Unit 3: Descriptive statistics (2)

Numerical measures

A variety of numerical measures are used to summarize data. The proportion, or percentage, of data values in each category is the primary numerical measure for qualitative data. The mean, median, mode, percentiles, range, variance, and standard deviation are the most commonly used numerical measures for quantitative data. The mean, often called the average, is computed by adding all the data values for a variable and dividing the sum by the number of data values. The mean is a measure of the central location for the data. The median is another measure of central location that, unlike the mean, is not affected by extremely large or extremely small data values. When determining the median, the data values are first ranked in order from the smallest value to the largest value. If there is an odd number of data values, the median is the middle value; if there is an even number of data values, the median is the average of the two middle values. The third measure of central tendency is the mode, the data value that occurs with greatest frequency.

Percentiles provide an indication of how the data values are spread over the interval from the smallest value to the largest value. Approximately p percent of the data values fall below the pth percentile, and roughly $100-p$ percent of the data values are above the pth percentile. Percentiles are reported, for example, on most standardized tests. Quartiles divide the data values into four parts; the first quartile is the 25th percentile, the second quartile is the 50th percentile (also the median), and the third quartile is the 75th percentile.

The range, the difference between the largest value and the smallest value, is the simplest measure of variability in the data. The range is determined by only the two extreme data values. The variance (s^2) and the standard deviation (s), on the other hand, are measures of variability that are based on all the data and are more

commonly used. Equation 1 shows the formula for computing the variance of a sample consisting of n items. In applying equation 1, the deviation (difference) of each data value from the sample mean is computed and squared. The squared deviations are then summed and divided by n − 1 to provide the sample variance.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \qquad (1)$$

The standard deviation is the square root of the variance. Because the unit of measure for the standard deviation is the same as the unit of measure for the data, many individuals prefer to use the standard deviation as the descriptive measure of variability.

Outliers

Sometimes data for a variable will include one or more values that appear unusually large or small and out of place when compared with the other data values. These values are known as outliers and often have been erroneously included in the data set. Experienced statisticians take steps to identify outliers and then review each one carefully for accuracy and the appropriateness of its inclusion in the data set. If an error has been made, corrective action, such as rejecting the data value in question, can be taken. The mean and standard deviation are used to identify outliers. A z-score can be computed for each data value. With x representing the data value, $\bar{x}$ the sample mean, and s the sample standard deviation, the z-score is given by $z = (x - \bar{x})/s$. The z-score represents the relative position of the data value by indicating the number of standard deviations it is from the mean. A rule of thumb is that any value with a z-score less than −3 or greater than +3 should be considered an outlier.

Exploratory data analysis

Exploratory data analysis provides a variety of tools for quickly summarizing and gaining insight about a set of data. Two such methods are the five-number summary

and the box plot. A five-number summary simply consists of the smallest data value, the first quartile, the median, the third quartile, and the largest data value. A box plot is a graphical device based on a five-number summary. A rectangle (i.e., the box) is drawn with the ends of the rectangle located at the first and third quartiles. The rectangle represents the middle 50 percent of the data. A vertical line is drawn in the rectangle to locate the median. Finally lines, called whiskers, extend from one end of the rectangle to the smallest data value and from the other end of the rectangle to the largest data value. If outliers are present, the whiskers generally extend only to the smallest and largest data values that are not outliers. Dots, or asterisks, are then placed outside the whiskers to denote the presence of outliers.

Comprehension Exercises

Choose a, b, c or d which best completes each item.

1) The mean, median and mode are measures of _____.

a) variability b) shape c) central location d) analysis

2) Approximately *75* percent of the data values fall below the _____ and roughly 25 percent of the data values are above it.

a) median b) first quartile c) second quartile d) third quartile

3) Many individuals prefer to use the _____ as the descriptive measure of variability.

a) range b) standard deviation c) variance d) 95th percentile

4) In a box plot, _____ extend from one end of the rectangle to the smallest data value and from the other end of the rectangle to the largest data value..

a) vertical lines b) dots c) whiskers d) asterisks

Words to Learn

Find the Persian equivalents of the following terms and expressions.

mean	average	central	location
median	outliers	rank	position
mode	error	smallest	place
percentile	erroneous	largest	scale
quartile	erroneously	unit of measure	middle
percentile	equation	score	divide
range	formula	extreme	difference
variance	square	square root	deviation
standard deviation	standardized	accuracy	compute
variability	unusual	appropriateness	apply
exploratory	analysis	rectangle	identify
box plot	whiskers	asterisk	extend
percent	outside	dot	presence