

Vocabulary

Term	Definition
5-number summary	A 5-number summary for a variable consists of the minimum, the lower quartile, the median, the upper quartile, and the maximum.
68-95-99.7 Rule	In a normal model, about 68% of values fall within 1 standard deviation of the mean, about 95% fall within 2 standard deviations of the mean, and about 99.7% fall within 3 standard deviations of the mean.
Area principle	In a statistical display, each data value should be represented by the same amount of area.
Bar chart	Bar charts show a bar representing the count of each category in a categorical variable.
Bimodal	Distributions with two modes.
Boxplot	A boxplot displays the 5-number summary as a central box with whiskers that extend to the non-outlying data values. Boxplots are particularly effective for comparing groups of different sizes.
Case	A case is an individual about whom or which we have data.
Categorical data condition	Before making a bar or pie chart, be sure that the categorical data is in counts or percentages of individuals.
Categorical variable	A variable that names categories (whether with words or numerals) is called categorical.
Center	A value that attempts the impossible by summarizing the entire distribution with a single number, a "typical" value.
	We summarize the center of a distribution with the mean or the median.
Changing center and spread	Changing the center and spread of a variable is equivalent to changing its <i>units</i> .
Conditional distribution	The distribution of a variable restricting the <i>Who</i> to consider only a smaller group of individuals is called a conditional distribution.
Context	The context ideally tells <i>Who</i> was measured, <i>What</i> was measured, <i>How</i> the data were collected, <i>Where</i> the data were collected, and <i>When</i> and <i>Why</i> the study was performed.
Contingency table	A contingency table displays counts and, sometimes, percentages of individuals falling into named categories on two or more variables. The table categorizes the individuals on all variables at once, to reveal possible patterns in one variable that may be contingent on the category of the other.
Data	Systematically recorded information, whether numbers, or labels, together with its context.
Data table	An arrangement of data in which each row represents a case and each column represents a variable.

Vocabulary

Term	Definition
Distribution	The distribution of a variable gives the possible values of the variable and the relative frequency of each value.
Dotplot	A dotplot graphs a dot for each case against a single axis.
Experimental units	Animals, plants, web sites, and other inanimate objects upon whom we experiment
Frequency table	A frequency table lists the categories in a categorical variable and gives the count of observations for each category.
Gaps	Regions of a histogram that have no values for a given data set.
Histogram	A histogram uses adjacent bars to show the distribution of values in a quantitative variable. Each bar represents the frequency of values falling in an interval of values.
Independence	Variables are said to be independent if the conditional distribution of one variable is the same for each category of the other.
Interquartile range (IQR)	The difference between the first and third quartile ($IQR = Q3 - Q1$).
Lower Quartile	The lower quartile ($Q1$) is the value with a quarter of the data below it (the median of the lower half of the data set).
Marginal distribution	In a contingency table, the distribution of either variable alone is called the marginal distribution. The counts or percentages are the totals found in the margins (last row or column) of the table.
Mean	The mean is found by summing all the data values and dividing by the count.
Median	The median is the middle value of an organized data set, with half of the data above and half below it.
Midrange	The mean of the minimum and maximum values of a set of data.
Modes	A hump or local high point in the shape of the distribution of a variable is called a "mode." The apparent location of modes can change as the scale of a histogram is changed.
Multimodal	Distributions with more than two modes.
Nearly normal condition	The shape of the distribution of a data set is unimodal and symmetric.
Normal model	A useful family of models for unimodal, symmetric distributions.
Normal percentiles	The normal percentile corresponding to a z -score gives the percentage of values in a standard normal distribution found at that z -score or below.
Normal probability plot	A display to help assess whether a distribution of data is approximately normal. If the plot is nearly straight, the data satisfy the nearly normal condition.
Normality assumption	When using a normal model, we make the assumption that the distribution of the data is normal.

Vocabulary

Term	Definition
Outliers	Outliers are extreme values that don't appear to belong with the rest of the data. They may be unusual values that deserve further investigation, or just mistakes; there's no obvious way to tell. Don't delete outliers automatically - you have to think about them. Outliers can affect many statistical analyses, so you should always be alert for them.
Parameter	A numerically valued attribute of a model. For example, the values of μ and σ in a $N(\mu, \sigma)$ model are parameters.
Percentages	Multiplying proportions by 100 to express as percentages, or ratios compared to a total of 100.
Percentiles	The i th percentile is the number that falls above $i\%$ of the data.
Pie chart	Pie charts show how a "whole" divides into categories by showing a wedge of a circle whose area corresponds to the proportion in each category.
Proportion	A comparison of numbers. For our use, dividing the counts by the total number of cases.
Quantitative data condition	The data are values of a quantitative variable whose units are known.
Quantitative variable	A variable in which the numbers act as numerical values is called quantitative. Quantitative variables always have units.
Quartiles	The median and the quartiles divide data into four equal parts.
Range	The difference between the lowest and highest values in a data set (maximum - minimum).
Records	The rows in a database, usually identifying cases
Re-express / Transform	Applying a simple function to a set of data to make a skewed distribution more symmetric.
Relative frequency bar chart	A bar chart that shows the proportion of people in each category rather than counts.
Relative frequency histogram	A histogram uses adjacent bars to show the distribution of values in a quantitative variable. Each bar represents the relative frequency of values falling in an interval of values.
Relative frequency table	A frequency table lists the categories in a categorical variable and gives the percentage of observations for each category.
Rescaling	Multiplying each data value by a constant multiplies both the measures of position (mean, median, and quartiles) and the measures of spread (standard deviation and IQR) by that constant.
Respondents	Individuals who answer a survey
Segmented bar chart	A bar chart where each bar is treated as the "whole" and divides the bar proportionally into segments corresponding to the percentage in each group.

Vocabulary

Term	Definition
Shape	To describe the shapes of a distribution, look for single versus multiple modes and symmetry versus skewness.
Shifting	Adding a constant to each data value adds the same constant to the mean, the median, and the quartiles, but does not change the standard deviation or IQR.
Simpson's paradox	When averages are taken across different groups, they can appear to contradict the overall averages. This is known as "Simpson's paradox."
Skewed	A distribution is skewed if it's not symmetric and one tail stretches out farther than the other. Distributions are said to be skewed left when the longer tail stretches to the left, and skewed right when it goes to the right.
Spread	A numerical summary of how tightly the values are clustered around the "center."
	We summarize the spread of a distribution with the standard deviation, interquartile range, and range.
Standard deviation	The standard deviation is the square root of the variance.
Standard normal model / distribution	A normal model, $N(\mu, \sigma)$, with mean $\mu = 0$ and standard deviation $\sigma = 1$.
Standardized value	A value found by subtracting the mean and dividing by the standard deviation.
Standardizing	We standardize to eliminate units. Standardized values can be compared and combined even if the original variables had different units and magnitudes.
Statistic	A value calculated from data to summarize aspects of the data.
Stem-and-leaf display	A stem-and-leaf display shows quantitative data values in a way that sketches the distribution of the data.
Subjects/Participants	People on whom we experiment
Symmetric	A distribution is symmetric if the two halves on either side of the center look approximately like mirror images of each other.
Tails	The tails of a distribution are the parts that typically trail off on either side. Distributions can be characterized as having long tails (if they straggle off for some distance) or short tails (if they don't).
Timeplot	A timeplot displays data that change over time. Often, successive values are connected with lines to show trends more clearly.
Uniform	A distribution that's roughly flat.
Unimodal	Having one mode. This is a useful term for describing the shape of a histogram when it's generally mound-shaped.
Units	A quantity or amount adopted as a standard of measurement, such as dollars, hours, or grams.
Upper Quartile	The upper quartile (Q3) is the value with a quarter of the data above it (the median of the upper half of the data set).

Vocabulary

Term	Definition
Variable	A variable holds information about the same characteristic for many cases.
Variance	The variance is the sum of squared deviations from the mean, divided by the count minus one.
z-score	A z-score tells how many standard deviations a value is from the mean; z-scores have a mean of zero and a standard deviation of one.