Available online at http://www.ijims.com
ISSN: 2348 – 0343

Role of Statistics in Multidisciplinary Research – With Reference to Psychological Research

Sharma Hemant
Amity Business School, Amity University Haryana, Gurgaon, India

Abstract

Statistics is the study of the collection, organization, analysis, interpretation and presentation of data. It deals with all aspects of data. It is usually noticed that some routine words are given technical meanings in statistical parlance (e.g. "mean," "normal," "significance," "effect," and "power"). It is essential to resist the temptation of conflating their 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. As, statistics is used (i) to describe 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" caution 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 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. The present paper discusses the critical issues of statistics in psychological research.

Key words: probability, descriptive statistics, inferential statistics, random sampling distribution, statistical power , statistical significance.

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 (i) to ascertain the relationship between two or more variables, (ii) to derive the test statistic, or (iii) to extract factors. As multivariate statistics is introduced in The Construction and Use of Psychological Tests and Measures, this article is almost exclusively about univariatestatistics.

Before proceeding, there is a need of making a distinction between the substantive population and the statistical population. Suppose that an experiment is carried out to study the effects of specialized coaching on the performance of students. The substantive population consists of all students.The sample selected from the substantive population is divided into two sub-samples. The experimental sub-sample receives the specialized coaching and the control sub-sample receives an ordinary coaching. In this experimental context, the two groups are not samples of the substantive population, "all students." Instead, they are samples of two statistical populations defined by the experimental manipulation "students given specialized

coaching" and "students given ordinary coaching." 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 very important implications such as (i) statistics deal with methodologically defined statistical populations (ii) statistical conclusions are about data in their capacity to represent the statistical populations, not about substantive issues (iii) apart from very exceptional cases, research data (however numerous) are treated as sample data and (iv) testing the statistical hypothesis is not verifying the substantive theory. 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.

2. Descriptive Statistics

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) 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.

2.1 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.

2.2 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".

Table 1: Various ways of tabulating the data

Panel 1: Enumeration of all 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 fromthe 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 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.

2.3.1Utilities 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. 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). Thus, psychologists answer different question using different types of probability distribution. 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 Brief 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 (Figure1a) 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.

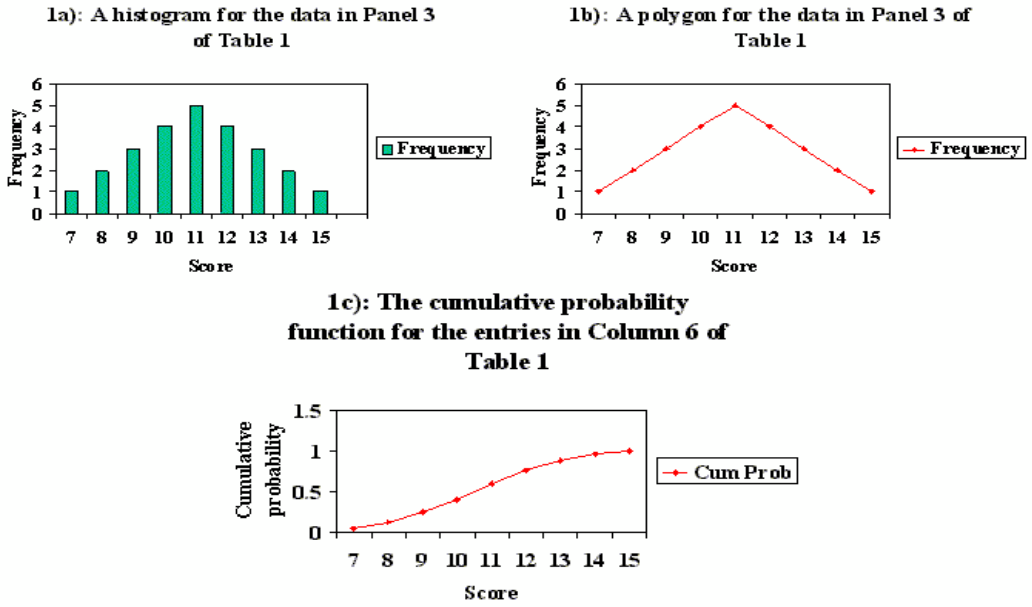

Figure 1. Graphical representations of the simple frequency (a & b) and cumulative probability distributions (c)

2.4.2 Measures of Central Tendency and Dispersion

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 thatoccurs 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. The mean is the center of gravity (or the balance point) of the aggregate may also be seen from panel 1 of Table 2.

Table 2:An illustration of the deviation score, sum of squares, variance, and standarddeviation of a set of scores

Panel 1: The deviation score

Score (X)	Deviation score	Deviation score times frequency	Score (X)	Deviation score	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 sumof the deviation scores =		∑= **–13**	The sumof the deviation scores =		∑= **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
Σ=	126	0	sumof squares = 94.00
s^2=			94 ÷ 8 = 11.75
s =			√(11.75) = 3.43

"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 is square 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).

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 one is 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 X. Hence, the parameter (n – 1) is the degrees of freedom associated with X.

2.4.4 Standardization

It is very difficult to compare the cost of electricity between two states when they have different costs of living. One solution is to express the cost of the electricity in terms of a common unit of measure, a process called "standardization." For example, we may quote the electricity's costs in the two states 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 – mean) and (X_B – mean) by their respective standard deviations (σ_A and σ_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 that a population parameter would lie within two specified limits (the confidence- interval estimate).

3.0 Correlation and Regression

Another important 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 is controlled by the researcher.

3.1 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 associated 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 relationship 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

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. The scatter plot assumes the form of a circle, which is indicative of no relationship between the two variables.

3.2 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.

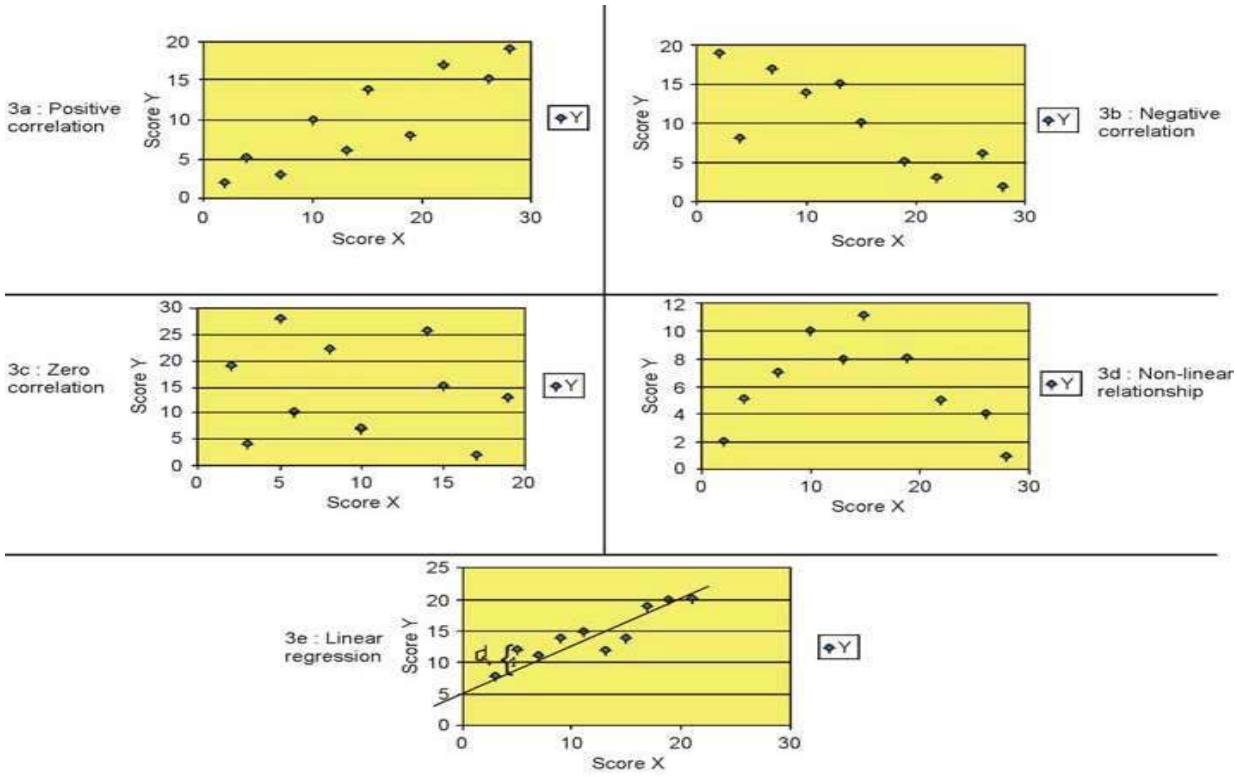

Figure 3. Graphical representation of some relationships between two variables

3.3 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 di 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 di, which is squared. The sum of the 10 $(di)^2$ 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 $(di)^2$ as there are data points. Moreover, each line gives rise to its own set of $(di)^2$. 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):

$Y = a + bX$(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.

4. 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 is practiced.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

4.5.1 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 practically ineffective, and (b) random chance is thecauseof 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 experimentaloutcome. Thisis less specific than saying that the experimental manipulation is effective because the significant result may be due to some confounding variables (see Experimentation in Psychology--Rationale, Concepts, and Issues).

4.5.2 Statistical Power

Ambiguity in significance tests is mainly because of the sample size. 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 small if a large enough sample size is used. The ambiguity is eliminated if psychologists know the probability of obtaining statistical significance. Power of the test is said to be the probability of obtaining statistical significance.Statistical power is considered such an indexof some aspects of decision making (e.g. the researchers' willingness or reluctance to choose H_0 in the face of uncertainty).

5.0 Conclusion

It is evident from the above discussion that statistics is the study of the collection, organization, analysis, interpretation and presentation of data. It is used (i) to describe 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. From the sample statistics, theoretical distributions: the distributions of various test statistics are derived which become the base of statistical decisions.Mostly, the calculated test statistic is compared to the theoretical distribution that is actually a cluster of an infinite number of tokens of the test statistic in question. Therefore, in every probabilistic statement based on inferential statistics, there is a need of caution in long run. (e.g. "the result is significant at the 0.05 level in the long run"). Apart from this, significance tests should also be conducted which are very helpful in clarifying some concepts, examining the role of statistics in empirical research and 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.

References

Chow SL. A précis of "Statistical Significance: Rationale, Validity and Utility."Behavioral and Brain Sciences, 1998;21: 169–194.

Cohen J. Statistical Power Analysis for the Behavioral Sciences, rev.edn. New York: Academic Press; 1987.

Meehl PE. Theory testing in psychology and physics: A methodological paradox. Philosophy of Science,1967; 34:103–115.

Siegel S. Nonparametric Statistics for the Behavioral Sciences, New York: McGraw-Hill; 1956.

Wilkinson L. Statistical Inference, APA Board of Scientific Affairs Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 1999; 54(8): 594–604.

Winer BJ. Statistical Principles in Experimental Design, New York: McGraw-Hill; 1962.