Naab, T. K. (2017). Statistics, Descriptive. In J. Matthes, C. Davis, C. & R. Potter (Hrsg.), *International Encyclopedia of Communication Research Methods.* Hoboken, NJ: Wiley. doi: 10.1002/9781118901731.iecrm0241

Title
Statistics, Descriptive

Your Name Teresa K. Naab
Affiliation University of Augsburg, Germany
Email Address Teresa.naab@phil.uni-augsburg.de

REPEAT FOR ANY CO-AUTHORS

Word Count
2195 words

Abstract

Descriptive statistics are procedures to summarize and present a set of observations or a set of data. This is achieved by displaying frequency distributions in tables and in graphical format (e. g., bar chart, histogram, polygon), by computing measures of central tendency (e. g., mode, median, mean), and measures of variability (e. g., range, mean absolute deviation around the mean, variance, standard deviation). The selection of appropriate measures depends on the measurement scale of the variables, which can be nominal, ordinal, interval, or ratio.

Main Text

Descriptive statistics are procedures to summarize and present a set of observations or a set of data. This is achieved by using tables, graphs, and various measures which provide descriptions of the observations. Descriptive statistics provide the basis for any further procedures of data analysis. Inferential statistics are procedures used to draw conclusions from sample observations to parameters of the population from which the sample is taken (see IECRM0242).

Measurement scales

In the process of measurement numbers are assigned to characteristics of objects according to a definite rule. Measurements can be classified into a hierarchy of measurement scales based on the relationship that exists between the objects having different score values. The measurement scale of a variable has important implications for the procedures of descriptive statistics that can be applied. The hierarchy includes the nominal scale, the ordinal scale, the interval scale, and the ratio scale. Nominal and ordinal scales are often called categorical and interval and ratio scales are often called metric (see also IECRM0199). The lowest scale in the measurement hierarchy is the *nominal scale*. Objects are classified into categories on the basis of some defined characteristic. Nominal data have two properties: 1) Data categories are mutually exclusive, i. e. an observation can belong only to one category. 2) Data categories have no logical order. Therefore any numerical designation to the categories is arbitrary and has no quantitative meaning (an example of a nominal variable would be to ask study participants about the brand of their cell phone). At the next level of measurement, the *ordinal scale*, data have an additional property: 3) Ordinal data categories are scaled according to the amount they possess of the characteristic being considered. That is, objects that possess a greater amount of a characteristic are assigned a greater number, which allows for ordering of the categories (e. g., ranking of most favorite radio songs). The third measurement scale is called *interval scale*, with an additional property that distinguishes it from the aforementioned measurement scales: 4) Equal differences in the characteristics are represented by equal differences on the scale. That is, an equal unit is established in the scale; equal differences between any two points on the scale are the same regardless of their position on the scale (e. g., temperature in degrees Fahrenheit). The highest level of measurement is the *ratio scale*: 5) On a ratio scale the point zero reflects an absence of the characteristic. While on an interval scale the point zero does not reflect the starting point of the scale, ratio data have a true zero. This enables to make statements about proportional amounts of the characteristic possessed by two objects (e.g., television usage in minutes per day). Nominal scales enable to depict frequency distributions, e.g. in frequency tabulations of a single variable or cross-tabulations of two or more variables. Ordinal data allow for statements that rely on logical ordering of the score values or class intervals. Interval data are amenable to mathematical manipulations of addition and subtraction. Ratio data additionally enable for multiplication and division. Statistical procedures that can be used at a low level of the measurement hierarchy can also be used to describe the distribution of variables at any higher level of the measurement hierarchy (Hinkle, Wiersma, & Jurs, 1982).

Frequency distribution

An initial way to summarize data is to show their frequency distribution (Hinkle et al., 1982). Frequency distributions indicate the number of times each score value or class interval (the width of a category as initially measured or later created by combining score values) occurs in the data. Frequency distributions can be portrayed as frequency tables or in graphical format (e. g., bar chart, histogram, polygon). A *frequency table* lists the absolute number of observations or the percentage of observations of each score value or class interval (relative frequency distribution) or the relative number of observations that lie above or below a particular score value or class interval (cumulative relative frequency distribution). A *bar chart* shows the number of observations or the percentage of observations of each score value or class interval. The horizontal axis (X-axis; abscissa) shows the range of scores of the variable. Each score value or each class interval receives an individual bar of identical width. The height of the bar on the vertical axis (Y-axis, ordinate) represents the frequency, relative frequency, or cumulative frequency. A *histogram* also represents the frequency, relative frequency, or cumulative frequency of score values or class intervals similar to a bar chart. Whereas in a bar chart the horizontal scale is discrete, the horizontal scale of a histogram is continuous. Usually, the bars in a histogram are of constant width representing equal intervals on the continuous scale. Interpretation of the frequency of observations in each interval depicted in the histogram depends on the area of the bar. Thus in a histogram the width of the bar is crucial. In a *frequency polygon* the height of a point on the vertical axis represents the frequency of a score value or a class interval. A point is plotted for each single score value or a point for each class interval (at the midpoint of the interval on the horizontal axis). The points are connected with straight lines. When a continuous variable is measured with very fine gradation and with many observations the polygon approaches the shape of a smooth frequency curve.

After a first insight into the distribution of a variable through tabulated and graphic frequencies, three characteristics are necessary for adequately describing a distribution: the shape of the distribution, the central tendency, and the variation of the scores. Several measures enable statements about these characteristics and allow for comparing different distributions.

Shapes of distribution

Shapes of distributions are usually described by referring to their skewness and kurtosis (Hinkle et al., 1982). *Skewness* is an indicator of the extent to which a distribution is asymmetric. A positively skewed distribution (skewed to the right) has more scores located toward the lower end of the scale of measurement and fewer scores toward the upper end; its tail extends more to the right than the left. A negatively skewed distribution (skewed to the left) has more scores located toward the upper end of the scale of measurement and fewer scores toward the lower end; its tail extends more to the left than the right. A distribution with a skewness of zero is symmetric. The *kurtosis* is an indicator of the heaviness of the tails of a distribution. When many scores are located at the center of the distribution and the distribution has heavier tails, it is called leptokurtic (high degree of peakedness, positive kurtosis). When scores still cluster at the center of the distribution, but the distribution is more uniformly shaped and has light tails, the distribution is called platykurtic (low degree of peakedness, negative kurtosis). Distributions of moderate degree of peakedness are called mesokurtic.

Measures of central tendency

Measures of central tendency (also measures of location) reflect the location of the distribution on the scale of measurement (Clayton, 1984). They indicate where the concentration of the scores is located. Commonly used measures of central tendency are mode, median, and mean. The *mode* is the most frequent score value in a distribution. When two or more scores have the same frequency and this frequency is greater than for any other scores, these distributions are referred to as multimodal. The mode is determined by inspection of the frequency table rather than by computation. It can be used with all measurement scale levels starting from nominal level and is not affected by extreme scores values (outliners). The *median* is the point on the measurement scale below which 50% of the score values in a distribution lie. Since depicting the median requires arranging the scores in a logical order, it can be used with all measurement scale levels starting from ordinal level. For an odd number of scores, the middle score is the median. For an even number of scores, the average of the two middle scores is the median. The median is less affected by extreme scores values and skewed data than the mean (see also IECRM0148). The median is the 50th percentile. Aside from the median, further Xth percentiles can be reported, which indicate the score value below which X percent of the scores lie. The *mean* is the arithmetic average of the scores in a distribution. It is the most frequently used measure of central tendency. It has two important properties: 1) The sum of all deviations around the mean (i. e. the difference between a given score and the mean of the distribution) is equal to zero. 2) The sum of squared deviations around the mean is smaller than the sum of squared deviations around any other value. Since indicating the mean requires the mentioned computation, it can be used with interval and ratio data only. In contrast to the mean, mode and median are relatively unaffected by changes in individual score values in a set of observations. Additionally, the mean is relatively sensitive to extreme score values. In symmetric, unimodal distributions mode, median and mean coincide.

In symmetric, bimodal distributions median and mean coincide. In asymmetric, negatively skewed distributions the mean is less than the median, which is less than the mode. In asymmetric, positively skewed distributions the mean is greater than the median, which is greater than the mode.

Measures of variability

Measures of variability (also measures of dispersion) reflect the dispersion of the distribution of scores on the measurement scale (Clayton, 1984). They indicate the degree to which the scores are dissimilar. Even distributions that are similar in shape and share similar central tendencies can be entirely different depending on their extent of dispersion. This is reflected in measures of variability. If the scores are compactly distributed around the central tendency, the distribution is called homogeneous (i. e. every case has a similar score value). If the scores are more widely dispersed throughout the scale of measurement, it is called heterogeneous. Commonly used measures of variability are range, mean absolute deviation around the mean, variance, and standard deviation. The *range* is the scaled distance between the highest and the lowest score. It tends to increase with the size of a set of observations. The range is strongly affected by extreme score values. When a set of observations contains exceptionally high or low outlines, the range is not typical of the variability within the set of data. The *mean absolute deviation around the mean* is the average deviation of all scores (in terms of absolute values) from the mean of scores in a distribution. The mean deviation can be used to compare distributions; distributions with greater mean deviation have a greater dispersion. The *variance* (also mean squared deviation) is the sum of the squared deviations around the mean of scores in a distribution divided by the total number of scores, i. e. it is the mean of the squared deviations. Since the deviations of scores from the mean are squared, the variance gives more weight to extreme scores that deviate from the mean by more than one unit on the measurement scale. Outliners increase the variance of a set of data (see also IECRM0006). The *standard deviation* is the positive square root of the variance. Thus the standard deviation has the same unit of measurement as the scores in the original set of data. The standard deviation is often used to report the percentage of scores in a distribution that are X standard deviations below and above the mean (see below for an example).

Normal distribution

A distribution often referred to is the normal distribution (also Gaussian distribution; Hays & Winkler, 1970). The normal distribution is a family of distributions sharing similar properties: 1) A normal curve is unimodal, symmetrical, and bell-shaped. Its mean, median, and mode coincide. Its skewness and kurtosis equal 0. 2) It is continuous. 3) It is asymptotic to the X-axis when graphed. Each curve of this family is determined by its mean and standard deviation. If score values have a normal distribution, the range between 1 standard deviation above the mean and 1 standard deviation below the mean contains 68.2% of all the scores. Likewise, 95.4% and 99.7% of all the scores lie within 2 and 3 standard deviation from the mean, respectively. For example, if the mean of a set of a data is 10 and the standard deviation is 2, 68.2% of the cases fall in the range between 8 and 12.

SEE ALSO:

IECRM0006

IECRM0148

IECRM0199

IECRM0242

References

Clayton, K. N. (1984). *An introduction to statistics for psychology and education.* Columbus, OH: Bell & Howell.

Hays, W. L., & Winkler, R. L. (1970). *Statistics. Probability, inference, and decision* (Vol. 1). New York, NY: Holt, Rinehart and Winston.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1982). *Basic behavioral statistics.* Boston, MA: Houghton Mifflin.

Further Readings

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2*, 292-307. doi: 10.1037/1082-989X.2.3.292

Hopkins, K. D., & Weeks, D. L. (1990). Tests for normality and measures of skewness and kurtosis: Their place in research reporting. *Educational and Psychological Measurement, 50,* 717-729. doi: 10.1177/0013164490504001

Author Mini-Biography:

Dr. Teresa K. Naab received her PhD at the Hanover University of Music, Drama and Media, Germany. She currently is post doc researcher at the University of Augsburg, Germany. Her research interests include audience and media effects studies and empirical methods of social science.

Keywords:

distribution; frequency; bar chart; histogram; polygon; nominal scale; ordinal scale; interval scale; ratio scale; measures of central tendency; measures of variability; mode; median; mean;

range; mean absolute deviation around the mean; variance; standard deviation; normal distribution