**An Overview of Descriptive Data Analysis**

Statistical analysis of descriptive data is conducted to provide a summary of data in published research reports. These data serve as the starting point for the reader to begin making decisions regarding the strength and applicability of the research as evidence for practice in specific populations. Descriptive data are derived from a data set to represent research variables for the purpose of summarizing information about the sample and do not involve generalization to a larger set of data such as the population. For example, a data set could contain variables including ages of participants, years of experience as a nurse, and scores on a job satisfaction instrument. Simply put, descriptive statistics use numbers narratively, in tables, or in graphic displays to organize and describe the characteristics of a sample ([**Polit, 2010**](https://jigsaw.vitalsource.com/books/9781284138887/epub/EPUB/xhtml/23_Chapter11.xhtml#ch11_bib10)).

Descriptive statistics involve a range of complexity, from simple counts of data to analyses of relationships. The most commonly encountered descriptive statistics are classified in the following ways:

* Counts of data, expressed as frequencies, frequency tables, and frequency distributions
* Measures of central tendency, expressed as the mean, median, and mode
* Measures of variability, including the range, variance, and standard deviation
* Measures of position, such as percentile ranks and standardized scores
* Measures of relationship such as correlation coefficients
* Graphical presentations in bar charts, line graphs, and scatter plots

The specific types of summary statistics that are applied to descriptive data are driven by the nature of the data and the level of measurement of the variable. It is important to apply as many descriptive techniques as necessary to provide an accurate picture of the variable, rather than giving only a single measure that may mislead a reader. For example, providing a percentage of 100% hand-washing compliance for a month may look very good, but if the frequency count is omitted—especially if that count is *n* = 1—a complete picture is not provided.

**Understanding Levels of Measurement**

The initial, and perhaps most vital, step in descriptive analysis is to identify the level of measurement for each variable so as to choose the appropriate statistical analysis. This decision is the responsibility of the researchers who create descriptive studies and is an important point for critique by nurses who read research reports. Data can be collected in one of four possible levels of measurement: nominal, ordinal, interval, or ratio. Each level has characteristics that make it unique, and each requires a particular type of statistical technique.[**Table 11.1**](https://jigsaw.vitalsource.com/books/9781284138887/epub/EPUB/xhtml/23_Chapter11.xhtml#ch11_tbl1) shows descriptive statistical techniques that are appropriate for each level of measurement.

Nominal-level data are those that denote categories and have no rank order; numbers given to these data are strictly for showing membership in a category and are not subject to mathematical calculations. Nominal data can be counted, but are not measured, so they can be summarized using statistics that represent counts. Fall precautions is an example of nominal data: Either a patient is on fall precautions or the patient is not. Summary statistics appropriate for this level of measurement are frequency, percentage, rates, ratios, and mode.

**Table 11.1** Levels of Measurement and Appropriate Summary Statistics

| **Level of Measurement** | **Distributions** | **Central Tendency Mode** | **Variability** | **Shape** |
| --- | --- | --- | --- | --- |
| Nominal | Frequency Percentage | Mode |  |  |
| Ordinal | Frequency Percentage | Mode Median | Range Minimum/maximum Range |  |
| Interval/ratio Mode |  | Mode Median Mean | Range Minimum/maximum Standard deviation Variance | Skew Kurtosis |

Ordinal data are also categories but have an added characteristic of rank order. These data differ from nominal data in that the categories for a variable can be identified as being less than or greater than one another. However, because the level of measure is still categorical, the exact level of difference cannot be identified. A pain scale, in spite of its representation as a series of numbers, is an example of ordinal data. For example, while we know a score of 7 on this scale is greater than a 5, the difference cannot be quantified. For example, we cannot conclude that the difference between a 7 and a 5 on the pain scale is the same as the difference between a 4 and a 2 on the same scale; we simply know that one number is higher than the other. Further, we do not know that a score of 5 for one patient is the same as a score of 5 for another patient. The fact that the pain scale is represented by numbers also does not necessarily mean the characteristic under study has been quantified. In our example, we do not know exactly how much pain exists; it is just rated against other experiences of pain for that patient. The patient’s score has simply been ranked against all other values of the variable. Statistical techniques appropriate for ordinal data include those appropriate for nominal data plus range, median, minimum, and maximum.

Interval and ratio data are recorded on a continuous scale that has equal intervals between all entries; length of stay is an example. Data collected on interval or ratio levels result in numbers that can be subjected to many mathematical procedures, including mean, standard deviation, variance, and evaluation of the distribution (skew and kurtosis).

**Identifying Shape and Distribution**

Initial analyses of data are meant to help the researcher identify the distribution, and therefore the shape, of the variable’s data. The outcome of this analysis, coupled with the level of measurement, guides the researcher in selecting the appropriate statistics to represent the variable’s center and spread.

**Summarizing Data Using Frequencies**

**Frequency** is a statistical term that means a count of the instances in which a number or category occurs in a data set. Frequencies are commonly used in clinical settings; for example, a frequency might be used to document the number of infections by surgery type, the number of patient falls by nursing unit, or the number of nurses who leave in the first 18 months of employment. In research, frequencies are used to count the number of times that a variable has a particular value or score. A researcher may collect data on nominal-level variables or ordinal-level variables and then generate a frequency count per category to summarize the data. For example, if information about gender were desired, two values (male and female) would be collected; the number of participants in the study who were male and the number who were female would be tallied.

**Frequency:** A count of the instances that an event occurs in a data set.

Frequency data can also be used to calculate percentages, rates, and derived variables. A percentage is a useful summary technique that shows the relative frequency of a variable. For example, if gender was measured as a variable and there were 180 female participants in a sample of 400, the percentage of female participants would be 45%. The number 45% is more meaningful as a summary value than the frequency count because readers can tell quickly that slightly less than half of the sample was female. To calculate the percentage in this example, the number of female participants is divided by the number of participants in the entire sample (180/400 = 0.45 or 45%).

Rates that are clinically important can be calculated to provide information about data trends over time. Like a percentage, a rate is calculated by dividing the frequency of an event in a given time period by all possible occurrences of the event during the same time period. The difference is that percentages by definition are “per 100,” whereas rates can have a different denominator, such as per 1000 patient-days. Monthly fall rates are an example; the number of falls in a month is divided by the total number of patient-days in that month and then multiplied by 1000 to give the number of falls per 1000 patient-days. This allows for comparison between units based on opportunities for falls (each day) instead of the raw count (number of falls). When calculated periodically, patient outcomes expressed as rates can be monitored as a basis for action planning to improve care ([**Altman, 2006**](https://jigsaw.vitalsource.com/books/9781284138887/epub/EPUB/xhtml/23_Chapter11.xhtml#ch11_bib1)).

**Derived variables** are created when data from other variables are put into a simple formula to produce a new piece of information—a new variable. An example of a derived variable is the hospital length of stay. It is calculated by summing the number of inpatient days on a nursing unit (a frequency) in a month and then dividing this sum by the number of patients in the unit. Rates and derived variables may be used to represent the effects of extraneous variables or to describe baseline performance. Operational definitions of these variables are important to include to facilitate consistency in the method of calculation.

**Rate:** A calculated count derived from dividing the frequency of an event in a given time period by all possible occurrences of the event during the same time period.

**Summarizing Data Using Frequency Tables**

Interval-level data can be summarized in a frequency table as well, but they must be grouped into categories first, which requires converting the data into ordinal data. Because interval-level data can fall anywhere on a continuous scale, each data point could theoretically be a unique number; as a result, there can be many different values in the data set, which makes interpreting the data difficult. As an example, [**Table 11.2**](https://jigsaw.vitalsource.com/books/9781284138887/epub/EPUB/xhtml/23_Chapter11.xhtml#ch11_tbl2) contains a data set that could benefit from summary as a frequency table. To create a frequency table for interval-level data, the data are sorted from lowest value to highest value, categories are created for the data, and the number of occurrences in each category is counted. Collapsing several years together provides an even clearer picture. [**Table 11.3**](https://jigsaw.vitalsource.com/books/9781284138887/epub/EPUB/xhtml/23_Chapter11.xhtml#ch11_tbl3) represents the same data in a frequency table that includes both counts and relative frequency (percentage). Representing complex data in this way enhances interpretation and understanding of the values. Although no patterns were readily apparent in the data in the original table, review of the frequency table makes it clear that most participants in the study had few years of nursing experience; in fact, the majority of participants had nine or fewer years of experience.

**Table 11.2** Unordered Data Set of Years as a Nurse

| **Years as a Nurse** | | | | |
| --- | --- | --- | --- | --- |
| 13 | 5 | 23 | 5 | 3 |
| 7 | 5 | 30 | 20 | 5 |
| 8 | 6 | 17 | 21 | 11 |
| 7 | 6 | 27 | 5 | 18 |
| 2 | 1 | 9 | 9 | 16 |
| 2 | 1 | 3 | 5 | 9 |
| 1 | 5 | 30 | 1 | 28 |
| 2 | 2 | 16 | 3 | 12 |
| 5 | 5 | 20 | 15 | 17 |
| 23 | 20 | 17 | 9 | 24 |

**Table 11.3** Frequency Table for Years as a Nurse

| **Years as a Nurse** | **Frequency** | **Percentage** |
| --- | --- | --- |
| 1–3 | 11 | 22 |
| 4–9 | 18 | 36 |
| 10–17 | 9 | 18 |
| 18–30 | 12 | 24 |
| Total | 50 | 100 |

**Summarizing Data Using Frequency Distributions**

In many ways, graphical presentation of data is easier to understand because the data are presented in summary fashion with colors, lines, and shapes that show differences and similarities in the data set. An easy chart to create is a **bar chart** for nominal or ordinal data. The most common way to design a bar chart is to have categories of the variable on the *x*-axis (horizontal) of the chart and the frequency for each category on the *y*-axis (vertical). Bar charts provide the reader with a quick assessment of which category has the most occurrences in a data set. For example, a bar chart of patient safety occurrences (nominal data) can show which occurrences happen the most, thereby providing information on which areas should be the focus of improvement efforts.

**Bar chart:** A graphic presentation for nominal or ordinal data that represents the categories on the horizontal axis and frequency on the vertical axis.

A bar chart showing the frequency per category for ordinal data enables evaluation of the shape of the distribution of values. Such graphs are called frequency distributions or **histograms**. For example, the values of subjects’ pain ratings can be represented in a graph with the value (pain rating) placed on the *x*-axis and the number of cases with that pain rating on the *y*-axis. [**FIGURE 11.1**](https://jigsaw.vitalsource.com/books/9781284138887/epub/EPUB/xhtml/23_Chapter11.xhtml#ch11_fig1) depicts a histogram of pain rating data. It shows categories for values of the variable across the horizontal dimension, with frequencies being displayed on the vertical dimension. In this histogram, the reader can see that most of the values of the variable “postop pain” were between 4 and 6.

**Histogram:** A type of frequency distribution in which variables with different values are plotted as a graph on *x*-axes and *y*-axes, and the shape can be visualized.

**FIGURE 11.1** A Normal Distribution

Another useful feature of the histogram is that it shows the curve of the data. The researcher can see whether the data in the study are normal, are skewed, or have an abnormal kurtosis. Based on the distribution of the data, the researcher will then select the statistic for describing the center and spread of the data. Figure 11.1 depicted an approximately normal distribution. A normal distribution, often called a bell curve, has a large proportion of values of the variable in the middle of the distribution and smaller proportions on the ends (tails). The shape of the distribution is symmetrical, meaning that the right and left sides of the distribution are mirror images. The shape of the distribution of a variable in a research study is important because many statistical procedures require the assumption of a normal distribution to yield reliable results ([**Field, 2013**](https://jigsaw.vitalsource.com/books/9781284138887/epub/EPUB/xhtml/23_Chapter11.xhtml#ch11_bib3)).

Some variables are not symmetric, so they do not have a normal distribution. Asymmetric distributions may be described as demonstrating skew or kurtosis. A skewed distribution has a disproportionate number of occurrences in either the right or left tail of the distribution. Describing the type of distribution is counterintuitive: The skew is described by the direction of the tail. For example, distributions with more values in the positive end of the distribution trail out to the negative end and so are described as negative skew; those with more values on the lower end of the scale will trail out to the positive end and so are called positive skew.[**FIGURES 11.2**](https://jigsaw.vitalsource.com/books/9781284138887/epub/EPUB/xhtml/23_Chapter11.xhtml#ch11_fig2) and [**11.3**](https://jigsaw.vitalsource.com/books/9781284138887/epub/EPUB/xhtml/23_Chapter11.xhtml#ch11_fig3) demonstrate skewed distributions. Kurtosis refers to an unusual accumulation of values in some part of the distribution, particularly in the size of the tails. Those distributions with large tails are called leptokurtic; those with small tails are called platykurtic. A leptokurtic distribution is depicted by the histogram in [**FIGURE 11.4**](https://jigsaw.vitalsource.com/books/9781284138887/epub/EPUB/xhtml/23_Chapter11.xhtml#ch11_fig4).

**FIGURE 11.2** A Negatively Skewed Distribution

Researchers who create data have a responsibility to check the descriptive data for compliance with the underlying assumptions of the statistical techniques selected for use. These assumptions often include a requirement for variables to have a normal distribution (not skewed or kurtotic) and to have equal variances (similar spread of scores around the mean). A visual inspection of the graphical display of values of a variable via a histogram can give a researcher a quick indication of whether the variable meets the requirements of the statistical test.

**FIGURE 11.3** A Positively Skewed Distribution

**Describing the Center and Spread**

By first checking the shape of the distribution of data for a variable, the researcher becomes able to select the most appropriate statistics to describe the variable’s center and spread. Measures of central tendency are used to describe the variable’s center; they include the mean, median, and mode. The spread, or variability, of the variable’s data is calculated via standard deviation, range, and percentiles.

**Summarizing Data Using Measures of Central Tendency**

A measure of central tendency is a single number that summarizes values for a variable. These measures represent the way the data tend toward the center; they comprise a single number used to reflect what may be a typical response in the data set. For example, when researching care of a geriatric population, the ages of the participants in a sample will be an important variable to evaluate in the study. Instead of listing all the ages of participants, a researcher can summarize the data using a measure of central tendency such as the average (mean) age of participants in the sample. If, for example, the mean age were 25 rather than 75, the reader would know that participants were young adults rather than elderly participants. Other measures of central tendency include the median and the mode.

**FIGURE 11.4** A Distribution with Kurtosis

The **mean** is commonly called the average. This number is calculated by adding all scores in the data set and dividing the sum by the number of scores in the data set. Therefore, only data measured at the interval and ratio levels are appropriate for calculating a mean score. The mean is an easily recognized and interpreted measure of central tendency. It is familiar to most readers and easy to calculate. The mean, however, is disproportionately affected by extreme values. For example, the mean length of stay of joint replacements would be 3 days if five patients each stayed in the hospital for 3 days—but it would also be 3 days if four patients each stayed 2 days and one patient stayed 7 days. This sensitivity to extreme scores is a weakness of the mean; thus data that are normally distributed are most appropriate to submit to the calculation of a mean score.

**Mean:** The average; a measure of central tendency.

The **median** is another measure of central tendency, but it is not a calculation; it is a location. To find the median, the values of a data set are arranged in sequence from smallest to largest, and the center of the data set is determined by finding the exact midpoint of the data. Using the example from the previous paragraph, patients who underwent joint replacements and had hospital lengths of stay of 3, 3, 3, 3, and 7 days would have a median length of stay of 3. When there is an even number of values in the data set, the two most central values are averaged to obtain the median. The median can be used with normal, skewed, or kurtotic distributions because it is not influenced by extreme values in the data set. It can be used with ordinal, interval, or ratio data, but its usefulness in inferential statistical procedures is limited. The primary weakness of the median is that it represents only the middle of the data set and can be used in only a narrow range of statistical procedures.

**Median:** A measure of central tendency that is the exact midpoint of the numbers of the data set.

An additional measure of central tendency is the **mode**, the most frequently occurring value in the data set. Some data sets will not have a mode; others may have multiple modes. The mode is an easy statistic to determine, but it has limited usefulness. It is not used in inferential statistics and provides little information about a data set. It is, however, the only measure of central tendency that can be applied to nominal data.

**Mode:** A measure of central tendency that is the most frequently occurring value in the data set.

The mean is a dependable measure of the center of the distribution because it uses all numbers in the data set for its calculation; however, when extreme values exist in the data set, the mean is drawn toward that extreme score. Thus researchers should use the mean statistic with caution in distributions that are badly skewed, particularly those involving small sample sizes. The median is the middle point of the data set; it is not sensitive to extremes, making it a good choice as a measure of central tendency in distributions that are skewed. The mode helps the reader understand if any value occurs more frequently than others. The minimum and the maximum help assess the spread of the data. The minimum value is the smallest number in the data set, and the maximum value is the largest. Using the mean, median, minimum, and maximum together helps the reader of the data used in research obtain a fuller, more accurate picture of the study findings. How these numbers come together to characterize a data set is demonstrated in [**Table 11.4**](https://jigsaw.vitalsource.com/books/9781284138887/epub/EPUB/xhtml/23_Chapter11.xhtml#ch11_tbl4). Such numbers give an indication of the breadth of variability in the data set, but more sensitive statistics are needed to evaluate the way individual values vary from the typical case.

**Summarizing the Variability of a Data Set**

Although a typical case can be described statistically, values for individual subjects will differ, sometimes substantially. Statistical techniques can demonstrate variability in ways that enhance understanding of the nature of the individual values represented by the variable scores. With skewed data, a researcher can use a variable’s **range** as well as percentiles to appropriately represent the variability of the data. To calculate the range, the analyst would first identify the minimum and maximum values in the data set. The range, which is determined by subtracting the minimum value from the maximum value of the variable, is the simplest way to represent the spread of the data. This single number provides an indication of the distance between the two most extreme values in the data set. The range is easy to calculate, which makes it useful for getting a quick understanding of the spread of scores. However, more powerful measures use every number in the data set to show the spread of the individual values.

**Range:** A measure of variability that is the distance between the two most extreme values in the data set.

**Variance** and standard deviation are the most commonly used statistics for measuring the dispersion of values from the mean. Just as the mean is used to represent the center of normally distributed data, so the standard deviation and variance are representations of the variability of data. The variance and standard deviation provide information about the average distance of values from the mean of a variable. Variance can be calculated using a calculator, a spreadsheet, or statistical software. The calculation of variance and standard deviation is explained in the box on page 305; it is illustrated here to provide an understanding of the concept.

**Variance:** A measure of variability that gives information about the spread of scores around the mean.

**Table 11.4** Interpreting Measures of Central Tendency

|  |
| --- |
| The variable of length of inpatient stay for adult community-acquired pneumonia has values of:   * Mean: 1.4 days * Median: 0.78 day * Minimum: 0.5 day * Maximum: 12.3 days |
| *Meaning* The median is smaller than the mean; this indicates a positively skewed distribution, because more scores will be in the lower half of the distribution. Thus most people stay in the hospital for shorter periods than this mean value, and some extreme values are likely artificially inflating the mean length of stay. The broad range between minimum and maximum reflects that the scores are spread out. The variable of length of inpatient stay for pediatric community-acquired pneumonia has values of:   * Mean: 2.2 days * Median: 2.4 days * Minimum: 0.5 day * Maximum: 4.6 days |
| *Meaning* The median and the mean are roughly equal; this indicates the distribution is likely a normal one. The range between minimum and maximum is not large, and the similarity between the measures of central tendency is a sign that extreme values are unlikely to have an effect on these statistics. The variable of length of inpatient stay for aspiration pneumonia has values of:   * Mean: 4.3 days * Median: 5.6 days * Minimum: 0.8 day * Maximum: 8.9 days |
| *Meaning* The median is larger than the mean; this indicates a negatively skewed distribution because more scores are in the upper half of the distribution. Thus more of these patients stay in the hospital longer than the average length of stay. There is a moderate range between the minimum and the maximum, and a moderate difference between the measures of central tendency, indicating there may be only a few extreme values in the data set. |

The calculation results in a single score that provides information about the spread of scores around the mean. When the variance increases, the distance of scores from the