openSAP
Introduction to Statistics for Data Science

Week 2 Unit 1

00:00:05	Welcome to week two, where we'll be learning about descriptive statistics.
00:00:11	In this first unit, we'll be learning all about key data types, status scales,
00:00:17	and how to appropriately use them. Data can be described according to data type
00:00:22	and also to measurement. Let's look at what these types and measures are.
00:00:30	We'll start with quantitative or numerical data. These data are numbers and can be quantified.
00:00:40	This kind of numerical data can be either discrete or continuous.
00:00:45	Discrete data is based on counts. Only a finite number of values is possible,
00:00:51	and the values cannot be subdivided meaningfully. For example, the number of parts damaged in a shipment,
00:00:59	or the number of students in a class. It's typically things counted in whole numbers.
00:01:07	Continuous data is information that can be measured on a continuum or scale.
00:01:13	So, continuous data can have almost any numeric value and can be meaningfully subdivided
00:01:20	into finer and finer increments depending upon the precision of the measurement.
00:01:28	An example would be weights. We have to watch the weight of one my cats
00:01:34	as it never stops eating. The accuracy of the weight we measure
00:01:38	is only limited by our needs, as we're not concerned about the tiniest microgram.
00:01:46	Quantitative data can be counted or measured and summarized using mathematical operations
00:01:54	such as addition or subtraction. Now, let's move to the other data type,
00:02:00	qualitative or categorical data. These data are not numbers or,
00:02:05	if they are, they cannot be quantified. Such data items can be placed into distinct categories
00:02:14	based on some attribute or some characteristic. These data can be summarized by frequency count or model.
00:02:24	No other mathematical operators can be applied. Now we understand the fundamental data types,
00:02:32	let's move to the nature of the measurement scales which underlies these data types.
00:02:39	Scales of measurement refer to ways in which variables or numbers are defined and categorized.
00:02:46	Each scale of measurement has certain properties which, in turn, determine the appropriateness for use
00:02:55	for certain statistical approaches and analyses. There are four kinds of measurement scale.
00:03:02	In order of increasing sophistication, there's the nominal scale, the ordinal scale,
00:03:07	the interval scale, and the ratio scale. There are four issues to consider when measuring variables.
00:03:16	Firstly, can the items be placed in separate categories? If yes, you should use a nominal or ordinal scale.
00:03:24	Can we rank or order the items from lowest to highest? In that case, you should use an ordinal scale.
00:03:31	Can we say how much one item is more in value than the other item?
00:03:36	For this you should use an interval scale. And, finally, can we say how many times

00:03:41	one item is more in value than the others? Here, we'd use a ratio scale.
00:03:49	Nominal data can be placed in separate categories to distinguish one from the other.
00:03:56	You cannot rank the categories on a value scale, therefore, you cannot say if one category
00:04:02	is higher or lower in value than the other. So, nominal data differ and can be distinguished
00:04:09	qualitatively, but not quantitatively. Also, data cannot be manipulated mathematically.
00:04:17	You can't add, subtract, divide, or multiply. The only statistical calculations that can be applied
00:04:24	are frequency count or mode. For example, you can count that there are 30 females
00:04:31	and 25 male students in a class, but not that a male is ranked on a higher or lower scale
00:04:37	than a female. Ordinal data can be placed in separate categories
00:04:44	to distinguish one item from the other. You can rank the data from lowest to highest in value
00:04:51	so the data can be placed in an order. Although data can be ranked,
00:04:58	you cannot establish the actual interval of difference between two categories.
00:05:04	There is actually no numerical difference in this situation, so you cannot say how much greater or smaller
00:05:11	one item is than the other. Data cannot be manipulated mathematically.
00:05:16	Again, that's add, subtract, divide, or multiply. The only statistical calculations that can be carried out
00:05:24	are frequency count, mode, or median. In contrast, interval data are numbers
00:05:31	and can be quantified. Data can be ranked from the lowest to the highest.
00:05:36	Not only can you say one item is greater or smaller than the other,
00:05:42	for example 15 degrees is warmer than 5 degrees, but you can also say by how much,
00:05:48	or how much warmer it is or how much cooler it is. You can establish the numerical interval difference
00:05:56	between two items, but you cannot calculate how many times one item is more or less in value
00:06:03	than the other. Again, for example, you cannot say 15 degrees
00:06:07	is three times warmer than five degrees C. This is because the starting point of zero
00:06:13	is an arbitrary point. So, zero degrees C does not mean that it's zero
00:06:17	in terms of temperature. In contrast to the Kelvin scale,
00:06:20	where zero degrees would mean there's no heat, it's absolute zero.
00:06:27	The arbitrary starting point can be confusing at first. For example, year and temperature do not have
00:06:35	a natural zero value. The year zero is arbitrary,
00:06:38	and it's not sensible to say that the year 2,000 is twice as old as the year 1,000.
00:06:45	Similarly, zero degrees centigrade does not represent the complete absence of temperature,
00:06:53	the absence of any molecular kinetic energy. In reality, the label zero is applied to its temperature
00:07:00	for quite accidental reasons connected to the freezing point of water,
00:07:06	so it's nothing to do with the absolute zero. So, it makes no sense to say 20 degrees
00:07:11	is twice as hot as 10 degrees, for example. However, again, zero on the Kelvin scale
00:07:16	would be absolute zero. Since an interval scale has no true zero point,
00:07:21	it does not make sense to compute ratios within this kind of scale.
00:07:25	You can only apply limited mathematical operations such as addition and subtraction
00:07:31	to manipulate interval data. And, you cannot use division or multiplication.
00:07:37	Examples include dress size, shoe size, IQ level, temperature, Celsius or Fahrenheit, and so on.

00:07:45 Data can be ranked from the lowest to the highest. You can say that one item is greater

00:07:51 or smaller than the other. For example, a person who is two years of age

00:07:56 is younger than a person who is 36 years of age. You can establish the numerical interval difference

00:08:05 between these two items. For example, the difference in age

00:08:09 between a 12-year-old and a 36-year-old person is 24 years, thus you can say how much greater or smaller

00:08:17 one item is when compared with the other. You can also establish how many times, or the ratio,

00:08:23 one item is more or less in value than the other. For example, you can say a person who is 36 years of age

00:08:31 is three times older than a person who is 12 years of age. The starting point of zero is an absolute point,

00:08:40 which means it represents absolute zero. As a result, all mathematical operations

00:08:47 can be performed, including addition, subtraction, division, and multiplication.

00:08:53 Other examples include, price, income, age in years, weight in kilos, distance,

00:08:59 miles, centimeters, and so on. The data type and measurement

00:09:04 dictate how data should be summarized using mean, median, and mode.

00:09:11 For instance, you cannot find an average for nominal data and ordinal data

00:09:17 because they're not numerical quantities. The inferences that can be drawn from a study

00:09:23 can only be related to the data being used. For instance, it's not correct to employ nominal data

00:09:32 and then draw greater than or less than conclusions. Neither is it correct to use ordinal data

00:09:39 and then summarize how much greater or smaller categories are.

00:09:46 Later in this training, you will see that the data types and measurement often dictate

00:09:51 the type of table and graph that should be used to organize and visualize information.

00:10:00 Understanding data types is fundamental to your goal to ensure the proper use of statistical methods

00:10:07 when analyzing data.

Week 2 Unit 2

00:00:06	Welcome back to week two, unit two, where we're going to be looking at tabular
00:00:11	and graphical methods to describe data. When conducting a statistical study,
00:00:19	a researcher must gather data for each of the particular variables to be studied.
00:00:25	In order to describe situations, draw conclusions, or make inferences about events,
00:00:33	the researcher needs to organize the data in some meaningful way.
00:00:39	Then, after organizing the data, the researcher must present the data so they can be understood
00:00:46	by those who will benefit from the results. A very useful method of organizing and then presenting data
00:00:56	is to construct tables. The type of table to use will depend on the scope
00:01:01	and object of the investigation. The process of placing classified data into tabular form
00:01:08	is known as tabulation. And this is the systematic arrangement
00:01:13	or orderly presentation of the statistical data in columns or rows to explain the problem that's under consideration.
00:01:24	Tabulation prepares the ground for analysis and interpretation.
00:01:29	It helps in drawing the inference from the statistical figures.
00:01:34	There are three main types of table: univariate, bivariate, and multivariate.
00:01:43	This simple univariate table example shows dress choice for 10 women.
00:01:49	The data that appears in one-way tables can easily be represented in a bar chart.
00:01:56	Both bar charts and one-way tables are used to visualize categorical data
00:02:02	in the form of frequency counts or relative frequencies. Frequency counts refer to the number of times
00:02:11	a specific event occurs. Relative frequencies refer to the number of times
00:02:18	a specific event occurs in relation to the total population. In this simple example, the relative frequency of a lady
00:02:28	preferring a red dress is 5 divided by 10, or 50%. Bivariate or two-way tables are ideal for analyzing
00:02:40	relationships between categorical variables. They are sometimes referred to as contingency tables
00:02:47	or as cross-tabulations. The bivariate table below shows data on the leisure activity
00:02:54	of 50 adults, with preferences broken down by gender. The relative frequency for an event is calculated
00:03:01	by dividing the number of times the event occurred by the total number of events.
00:03:08	So, for example, the relative frequency of men preferring football is 10 divided by 50, or 20%.
00:03:19	Often, the behavior you are analyzing is too complicated to be studied with only two variables.
00:03:26	Therefore, you'll want to consider sets of three or more variables,
00:03:31	and this is called multivariate analysis. Once you have conducted your bivariate analysis,
00:03:38	you identify a third variable that you want to consider. This is called the control or test variable.
00:03:47	You then separate the cases in your sample by the categories of the control variable.
00:03:53	In this example, an initial bivariate analysis was conducted to analyze if females were more likely
00:04:01	than males to say they were willing to vote for a woman. Then a control variable was introduced to see if age,

00:04:10	split into younger and older participants, had an effect. There are different approaches to visualize data
00:04:20	depending if the data are quantitative or qualitative. Some of the most commonly used visualizations
00:04:29	are shown in this table. So for quantitative data, for example,
00:04:35	you can use pie charts, histograms, and scatter plots. For qualitative data, for example,
00:04:42	you could use bar graphs, Pareto charts, and heat maps. A pie chart is best used for nominal data
00:04:53	where there are a small number of categories. Pie charts enable quick interpretation of the data
00:05:00	with few mathematical skills. However, it's not appropriate to use a pie chart
00:05:06	to compare two categories. Bar charts represent data size more accurately
00:05:13	and allow for easier comparisons between the datasets. Pie charts are relatively clumsy and lose visual clarity
00:05:22	when there are too many categories represented. As a rule of thumb, 10 or more categories
00:05:28	are usually too many for a pie chart. Pie charts are not very popular
00:05:34	because there are many problems associated with their interpretation.
00:05:40	The brain is not very good at comparing the size of the angles.
00:05:46	There is no scale, so reading accurate values can be difficult.
00:05:51	And as you add more segments and colors, the problem gets worse.
00:05:56	Labels can be hard to fit, especially to smaller segments, so often legends are required.
00:06:05	A histogram is a graphical representation of a frequency table, with percentage values plotted
00:06:12	on the vertical axis, and the class intervals shown along the horizontal axis.
00:06:21	A histogram looks very much like the bar graph used for qualitative data.
00:06:26	However, histograms are used for quantitative data. The vertical bars are placed side by side,
00:06:35	with no space in between them. This is done to reflect the continuous nature
00:06:41	of quantitative data as opposed to the categorical nature of qualitative data.
00:06:51	You can see how increasing the number of data points in the histogram can smooth out
00:06:57	the distribution of the data. A scatter plot uses Cartesian coordinates
00:07:06	to display values for typically two variables. If the points are color-coded,
00:07:11	one additional variable can be displayed. The data are displayed as a collection of points,
00:07:19	each having the value of one variable determining the position on the horizontal axis
00:07:26	and the value of the other variable determining the position on the vertical axis.
00:07:32	A scatter plot can suggest various kinds of correlations between variables, not causation, of course,
00:07:40	with a certain confidence interval. Categorical data have discrete groups,
00:07:48	such as months of the year, age group, shoe size, and animal type.
00:07:53	A bar chart or bar graph presents categorical data with rectangular bars with heights or lengths
00:08:01	proportional to the values that they represent. The bars can be plotted vertically or horizontally.
00:08:09	A vertical bar chart is sometimes called a line graph. Bar charts have a discrete domain of categories,
00:08:18	and are usually scaled so that all the data can fit on the chart.
00:08:24	When there is no natural ordering of the categories which you are comparing,

00:08:30	bars on the chart may be arranged in any order. Bar charts arranged from highest to lowest incidence
00:08:37	are called Pareto charts. A Pareto chart contains bars and a line graph
00:08:46	where individual values are represented in descending order by bars,
00:08:52	and the cumulative total is represented by the line. The purpose of the Pareto chart
00:08:58	is to highlight the most important of a set of factors. In quality control,
00:09:05	it often represents the most common sources of defects, the highest occurring type of defect,
00:09:13	or the most frequent reasons for customer complaints. A heat map is a graphical representation of data
00:09:24	where the individual values contained in a matrix are represented as colors.
00:09:30	There are different kinds of heat maps. There are web heat maps, which are used for displaying areas
00:09:37	of a webpage that are most frequently scanned by visitors. Biology heat map are typically used in molecular biology.
00:09:47	A mosaic plot is a tiled heat map for representing a two-way or higher-way table of data.
00:09:55	So, to summarize, you've learned about the different types
00:10:00	of table you can use. There's univariate, bivariate, and multivariate tables
00:10:06	that organize and present your data. You've also seen which visualizations you should choose
00:10:13	if the data are quantitative or qualitative. In the next unit, you are going to learn
00:10:19	about samples and populations.

Week 2 Unit 3

00:00:05	This unit will introduce you to sampling and populations
00:00:09	and some of the terminology that's commonly used. Population.
00:00:16	This refers to the entire group of items or individuals being studied.
00:00:23	Sample is a part of the population being studied. A representative sample of the population
00:00:31	is needed in order to make an accurate prediction based on the data or to make a valid inference.
00:00:40	Unbiased sample is a sample that is selected so that it's representative
00:00:45	of the entire population. An unbiased sample is selected at random
00:00:51	and is large enough to provide accurate data. A biased sample, in contrast,
00:00:59	is a sample that's drawn so that one or more parts of the population
00:01:03	are favored over others. Often with large populations,
00:01:09	it's not possible or practical to measure the parameters of the whole population.
00:01:14	There are constraints such as cost and time that prohibit the collection of data
00:01:20	from the whole population. Therefore, the mean and standard deviation values
00:01:26	for the whole population might never be known. The population is described by the number capital N,
00:01:36	the mean is mu and standard deviation is sigma.
00:01:41	Okay, let's move now to look at different kinds of sampling. Haphazard sampling: This kind of sampling
00:01:48	is based on convenience and/or self-selection.
00:01:53	Examples include street corner interviews, television call-in surveys, questionnaires published
00:02:00	in newspapers or magazines, or even online. Quota sampling: In this case,
00:02:07	categories and proportions of the sample are predefined. Probability sampling is a sample
00:02:17	of the population in which each person has a known chance of being selected.
00:02:24	In addition, there are a number of different sampling approaches.
00:02:29	First N, last N, which takes a predefined number or percentage; every Nth, which samples,
00:02:37	certain numbers of data, for example, only taking every fifth sample.
00:02:43	And there's simple random. This is a number or percentage.
00:02:47	Systematic random takes buckets and randomly samples from each bucket.
00:02:53	This example shows the results of sampling from a table with 10 records,
00:02:59	first with a bucket size of five which gives two buckets
00:03:03	and then with a bucket size of two that will give five buckets
00:03:07	to be randomly sampled. Stratified sampling, this is where the population
00:03:14	is separated into groups called strata. Then a probability sample,
00:03:21	which is a simple random sample, is drawn from each group.
00:03:27	Since we cannot analyze the whole population, we often choose to draw a random sample
00:03:33	from that population. In this case, the sample size is small n,
00:03:39	the sample mean is represented by x, and the sample standard deviation by small case s.
00:03:46	Sampling error can occur such that there's a difference between the population mean
00:03:51	and the mean of the sample. Sampling bias is a bias
00:03:57	in which a sample is collected in such a way that some members
00:04:01	of the intended population are less likely to be included than others.
00:04:07	This results in a biased sample. And this would be a non-random sample
00:04:13	of a population in which all individuals or instances were not equally likely

00:04:18	to have been selected. If the sample selection process
00:04:23	is based on personal prejudice or bias of the analyst, then the results will clearly be prone to bias errors,
00:04:31	or systematic errors. I'll illustrate this with a well-known example.
00:04:37	In 1936, the "American Literary Digest" magazine collected over two million postal surveys
00:04:45	and predicted that the Republican candidate in the US presidential election,
00:04:51	Alf Landon, would easily beat Franklin Roosevelt, who was the incumbent at the time.
00:04:58	The result was the exact opposite. The "Literary Digest" survey represented a sample
00:05:05	collected from readers of the magazine, and this was supplemented by records
00:05:10	of registered automobile owners and telephone users.
00:05:15	This sample included an overrepresentation of individuals who were rich
00:05:21	and more likely to vote for the Republican party. In contrast, a poll of only 50,000 citizens selected
00:05:29	by George Gallup, George Gallup's organization, successfully predicted the result
00:05:35	and this led to the popularity of the Gallup poll.
00:05:42	This was a short introduction to sampling and populations and the challenges
00:05:47	that are associated with creating representative samples. It also introduced you to some of the terminology
00:05:56	that's used in this area. We'll cover more on this important topic later
00:06:02	in the course.

Week 2 Unit 4

00:00:05	Welcome to week two, unit four where we're going to look
00:00:09	at measures of central tendency. Summary statistics are used to summarize
00:00:16	a set of observations in order to communicate the largest amount of information
00:00:23	as simply as possible. Statisticians commonly try
00:00:27	to describe the observations in a measure of location, or central tendency,
00:00:34	such as the arithmetic mean, a measure of statistical dispersion,
00:00:39	such as the standard deviation, a measure of the shape of the distribution,
00:00:46	such as the skewness or kurtosis. If more than one variable is measured,
00:00:51	a measure of statistical dependence, such as the correlation coefficient.
00:00:57	Measures of central tendency refer to techniques that inform us about the center value of a distribution
00:01:06	or the central point around which other values tend to cluster.
00:01:11	We want to know what value tends to lie in the middle or center of a distribution.
00:01:17	This value could be the most common or the most typical value.
00:01:22	Mean, median, and mode are the most commonly used measures
00:01:27	of central tendency. These methods are used to summarize
00:01:33	an entire distribution into a single number.
00:01:37	The three methods share a common purpose. However, they provide very different approaches
00:01:45	to find the central point in a distribution. The right method to use
00:01:50	to summarize a distribution will depend on the type of data
00:01:55	and how the data was measured. Mode refers to the value in the distribution
00:02:04	that appears most frequently. It's the most commonly occurring value
00:02:10	or the value with the highest frequency. It's used when you want to report
00:02:16	on the most popular or common value in a distribution.
00:02:20	It's most useful when you want a quick and easy indicator of central tendency.
00:02:27	Mode is simple and easy to find as it involves no mathematical calculation,
00:02:33	but it is the least powerful of the three measures of central tendency.
00:02:40	It can be applied to all types of data, numerical and categorical,
00:02:45	and all scales of measurements, nominal, ordinal, interval, and ratio.
00:02:53	For example, in this distribution shown here, the mode is six.
00:02:57	In some cases, the mode may not be central to the distribution as a whole.
00:03:03	So that is, the most common value may actually not be necessarily
00:03:08	the most typical value. Mean, also known as the average,
00:03:16	is the most commonly used technique to summarize a distribution.
00:03:21	The mean is easy to calculate. Add up all the numbers, then divide it
00:03:26	by how many numbers there are. In other words it's the sum
00:03:31	divided by the count. It's closer to all values in a distribution
00:03:36	than any other measures of central tendency. As a result, it's the point in a distribution
00:03:43	around which the variation of the values is minimized.
00:03:48	It can only be calculated when you have numerical data
00:03:53	and measured in interval or ratio scale. You can't calculate a mean
00:03:59	when the data type is categorical or the data is measured in nominal
00:04:03	or ordinal scale. It's affected by every single value in the distribution,

00:04:09	including extreme values. So, the mean is not the best measure
00:04:15	to use for data with extreme values or outliers. Instead the median is the preferred method
00:04:22	when data has an extreme value. While the mean takes into account
00:04:28	every value in the distribution, the mode and the median, on the other hand,
00:04:34	take into account only one or two values in the distribution.
00:04:41	The median is the middle value of a ranked population.
00:04:45	It's the value in the middle position of a distribution.
00:04:49	To find the median, place the numbers you are given in value order and find the middle number.
00:04:56	Median is always at the exact center of a distribution of values,
00:05:01	so it splits the distribution into two equal parts.
00:05:05	Because it's the middle value, half of the data have values
00:05:10	less than the median and the other half
00:05:13	have values more than the median. For example, if the median family income
00:05:21	of a community is $35,000 then half of the families earned less than $35,000
00:05:27	and the other half earned more than $35,000. Median is applicable to ordinal, interval,
00:05:35	and ratio scales of measurement. However, it can't be used for nominal data
00:05:41	because you can't rank nominal data. It's the preferred method
00:05:47	when your data is skewed. To calculate the median arrange the measurements
00:05:53	from the smallest to the largest. And then, if the number of measurements is odd,
00:05:59	the median is the middle number. However, if the number of measurements is even,
00:06:05	the median is the mean of the middle two numbers. Here you can see a summary
00:06:15	of when to use each of the three averages, mean, median, and mode.
00:06:21	Use the mode if the data type is numerical or categorical
00:06:25	and the measurement scale is nominal, ordinal, interval, or ratio.
00:06:32	Use the median if the data type is numerical or categorical
00:06:36	and the measurement scale is nominal, ordinal, interval, or ratio.
00:06:42	And use the mean if the data type is numerical and the measurement scale is interval or ratio.
00:06:51	The data has also no extreme data points. Your knowledge of the mean and median
00:07:03	can help you determine the shape of the distribution, so you can identify if the data
00:07:09	are skewed by an extreme value. Skewness refers to the direction of a distribution.
00:07:17	Data can be skewed in three different ways. Firstly, data skewed to the left,
00:07:23	a negative skew, indicates that it's concentrated
00:07:28	at the high end of the distribution. Secondly, data could be skewed to the right,
00:07:34	a positive skew. That indicates that it's concentrated
00:07:38	at the low end of the distribution. Thirdly, it could be symmetric.
00:07:44	That indicates that it's concentrated in the middle of the distribution,
00:07:49	and it's a bell-shaped curve or bell-shaped distribution.
00:07:55	If the data are skewed to the right, then the median is less than the mean.
00:08:01	If the data are symmetric, then the mean equals the median.
00:08:07	If the data are skewed to the left, then the mean is less than the median.
00:08:15	So in summary, you've learned that measures of central tendency
00:08:21	refer to techniques that inform you about the center value of a distribution.
00:08:27	Mean, median, and mode are the most commonly used measures
00:08:33	of central tendency. These methods are used to summarize

00:08:39 an entire distribution into a single number.
00:08:43 In the next lesson, you're going to learn about measures of dispersion.

Week 2 Unit 5

00:00:05	Hello and welcome to unit five, where we're looking at measures of dispersion.
00:00:12	Dispersion, which is also called variability, or scatter or spread,
00:00:16	is the extent to which a distribution is stretched or squeezed.
00:00:21	Common examples of measures of statistical dispersion are variants, standard deviation,
00:00:28	and interquartile range, which is discussed in a later lesson.
00:00:34	In this simple example, measuring along the horizontal axis,
00:00:39	you can see that Distribution A has the lowest dispersion.
00:00:44	And Distribution C had the highest. Knowing the average provides no information
00:00:53	about variability. For example, in both of these two groups,
00:00:57	the average age is 36 years. There were ages between 34 and 38, the average is 36 years
00:01:05	and in the second group, the range is from the ages of five to 60, again the average is 36 years.
00:01:13	The average does therefore not give you information about variation in the age in the group.
00:01:20	And by knowing only the average, you don't know if the individuals
00:01:24	within a group have similar ages. The range is the difference
00:01:31	between the largest and the smallest value in a distribution.
00:01:36	It's easy to calculate the range, but it has the limitation of only considering the high
00:01:42	and low values, the two extreme values, and it ignores all the values in between.
00:01:49	The three datasets at the bottom of the slide may seem to have the same variability,
00:01:55	but, in fact, the variation is not the same. When looking for variation,
00:02:00	we need to look for a better method that considers all the values in the distribution
00:02:07	and not just the highest and the lowest. Standard deviation is a method used
00:02:14	to quantify how values in a distribution fluctuate from the center,
00:02:19	that is, from the mean. It's the most widely used technique
00:02:24	to measure variation in data. It takes into account all the values
00:02:30	within the dataset within a distribution. Let's compare the formulas
00:02:35	to calculate the standard deviation for the population or for a sample
00:02:40	which is taken from a population. If the data is being considered a population on its own,
00:02:46	you divide by the number of data points n. However, if the data is a sample from a larger population,
00:02:54	you divide by one fewer than the number of data points in the sample, so n minus one.
00:03:02	You are normally interested in knowing the standard deviation,
00:03:05	of course, for the population because the population contains all the values
00:03:10	that you have an interest in. However, in statistics,
00:03:14	you're usually presented and faced with only a sample,
00:03:17	from which you wish to estimate and generalize to the population.
00:03:22	Therefore, if all you have is a sample, but you wish to make a statement
00:03:27	about the population standard deviation from which the sample is drawn,
00:03:34	you need to use the sample standard deviation. Confusion can often arise
00:03:40	as to which standard deviation to use due to the name sample standard deviation.
00:03:48	It incorrectly being interpreted as meaning the standard deviation of the sample itself
00:03:55	and not the estimate of the population as a whole. The empirical rule, also known as the three-sigma rule,
00:04:06	or the 68-95-99.7 rule, provides a quick estimate

00:04:11	of the spread of data in a normal distribution. Those are distributions that are bell shaped
00:04:16	and symmetrical about the mean, given the mean and standard deviation.
00:04:24	Approximately 68% of the measurements will fall within one standard deviation of the mean.
00:04:31	Approximately 95% of the measurements will fall within two standard deviations of the mean.
00:04:37	And approximately 99.7 of the measurements will fall within three standard deviations of the mean.
00:04:44	In other words, within the interval. Another term that you'll come across is variance.
00:04:53	Variance is the average squared and standard deviation of values from the mean.
00:04:59	So standard deviation equals the square root of the variance.
00:05:04	The variance of a dataset measures the mathematical dispersion
00:05:08	of the data related to the mean. Calculating variance involves squaring deviations.
00:05:16	Therefore, it does not have the same unit of measurement as the original observations.
00:05:23	For example, lengths measured in meters have a variance measured in meters, meters squared.
00:05:32	Taking the square root of the variance gives us the units used in the original scale,
00:05:39	and this is the standard deviation. So the standard deviation is expressed in same units
00:05:46	as the mean is whereas the variance is expressed in squared units.
00:05:53	However, whenever you're looking at a distribution, you can use either,
00:05:57	so either standard deviation or variance, as long as you are clear about what you're using.
00:06:06	In summary, you've seen why looking at the range or average values of a distribution
00:06:12	without considering the dispersion will not give you a full understanding of the data.
00:06:21	Common measures include variance and standard deviation.
00:06:27	You've also been introduced to the empirical rule that helps explain the spread in a distribution.
00:06:35	Next, you will look at detecting outliers and the interquartile range.

Week 2 Unit 6

00:00:05	Welcome to week two, unit six, where we're going to be looking at outliers.
00:00:12	An anomaly is something that deviates from what is standard, normal, or expected.
00:00:20	In statistics, an outlier is an observation that is numerically distant from the rest of the data.
00:00:28	Some statistics and algorithms can be heavily biased by outliers.
00:00:33	For example, the simple mean correlation and linear regression.
00:00:38	In contrast, the trimmed mean and median are not so affected.
00:00:44	Outliers can be detected visually with scatter plots or box plots.
00:00:50	Although data volumes can limit these approaches and by using statistical and algorithmic techniques.
00:00:59	Outliers can occur because of errors and might need to be removed from the data or corrected.
00:01:07	They can occur naturally and therefore must be treated carefully.
00:01:13	The outlier can be the most interesting thing in the dataset, for example in fraud analysis,
00:01:19	where you might be trying to specifically detect unusual behavior.
00:01:29	To understand one of the most popular tests for outliers, you will need to understand quartiles.
00:01:37	Quartiles are referred to as measures of position, because they give the relative ranked position
00:01:44	of a specific value in a distribution. To create quartiles, you simply order the data by value
00:01:52	and then split it into four equal parts. The second quartile, Q2, is the median of the data.
00:02:01	These data show the number of volunteer hours performed by 15 students in a year.
00:02:09	These values have been ranked, with the lowest value on the left
00:02:13	and the highest on the right. The box plot shown here describes the distribution,
00:02:22	as you can see, where the lowest to the highest range is drawn out within a clear range description.
00:02:33	The interquartile range is the difference between the highest and lowest values for the middle 50%
00:02:40	of the ranked data. This refers also to the spread of the middle 50%
00:02:45	of the data. The interquartile range is Q3 minus Q1.
00:02:50	In this example, half the values lie between 24 and 46. So the interquartile range is 22.
00:02:58	This refers to the distance or the difference between the bottom 25%
00:03:02	and the upper 25% of the ranked data. Upper and lower fences cordon off outliers
00:03:10	from the bulk of the data. A point beyond an inner fence on either side
00:03:15	is considered a mild outlier. A point beyond an outer fence is considered
00:03:21	an extreme outlier. Fences are usually found with the following formulas.
00:03:27	The upper fence is Q3 plus 1.5 times the IQR, and the lower fence is Q1 minus 1.5 times IQR.
00:03:36	Sometimes you'll see references to inner and outer fences and the formulas are given here for you
00:03:42	in the slide for your reference. The values that are contained or lie within the inner
00:03:51	fences are the usual values. For usual values, the lowest value here is 12,
00:03:57	and the highest value is 48. The values that lie between the inner and outer fences
00:04:04	are called outliers or suspect outliers and are denoted by a circular symbol.
00:04:11	Here it's 86. The values that lie outside the outer fence
00:04:15	are called extreme or highly suspect outlier values and are denoted by a star symbol.

00:04:23	In this case, it's 128. In the box plot, you can see that the central point
00:04:30	of the distribution is determined by the location of the median.
00:04:36	This central could be nearer to the first or third quartile, or equidistant.
00:04:42	In this example, the central point is 30. The spread of the data is determined by the length
00:04:48	of the box. It represents the spread of the middle 50% of the data. It's also known as the interquartile range.
00:04:56	The wider the box, the more spread or variation in the data.
00:05:01	In this example, the data spread is 22. The shape of the distribution is determined by
00:05:07	comparing the length of the right and left whiskers. If the left whisker is the longer one,
00:05:15	then the distribution is left skewed. However, if the right whisker is the longer one
00:05:21	then the distribution is right skewed. An outlier or extreme value refers to any value
00:05:27	that lies below the left inner fence or above the right inner fence.
00:05:32	In this example, 86 is an outlier, and 128 is an extreme value.
00:05:39	The normal distribution is a theoretical distribution with the mean, median, and mode positioned at the same point,
00:05:48	the exact center of the distribution. The location of a normal distribution is determined by
00:05:55	the mean, and the spread is determined by the standard deviation.
00:06:02	Distance away from the mean is measured in standard deviation units known as zed scores.
00:06:09	You'll be learning more about the normal distribution later in this course.
00:06:15	However, this is a brief introduction because the properties of a normal distribution can help to detect outliers.
00:06:24	The empirical rule states that for a normal distribution nearly all of the data will fall within three standard
00:06:32	deviations of the mean. The rule is also called the 68-95-99.7 rule,
00:06:39	or the three-sigma rule. The rule applies generally to a random variable X,
00:06:45	following the shape of a normal distribution. The average score in this normally distributed data
00:06:53	is 65, and there is a standard deviation of 10. If the scores aren't normally distributed,
00:07:00	then it means one standard deviation is measured between 55 and 75, two standard deviations measured
00:07:08	between 45 and 85, and three standard deviations between 35 and
00:07:14	Observations with zed scores greater than three in absolute values are considered outliers.
00:07:21	For some highly skewed data sets, observations with zed scores greater than two
00:07:27	in absolute value may be outliers. This lesson has introduced you to some simple
00:07:35	but very powerful methods to detect outliers. You've seen how the interquartile range, box plot,
00:07:43	and empirical rule can be used to test for outliers. The empirical rule and box plot methods both establish
00:07:52	rule-of-thumb limits, outside of which a measurement is deemed to be an outlier.
00:07:58	Usually, the two methods produce similar results. However, the presence of one of more outliers in a dataset
00:08:06	can inflate the computed value of the standard deviation. Consequently, it will be less likely
00:08:14	that an errant observation would have a zed score larger than the absolute value three.
00:08:21	In contrast, the values of the quartiles used to calculate the intervals for a box plot are not affected
00:08:28	by the presence of outliers.

Week 2 Unit 7

00:00:06	Welcome to this unit on distorting the truth using descriptive statistics.
00:00:12	Whatever types of data visualization you choose to use, they must show that the correct scales used,
00:00:19	the starting value, often zero but not necessarily, the method of calculation,
00:00:24	the dataset used, and the time period, the analysis covered.
00:00:28	Also going to show if any of these elements are missing. Then otherwise, the visualization risks
00:00:34	distorting the truth. There are a number of ways the media
00:00:40	and unscrupulous politicians present data and distort the truth.
00:00:46	Common approaches include using area to equate to value, not providing a relative basis to compare datasets,
00:00:56	compressing the vertical axis, ignoring the zero point on the vertical axis,
00:01:03	using a gap on the vertical axis, using misleading wording,
00:01:09	providing a central tendency value but not the variability, omitting data, confused visualization.
00:01:17	Let's look at a few of these in a bit more detail. A visualization that uses area to equate to a value
00:01:26	as in the example here can easily be used to distort the truth.
00:01:32	The size of the dollar note from 1990 looks far larger than 3.8 times the size of the 1960 note.
00:01:40	This example shows how simple frequency charts can be used to distort the truth.
00:01:48	The graph on the left shows the frequency of A grades for each category.
00:01:54	However, it does not take into consideration the total volume in each category.
00:02:01	However, the graph on the right shows the percentage of A grades per category,
00:02:06	which shows performance much more clearly. This example shows how extending the vertical axis
00:02:16	can be used to hide information by compressing the differences in the bars.
00:02:24	The presentation on the left has the range of the vertical axis extended
00:02:29	so that the differences in the quarterly figures are hidden.
00:02:34	The presentation on the right shows the quarterly differences much more clearly.
00:02:44	The conventional way of representing the data on the y-axis is to start at zero and then go up
00:02:50	to the highest data point in your set. By not setting the origin of the y-axis at zero,
00:02:57	small differences become exaggerated. The exception to this practice would be
00:03:04	if these small differences themselves actually mean something significant.
00:03:10	For example, in global climate change data, an increase in temperature of only one degree
00:03:18	can be very significant. Even though, in this case,
00:03:22	it's a very small increase in temperature, this is very important and needs to be highlighted.
00:03:29	However, when small changes are not correlated with a big impact,
00:03:36	then you should start your y-axis at zero. Distorting the vertical axis is a classic way
00:03:43	to visually mislead. This famous example from Fox News
00:03:47	shows a graph comparing people with jobs versus people on welfare.
00:03:54	The height of the bar suggests people on welfare are four times as many as those with jobs.
00:04:02	However, the actual numbers, 108.6 million versus 101.7 million,
00:04:08	are much less sensational than the data visualization would suggest.
00:04:12	Also, this graph shows how selective data collection criteria can be used to deceive.
00:04:19	Fox's 108.6 million figure for the number of people on welfare

00:04:24	comes from a Census Bureau's account of participation in means-tested programs,
00:04:31	and this includes anyone residing in a household in which one or more people receive benefits.
00:04:41	This, therefore, included individuals who did not themselves receive government benefits.
00:04:49	On the other hand, for the people with a full-time job, Fox included only individuals who worked,
00:04:58	not individuals residing in a household where at least one person worked.
00:05:04	In other words, if you lived with your parents or any member of your family,
00:05:10	and the family member was briefly on some kind of welfare,
00:05:16	that counted against you, and everyone in the household.
00:05:22	So avoid the use of gaps in the vertical axis. This is another method of focusing attention
00:05:29	on small differences. It's very easy to influence a reader
00:05:36	by changing the title of a visualization. Compare the titles on the left
00:05:41	and the right charts in this example. Both charts show exactly the same information,
00:05:47	but the title changes your interpretation. You should use non-emotive titles
00:05:53	that will not influence the reader. It's very easy to influence the interpretation
00:06:00	by only presenting the average of a range of values without referring to the spread of the data.
00:06:09	In this example, the average miles per gallon for two cars are compared.
00:06:14	Knowing only the central tendency, the mean, might lead you to purchase model A.
00:06:21	However, knowing the variability as well might change your decision.
00:06:29	If certain data points are omitted in the visualization, it's possible to create a trend
00:06:34	that doesn't actually exist or hide or highlight something that goes unnoticed.
00:06:40	In this example, compare all of the data shown in a graph on the left to that on the right,
00:06:46	where some of the data are omitted. By only plotting every second year instead of every year,
00:06:54	the right-hand graph appears to have a steady, more stable increase
00:06:59	while the real data shown in the left-hand graph is much more variable.
00:07:06	It's possible for companies to then take advantage of this by simply omitting years which show significant changes
00:07:13	in sales, for example, or costs so that they make their profit look constant
00:07:19	and therefore predictable. Many statistical tests
00:07:23	calculate correlations between variables. When two variables are found to be correlated,
00:07:30	it's tempting to assume that this shows that one variable causes the other.
00:07:37	A correlation between two variables does not automatically mean
00:07:42	that the change from one variable is the cause of the other. Causation indicates that one event
00:07:52	is the result of the occurrence of the other event. In other words, there is a causal relationship between the two events.
00:08:02	In the example, you can see the correlation between murders and sales of ice creams.
00:08:08	This does clearly not indicate that ice cream sales or ice cream consumption leads to murder.
00:08:14	We'll return to this subject of correlation and causation later in the course.
00:08:22	These two graphs represent the same data. The graph on the left is confused and difficult to read.
00:08:29	The graph on the right is much easier to understand, to visualize the trends,
00:08:36	and more aesthetically pleasing. You need to create your visualizations
00:08:42	so they are as clear as possible so that the information they contain

00:08:46	can be understood very easily. This is also a famous example
00:08:52	of the use of a visualization to distort the truth, apparently showing a comparison of the increase in abortions
00:09:02	against the decrease in life-saving procedures. On September 29, 2015, Republicans from the U.S. Congress
00:09:10	questioned Cecile Richards, the president of Planned Parenthood,
00:09:15	and this was regarding misappropriation of $500 million in annual federal funding.
00:09:21	This chart was presented as a point of emphasis. Representative Jason Chaffetz from Utah explained,
00:09:30	"In pink, that's the reduction in the breast exams, and the red is the increase in the abortion rate.
00:09:38	That's what's going on in your organization," he was claiming.
00:09:42	This chart is designed to show that the number of abortions since 2006
00:09:47	experienced substantial growth while the number of cancer screenings
00:09:52	substantially decreased. This is an attempt to show that there was a shift in focus
00:09:58	from cancer screenings to abortions. The chart appears to indicate the 327,000 abortions
00:10:06	are greater in value than 935,000 cancer screenings. However, closer examination reveals
00:10:13	that the chart has no defined y-axis. Therefore, there is no justification
00:10:19	for the placement of the measurement lines. This graph was designed to mislead and distort.
00:10:28	PolitiFact, a fact-checking advocacy website, reviewed Representative Chaffetz's numbers
00:10:35	using a comparison with Planned Parenthood's own annual reports.
00:10:41	In blue, you can see the cancer screenings and in red, the abortion procedures.
00:10:47	That's on the left-hand side. You can see the actual numbers
00:10:52	clearly shown on the y-axis, and on the right-hand chart,
00:10:56	the year-on-year percentage change. Once placed within a clearly defined scale,
00:11:02	it becomes evident that while the number of cancer screenings has, in fact, decreased,
00:11:10	it still far outnumbers the quantity of abortions. These two graphs are very clear
00:11:19	and are designed to give you a truthful account of the situation.
00:11:28	This is another famous example that has been used in the global warming debate.
00:11:35	It's generally agreed that the global mean temperature in 1998 was 58.3 degrees Fahrenheit.
00:11:43	This is according to NASA's Goddard Institute for Space. In 2012, the global mean temperature
00:11:51	was measured at 58.2 degrees. It's therefore argued by global warming opponents
00:12:00	that as there was only a 1.1-degree decrease in the global mean over a 14-year period,
00:12:08	global warming is disproved. This growth is one of the most often referenced graphs
00:12:14	to disprove the global warming theory. It demonstrates the change in air temperature (Celsius)
00:12:20	from 1998 to 2002. However, only looking at this figure over this short time
00:12:26	is designed to mislead and distort the truth. It's worth mentioning that 1998
00:12:34	was one of the hottest years on record due to an abnormally strong El Nino
00:12:41	with wind currents, and so on. It's also worth noting that
00:12:44	as there is a large degree of variability within the climate system,
00:12:51	the temperatures are typically measured with at least a 30-year cycle.
00:12:57	This chart expresses a 30-year change in global mean temperatures.
00:13:02	This graph shows a long-term view from 1900 to While the long-term data may appear to reflect a plateau,
00:13:12	it clearly indicates gradual warming. Therefore, using the first graph

00:13:18	that plots data from 1998 to 2012 only is designed to disprove global warming.
00:13:25	It's misleading and distorting the truth. Visualization guru Edward Tufte explains,
00:13:34	"Excellence in statistical graphics consists of complex ideas communicated with clarity,
00:13:41	precision, and efficiency." Because of the variation that inevitably crops up
00:13:47	in graphical representations of data, Tufte came up with six principles
00:13:52	that are meant to ensure high graphical integrity. These are outlined in his book,
00:13:59	"The Visual Display of Quantitative Information". These are as follows:
00:14:06	Representation of numbers should match the true proportions. Labeling should be clear and detailed.
00:14:14	Design should not vary for some ulterior motive. To represent money, well known units are best.
00:14:24	The number of dimensions represented should be the same as the number of dimensions in the data.
00:14:33	Representations should not imply unintended context. We hope you've enjoyed this week's content.
00:14:42	After you have had the opportunity to complete the assignment,
00:14:47	we'll move to next week's lessons covering correlation and linear regression.
00:14:54	Thank you very much.

www.sap.com/contactsap

THE BEST RUN