Spring 1971

The Importance Of Interpretation In Research Statistics: An Overview

Bill Cook
Carroll College

Follow this and additional works at: https://scholars.carroll.edu/mathengcompsci_theses

Part of the Applied Statistics Commons

Recommended Citation
https://scholars.carroll.edu/mathengcompsci_theses/118

This Thesis is brought to you for free and open access by the Mathematics, Engineering and Computer Science at Carroll Scholars. It has been accepted for inclusion in Mathematics, Engineering and Computer Science Undergraduate Theses by an authorized administrator of Carroll Scholars. For more information, please contact tkratz@carroll.edu.
THE IMPORTANCE OF INTERPRETATION IN RESEARCH STATISTICS:

AN OVERVIEW

by Bill Cook

Carroll College
Helena, Montana
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRODUCTION</td>
<td>2</td>
</tr>
<tr>
<td>RULES OF NOTATION</td>
<td>4</td>
</tr>
<tr>
<td>CHAPTER I. COLLECTION OF DATA: FIELD SURVEYS AND EXPERIMENTAL PROCEDURES</td>
<td>5</td>
</tr>
<tr>
<td>CHAPTER II. GROUPING AND COMPILING DATA</td>
<td>9</td>
</tr>
<tr>
<td>CHAPTER III. DESCRIBING THE FREQUENCY DISTRIBUTION</td>
<td>12</td>
</tr>
<tr>
<td>PART 1: MEASURES OF CENTRAL VALUE</td>
<td>12</td>
</tr>
<tr>
<td>PART 2: MEASURES OF DISPERSION (VARIATION)</td>
<td>14</td>
</tr>
<tr>
<td>PART 3: MEASURES OF SKEWNESS</td>
<td>16</td>
</tr>
<tr>
<td>PART 4: MEASURES OF KURTOSIS</td>
<td>17</td>
</tr>
<tr>
<td>CHAPTER IV. STATISTICAL INFERENCE</td>
<td>19</td>
</tr>
<tr>
<td>CHAPTER V. THE LOGIC OF DISTRIBUTION-FREE AND NON-PARAMETRIC METHODS</td>
<td>28</td>
</tr>
<tr>
<td>CONCLUSION</td>
<td>32</td>
</tr>
<tr>
<td>BIBLIOGRAPHY</td>
<td>33</td>
</tr>
</tbody>
</table>
INTRODUCTION

This paper was originally conceived of as a real research project, as a way of putting theoretical statistical knowledge to work in the area of real life. The author has access to a suitable population for study (institutionalized retardsates), but it was soon discovered that the problems incurred in gathering the needed information were too great to be pursued on an undergraduate level. Medical and social records were available, but they were filled with ambiguities or generally lacking in needed information. For example, the socio-economic factors involved in retardation were of prime consideration, but the needed information could probably have been acquired only through interviews with parents and they may well have been hesitant to divulge such information. Such problems would not be nearly as considerable if Medical records contained complete, standardized information. At the present time, they do not.

The intent to deal with the real world will be pursued, then, by looking beyond the statistics themselves to an analysis of what can be said after the data is compiled. A discussion of the tools of statistics is included as a way of presenting the meaning, interpretation and limitation of each tool. The framework will follow a sort of handbook fashion.

The author assumes the readers will have taken an elementary statistics course and the intent is to expand this basic knowledge into an awareness of how it is to be interpreted. The importance of such an awareness is twofold; it provides guidelines for conducting the research
and it also determines what can be said once the research itself is completed.

The presentation will move in the same sequence as an actual research project would proceed:

1) Method of Collecting Data (Chapter 1)
2) Grouping and Compiling Data (Chapter 2)
3) Describing the Distribution Mathematically (Chapter 3)
4) Statistical Inference: Interpretation of Results (Chapter 4)
5) Non-Parametric Methods (Chapter 5)

Whenever possible, real examples are included in order to illustrate a given subject. Explanations are short and to the point so as to present an overview rather than a detailed analysis. Since rigorous computation formulas are readily available, the intent here is to explore the logic of the methods to be used.
NOTATION RULES

Because of the confusion which can arise due to notation, the following symbols have been used consistently throughout the text:

\(\Xi\) -- this is a variable representing all of the various scores, persons or observations in our data; \(i\) varying from 1 to \(N\). Thus \(\Xi\) represents \((\Xi_1, \Xi_2, \ldots, \Xi_N)\).

\(N\) -- The total of scores, persons, etc. in a given sample. Other uses of \(N\) will be clear in their context.

\(\Sigma\) -- The summation sign is always understood to represent \(\sum_{i=1}^{N}\) the sum of all the scores.

\(l\) -- \(l\) is used for the length of an interval. \(i\) is also used, but is confusing because the subscript \(i\) is also used for other purposes.

\(u\) -- \(\mu\), the population mean.

\(\bar{X}\) -- mean of a given sample.

\(S\) -- standard deviation of a sample distribution.

\(\sigma\) -- standard deviation of the population.

\(P\) -- Theoretical probability of a given population.

\(p\) -- Observed proportion in a sample from the population.

All other symbols are defined clearly in their context. The above are given because of the ambiguities which can surround them.

The sources used are given in parentheses (Author and page number) where applicable. This is standard procedure in all research journals and is employed because of its practical convenience to the reader.
CHAPTER 1

Collection of Data: Field Surveys
And Experimental Procedures

Many errors in field studies or experimental testing of hypotheses are due directly to sampling techniques which do not yield an unbiased sample. Obviously no amount of careful interpretation will make up for an inappropriate sample. First, we will consider three sampling methods for field surveys, and then we will note models for experimental purposes.

Types Of Field Surveys

1) Random Sampling — There are two conditions for random sampling:
   a) Each person in the defined population will have an equal chance of being included.
   b) Drawings must be independent. (McNemar, p. 55).

The first condition is not easily met in practice; however the general aim is to obtain a sample which will be representative of the population from which it is drawn. For example, a systematic sampling of the nth file in medical records does not exactly satisfy the conditions of random sampling, but it will assure a random sample unless the files have been prearranged in some particular fashion. Note that the above concerns are meant for a catalogued population only (usually files).

The use of random sampling in an uncatalogued population is replete with problems. Actually, increasing the sample size is the only way to reduce chance errors when the random method is used. Sometimes samples
are believed to be random, but they are not. For example, certain questionnaires will be returned only by those interested, usually providing for a biased sample.

2) **Stratified Sampling** -- This method involves grouping the population to be studied into strata which are really different with regard to the attribute being studied. Then cases within the strata are drawn randomly and the number within each stratum is proportional to its representation in the entire population. The precision in such a sampling method is greater than in mere random sampling alone.

3) **Area Sampling** -- If extensive facilities are available, it seems that area or "pin-point" is the best method devised for drawing samples in survey studies. This method is often used in governmental surveys. For example, the regions could be divided into Northeast, North Central, South and West. Distinct boundaries are then set up for even further divisions, such as blocks or streams or roads. Every nth household in a given area is then interviewed. This is actually a special case of stratified sampling, where geography is the criterion for division.

Another special kind of stratified sampling is called quota sampling. Here the interviewer is given a strata to cover, but it is up to him to make the selection. The only criterion is that he have so many professionals, so many housewives, etc. Such samples may be quite accurate: the biggest problem is that there is no way to calculate just how great the error is. Errors can be calculated only on the assumptions of random sampling and there is no alternative model for non-random sampling.

Now our attention shifts to experimental research. Suppose that the
effect of a particular product, atmosphere, etc. is to be tested. This involves the use of control groups. We will now attempt to present some appropriate experimental procedures.

**Experimental Procedures**

For experimental purposes, the formation of groups can be accomplished in five ways:

1) Random sampling -- random assigning of individuals to the groups.
2) Pairing -- making comparable the groups or certain variables which might affect the outcome of the experiment.
3) Use of siblings or littermates.
4) Matched distribution -- instead of pairing persons, distributions are matched for the given variables.
5) Usage of the same person under all of the experimental conditions.

The question of which method is preferable is an open one. Greater precision is allowed in methods 2, 3, or 5. However, the advantages of pairing depends completely on how highly the variables so controlled are correlated with the dependent variable. If there is a low correlation, there is little gain in using the pairing method.

Method four has a particular advantage if the cost for each case is higher in the experimental group than in the control group, because unlike ordinary pairing procedures, there is no need to have an equal number of cases in the groups. (The standard error of difference can be obtained without this restriction.)

The use of paired individuals for experimental and control groups has long been advocated. However, as McNemar points out, "The gain in error reduction may not be appreciable." (McNemar, p. 365). Often times,
this is because the advocates of pairing feared randomization as a way of setting up groups. However, even pairing must use the principles of randomization. "Random differences between groups never have more than a random effect on the results; the error formulas always include all random variables." (McNemar, p. 365). Thus often the error reduction involved in the pairing procedures may not be worth the needed effort.

Finally, no matter what method is used for establishing the experimental groups, it must be remembered that generalizations can be made only if the defined sample is truly representative of the generality of mankind being studied. For example, to generalize from a group of college sophomores to mankind in general may well involve a leap that is simply not permissible. Thus the original sampling techniques used may well be more important than the correlation between experimental and control groups.
CHAPTER 2

Grouping and Compiling Data

Once the data is collected, it must be organized and put into meaningful form. Ungrouped data can be placed in an array and ordered from the lowest to the highest. A more common practice involves grouping. With ungrouped data there is obviously greater precision, but with large numbers of observations and a wide range of values, a frequency distribution is necessary.

A frequency distribution involves nothing more than establishing arbitrary intervals (usually equal, but this is not theoretically necessary) and listing the frequency or number of observations in each interval. The customary number of intervals is between ten and fifteen, and a multiple of three, five or ten is convenient for the size of each class interval.

The following is an example of frequency distribution:

Socio-economic Score of 909 Southern Farm Failures (Zelditch, p. 21).

<table>
<thead>
<tr>
<th>Interval</th>
<th>Frequency</th>
<th>Midpoint</th>
</tr>
</thead>
<tbody>
<tr>
<td>35-39</td>
<td>3</td>
<td>37</td>
</tr>
<tr>
<td>40-44</td>
<td>33</td>
<td>42</td>
</tr>
<tr>
<td>45-49</td>
<td>191</td>
<td>47</td>
</tr>
<tr>
<td>50-54</td>
<td>229</td>
<td>52</td>
</tr>
<tr>
<td>55-59</td>
<td>169</td>
<td>57</td>
</tr>
<tr>
<td>60-64</td>
<td>90</td>
<td>62</td>
</tr>
<tr>
<td>65-69</td>
<td>56</td>
<td>67</td>
</tr>
<tr>
<td>70-74</td>
<td>54</td>
<td>72</td>
</tr>
<tr>
<td>75-79</td>
<td>52</td>
<td>77</td>
</tr>
<tr>
<td>Interval</td>
<td>Frequency</td>
<td>Midpoint</td>
</tr>
<tr>
<td>----------</td>
<td>-----------</td>
<td>----------</td>
</tr>
<tr>
<td>80-84</td>
<td>26</td>
<td>82</td>
</tr>
<tr>
<td>85-89</td>
<td>5</td>
<td>87</td>
</tr>
<tr>
<td>90-94</td>
<td>1</td>
<td>92</td>
</tr>
</tbody>
</table>

A rough description of the data is often helpful before a precise, mathematical one. For example, consider the following histogram which is a rough graph of the grouped data. (Zelditch, p. 22).

![Figure 1.](image)

Class Intervals are marked off on baseline.

A second common visual representation is the frequency polygon.

(Zelditch, p. 22):
Through grouping and graphing procedures, a beginning understanding of the data is possible. A mathematical description is also now possible.
CHAPTER 3

Describing The Frequency Distribution

In this chapter, four methods of describing the frequency distribution are noted and the importance and limitations of each are considered.

Part 1: Measures of Central Value

Actually, central value is only another way of saying average, and "an average is a measure designed to reduce a set of measurement data to a single representative number." (Zelditch, p. 41). An analysis of three types of averages is now provided.

The Mode

The value of \( Xi \) which has the greatest frequency is defined as the mode. For grouped data, the mode is taken as the midpoint of the interval which has the greatest frequency. The limitations of the mode are several. For instance, the mode depends too heavily on the size of the interval. Often a distribution may appear to be bimodal, but this may be due to the size of the interval chosen and therefore accidental and not real. The mode is also subject to greater sampling fluctuations than the mean or the median. (McNemar, p. 14).

The Median

The median is defined in two ways:

1) "If the individual scores are arranged in order, the median is the value of the midmost individual if \( N \) is odd, or lies midway between the
two middle individuals if \( N \) is even.

2) When a distribution has been made, the median is defined as the point on the scale such that the frequency above or below the point is 50% of the total frequency." (McNemar, p. 15).

For grouped data, the median can be found by means of a cumulative frequency column and either interpolation or a formula method. (See Spiegel, p. 57). The chief merits of the median are its ease of computation, its independence of extremes, and the fact that it is not affected by the size of the extremes.

**The Mean**

The mean is defined simply as the sum of the scores divided by their number.

\[
\bar{X} = \frac{\sum X}{N}
\]

For grouped data, the procedure is simplified by means of coded midpoints and a guessed mean as outlined in Zelditch, p. 43. Here:

\[
\bar{X} = \frac{1}{N} \sum Fdil
\]

Where \( X_0 = \) guessed mean, \( F = \) frequency, \( d = \) coded midpoint and \( l = \) size of interval.

The application is of considerable importance. The mean of a sample is the expected value of a member of the sample. Of all statistics estimating the population mean, the sample mean provides the best or most effective estimate. Thus, generally, for \( N = 30 \), since the sampling distribution is approximately normal, the Population mean, \( \mu = \bar{X} \pm 1.96S_x \) with 95% confidence level is computed. Other descriptions of the normal curve are
also applicable. This is a very important statement, and will be further clarified in our discussion of the Central Limit Theorem in Chapter four.

The value of the mean depends on the nature of the data. If the data is nearly symmetrical, then the mean and median will be nearly equal and the mean is to be used because it is usually a more stable measure (the means of two samples will show closer agreement than the medians) and it can be handled arithmetically and algebraically (for two samples, the mean of the combined sample = \( \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2}{N_1 + N_2} \)).

If, on the other hand, there is an extreme measure in one direction (skewness), then the mean is inflated and does not provide as typical a representation of the average as does the median.

Part 2 Measure of Dispersion (Variation)

Measures of variation may be described as the extent of scatter (or cluster) about the central value. Several measures are possible and they differ markedly in their interpretation and usefulness.

Range

The range is the difference between the highest and lowest scores. Since it depends only on two points of data which could be greatly influenced by chance it does not say much about the data as a whole, but does give an idea of the extremes.

Quartile Deviation

The quartile deviation, \( Q = \frac{Q_3 - Q_1}{2} \), where \( Q_3 \) is the point above which \( \frac{3}{4} \) of the data are found and \( Q_1 \) is the point below which \( \frac{1}{4} \) of the
data are found. Q₁ and Q₃ are computed exactly as the median.

The quartile deviation, or the semi-interquartile range, is helpful only in symmetrical distributions, for it assumes that 50% of the data will fall in this range. For distributions which lack symmetry, it is better to give Q₁ and Q₃ as values along with the median; this will enable one to picture whether or not the cluster above the median is different from that below.

**Average Deviation**

The average deviation or mean deviation is the average of the deviations of the scores from the mean. The absolute value of the deviations is used. Thus, \( MD = \frac{\sum X_i - \bar{X}}{N} \) The average deviation is currently little used, but is helpful in understanding earlier psychological literature.

**The Variance and Standard Deviation**

The variance is the square of the standard deviation. The variance of the sample distribution is defined as \( S^2 = \frac{\sum (X_i - \bar{X})^2}{N} \). A method for computing the standard deviation from grouped data is similar to that for the mean and is outlined in Spiegel, p. 78, or Zelditch, p. 72. Because of the error introduced due to grouping, the standard deviation thus obtained should be corrected by means of Sheppard's correction: \( \frac{1}{12} \).

The correction is small if twelve or fifteen intervals are used, but can be appreciable for ten or less. The grouping error is due to the fact that all data within an interval is given the same value, i.e., the midpoint of the interval.

In the interpretation of the standard deviation, it should be realized that it can be of little help if the distribution is not symmetrical. Other-
wise it is the best measure of dispersion. Besides being stable and able to be handled algebraically, it is used in advanced work with determination of sampling errors and analysis of variance.

Part 3: Measures of Skewness

If a distribution is not of the symmetrical, bell-shaped type, it is not sufficient just to give the mean and the standard deviation. Skewness is the measure of the lack of symmetry. In computation of the degree of skewness, we first need to define the first four moments about the mean.

\[
\begin{align*}
U_1 &= \frac{\sum(X_i - \bar{X})}{N} = 0 \\
U_2 &= \frac{\sum(X_i - \bar{X})^2}{N} = s^2 \\
U_3 &= \frac{\sum(X_i - \bar{X})^3}{N} \\
U_4 &= \frac{\sum(X_i - \bar{X})^4}{N}
\end{align*}
\]

This allows us to compute \( g \), a measure of skewness defined in terms of moments:

\[
g_1 = \frac{U_3}{U_2 N U_2^{1/2}}
\]

The interpretation is as follows: for symmetrical distributions, \( g_1 \) will be 0. Deviations from zero provide a measure of skewness, it being positive when \( g_1 \) is positive, negative when \( g_1 \) is negative. In visual terms, if the longer tail is to the right, the curve is positively skewed. If the longer tail is to the left, it is negatively skewed. This is determined by drawing a line from the maximum and comparing areas. Figure 3 provides a visual interpretation of \( g_1 \). (McNemar, p. 30).
The importance of the degree of skewness should be obvious due to the limitations of other measures when symmetry is lacking.

**Part 4: Measure of Kurtosis**

Kurtosis is defined as the degree of "peakedness" of a curve. The degree of kurtosis can be described as follows:

\[ g_2 = \frac{U_4}{U_3} - 3 \]

where \( U_3 \) and \( U_4 \) are the third and fourth moments about the mean respectively. In the interpretation, we note that when \( g_2 \) is less than zero, the distribution tends to be flat-topped (platykurtic), and when \( g_2 \) is greater than zero, it is steep or peaked (leptokurtic).

It is not usually necessary to compute \( g_1 \) or \( g_2 \) unless the curve appears to be skewed or peaked or flat. It is seldom advisable to compute these measures when \( N \) is less than 100, for the graph would not be meaningful due to the sparseness of the data.
It should be seen that the occurrence of a skewed distribution is closely related to the measuring units used, and since often equal scale units are not used, it can only be said that in terms of the units a distribution has a particular shape. However, knowledge of skewness and kurtosis can be most helpful in determining what happens when a given scale of measurement is applied to a given group.
CHAPTER 4

Statistical Inference

In the methods described in the first three chapters it was seen how to collect data, organize it and describe it mathematically. Now it is time to make statements about the data and draw inferences from it.

Regression Techniques

Suppose that we have two simultaneous observations about each individual in our sample. Regression techniques are used to predict the values of one variable when the values of the other are known. Only a linear relationship will be shown here. This involves a method of finding a and b, such that \( Y = a + bX \). The best fitting line for ungrouped data can be found by the method of least squares and computation formulas for a and b are readily available, as in a method of error variance. (Zelditch, p. 97). Regression techniques are applicable only when the relationship is linear. A method of determining this relationship is outlined below.

Correlation

Frequently, it is the association of two variables which concerns us (not which variable is independent and which dependent). In this case it is common to measure the degree of mutual relationship between two variables by a correlation coefficient, \( r \). The correlation coefficient is defined as:
\[ r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}, \]

where \((Y_i - \bar{Y})\) represents the deviations of the observed \(Y\) values from the mean \((\bar{Y})\) of the sample, defined analogously to our standard definition for \(X\). The value of \(r\) ranges from -1.00 to +1.00. When \(r = 0\), there is no linear relationship between \(X\) and \(Y\). When \(r\) is -1.00, there is perfect negative correlation (as \(X\) increases, \(Y\) decreases). When \(r\) is +1.00 there is perfect positive correlation (as \(X\) increases, \(Y\) increases). Thus if \(|r|\) is near 1, then the relationship is linear or approximately linear and regression techniques are then applicable to predict one value when the other is known. Again, computation formulas for grouped or ungrouped data are readily available. The important matter in dealing with \(r\) is interpretation.

**Interpretation of \(r\) (Zelditch, p. 108).**

1) \(r\) measures association only; it tells you nothing of cause and effect.

2) \(r^2\) expresses how much variation in \(X\) and \(Y\) is accounted for by their mutual relationship. If \(r = .90\) then \(r^2 (.81\) or \(81\%)\) tells us that 81\% of the total variation in \(X\) and \(Y\) is explained by their mutual relationship.

3) It measures only linear relationships. Some important relationships in sociology are not linear.

4) The correlation of properties of areas or groups (called ecological correlations) involves information about groups rather than individuals and is not a legitimate use of correlation techniques. Correlation of
the median income of a tract and the delinquency rate is an example.
Such information tells us nothing about which individuals are delinquent.

Association

Actually, correlation is just a special case of association. If two attributes are independent, their correlation coefficient is 0; however, the converse is not necessarily true. Two attributes A and B are associated when A effects B or vice versa. (See Meyer, p. 72).

Degree of Effect (Association)

A rough measure, but still a common one in routine analysis, is called E (Greek letter epsilon). E is the percentage difference as seen below. Suppose that we have the following attributes. For a given sample we know the party affiliation (A) and the person last interviewed about politics (B). These facts are first placed in a table, (Zelditch, p. 137).

A (Affiliation of Respondent)

<table>
<thead>
<tr>
<th>Republican</th>
<th>Democrat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rep.</td>
<td>120</td>
</tr>
<tr>
<td>Dem.</td>
<td>40</td>
</tr>
<tr>
<td>160</td>
<td>75</td>
</tr>
</tbody>
</table>

Next they are percentaged thusly,

A (Affiliation of Respondent)

<table>
<thead>
<tr>
<th>R</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>75%</td>
</tr>
<tr>
<td>D</td>
<td>25%</td>
</tr>
<tr>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>
Note, to find the effect of A on B, we have percentaged down. To find the effect of B on A, percentage across.

Now, \( E = 0.75 - 0.33 = 0.42 \). E ranges between -1.00 and +1.00 and is interpreted the same way as the correlation coefficient.

Other measures of association are also available and the choice is dependent on a particular situation.

Control of Confounding Factors.

When two variables appear to be related, there is a danger that they are really not related to each other, but to a third variable independently. To test to see if a third factor, C, might be responsible for a supposed relation between A and B, we stratify A and B by C. This means that we test A and B under different conditions of C. If A and B are genuinely related, then the degree of association will lessen, but will not vanish. If the relation is spurious, due to the effect of the third variable, then the relationship will disappear in the different conditions of C.

The second possibility is that while A and B are not directly related to C, their association may be greater under certain conditions of C. For example, the Catholic suicide rate is about half that of the Protestant. But if we control urban versus rural residence, the rate is about the same in cities and much less in rural areas.

Tests of Significance.

T-test: the t-test has its greatest use in testing whether the means obtained in two different samples could have come from the same population or whether they differ significantly. To do this, we actually test the null hypothesis which says that the means are in fact equal. A t table then gives you a confidence level for when t could have particular values with certain
degrees of freedom. By degrees of freedom, it is meant that if N-1 deviations are known, then the Nth is determined. Only N-1 are free to vary. Because \( \sum (X_i - \bar{X}) = 0 \), t is defined as:

\[
t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}
\]

where \( s_{\bar{X}_1 - \bar{X}_2} \) is the standard error of the mean difference for independent samples. The standard error for any given statistic is the standard deviation of the sampling distribution of that statistic. These are defined for given sampling distribution. (See Spiegel, p. 144).

Thus for a given t and N-1 degrees of freedom, a probability of a t that large is read at the top of a t table. This confidence level gives us a measure of whether the means differ significantly. For a given confidence level, if t is larger than the value given, the null hypothesis can be rejected. If the hypothesis is one-sided, \( \bar{X}_1 \leq \bar{X}_2 \) or \( \bar{X}_1 \geq \bar{X}_2 \), use half of the tabled probability.

The interpretation is as follows: 1) the population from which the samples are drawn must be normal. (Otherwise, non-parametric or distribution-free tests must be used. Such tests are cruder, but more general in application). 2) To use the t-test, the two samples must have equal variances. To test the variances, use the F test. The steps involved are three: a) divide the larger variance by the smaller, and this is \( F \) (variance ratio); b) look at an F table which has degrees of freedom for both samples; c) if the F has a value larger than the tabled F, the variances differ significantly.

3) Observations must be independent. This concerns the research design. For example, tests involving the same individual or paired observations on individuals who are alike in some respect involve correlated observations, not independent
ones. 4) The t-test tells when to reject the null hypothesis only. Acceptance of the alternate hypothesis is dependent on the entire research design, such as the possibility of confounding variables, etc.

Tests of Significance of Proportions.

Suppose that one obtains a given proportion in a sample, call it $p$. Three methods will be discussed for testing whether or not $p$ could have been obtained from a population with theoretical probability $P$ for a given $N$. $p$ is the smallest proportion, be it $p$ or $q$. (Note: $p + q = 1$). Generally, the choice of which distribution to use depends on a rule of thumb. If $Np > 5$, the normal distribution is used. A slightly more sensitive rule is:

$Np + 9p > 9$. If the values obtained in these expressions are less than their right side, the Binomial distribution is probably the correct one. If, however, $N \geq 50$ and $Np < 5$, the Binomial tends to approximate the poisson distribution.

Binomial Distribution.

The Binomial distribution gives the exact probability of obtaining a $p$ for a given $P$. The procedure is quite tedious for large $N$. Similar to the t-test, a null hypothesis is established and an attempt is made to disprove it. Example 1. Suppose that the theoretical probability were $3/6$ and the relative frequency, $4/6$. To test the null hypothesis, one simply looks in a Binomial distribution table for $N = 6$, $p = \frac{1}{2}$ and sees that the $p (X = 4) = .2344$, which is significant and one cannot reject the hypothesis that $p = 4/6$.

Example 2. Suppose one wants to know the chances of obtaining a value other than $p = q = \frac{1}{2}$. This involves a two-tailed test as illustrated below. The null hypothesis is that $p \neq \frac{1}{2}$. From binomial tables, the chances of $p$ falling
anywhere but $p = \frac{1}{2}$ is $1 - P(p = \frac{1}{2}) = .6876$. Thus one cannot reject the null hypothesis and it is quite common, for example, that one could have four males in a sample of six persons who have an equal probability of being male or female.

**Normal Distributions.**

For values when $Np > 5$, one can approximate binomial probabilities with the Normal distribution in a way analogous to the t-test. (See Spiegel, p. 133).

**Chi-Square Test.**

"Chi-square evaluates the probability of obtaining a set of observed frequencies from a population having certain theoretical or assumed frequencies." (Zelditch, p. 280). There are two general uses: Goodness-of-fit tests and tests of independence for contingency tables. In the Goodness-of-fit case, the theoretical frequency may come from anywhere such as a theoretical distribution, statistical data, even the symmetrical nature of the sample (coin, die, etc.). The Chi-square distribution, written $x^2$, is defined as:

$$\sum \frac{(0 - E)^2}{E}$$

where $0$ = the observed frequency, and $E$ = the Expected frequency. Recall that $E = Np$ established by the null hypothesis. The degrees of freedom are still $N-1$ and one can simply compute $x^2$ and see at what confidence level this value or a larger value of $x^2$ would occur. If the value is equal to or larger than the value for a 5% or 1% confidence level, the null hypothesis is rejected. **Example:** Suppose that we take a sample of the number of people living in towns and on farms. We wish to test the null hypothesis that the two proportions are actually equal, the difference being due entirely to sampling error. In table form, the results are as follows

---

*The .5 factor is a correction for continuity made in discrete cases.*
(Zelditch, p. 281):

<table>
<thead>
<tr>
<th></th>
<th>On Farms</th>
<th>In Towns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>65</td>
<td>119</td>
</tr>
<tr>
<td>Expected</td>
<td>92</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>184</td>
<td>184</td>
</tr>
</tbody>
</table>

For the first cell, \( 0 - E = -27 \), i.e. \((65-92)\), and for the second cell, \( 0 - E = 27 \), i.e. \((119-92)\). Computing \( x^2 \) yields \( x^2 = 15.26 \) and chi-square tables tell us that we must reject the null hypothesis.

In tests of independence, the theory is the same as in the goodness-of-fit test. Here one assumes in the contingency table that there is no association in the table. If there is no association in a table, the expected value for a given cell is a product of the marginal proportions (recall the theory of independent probabilities). In terms of frequencies, the expected value in a cell is \( n_i n_j / N \). Using the same formula for \( x^2 \), one can make the same use of the tables as in the goodness-of-fit case. If the value of \( x^2 \) is less than the tabled values for a 5% or 1% level, then the attributes in the tables are indeed independent. Otherwise, there is some degree of association.

Summary of Statistical Inference.

The idea of statistical inference rests firmly on the three following theorems:

"1) The mean of the distribution of \( \bar{X} \) on repeated samplings from the same population will tend toward \( \mu \), the population mean.

2) The standard deviation of the sampling distribution of \( \bar{X} \) will be reasonably estimated by \( s/\sqrt{N} \).

3) The distribution of \( \bar{X} \) will tend toward a normal distribution form."

(Zelditch, p. 313).
These theorems in turn are true only under one of the following conditions:

1) "They will be true if the population from which the samples are drawn are known to be normal.

2) They may be true if the size of the sample is sufficiently large."

(Zelditch, p. 313).

What happens in research is that the first principle is usually justified on the grounds of the second. This is in turn justified by two theorems:

Central Limit Theorem -- The sampling distribution of any statistic will approach the Normal one (regardless of the distribution of the population from which samples are taken) if the population has a finite variance.

Law of Large Numbers -- As $N$ increases, the relative frequency of an event approaches the theoretical probability of that event. The question ultimately arises, however, "just how large must $N$ be so that the Normal approximation is valid?" If one is certain $N$ is not large enough, or is in doubt as to whether or not $N$ is large enough, there are other methods available. It is these methods which will now be discussed.
The Logic of Distribution-free 
and Non-parametric Methods

In recent years, statisticians have developed new methods for the case when \( N \) is small which require no assumptions about the parameter (non-parametric) and no, or few, assumptions about the distribution (distribution-free). It is obvious that such methods have wider application, but as shall be seen, these methods lead to far less precise conclusions about the distribution. They will generally sacrifice some of the information contained in the data on hand.

For example, the Tchebyscheff's Inequality is an example of a distribution free method. "It states that the area which is more than \( k \) sigmas from the mean under any distribution curve, whatever its form, is less than \( 1/k^2 \)." (Zelditch, p. 314).

Thus \( P (|x - u| > k \sigma) < 1/k^2 \). Thus, if \( k = 2 \), at most 25% of the area under any curve is more than \( 2\sigma \)'s away from the mean. Note that this compares with 5% if one knows that the curve is Normal. The point is obvious, if anything is known about the population form or parameter, one is better off using parametric methods.

Order Statistics.

By concerning oneself with the order of a set of data rather than its measure, it is possible to set up a confidence level for the Median, which is used as a measure of central tendency for order statistics. The method comes
directly from probability theory. For example, the Median, \( M \), for a set of ten numbers is the average of the fifth and sixth numbers. Then to find the confidence level that the median is between say the second and ninth observations, we just take one minus the sum of the outside probabilities or \( P \left( X_2 < M < X_9 \right) = \sum_{r=2}^{8} \binom{10}{r} \left( \frac{1}{2} \right)^{10} \) and can be read from Binomial tables to give us a confidence level that \( (X_2 < M < X_9) \).

**Rank Correlation Methods.**

Suppose that there are two kinds of information about a given sample. It is possible to convert the observations into ordered sets and rank the position of each person from highest to lowest for each characteristic. Perfect correlation of the two characteristics would mean that the person who is first for characteristic A will also be first for characteristic B, etc. Speaman's \( r \) gives a way of testing what degree of correlation there is between A and B. Simply square the difference for each person. For example:

**Set 1.**

**Person 3**
Rank on A -- 1  
Rank on B -- 4  

The difference for person 3 is 3, and squared is 9.

Sum the squared differences and compute \( r \) by the following formula:

\[
 r = 1 - 6 \sum \frac{d^2}{N^3 - N}, \quad \text{where } d = \text{difference in each set.} 
\]

\[
-1 \leq r \leq +1 
\]

The Null hypothesis is that the true correlation is 0. There are tables to obtain a confidence level that a given \( r \) value could have happened by chance. If the value of \( r \) is larger than the tabled value for a 5\% confidence level, the null hypothesis is rejected. This table is for \( 4 \leq N \leq 10 \). For \( N \) greater than 10, the \( t \) distribution for an approximation to the sampling distribution of \( r \) is used. In this case, \( t \) is defined thusly:
\[ t = r \sqrt{\frac{(N-2)}{1-r^2}}, \quad d. F = N-2 \]

**Median Test**

For the median test, one has observations on two groups and wants to test the hypothesis that they both came from the same population. Thus, conditions are similar to those involving the t-test in parametric methods, but the assumptions of the t-test are not warranted. The procedure is as follows:

1) Pool both groups and find the common median.

2) Separate groups again and see how many in each group are above, and how many below the median.

3) Put the results in a 2 x 2 table:

<table>
<thead>
<tr>
<th></th>
<th>Sample 1</th>
<th>Sample 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>N above</td>
<td>N above</td>
</tr>
<tr>
<td>-</td>
<td>N below</td>
<td>N below</td>
</tr>
</tbody>
</table>

4) Test the independence by using the Chi-square method. See page 25.

**Sign Test.**

The sign test provides a method for testing the correlation between two sets of data without using the median. Pair up the observations and consider the N differences, \( X_1 - X_2 \). If there are no differences between the two sets of scores, there should be an equal number of pluses and minuses (disregard zeroes). The binomial, \( (p + q)^N \) where \( p = .5 \), then gives a way of using tables to find confidence levels in the same way as shown in previous tests. Note that \( p \) equals the number of pluses, and \( q \) the number of minuses. The Normal approximation for \( N > 10 \) or the Chi-square approximation are applicable.

**Summary of Order Statistics.**
The use of order methods may sacrifice some information, but they are applicable in three areas:

1) Data are inherently in form of a rank order.

2) Data are measured (have number value), but measurements are ambiguous or are not exact.

3) Distribution is not Normal or N is small and the distribution of the sample statistic cannot be assumed to be normal. (Zelditch, p. 331).
CONCLUSION

With this rather limited discussion of non-parametric methods, the purpose intended here is completed. The author hopes that his readers now sense the greater value of knowing how to make appropriate use of statistics, rather than simply how to compute a given statistic. The importance of interpretation cannot be overstressed. A t-test in an inappropriate situation means nothing even if the confidence level is 99.9%. This means that the whole investigation, from data collection to hypothesis testing is equally important. It is hoped that this overview has brought that entire investigation into sharper focus and shed some light on the implications and the interpretations of its many details.
BIBLIOGRAPHY


