Introduction

In the theory of statistics the normal distribution occupies a very special place. We meet it early in our study of the subject, and we find that it is particularly amenable to mathematical treatment. Most of the statistical tests in common use are based on the assumption that the populations' values are normal. Although experience of the normal distribution often gives us good grounds for believing that this assumption is approximately true, yet we often meet with cases where it is unjustified, and we have to face the question: 'How can we do with non-normal populations?'

It may happen that although a chance variable $X$ is not normal or approximately normal, yet some simple function of $X$ such as $Y = g(X)$ may be so. Such transformations are sometimes used in the analysis of variance. Now, if we confine our attention to continuous chance variables whose frequency functions are of known forms but involve unknown location and scale parameters, we can do almost as much with normal variables, i.e. for continuous chance variables with frequency function

$$\frac{1}{z} f \left( \frac{x - a}{c} \right)$$

when the function $f(x)$ is known, but the location parameter $a$ and the scale parameter $c$ are unknown, there are analogues of the $t$-test, the variance ratio test, and other such tests.

But often we have no knowledge of the nature of the distribution except what is supplied by the sample. We may have to decide
From samples whether the distributions of two chance variables \( x \) and \( y \) are the same or different, without any a priori knowledge about the distribution of either. Valid tests have been devised to deal with such cases; the theory of them is the subject of this course. Because in developing them and applying them we usually do not start from an assumption that a chance variable has a distribution function of known form but involving certain unknown parameters, they have been called non-parametric tests; but the name is not a good one and it is to be hoped that someone will soon think of a much better. As will be seen, we can use these tests to estimate parameters and so test hypotheses concerning parameters.

2. The Sign Test

In the simplest application of the \( t \) test we have a set of \( n \) numbers \( x_1, \ldots, x_n \), which are observed values of a chance variable \( X \). We assume that \( X \) is normally distributed with unknown mean and unknown standard deviation, or, as we often express it, that the \( x \) are a random sample from a normal population of unknown mean and standard deviation. The question we wish to decide is "Is the mean of the population zero or not? Does the mean of the sample differ significantly from zero?" The hypothesis which we test, the null hypothesis, is that the mean of the population is zero (or some other specified value). The \( t \) test does this, and does it more efficiently than any other test would do it, but always with the assumption that the distribution of \( X \) is normal. What can we do if we have no grounds at all for assuming that the distribution of \( X \) is normal, and may even suspect or perhaps know that its distribution is far from normal?

The sign test makes no assumptions about the form of the population sampled. It tests hypotheses, not about the mean of a
population, but about the median of the population. By a change of origin we can always bring a hypothesis about the median having any specific value to the form that the median of the population is zero. To test this from the sample $X_1, X_2, \ldots, X_n$, we note only the numbers of positive and negative signs, say $n_1$, $n_2$, etc. $n_1 + n_2 = n$

We are assuming that there are no zero values. The probability of zero deviations occurring is zero if the population is continuous; for simplicity we assume this.

If the null hypothesis, population median $= 0$, is true, a positive and a negative sign for any $X$ are equally probable, and so positive and negative signs would occur with about equal frequency. The chance variables $n_1$ and $n_2$ will be symmetrical binomial variables, each with mean $n/2$ and variance $n/4$. Values of $n_2$ (or $n_1$) deviating widely from the expected value $n/2$ will be regarded as significant, i.e., as indicating that the null hypothesis is untrue. The probability of a deviation as large as, or larger than that observed, can be computed. The normal approximation to the binomial distribution will give sufficient accuracy at the ordinary levels (0.05 or 0.01) of significance. This can be conveniently used by means of the probability table of $\chi^2$

<table>
<thead>
<tr>
<th>$\chi^2$</th>
<th>1.6</th>
<th>2.7</th>
<th>3.8</th>
<th>5.4</th>
<th>6.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$</td>
<td>0.20</td>
<td>0.10</td>
<td>0.05</td>
<td>0.02</td>
<td>0.01</td>
</tr>
</tbody>
</table>

For a sample of 14 we have the following results:
\[ z^2 = \frac{2.35^2}{7} = 3.3 \quad 0.10 < \rho < 0.01 \]
\[ z^2 = \frac{3.2^2}{7} = 4.3 \quad 0.05 < \rho < 0.01 \]
\[ z^2 = \frac{4.8^2}{7} = 5.8 \quad 0.01 < \rho < 0.001 \]

Thus, at the 0.05 level, \( z = 3.2 \) is just not significant, but \( z = 4.8 \) is significant at the 0.01 level, and similarly \( z = 2.35 \), \( z = 4.3 \), and \( z = 5.8 \) as indicating that the median of the populations is \( \geq 0 \) and the latter as indicating median \( \leq 0 \).

It is easy to make a table showing \( r \), the minimum number of more numerous signs required to reach the 5% level of significance. Cochran, J.R.S.S. (1937) 62, 72, gives a table \( r = 3 \) for \( n = 5 \) to 50. By the normal approximation

\[ r = \frac{1}{2} n + \frac{1.96 \sqrt{n}}{2} + \frac{1}{2} \approx \frac{1}{2} n + \sqrt{n} + \frac{1}{2} \quad \rho \approx \text{approximately} \]

So far we have been discussing the two-way test. If \( n \) wish to test the hypothesis, median of population \( \neq 0 \), against the alternative median \( \neq 0 \), large values of \( n \), are significant, but not small values. Remembering that \( P \left\{ z_2 > \frac{1}{2} n + k \right\} = \frac{1}{2} P \left\{ |z_1| > k \right\} \), we can still use the \( z^2 \) method provided that we have the tabulated probabilities. Considering again the sample of \( n = 14 \), the probabilities of obtaining as many or more positive signs when the median is really zero are

\[ z^2 = \frac{10}{14} \]
\[ 0.05 < \rho < 0.1 \]
\[ z^2 = \frac{11}{14} \]
\[ 0.025 < \rho < 0.05 \]
\[ z^2 = \frac{12}{14} \]
\[ 0.005 < \rho < 0.01 \]

Note that these probabilities will be smaller if the population median is \( < 0 \). The result is now significant at the 5% level, and the 1% nearly so.
We shall consider the one-sided test of the hypothesis: median $\leq 0$ against the alternative: median $> 0$. Suppose we have $n = n_1 + n_2$, observed values of $X$ of which $n_1$ are negative and $n_2$ positive. Let $\lambda$ be a fixed positive number. Put

$$H = \frac{1}{\sqrt{n_2}} \int_{-\infty}^{\infty} \frac{1}{2\sqrt{\pi}z^2} \, dz = o(h)$$

If the population median is zero

$$P\{n_2 \geq H\} \rightarrow \alpha(\lambda) \quad \text{as } n \rightarrow \infty$$

determine $\lambda$ so that $\alpha(\lambda) = \alpha$, the level of significance (e.g., 0.05 or 0.01). If we take values of $n_2$ which are $\geq H$ as significant, the level of significance (i.e., the probability of obtaining a significant result when the median $= 0$) will $\rightarrow \alpha$ as $n \rightarrow \infty$.

Suppose now that the median $> 0$, so that

$$P(X > 0) = p > \frac{1}{2}$$

It is easy to show, either by the normal approximation or by use of Chebyshev's inequality, that for fixed $p$

$$P\{n_2 \geq H\} \rightarrow 1 \quad \text{as } n \rightarrow \infty$$

So therefore put

$$p = 1 + \frac{K}{\sqrt{n}}$$

$n_2$ will be a binomial variable of mean $n_p = \frac{n}{2} (1 + \frac{K}{\sqrt{n}})$ and variance $\frac{n}{4} \frac{(1 - \frac{K}{n})^2}{n}$. The power of the test will be

$$P\{n_2 \geq H\} = C(\lambda) + \xi_n,$$

where $\xi_n \rightarrow 0$ as $n \rightarrow \infty$, and
\[ \frac{1}{\sqrt{n}} \left[ \frac{1}{\frac{k}{\sqrt{n}}} \right] = \frac{(\lambda - k) \sqrt{n}}{\sqrt{n} \left( \frac{k}{\sqrt{n}} \right)} \]

For fixed \( k \) this \( \rightarrow \lambda - k \) as \( n \rightarrow \infty \), and so the power of the test will \( \rightarrow \beta(\lambda - k) \) as \( n \rightarrow \infty \). \( \lambda, k \) fixed. If in place of \( \lambda \) we have \( k_n \), where \( k_n \rightarrow k \) as \( n \rightarrow \infty \), the power of the test will tend to the same limit.

4. Comparison of the Sign Test with the \( t \) Test:

If the population is normal with standard deviation \( \sigma \) and mean \( m \), \( \sigma /\sqrt{n} \neq 0 \),

\[ P(X > 0) = \Phi(-m/\sqrt{n}) = \Phi(1 + k_n/\sqrt{n}) \]

where

\[ k_n = \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-\frac{1}{2}x^2} dx \]

Therefore when \( n \rightarrow \infty \), \( m \) fixed,

\[ \frac{k_n}{2(\frac{m}{\sqrt{n}})} \sim \frac{1}{\sqrt{2\pi}} \frac{m}{\sqrt{n}} \]

and so \( k_n \rightarrow 2m/\sqrt{2\pi} \). Hence the power of the \( t \) test \( \rightarrow \beta(\lambda - \frac{2m}{\sqrt{2\pi}}) \).

If we apply the \( t \) test to a sample of \( n' \) from a normal population of standard deviation \( \sigma \) and mean \( m', \sigma /\sqrt{n'} \), we consider

\[ t = \frac{x/\sqrt{n'}}{\sqrt{s/(n'n')}} = \frac{X/\sqrt{n'}}{\sqrt{n'/s(n'n')}} \]

The numerator is a normal variable of mean \( m' \) and unit standard deviation, and the denominator \( \rightarrow 1 \) in probability.
as \( n \to \infty \). Therefore as \( n \to \infty \) the distribution of \( t \to \) normal of mean \( 0 \) and unit standard deviation. On the null hypothesis \( m = \bar{m} \) and values of \( t \) are significant which are 
\[
\geq t_{\alpha} \quad \text{where} \quad 3(t_{\alpha}) \to \alpha \quad \text{as} \ n \to \infty ,
\]
and therefore \( t \to \lambda \). The power of the \( t \) test is
\[
P\left( t \geq t_{\alpha} \right) \quad \text{which} \to \quad \left( \lambda - m' \right) \quad \text{as} \ n \to \infty .
\]
The population tests will be the same in each case if
\[
\frac{m}{\sqrt{n}} = \frac{m'}{\sqrt{n'}}
\]
and the limiting powers will be the same if
\[
\frac{2m}{\sqrt{2n}} = m'.
\]
These require
\[
\sqrt{\frac{n}{n'}} = \frac{2}{\sqrt{2\pi}}
\]
\[
m' = \frac{2}{\pi}
\]
Then applied to \( n \) normal population the sign test as compared with the \( t \) test, which is the most efficient test for such a population, has an asymptotic efficiency of \( 2/\pi \). For small samples the efficiency is greater.

One disadvantage of the sign test is that it has a very limited choice of significance level when the sample is small. Inert from its application to non-normal populations, it is useful even with normal populations as a rough test which can be carried out in the head, remembering \( \frac{1}{2n} + \sqrt{\frac{n}{2\pi}} + \frac{1}{2} \) for the
First I shall discuss what may be called the closed form of the test, in which the observed numbers are not really regarded as a sample from a larger population.

Fisher first proposed the test in discussing (Design of Experiments, p. 53-54) an experiment of Darwin's as an illustration of the method of paired comparisons, as much used in biological work. The object of the experiment performed by Darwin was to determine whether the difference in origin between inbred and cross-bred plants influences their growth rate as measured by height at a given date. The plants used were Zea mays = maize = corn in United States. The seeds were planted in 5 pots, each pair consisting of a seed from a self-fertilized parent and a seed from a cross-fertilized plant. The seeds so chosen for comparison had germinated at the same time and the soil conditions in which they grew were largely equalized by planting in the same pot.

Differences in heights of an inch between cross- and self-fertilized plants of the same pair

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>49</td>
<td>23</td>
<td>56</td>
</tr>
<tr>
<td>57</td>
<td>28</td>
<td>24</td>
</tr>
<tr>
<td>8</td>
<td>41</td>
<td>75</td>
</tr>
<tr>
<td>16</td>
<td>14</td>
<td>60</td>
</tr>
<tr>
<td>6</td>
<td>23</td>
<td>48</td>
</tr>
</tbody>
</table>
The hypothesis which we wish to test, the null hypothesis, is that origin, cross or self-fertilized parent, has no influence on the growth rate of the resulting plant. If the null hypothesis is really true, the cross bred or inbred nature of a plant has no more effect than a label with that name tied on the plant would have. Interchange of labels between members of a pair would simply reverse the sign of the corresponding difference. Denote the sum of the observed differences by T (its numerical value above is 314). We may imagine repetitions of the experiment with the same plants growing to the same heights but with different allocats one of the labels to the individuals in the various pairs. The labels may be allocated in 2^15 ways; there are the same number of corresponding sums values of T, some of which happen to coincide with one another. In order to test the null hypothesis in this way we must know the distribution of T when all these allocations are equally probable. This distribution is entirely determined by the observed numbers. We can then determine the probability that, if the null hypothesis were true, a value (positive or negative) of T as large as, or larger than, that observed would be obtained. If this probability is small, say less than 0.05, we conclude that the observed value of T is significant and that there is a difference in growth rate between inbred and cross bred plants.

* Throughout these notes the terms "large", "larger", "largest", "small", "smaller", "smallest" always refer to modulus. x is larger than y if |x| > |y|.
Instead of using the value of \( z \) for a test, we merely note that the numbers of positive and negative signs in the observed differences and use a closed form of the binomial test. If the null hypothesis is true we may consider the "binomial" distribution of the number of positive and negative signs and calculate the probability of a discrepancy between these as great as or greater than that observed. This is

\[
\frac{2(1-15+\frac{15^2}{2})}{15} = \frac{121}{121} = 0.0071
\]

and the final result would be regarded as significant at the 1% level.

I call this form of test "closed" because in applying it we do not go outside the observations and make suppositions about populations from which they have been drawn. The probability distributions which we consider are determined completely by the observations themselves. We shall discuss several different tests of this closed form but they all have this in common. We may express it thus: A crime has been committed. Various natural agencies and the statistician are known to have been present. The question is who did it?

Now the statistician is always regarded as the villain of the piece. Suspicion always falls on him. In contrast to the best principles of political and social justice he is assumed to be guilty until he can establish his innocence. He is accused of tying irrelevant labels around the necks of supposed causes and taking up the appearance of crime, then, in fact nothing of any importance has really occurred. The null hypothesis is that the
statistician is guilty.

The method of detection is to reconstruct the crime with all the actors present. All except the statistician are asked to repeat their actions exactly as before (in Darwin's experiment the sunlight, soil, humidity, etc.). Only the part of the statistician is varied. Fortunately, such an impossible repetition is unnecessary. If only the statistician's part is varied we can work out the possible consequences of that with pencil and paper, and not bother the other actors.

The various consequences of the different alternatives which the statistician might have taken are worked out and classified as significant or non-significant, the former being few in number compared with the latter and such as would be more likely to happen if the null hypotheses were false. Assuming that the particular alternative actually taken by the statistician was determined by chance we can calculate the probability of a significant result from mere labelling. It is arranged so that this is small, say 0.05. A significant result exonerates the statistician and fastens the crime on one or more of the other actors. The probability of the statistician's being acquitted when he is really guilty of irrelevant labelling is small, say 0.05.

The most general statement of the closed form of Fisher's test is as follows. We wish to decide whether the
distributions of two chance variables are the same or different. We are not able to observe pure values of \(x\) or \(y\), but only values of these chance variables mixed with the values of other variables, so that our observations consist of \(n\) pairs of numbers

\[ a_1, b_1; a_2, b_2; \ldots; a_n, b_n, \]

where

\[ a_r = x_r + u_r, \quad b_r = y_r + v_r, \quad (r = 1, 2, \ldots, n) \]

and the \(x\) are a random sample of values of \(X\), the \(y\) a random sample of \(Y\), and the \(u\) and \(v\) are values of variables about which we know little or nothing. A well designed experiment will ensure that the values of \(u_r v_r\) are as small as possible, e.g., planting two seeds in a pair in the same plot etc.

Put

\[ w_r = a_r - b_r = (x_r + u_r) - (y_r + v_r) \]

\((r = 1, 2, \ldots, n)\)

\[ W = w_1 + w_2 + \ldots + w_n \]

The null hypothesis which we wish to test is that \(X\) and \(Y\) are experimentally indistinguishable. If this is true, the distinction between the numbers \(a_r\) and \(b_r\) is simply a case of labelling, and the experiment is so designed that when the null hypothesis is true this labelling has been done by chance. Relabelling would interchange \(a_r\) and \(b_r\) and give \(w_r\) in place of \(w_r\). We base our test on the value of \(T\), large values of which (positive or negative) are regarded as significant and as indicating that the null hypothesis is false. \(T\) is a suitable statistic for the test because we have in mind as
a possible alternative hypothesis that \( X - h \) and \( Y \) have the same distribution, where \( h \) is some constant not zero. The larger \( h \) is, the larger we should expect \( W \) to be.

Consider the \( 2^n \) sets of numbers

\[ \pm \omega_1, \pm \omega_2, \ldots, \pm \omega_n \]

and the corresponding values of their sum \( W \). Let \( k \) be a fixed integer less than \( 2^n \). Denote any particular set by \( I' \) and its sum by \( W' \). If the number of sets with a sum as large as or larger than \( W' \) is less than or equal to \( k \), we shall call \( I' \) significant. Otherwise \( I' \) will be called non-significant. A non-significant set such that the number of sets with a larger sum is less than \( k \) will be called borderline. Then there are borderline sets the number of significant sets will be less than \( k \) and the number of non-significant sets will be greater than \( 2^n - k \). With no borderline sets the numbers will be \( k \) and \( 2^n - k \) respectively. Since the sets occur in pairs of opposites with the same value of \(|I|\), we always take \( k \) to be even. The significant sets are most easily picked out as those whose negative members have the smallest sums, and the opposites of these.

The probability of obtaining a significant set by chance labelling, if all sets are equally probable is \( k 2^{-n} = P \), say, when there are no borderline sets, and the probability of a non-significant set is \( 1 - P \). Then there are borderline sets, the probabilities are less than \( P \) and greater than \( 1 - P \) respectively.
If we decide that a significant set of observed differences
\[ \frac{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n}{\bar{X}} \]
disproves the null hypothesis, the probability of error when the null hypothesis is true, is not greater than \( P \). In practice we choose \( n \) so that \( P \) is approximately equal to our usual working value, such as 0.05 or 0.01, for the permissible probability of an error of this kind. Then \( n \) is greater than 6 a convenient value of \( n \) is \( 3 \times 2^{n-6} \). This gives \( n = 3 \times 2^{-6} = \frac{3}{64} = 0.0469 \). For \( n > 8 \), \( n = 3 \times 2^{n-6} \) gives \( P = 3 \times 2^{-8} = 0.0117 \). We could avoid border-line sets by adjusting the value of \( n \) to the particularly as set of observed differences, but at present we are thinking of determining \( \bar{X} \) and \( P \) before the experiment is analysed, of formulating a definite rule which we automatically follow in all cases.

This test can give no significant result at the 5% level when \( n \) is less than 6. Then \( n > 6 \) there are two significant sets, viz. all positive and all negative. These are significant with \( P = 1/32 = 0.03125 \) if no \( n \) is zero. Half of the significant sets have positive sums and half have negative sums. We only need to pick out one of these groups since the other group consists of the opposite sets. Corresponding values of \( n \) and \( \bar{X} \) are given below.

\[
\begin{array}{cccc}
  n & 3/2^6 & 0.0469 & n & 3/2^6 & 0.0117 \\
 7 & 3 & 1 & 9 & 3 & 1 \\
 8 & 6 & 2 & 10 & 6 & 2 \\
 9 & 12 & 3 & 11 & 12 & 3 \\
 10 & 24 & 4 & 12 & 24 & 4 \\
\end{array}
\]
The number in the last column is the maximum number of \( w \) opposite in sign to the sum of the set consistent with significance, e.g., a set of 8 \( w \) of which 3 are negative and which has a positive sum cannot be significant at the level \( P = \frac{3}{2^6} \), for alteration of the signs of all or some or none of these negative \( w \), while leaving the other 5 unaltered, will give 8 sets with a positive sum as large as or larger than that observed, and \( n \) is greater than \( \frac{1}{2} N \).

Let \( V \) be the sum of the moduli of those \( w \) of a set which are opposite in sign to the sum \( S \), and let \( I \) denote the number of combinations (including the zero combination) of

\[
\left| w_1 \right|, \left| w_2 \right|, \ldots, \left| w_n \right|
\]

which have a sum less than or equal to \( V \). The necessary and sufficient condition for the given set to be significant is

\[
N \leq \frac{\left| w \right|}{2V} \quad \text{Note:} \quad \left| w \right| = \left| \frac{1}{2} \sum w - 2V \right|
\]

**Examples:**

\( 0.9, -1.9, 2.6, 5, 7.1, 4.6, 11.2, 10.3 \)

\( n = 7, \frac{1}{2} n = 3 \) for \( P = 0.0469 = \frac{3}{2^6} \).

The combinations with the three smallest sums are 0, 0.9, 1.9, and \( N = 3 = \frac{1}{2} n \), and the set is significant.

\( -1.0, 4.6, 3.2, 21.2, -4.1, 8.5, 1.2, 0.9 \)

Here \( n = 8, \frac{1}{2} n = 6 \) for \( P = 0.0469 \), \( V = 5.1 \). Obviously \( N > 6 \), for the 8 combinations from the three smallest members 0.9, 1.2, 1.2 have sum smaller than \( V \). The set is therefore non-significant.

\( \)
The approximate distribution of \( Z \).

Denote by \( Z \) a chance variable which takes the values \( r \) and \( -r \) each with probability \( \frac{1}{2} \). Its mean value is 0 and its variance is \( r^2 \). The chance variable \( Z \) is the sum of \( n \) independent chance variables

\[ Z = Z_1 + Z_2 + \ldots + Z_n. \]

Its mean value is 0 and its variance is \( \sum r^2 \). If the maximum of \( r^2 \) for \( s = 1, 2, \ldots, n \) is small, the distribution of \( Z \) will be approximately normal.

\[ Z = \frac{Z}{\sqrt{\sum r^2}} \]

will be approximately standard normal variable (Gibb).

Fisher has discussed in detail the analysis of Darwin's experiment. Let us consider another of Darwin's experimental results.

Differences in eighths of an inch between heights of cross- and self-fertilized plants of the grass Phalaris canariensis

8, -46, 42, 93, -107, 17, 78, 60, 72, 41, 29

Arranging in order of absolute magnitude, we have

8, 17, 20, 41, 42, 46, 60, 72, 78, 93, 107

Here \( n = 11 \) and \( \frac{1}{n} = 48 \) for \( P = 0.0469 \). The sum of the moduli of negative differences is 153. The single numbers and 0 are three less than 98 give smaller sums. The number of these is \( 11 + 1 + 36 = 48 \). Also any three less than 72 give a smaller sum. Hence \( n > 3! \), and the set is not significant at level 0.0469.
The value of \( T \) is 283. Let us calculate the probability of obtaining a value as large as or larger than this. \( z^2 = 42235 \), the square root of which is 205.5. If the normal approximation is good, as the numbers themselves suggest, \( z = 7/205.5 \) will be approximately 3.37. \( T = 283 \) gives \( z = 1.377 \). The probability of \( z \) being as large as or larger than this is 0.169.

To determine the exact value of the probability of obtaining a value of \( z \) as large as or larger than 283 we set down the possible values of \( V \) in order. First 0, then 8, then the values of \( V \) for combinations involving no \( w \) larger than 17, then those involving no \( w \) larger than 27, and so on, up to \( V = 153 \).

<table>
<thead>
<tr>
<th></th>
<th>8</th>
<th>17</th>
<th>20</th>
<th>41</th>
<th>42</th>
<th>46</th>
<th>60</th>
<th>72</th>
<th>75</th>
<th>98</th>
<th>177</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>28</td>
<td>37</td>
<td>45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>49</td>
<td>53</td>
<td>62</td>
<td>61</td>
<td>69</td>
<td>78</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>50</td>
<td>59</td>
<td>67</td>
<td>62</td>
<td>70</td>
<td>79</td>
<td>87</td>
<td>93</td>
<td>91</td>
<td>107</td>
<td>108</td>
</tr>
<tr>
<td>46</td>
<td>54</td>
<td>63</td>
<td>71</td>
<td>66</td>
<td>74</td>
<td>83</td>
<td>91</td>
<td>87</td>
<td>95</td>
<td>107</td>
<td>108</td>
</tr>
<tr>
<td>60</td>
<td>68</td>
<td>77</td>
<td>85</td>
<td>80</td>
<td>88</td>
<td>97</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>82</td>
<td>89</td>
<td>97</td>
<td>92</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>73</td>
<td>86</td>
<td>95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>98</td>
<td>(13)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>107</td>
<td>(17)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The numbers in parentheses are the numbers of values of \( V \) which are maximum not greater than 153 but too large to be required for subsequent entries. Those less than 171 are all listed.

\[ N = 182 \] and so the true value of \( P \) is \( 182/2^{10} = 0.178 \). The normal approximation is thus fairly good at this point.
The normal approximation also gives the 1% and 5% points fairly accurately as shown by the table below.

| N  | V  | |z|  | z  | P  | P' |
|----|----|---|---|---|---|---|
| 12 | 49 | 491 | 2.389 | 0.0168 | 0.0117 |
| 35 | 80 | 429 | 2.088 | 0.0368 | 0.0342 |
| 48 | 92 | 405 | 1.971 | 0.0488 | 0.0469 |
| 182| 153| 283 | 1.377 | 0.1686 | 0.1777 |

P is the true probability and P' the value given by the normal approximation.

In the ordinary t test based on the assumption of sampling from a normal population we consider

\[
T = \frac{W \sqrt{n}}{\sqrt{(\sum x^2 - W^2)}} \sqrt{n-1} = \frac{W \sqrt{n-1}}{\sqrt{n \sum x^2 - W^2}}
\]

\[
= \frac{Z \sqrt{n-1}}{\sqrt{n \sum x^2}}
\]

\[
Z = \frac{T}{\sqrt{1 + \frac{Z^2}{n}}}
\]

Then n tends to \( \infty \), z tends to \( t \) uniformly with respect to \( t \) in any finite interval, and the distribution of \( t \) tends to standard normal. Therefore the z test and the t test are asymptotically equivalent. For values of \( n \) not too small they will be approximately equivalent. They would be exactly equivalent if the distribution of

\[
\frac{T}{\sqrt{1 + (z^2 - 1)/n}}
\]

were standard normal. This is a fairly good approximation to the truth when \( n > 9 \) and \( t^2 > 1 \). The tests are based on very different assumptions. The t test assumes that the observed differences
$w_1, w_2, \ldots, w_n$

are a sample from a normal population. The z test assumes merely that the values of these $w$ are such that the distribution of $T$ is approximately normal, and assumption which can be checked from the numbers themselves.

In order to investigate further the relation of the $w$ test and the $t$ test let us find the moments of the $w$ distribution. Denote the moments about 0 of the $n$ numbers $w_1, w_2, \ldots, w_n$ by $M_1, M_2, \ldots, M_n$. The m.g.f. of $w$ is

$$E(\exp \ w u) = \frac{1}{2} \exp (w u) + \frac{1}{2} \exp (-w u) = \cosh (w u)$$

The cumulant function is therefore

$$\log \cosh (w u) = \frac{w^2 u^2}{12} - \frac{2 w^4 u^4}{144} + \ldots$$

Hence the cumulant function of $T$ is

$$n M_2 u^2 \frac{1}{12} - 2 n M_4 u^4 \frac{1}{144} + \ldots$$

The odd moments of $T$ are all zero, and

$$E(T^2) = n M_2$$

$$E(T^4) = 3 n^2 M_4 - 2 n M_6$$

Put

$$U = \frac{M_2}{\sqrt{M_4}} = \frac{W^2}{n^2 M_2}$$

Then $0 \leq U \leq 1$

$$E(U) = \frac{1}{n}$$

$$E(U^2) = \frac{3}{n^2} \left( 1 - \frac{2 M_4}{3 n M_2^2} \right)$$

The $B(\frac{3}{2}, \frac{n-1}{2})$ distribution has the same range as $U$. Its first two moments are

$$\frac{1}{n} \quad \frac{\frac{3}{2}}{n(n+2)}$$
The first moment is exactly the same as that of $U$, and the second moment is approximately the same as the second moment of $U$ if $\frac{\overline{w}_2}{\sqrt{n} \overline{w}_2}$ is small. This quantity necessarily lies between $\frac{1}{n}$ and 1. It has the former value when the $w_p$ are all equal and the latter value when all but one $w_p$ are zero. It will be approximately 1 if one $w_p$ is much larger than all the rest. If $m$ of the $w_p$ have about the same value which is much larger than all the rest $\frac{\overline{w}_2}{\sqrt{n} \overline{w}_2}$ will be approximately $\frac{1}{n}$.

Hence if $\frac{\overline{w}_2}{\sqrt{n} \overline{w}_2}$ is small we may expect the distribution of $U$ to be approximately a $B(\frac{1}{2}, \frac{n-1}{2})$ distribution. Now

$$U = \frac{\overline{w}_2}{\overline{w}_2} = \frac{\overline{Z}^2}{Z^2 + n - 1}$$

But if we are sampling from a normal population of mean 0,

$$\frac{\overline{Z}^2}{Z^2 + n - 1}$$

has a $B(\frac{1}{2}, \frac{n-1}{2})$ distribution. Therefore the use of this approximation for the distribution of $U$ is exactly equivalent to the use of the $t$ test in the ordinary way. But in this use of the $t$ test we are not making any assumption of normality. We are merely assuming that the distribution of $U$ is approximately a $B(\frac{1}{2}, \frac{n-1}{2})$ distribution. This can, if necessary, be checked.
7. Fiducial Limits (or Confidence Intervals) for the Difference in Location of Two Populations of the Same Form.

Our definition of the numbers \( w \) in
\[
\omega_r = a_r - b_r = x_r + u_r - y_r - v_r
\]
where the \( x \) are a random sample of a chance variable \( X \), and the \( y \) are a sample of a chance variable \( Y \). To now make the assumption that the distributions of \( X \) and \( Y \) differ only in location, i.e., that \( X - h \) and \( Y \) have the same distribution, where \( h \) is an unknown constant which we wish to estimate.

On this assumption
\[
(1) \quad \omega_1 - h, \quad \omega_2 - h, \quad \ldots, \quad \omega_n - h
\]
take the place of \( x_1, \omega_2, \ldots, \omega_n \). The probability that they form a significant set is \( \leq P \), and the probability that they are not significant is \( \geq 1 - P \).

If the set (1) is significant or non-significant but not borderline, it will remain so when \( h \) is altered provided the change in \( h \) is sufficiently small. If it is significant and has a positive sum it will still be significant with a positive sum when \( h \) is decreased, for its sum will increase and the increase will be greater than the increase of the sum of any of the \( 2^n - 1 \) other sets with which it is compared. A borderline set with a positive sum will become significant when \( h \) is decreased. Similarly if (1) is significant with a negative sum, it will remain so when \( h \) is increased. A borderline set will become significant.

Let \( h_1 \) be the supremum of values of \( h \) for which the set (1) is significant with a positive sum, and \( h_2 \)
the minimum of $h$ values for which (1) is significant with a negative sum. From the preceding paragraph it follows that (1) is significant if $h < h_1$ or $h > h_2$, non-significant if $h_1 < h < h_2$, and borderline if $h = h_1$ or $h = h_2$.

The statement
\[ h_1 < h < h_2 \]
has fiducial probability $\geq 1 - P$, for it is equivalent to the statement that the set (1) is non-significant.

The statement
\[ h < h < h_2 \]
has fiducial probability $\leq 1 - P$. If $X$ and $Y$ are continuous chance variables, the probability of a borderline set is 0, and so the fiducial probability of both statements is exactly $1 - P$. In future we shall assume this unless the contrary is stated.

If the mean values of $X$ and $Y$ both exist
\[ E(Y) = E(X - h) = E(X) - h \]
and
\[ h = E(X) - E(Y) \]

Consider the set
\[ (2) \quad -94, -27, 4, 6, 19, 20, 23, 24 \]
We shall determine the values of $h$ for which the set
\[ (3) \quad -94-h, -27-h, 4-h, 6-h, 19-h, 20-h, 23-h, 24-h \]
is non-significant at the level $P = \frac{\chi^2}{2} = 0.0469$, $1 - P = 0.9531$.
For the set (2)

\[ \sum w = -2.7, \quad \sum w^2 = 11.403, \quad n = 8, \]

\[ \mu_1 = -3.157, \quad n\mu_1 = 7.8, \quad S = \sum w^2 - n\mu^2 = 11.405 \]

For the set (3)

\[ z = \frac{\sum (w_i - h_i)}{\sqrt{\sum (w_i - h_i)^2}} = \frac{n(\mu - h)}{\sqrt{\sum (w_i - h_i)^2}} \]

therefore

\[ \mu_i - h_i = \frac{Z\sqrt{n}}{\sqrt{\sum (w_i - h_i)^2}} \]

Assuming that \( z \) is \( \text{SNV}_0 \)

\[ P\{ |z| \leq 1.967 \} = 0.9531 \]

This gives

\[ P\{ |\mu_i - h_i| \leq 1.967 \frac{\sqrt{11.405}}{\sqrt{8 - 8.1987}} = 37.27 \} = 0.9531 \]

100 \[ P\{ -40.40 \leq h_i \leq 34.14 \} = 0.9531 \]

Values of \( h \) outside the range \(-40.40, 34.14\) will correspond to larger values of \( z \) and so will be significant. Thus the normal approximation gives

\[ h_i = -40.40, \quad h_z = 34.14 \]
Now
\[ t = \frac{(\mu_i - h) \sqrt{n}}{\sqrt{\frac{s}{(n-1)}}} \]

\[ \mu_i - h = \frac{t \sqrt{s}}{\sqrt{(n)(n-1)}} \]

Using the t distribution (T.D.W.) we have
\[ P\{ |t| < 2.408 \} = 0.9531 \]

and therefore
\[ P\left\{ \left| \mu_i - h \right| < 3.436 \right\} = 0.9531, \]

which gives
\[ h_1 = -37.49, \quad h_2 = 31.23. \]

Note that in using these approximations we are assuming, not that the original numbers w satisfy the requirements for a good approximation, but that the resultant numbers w - h do so. Let us determine the true values of \( h_1 \) and \( h_2 \). These approximate values will be a guide.

(3) -94-h, -27-h, 4-h, 6-h, 19-h, 20-h, 23-h, 24-h

\[ P = \frac{3}{2}, \quad n = 8, \quad \frac{1}{2}M = 6 \]

Then \( h = -40 \), the set becomes
-54, 13, 44, 46, 59, 60, 63, 64

and \( N = 5 \).
h = -39
-5.5, 12, 4.3, 4.5, 5.8, 5.9, 6.2, 6.3  N = 6

h = -38 1/3
-5.5 1/3, 11 1/3, 4.2 1/3, 4.4 1/3, 5.7 1/3, 5.8 1/3, 6.1 1/3, 6.2 1/3  N = 7

and the set is borderline, for there are only 5 sets with a larger positive sum. Thus h = -38 1/3. With these resultant numbers we should expect the normal or the t approximation to be fairly good; but it must be remembered that we are working on the tails of the distributions.

h = 30
-12, -5.7, -2.6, -2.4, -1.1, -1.0, -7, -6  N = 1

h = 20
-11.4, -4.7, -1.6, -1.5, -1, 0, 3.4  N = 15

h = 22
-11.6, -4.9, -1.8, -1.6, -3, -2, 1, 2  N = 7

This set is borderline for there are only 4 sets with a larger negative sum. Hence h = 22. With the resultant values of α-h we should not expect either the normal or the t approximation to be good. The true result is

P{ -38 1/3 ≤ h ≤ 22 } = 0.9531
In testing the null hypothesis which shall we use, the normal or the $t$ distribution approximation? To the order to which either has been justified they coincide.

In practice we may use the usual $t$ test provided that there are not just a few $w$ values much larger than all the others. If there are, we must use the exact test unless we are prepared to reject these values or the whole experiment. If we retain these outstanding $w$ they will dominate the experiment. In determining fiducial limits we must use the exact test if the resultant values of $w - h$ show that the approximation is unreliable.

3. TEST FOR THE CENTRE OF A SYMMETRICAL POPULATION

In the simplest application of the $t$ test we have a sample from a normal population of unknown mean and unknown standard deviation, and we test the hypothesis that the population mean is $0$ or some other specified value. Fisher's test can be applied in a similar way to a sample from any symmetrical population.

Suppose that the numbers

$w_1, w_2, \ldots, w_n$

are a sample from a population which is symmetrical about the value $h$. If $h = 0$, the sets

$\pm w_1, \pm w_2, \ldots, \pm w_n$

are equally probable. Hence the conditional distribution of $t$, for

$|w_1|, |w_2|, \ldots, |w_n|$
all fixed, will be the same as before, and we can classify a sample as significant or non-significant by comparing its \( t \) value with those of other samples

\[ \pm w_1, \pm w_2, \ldots, \pm W \]

If the null hypothesis, \( h = 0 \), is true, the probability of obtaining a significant sample is \( P \), assuming for simplicity that \( w \) are sampling from a continuous population. To test the hypothesis that \( h \) has the value \( 0 \) or any other specified value, and we can determine fiducial limits for \( h \). In fact, this is really a special case of the preceding. Then all the \( u \) and \( v \) are zero, the \( w \) are a sample of values of the chance variable \( X-Y \), which has a distribution symmetrical about \( h \) if \( X-h \) and \( Y \) have the same distribution. But there is a difference in the point of view. In the previous discussion attention was focussed on the activities of the statistician, here it is not. We are now thinking of repeated sampling from the population and working out probabilities by considering conditional distributions. To indicate the difference we may say that this is the open form of the test while the other is the closed form.

9. THE USE OF RANKS IN FISHER'S TEST

The \( F \) test is an exact test. It has the great theoretical advantage that the distribution of \( F \) is completely known, because it is completely determined
by the observed numbers \( w_r \); but it has the practical
disadvantage that this distribution varies from set to
set of observed numbers and so cannot be tabulated.
We can get rid of this disadvantage by using in place
of the observed numbers their signed ranks. Let \( R(\pm w_i) \)
be the rank of \( |w_r| \) in the set
\[
|w_1|, |w_2|, \ldots, |w_n|
\]
then we replace \( w_r \) by \( R(\pm w_i) \) if \( w_r \) is
positive and by \( -R(\pm w_i) \) if \( w_r \) is negative.
If no two \( w_r \) have the same modulus and if none is
zero, our set of observed numbers is replaced by one
of the sets
\[
\pm 1, \pm 2, \ldots, \pm n
\]
We apply the \( T \) test to this set. Since the distribution
used in the test is the distribution of \( T \) when all those
sets are equally probable, this distribution will be
independent of the observed numbers \( w_r \) and will depend
only on \( n \); it can therefore be tabulated. Wilcoxon
(Biometrics Bulletin 1 (1945) p. 32) gives a table show-
ing values of \( V \) (see \( \gamma \) above) and \( P \), the probability
of obtaining a value of \( T \) as large as, or larger than
that corresponding to \( V \) for values of \( n \) from 7 to 15.

Taking Darwin's results on *P. calavis canaliculatus*
discussed in \( \gamma \),
8, 17, 20, 41, 42, 46, 60, 78, 78, 98, 103
We replace these numbers by

1, 2, 3, 4, 5, -6, 7, 8, 9, 10, -11.

\( v = 17 \), \( n = 11 \). Wilcoxon's table gives \( P = 0.053 \) corresponding to \( v = 11 \) for \( n = 11 \) and therefore the set is not significant at this level.

When two or more \( |w_i| \) have the same rank they are assigned the mean rank value. This will not affect the distribution of \( v \) very much if there are only a few of these compared with \( n \), and it may have no effect on the part of the distribution we are interested in. If it appeared that there might be considerable effect we should have to work out the distribution instead of using the table. Zero values of \( w_i \) are simply discarded -- they give no evidence either way -- and we deal with the smaller sample consisting of non-zero values. If the chance variables are continuous the probability of either of these occurrences is zero.
10. CONSISTENCY OF FISHER'S TEST

Consider the open form of Fisher's test.

Suppose that \( \omega_1, \omega_2, \ldots, \omega_r \)
is a sample and that we wish to test the hypothesis that
the population is symmetrical about 0. The alternatives
are that the population is not symmetrical or that it is
symmetrical about some other point.

\[
Z = \frac{\sum \omega_r}{\sqrt{\sum \omega_r^2}}
\]

has mean 0 and standard deviation 1. If

\[
\max_s \frac{\omega_s^2}{\sum \omega_r^2}
\]
is small then \( Z \) has approximately \( N(0, 1) \) distribution.

Now if \( \omega_1, \omega_2, \ldots, \omega_r \) is a sample
from a population of finite variance it can be shown
that

\[
\max_s \frac{\omega_s^2}{\sum \omega_r^2} \rightarrow 0 \text{ in probability}
\]
as \( n \rightarrow \infty \). Therefore the distribution of \( Z \) tends
in probability to standard normal as \( n \rightarrow \infty \). The
true distribution of \( Z \) gives an exact test. Using
the normal approximation, the approximate \( Z \) test and the
exact \( t \) test are asymptotically equivalent.

Determine \( k \) so that

\[
\frac{1}{\int_{-\infty}^{\infty}} e^{-\frac{x^2}{2}} \; dx = \beta
\]
then values of \( z \geq k \) are significant.

Consider now the distribution of the observed value of \( z \). (Do not confuse this distribution with the distribution of \( z \) determined by the observed values.) Suppose that the population has finite variance \( \sigma^2 \) and mean \( \mu \neq 0 \).

\[
Z = \frac{\sum \frac{X_i}{\sigma} \sqrt{n}}{\sqrt{\sum \frac{X_i^2}{n}}}.
\]

As \( n \to \infty \), the denominator tends in probability to \( \sqrt{\sigma^2} \). The distribution of the numerator tends to normal with mean \( \frac{\mu}{\sigma} \) and standard deviation \( \frac{\sigma}{\sigma^2} \).

Therefore \( |\frac{\sum w_i}{n}| \to \infty \) in probability as \( n \to \infty \). Hence \( |Z| \to \infty \) in probability as \( n \to \infty \). Thus the test is consistent against the alternative of finite variance and mean \( \neq 0 \).

If \( \mu = 0 \), this distribution of \( z \) tends to standard normal even if the population is unsymmetric and the probability of rejecting the hypothesis will \( \to \) \( P \). The test is ineffective against this alternative.

11. **SIGNIFICANT AND NOSIGNIFICANT SEPARATIONS**

We now go on to discuss tests of the hypothesis that two given samples came from the same population. First we shall consider a test analogous to the \( t \) test for the significance of difference of means.

\( \neq \text{ and so the probability of a significant result} \to \infty \) as \( n \to \infty \).
Suppose that we have \( m+n \) real numbers (not necessarily all different), and that their mean is \( \overline{z} \). Numbers which are equal are supposed to be distinguishable from one another -- we may think of the \( m+n \) numbers as painted on \( m+n \) marbles. Consider a separation of the numbers into two different classes of \( m \) and \( n \):

\[ u_1, u_2, \ldots, u_m, \quad \text{with mean } \overline{u}, \]
\[ v_1, v_2, \ldots, v_n, \quad \text{with mean } \overline{v}. \]

The number, \( N \), of such separations is \( \binom{m+n}{m} \), provided that when \( m = n \) the two classes, though equal in size, are regarded as different, so that two separations in which the classes are simply interchanged are regarded as different separations.

We shall call \( |\overline{u} - \overline{v}| \) the spread of the separation, since

\[ m \overline{u} + n \overline{v} = (m+n) \overline{z}, \]

an alternative expression for the spread is

\[ \frac{(m+n) |\overline{u} - \overline{v}|}{mn} = \frac{(m+n)}{n} |\overline{u} - \overline{z}|. \]

Let \( \nu \) be a fixed integer less than \( N \). Denote any particular separation by \( R \). If the number of separations with a spread equal to or greater than that
of $R$ is less than or equal to $1$, we shall call $R$ significant. Otherwise $R$ will be non-significant. A separation which is non-significant but whose spread is exceeded by fewer than $t$ other separations will be called borderline. If $m = n$ the separations occur in pairs with equal spreads, and in that case we shall always take $t$ to be even. Then there are borderline separations, the number of significant separations is less than $t^*$. If there are no borderline separations, the number of significant separations is $t^*$. If $m \leq n$ the significant separations are most easily picked out as those with the largest values of \\

$$|z^2 - m \bar{z}|$$

If the separation is arrived at by chance, so that all separations are equally probable, the probability of a significant set is $\frac{M}{N} = P$ say, then there are $n$ borderline separations, and the probability of a non-significant separation is $1 - P$. Then there are borderline separations these numbers are respectively less than $P$ and greater than $1 - P$. It should be noted that throughout any discussion $m, n, t^*$ are supposed to be fixed.

Let $x_1, x_2, \ldots, x_m, y_1, \ldots, y_n$ be $m + n$ given numbers, and consider the separation

$$x_1 + 1, \ldots, x_m + 1$$

(1)

$$y_1, \ldots, y_n$$
When $h$ is altered a significant separation will remain significant, and a non-significant separation which is not border-line will remain so, provided that the change in $h$ is sufficiently small. If $\bar{x} + h > \bar{y}$, a significant separation will remain significant when $h$ is increased, but a border-line separation will become significant. Similarly, if $\bar{x} + h < \bar{y}$, a significant separation will remain significant when $h$ is decreased, but a border-line separation will become significant.

Consider the values of $h$ for which the separation (1) is significant. Denote by $h_1$ the supremum of those which satisfy the inequality

$$\bar{x} + h < \bar{y},$$

and by $h_2$ the infimum of those which satisfy the inequality

$$\bar{x} + h > \bar{y}.$$

From the results given above it follows that the separation (1) is significant if

$$h < h_1 \quad \text{or} \quad h > h_2,$$

non-significant if

$$h_1 \leq h \leq h_2,$$

and border-line if

$$h = h_1 \quad \text{or} \quad h = h_2.$$
Suppose now that we have two samples of size \( m \) and \( n \), and that the sample values are

\[
\begin{align*}
x_1, & \quad x_2, \quad \ldots, \quad x_m \\
\frac{x_1}{m}, & \quad \frac{x_2}{m}, \quad \ldots, \quad \frac{x_m}{m} \\
\frac{y_1}{n}, & \quad \frac{y_2}{n}, \quad \ldots, \quad \frac{y_n}{n}
\end{align*}
\]

Whether the samples are from the same population or not, we shall say that they are significant, non-significant, or border-line, according as the separation of the \( m \times n \) numbers

\[
\begin{align*}
x_1, & \quad x_2, \quad \ldots, \quad X_m, \quad \frac{x_1}{m}, \quad \frac{x_2}{m}, \quad \ldots, \quad \frac{x_m}{m} \\
\frac{y_1}{n}, & \quad \frac{y_2}{n}, \quad \ldots, \quad Y_n
\end{align*}
\]

into the two classes

\[
\begin{align*}
x_1, & \quad x_2, \quad \ldots, \quad X_m, \quad \frac{x_1}{m}, \quad \frac{x_2}{m}, \quad \ldots, \quad \frac{x_m}{m} \\
\frac{y_1}{n}, & \quad \frac{y_2}{n}, \quad \ldots, \quad Y_n
\end{align*}
\]

is significant, non-significant, or border-line.

If the samples are drawn from the same population, the probability of obtaining a significant pair is \( < P \). For instead of drawing separate samples of \( m \) and \( n \) we might, without affecting any probabilities, first draw a sample of \( m + n \) and then separate it into a sample of \( m \) and a sample of \( n \) by a process which makes all such separations equally probable. For simplicity we shall assume that the population is continuous. The probability of a border-line pair is then zero and the probability of a significant pair is \( P \).
To test the hypothesis that the two samples are from the same population we determine whether the samples are significant or not. If they are significant we reject the hypothesis, if non-significant we accept it. The probability of an error of the first kind, of rejecting the hypothesis when it is really true is $P$.

**Example**: Are the following samples significantly different?

$1.2, 2.3, 2.4, 3.2$ and $2.9, 3.1, 3.4, 3.6, 4.1$

For convenience we may without affecting the test, subtract 1.2 from each sample value so as to reduce the smallest number to 0, and then multiply each by 10 to get rid of the decimal points. We then have

$0, 11, 12, 20$ and $19, 19, 22, 24, 29$

Arranging these in order of magnitude, we have

$0, 11, 12, 19, 19, 20, 22, 24, 29$

The mean value is 17, so that $m = 68$. There are 126 separations of the numbers into classes of 4 and 5. Hence if we take $N = 6$, we shall have $P = 1/21$.

The groups of 4 which give the largest values of $|\Sigma u - 68|$ are:
The group 0,11,12,29 gives the smallest \( T^2 = 681 \) and so with \( k = 5 \) the corresponding separation is significant. Hence at the level \( P = 0.021 \) the samples must be regarded as significantly different. \( k = 5 \) would give \( P = 0.0397 \) and at this level also the samples are significantly different.

13. **FIDUCIAL LIMITS FOR THE DIFFERENCE IN LOCATION OF THE POPULATIONS OF THE SUTT POOL**

If \( x_1, x_2, \ldots, x_m \)

are \( m \) observed values of a chance variable \( X \),

and \( y_1, y_2, \ldots, y_n \)

\( n \) values of \( Y \), and if the distributions of \( X \) and \( Y \) are the same, then fiducial limits for
Our mean numbers are not a sample from one population, and so the probability of the separation being non-significant is 1 - \( \gamma \),

i.e., \( P \left\{ \begin{array}{c}
  h_1 + h = 82 \end{array} \right\} = 1 - \gamma \)

then \( h_1, h_2 \) are the values of \( h \) for which the separation is borderline.

Considering the sample

\[ 6.7, 2.4, 3.2, 8.7, 9.5, 14.6, 4.2 \]

and using the modified values

\[ 6, 11, 12, 20 \quad \text{and} \quad 16, 19, 22, 24, 29 \]

we have to determine the values of \( h \) which make

\[ h_1, h + 11, h + 12, h + 20 \]

borderline. Take \( h = 6 \), \( P = 1/21 \), then we find

\[ h_1 = .5 \quad h_2 = 22 \]

With the original numbers we should have

\[ h_1 = 6.15 \quad h_2 = 22 \]

and so

\[ P \left\{ \begin{array}{c}
  6.15 \leq h \leq 22 \end{array} \right\} = 1 - 1/21 \]
Sometimes our idealized picture of the experiment takes
the form that \( x, y \) each take only one value, i.e. they
are constants, and we want to know whether their values
are the same or not.

As an example, consider an agricultural experiment
in which a block of ground is divided into \( m \times n \)
equal plots which are all sown at the same rate with the
same variety of seed. To \( m \) of these plots, chosen at
random, treatment \( A \) is applied, and to the remaining \( n \)
treatment \( B \). The yields are as follows:

\[
\text{Treatment } A: \quad x_1, x_2, \ldots, x_m \\
\text{Treatment } B: \quad y_1, y_2, \ldots, y_n
\]

Now the yield from any plot may be regarded as the
sum of three terms, thus

\[
x_\nu = x'_\nu + a + e \\
y_\nu = y'_\nu + b + e
\]

where \( a \) is a constant which denotes the effect of
treatment \( A \) and \( b \) the effect of \( B \). \( x'_\nu \) denotes the
effect of the fertility and situation, etc., of the
particular plot, and \( y'_\nu \) is a random error affecting
all the \( x'_\nu \) and the \( y'_\nu \). Put

\[
x_\nu = h_\nu + a \\
y_\nu = h'_\nu + b
\]

The separation

\[
\begin{align*}
h_1 & \quad h_1 \\
h_m & \quad h_m \\
\end{align*}
\]

\[
x_1 - a_1, \ldots, x_m - a_m \\
y_1 - b_1, \ldots, y_n - b_n
\]
in a more or less separation. Hence the probability of a
both significant and P. This will not affect the
significance or non-significance by adding b to
each number.

As then have.

\[ x_i + (b-a) \geq \cdots \geq a + (b-a) \]

To test the hypothesis \( b = a \) to simplify put
\( b - a = 0 \) and determine whether the separation
is significant or not. To determine fiducial
limits for \( x_i - a \), to know what the separation
is non-significant, for

\[ b_i \leq b \leq b_i \]

and so the probability of this \( b \geq a \). In
such a case it seems wrong to ignore the
possibility of discontinuous variables.
So far we have been discussing the "open" application of the test, in which the two sets of observed numbers are really thought of as random samples of two chance variables in the "closed" application we wish to test as before the hypothesis that the chance variables $X, Y$ have the same distribution, or, alternatively that $X + h, Y$ have the same distribution, to determine fiducial limits for $h$.

The difference is that we now have error terms about which we know nothing and which may not be independent of one another. Our observations consist of two sets of numbers

$$q_1, \ldots, q_m$$

$$b_1, \ldots, b_n$$

where

$$q_r = x_r + k_r, r = 1, 2, \ldots, m$$

$$b_s = y_s + k_s, s = 1, 2, \ldots, n$$

and the $x$ are a random sample of $X$ and the $y$ of $Y$; but about the $k_r$ we know little or nothing. They may not be independent, but the experiment is so designed that the allocation of $k_r$ to the $x$ and the $y$ is by chance so that all such allocations are equally probable. Then the null hypothesis is true, $X$ and $Y$ are experimentally indistinguishable and the separation of the $m + n$ numbers $a$ and $b$ is simply labelling, and the probability of a significant
Instead of the error \( |\bar{u} - \bar{v}| \) we could use \( \bar{u} - \bar{v} \) and obtain a new test, which might be more suitable for some purposes. If we have positive values of \( \bar{u} - \bar{v} \), the difference between the values of \( \bar{u} \) and \( \bar{v} \) would be significant, and vice versa.

The test of the significance of the difference of the two means is in later involved in deriving \( t \). This will be greatly reduced to the use of one of the standard correlated distributions as an approximation to the true distribution of the errors.

25. THE PROBABILITY DISTRIBUTION OF \( |\bar{u} - \bar{v}| \)

Instead of the spread \( |\bar{u} - \bar{v}| \) we use in practice \( |\bar{u} - \bar{z}| \), which is proportional to it, or we may use any monotonic function of this such as \( (\bar{u} - \bar{z})^2 \).

Let

\[ y_i = z_i - \mu \]

be some numbers with mean \( \bar{z} \), and let their moments about their mean be \( \mu_2 \), \( \mu_3 \), etc. Suppose that a sample of \( z \) is drawn without replacement. Denoting the mean of the sample by \( \bar{u} \), we have...
\((1) \quad E (\bar{U} - \bar{T}) = 0\)

\((2) \quad E \left\{ (\bar{U} - \bar{Z})^2 \right\} = \frac{\eta M_z}{(m-n-1)m}\)

since \(\bar{U}, \bar{Z}\) is independent of the \(z\) origin, we may in working these out, temporarily change the origin so that \(\bar{T} = 0\); \(\mu_z, \sigma_z\), etc. will not be affected. We then have

\[E (\bar{U} - \bar{Z})^2 \right\} = E \left\{ \left( \frac{\bar{U} - \bar{Z}}{m} \right)^2 \right\} = \frac{\eta M_z}{m} + \frac{\sigma_z^2}{m} E (\bar{Z}, \bar{Z}),\]

where temporarily, \(z_1, z_2\) denote values of the modified \(z\) taken at random.

\[E (z, z) = \frac{(\bar{Z} - \bar{Z})^2 - \bar{Z} \bar{Z}}{(m)(m+n-1)} = -\frac{M_z}{m+n-1},\]

from which the result stated follows. Higher moments may be obtained in the same way.

Returning now to the original \(z\) values, we have

\[M_z = E \left\{ \frac{(\bar{U} - \bar{Z})^2 + \eta (\bar{Z} - \bar{V})^2}{m + n} \right\}\]

for the expression on the right is the value the variances of \(z\) would have if \(n\) of them were replaced by their mean
\( \bar{u} \), and the other \( n \) by their mean \( \bar{v} \).

\[ M_2 = \frac{m(\bar{u} - \bar{v})^2}{n} \]

Therefore

\[ (\bar{u} - \bar{v})^2 \leq \frac{\eta M_2}{m} \]

Put

\[ w = \frac{m(\bar{u} - \bar{v})^2}{\eta M_2} \]

then

\[ 0 \leq w \leq 1 \]

\[ E(w) = \frac{1}{m+n-1} \]

It can be shown (PIOTRA, Supplement to Journal of Royal Stat. Soc. 1937 pp 119-136) that

\[ E(w^2) = \frac{3}{(m+n-1)(m+n+1)} \]

approximately

if \( \frac{M_2}{\eta} \) is not large and \( m, n \) large. In the

\( B(\frac{1}{2}, \frac{1}{2}(m+n-2)) \) has the same range and the same

first moment as \( w \). Its second moment is

\[ \frac{3}{(m+n-1)(m+n+1)} \]

and so we may expect the distribution of \( w \) to be approximately a

\( B(\frac{1}{2}, \frac{1}{2}(m+n-2)) \) distribution.

If we have a pair of samples

\( u_1, \ldots, u_m \)

\( v_1, \ldots, v_n \).
an alternative expression for \( w \) is

\[
  w = \frac{(\overline{u} - \overline{v})^2 m n}{\mu_s (m+n)^2} = \frac{(\overline{u} - \overline{v})^2 \frac{m n}{m+n}}{S_1 + S_2 + \frac{m n}{m+n} \overline{u} - \overline{v})^2}
\]

where \( S_1, S_2 \) are the variances of the samples.

The statistic \( t \) which is used for testing the hypothesis that the two samples come from the same normal population is

\[
  t = \frac{(\overline{u} - \overline{v}) \sqrt{\frac{m+n}{m+n-2}}}{\sqrt{S_1 + S_2}}
\]

We can then easily show that

\[
  w = \frac{t^2}{t^2 + m+n-2}
\]

But when the two samples do come from the same normal population the distribution of

\[
  \frac{t^2}{t^2 + m+n-2}
\]

is exactly a \( B\left(\frac{1}{2}, \frac{1}{2}(m+n-2)\right) \) distribution. Therefore the use of this approximate distribution of \( w \) is exactly equivalent to the use of the \( t \) test; but we are here making no assumptions about normality. We are merely assuming that the distribution of \( w \) for the given values of \( u \) and \( v \) is approximately a \( B\left(\frac{1}{2}, \frac{1}{2}(m+n-2)\right) \) distribution.
16. CONSISTENCY OF THE SAMPLING TEST

Suppose that
\[ x_1, \ldots, x_m \]
\[ y_1, \ldots, y_n \]
are a pair of samples of chance variables \( X, Y \) whose means are \( \mu_1, \mu_2 \) and whose standard deviations are \( \sigma_1, \sigma_2 \) respectively.

For given \( x, y \) the mean value of \( \bar{u} - \bar{v} \) is 0, and the mean value of \( (\bar{u} - \bar{v})^2 \) is (see § 15 (2) and use
\[ m(\bar{u} - \bar{v}) = n(\bar{v} - \bar{u}) = \frac{mn(\bar{u} - \bar{v})}{m+n} \]

\[ (m+n)^2 / \frac{m+n}{m+n-1} \frac{m}{m+n} \]

Therefore
\[ \begin{align*}
\Pr \{ |\bar{u} - \bar{v}| \geq k \} & = \frac{(m+n)^2 / \frac{m+n}{m+n-1} \frac{m}{m+n} \alpha}{k^2} \\
\end{align*} \]

Hence if we take \( k \) so that the last expression is equal to \( \alpha \), i.e.

\[ k = (m+n) \sqrt{\frac{\mu_2}{(m+n-1)mn \alpha}} \]

\[ \Pr \{ |\bar{u} - \bar{v}| \geq k \} \leq \alpha, \]

and values of \( |\bar{u} - \bar{v}| \) which are \( \geq k \) will certainly be significant at the level \( \alpha \).

\[ \begin{align*}
M_2 & = \frac{\sum x_i^2 + \sum y_i^2 - (\frac{m \bar{x} + n \bar{y}}{m+n})^2}{m+n} \\
& = \frac{m}{m+n} \frac{\sum x_i^2}{m} + \frac{n}{m+n} \frac{\sum y_i^2}{n} - (\frac{m \bar{x} + n \bar{y}}{m+n})^2
\end{align*} \]
If $m \rightarrow \infty$ in constant ratio $\frac{m}{a} = \frac{n}{b}$, $\mu_2$ will tend in probability to

$$\frac{a}{a+b} (s_1^2 + k_1^2) + \frac{b}{a+b} (s_2^2 + k_2^2) - \left(\frac{a k_1 + b k_2}{a+b}\right)^2$$

$$= \frac{a}{a+b} s_1^2 + \frac{b}{a+b} s_2^2 + \frac{a}{a+b} (k_1 \cdot k_2)^2$$

i.e., to a finite limit. Therefore $k \rightarrow 0$ in probability as $m, n \rightarrow \infty$.

On the other hand, the observed spread,

$$|\bar{x} - \bar{y}| \rightarrow |h_1 - h_2| \quad \text{in probability},$$

hence if $h_1 \neq h_2$, the probability of a significant result

$$\rightarrow 1 \quad \text{as } m, n \rightarrow \infty.$$ Thus the spread test of the hypothesis

"the distribution of $X$ is the same as the distribution of $Y$" is consistent with respect to the alternative "$X, Y$ have finite variances but different mean values".

Then $h_1 = h_2$, the observed spread $|\bar{x} - \bar{y}| \rightarrow 0$ in probability as $m, n \rightarrow \infty$; but so does $k$, and the above argument is not delicate enough to show that the test is not then consistent.

Let us consider $\sqrt{m+n} (\bar{u} - \bar{v})$

For given values of the $x$ and the $y$, its mean value is $\gamma$, and its variance is

$$\frac{(m+n)^3 M_2}{(m+n-1) mn}$$
If \( x = b \) or \( y = c \), this is \( \lambda \) and so the probability of a significant result \( \to \alpha \). In any case, the probability of a significant result will not \( \to 1 \) as \( n \to \infty \) and so the test is not consistent if \( x \), \( y \) have different distributions with finite variances and the same mean values.

17. THE USE OF RANKS IN THE SPEP'T TEST

In the exact spread test the distribution of \( |\bar{U} - \bar{T}| \) or of \( W \) depends on the sample values, and so varies from sample to sample. If we replace each observed number \( x_1, x_2, \ldots, x_n \),

\( y_1, y_2, \ldots, y_n \)

by its rank in the whole set of \( m + n \) numbers, we obtain the separation

\[ r_1, r_2, \ldots, r_m, s_1, s_2, \ldots, s_n, \]

where \( r_k \) is the rank of \( x_k \) and \( s_k \) the rank of \( y_k \). This can be treated as before, but since the \( r \) and the \( s \) together consist of the \( m + n \) integers from 1 to \( m + n \), the "labelling" distribution will be the same for all samples of the same sizes. It can therefore be worked out once for all and tabulated. We have assumed no ties.

The test is then identical with that proposed by Wilcoxon (Biometrics Bulletin 1945 pp. 80-83) and by Mann and Whitney (Annals of Math. Stat. 1947 pp. 50-60). Wilcoxon uses \( T^- = \sum_{i=1}^{n} s_i = n \bar{V} \)
in our notation. In this and $U_i$ they use

$$U = \text{number of times } y \text{ precedes an } x.$$  

$$= \sum_{i=1}^{N} (v_i - i) = m \left( \bar{u} - \frac{1}{2} (m+1) \right)$$

Also

$$\sum_{i=1}^{N} W_{ij} = u_{ij}$$

where $u_{ij} = 0$ if $x_i \leq y_j$

$$= 1 \quad \text{if } x_i > y_j$$

For given $x,y$ the distribution of $U$ is symmetrical, because that of $U_i$ is. Also

$$E(U) = m \left[ E(\bar{u}) - \frac{1}{2} (m+1) \right] = m \left( \frac{m+n+1}{2} - \frac{m+1}{2} \right)$$

$$= \frac{mn}{2}$$

$$V(U) = m^2 \quad V(\bar{u}) = \frac{mn M_2}{mn - 1} \quad (\text{15.12})$$

Here $M_2 = \frac{1}{2} (m+n+1)(2m+3n+1) - \frac{1}{4} (m+n+1)^2$

$$= \frac{(m+n+1)(m+n-1)}{12}$$

$$V(U) = \frac{mn (m+n+1)}{12}$$

$$E\left( \frac{U}{mn} \right) = \frac{1}{2}, \quad V\left( \frac{U}{mn} \right) = \frac{(m+n+1)}{12mn} \to 0$$

As $m, n \to \infty$.

If we use $\frac{U}{mn}$ for the test, and if values $\geq \chi$ are significant for a level $\alpha$, $\chi \to \frac{1}{2}$ as $m, n \to \infty.$
Now let us consider the population distribution of the observed value of $U/mn$ in the general case when the null hypothesis is not necessarily true.

Let $p = \{X_i > Y_j\} = P\{X_i - Y_j > 0\}$. 

$$ U = \sum \sum U_{ij} $$

where $U_{ij}$ is a chance variable with values 0, 1 and corresponding probabilities $1-p$, $p$. Therefore

$$ E(U_{ij}) = p = E(U_{ij})^* $$

$$ E(U) = \sum \sum E(U_{ij}) = mn\bar{p} $$

$$ E(U^2) = \sum \sum E(U_{ij}^2) = \sum \sum E(U_{ij} U_{ik}) + \sum \sum E(U_{ij} U_{ji}) $$

$$ = mn\bar{p} + m(m-1)n(n-1)\bar{p}^2 + m(m-1)n\bar{p}^2 + mn(n-1)\bar{p} $$

where $\bar{p}^1 = P\{(X_i > Y_j)(X_i > Y_j)\}$, $\bar{p}^2 = P\{(X_i > Y_j)(X_i < Y_j)\}$, $\bar{p}^3 = P\{(X_i < Y_j)(X_i > Y_j)\}$, $\bar{p}^4 = P\{(X_i < Y_j)(X_i < Y_j)\}$.

$$ V(U) = mn\left\{p + (m-1)(n-1)\bar{p}^2 - mn\bar{p}^2 + (m-1)\bar{p} + (n-1)\bar{p}^4\right\} $$

$$ V\left(\frac{U}{mn}\right) = \frac{1}{mn}\left\{(m-1)(\bar{p}^1 - \bar{p}^2) + (n-1)(\bar{p}^3 - \bar{p}^4) + \bar{p} - \bar{p}^2\right\} $$

$$ \rightarrow 0 \text{ as } m, n \rightarrow \infty. $$

Thus $U/mn \rightarrow p$ in probability as $m, n \rightarrow \infty$.

Hence the one-way test $U/mn > \kappa$ is consistent if $p > \frac{1}{2}$, and the two-way test $|U/mn| > \kappa$ is consistent if $p \neq 1/2$. 
\[ G(\lambda) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2} x^2} dx. \]

Therefore if
\[ \lambda_n = \frac{k_n - \psi_n(\theta_0)}{\sigma_n(\theta_0)} \]

\[ G(\lambda_n) \rightarrow \alpha \quad \text{as} \quad n \rightarrow \infty. \]

Hence \( \lambda_n \rightarrow \lambda \) when \( G(\lambda) = \alpha. \)

Now put \( \theta = \theta_0 + \frac{k}{\sqrt{n}}, \quad \kappa > 0, \) constant.

\[ P \left\{ T_n \geq k_n \mid \theta_0 + \frac{k}{\sqrt{n}} \right\} \sim G(\lambda'_n) \]

where \( \lambda'_n = \frac{k_n - \psi_n(\theta_0 + \frac{k}{\sqrt{n}})}{\sigma_n(\theta_0 + \frac{k}{\sqrt{n}})} \)

\[ = \frac{\psi_n(\theta_0) - \frac{k}{\sqrt{n}} \psi'_n(\theta_0 + \varepsilon \frac{k}{\sqrt{n}})}{\sigma_n(\theta_0 + \frac{k}{\sqrt{n}})}, \quad 0 \leq \varepsilon \leq 1 \]

\[ \rightarrow \lambda - \kappa c. \]

Thus power for \( \theta = \theta_0 + \frac{k}{\sqrt{n}} \rightarrow G(\lambda - \kappa c) \), and this limit for given \( \lambda, \kappa \) depends only on \( \theta_0. \)

If we have two tests of same hypothesis, and for the same power with respect to the same alternative one requires a sample of \( n \) and the other a sample of \( n' > n \), the relative efficiency of the second test with respect to the first is \( n/n' \).
Put \[ \frac{\psi_n'(\theta_0)}{\sigma_n(\theta_0)} = H_1(n). \]

Denote the corresponding function for the second test by \( H_2(n) \).

Suppose the second test satisfies the same conditions but that \[ \frac{H_2(n)}{\sqrt{n}} \to c' \]

**1st test**
\[ \theta = \theta_0 + \frac{k}{\sqrt{n}} \]

**2nd test**
\[ \theta = \theta_0 + \frac{k'}{\sqrt{n'}} \]

\[ \therefore \frac{k}{\sqrt{n}} = \frac{k'}{\sqrt{n'}} \quad \text{for some} \ \theta. \]

For the limiting powers to be equal \( kC = k'C' \).

Therefore
\[ \frac{\sqrt{n}}{\sqrt{n'}} = \frac{k}{k'} = \frac{c'}{c} = \frac{\lim_{n \to \infty} H_2(n)}{\lim_{n \to \infty} H_1(n)} \]

\[ = \lim_{n \to \infty} \frac{H_2(n)}{H_1(n)} \]

and
\[ \frac{n}{n'} = \lim_{n \to \infty} \frac{H_2(n)^2}{H_1(n)^2}. \]

The function \[ H_1(n)^2 = \frac{\psi_n'(\theta_0)^2}{\sigma_n(\theta_0)^2} \]

may be called the **efficacy** of the statistic \( T_n \). The asymptotic relative efficiency is the limit of the ratio of the efficacies.
If \( X, Y \) are normal variables with same variance \( \sigma^2 \) and
\[ Z(X) - \mathbb{E}(Y) = \theta, \quad X-Y \text{ has mean } \theta \text{ and variance } 2\sigma^2. \] If we use the rank spread test to test the hypothesis \( \theta = 0 \), it can be shown that the statistic \( \frac{U}{mn} \) satisfies the required conditions.

\[ E\left( \frac{U}{mn} \right) = \frac{1}{\sqrt{2\pi}} = \mathbb{P}(X-Y > 0) = \int_{-\theta}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \]

\[ \frac{dE}{d\theta} = \frac{1}{\sqrt{2\pi}m} e^{-\frac{\theta^2}{4\sigma^2}} \]

which is \( \frac{1}{\sqrt{m+n}} \) when \( \theta = 0 \).

\[ \sqrt{\frac{U}{mn}} = \frac{m+x+1}{12 mn} \quad \text{when } \theta = 0 \]

and so the efficacy of the rank spread test is

\[ \frac{3 mn}{\pi (m+n+1) \sigma^2}. \]

If we use the \( t \) test

\[ t = \frac{(\bar{X} - \bar{Y}) \sqrt{mn}}{\sqrt{S_1 + S_2 \over m+n-2}} \]

\[ = \frac{(\bar{X} - \bar{Y} - \theta) \sqrt{mn}}{\sqrt{S_1 + S_2 \over m+n-2}} + \frac{\theta \sqrt{mn}}{\sqrt{S_1 + S_2 \over m+n-2}} \]

The first term is asymptotically standard normal, while the denominator of second term tends in probability to \( \sigma^2 \).
hence the efficacy of the t-test

\[ \sim \frac{m n}{(m+n) \sigma^2} \]

and so relative asymptotic efficiency of the rank spread test is

\[ \frac{3}{\pi} = 0.95. \]

Let \( Y, Z \) be independent continuous chance variables with frequency functions in \( x, f(x), g(x) \).

\[ \frac{\partial}{\partial \theta} \mathbb{P} \left\{ Y - Z \leq \theta \right\} = \text{frequency function of } Y - Z \text{ at } Y - Z = \theta = \int f(x) g(x - \theta) \, dx. \]

\[ \frac{\partial}{\partial \theta} \mathbb{P} \left\{ Y - Z \leq \theta \right\} \bigg|_{\theta = 0} = \int_{-\infty}^{\infty} f(x) g(x) \, dx. \]

Suppose now that \( X, Y \) are continuous chance variables with distributions which differ only in location and that their frequency functions are \( f(x - \theta), g(x) \). If we apply the rank spread test to test \( \theta = 0 \),

\[ E \left( \frac{U}{mn} \right) = \mathbb{P} \left\{ X \geq Y \right\} = \mathbb{P} \left\{ Z \geq Y - \theta \right\} = \mathbb{P} \left\{ Y - Z \leq \theta \right\} \]

where \( Z = X - \theta \) has the same distribution as \( Y \).

Hence

\[ \frac{\partial f}{\partial \theta} \bigg|_{\theta = 0} = \int_{-\infty}^{\infty} f(x)^2 \, dx. \]

\[ V \left( \frac{U}{mn} \right)_{\theta = 0} = \frac{m+n+1}{12 mn} \]

and the efficacy = \[ \frac{12 mn}{m+n+1} \left( \int_{-\infty}^{\infty} f(x)^2 \, dx \right)^2. \]

For the t test applied to same populations of finite variance \( \sigma^2 \)

efficacy \( \sim \frac{m n}{(m+n) \sigma^2}. \)

Thus the asymptotic relative efficiency of the rank spread test compared to the t test is

\[ 12 \left( \sigma \int_{-\infty}^{\infty} f(x)^2 \, dx \right)^2. \]
Note that \( \int_{-\infty}^{\infty} f(x)^2 \, dx \) is independent of scale.

For a rectangular population
\[
f(x) = \begin{cases} 1, & -\frac{1}{2} \leq x \leq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}
\]
\[
\int_{-\infty}^{\infty} f(x)^2 \, dx = 1, \quad \sigma^2 = \frac{1}{12}
\]
and the asymptotic relative efficiency is 1.

For a \( \Gamma(\alpha) \) population
\[
f(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}, \quad 0 \leq x
\]
the asymptotic relative efficiency is \( \frac{81}{64} \). Here the rank spread test is more efficient than the spread test which is asymptotically equivalent to the t test.

It should be noted that the rank spread test can be applied to populations like the Cauchy population which have infinite variance. It was only while comparing it to the t test and the spread test that we assumed finite variance.

19 THE WALD AND WOLKOWITZ TEST

We have a pair of samples
\[
x_1, \ldots, x_n
\]
\[
y_1, \ldots, y_n
\]
and we wish to test the hypothesis that the distribution of \( X \) is the same as the distribution of \( Y \). Arrange the \( n+m \) numbers in order of magnitude
\[
z_1, z_2, \ldots, z_{m+n}
\]
then count the number of runs, a run being defined as a succession of one or more \( X \)'s or of one or more \( Y \)'s. We assume that no two \( Z \)
are equal, which will be true with probability 1 when we are sampling from continuous populations. In counting the runs we only need to know whether any \( z \) is an \( x \) or a \( y \), and so we may replace it by \( x \) or \( y \) as the case may be. For the pair of samples

\[
\begin{align*}
 x & \quad 0.8, 1.0, 1.1, 1.6 \\
 y & \quad 0.9, 1.2, 1.4, 1.5, 1.8, 2.0
\end{align*}
\]

we have the arrangement

\[
0.8, 0.9, 1.0, 1.1, 1.2, 1.4, 1.5, 1.6, 1.8, 2.0
\]

i.e.,

\[
x y x x y y x y y
\]

which has 6 runs.

If the null hypothesis is true all permutations of the \( n \) \( x \)'s and \( n \) \( y \)'s are equally probable. The number of such arrangements is \( m+n \binom{m+n}{m} \). We can determine the exact distribution of \( U \), the number of runs, when the null hypothesis is true. If the distributions of \( X \) and \( Y \) are different we should expect bunching of \( X \) values and \( Y \) values, and fewer runs.

20. The distribution of \( U \) under the null hypothesis is true.

The number of ways in which \( n \) like objects can be separated into \( k \) different sets, each set non-empty, is \( m! \binom{m}{k-1} \). Make a row of \( n \) crosses to represent the \( n \) like objects. Then mark the first object in each set by putting a tick above. A tick must go above the first cross, and therefore the places for the ticks can be \( x \) chosen in \( m! \binom{m}{k-1} \) ways.

If we are to arrange \( m \) \( x \)'s and \( n \) \( y \)'s in a row so as to have 2 \( k \) runs, we may first separate the \( x \) into \( k \) different non-empty sets and the \( y \) also into \( k \) different non-empty sets, and then interlace the two \( m+n \) sets. The latter operation may be done in \( 2^m \) ways. Hence

\[
\Pr\{ U = 2k \} = \frac{2^{m-1} \binom{m-1}{k-1} \cdot n^{-1} \binom{n-1}{k-1}}{m+n \binom{m+n}{m}}
\]
Similarly

\[ P(U = 2k - 1) = \frac{m-1 \binom{C_{k-2}}{k-1} + m-1 \binom{C_{k-1}}{m-1} \binom{C_k}{k-2}}{m+n \binom{C_m}{m}} \]

since we must have either \(k\) sets of \(x\) and \(k\) sets of \(y\) or \(k\) sets of \(x\) and \(k-1\) sets of \(y\).
Let $L = u_1, u_2, \ldots, u_n$ where $u_r = 1$ if a run starts at place $r$, $u_r = 0$ otherwise. Let $r > 1$. It is a chance function which takes the values 0, 1. If $u_r = 1$, the places $r-1,$ $r$ such that filled by $y$ or by $y$, $x$. Hence

$P(u_r = y) = \frac{2n}{(m+n)(m+n-1)}$, $r > 1$. 

$E(u_r) = \frac{2n}{(m+n)(m+n-1)} = \frac{E(U_r^2)}{r > 1}$

$\sum_{r=1}^{n} E(u_r) = 1 + \frac{2mn}{m+n}$. 

$\sum_{r=1}^{n} (u_r - 1) = \sum_{r=1}^{n} u_r$. 

$E(u_r^2) = \sum_{r=1}^{n} E(u_r^2) = 2 \sum_{r=1}^{n} E(u_r U_{r+1}) + \sum_{r=1}^{n} \sum_{s=r+1}^{n} E(u_r U_s)$. 

$\sum_{r=1}^{n} E(u_r^2) = \frac{2mn}{m+n}$. 

$U_r U_{r+1}$ takes the values 0, 1; therefore

$P(U_r U_{r+1}) = P(U_r = 1, U_{r+1} = 1) = P(U_r = 1) P(U_{r+1} = 1)$. 

For both $u_r$, $U_{r+1}$ to be 1 we must have one of the following arrangements in places $r-1$, $r$, $r+1$:

- $r-1$ $r$ $r+1$ 
  - $x$ $y$ $x$ 
  - $y$ $x$ $y$ 

The multiplicity of this is

$$\frac{m(m-1)n^2 \cdot \frac{n}{m+n}}{m+n} \cdot (m+n-2) = \frac{mn}{(m+n)(m+n-1)}.$$
\[
\begin{align*}
\text{(2)} & \quad \sum_{r=0}^{\min(m, n)} 2 \mathbb{E}(u_r | u_{r+1}) = \frac{2 \min(m, n, n-2)}{(m+n)(m+n-1)}. \\
\end{align*}
\]

\[
\begin{align*}
\mathbb{E}(u_r | u_{r+1}) &= \mathbb{P}(u_r = 1 | u_{r+1}) \\
&= \frac{m(m-1)}{(m+n)(m+n-1)(m+n-2)(m+n-3)}. \\
\end{align*}
\]

By using the above-represented equations, we must have one of the following:

\[
\begin{align*}
\vdots & \quad \vdots \\
z & \quad y \\
\vdots & \quad y \\
\vdots & \quad y \\
\end{align*}
\]

Number of cases:

\[
\begin{align*}
\text{for even } u_r, u_{r+1} &= \frac{\min(m, n-2)}{2} (m+n-r) \\
&= (m \cdot r - 3)(m+n-2) \\
&= \frac{(m+n-3)(m+n-2)}{2}. \\
\end{align*}
\]

\[\sum \mathbb{E}(u_r | u_{r+1}) = \frac{4 \min(m, n-1)(m+n-1)}{(m+n)(m+n-1)}.\]

From (1) (2) (3) we get:

\[
\begin{align*}
\mathbb{E}(u_r | u_{r+1}) &= \frac{2 \min(m, n-1)}{(m+n)(m+n-1)}. \\
\mathbb{E}(u_r | u_{r+1}) &= \frac{2 \min(2m-n, n)}{(m+n)(m+n-1)}. \\
\end{align*}
\]

\[
\mathbb{E}(u_r | u_{r+1}) = \frac{2 \min(2m-n, n)}{(m+n)(m+n-1)}. \\
\]
Then \[ \frac{m \cdot n}{m+n} + \frac{1}{m+n} \rightarrow \frac{2ab}{(ab)^2}, \]
\[ \sqrt{\frac{m+n}{m^2}} \rightarrow 0. \]

21. The distribution of \( U \) when the null hypothesis is \( H_0 \) true.

Suppose that \( x, y \) are continuous chance variables with frequency functions \( f(x), g(x) \), and that we have a sample of \( m \) values of \( x \) and \( n \) values of \( y \), and that the number of runs is \( U \).

We shall assume that the number of points of discontinuity of \( f(x) \) is finite, not counting an end point of an interval throughout which \( f(x) \equiv 0 \) if \( f(x) \) is continuous at that end-point in the adjacent interval. We make the same assumption about \( g(x) \).

Corresponding to any positive \( \varepsilon \), there exists a set \( I \) of a finite number of intervals such that

\[ P\{X \in J_0\} \leq \varepsilon, \quad \varepsilon = P\{Y \in J_2\} \leq \varepsilon, \]

and \( f, g \) are bounded outside \( I_0 \). If \( U_0 \) denotes the number of runs starting in \( I_0 \), \( U_0 \leq m' + n' \), where \( m', n' \) are the numbers of \( x \) value's and \( y \) values in \( I_2 \). Therefore

\[ E(U_0) \leq E(m' + n') = pm + qn \leq E(m + n). \]

Thus

\[ \frac{E(U_0)}{m+n} < \varepsilon. \]

Let \( I_1 \) be any interval of length \( \frac{1}{m+n} \) belonging to \( I \), the \( \text{im} \) complement of \( I_0 \). Denote by \( U_1 \) the number of runs starting in \( I_1 \), and by \( U_1' \) the number of runs starting in \( I_1 \) when we do not count a run which starts at the first \( (x \text{ or } y) \) value in \( I_1 \). Obviously

\[ U_0 = U_1' \quad \text{or} \quad U_1' + 1. \]
First suppose that \( f, g \) are both constant over \( I_r \). For given numbers \( m', n' \) of \( x \) values and \( y \) values in \( I_r \), all permutations of the \( x \) and \( y \) values are equally probable, and so for fixed \( m', n' \) the mean value of \( U_r' / (m+n) \) is

\[
\frac{2 m' n'}{(m'+n')(m+n)} = \frac{2 \frac{m'}{m} \cdot \frac{n'}{n}}{(\frac{m'}{m} m + \frac{n'}{n} n)} \cdot \frac{m n}{m+n} = \frac{m'}{m} \cdot \frac{n'}{n} \cdot \frac{2 a b}{(a+b)}.
\]

The chance variable

\[
\frac{m' \cdot n'}{m a + n' b}
\]

is bounded, being always non-negative and \( \leq 1/a \). Also as \( m, n \to \infty \) it tends in probability to

\[
\frac{h_r \cdot g h_r}{f h_r a + g h_r b} = \frac{f g h_r}{a f + b g}.
\]

Hence its mean value must tend to the same limit. Thus

\[
lim E \left\{ \frac{U_r'}{(m+n)} \right\} = \lim E \left\{ \frac{U_r'}{m+n} \right\} = \frac{2 a b}{a f + b g} \cdot h_r.
\]

In the same way we can show that

\[
E \left\{ \left( \frac{U_r'}{m+n} \right)^2 \right\} = E \left\{ \frac{2 m' n' (2 m' n' - 1)}{(m'+n')(m'+n'-1)} \right\} \frac{1}{(m+n)^2}
\]

\[
= \frac{4 a^2 b^2}{(a+b)^2} E \left\{ \frac{m n (n - \frac{1}{2mn})}{(m a + n b)(m a + n' b - a/m)} \right\}
\]

and that the limit of this is

\[
\frac{4 a^2 b^2}{(a+b)^2} \cdot \frac{f g^2}{(a f + b g)^2} \cdot h_r^2
\]

\[
= \lim E \left\{ \left( \frac{U_r}{m+n} \right)^2 \right\}.
\]
Consider now the general case where \( f, g \) are not necessarily constant over \( I_r \). If the value of \( f(x) \) is increased throughout \( I_r \)--with a compensating decrease elsewhere--both \( E(U_r) \) and \( E(U_r^2) \) will be increased, for this is equivalent to throwing in more \( x \) values in this interval, which can never decrease and may increase the number of runs. A similar statement applies to \( g \). In particular, the mean values of \( U_r \) and \( U_r^2 \) will be increased when \( f(x), g(x) \) are replaced over \( I_r \) by their supremum values \( \tilde{f}_r, \tilde{g}_r \) in this interval. Therefore
\[
\lim \sup E \left\{ \frac{U_r}{m+n} \right\} \leq \frac{2ab}{a+b} \frac{\tilde{f}_r \tilde{g}_r}{af_r + bg_r} h_r
\]
\[
\lim \sup E \left\{ \frac{(U_r^2)}{m+n} \right\} \leq \frac{4a^2b^2}{(a+b)^2} \frac{\tilde{f}_r^2 \tilde{g}_r^2}{(af_r + bg_r)^2} h_r^2
\]
Subdividing \( I \) into intervals, we get
\[
\lim \sup E \left\{ \frac{U-U_k}{m+n} \right\} \leq \frac{2ab}{a+b} \sum \frac{\tilde{f}_r \tilde{g}_r}{af_r + bg_r} h_r.
\]
Similarly
\[
\lim \inf E \left\{ \frac{U-U_k}{m+n} \right\} \geq \frac{2ab}{a+b} \sum \frac{\tilde{f}_r \tilde{g}_r}{af_r + bg_r} h_r.
\]
where \( f, g \) are the infimum values of \( f(x), g(x) \) in \( I_r \).
Taking each \( h_r \to 0 \), we have
\[
\lim E \left\{ \frac{U-U_k}{m+n} \right\} = \frac{2ab}{a+b} \int_1^\infty \frac{f(x)g(x)}{af(x) + bg(x)} \, dx.
\]
Now making \( \varepsilon \to 0 \), we obtain
\[
\lim E \left\{ \frac{U}{m+n} \right\} = \frac{2ab}{a+b} \int_0^\infty \frac{f(x)g(x)}{af(x) + bg(x)} \, dx.
\]
Note that
\[
\frac{(a+b)^2f+ab(f-g)^2}{af+bg} = \frac{abf^2 + (a^2+b^2)fg + abg^2}{af + bg} = bf + ag.
\]

Therefore
\[
\int_{-\infty}^{\infty} \frac{fg}{af+bg} \, dx = \int_{-\infty}^{\infty} \frac{bf + ag}{(a+b)^2} \, dx - \frac{ab}{(a+b)^2} \int_{-\infty}^{\infty} \frac{(f-g)^2}{af+bg} \, dx.
\]

Hence
\[
lm E \left\{ \frac{V}{m+n} \right\} = \frac{2ab}{(a+b)^2} - \frac{2a^2b}{(a+b)^3} \int_{-\infty}^{\infty} \frac{(f-g)^2}{af+bg} \, dx < \frac{2ab}{(a+b)^2}
\]

unless \( f(x) \equiv g(x) \). Put
\[
H(x) = \frac{2ab}{a+b} \frac{f(x)g(x)}{af(x) + bg(x)},
\]
then \( H(x) \) is continuous at all points where both \( f, g \) are continuous.

Corresponding to any positive \( \varepsilon \) there exists a set \( I_0 \) of intervals such that
\[
\mu(x) \, dx < \varepsilon
\]
and \( H(x) \) is continuous and bounded outside \( I_0 \) with a positive infimum. Let \( K, K \) be its supremum and infimum values outside \( I_0 \). Denote by \( U \) the number of runs starting in \( I_0 \), and by \( \sigma^2 \) the variance of \( U/(m+n) \), then
\[
lm E \left\{ \frac{U}{m+n} \right\} = \mu(x) \, dx < \varepsilon.
\]

Since
\[
0 \leq \frac{U}{m+n} \leq 1
\]
\[
lm \sup E \left\{ \left( \frac{U}{m+n} \right)^2 \right\} < \varepsilon.
\]
Hence when \( m, n \) are large, \( \sigma \leq 2 \sqrt{\delta} \).

Let \( I_r \) be any interval of length \( h_r \) in \( I \), the complement of \( I_\alpha, U_r \) the number of runs starting in \( I_r \), \( m_r, \sigma_r \) the mean and variance of \( U_r/(m+n) \).

\[
\lim m_r = \int_{I_r} H(x) \, dx \geq k h_r
\]

and also
\[
\sigma_r \geq H, h_r
\]
where \( H \) is the \( \inf \) infimum of \( H(x) \) in \( I_r \). We have already shown that

\[
\lim \sup \mathbb{E} \left\{ \left( \frac{U_r}{m+n} \right)^2 \right\} \leq H_r^2 h_r^2
\]

where \( H_r \) is the supremum of \( H(x) \) in \( I_r \), hence

\[
\lim \sup \frac{\sigma_r^2}{m_r} \leq \frac{H_r^2 - H^2}{k^2} \leq \frac{2 k (H_r^2 - H^2)}{k^2} < \epsilon^2
\]

for all \( r \) when the \( h_r \) are sufficiently small because of the uniform continuity of \( H(x) \) in \( I \). Thus

\[
\lim \sup \left\{ \sup \left\{ \frac{\sigma_r}{m_r} \right\} \right\} < \epsilon
\]

when the \( h_r \) are sufficiently small. But

\[
\frac{\sigma_1 + \sigma_2 + \cdots}{m_1 + m_2 + \cdots} \leq \sup \frac{\sigma_r}{m_r}
\]

and so

\[
\lim \sup \frac{\sigma_1 + \sigma_2 + \cdots}{m_1 + m_2 + \cdots} < \epsilon.
\]

If \( \sigma \) is the \( \text{D. D.} \) of \( U/(m+n) \),

\[
\sigma \leq \sigma_0 + \sigma_1 + \sigma_2 + \cdots
\]

and

\[
\mathbb{E} \left( \frac{U}{m+n} \right) \geq m_1 + m_2 + \cdots
\]

Therefore

\[
\sigma \leq 2 \sqrt{\delta} \sigma_0 + \frac{\sigma_1 + \sigma_2 + \cdots}{m_1 + m_2 + \cdots} \mathbb{E} \left( \frac{U}{m+n} \right).
\]

Hence

\[
\lim \sup \sigma \leq 2 \sqrt{\delta} \epsilon + \epsilon \lim \mathbb{E} \left( \frac{U}{m+n} \right),
\]

and so \( \sigma \to \sigma_0 \) as \( m, n \to \infty \). We therefore have that

\[
\frac{U}{m+n} \to \lim \mathbb{E} \left\{ \frac{U}{m+n} \right\} = \frac{2ab}{a+b} \int_{-\infty}^{\infty} \frac{f(x)g(x) \, dx}{af(x)+bg(x)}
\]
in probability as \( m, n \rightarrow \infty \), and this is less than \( \frac{2ab}{(a+b)^2} \) when \( f(x) \neq r(x) \).

If the null hypothesis is true \( \frac{U}{m+n} \rightarrow \frac{2ab}{(a+b)^2} \) in probability. Hence if we determine \( \lambda(m,n) \) such that when the null hypothesis is true
\[
\rho \left\{ \frac{U}{m+n} \leq \lambda(m,n) \right\} = \alpha_n \rightarrow \alpha
\]
as \( m, n \rightarrow \infty \), \( \lambda(m, n) \) will \( \rightarrow \frac{2ab}{(a+b)^2} \). Therefore, when \( f(x) \neq r(x) \)
\[
\lim \rho \left\{ \frac{U}{m+n} \leq \lambda(m,n) \right\} = 1
\]
and so the test is consistent against any alternative \( f \neq r \), when values of \( U/(m+n) \) less than \( \lambda(m,n) \) are taken as significant.

A table of the distribution of \( U \) for \( m \leq n \leq 20 \) is given by Swed and Eisenhart, J.M.S. (1943), pp. 66-87. Wald and Wolfowitz show that when the null hypothesis is true the distribution of \( U \) is asymptotically normal.
Suppose that the frequency functions of \( x, y \) are respectively
\[
\begin{align*}
\tilde{f}(x, \theta, \theta_1, \theta_2, \ldots), \quad f(x, \theta, \theta_1, \theta_2, \ldots),
\end{align*}
\]
where \( \theta, \theta_1, \theta_2, \ldots \) are unknown. We shall write these \( \tilde{f}(x, \theta), f(x, \theta) \).

We wish to use the \( U \) test to test the hypothesis \( \theta = 0 \). Put
\[
\Psi(\theta) = \lim E\left( \frac{U}{\sqrt{n}} \right) = \frac{2a\xi}{\alpha + \xi} \int_{-\infty}^{\infty} \frac{\tilde{f}(x, \theta) f(x, \theta) \, dx}{\alpha \tilde{f}(x, \theta) + \xi f(x, \theta)}
\]
\[
= \frac{2a\xi}{\alpha + \xi} \int_{-\infty}^{\infty} \left\{ \frac{\tilde{f}(x, \theta)}{\alpha \tilde{f}(x, \theta) + \xi f(x, \theta)} \right\} \, dx
\]
\[
\Psi'(\theta) = \frac{2a\xi^2}{\alpha + \xi} \int_{-\infty}^{\infty} \frac{\tilde{f}(x, \theta)^2 f'(x, \theta) \, dx}{\alpha \tilde{f}(x, \theta) + \xi f(x, \theta)}
\]
if differentiation under the integral sign is permissible.

\[
\Psi''(\theta) = \frac{2a\xi^2}{(\alpha + \xi)^3} \int_{-\infty}^{\infty} \tilde{f}'(x, \theta) \, dx = 0
\]
since
\[
\int_{-\infty}^{\infty} \tilde{f}(x, \theta) \, dx = 1
\]
[Note that the primes denote always differentiation with respect to \( \theta \).] Since \( \Psi(\theta) \) has a maximum at \( \theta = 0 \), \( \Psi'(\theta) \) must be 0 if the derivative exists at \( \theta = 0 \).

Put \( m = Na \), \( n = Nb \). We know that when \( \theta = 0 \) the distribution of \( U \) is asymptotically normal when \( N \to \infty \). Presumably this is still true when \( \theta \neq 0 \); but this has not been proved. For the investigation of the test for values of \( \theta \) differing infinitesimally from 0 the following weaker theorem would suffice.

If \( \theta \to 0 \) as \( N \to \infty \) the \( \Theta_N \) distribution of
\[
\frac{U - E_{\Theta_N}(U)}{\sqrt{V_{\Theta_N}(U)}}
\]
→ standard normal and
\[ \frac{V_{\theta N}(U)}{V_0(U)} \rightarrow 1, \]

so that the distribution of
\[ \frac{U - \mathbb{E}_{\theta N}(U)}{\sqrt{V_0(U)}} \]

\( \rightarrow \) standard normal.

We shall assume this although it has not been proved. If we calculate the efficacy of the statistic \( U \), we find that it is 0 because \( \psi'(0) = 0 \). This means simply that if \( \Theta_N = \frac{\theta}{\sqrt{N(a+\theta)}} \), where \( \Theta \) is constant and if
\[ \Pr\{U \leq \lambda_N | \Theta = 0\} = \alpha_N \rightarrow \alpha \]

then
\[ \Pr\{U \leq \lambda_N | \Theta_N\} \rightarrow \alpha, \]

i.e., when \( U \) is small the test is ineffectual against the alternative \( \Theta = \frac{\theta}{\sqrt{N(a+\theta)}} \). This is in contrast to the usual parametric tests, the limiting probability of a significant result from which can be made as near to 1 as we please by making \( \Theta \) sufficiently large.

Thus the \( \chi^2 \) test, which is consistent against any alternative \( f(x) \neq g(x) \), is not so effective as a properly designed parametric test, for the alternative
\[ f(x) \equiv f(x, \Theta, \Theta_1, \Theta_2, \ldots), \quad g(x) \equiv f(x, 0, \Theta_1, \Theta_2, \ldots), \quad \Theta > 0. \]

But it must be noted that we have assumed certain conditions on \( f(x, \Theta) \). These conditions include discontinuities whose positions vary with \( \Theta \), e.g., variable end-point, where \( f(x, \Theta) \) is not zero.

\[ \psi''(\Theta) = \frac{2\alpha \beta^2}{(a+\beta)} \int_{-\infty}^{\infty} \left\{ \frac{\psi(x, \Theta')}{\psi(x, \Theta)} - \frac{2\alpha \beta f(x, \Theta')^2 f'(x, \Theta')}{(\alpha f(x, \Theta) + \beta f(x, \Theta))^3} \right\} \Theta' \]
\[ \Psi'(0) = -\frac{4a^2e^2}{(a+e)^4} \int_{-\infty}^{\infty} \frac{f''(x,0)}{f(x,0)} \, dx = -\frac{4a^2e^2i}{(a+e)^4} \]

since \[ \int_{-\infty}^{\infty} f''(x,0) \, dx = 0. \]

Put \[ \Theta_N = \frac{\Theta}{\{N(a+e)\}^\frac{1}{2}} \]

\[ \Psi(\Theta_N) = 2 \frac{\Theta}{(a+e)^2} - \frac{1}{2} \left( \frac{\Theta^2}{N(a+e)} + O\left(\frac{1}{N}\right) \right) \frac{4a^2e^2i}{(a+e)^4} \]

\[ V_0\left(\frac{\Theta}{N(a+e)^2}\right) \sim \frac{4a^2e^2}{N(a+e)^5} \]

\[ \frac{U - E_{\Theta_N}(U)}{\sqrt{V_0(U)}} \to \frac{
abla_f}{m+n} - \frac{2a \ell}{(a+e)^2} + \frac{a \ell \kappa}{(a+e)^2} = V + \mu k^2 \text{ a.s.} \]

Then \( \Theta = 0 \) the limiting distribution of \( V \) is standard normal and

\[ \lim \mathbb{P}\{ V \leq \lambda \} = \alpha \text{ where } \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\lambda} e^{-\frac{1}{2}x^2} \, dx = \alpha \]

Then \( \Theta = \Theta_N \), the limiting distribution of \( V + \mu k^2 \) is standard normal, therefore

\[ \lim \mathbb{P}\{ V \leq \lambda | \Theta_N \} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\lambda + \mu k^2} e^{-\frac{1}{2}x^2} \, dx \]

which \( \to 1 \) as \( k \to \infty \).
22. Consistency and asymptotic power of Fisher's Test Using Ranks.

We have \( n \) values of a continuous chance variable \( x \) of which \( n_1 \) are positive and \( n_2 \) negative. Let

\[
p_1 = P(x > 0), \quad p_2 = P(x < 0), \quad p = P(Z > Y),
\]

where \( Z \) is a random positive value of \( x \) and \(-Y\) a random negative value.

\[
T = \text{sum of ranks of positive values} - \text{sum of ranks of negative values}
\]

\[
= 2 \left( U + \frac{1}{2} n_1 \left( n_1 + 1 \right) \right) - \frac{1}{2} n(n+1),
\]

where \( U \) is the number of times a negative member of the \( n \) observed values of \( x \) is smaller than a positive member (§ 17).

For fixed \( n_1, n_2 \), \( U = n_1 n_2 p \), therefore

\[
T = 2 \left( n_1 n_2 p + \frac{1}{2} n_1 \left( n_1 + 1 \right) \right) - \frac{1}{2} n(n+1).
\]

\[
E\left( \frac{T}{n^2} \right) \rightarrow P_1 P_2 \left( 2p-1 \right) + p_1 - \frac{1}{2} \quad \text{when} \ n \rightarrow \infty.
\]

It can be shown that \( V \left( \frac{T}{n^2} \right) \rightarrow 0 \) as \( n \rightarrow \infty \). Hence \( \frac{T}{n^2} \rightarrow \) in probability to \( P_1 P_2 \left( 2p-1 \right) + p_1 - \frac{1}{2} \). When the null hypothesis is true, the distribution of \( x \) is symmetrical about 0, and \( P_1 = P_2 = \frac{1}{2} + p \), and so

\[
\frac{T}{n^2} \rightarrow \frac{1}{2} \quad \text{in probability}.
\]

Hence the test is consistent against the alternative

\[
P_1 P_2 \left( 2p-1 \right) + p_1 = \frac{1}{2}
\]
The left hand expression may be written

\[ \sum_{i=1}^{2} p_i p_{ij} + p_{ij}^2 \]

which shows that it is the probability that in a sample of two has a positive value, i.e., the probability that a sample of two consists of either two positive values or one positive value and a smaller negative value.

Suppose now that the distribution of \( x \) is symmetrical about \( \theta \), so that the frequency function of \( x \) is \( f(x-\theta) \) where \( f(x) \) is even in \( x \). Then \( \theta = 0 \)

\[ p_1 = p_2 = \frac{1}{2} = p \]

\[ \frac{d\gamma(\theta)}{d\theta} |_{\theta=0} = \gamma \left\{ \frac{1}{2} \frac{d \phi}{d \theta} + \frac{1}{2} \frac{d \phi_1}{d \theta} \right\} + \mu \left( -\frac{1}{2} \frac{d \phi}{d \theta} + \frac{d \phi_1}{d \theta} \right) |_{\theta=0} \]

\[ p_1 = \int_{-\infty}^{\infty} f(x-\theta) dx = \int_{-\infty}^{\infty} f(x) dx \]

\[ \frac{dp_1}{d\theta} = \theta f(-\theta), \quad \frac{d\phi_1}{d\theta} |_{\theta=0} = \phi'(0) \]

\[ p_1 p_2 = \frac{1}{4} \left\{ 1 - (p_1 + p_2)^2 \right\} \]

\[ \frac{d(f_1 f_2)}{d\theta} = -\frac{1}{2} (p_1 - p_2) \frac{\partial (p_1 - p_2)}{\partial \theta} \]

\[ = 0 \text{ at } \theta = 0 \text{ since then } p_1 + p_2 = \frac{1}{2} \]

\[ p = \int_{-\infty}^{\infty} f(x-\theta) \int_{-\infty}^{\infty} f(y-\theta) \frac{dy}{p_1} \frac{dy}{p_2} \]

\[ = \frac{1}{p_1 p_2} \int_{-\infty}^{\infty} f(x-\theta) \int_{\theta}^{x+\theta} f(z) dz dx \quad (y-\theta = -z) \]

\[ = \frac{1}{p_1 p_2} \int_{-\infty}^{\infty} f(x) \int_{\theta}^{x+\theta} f(z) dz dx \]

\[ \frac{d\phi}{d\theta} |_{\theta=0} = \frac{1}{p_1 p_2} \left\{ \phi(-\theta) \int_{0}^{\theta} f(z) dz + \int_{c}^{\phi(\theta)} 2 \phi'(\theta) f(x+2\theta) - \phi(\theta) f(x) \right\} |_{\theta=0} \]
\[ = 4 \left\{ \int_0^\infty f(x)^2 \, dx - \frac{1}{2} f(0) \right\} \]

Therefore

\[ \frac{\partial E(Q)}{\partial \Theta} \bigg|_{\Theta=0} = n^2 \left\{ \int_0^\infty f(x)^2 \, dx \right\} + n \left\{ 2 f(0) - 2 \int_0^\infty f(x) \, dx \right\} \]

\[ V(Q) \big|_{\Theta=0} = \frac{1}{6} n(n+1)(2n+1) \]

Hence efficacy \( \sim 12n \left\{ \int_0^\infty f(x)^2 \, dx \right\} \)

For the t test, \( t = \frac{\bar{x} \sqrt{n}}{\sqrt{\frac{S}{n-1}}} \), and efficacy \( \sim \frac{n}{\sigma^2} \).

Thus asymptotic relative efficiency of Fisher's test by ranks compared to the t test for testing \( \Theta = 0 \) is

\[ 12 \sigma^2 \left\{ \int_0^\infty f(x)^2 \, dx \right\} \]

For a normal distribution

\[ V \left\{ \int_0^\infty f(x)^2 \, dx \right\} = \frac{1}{4 \pi} \]

and so in this case the asymptotic relative efficiency is \( \frac{12}{4 \pi} = \frac{3}{\pi} \).

Analysis of Variance in Randomized Blocks.

The principles of this test may be briefly summarized as follows. Several batches, each consisting of \( n \) individuals, are taken, and the individuals of a batch subjected to \( n \) different treatments, the allocation of the treatments to the individuals of a batch being determined by chance. Each individual is then measured, and we wish to determine whether the differences in treatment have produced any real differences in the character measured. The batches might, for example, be the blocks in an agricultural experiment, and the individuals the plots into which each block is subdivided, and the quantity measured might be the
"yield" from each plot. For convenience we shall consider this case, which is in no way special. Suppose that there are m blocks.

Let \( x_{ijk} \) be the yield from the \( i^{th} \) treatment in the \( k^{th} \) plot in the \( j^{th} \) block to which it was assigned by the randomization process. We assume that

\[
x_{ijk} = T_i + B_j + y_{jk}
\]

where \( T_i \) denotes the "effect" of treatment \( i \), \( B_j \) the "effect" of block \( j \), and the third term \( y_{jk} \) arises from variability among plots in a block, errors in measurement, and other accidents affecting particular plots. The null hypothesis, which we wish to test, states that

\[
T_1 = T_2 = \ldots = T_m
\]

Put \( x_{i.} = n^{-1} \sum_j x_{ij} \), \( x_{.j} = m^{-1} \sum_i x_{ij} \), \( x_{ij} = \text{yield of } i^{th} \text{ treatment in } j^{th} \text{ block} \).

The analysis of variance is

\[
S = S_T + S_B + S_E
\]

where \( S = \sum (x_{ij} - x)^2 \), is the total squarariance,

\( S_B = n \sum (x_{ij} - x)^2 \), the squarariance due to blocks,

and is independent of \( T_1, T_2, \ldots \)

\( S_T = m \sum (x_{i.} - x)^2 \), the squarariance due to treatments,

and is independent of \( B_1, B_2, \ldots \)

\( S_E = \sum \sum (x_{ij} - x_{.j} - x_{i.} + x)^2 \), the residual squarariance,

and is independent of both \( B_1, B_2, \ldots \) and \( T_1, T_2, \ldots \)

Differences in the values of the \( T \) tend to increase the value of \( S_T \) without affecting the value of \( S_E \).

It can be shown that if the \( y_{jk} \) are independent chance variables, each with mean 0 and S.D. \( \sigma \),
\( T(T) = \frac{m}{n} \sum (T_i - \bar{T})^2 + (n-1) \sigma^2 \)

where \( \bar{T} = \frac{n}{n} \sum T_i \),

\( T(\mathbf{S}) = (m-1)(n-1) \sigma^2 \)

\[ E\left( \frac{s_T}{m-1} \right) = \sigma^2 + \frac{\eta \sum (T_i - T)^2}{m-1} \]

\[ E\left( \frac{s_S}{(m-1)(n-1)} \right) = \sigma^2 \]

Therefore to test the null hypothesis we consider

\[ F = \frac{s_T/(n-1)}{s_S/(m-1)(n-1)} \]

large values of which are significant. If, further, the \( y \) are normal variables, then when the null hypothesis is true, \( s_T/\sigma^2 \) and \( s_S/\sigma^2 \) are independent chance variables distributed like \( \chi^2 \) with \( n-1 \) and \( (m-1)(n-1) \) degrees of freedom respectively.

From this the exact distribution of \( F \) under the null hypothesis can be obtained.

We can, however, make an exact test of the null hypothesis without any assumption of normality or even independence of the \( y \). If the null hypothesis is true, the observed value of \( F \) is the result of the chance allocation of the different treatments to the different plots in the blocks. This allocation may be done in \( (m!)^{m-1} \)

ways, but these occur in sets of \( n! \) which differ only by interchange of treatment names. Hence there are in general \( (m!)^{m-1} \)

values of \( F \), some of which may happen to coincide with one another. As the allocation of treatments to plots is determined by chance, all such allocations are equally probable, and so, therefore, are
the corresponding values of $F$. We can obtain a test by comparing the observed value of $F$ with the other possible values.

Gatenby & Yates (J. Agric. Sci. 1933 Vol. 23 pp 6-16), by a sampling process applied to data from a uniformity trial consisting of 8 blocks each of 5 plots, showed that in that case there was good agreement between the randomization distribution of $F$ and the sampling distribution of $F$ on the assumption of normality.

See also Welch (Biometrika 1937, 30 pp 21-52) and Pitman (Biometrika 1938, 30 pp 322-335).

We obtain a considerable simplification if, instead of $F$, we use

$$W = \frac{S_T}{S_T + S_E} = \frac{S_T}{S} = \frac{F}{F + m - 1}$$

for now the denominator is constant for permutations within each block. $W$ is a monotonic increasing function of $F$ and therefore if we take large values of $W$ as significant the test based on $W$ will be equivalent to the test based on $F$.

Note that in the usual test assuming normality, when the null hypothesis is true $W$ has a $t\left(\frac{1}{2}(n-1), \frac{1}{2}(m-1)(n-1)\right)$ distribution and that $\frac{1}{2}(S_T + S_E)/\sigma^2$ is a $\Gamma\left(\frac{m(n-1)}{2}\right)$ variable.

Hence

$$\frac{1}{2}(S_T + S_E)/\sigma^2 \rightarrow 1 \quad \text{in probability as } m \rightarrow \infty.$$ 

Therefore the sampling distribution of

$$m(m-1)W = \frac{S_T/\sigma^2}{(S_T + S_E)/m(m-1)\sigma^2}$$

tends to the $\chi^2$ distribution with $n-1$ degrees of freedom.
Let \( c_{jk} \) = yield in plot \( k \) of block \( j \). Let

\[
\begin{align*}
n^{-1} \sum_k c_{jk} &= a_j, \\
n^{-1} \sum_k (c_{jk} - a_j)^2 &= b_j
\end{align*}
\]

Now

\[ 0 \leq a_j \leq 1 \]

and it can be shown that for the randomization distribution

\[
\begin{align*}
\mu(\cdot) &= \frac{1}{m}, \\
\nu(\cdot) &= \frac{1}{m} \sum_{\alpha=1}^m \frac{\prod_{j=1}^{n-1} \frac{\alpha_j}{\alpha_j}}{\alpha_j^{n-1}}
\end{align*}
\]

If the \( b_j \) are all equal \( V(\cdot) \) becomes

\[
\frac{2(n-1)}{m^2(n-1)}
\]

and it will be asymptotically equal to this when \( n \to \infty \) if \( 0 < b' \leq b_j \leq c' \) for all \( j \). The \( \mathcal{B} \left( \frac{1}{2}, \frac{1}{2} \right) \) \( \mathcal{B} \left( \frac{1}{2}, \frac{1}{2} \right) \) distribution has the same range and the same mean value as \( V \). Its variance is

\[
\frac{2(n-1)}{m^2(mn-m+2)}
\]

which is approximately that of \( V \) when \( n \) is large and the \( b_j \) have all about the same value. Hence we may expect this distribution of \( V \) to be approximately a \( \mathcal{B} \left( \frac{1}{2}, \frac{1}{2} \right) \) distribution.

We can get a more definite result by following Yeld & Wolfovitz (Annals of Math. Stat. 1944 p. 365), and showing that under certain weak conditions on the \( b_j \) the limiting distribution of \( \frac{1}{m}(n-1) \) \( \chi^2 \) distribution with \( n-1 \) degrees of freedom. \( x_{ij} \), the yield of treatment \( i \) in block \( j \) takes the values

\[ a_{ij}, a_{ij}, \ldots, a_{ij} \]
each with probability 1/n. Hence
\[
\tau(x_{ij}) = n^{-1} \sum_{k} c_{ik} a_k = a_j
\]
\[
\nu(x_{ij}) = n^{-1} \sum (a_{ik} - a_j)^2 = b_j
\]
\[
\gamma(x_{ij}; x_{ij}) = \frac{\left( \sum x_{ij} \right)^2 - \sum x_{ij}^2}{m(m-1)}
\]
\[
= \frac{m^2 a_j^2 - m \left( b_j + a_j^2 \right)}{m(m-1)} = \frac{-b_j}{m-1} + a_j
\]
\[
\text{cov}(x_{ij}, x_{ij}) = \frac{-b_j}{m-1}
\]
\[
\omega = \frac{\sum x_{ij}^2}{m} = \frac{m \sum (x_{ij} - \bar{x})^2}{\sum x_{ij}^2} - \frac{\sum (x_{ij} - \bar{x})^2}{\sum (x_{ij} - x_{ij})^2}
\]
Let
\[
x_{ij}^* = \sum_{\nu} \lambda_{\nu} x_{\nu j} \quad (i = 1, \ldots, n)
\]
where \(\lambda_{\nu}\) is an orthogonal matrix with \(\lambda_m = \lambda_m = \cdots = \lambda_m = n^{-\frac{1}{2}}\), so that
\[
x_{ij}^* = \frac{\sum x_{ij}}{\sqrt{m}} = \sqrt{m} \alpha_j .
\]
Then
\[
\gamma(x_{ij}^*) = \left( \sum \lambda_{\nu} \right) \alpha_j = 0 \quad \text{if } i = 1, 2, \ldots, n-1
\]
\[
\nu(x_{ij}^*) = \sum (x_{ij}^*)^2
\]
\[
= \left( \sum \lambda_{\nu}^2 \right) E(x_{\nu j}^2) + 2 \sum (\lambda_{\nu_1} \lambda_{\nu_2} E(x_{\nu_1} x_{\nu_2}))
\]
\[
= b_j + (0-1) \left( -\frac{b_j}{m-1} \right) = \frac{b_j}{m-1} .
\]
\[
6 \text{cov}(x_{ij}^*, x_{ij}^*) = E \left\{ \left( \sum \lambda_{\nu} x_{\nu j} \right) \left( \sum \lambda_{\nu} x_{\nu j} \right) \right\}
\]
\[
= \left( \sum \lambda_{\nu} \right) \left( \sum \lambda_{\nu} \right) - \sum \lambda_{\nu} \lambda_{\nu} \sum \lambda_{\nu} \sum \lambda_{\nu} \right) E(x_{\nu j} x_{\nu j}) + \left( \sum \lambda_{\nu} \lambda_{\nu} \right) E(x_{\nu j}^2) = 0
\]
Put
\[
x_{ij}^* = m^{-1} \sum x_{ij}^* , \quad i = 1, \ldots, n .
\]
The relation of the \( x_i^* \) to the \( x_i \) is the same as the relation of the \( x_i^* \) to \( \lambda_i \), and \( x_i^* = x \sqrt{n} \).

\[
\nu(x_i^*) = \frac{\sum b_j}{m} \cdot \frac{M}{m-1} = \frac{b}{m-1},
\]

where \( b = m \sum b_j \).

\[
\omega = \frac{m \sum (x_i^* - x)^2}{\sum \sum (a_j \lambda - a_i \lambda)^2} = \frac{m \sum x_i^2 - m x^2}{\sum m b_j} = \frac{\sum x_i^2}{m \sum b_j}.
\]

\[
m(n-1)\nu = \frac{\sum x_i^2}{m-1}.
\]

Put \( X = (x_{i_1}^*, x_{i_2}^*, \ldots, x_{i_m}^*) \) \( r = 1, \ldots, m \)

\[
X = \left( \frac{x_{i_1}^* + x_{i_2}^* + \ldots + x_{i_m}^*}{\sqrt{m \sum b_j}} \right) = \frac{x_1 + x_2 + \ldots + x_m}{m \sqrt{\sum b_j}}.
\]

Now the \( X_i \) are independent chance variables. Hence if the \( a_j \) satisfy certain conditions, the distribution of \( X \) will tend to the distribution of \( n-1 \) independent standard normal variables when \( m \to \infty \). Hence the distribution of

\[
m(n-1)\nu = \frac{x_1^2 + \ldots + x_m^2}{m \sum b_j}.
\]

will tend to a \( \chi^2 \) distribution with \( n-1 \) degrees of freedom.

Simple sufficient conditions are

\[0 < b_i \leq c \] for all \( j \), \( b \) and \( c \) being fixed.

Suppose that we are given n pairs of observations from a bivariate distribution,

\((x_1, y_1), \ldots, (x_n, y_n)\),

and that we wish to test the hypothesis \(H_0\) that \(X\) and \(Y\) are independent chance variables.

Consider the \(n!\) possible associations or pairings:

\((x_1, y_{q_1}), \ldots, (x_n, y_{q_n})\),

where \(q_1, \ldots, q_n\) is a permutation of the integers \(1, \ldots, n\). Under \(H_0\) all such associations are equally probable. For any such association we may calculate a correlation coefficient \(r\) defined by

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

where \(\bar{x}\) is the mean of the \(x\) values and \(\bar{y}\) the mean of the \(y\) values. We may then define a critical region of size \(N/n! = \alpha\) to test \(H_0\) by declaring the \(\alpha\) largest values of \(|r|\) to be critical, or significant. Since for all associations the denominator of the expression for \(r\) is the same, the significant associations are those which give the largest values of

\[
|\sum xy - nx\bar{y}|
\]

Example. The following pairs of values of the variables \(X\) and \(Y\) are observed:

\[
\begin{array}{ccccccc}
X & 1.1 & 1.2 & 1.3 & 1.5 & 1.9 \\
Y & 1.7 & 1.6 & 1.9 & 1.3 & 1.0
\end{array}
\]
Is there any evidence of dependence? Since \( r \) is independent of scales and origins, we may take the origins at the lowest observed values of \( \chi \) and \( \gamma \), and then drop the decimal points. We have

\[
\chi = 0, 1, 2, 4, 8 \quad \bar{\chi} = 3
\]

\[
\gamma = 7, 6, 9, 3, 0 \quad \bar{\gamma} = 5, \quad n\bar{\gamma} = 75.
\]

The pairings which give the largest values of \( |\xi_{xy} - 75| \) are:

\[
\begin{array}{cccc}
\chi & 0 & 1 & 2 & 4 & 3 \\
\gamma & 9 & 7 & 6 & 3 & 0 \\
& 9 & 6 & 7 & 3 & 0 \\
& 7 & 9 & 6 & 3 & 0 \\
& 6 & 9 & 7 & 3 & 0 \\
& 0 & 3 & 6 & 7 & 9 \\
& 7 & 6 & 9 & 3 & 0 \\
& 9 & 7 & 3 & 6 & 0 \\
& 6 & 7 & 9 & 3 & 0 \\
& 0 & 3 & 7 & 6 & 9 \\
\end{array}
\]

\[
|\xi_{xy} - 75| = 31, 32, 33, 35, 115, 36, 37, 37, 113.
\]

If we take \( \hat{p} = 0.05, \quad M = 5\sqrt{20} = 5 \). The pairing determined by the observations gives the sixth largest value of \( |r| \), and is therefore significant, and we conclude that \( \chi \) and \( \gamma \) are not independent.

For the test distribution of \( r \)

\[
F(r) = 0, \quad 3(r^2) = \frac{1}{n-1}.
\]

From this it is easy to show that the test is consistent if \( \chi \) and \( \gamma \) have finite variances and a non-zero correlation coefficient.

For large \( n \) we can use an approximate distribution. Let
\( \mu_x \) be the \( i^{th} \) moment about the mean of the \( x \) values
\( \mu'_x \) be the \( i^{th} \) moment about the mean of the \( y \) values

Then if \( \frac{1}{n} \left( \frac{\mu_x}{\mu_x^2} \frac{\mu'_x}{\mu'_x} \right) \) and \( \frac{1}{m} \left( \frac{\mu_y - 3\mu_y^2}{\mu_y^2} \right) \left( \frac{\mu'_y - 3\mu'_y^2}{\mu'_y^2} \right) \) are small, a good approximation to the distribution of \( r \) is the continuous distribution with frequency function

\[
\frac{1}{B \left( \frac{1}{2}, \frac{n-2}{2} \right)} (1 - x^2)^{\frac{n-2}{2}}
\text{ for } -1 \leq x \leq 1
\]

If this were the exact distribution of \( r \), \( r^2 \) would have frequency function

\[
\frac{1}{B \left( \frac{1}{2}, \frac{m-1}{2} \right)} x^{-\frac{1}{2}} (1-x)^{\frac{m-2}{2}}
\text{ for } 0 \leq x \leq 1
\]

and \( \frac{1}{2}(r+1) \) would have frequency function

\[
\frac{1}{B \left( \frac{1}{2}, \frac{m-1}{2} \right)} x^{\frac{1}{2}} (1-x)^{\frac{m-2}{2}}
\text{ for } 0 \leq x \leq 1
\]

The use of any one of these is exactly equivalent to the ordinary test for significance of a correlation coefficient based on the assumption of normality.

I think it can be shown that if \( \chi \) and \( \gamma \) have finite variances, the test distribution of \( \sqrt{n-1} \, r \) tends in probability to standard normal when \( n \to \infty \).