Chow, S. L. (2002). STATISTICS AND ITS ROLE IN PSYCHOLOGICAL RESEARCH. In *Methods in Psychological Research*, In *Encyclopedia of Life Support Systems (EOLSS)*, Eolss Publishers, Oxford, UK, [http://www.eolss.net]

**Siu L. Chow**
*Department of Psychology, University of Regina, Canada*
**Keywords:** associated probability, conditional probability, confidence-interval estimate, correlation, descriptive statistics, deviation score, effect size, inferential statistics, random sampling distribution, regression, standard deviation, standard error, statistical power, statistical significance, sum of squares, test statistic, Type I error, Type II error

## Contents



## Summary

As readers will have noticed, some everyday words are given technical meanings in statistical parlance (e.g. “mean,” “normal,” “significance,” “effect,” and “power”). It is necessary to resist the temptation of conflating their vernacular and technical meanings. A failure to do so may have a lot to do with the ready acceptance of the “effect size” and “power” arguments in recent years.
To recapitulate, statistics is used (i) to describe succinctly data in terms of the shape, central tendency, and dispersion of their simple frequency distribution, and (ii) to make decisions about the properties of the statistical populations on the basis of sample statistics. Statistical decisions are made with reference to a body of theoretical distributions: the distributions of various test statistics that are in turn derived from the appropriate sample statistics. In every case, the calculated test statistic is compared to the theoretical distribution, which is made up of an infinite number of tokens of the test statistic in question. Hence, the “in the long run” caveat should be made explicit in every probabilistic statement based on inferential statistics (e.g. “the result is significant at the 0.05 level in the long run”).
Despite the recent movement to discourage psychologists from conducting significance tests, significance tests can be (and ought to be) defended by (i) clarifying some concepts, (ii) examining the role of statistics in empirical research, and (iii) showing that the sampling distribution of the test statistic is both the bridge between descriptive and inferential statistics and the probability foundation of significance tests.

## 1. Introduction

Statistics, as a branch of applied mathematics, consists of univariate and multivariate procedures. Psychologists use univariate procedures when they measure only one variable; they use multivariate procedures when multiple variables are used (a) to ascertain the relationship between two or more variables, (b) to derive the test statistic, or (c) to extract factors (or latent variables). As multivariate statistics is introduced in *The Construction and Use of Psychological Tests and Measures*, this article is almost exclusively about univariate statistics. The exception is the topic of linear correlation and regression.

The distinction needs to be made before proceeding between the substantive population and the statistical population. Suppose that an experiment is carried out to study the effects of diet supplements on athletic performance. The substantive population consists of all athletes. The sample selected from the substantive population is divided into two sub-samples. The experimental sub-sample receives the prescribed diet supplements and the control sub-sample receives a placebo.
In this experimental context, the two groups are not samples of the substantive population, "all athletes." Instead, they are samples of two statistical populations defined by the experimental manipulation "athletes given diet supplements" and "athletes given the placebo." In general terms, even if there is only one substantive population in an empirical study, there are as many statistical populations as there are data-collection conditions. This has the following five implications.
First, statistics deal with methodologically defined statistical populations. Second, statistical conclusions are about data in their capacity to represent the statistical populations, not about substantive issues. Third, apart from very exceptional cases, research data (however numerous) are treated as sample data. Fourth, testing the statistical hypothesis is not corroborating the substantive theory. Fifth, data owe their substantive meanings to the theoretical foundation of the research (for the three embedding conditional syllogisms, see *Experimentation in Psychology--Rationale, Concepts, and Issues)*.
Henceforth, "population" and "sample" refer to statistical population and statistical sample, respectively. A parameter is a property of the population, whereas a statistic is a characteristic of the sample. A test statistic (e.g. the student-*t*) is an index derived from the sample statistic. The test statistic is used to make a statistical decision about the population.
In terms of utility, statistics is divided into descriptive and inferential statistics. Psychologists use descriptive statistics to describe research data succinctly. The sample statistic (e.g. the sample mean, $\overline{X}$) thus obtained is used to derive the test statistic (e.g. the student-*t*) that features in inferential statistics. This is made possible by virtue of the "random sampling distribution" of the sample statistic. Inferential statistics consists of procedures used for (a) drawing conclusions about a population parameter on the basis of a sample statistic, and (b) testing statistical hypotheses.

## 2. Descriptive Statistics

To measure something is to assign numerical values to observations according to some well-defined rules. The rules give rise to data at four levels: categorical, ordinal, interval, or ratio. A preliminary step in statistical analysis is to organize the data in terms of the research design. Psychologists use descriptive statistics to transform and describe succinctly their data in either tabular or graphical form. These procedures provide the summary indices used in further analyses.

### 2.1. Four Levels of Measurement

Using numbers to designate or categorize observation units is measurement at the nominal or categorical level. An example is the number on the bus that signifies its route. Apart from counting, nominal data are amenable to no other statistical procedure.
An example of ordinal data is the result of ranking or rating research participants in terms of some quality (e.g. their enthusiasm). The interval between two successive ranks (or ratings) is indeterminate. Consequently, the difference between any two consecutive ranks (e.g. Ranks 1 and 2) may not be the same as that between another pair of consecutive ranks (e.g. Ranks 2 and 3).
Temperature is an example of the interval-scale measurement. The size of two successive intervals is constant. For example, the difference between 20°C and 30°C is the same as that between 10°C and 20°C. However, owing to the fact that 0°C does not mean the complete absence of heat (i.e. there is no absolute zero in the Celsius scale), it is not possible to say that 30°C is twice as warm as 15°C.
In addition to having a constant difference between two successive intervals, it is possible to make a definite statement about the ratio between two distances by virtue of the fact that 0 m means no

distance. Hence, a distance of 4 km is twice as far as 2 km because of the absolute zero in the variable, *distance*. Measurements that have properties like those of *distance* are ratio data.

### 2.2. Data—Raw and Derived

Suppose that subjects are given 60 minutes to solve as many anagram problems as possible. The scores thus obtained are raw scores when they are not changed numerically in any way. In a slightly different data collection situation, the subjects may be allowed as much time as they need. Their data may be converted into the average number of problems solved in a 30-minute period or the average amount of time required to solve a problem. That is, derived data may be obtained by applying an appropriate arithmetic operation to the raw scores so as to render more meaningful the research data.

### 2.3. Data Tabulation and Distributions

Data organization is guided by considering the best way (i) to describe the entire set of data without enumerating them individually, (ii) to compare any score to the rest of the scores, (iii) to determine the probability of obtaining a score with a particular value, (iv) to ascertain the probability of obtaining a score within or outside a specified range of values, (v) to represent the data graphically, and (vi) to describe the graphical representation thus obtained.

#### 2.3.1. Simple Frequency Distribution

The entries in panel 1 of Table 1 represent the performance of 25 individuals. This method of presentation becomes impracticable if scores are more numerous. Moreover, it is not conducive to carrying out the six objectives just mentioned. Hence, the data are described in a more useful way by (a) identifying the various distinct scores (the "Score" row in panel 2), and (b) counting the number of times each score occurs (i.e. the "Frequency" row in panel 2). This way of representing the data is the tabular "simple frequency distribution" (or "frequency distribution" for short).

*Table 1. Various ways of tabulating data*

**Panel 1: A complete enumeration of all the scores**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 15 | 14 | 14 | 13 | 13 | 13 | 12 | 12 | 12 | 12 |
| 11 | 11 | 11 | 11 | 11 | 10 | 10 | 10 | 10 | 9 |
| 9 | 9 | 8 | 8 | 7 | | | | | |

**Panel 2: The simple frequency distribution**

| Score | 15 | 14 | 13 | 12 | 11 | 10 | 9 | 8 | 7 |
|---|---|---|---|---|---|---|---|---|---|
| Frequency | 1 | 2 | 3 | 4 | 5 | 4 | 3 | 2 | 1 |

**Panel 3: Distributions derived from the simple frequency distribution**

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| **Score value** | **Frequency** | **Cumulative frequency** | **Cumulative percentage** | **Relative frequency** | **Cumulative relative frequency** |
| 15 | 1 | 25 | 100 | 0.04 | 1.00 |
| 14 | 2 | 24 | 96 | 0.08 | 0.96 |
| 13 | 3 | 22 | 88 | 0.12 | 0.88 |
| 12 | 4 | 19 | 76 | 0.16 | 0.76 |

| | | | | | |
|---|---|---|---|---|---|
| 11 | 5 | 15 | 60 | 0.20 | 0.60 |
| 10 | 4 | 10 | 40 | 0.16 | 0.40 |
| 9 | 3 | 6 | 24 | 0.12 | 0.24 |
| 8 | 2 | 3 | 12 | 0.08 | 0.12 |
| 7 | 1 | 1 | 4 | 0.04 | 0.04 |
| | Total = *25* | | | | |

### 2.3.2. Derived Distributions

The frequency distributions tabulated in panel 2 of Table 1 have been represented in columns 1 and 2 of panel 3. This is used to derive other useful distributions: (a) the cumulative percentage distribution (column 3), (b) the cumulative percentage (column 4), (c) the relative frequency (probability) distribution (column 4), and (d) the cumulative probability distribution (column 6).
Cumulative frequencies are obtained by answering the question "How many scores equal or are smaller than *X*?" where *X* assumes every value in ascending order of numerical magnitude. For example, when *X* is 8, the answer is 3 (i.e. the sum of 1 plus 2) because there is one occurrence of 7 and two occurrences of 8. A cumulative percentage is obtained when 100 multiply a cumulative relative frequency.

A score's frequency is transformed into its corresponding relative frequency when the total number of scores divides the frequency. As relative frequency is probability, the entries in column 5 are the respective probabilities of occurrence of the scores. Relative frequencies may be cumulated in the same way as are the frequencies. The results are the cumulative probabilities.

### 2.3.3. Utilities of Various Distributions

Psychologists derive various distributions from the simple frequency distribution to answer different questions. For example, the simple frequency distribution is used to determine the shape of the distribution (see *Section 2.4.1. The Shape of the Simple Frequency Distribution)*. The cumulative percentage distribution makes it easy to determine the standing of a score relative to the rest of the scores. For example, it can be seen from column 3 in panel 3 of Table 1 that 22 out of 25 scores have a value equal to or smaller than 13. Similarly, column 4 shows that a score of 13 equals, or is better than, 88% of the scores (see column 5).
The relative frequencies make it easy to determine readily what probability or proportion of times a particular score may occur (e.g. the probability of getting a score of 12 is 0.16 from column 5). Likewise, it is easily seen that the probability of getting a score between 9 and 12, inclusive, is 0.64 (i.e. 0.12 + 0.16 + 0.20 + 0.16). The cumulative probability distribution in column 6 is used to answer the following questions:
(a) What is the probability of getting a score whose value is *X* or larger?
(b) What is the probability of getting a score whose value is *X* or smaller?
(c) What are $X_1$ and $X_2$ such that they include 95% of all scores?
The probability in (a) or (b) is the associated probability of *X*. In like manner, psychologists answer questions about the associated probability of the test statistic with a cumulative probability distribution at a higher level of abstraction (see *Section 3.2. Random Sampling Distribution of Means)*. The ability to do so is the very ability required in making statistical decisions about chance influences or using many of the statistical tables.

## 2.4. Succinct Description of Data

Research data are described succinctly by reporting three properties of their simple frequency distribution: its shape, central tendency, and dispersion (or variability).

### 2.4.1. The Shape of the Simple Frequency Distribution

The shape of the simple frequency distribution depicted by columns 1 and 2 in panel 3 of Table 1 is seen when the frequency distribution is represented graphically in the form of a histogram (Figure 1a) or a polygon (Figure 1b). Columns 1 and 6 jointly depict the cumulative probability distribution whose shape is shown in Figure 1c. In all cases, the score-values are shown on the X or horizontal axis, whereas the frequency of occurrence of a score-value is represented the Y or vertical axis.

A frequency distribution may be normal or non-normal in shape. The characterization "normal" in this context does not have any clinical connotation. It refers to the properties of being symmetrical and looking like a bell, as well as having two tails that extend to positive and negative infinities without touching the X axis. Any distribution that does not have these features is a non-normal distribution.

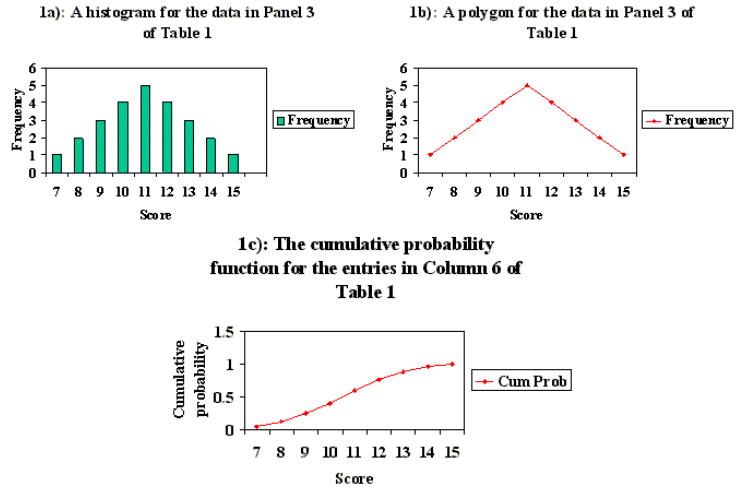


Figure 1. Graphical representations of the simple frequency (a & b) and cumulative probability distributions (c)

### 2.4.2. Measures of Central Tendency

Suppose that a single value is to be used to describe a set of data. This is a request for its typical or representative value in lay terms, but a request for an index of central tendency in statistical parlance. There are three such indices: mode, median, and mean. The mode is the value, which occurs the most often. For example, the mode of the data in Table 1 is 11 (see panel 2). The median of the data set is the value that splits it into two equally numerous halves. It is 11 in the data in Table 1.

The mean is commonly known as the average. Consider the following set of data: 18, 12, 13, 8, 18, 16, 12, 17, and 12. The mean is 14. Introduced in panel 1 of Table 2 is $x$ (i.e. the deviation score of $X$), which is the distance of $X$ from the mean of the data (i.e. $\overline{X}$). That the mean is the center of gravity (or the balance point) of the aggregate may also be seen from panel 1 of Table 2 and the open triangle in Figure 2 in terms of the following analogy.

*Table 2. An illustration of the deviation score $x = (X - \overline{X})$, sum of squares, variance, and standard deviation of a set of scores*

**Panel 1: The deviation score**

| **Scores to the** left **of the mean** = negative **scores vis-à-vis the mean** | **Scores to the** right **of the mean** = positive **scores vis-à-vis the mean** |
|---|---|

| Score ($X$) | Deviation score $x = (X - \bar{X})$ | Deviation score *times* frequency | Score ($X$) | Deviation score $x = (X - \bar{X})$ | Deviation score *times* frequency |
|---|---|---|---|---|---|
| 8 | 8 – 14 = –6 | –6 × 1 = –6 | 16 | 16 – 14 = 2 | 2 × 1 = 2 |
| 12 | 12 – 14 = –2 | –2 × 3 = –6 | 17 | 17 – 14 = 3 | 3 × 1 = 3 |
| 13 | 13 – 14 = –1 | –1 × 1 = –1 | 18 | 18 – 14 = 4 | 4 × 2 = 8 |
| | | | | | |
| The sum of the deviation scores = | | $\Sigma$ = **–13** | The sum of the deviation scores = | | $\Sigma$ = **13** |

**Panel 2: The sum of squares, variance, and standard deviation**

| **1** | **2** | **3** | **4** |
|---|---|---|---|
| | X | $x = (X - \bar{X})$ | $x^2 = (X - \bar{X})^2$ |
| | 18 | 4 | 16.00 |
| | 12 | –2 | 4.00 |
| | 13 | –1 | 1.00 |
| | 8 | –6 | 36.00 |
| | 18 | 4 | 16.00 |
| | 16 | 2 | 4.00 |
| | 12 | –2 | 4.00 |
| | 17 | 3 | 9.00 |
| | 12 | –2 | 4.00 |
| $\Sigma$ = | 126 | 0 | sum of squares = 94.00 |
| $s^2$ = | | | 94 ÷ 8 = 11.75 |
| $s$ = | | | $\sqrt{}$(11.75) = 3.43 |

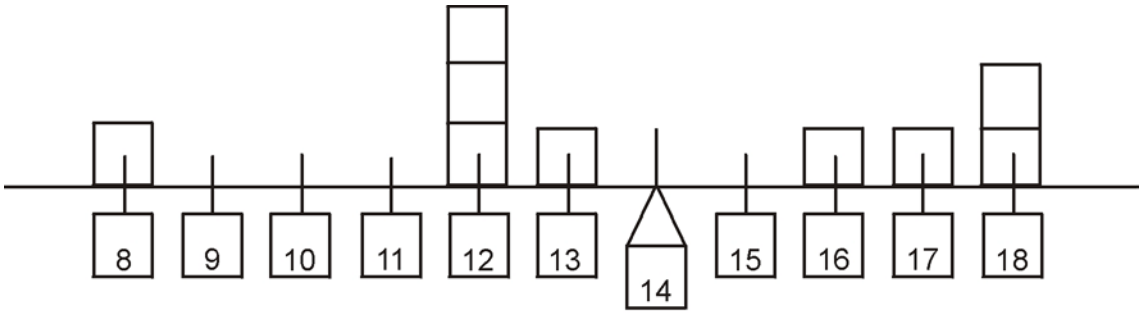


Figure 2. The graphical representation of the mean as the point of balance

Suppose that the scores are the weights of nine children in arbitrary units. It is assumed in Figure 2 that the distance between two successive units of weight is constant. A square represents a child, and the position of the child on the seesaw represents the child's weight. Hence, the three occurrences of 12 are represented by three squares at location 12. The task is to balance the children on the seesaw by (a) arranging them, from left to right, in ascending order of weights, and (b) placing the fulcrum at the place that keeps the seesaw level (i.e. the open triangle in Figure 2).
In order for the seesaw to remain level, the sum of the moments (*mass* × *distance from fulcrum*) on the left should equal that on the right. The location of the fulcrum is 14, which is also the mean of

the scores. This prerequisite state of affairs at the numerical level may be seen from panel 1 of Table 2 by the fact that the sum of the negative deviation scores equals that of the positive deviation scores.

Of importance is the fact that the mean is used as the reference point for transforming the raw scores into their respective deviation scores. The deviation score of $X$ ($x$) shows how far, as well as in what direction, it is away from $\overline{X}$ (2 units above 14 in the case when $X = 16$). This foreshadows the fact that these deviation scores are the basis of all indices of data dispersion, the topic of *Section 2.4.4. Measures of Dispersion*. Meanwhile, it is necessary to introduce the degrees of freedom associated with $\overline{X}$.

### 2.4.3. Degrees of Freedom (*df*)

As the sample size is nine in the example in Table 2, there are nine deviation scores. Suppose that we are to guess what they are. We are free to assume any value for each of the first eight deviation scores (e.g. –1, –2, –2, –2, 2, 3, 4, and 4). These eight deviation scores sum to 6.

Given that the deviation scores of the sample must sum to 0, we are not free to assign any value other than –6 to the ninth deviation score. This means that the ninth score is also not free to vary. In other words, only ($n$ – 1) of the sample of $n$ units are free to assume any value if the deviation scores are derived with reference to $\overline{X}$. Hence, the parameter ($n$ – 1) is the degrees of freedom associated with $\overline{X}$. Such a constraint is not found when the deviation scores of the sample are derived with reference to $u$.

### 2.4.4. Measures of Dispersion

The frequency distribution in panel 2 of Table 1 makes explicit the fact that the largest score value in condition E is 15, whereas the smallest score value is 7. These two values define the range of the scores. The range is an index of data dispersion (or the variation in values among the data). A larger numerical value means greater variability. The range in the example is 8. However, the range gives only a rough indication of data dispersion. Moreover, it is not useful for transforming data or making statistical decisions. For more sophisticated purposes, the index of data dispersion to use is the standard deviation.

Of interest at the conceptual level are that (a) "deviation" in "standard deviation" refers to the deviation score illustrated in panel 1 of Table 2, and (b) "standard" refers to a special sort of pooling procedure. For example, to calculate the standard deviation of the scores in question, each of the deviation scores [i.e. $x = (X - \overline{X})$] is squared [i.e. $x^2 = (X - \overline{X})^2$] (see columns 3 and 4 in panel 2 of Table 2), and all the squared deviation scores are summed together. The sum of all squared deviation scores is called the "sum of squares" (94 in the example; see row 11).

The variance is obtained when the sum of squares is divided by the degrees of freedom ($df = n - 1$), where $n$ is the sample size ($s^2 = 11.75$ in the example; row 12). In other words, the variance is the average squared deviations. The standard deviation is the result of taking the square root of the variance ($s = 3.43$; row 13). It is in this sense that the standard deviation is the result of pooling all deviation scores. In such a capacity, the standard deviation is an index of data dispersion.

## 2.5. Standardization

It is not easy to compare the costs of an automobile between two countries when they have different costs of living. One solution is to express the cost of the automobile in terms of a common unit of measure, a process called "standardization." For example, we may quote the automobile's costs in the two countries in terms of the number of ounces of gold.

Similarly, a common unit of measure is required when comparing data from data sets that differ in data dispersion. Specifically, to standardize the to-be-compared scores $X_A$ and $X_B$ is to transform

them into the standard-score equivalent ($z$), by dividing ($X_A - u_A$) and ($X_B - u_B$) by their respective standard deviations ($\sigma_A$ and $\sigma_B$).

If standardization is carried out for all scores, the original simple frequency distribution is transformed into the frequency distribution of $z$ scores. The mean of the $z$ distribution is always zero and its standard deviation is always one. Moreover, the distribution of $z$ scores preserves the shape of the simple frequency distribution of the scores. If the original distribution is normal in shape, the result of standardizing its scores is the "standard normal distribution," which is normal in shape, in addition to having a mean of zero and a standard deviation of one.

The entries in the $z$ table are markers on a cumulative probability or percentage distribution derived from the standard normal curve. It is in its capacity as a cumulative probability distribution that the distribution of the test statistic (e.g. z, $t$, F, or $\chi^2$) is used to provide information about the **long-run probability** (a) that a population parameter would lie within two specified limits (the confidence-interval estimate), or (b) that the sample statistic has a specific associated probability (for the role of the long-run probability in tests and measurements, see *The Construction and Use of Psychological Tests and Measures)*.

### 2.6. Correlation and Regression

Another major function of descriptive statistics is to provide an index of the relationship between two variables. The correlation coefficient is used to describe the relationship between two random variables. The regression coefficient is used when only one variable is random and the other in controlled by the researcher.

#### 2.6.1. Linear Correlation

Suppose that 10 individuals are measured on both variables $X$ and $Y$, as depicted in each of the three panels in Table 3. Depicted in panel 1 is the situation in which increases in $Y$ are concomitant with increases in $X$. While a perfect positive correlation has a coefficient of 1, the present example has a positive correlation of 0.885. The data show a trend to move from bottom left upwards to top right, as may be seen from Figure 3a.

*Table 3. Some possible relationships between two variables*

**Panel 1: Positive correlation**

| | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** |
|---|---|---|---|---|---|---|---|---|---|---|
| X | 7 | 13 | 2 | 4 | 15 | 10 | 19 | 28 | 26 | 22 |
| Y | 3 | 6 | 2 | 5 | 14 | 10 | 8 | 19 | 15 | 17 |

**Panel 2: Negative correlation**

| | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** |
|---|---|---|---|---|---|---|---|---|---|---|
| X | 22 | 26 | 28 | 19 | 10 | 15 | 4 | 2 | 13 | 7 |
| Y | 3 | 6 | 2 | 5 | 14 | 10 | 8 | 19 | 15 | 17 |

**Panel 3: Zero correlation**

| | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** |
|---|---|---|---|---|---|---|---|---|---|---|
| X | 10 | 19 | 17 | 3 | 15 | 6 | 2 | 5 | 14 | 8 |
| Y | 7 | 13 | 2 | 4 | 15 | 10 | 19 | 28 | 26 | 22 |

**Panel 4: A non-linear relationship**

| | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** |
|---|---|---|---|---|---|---|---|---|---|---|
| X | 7 | 13 | 2 | 4 | 15 | 10 | 19 | 28 | 26 | 22 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Y | 7 | 8 | 2 | 5 | 11 | 10 | 8 | 1 | 4 | 5 |

**Panel 5: Data used to illustrate linear regression**

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| X | 3 | 5 | 7 | 9 | 11 | 13 | 15 | 17 | 19 | 21 |
| Y | 8 | 12 | 11 | 14 | 15 | 12 | 14 | 19 | 20 | 20 |

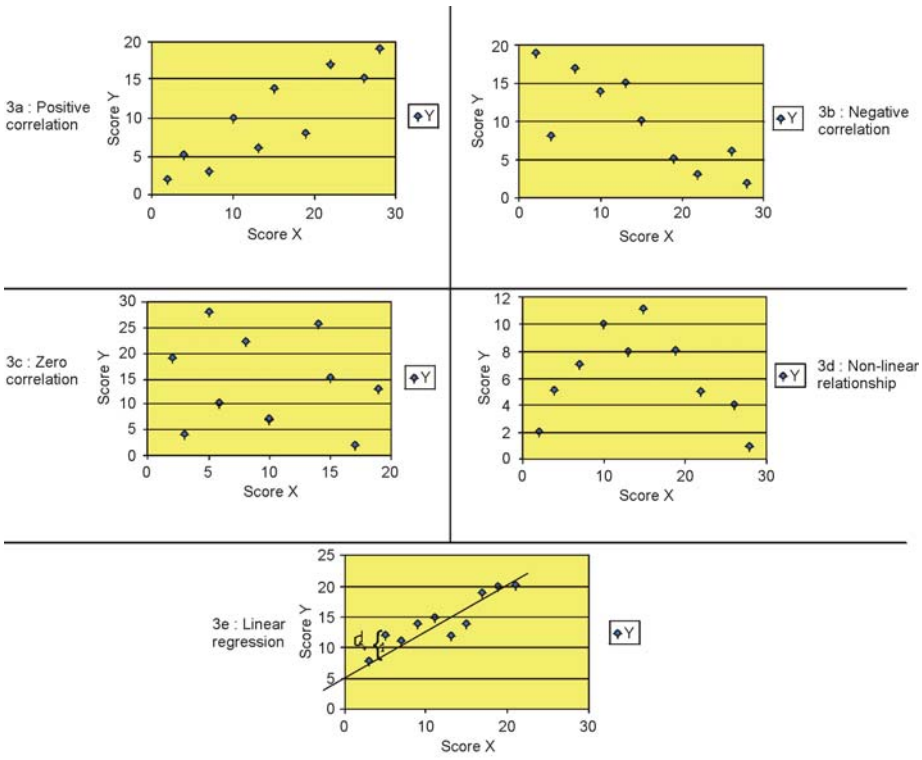


Figure 3. Graphical representation of some relationships between two variables

The data tabulated in panel 2 of Table 3 have been depicted in Figure 3b. The data have a trend of moving from top left downward to bottom right. This pattern is typical of a negative correlation: *X* and *Y* are inversely related (a coefficient of –0.81 in the present example). A perfect negative correlation has a coefficient of –1.

Figure 3c depicts the data tabulated in panel 3 of Table 3. The data show a correlation coefficient of –0.161, which does not differ significantly, from 0 (see *Section 4.5. The Meaning of Statistical Significance)*. The scatter plot assumes the form of a circle, which is indicative of no relationship between the two variables.

### 2.6.2. Non-Linearity

Although the correlation is not perfect in either Figure 3a or 3b, the data nonetheless show a linear trend in the sense that, when a straight line is drawn through the main body of the data points, the resultant line gives a good representation of the points. Such is not the case with the plot in Figure 3d, which represents the data shown in panel 4 of Table 3. The correlation coefficient in Figure 3d is –0.204, which does not differ significantly from 0. However, it would be incorrect to conclude that there is no relationship between *X* and *Y*.

The non-linear trend in the data in Figure 3d means that the exact relationship between *X* and *Y* in panel 4 of Table 3 depends on the range of *X*. Specifically, there is a positive relationship between *X*

and $Y$ when the value of $X$ is small. A negative relationship is found with larger values of $X$. There may be no relationship between $X$ and $Y$ in the medium range of $X$ values.
Taken together, Figures 3c and 3d make clear that the correlation coefficient alone is not sufficient for interpreting correlational data. A scatter plot of the data is necessary. Moreover, Figure 3d shows that correlational data based on a limited range of either of the two variables is ambiguous.

### 2.6.3. Linear Regression

The correlation coefficient informs researchers the extent to which variables $X$ and $Y$ are related. However, it conveys only ordinal information. For example, given three correlation coefficients 0.7, 0.6, and 0.5, we can only say that (a) the first one indicates a closer relationship than the second one, and (b) the second one signifies a closer relationship than the third one. However, we cannot know that the difference between the first two is the same as that between the second and third coefficients. Moreover, the correlation coefficient does not enable us to tell how much change there is in $Y$ per unit change in $X$, or vice versa.
Suppose that the data in panel 5 of Table 3 are obtained by manipulating $X$ and measuring $Y$. Recall that the mean is the point of balance of the data. Likewise, we may draw a line through the data depicted in Figure 3e to represent the relationship between $X$ and $Y$. To the extent that the line is a valid representation of the scatter plot, it is possible to tell the amount of change in $Y$ per unit change in $X$. In such a capacity, the solid line is the regression line (or the line of prediction).
At first glance, drawing such a prediction line seems a non-exact task because many such lines may be drawn. However, the method of least squares is used to decide the best fitting line. Specifically, the dotted line marked $d_i$ in Figure 3e represents dropping a line perpendicular to the X axis from the datum, cutting the solid line at $Y'$. The difference between $Y$ and $Y'$ is $d_i$, which is squared. The sum of the 10 $(d_i)^2$ in the present example is the "sum of squares of prediction." It is an index of the error of prediction.
Given any such line, there are as many $(d_i)^2$ as there are data points. Moreover, each line gives rise to its own set of $(d_i)^2$. The line that gives rise to the smallest error of prediction is chosen as the best fitting line (hence, the "least squares" characterization of the method). The method of least squares gives rise to Equation (1):

$$Y' = a + bX, \tag{1}$$

where $Y'$ is the predicted value of $Y$; $a$ is the zero intercept and $b$ is the regression coefficient. Specifically, $b$ describes the amount of change in $Y$ per unit change in $X$. Numerically, the zero intercept ($a$) represents the value of $Y$ when $X$ is zero. Its conceptual meaning depends on the substantive meaning of the research manipulation. Suppose that $Y$ represents examination grade and $X$ represents the number of hours of extra tutoring. The zero intercept represents the examination grade when there is no extra tutorial. However, researchers sometimes carry out regression analysis even though $X$ is not a manipulated variable. The zero intercept may not have any substantive meaning under such circumstances.

## 3. Bridging Descriptive and Inferential Statistics

Bridging descriptive and inferential statistics are various theoretical distributions: the random sampling distributions of various test statistics. In what follows, the meanings of "random sampling" and "all possible samples" are introduced. An empirical approximation to the "random sampling distribution of the differences between two means of samples of sizes $n_1$ and $n_2$" (or "sampling distribution of differences" henceforth) will be used (a) to describe the theoretical properties, as well as the utility, of the theoretical distribution, and (b) to introduce the rationale of statistical hypothesis testing.

### 3.1. Random Sampling

Suppose that the population of interest consists of the following scores: 1, 2, 3, 4, 5, 6, and 7. The population size (N) is 7; its mean ($u$) is 4; and its standard deviation ($\sigma$) is 2. Shown in panel 1 of Table 4 are the 49 possible combinations of two units selected with replacement from the population. By "with replacement" is meant the procedure in which the item selected on any occasion is returned to the population before the next item is selected. That is, the same item may be selected again.

*Table 4. All possible samples of size 2 from a population of size 7*

**Panel 1: Population P: 1, 2, 3, 4, 5, 6, 7; N = 7; $u$ = 4; $\sigma$ = 2**

| | | | | | | |
|---|---|---|---|---|---|---|
| 1, 1 | 2, 1 | 3, 1 | 4, 1 | 5, 1 | 6, 1 | 7, 1 |
| 1, 2 | 2, 2 | 3, 2 | 4, 2 | 5, 2 | 6, 2 | 7, 2 |
| 1, 3 | 2, 3 | 3, 3 | 4, 3 | 5, 3 | 6, 3 | 7, 3 |
| 1, 4 | 2, 4 | 3, 4 | 4, 4 | 5, 4 | 6, 4 | 7, 4 |
| 1, 5 | 2, 5 | 3, 5 | 4, 5 | 5, 5 | 6, 5 | 7, 5 |
| 1, 6 | 2, 6 | 3, 6 | 4, 6 | 5, 6 | 6, 6 | 7, 6 |
| 1, 7, | 2, 7 | 3, 7 | 4, 7 | 5, 7 | 6, 7 | 7, 7 |

**Panel 2: Various distributions of all possible sample means from panel 1**

| **1** | **2** | **3** | **4** |
|---|---|---|---|
| Value of Mean | Freq of Value | Long-run probability | *Long-run cumulative probability* |
| **7** | 1 | **0.0204** | *1.0000* |
| **6.5** | 2 | **0.0408** | *0.9792* |
| **6** | 3 | **0.0612** | *0.9384* |
| **5.5** | 4 | **0.0816** | *0.8772* |
| **5** | 5 | **0.102** | *0.7956* |
| **4.5** | 6 | **0.1224** | *0.6936* |
| **4** | 7 | **0.1428** | *0.5712* |
| **3.5** | 6 | **0.1224** | *0.4284* |
| **3** | 5 | **0.102** | *0.306* |
| **2.5** | 4 | **0.0816** | *0.204* |
| **2** | 3 | **0.0612** | *0.1224* |
| **1.5** | 2 | **0.0408** | *0.0612* |
| **1** | 1 | **0.0204** | *0.0204* |

$u_{\bar{X}} = u = 4;\ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{2}} = 1.4$

It is important to note that the 49 samples in panel 1 are not necessarily the outcomes of 49 selection trials. They are logical possibilities only. However, they are useful for defining the crucial term "random sampling." The selection is random if all of the logically possible samples have an equal chance of being selected.

### 3.2. Random Sampling Distribution of Means

Each of the 49 possible samples gives a mean. As may be seen from column 1 in panel 2 of Table 4, some samples have the same mean. Depicted in column 2 are the frequencies of various values of the mean. Shown in column 3 are their probabilities of occurrence **in the long run** (i.e. repeating the random selection process **an infinite number of times**), and their **long-run** cumulative probabilities (column 4). The "in the long run" caveat highlights the fact that an empirical approximation to the logical possibilities depicted in columns 1 and 2 of panel 2 requires carrying out the random sampling procedure an infinite number of times.

Taken together, the entries in columns 1 and 2 of panel 2 form a frequency distribution. In such a capacity, it has a mean ($u$ = 4) and a standard deviation ($\sigma$ = 1.4). However, this is not a frequency distribution of raw scores. Instead, it is the distribution of the means of random samples of two scores. That is, it is a frequency distribution at a higher level of abstraction. This more abstract distribution is called the "random sampling distribution of means." Its mean is the "mean of means" ($u_{\overline{X}}$) and its standard deviation is the "standard error of means" ($\sigma_{\overline{X}}$). The parameters $u_{\overline{X}}$ and $\sigma_{\overline{X}}$ are indices of the central tendency and dispersion, respectively, of the sampling distribution of means. Note that $u_{\overline{X}}$ is numerically equal to $u$, and that $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$.

To reiterate, the random sampling distribution of means is a distribution at the higher level of abstraction than the distribution of the scores. It is a theoretical distribution of the chosen statistic ($\overline{X}$ in the present example). The '"theoretical" characterization signifies that the distribution is based on the mathematical derivation of what should be the result if an infinite number of samples of the same size is selected randomly. However, its parameters bear a systematic relationship with the population parameters. This relationship makes it possible to talk about the population's parameter on the basis of what is known about its corresponding statistic in a randomly selected sample. This point may be explicated with the sampling distribution of differences.

### 3.3. The Random Sampling Distribution of Differences

Shown in Table 5 is an empirical approximation to the theoretical distribution. Underlying the scenario are two statistical populations that have the same mean ($u_1 = u_2 = 4.812$) and the same standard deviation ($\sigma_1 = \sigma_2 = 0.894$). Hence, the difference between the two population means is zero. The following procedure was carried out:

(a) A sample of 25 was selected randomly with replacement from each of the two statistical populations.
(b) The means of the two samples were ascertained (e.g. $\overline{X}_1 = 4.96$ and $\overline{X}_2 = 4.84$ in row 1 of Table 5).
(c) The difference between $\overline{X}_1$ and $\overline{X}_2$ was then determined (i.e. 4.96 – 4.84 = 0.12 in row 1 of Table 5).
(d) Units of the two samples were returned to their respective populations.
(e) Steps (a) through (d) were repeated 30 times.

T*able 5. An empirical approximation to the random sampling distribution of differences*

| | **$N_1 = 1328$<br>$u_1 = 4.812$<br>$\sigma_1 = 0.894$<br>$n_1 = 25$** | **$N_2 = 1328$<br>$u_2 = 4.812$<br>$\sigma_2 = 0.894$<br>$n_2 = 25$** | **$u_1 - u_2 = 0$** |
|---|---|---|---|
| Sample-pair | $\overline{X}_1$ | $\overline{X}_2$ | $(\overline{X}_1 - \overline{X}_2)$ |
| 1 | 4.96 | 4.84 | 0.12 |
| 2 | 4.76 | 4.92 | –0.16 |

| | | | |
|---|---|---|---|
| 3 | 4.44 | 4.72 | –0.28 |
| 4 | 4.56 | 5.00 | –0.44 |
| 5 | 4.84 | 4.68 | 0.16 |
| 6 | 4.80 | 5.00 | –0.20 |
| 7 | 4.88 | 4.72 | 0.16 |
| 8 | 4.96 | 4.88 | 0.08 |
| 9 | 4.56 | 4.64 | –0.08 |
| 10 | 4.84 | 4.96 | –0.12 |
| 11 | 4.80 | 4.72 | 0.08 |
| 12 | 4.80 | 4.76 | 0.04 |
| 13 | 4.88 | 4.72 | 0.16 |
| 14 | 4.80 | 4.72 | 0.08 |
| 15 | 4.80 | 4.48 | 0.32 |
| 16 | 4.96 | 4.96 | 0 |
| 17 | 4.88 | 4.72 | 0.16 |
| 18 | 4.76 | 4.64 | 0.12 |
| 19 | 4.76 | 5.08 | –0.32 |
| 20 | 4.96 | 5.12 | –0.16 |
| 21 | 5.20 | 4.76 | 0.44 |
| 22 | 4.60 | 4.72 | –0.12 |
| 23 | 4.76 | 4.76 | 0 |
| 24 | 4.68 | 5.04 | –0.36 |
| 25 | 4.84 | 4.84 | 0 |
| 26 | 4.68 | 5.04 | –0.36 |
| 27 | 4.76 | 4.72 | 0.04 |
| 28 | 4.76 | 4.84 | –0.08 |
| 29 | 4.84 | 5.08 | –0.24 |
| 30 | 4.60 | 5.04 | –0.44 |

| | |
|---|---|
| $u_{(\overline{X}_1 - \overline{X}_2)} = -0.047$ | $\sigma_{(\overline{X}_1 - \overline{X}_2)} = 0.220$ |

Shown in columns $\overline{X}_1$ and $\overline{X}_2$ of Table 5 are the means of the 30 samples randomly selected from two statistical populations, 1 and 2, respectively. Their differences are shown in the $(\overline{X}_1 - \overline{X}_2)$ column. An examination of either the $\overline{X}_1$ or the $\overline{X}_2$ column shows that (a) samples of the same size selected randomly from the same population may have different means, and (b) samples selected randomly from their respective statistical populations that have the same *u* may not have the same $\overline{X}$.

When the 30 differences shown in the $(\overline{X}_1 - \overline{X}_2)$ column are represented graphically, it is an empirical approximation to the theoretical sampling distribution of differences. As such, it has a mean, called the "mean difference" $(u_{(\overline{X}_1 - \overline{X}_2)})$, and a standard deviation, called the "standard error of differences" $(\sigma_{(\overline{X}_1 - \overline{X}_2)})$.

### 3.4. Theoretical Properties of the Sampling Distribution of Differences

The sampling distribution of differences is neither of the two statistical populations. It is the theoretical distribution of differences between an infinite number of pairs of samples drawn randomly from the two statistical populations. Nonetheless, the properties of this theoretical distribution are related to those of the two underlying statistical populations. Specifically, the *mean difference* equals the difference between the means of the two statistical populations (i.e. $u_{(\overline{X}_1 - \overline{X}_2)} = (u_1 - u_2)$), and the exact relationship between the standard error of the difference and the standard deviations of the two statistical populations depends on whether or not the two methodologically defined statistical samples are independent. These theoretical properties make it possible to make a decision about the difference between the means of two statistical populations on the basis of that between two sample means.

Represented in the sampling distribution of differences are the frequencies of all possible differences that can occur by random chance, given the two statistical populations. The test statistic, *t*, is the result of standardizing the difference between two sample-means in terms of the estimated standard error of the difference. Hence, the *t*-distribution is obtained when all possible differences between two sample-means are standardized. Consequently, given a particular difference between two sample-means, it is possible to determine its associated probability with reference to the *t*-distribution.

In sum, the foundation of probability statements in inferential statistics is ultimately the sampling distribution of the test statistic that is contingent on chance. The applicability of these probability statements is based on the assumption that chance is the sole determinant of data dispersion.

## 4. Inferential Statistics

Psychologists apply inferential statistics to decide whether or not there is statistical significance with reference to a criterion value set in terms of the distribution of the test statistic. As an example, consider the case in which the experimental and control conditions are the partial-report and whole-report tasks (see Appendix 1 of *Experimentation in Psychology--Rationale, Concepts, and Issues)*. The reasoning that gives rise to the decision criterion is shown in panel 1 of Table 6.

*Table 6. The conditional syllogisms implicated in testing $H_0$*

**Panel 1: The reasoning that gives rise to the decision criterion used in panel 2 (adopted from Table 7 of "Laboratory Experimentation")**

| | |
|---|---|
| Experimental hypothesis | **If** only raw sensory information is available in the buffer, **then** partial-report superiority is found when a spatial cue is used |
| *Complement of experimental hypothesis* | *If the storage format envisaged in the theory is false, then there is no partial-report superiority with a spatial cue* |
| Statistical alternative hypothesis ($H_1$) | **If** the experimental hypothesis is true, **then** $H_1$: $u_{\text{partial report}} > u_{\text{whole report}}$ |
| *Statistical null hypothesis ($H_0$)* | *If the experimental hypothesis is false, then $H_0$: $u_{\text{partial report}} \leq u_{\text{whole report}}$* |
| Sampling distribution based on $H_1$ | If $H_1$ is used, the probability associated with a *t*-value as extreme as 1.729 is **not known** |
| *Sampling distribution based on $H_0$* | *If $H_0$ is used, the probability associated with a t-value as extreme as 1.729 is 0.05 in the long run* |

**Panel 2: Two conditional syllogisms involving the sampling distribution**

| | Criterion exceeded | Criterion not exceeded |
|---|---|---|
| Major premise | *If* calculated $t >$ (criterion = 1.729), *then* not $H_0$ | *If* calculated $t \leq$ (criterion = 1.729), *then* $H_0$ |
| Minor premise | $t >$ (criterion = 1.729) [e.g., $t = 2.05$] | $t \leq$ (criterion = 1.729) [e.g.. $t = 1.56$] |
| Conclusion | Not $H_0$ | $H_0$ |

**Panel 2: The disjunctive syllogism that decides between $H_1$ and $H_0$**

| | Statistical significance | No statistical significance |
|---|---|---|
| Major premise | $H_1$ or $H_0$ | $H_1$ or $H_0$ |
| Minor premise | Not $H_0$ | $H_0$ |
| Conclusion | $H_1$ | Not $H_1$ |

### .1. Experimental Hypothesis versus Statistical Hypothesis

To begin with, the experimental expectation of partial-report superiority shown in row 1 of Table 6 is translated into a directional difference between two parameters in the statistical alternative hypothesis ($H_1$: $u_{\text{partial report}} > u_{\text{whole report}}$ in row 3). As may be recalled, the *mean difference* equals the difference between two population means. For this reason, there is no information about the mean difference because the researchers do not know the means of the two statistical populations. If the *mean difference* is not known, it is not possible to determine which sampling distribution of differences to use. It is for this reason that psychologists appeal to the logical complement of the experimental hypothesis, which denies the hypothetical property, envisaged in the theory.

### 4.2. The Implication of $H_0$

The statistical null hypothesis ($H_0$: $u_{partial\ report} \leq u_{whole\ report}$) is the statistical representation of the logical complement of the experimental hypothesis (see rows 1 and 4). This $H_0$ stipulates that the sampling distribution with a *mean difference* of zero be used. Recall from Table 5 that the difference between two sample means may not be zero even though the samples are selected randomly from two populations that have the same mean. In other words, the difference between $\overline{X}_{partial\ report}$ and $\overline{X}_{whole\ report}$ need not be zero even if $H_0$ is true (i.e. chance is the sole determinant of data dispersion). Nonetheless, most of the non-zero differences are close to zero. Moreover, the sampling distribution with a *mean difference* of zero is informative as to the probability of obtaining a difference that equals, or is larger than, 95% of all possible differences. Such a difference gives rise to $t_{(df = 19)} = 1.729$ in the long run when $n = 20$ (with the repeated-measures design). That is, the probability of obtaining a difference that gives rise to $t_{(df = 19)} \geq 1.729$ is 0.05 or lower, as may be seen from row 6 in Table 6.

### 4.3. The Decision Rule—Criterion and Conditional Probability

Suppose that it is rare that Event E would occur when chance is the cause. It is not unreasonable to ignore chance as an explanation of E under such circumstances, particularly when there is a well-defined way of stipulating the meaning of "rare." Suppose further that Event E is the data that give rise to a calculated $t_{(df = 19)}$ of 1.85, with $df = 19$. In exceeding the critical $t$ value of 1.729, the associated probability of E is lower than 0.05 with reference to the sampling distribution of differences based on $H_0$.

At the same time, psychologists adopt in most research the convention that anything that happens five out of 100 times **in the long run** is a rare event. Hence, it is deemed unlikely that the data (whose associated probability is 0.05 or smaller) have been obtained from the sampling distribution

based on $H_0$. The outcome is characterized as "statistically significant" in such an event. By the same token, data that produce a calculated *t* that is smaller than 1.729 has a probability higher than 0.05 occurring **in the long run**, given the sampling distribution based on $H_0$. The decision is that such an outcome occurs frequently enough for maintaining that $H_0$ is true. Such an outcome is a statistically non-significant result.

### 4.4. The Level of Significance

The demarcation between significant and non-significant results is based on the critical *t* value identified in terms of the "five out of 100 times in the long run" criterion in the present example. The 0.05 value is called the "level of significance." The significance level is an index of the strictness of the decision in the sense that it stipulates the probability of committing the Type I error, the error of rejecting $H_0$ **when** $H_0$ is true (see panel 1 of Table 7). The boldface "**when**" emphasizes the fact that the Type I error is a conditional event. Hence, the probability of the Type I error is a conditional probability (for the conditional nature of many statistical indices, see *The Construction and Use of Psychological Tests and Measures)*.

*Table 7. Two types of errors in statistical hypothesis testing*

**Panel 1: Statistical decision in terms of $H_0$ only**

| **Decision made with reference to the $\alpha$ level** | **State of affairs** | | **Power** | **Underlying distribution** |
|---|---|---|---|---|
| | $H_0$ is true | $H_0$ is not true | | The random sampling distribution of differences |
| Accept $H_0$ | Correct acceptance of $H_0$ | Type II error<br>*p*(Type II error) = $\beta$ | Not possible | |
| Reject $H_0$ | Type I error<br>*p*(Type I error) = $\alpha$ | Correct rejection of $H_0$ | | |

**Panel 2: $H_1$ introduced in the statistical decision in power analysis**

| **Decision made with reference to the $\alpha$ level** | **State of affairs** | | **Power** | **Underlying distributions** |
|---|---|---|---|---|
| | $H_0$ is true | $H_1$ is true | | Two distributions of population scores predicated on $H_0$ and $H_1$ |
| Accept $H_0$ | Correct acceptance of $H_0$ | Type II error<br>*p*(Type II error) = $\beta$ | 1– $\beta$ | |
| Reject $H_0$ | Type I error<br>*p*(Type I error) = $\alpha$ | Correct acceptance of $H_1$ | | |

The significance level is arbitrary in the sense that psychologists may adopt the "one out of 100 times in the long run" value as the significance level (i.e. the 0.01 level). The choice of the significance level depends on common practice or some non-statistical considerations. Be that as it may, although the 0.01 level is stricter than the 0.05 level, it does not follow that a decision based on the former is superior to one based on the latter. The theoretical foundation of the research, the validity of the research design, and the appropriateness of the experimental procedure collectively determine data quality, not the decision about statistical significance.

For example, the bulk of experiments in support of the partial-report superiority described in Appendix 1 of *Experimentation in Psychology--Rationale, Concepts, and Issues* is established with $\alpha = 0.05$ when subjects were given extensive training on the partial-report task (e.g. 100 trials of training). Hence, it is necessary to use a more stringent criterion (e.g. $\alpha = 0.01$) if subjects are not given a comparable amount of training on the partial-report task (e.g. fewer than 50 training trials). The quality of the data is determined by how well trained the subjects are, as well as the number of trials in a session. The alpha level has nothing to say about how the data are collected.

### 4.5. The Meaning of Statistical Significance

It may be seen that "statistical significance" owes its conceptual meaning to the sampling distribution of the test statistic. The said theoretical distribution is based on the assumptions that (a) the research manipulation is substantively ineffective, and (b) random chance is the cause of variation in the score-values. Hence, to adopt the sampling distribution based on $H_0$ is to adopt chance influences as an explanation of the data. In its capacity as the logical complement of $H_0$, the conceptual meaning of $H_1$ is that chance influences may be ruled out as an explanation of the experimental outcome. This is less specific than saying that the experimental manipulation is efficacious because the significant result may be due to some confounding variables (see *Experimentation in Psychology--Rationale, Concepts, and Issues)*.

### 4.6. "$H_0$ Is Never True" Revisited

Critics find the significance test wanting because it is based on $H_0$, and critics assert that $H_0$ is never true in the substantive population. However, as has been shown in the introduction, statistics is about statistical populations, not any substantive population. The second difficulty with the critics' stance is that $H_0$ is not (and should not be) stated as a categorical proposition, as may be seen from the foregoing discussion. Instead, it appears in the two conditional propositions [P1] and [P2]:

If chance is the cause of data dispersion, then $H_0$ is true [P1]

If $H_0$ is true, then the sampling distribution of differences has a mean difference of zero [P2]

In view of the fact that $H_0$ is the consequent of [P1], the reservation about $H_0$ is actually a question about the antecedent of [P1]. Such a concern is about the data-collection procedure, not about $H_0$ per se. In sum, it is misleading to say "$H_0$ is never true."

## 5. Effect Size and Statistical Power

Having a utilitarian goal in mind (see *Experimentation in Psychology--Rationale, Concepts, and Issues)*, critics find significance tests wanting. For example, they note that the difference between subjects' partial-report and whole-report performance is used to obtain the *t*-statistic, and the binary decision about statistical significance is made on the basis of the *t*-statistic. At the same time, statistical significance is not informative of the practical importance of the result. Worse still, statistical significance is ambiguous or misleading because of its anomalous relationship with the effect size, as may be seen from Table 8.

*Table 8. The putative anomalous relationship between statistical significance and effect size*

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Study | $u_E$ | $u_C$ | Effect size = $d = (\overline{X}_E - \overline{X}_C) \div s_E$ | *t* significant? | *df* |
| I | 8 | 2 | 0.5 | Yes | 22 |
| II | 17 | 8 | 0.9 | No | 5 |

| | | | | | |
|---|---|---|---|---|---|
| III | 25 | 24 | 0.1 | No | 8 |
| IV | 6 | 5 | 0.2 | Yes | 20 |

### 5.1. The Putative Anomaly of Significance Tests

A commonly used index of effect size is the result of dividing the difference between the two sample means by the standard deviation of the control sample (see column 4 of Table 8). Study IV is significant, whereas Study II is not. Yet, the effect size of Study II is larger than that of Study IV. Although both Studies I and IV are significant, the effect size of Study I is larger than that of Study IV. Similarly, the two non-significant studies have different effect sizes. Specifically, Study II has a larger effect size than Study III. The critics make the point that, in concentrating only on statistical significance, psychologists are losing the important information conveyed by the effect size. To the critics, non-significance results may lead psychologists to ignore important (i.e. large-effect) results. By the same token, significance results may lead psychologists to accept trivial (i.e. small-effect) results.

### 5.2. "Effect"—Statistical versus Substantive

A simple reflection will show that statistical statements are about data, not substantive issues. This may be seen from the fact that psychologists use the *t*-test in a way indifferent to the nature of the research manipulation. For example, the *t*-test is used to assess the difference between two sample means, be they drug D and the placebo or a new teaching method and the traditional teaching method or acoustically similar words and acoustically dissimilar words. That such a practice is legitimate means that the research manipulation or substantive issue and statistics belong to different domains.

In a similar vein, "effect" is simply the difference between two means in cases where the *t*-test is used (e.g. $\overline{X}_C - \overline{X}_E$). Its numerical magnitude says nothing about its being trivial or important in the substantive domain. The statistical decision about the effect of the research manipulation is one about $(\overline{X}_C - \overline{X}_E)$ as a numerical difference, not as the product of a substantive causal agent. In other words, the anomalies suggested by critics of significance tests are more apparent than real.

### 5.3. Statistical Power

Table 8 also suggests to critics that sample size is responsible for the ambiguity in significance tests. The critics' argument is that too small a sample size will produce non-significant results despite a large effect size. At the same time, statistical significance is assured even though the effect size is trivial if a large enough sample size is used. The ambiguity is eliminated if psychologists know the probability of obtaining statistical significance. Statistical power is considered such an index.

It is necessary to refer to Table 7 again to present the power-analytic argument. The two panels of Table 7 would be saying the same thing if $H_0$ and $H_1$ are mutually exclusive, as is the case when they are identified with "chance" and "not chance" influences, respectively, as in *Section 4.5. The Meaning of Statistical Significance.* However, $H_1$ and $H_0$ are not mutually exclusive in the power-analytic account. To power analysts, there are as many $H_1$ as there are possible differences between two sample means expressed in units of the standard deviation of the control sample. Hence, a specific numerical value is assumed in "$H_1$ is true" in panel 2 of Table 7. Consequently, the statistical hypothesis testing is represented graphically with two distributions by power analysts, one based on $H_0$ and the other on $H_1$. Once, the decision is made about the desired statistical power, desired effect size, and the level of significance the appropriate sample size may be obtained by consulting the tables prepared for that specific purpose.

### 5.4. Reservations about Statistical Power

Before accepting the scenario in panel 2 of Table 7, it is necessary to settle a few important questions. The first is about the level of abstraction. As may be seen from column 4 of Table 8, the effect size is defined at the level of scores. Consequently, the two distributions envisaged in panel 2 of Table 7 are distributions of population scores. However, the statistical decision is made with reference to the sampling distribution of differences, not on the basis of the substantive population. Hence, the $\alpha$ level envisaged in panel 2 of Table 7 is not (and cannot be) that depicted in panel 1 of the same table.

Second, the power of the test is said to be the probability of obtaining statistical significance. One gets the impression that statistical power is an exact probability about $H_0$. This state of affairs may be responsible for the ready and uncritical acceptance of the power-analytic argument. However, as $\beta$ is conditional probability, so should be statistical power $(1 - \beta)$. How can statistical power be the exact probability of obtaining statistical significance?

The power-analytic stance also owes its third difficulty to the conditional nature of the Type II error. Recall that the probability of the Type I error is an index of the strictness of the statistical decision. That is, the $\alpha$ level says nothing about the data, but the researchers' decision about the data. By the same token, specifying the Type II error is to specify something about the decision makers, not the data. Consequently, being the one's complement of the probability of the Type II error, statistical power is, at best, an index of some aspects of decision making (e.g. the researchers' willingness or reluctance to choose $H_0$ in the face of uncertainty). It cannot be about data; nor can it be about $H_0$.

### Glossary

**Alternative hypothesis ($H_1$):** The statistical hypothesis implied by the hypothesis that the research data are not the result of chance influences.

**Associated probability:** The probability of obtaining a score as extreme as, or more extreme than, *X*.

**Conditional probability:** The probability of obtaining *X*, on the condition that another event occurs.

**Confidence interval:** The interval that would include the population parameter with the specified long-run probability.

**Correlation:** The relationship between two variables.

**Data dispersion:** The variability among the scores.

**Degrees of freedom:** The number of scores in a sample that are free to vary when the deviation scores are calculated with reference to the sample mean.

**Deviation score:** The difference between *X* and the sample mean.

**Level of significance:** The long-run probability that the decision to exclude chance influences as an explanation of the data is wrong.

**Long-run probability:** The probability based on the theoretical exercise of carrying out the random sampling procedure an infinite number of times.

**Mean:** The point of balance or the center of gravity of the scores.

**Null hypothesis ($H_0$):** The statistical hypothesis implied by the hypothesis that the research data are the result of chance influences.

**Parameter:** A summary of characteristics of a population.

**Population:** The entire collection of units about which the research is carried out.

**Random sampling:** The procedure used to select samples such that all possible samples of the same size have an equal chance of being selected.

**Random sampling distribution:** The theoretical distribution of all possible values of a sample statistic based on an infinite number of random samples of the same size.

**Random sampling distribution of means:** The theoretical distribution of the means of an infinite number of random samples of the same size.

**Random sampling distribution of the difference between two means:** The theoretical distribution of the differences between an infinite number of pairs of means from random samples of size $n_1$ and $n_2$.

**Regression:** The functional relationship between a manipulated and a random variable.

**Sample:** A subset of a population.

**Sampling:** The selection of a sample with reference to a well-defined rule.

**Standard deviation:** The square root of the mean-squared deviation scores.

**Statistic:** A summary characteristic of a sample.

**Statistical hypothesis:** An hypothesis about the parameters of the methodologically defined statistical populations.

**Statistical power:** The one's complement of the probability of committing the Type II error.

**Statistical significance:** The decision that chance influences can be excluded as an explanation of the data at the level of significance specified.

**Test statistic:** A statistic derived from a sample statistic for the purpose of testing a hypothesis about the population parameter (e.g. *t* of *F*).

**Type I error:** The error committed in rejecting $H_0$, given that $H_0$ is true.

**Type II error:** The error committed in not rejecting $H_0$, given that $H_0$ is false.

**Bibliography**


Chow S.L. (1998). A précis of "Statistical Significance: Rationale, Validity and Utility." *Behavioral and Brain Sciences* **21**, 169–194. [This open-peer review consists of a précis of Chow's (1996) *Statistical Significance: Rationale, Validity and Utility*; 39 commentaries on the book, and Chow's reply to the commentaries. The set is a good place to review the arguments for and against significance tests.]

Cohen J. (1987). *Statistical Power Analysis for the Behavioral Sciences*, rev. edn. New York: Academic Press. [This book provides much of the impetus for the adoption of power analysis.]

Meehl P.E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science* **34**, 103–115. [This is one of the first articles that remind psychologists not to conflate the statistical hypothesis with the substantive hypothesis. By implication, readers will learn that making the statistical hypothesis is not corroborating the substantive hypothesis.]

Siegel S. (1956). *Nonparametric Statistics for the Behavioral Sciences*, 312 pp. New York: McGraw-Hill. [This book is a classic introduction to nonparametric procedures.]

Wilkinson L. and Task Force on Statistical Inference, APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist* **54**(8), 594–604. [This report is an advocacy document, in which a set of research agenda is set forth. Despite the Task Force's disavowal, much of what the report says is about research methods in general, and psychologists' putative departure from good research practice in particular. The report sets in high relief the importance of the philosophical and methodological issues.]

Winer B.J. (1962). *Statistical Principles in Experimental Design*, 672 pp. New York: McGraw-Hill. [This book is a classic introduction to analysis of variance.]