DOCUMENT RESUME

ED 069 745 TM 002 200

AUTHOR Blai, Boris, Jr.
TITLE Simple Statistics: - Summarized!
INSTITUTION Harcum Junior Coll., Bryn Mawr, Pa.
PUB DATE Apr 71
NOTE 10p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Analysis of Variance; Correlation; Hypothesis Testing; *Mathematical Logic; Probability Theory; Reliability; *Research Methodology; Standard Error of Measurement; *Statistics; Technical Reports

ABSTRACT

Statistics are an essential tool for making proper judgement decisions. It is concerned with probability distribution models, testing of hypotheses, significance tests and other means of determining the correctness of deductions and the most likely outcome of decisions. Measures of central tendency include the mean, median and mode. A second important class of descriptive statistics is measures of variability described by the standard deviation. Inferential statistics estimates what a total set of measures would be like on the basis of a sample. The standard error of the mean is a statistic that indicates an estimate of the reliability of the sample mean. The t-ratio is the ratio of the expected difference in a set of scores to the obtained difference. This can also be interpreted in relation to the normal curve. The coefficient of correlation is a measure showing the relationship of measurements on one set of variables to the measurements on another set of variables. Two variables can be highly correlated without being causally connected. (DJ)

Harcum Junior College
Research Memorandum

TO: Faculty and Staff
FROM: Office of Research
SUBJECT: Simpler Statistics:- Summarized!

All you have to do to convince yourself of the importance of statistics today is to pick up any current professional journal or even elementary text book in any of the behavioral sciences. All are full of statistical language such as the following: the correlation between the intelligence test scores of offspring and parent is about .50; 30% of sixth grade boys exceed the median of sixth grade girls in reading; the chi-square of 2.5, for one degree of freedom, yields a P which lies between .20 and .10, and hence is not significant; scores were normalized (expressed as T-scores) in order to make them equivalent; this survey employed stratified sampling, the sampling within the strata being random. Obviously, statements like these are well-nigh incomprehensible without a basic working-knowledge of statistics.

Those who object to "memory work" should find statistics a delight, for there are only a few basic relationships to learn, and a considerable amount of reasoning to follow. Since you can add, subtract, divide and multiply, you have all the mathematical background you need to understand what follows. (A descriptive summary of what is discussed in this memorandum begins on page 7.)

There are two general reasons to use statistics: (1) to describe a large and unwieldy mass of numbers, and (2) to make inferences from these data. It is absolutely essential for the data-gatherer to have his figures summarized, reduced to something meaningful and manageable. Descriptive statistics perform this function.

However, in addition to simply describing data, what conclusions may be drawn from them? For example, in noting a difference in the performance between two groups does one conclude that it is a real difference or was it due to chance? The process by which an answer is worked out to a question such as this is known as statistical inference.

Descriptive Statistics

Measures of Central Tendancy

These provide information about the typical performance of the group members being measured. The three most commonly used are the mean, median and mode.

1 - To compute a mean, add up all the raw scores and divide this sum by the total number of scores. The formula is:

$M = \frac{\Sigma X}{N}$ where M = the mean, X = raw score

N = the number of scores (the frequency)
Σ = the Greek capitol letter "sigma," meaning "sum of"

1 - Use the mean when the most stable measure of central tendency is wanted. It is less variable from sample to sample.

2 - When the size of each score should enter in and influence the central tendency.

3 - When correlation coefficients and standard deviations are to be determined from the data obtained.

2 - The median is that point in a distribution of scores above which (or below which) lies 50% of the frequency (N). The formula calculated from a frequency distribution of scores is:

$$\text{Mdn} = l + i \left(\frac{\frac{N}{2} - \text{cum } f_l}{fm} \right)$$

where

i = length of interval
l = lower limit of i upon which the Mdn lies
N/2 = 1/2 the total number of scores (N)
cum f_l = sum of scores on i's below l
fm = frequency on the i containing the Mdn

Use the median when -

1 - there are extreme scores at either end of the series

2 - when certain scores should influence the measure of central tendency, but all that is known about them is that they lie outside the distribution.

3 - The mode is defined as the most oft-recurring measure or score in a series. A formula for approximating the true mode is:

$$\text{Mode} = 3 \text{ Mdn} - 2 \text{ Mean}$$

The mode is often employed as a simple inspectional average - to provide a rough notation of the concentration of scores.

Measures of Variability

To describe a set of data adequately, it is necessary to know, beyond a measure of central tendency, how the individual scores are dispersed about the mean; whether they are all close to the mean or spread over a wide range away from the mean. Measures of variability give an indication of this dispersal.

1 - The Range - The simplest variability measure, it is the difference between the highest and lowest score of a distribution. The range is not a very satisfactory measure of variability because it is based on only two scores. It does not reflect all the data available.

2 - The Standard Deviation - Unlike the range, it is based on all the scores in the distribution. It is a widely used statistic; not only as a measure of variability, but also as a fundamental part of more complicated computations.

The formula for the standard deviation is:

$$s = \sqrt{\frac{\Sigma x^2}{N-1}}$$

where

s = standard deviation
Σ = take the sum of
x = deviation score or the amount a raw score deviates from the mean ($x = X-M$)
N = number of raw scores

3 - The Normal Curve - For many characteristics of living organisms, when a large amount of data are arranged along a scale according to their frequency of occurrence, they tend to approximate a bi-laterally symmetrical shape (one that looks the same on both sides). This shape approximates what is called "normal" and the line that is drawn which joins the tips of all the scores graphically portrayed, is the so-called "normal curve." A distribution of scores that approximates the "normal curve" is called a normal distribution. This concept of the normal curve is extremely useful in statistics.

The normal curve has certain mathematical properties which yield a great deal of information about the variability of scores once the standard deviation for the particular distribution has been computed. For example, between the mean and one standard deviation exactly 34.13% of all the measurements that make up the distribution will occur. Because of the properties of normal distribution, 95.44% of the cases will always be between plus and minus two standard deviations from the mean.

Knowing that fixed values always apply with reference to different points along a normal distribution curve it is then possible to consider the final measure of variability - the standard score.

4 - The Standard Score - The normal curve and its fixed characteristics is very useful to ascertain precisely how one score stands in relation to the others in a distribution. The formula for the standard score is:

$$\text{Standard Score} = \frac{X - M}{s}$$

where

X = Score
M = Mean
s = Standard deviation

Thus, in a distribution of examination scores with a mean of 65 and a standard deviation of 5, someone scoring a 74 has a standard score of 1.80 which may then be interpreted from a Table of Values for the Normal Curve to mean that more than 96% of the scores in that particular distribution lie below the raw score of 74, and that only 4% of those who took the examination did better. In short, we now know, rather precisely, how the 74 score compares with the others who took the examination.

Inferential Statistics

Having discussed how to describe various characteristics of a set of measurements, how does one infer, or make predictions, based on the data collected? And of paramount importance, how does one ascertain the reliability of such predictions?

By definition - the total number of individuals or objects having a defined characteristic are known as the population. A lesser number from among the population are known as a sample of that population. Since populations are usually too large and unwieldy to measure all, a sample (fewer measures) is taken with the hope that it is truly representative of all (the population).

By making a number of "sampling error" statistical calculations, it is possible to obtain a probability estimate of the extent to which the sample measurements differ from the total population.

1 - The Standard Error - When one sample is taken from a population, the standard error is a statistic that estimates the reliability of a mean obtained from this single sample. It tells with what degree of probability the sample represents the population mean. More precisely, it tells how likely an error will occur if the sample mean is treated as if it were the population mean. The standard error formula is:

$$SEm = \frac{s}{\sqrt{N-1}}$$

where

SEm = The standard error of the mean
s = the standard deviation of the sample
N = the number of cases in the sample

The larger the standard error of the mean, the less certainty that the sample is a true representation of the population.

2 - The standard error of the difference between two means

A problem very often faced in education is to determine whether two or more groups differ significantly from each other. Any such difference found between the means of the two groups may be the result of an accident in sampling rather than of any real difference between the groups. In other words, the difference may have been caused simply by chance - sometimes called "sampling fluctuation." The problem, therefore, is to determine how likely it is that any difference between the two groups could be owing simply to chance.

The equation that yields a statistic which estimates the amount of chance variability in a sample of a difference between two means is:

$$SE_D = \sqrt{SE^2_{m_1} + SE^2_{m_2}}$$

where

SE_D = standard error of the difference between two means

$SE^2_{m_1}$ & $SE^2_{m_2}$ = the squared standard error of the mean for samples 1 (m_1) and 2 (m_2)

where

$$SE_m = \frac{s}{\sqrt{N-1}}$$

(see formula in paragraph 1 above)

If the standard error of the difference is great, then the difference between the sample pair would hardly represent the true mean difference.

3 - The t - ratio

To determine if the obtained difference between the means of two groups is likely to be a chance or true one, the following formula, the so-called t-ratio, is used:

$$t = \frac{D\ -}{SE_D}$$

where

t = t-ratio
D = the obtained difference between the means of the two groups
SE_D = the standard error of the means of the two groups

How large must a t-ratio be before an obtained difference between two means can be accepted as significant - that is, before one can assume that it is unlikely to occur by chance? A "convention" or arbitrary agreement stipulates that the obtained difference must be at least large enough so that it could arise by chance alone only 5% of the time if there were no true difference between means. This is called the 5 per cent level of confidence.

A t-ratio of 3 is a virtual certainly (99.87 chances out of 100) that a true difference in means exists; one larger than 3 is that much more assurance of a true difference; and a t-ratio of 1.6 is near certainty (95 chances out of 100- i.e. the so-called 5% level of confidence) that a true difference exists.

Speaking literally - there is never any absolute certai nty in science. All conclusions drawn from experiments contain an element of risk. What the t-ratio procedure permits one to state, quite precisely, is the extent of this risk. By convention, the 5 per cent level of confidence is usually adopted as the cut-off point. However, very frequently, the more exacting 1 per cent or even the .1 per cent or .01 per cent level are used; each one being 10 times less of a "chance" probability e.g. a 1% level indicates 1 chance out of 100 of the difference being due to sampling fluctuations ("chance"); a .1% level = 1 chance out of 1000, and a .01% level = 1 chance out of 10,000.

4 - The Chi-Square Test

This test represents a useful method of comparing experimentally obtained results with those to be expected theoretically on some hypothesis. The equation for chi-square (X^2) is stated as follows:

$$X^2 = \Sigma \left[\frac{f_o - f_e^{\,2}}{fe} \right]$$

where

X^2 = chi-square formula for testing agreement between observed and expected results.
f_o = frequency of occurrence of observed or experimentally determined facts.
f_e = expected frequency of occurrance on some hypothesis

The more closely the observed results approximate to the expected, the smaller the chi-square and the closer the agreement between observed data and the hypothesis being tested. Contrariwise, the larger the chi-square the greater the probability of a real divergence of experimentally observed from expected results. To evaluate the significance of a chi-square, it is determined by reference to a X^2 table of probabilities for the appropriate "degrees of freedom" (df). The number of df = (r-1) (c-1) in which r is the number of rows and c is the number of columns in which the data are tabulated. This reveals for the particular observed measurements the probability of the chi-square being significant, i.e. a P of .05 or less is generally considered as being significant - since this indicates that there is only one chance in 20 (or 5 in 100) of that large or larger a . X^2 occurring through chance variation.

5 - Coefficient of Correlation

The coefficient of correlation is a statistic that is both descriptive and predictive. It is a single number that describes the relationship between two variables, and it also permits the prediction of the score on one variable from the score on the other. There are a number of different kinds of correlational procedures, however, the most commonly used one is the Pearson product-moment correlation, symbolized by r. The formula for r is:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}$$

where

r = the coefficient of correlation
Σxy - Sum of the products of the deviations 'x' & 'y' from their respective means
Σx^2 = sum of the squared deviations 'x'
Σy^2 = sum of the squared deviations 'y'

Correlation coefficients may be either positive, negative, or zero. A positive correlation indicates that as the scores on one variable increase, the scores on the other variable also increase. A negative correlation means that as the scores from one variable increase, the scores from the other variable decrease. A zero correlation means that there is no relationship between the two.

Coefficient correlations range from a perfect positive correlation of +1.00 in descending order to a perfect negative correlation of -1.00; e.g. +1.00, +.99, +.98... +.01, 0.0, -0.1, -.02, -.03... -.98, -.99, -1.00.

It is possible, of course, to get a high correlation simply by chance, and so once again it is necessary to determine the liklihood that the obtained correlation is a chance one. This is done by estimating the amount of variability for a population of correlations by computing the standard error of correlation.

6 - The Standard Error of Correlation

The classical formula for the standard error of correlation is:

$$SE_r = \frac{(1-r^2)}{N}$$

where
SE_r = the standard error of correlation
r = coefficient of correlation
N = number in the sample

After the standard error of correlation is determined, the next and final step, is to determine its significance.

7 - The Significance of the Coefficient of Correlation

As in paragraph 3 above (the t-ratio), a "t" score is computed by dividing the standard error of correlation into the obtained correlation. the Formula is:

$$t = \frac{r}{SE_r}$$

As noted in paragraph 3 above, a t-ratio must be at least 1.6 for it to be significant at the 5 per cent level. If the t-ratio is 1.6, the r could have occurred by chance only 5 times out of 100, and it is therefore considered to be a significant one; i.e., highly unlikely that it would have occurred through chance variation.

A final correlation-causation comment. Just because two variables are highly correlated does not mean that there is a causal connection between them. There is no necessary implication of causality in correlation. In the past twenty years there has been a positive correlation between the number of Churches in the United States and the amount of liquor consumed. Both have increased; the correlation is high and positive. However, this correlation in no sense implies that churches cause liquor consumption, nor does it mean that increased liquor consumption causes an increase in the number of churches built. The reason (or cause) for this correlation is undoubtedly the general increase in both the population and wealth of the country. With more people and more money, the purchase of many things has increased; among them, churches and liquor consumption - two un-causually related facts.

SUMMARY

When we have hundreds, thousands, or even millions of numbers we must make some sense out of them - order them in some way. Statistics enables us to do this, as well as being a major part of decision mathematics which has emerged in our world of uncertainty as an essential tool for making proper judgment decisions.

"Because uncertainty is implicit in nearly everything we do, statistics is concerned with probability distribution models, testing of hypotheses, significance tests, and other means of determining the correctness of our deductions and the most likely outcome of our decisions". (Longley-Cook, 1970, p. 1)

"Measures of central tendency make up one important class of descriptive statistics. Among the measures of central tendency are the mean, median, and mode. These measures are also called averages. The mean is computed by adding all figures with which we are dealing and dividing the sum by the number of figures. The median is a score value that exactly splits the number of scores. One half of the scores lies above this point, and one half lie below this point. The mode is that score which occurs most frequently.

"A second important class of descriptive statistics consists of the measures of variability. The standard deviation is the most useful of the measures of variability. The standard deviation is a sort of average deviation, an average amount that all raw scores deviate from the mean of the raw scores. To compute the standard deviation, we subtract the mean from each raw score. This yields a deviation score, which is then squared. All the squared deviation scores are added together and divided by the total number of deviation scores. The square root of this quotient is then determined. This value is the standard deviation. It tells us whether the scores are grouped close to the mean, or whether they are spread out widely away from the mean. A precise interpretation of the standard deviation can be obtained in terms of the normal curve.

"Most measures obtained in psychology distribute themselves normally. We know rather exactly, from the properties of the normal curve, and the fact that many variables are normally or approximately distributed, what percentages of the total scores are within given score ranges on our measuring dimension. We know, for example, that within plus and minus one standard deviation of the mean, approximately 68 per cent of the total number of cases will lie; that within plus and minus two standard deviations of the mean, approximately 95 percent of all the cases will lie; and that within plus and minus three standard deviations, practically all the cases occur. Thus, by means of the measure of central tendency and variability, and in conjunction with our knowledge of the normal curve, we can get a pretty good descriptive picture of what even a long list of numbers looks like as a whole.

"Frequently, we want to do more than just describe a set of numbers. For practical reasons we are often forced to work with a few measures and then to estimate what the total set of measures would be like; we must estimate population values from sample values, for example. Such estimates fall within the realm of statistical inference. Our estimate of a population value - the mean, for example - will depend on the reliability of our sample mean. If the sample mean is reliable - that is, if we are likely to get the same mean value from successive samples - then we can be reasonable sure that our sample mean closely represents the population mean. But if the sample mean is unreliable so that on successive samples we would be apt to get very different means, then our estimate of the population mean from the sample mean is not likely to be very accurate.

The standard error of the mean is a statistic that indicates an estimate of the reliability of the sample mean. The standard error of the mean tells us the expected variability, or error, in a sample of means. It is a measure of variability, and is interpreted in the same fashion that the standard deviation is by means of the normal curve.

"We frequently must decide whether two samples differ from one another. We need to know whether the means of the two samples are significantly different, for example, or whether the difference could arise only by chance. In order to get an estimate of the possibility that our mean difference is owing to chance, we must first compute the standard error of the difference between the means. The standard error of the difference gives us an estimate of the amount of difference between means that could be expected to arise by chance.

"The next step is to compare this expected chance difference with the difference we obtained with our actual measurements. We perform this operation by dividing the expected chance difference into the actual obtained difference. The result is called a t-ratio, which is the ratio of the expected difference to the obtained difference. The t ratio can also be interpreted by means of the normal curve. If, for example, we obtain a t ratio of 3.00, we know, from the normal curve, that a difference of the size we actually obtained would occur only 1 per cent of the time by chance. Under such circumstances we are likely to conclude that the difference is a real one. Nevertheless, 1 per cent of the time, when there is no real difference, we will still be making a mistake. We are never absolutely certain in any area of science, but we can to a considerable degree state the amount of uncertainty for many of our conclusions. This statement is called a level of confidence.

"If we have one set of measurements on one variable and another set of measurements on another variable, we may want to know the relationship that exists between these two variables. Such a relationship we can determine by means of a statistic called the coefficient of correlation. If the two variables are perfectly related, we will get a coefficient of correlation either of +1.00 of -1.00. Either a perfect positive or a perfect negative correlation indicates that the two variables are perfectly related. If the relationship is perfect and positive, it means that as one variable increases, the other variable also increases proportionately. If the correlation is perfect and negative, it means that as one variable increases the other variable decreases proportionately. If there is no relationship between our two variables, we obtain a coefficient of correlation of 0.00.

"Two variables can be highly correlated without being causally connected. There is, for example, a positive correlation between the number of cars produced on successive years over the past half century and the amount we pay our state legislators per year over the same period. This does not mean that producing more cars causes our legislators to receive more money. The two are correlated, but not causally related."

(Lewis, 1963; pp 46-48)

Boris Blai, Jr. Ed.D.
Director of Research

April 1971

References

Garrett, Henry E. 1962, Elementary Statistics
David McKay Co., Inc.

_______ 1966, Statistics in Psychology and Educat ion
Sixth Edition, David McKay Co., Inc.

Lewis, Donald J. 1963. Scientific Principles of Psychology
Prentic - Hall, Inc.

Longley-Cook, L.H. 1970. Statistical Problems and How to Solve
Them. Barnes and Noble, Inc.