \bar{X}$ is observed under random conditions the only sense in which there is an objective difference is as a statistical limit L_s as the sample size n approaches infinity -

$$\underset{n \to \infty}{L_s} = \Delta \bar{X}'$$

This is of much importance from the viewpoint of practical significance of differences. For example, in a previous section, we have pointed out four ways in which $\Delta \bar{X}'$ may be practically significant. In this case, there isn't a true difference which is observable. The only kind of observable difference is either, (a) between an observed difference and an assumed true difference or (b) between an average of, let us say, n_1 observations and an average of, let us say, n_2 observations made under presumably the same essential conditions as we say. In either case, the observed average or

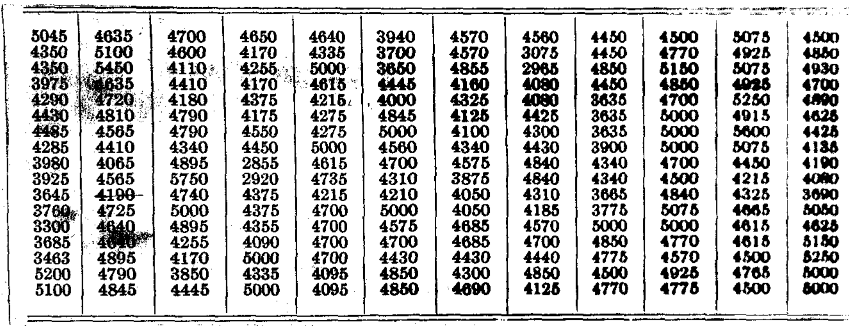
[1] Of course, it should be pointed out that distribution of numbers in the bowl may not constitute the expected distribution of numbers

averages are ~~supposed~~ assumed to be obtained under random or the same essential conditions.

What is mean't here by random? Did the ten patients constitute a random sample? Suppose you say yes and I say no. How may we examine the data to see who is right or at least how can we settle the argument?

Let me illustrate with another set of data, this time, measurements of insulation resistance on a new kind of material. Table 4 gives the results of measurements of insulation resistance on succession of its many sample pieces of material. Do these data constitute a random sample? The man who made them tried to keep his conditions essentially the same. In

<u>TABLE 4</u>



5045	4635	4700	4650	4640	3940	4570	4560	4450	4500	5075	4500
4350	5100	4600	4170	4335	3700	4570	3075	4450	4770	4925	4850
4350	5450	4110	4255	5000	3650	4855	2965	4850	5150	5075	4930
3975	4635	4410	4170	4615	4445	4160	4080	4450	4850	4925	4700
4290	4720	4180	4375	4215	4000	4325	4080	3635	4700	5250	4890
4430	4810	4790	4175	4275	4845	4125	4425	3635	5000	4915	4625
4485	4565	4790	4550	4275	5000	4100	4300	3635	5000	5600	4425
4285	4410	4340	4450	5000	4560	4340	4430	3900	5000	5075	4135
3980	4065	4895	2855	4615	4700	4575	4840	4340	4700	4450	4190
3925	4565	5750	2920	4735	4310	3875	4840	4340	4500	4215	4090
3645	~~4190~~	4740	4375	4215	4210	4050	4310	3665	4840	4325	3690
3760	4725	5000	4375	4700	5000	4050	4185	3775	5075	4665	5050
3300	4640	4895	4355	4700	4575	4685	4570	5000	5000	4615	4625
3685	4640	4255	4090	4700	4700	4685	4700	4850	4770	4615	5150
3463	4895	4170	5000	4700	4430	4430	4440	4775	4570	4500	5250
5200	4790	3850	4335	4095	4850	4300	4850	4500	4925	4765	5000
5100	4845	4445	5000	4095	4850	4690	4125	4770	4775	4500	5000

the table, the order of taking the measurements is from the top of the first column downward and beginning again with the top of the second column and so on. Do these data constitute a random sample? If they do, we may apply a test such as the Student-Fisher t test to test the significance of the difference of the observed mean from any specified value subject to the interpretation already considered. If, however, the sample is not random, the test is not supposed to apply or if applied there is some question as to how the results may be interpreted. Hence it is important to know from practical [illegible] whether or not the data are random.

Early in our work we devised a criterion to apply to such data: Break up the total set of n data into subsamples of 4 taken in the order in which the data were taken. Calculate the grand average and certain limits on either side of this average. Plot the succession of points corresponding to the averages of 4. If all points, when there is at least 25 averages of 4, fall within these limits, then apply tests for significance of differences in the observed mean from some specified value with a clear conscience and an expectation that the interpretation of P under "random" conditions will be found to be justified.

Applying this criterion to the 204 observations, we get the results shown in Fig. 3. There are 8 points outside.

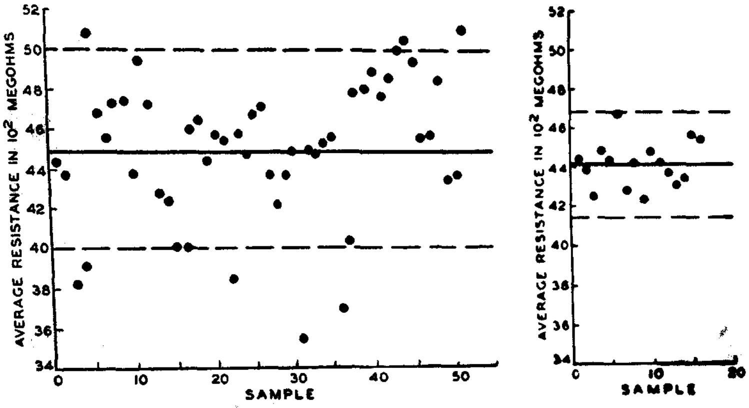


Fig. 3

I wouldn't feel justified in applying the Student-Fisher tests to these data.

What happens if we apply this Control chart test to the 56 values of the charge on an electron as observed by Millikan? Fig. 4 is the answer. All points are inside.

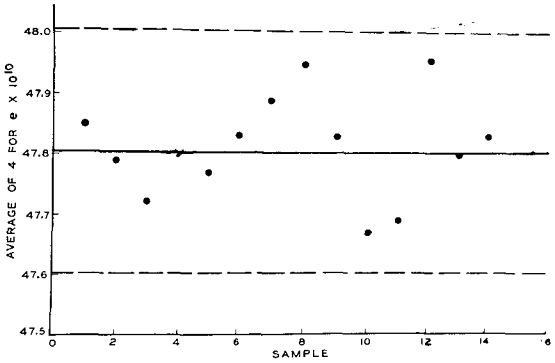
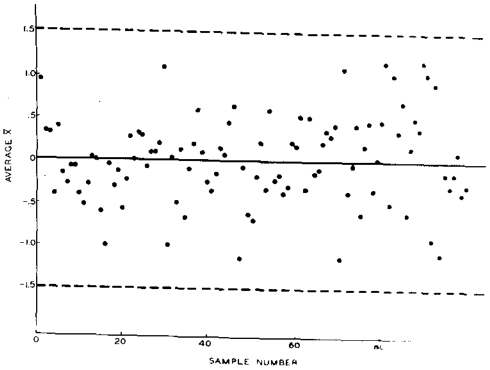


Fig. 4

Likewise what happens when the test is applied to succession of averages of 4 drawn from bowl #1 used above. Fig. 5 is the answer. All points are inside.



#. 15 - 1

Fig. 5

In the case of the Millikan data there are only 14 instead of 25 averages of 4. I would feel pretty safe in applying the test for significance here but in general our experience shows the desirability of going to 25.

May we consider the above criterion a test for randomness? Well yes. We may call it that if we please. From a practical viewpoint, however, it is simply an operation. Personally I do not know what "random" means. My experience has shown me, however, that a test for significance applicable to a "random" sample works in general when applied to data that have been previously shown to satisfy the above criterion. Otherwise it does not. There may exist other operations[1]

[1] Another "test" might be as in the case of a die, a stud observation of its physical structure.

Interpretation of Test for Significance - Slightly Different Case

Let us consider the problem: A large brewery recommends to farmers a special variety of barley, say V_0. A breeder recommends a certain new variety, say V_1, which in his opinion is able to give a larger value of a characteristic X desirable for the brewery. The brewery carries out experiments to test the advantage of the new variety. We are interested from a practical viewpoint on knowing whether or not

$$\bar{X}'_{V_1} - \bar{X}'_{V_0} \leq 0$$

This problem has been recently proposed by Neyman and is typical of many arising in industrial research when we want t compare the new with the old.

We could, of course, compare the twc by taking samples of each but it is perhaps better to consider the comparison of the average of a sample of the new with the accepted (or hypothetical) average for the old. Assume that such a comparison is made and that the observed difference

$$\bar{X} - \bar{X}'_{V_0} = \Delta \bar{X}$$

is found to give a t corresponding to P = .08. Would one change to the new on a basis of such evidence if the sample was of the order of 4 let us say? In the previous case, when we assumed that all we knew about the two averages under comparison was revealed by the sample, it was argued that if forced to a decision it was reasonable to choose the one with the larger average. Would one do the same thing here? I think the answer should be no! One can't afford to throw over the old for the new about which we can judge only from a small sample which tells us so little. I well remember, before I became interested in ~~statistics~~ statistical theory and practice, of an industry, taking a chance of this character based on a sample of n = 13. It turned out that the new was not even as good as the old and in the end they lost over $50,000 in their hasty decision. Possibly it was the unlucky number! However, I have seen other examples which turned out about as bad.

Three Kinds of Errors

Consumer Risk - Error of First Kind

We may look upon the application of the Student-Fisher test in the manner indicated above as a test for a certain kind of error: thus giving the probability of rejecting an hypothesis when true. For example The hypothesis H_1 is that the difference between the true average is not zero, If the true difference is zero, we shall make a mistake on the average of PN times in every N applications of the test. This is an exceedingly useful conception because it makes it possible to calculate the limits within which it is likely that a certain statistic will fall in a series of samples drawn at random from a given universe.

A simple example will serve as illustration. A sample of n units is drawn at random from a very large lot of product in which the proportion of units having non-conforming quality is an unknown value p'. In the sample X = pn individuals are found to be non-conforming. The problem is to obtain limits p_1' and p_2' such that we may expect (1-P)N lots for which p = nx is found to be non-conforming in samples of n to have true values of p' between p_1' and p_2'. Fig. 6 shows[5] such limits for a sample of 10, and a value of P = .05.

Subject to certain approximations, the prediction that $p_1' < p' < p_2'$ will be correct in 95% of the cases met with in a long run of experience and wrong in 5%. In 2.5% because $p' \leq p_1$ and 2.5% because $p \geq p_2$.

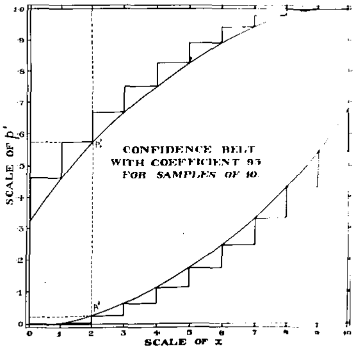


Fig. 6

[5] Clopper, C.J. and Pearson, E.S. "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial", Brimetrika Dec. 1934, pp. 404-413.

In general let us consider a universe

$$df = f(\theta_1', \theta_2', \ldots \theta_m')\, dx$$

Let us take a random sample of n values of X and let us consider a statistic

$$\theta_i = \psi(X_1, X_2, \ldots . X_n)$$

In certain cases the distribution of θ_i in samples of n can be expressed solely as a function of a parameter θ_i' of which θ_i is the estimate found by the method of maximum likelihood. <u>If</u> θ_i is such a statistic of continuous variation and P the probability that θ should be less than any specified value, Fisher[6] has pointed out that we have a relation of the form

$$P = F(\theta_i, \theta_i')$$

A case in point is the distribution of the maximum likelihood estimate θ of the standard deviation $\sigma' = \theta'$ of a normal universe from a sample of size n. Given ~~n~~, we may set up a range

$$k_1 \sigma' < \theta < k_2 \sigma'$$

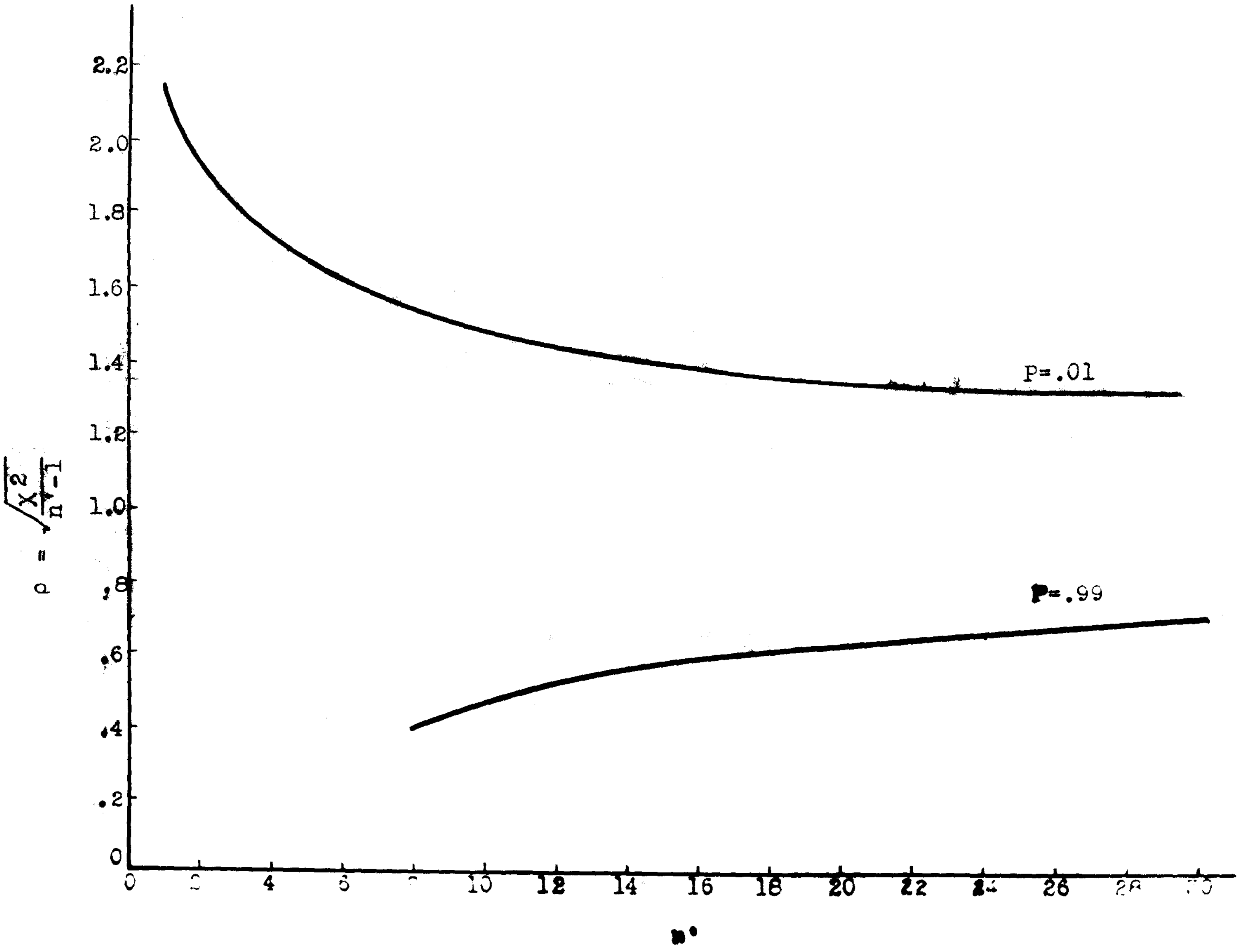
such that for any value σ' the probability of getting a value of θ within this interval is any previously specified value P. Now, if every time in practice that we take a

[6] Fisher, R.A. "Inverse Probability", <u>Proc. of Camb. Phil. Soc.</u> Vol. XXVI, Oct. 1930 pp. 528-535.

sample of n under such conditions, we take the value of θ found from this sample as being σ', we will obviously have P% of our estimates lying within the above range. Fig. 7 gives such ranges for a value P = .98: 1% below $k_1\sigma'$ and 1% above $k_2\sigma'$.

Fig. 7

In our own work, as far back as 1924 we ran into the need of considering an error of this character. In sampling product there is always a chance of rejecting product even though at standard level. That is to say a producer runs a certain risk of having product rejected even though satisfactory. This producer risk is of the nature of an error which has more recently been referred to as the first kind by E. S. Pearson and J. Neyman. The same type of error arises


2.2
2.0
1.8
1.6
1.4
1.2
1.0
.8
.6
.4
.2
0
$\rho = \sqrt{\frac{\chi^2}{n'-1}}$
P=.01
P=.99
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
n'

in the application of the quality control chart in that we run a certain risk of looking for trouble when it is not present.

Errors of the Second Kind and Consumer Risk

If in testing an hypothesis such for example as: The true difference in averages of two universes is zero, we may reject the hypothesis even when true and we may accept the hypothesis when not true. The probability of getting this general type of error which Pearson and Neyman have termed an error of the second kind, we had previously called a consumer risk.

Neyman in two recent papers has discussed in a very interesting manner the error of the second kind in testing Students Hypothesis.[7] We wish here to consider certain aspects of this kind of a test.

Let us consider two normal universes with true means $\bar{X}'_1$ and $\bar{X}'_2$ respectively. Let $\Delta \bar{X}' = \bar{X}'_2 - \bar{X}'_1$, and let us consider the hypothesis

$$H \text{ that } \bar{X}'_2 \leq \bar{X}'_1$$

The Student-Fisher test gives us a method of testing this hypothesis. In general we reject the hypothesis H when the observed

[7] Neyman, J. Loc. cit. and Statistical Problems in Agricultural Experimentation. Supplement to the Journal of the Royal Statistical Society pp. 107-180, Vol. II #2, 1935.

value of

$$t = \frac{\Delta \bar{X}'}{S} \geq t_P$$

where t_P is usually taken as either .01 or .05, and we accept it in other cases. We can get t_P from the Fisher table by taking twice the probability associated with a given value of t. Under such conditions, we have:

1. Errors of the first kind: Those that consist of rejecting the hypothesis H when it is in fact true.
2. Errors of the second kind: Those that consist in accepting the hypothesis H when it is in fact false.

To illustrate suppose we make t_P such that the probability P_1 of getting a value of $t \geq t_P$ is .01. This means that the probability of rejecting the hypothesis H, if in fact it is true or $\bar{X}'_2 \leq \bar{X}'_1$ is such that

$$P_1 \leq .01$$

Under these conditions Neyman has given tables which make it possible to answer the questions:

1. What is the probability P_2 of accepting the hypothesis H when the true value $\bar{X}'_2$ is

$$\bar{X}'_2 = \bar{X}'_1 + K\,\sigma'_{\Delta\bar{X}'},$$

and

2. What is the standardized size $K = \frac{\Delta \bar{X}'}{\sigma'_{\Delta\bar{X}}}$, of the difference $\Delta\bar{X}'$ between the true population mean

and the hypothetical limit which will be undetected with the given frequency P_2.

For example, Fig. 8 reproduces Neyman's curves showing the dependence of the probability P_2 of the second kind of errors on the ratio K and the number of degrees of freedom n' when P_1 = .01.

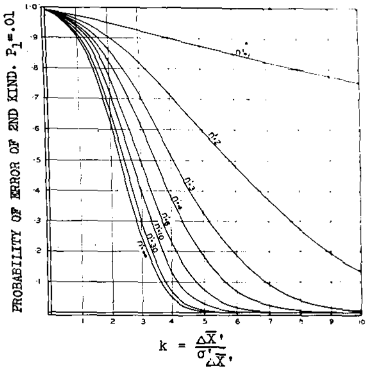


Fig. 8

Suppose we apply these curves to the case where P_1 = .01, and the difference $\Delta\bar{X}$ is determined as in the case of the comparison of the effects of the two drugs in a controlled experiment by let us say a sample of n = 1000 pairs. Fig. 9 shows the case where K is taken to be 3. From Neyman's calculations P_2 is .250 for this condition.

Hypothesis $\delta_2 \leq \delta_1$

1st kind – reject hyp [illegible]

2nd accept [illegible]

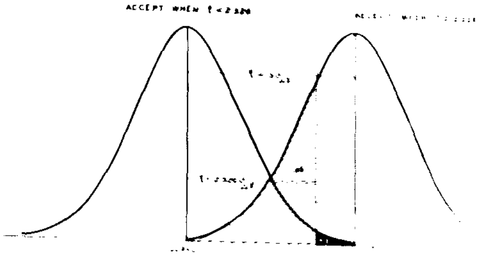


Fig. 9

That is to say, even when the true difference is greater than zero by three times the standard error of the difference, P_2 is 0.250.

Now, let us look again at the curves in Fig. 8. Note how little difference there is between those corresponding to $n' = 30$ and $n' = \infty$. Does this mean that the practical significance attached to P_2 for small samples is just the same as that for large samples? Well if we look at the ratio K we see that this is in terms of the standard error of the difference which at least in the case just considered is inversely proportional to the square root of the sample size! Hence we can narrow down the band within which we work at will by increasing the sample size in such a case.

Errors of the Third Kind

We have already mentioned the errors of the third kind - the effect of lack of normality and lack of randomness.

As noted, the effect of lack of normality can be and has been investigated mathematically and experimentally.

The effect of lack of randomness, however, is not so easily investigated. One difficulty is that the term is not so easily definable in an exact sense so far as observable values are concerned.

Let us consider a very simple case of a bowl containing N chips supposedly as near alike as we can make them. Suppose that the number X_1 is written on p'_1 N of these chips, the number X_2 on p'_2 N of these chips and so on until we have a discrete universe

$$p'_1, X_1; p'_2, X_2; \ldots; p'_i, X_i; \ldots p'_m, X_m ,$$

where

$$p'_1, + p'_2, + \ldots + p'_m = 1.$$

Then in discussing the theory of random samples of n where $n < N$ drawn from such a universe with replacement, we mean by random that we shall consider all possible samples of n numbers that can be obtained by taking n *different* chips. There are, of course, N^n ways of taking n chips from N where each of the n chips may be any one of the N chips. This constitutes a perfectly definite operation for defining *mathematically random*.

I do not see, however, how this can define drawing a sample at random. In fact this is usually more like requiring that the sample of n shall be drawn one at a time with replacement by one that is blindfolded where the operation mixes the chips thoroughly after each replacement. Another way of defining random drawing in this case is to specify that the drawing shall be made with the aid of some kind of "random" machine ~~maxbine~~. However, when it comes to defining the operation of making n measurements of a length, or of a charge on an electron or any other physical quality, there is a real difficulty. One usually ends up by saying that the measurements shall be made under the "same essential conditions". This, however, doesn't go very far toward fixing an operation in an operation [illegible] has to be improved further.

It is under such conditions that we impose a requirement on the numbers obtained under presumably the same essential conditions: that is, what we term Criterion 1 should give no points outside limits for something like 25 sets of 4.

COMMENT ON THE MEANING OF PROBABILITY

Before considering further the practical significance of tests for significance let us consider briefly three concepts of probability.

1. Mathematical Probability

The mathematical probability of a proposition or an event is something fixed and objective. Thus we say that the mathematical probability of throwing an ace with a perfect die is 1/6 - and the probability of throwing an ace (0,1,2,3...n) times in n throws is given by the terms of the expansion

$$(5/6 + 1/6)^n$$

The rules of operations of calculating probabilities in this sense are formal. They are the same today, tomorrow and the next day, at least so far as any observed data are concerned.

Sometimes we talk about probability distributions as universes. For example, we say: Let us assume a normal "universe" such that the probability of a value X lying within the range X ± 1/2 d X is given by

$$dy_1 = \frac{1}{\sigma' \sqrt{2\pi}} e^{-\frac{(X-\bar{X}')^2}{2\sigma'^2}} dX \qquad (1)$$

Then we say that the probability that an average of a sample of n drawn from this universe will have a value within the range $\bar{X}$ ± 1/2 d $\bar{X}$ is

$$dy_2 = \frac{\sqrt{n}}{\sigma' \sqrt{2\pi}} e^{-\frac{n(X-X')}{2\sigma'^2}} dX \qquad (2)$$

The rule of going from (1) to (2) is a purely formal operation independent of any data.

Now let us ask how the probability as used in the Student Fisher test of significance differs from mathematical probability. Recently, of course, it has been referred to as "fiducial" because it is a probability based upon a sample. So far as I can see, however, there is no difference. Both represent mathematical distributions that are derived from other distributions in a purely formal manner.

Statistical Probability

If probability theory is to be used in practice, with a perfectly definite operational meaning in respect to the observables of nature, it is necessary, in so far as I can see, that we give to it the meaning of an observable ratio p of the number n_1 of times a given event E occurs in n trials to the total number n of trials. In order to give definite meaning to a probability in this sense we must fix (a) the event E, (b) the operation constituting a trial, (c) the number n of trials, and (d) the ratio.

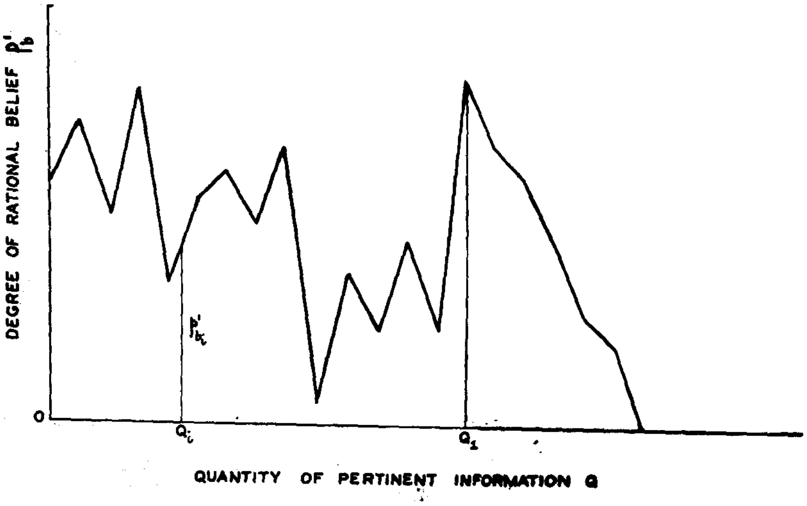
Now, of course, the experienceable fiducial probability given by the t test as interpreted in previous sections differs from the probability distribution (2) of averages in the event and in the operation constituting a trial.

I like to think of this observable fraction p as a statistical probability. In so far as we try to substitute this kind of probability for the mathematical probability we are trying to apply mathematics. Whether or not such applications will work can only be found out through experiment. Whether or not there are objective p's in nature is to be judged on experience. Mathematical probability deductions are valid irrespective of such experience.

Degree of Belief Probability

Having defined a statistical probability in a definite operational way the next thing is to find out if it exists in this verifiable sense. In this case we generally assume that the objective degree of belief p'_b in an inference, proposition or event E is not an intrinsic property like truth but inheres in the inference through some relation to evidence

Whereas mathematical probability is independent of data or evidence, degree of belief probability depends upon evidence Q. This is naturally a very vital distinction. From this viewpoint, not only the original (1) set of data but also the (2) results of an analysis constitute evidence.


DEGREE OF RATIONAL BELIEF p'_b
0
p'_{b_i}
Q_i
Q_1
QUANTITY OF PERTINENT INFORMATION Q

An Illustration

To illustrate the difference between the three kinds of probability, let us consider the simple case of drawing numbered chips with replacement from a normal distribution in a bowl. The mathematical statistician, or perhaps it would be better to say, the mathematician gives us two distributions:

$$dy_2 = f_2\ (\bar{X},\ \bar{X}',\ \sigma')\ d\bar{X} \qquad (2)$$

and

$$dy_3 = f_3\ (t)\ dt \qquad (3)$$

Now the statistician comes along and defines two kinds of operations and makes two predictions.

Operation 1

Make sure that the distribution of numbers on the chips is normal. Draw a sample of let us say n=4, calculate a range $\bar{X}_1 \pm .44\ \sigma_1$, draw another sample and calculate a similar range. Repeat process let us say N=100 times. Now calculate average of the numbers on the chips in the bowl. Then he predicts that you should find about 50 ranges including this true average.

Operation 2

Make sure the universe in bowl is normal, calculate the true average $\bar{X}'$ and standards deviation σ'. Now calculate a range $\bar{X}' \pm .6745\ \sigma'$. Draw a sample of 4 with replacement.

Calculate its average $\bar{X}_1$. Draw another sample of 4 and calculate its average $\bar{X}_2$. Repeat process let us say N=100 times. Now see how many such averages fall within the established range. Then he predicts that you should find $pN \doteq 50$ within this range.

Fig. 10 shows the results of applying these two operational tests. As a check on the first test, we find 51 instead of 50. As a check on the second, we find - instead of 50. The points ~~represent~~ the averages ~~found~~ within the limits $\bar{X}' \pm .6745\ \sigma'$.

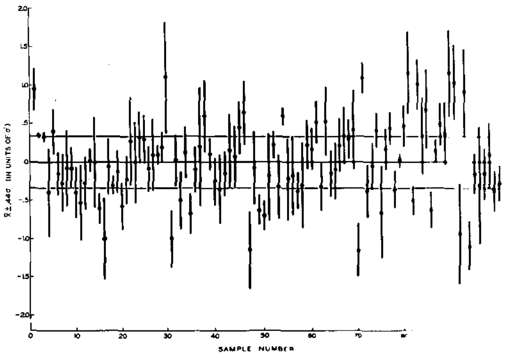


Fig. 10

In both cases we predict the outcome of a series of N operations before a sample is drawn. The operations in one case, however, are distinctly different from those in the other. Behind each is a mathematical equation involving mathematical probabilities.

Now, let us see where errors of the third kind come in. Suppose that while you are not looking and before you begin to draw a series of samples as in the cases cited above, someone with sticky or even just wet fingers stirs the chips in the bowl, your prediction probably would not be so successful as the illustration. In most cases, we are not drawing chips from a bowl except in the sense of a "bowl of unknown causes." For example, if we apply both of these tests to the 204 measurements given in Table 2, both of them give observed values of statistical ratios outside limits that hold in the case of the bowl.

Our belief in the successful outcome of either of the predictions depends upon evidence q which we have about (a) the distribution of the numbers on the chips in the bowl and (b) the physical similarity of the chips and the conditions of drawing. Of course, the prediction in terms of the ranges can be made justifiably upon the basis of less evidence than the prediction that 50% of the averages of samples of four will fall within the range $\bar{X}' \pm .6745\,\sigma'$.

COMMENT ON FUNDAMENTAL NATURE OF A TEST FOR SIGNIFICANCE

Fundamentally there is always a mathematical distribution of a statistic θ of a sample of size n behind every test for significance, i.e.

Bridgman (6)

1. I shall seek the meaning of our statements and concepts by trying to analyze what it is that we do when we are confronted with any concrete physical situation to which we attempt to apply the concept or about which we make a statement.

"and in fact the notion of probability is meaningless when applied to an individual event the proof of this is given by mere observation of what we do in applying the notion of probability. Suppose that I show you a die and remark that I intend to throw it in a minute. You volunteer the information that the probability is one-sixth that the throw will be a six. I am sceptical and ask you to justify your statement by pointing out the property of the event, when it takes place, that

Bridgman (6)

(4)

1. I shall seek the meaning of our statements and concepts by trying to analyze what it is that we do when we are confronted with a concrete physical situation to which we attempt to apply the concept or about which we make a statement.

"and in fact the notion of probability is meaningless when applied to an individual event, the proof of this is given by mere observation of what we do in applying the notion of probability. Suppose that I show you a die and remark that I intend to throw it in a minute. You volunteer the information that the probability is one-sixth that the throw will be a six. I am sceptical and ask you to justify your statement by pointing out the property of the event, when it takes place that

lity. From

mathematical

θ_1 to θ_2.

$$dy = \varphi(\theta, n)\, d\theta \quad \text{where}$$

Can be described as a probability one-sixth for a six. I then make the throw and get a six. What possible characteristics has the single event that can justify the statement? Your statement has a meaning only when applied to a sequence of events or when applied to the <u>construction</u> of the die and the method of throwing it. Even when applied to a sequence of events there is always an unbridgeable logical chasm thwarting a precise application of the notion of probability to any actual sequence."

"I believe that an examination, as the operational point of view prescribes,

2 of what one does, will show that the meaning to be ascribed to the probability of individual happenings is to be found in the rules of the <u>mental</u> game that one plays in thinking about and trying to understand the individual event."

$$dy = \varphi_1(\theta,\ n)\ d\theta \quad \text{where}$$
$$\theta = \varphi_2(X_1, X_2, \ldots X_n) \qquad (4)$$

This is of the nature of a mathematical probability. From this function we may determine the fraction or mathematical probability p of θ's lying within any interval θ_1 to θ_2. Thus far we are dealing only with mathematics. Now for any level of significance there is some such distribution and the numerical value of the level is simply a probability corresponding to a certain interval θ_1 to θ_2.

Such a level does not depend upon any experience with measurements, and hence has no practical significance until the statistician comes upon the scene. He sets up an operation which gives him a θ for a given sample. If he can show that in repeating this operation, the observed fraction p approaches p' as a statistical limit he can legitimately use the test for significance. The statistican cannot do this mathematically - for example he cannot determine that the chips in the bowl are essentially the same by mathematics. To interpret the statistician's level of significance we must therefore consider both the operation and his evidence Q that the observed fraction p will approach p' as a statistical limit.

PRACTICAL SIGNIFICANCE OF TESTS FOR SIGNIFICANCE

1. The practical significance of a test for significance must take into account the quantity and kind of evidence behind the test to indicate that errors of the third kind have been eliminated.

2. A test for significance is fundamentally an operation for testing an hypothesis. In considering the results of such a test we must therefore take into account the nature of the hypothesis tested. For example, in testing the significance of an observed average, we simply test this in respect to a chosen or hypothetical value. By choosing the hypothetical value, we can really make the so-called level of significance what we please.

3. There are at least two important levels of significance in testing the significance of any observed statistic corresponding to errors of the first and second kinds. Both of these levels are of importance in fixing the practical significance of a result.

4. Beyond everything else an understanding of the tests for significance should make every experimentalist cautious in accepting the results of small numbers of observations. In the first place, a comparatively large number of observations is always required except in simple bowl experiments to get

adequate evidence that errors of the third kind have been eliminated. Furthermore in commercial work the decision that an observed difference is significant may involve the outlay of a large sum of money in taking advantage of the difference assumed to exist. Hence even after one has assured himself that errors of the third kind have been eliminated he must take enough measurements to increase his precision to the desired level. In other words, practical significance usually depends upon more than one parameter of the distribution of the measurement under consideration. For example, we considered above only the test for the average. We usually need to consider tests for at least the variance. This will increase the number of our tests for significance by at least two, [redacted] corresponding to errors of the first and second kinds in the variance.

5. Personally I am of the opinion that the greatest practical significance of tests for significance are as a guide to laying out experiments to test a given hypothesis with a given number of trials - as guides to making the best use of experimental effort. This is a <u>very</u> important use. But we must not lose sight of the fact that even with the best layout the practical significance to be attached to results usually increases with an increase in the number of repetitions.

6. A study of such tests certainly shows the fallacy in the common practice of interpreting data when tabulated in the form of the observed average plus or minus a quantity such as a "probable error." A case in point is the discussion of Eddington's previously referred to. Planck's h either does or does not lie within a given range. The only thing that can be observed is further measurements of h. What one is interested in is: How many of these are likely to fall within the given fixed range. The probability given by the t test for example is not the probability of this kind of event. In tabulating data in terms of an average and range it is desirable that the number of measurements also be tabulated.