openSAP
Introduction to Statistics for Data Science

Week 1 Unit 1

00:00:06	Hello and welcome to the openSAP course, Introduction to Statistics for Data Science.
00:00:11	My name's Mike Jordan, and I'm an Education Portfolio Manager and Academic Ambassador.
00:00:17	I teach data science to master students for SAP. I've developed this course together
00:00:23	with my colleague Stuart Clarke, who is a consultant and trainer also delivering
00:00:27	data science projects around the world, again, for SAP. So now, what should you expect over the next six weeks?
00:00:35	Well, this week, in week one, we'll introduce you to statistics, looking at how we assess numbers
00:00:41	in everyday life, some key statistical terms, and different types of analytical approaches.
00:00:47	Next week, in week two, we'll look at descriptive statistics, measures of central tendency,
00:00:53	and measures of dispersion. In week three,
00:00:56	we'll look at correlation and linear regression. And moving to week four, we'll introduce you
00:01:01	to probability and this thing called Bayes' theorem. Then, in week five, probability distributions,
00:01:06	looking at the normal distribution in a little bit more detail and discuss hypothesis testing.
00:01:11	Finally, in week six, we'll look at some of the SAP solutions that provide statistical functionality.
00:01:18	After having successfully completed the first six weeks, you'll have one further week to prepare for
00:01:24	and to participate in the final exam to earn a record of achievement.
00:01:29	Throughout the course, your feedback, your questions, and your ideas are very welcome, very appreciated.
00:01:36	If you put them in your discussion forum, we'll appreciate that.
00:01:38	So how do you get points and successfully complete the course?
00:01:43	Well, there are six graded assignments throughout the first six weeks of the instructional content.
00:01:49	Each assignment is worth 30 points for a total of 180 points, which is half of the total
00:01:55	of points available in the course. The other half of the available points
00:01:59	come from the final exam. And just like every openSAP course,
00:02:04	you need at least half of the maximum points available, in this case, 180 points, to pass the course
00:02:11	and receive your record of achievement. The goal of this course is to provide you
00:02:16	with an introduction to statistics for data science. Statistics covers a wide range of numbers
00:02:22	and topics that many people find difficult to understand. So we'll try to demystify this for you,
00:02:28	so that you have a good, clear understanding of the major topics.
00:02:33	By the end of the course, you should have a basic understanding that will enable you
00:02:38	to take the next steps in your statistics adoption journey. We'll teach you all you need to know
00:02:44	about the fundamentals of statistics. Numbers are everywhere in our everyday lives.
00:02:51	During every hour of every day, we make decisions and judgments based on data.

00:02:57	For example, when you are making a decision to purchase a house, you will need to decide on the location,
00:03:04	the size of the house, the town where you want to live, also, proximity to any services, shops, and the seaside,
00:03:11	crime rates in your local area, property prices, size and number of rooms in the house,
00:03:17	and, of course, the condition of the house itself on top of everything else.
00:03:22	So typically, you're faced with claims and numbers at many moments during every day.
00:03:29	You even face them when you're doing the shopping. Do you buy the offer?
00:03:34	Do you take a bunch of options or not? The three-pack of Jaffa Cakes,
00:03:36	is that better value than the one-pack? Which is the best value?
00:03:40	Which is the best value supermarket? Before we get into how we can use statistics
00:03:44	to get underneath these claims, I'll introduce you to some of the jargon.
00:03:49	None of it is especially complicated to understand, but there are some initial key terms
00:03:56	that you do need to understand. In our personal lives, we consider many
00:04:03	statistical problems: For example, how much sleep should the average person get?
00:04:09	Is there a difference by age, gender, and so on? What lifestyle characteristics
00:04:14	influence sleep quantity and quality? How do I test whether a sleep intervention
00:04:20	is effective or not? Also, in our business lives,
00:04:25	we consider statistical problems. How much profit are we making?
00:04:30	How many product defects have we come across or are we discovering?
00:04:34	What's the trend? Has the process change led to a significant increase
00:04:39	or decrease in employee satisfaction or productivity? What is our customer churn rate?
00:04:46	From now, we're going to introduce you to some basic terms that will help you understand some of the important concepts
00:04:53	that we'll be discussing later on; population versus sample, randomness,
00:05:00	descriptive statistics, distributions, inference, probability, correlation.
00:05:08	Let's first consider population datasets and sample datasets.
00:05:13	A population dataset contains all possible members of a specified group,
00:05:17	the entire list of possible data values. A sample dataset, in contrast, contains only a part,
00:05:23	or a subset, of a population. The size of the sample is always less than
00:05:27	the size of the population from which it is drawn. It's often impractical
00:05:32	and just too expensive to collect data for your whole population.
00:05:37	Therefore, sometimes it's only practical just to take a sample set representative of that whole population.
00:05:43	But what is representative? What does that mean?
00:05:45	Descriptive statistics attempt to summarize a large body of data so that you can highlight key information.
00:05:53	This is mainly through measures of central tendency and measures of dispersion.
00:05:58	Measures of central tendency try to find the center of the data: for example,
00:06:03	using the average or using quartiles. Measures of dispersion provide insight
00:06:08	into the spread of the data. Example: A measure of the average salary where you work
00:06:14	may not reflect yours at all. There may be a huge spread of salaries,
00:06:18	so you need to see the dispersion. A frequency distribution provides a way of viewing
00:06:24	all the values of a sample in a table view or a histogram. Typically, the data bunches around the center.
00:06:32	This is because this is where most of the data values are to be found.
00:06:35	For example, for my wife, she often complains that she's very average in size.

00:06:40	She can't find size 12 dresses because they're the ones that are always gone, but those are the most common sizes.
00:06:48	A probability distribution is a mathematical function that describes the probability of getting
00:06:53	any particular result, such as rolling two dice. We'll explain this in more detail later.
00:07:00	It's important to note that there are many types of specialty distribution,
00:07:05	some of which we will learn about later in the course, and some of which we won't go through.
00:07:11	Probability is a statistical measure. It shows the likelihood of any event happening.
00:07:16	It's typically measured between zero and one, so that an event that definitely
00:07:22	won't happen is at zero and an event that absolutely will happen is represented,
00:07:26	at one, but most probabilities are in the middle. So probability for all possible events
00:07:32	will always add up to one. For example, a coin toss as heads or tails would each have
00:07:37	a probability of 0.5 and that would add up to one. In statistics, dependence or association
00:07:47	is any statistical relationship, whether causal or not, between two random variables,
00:07:52	and that's described as correlation. For example, think about people's height and weight.
00:07:57	What kind of relationship exists between those two variables?
00:08:01	Do you think that there is a perfect relationship, so any increase in weight is followed by
00:08:06	an increase in height? Probably not.
00:08:08	Or there's no relationship? Also, probably not.
00:08:11	It's more likely that that relationship is in the middle, so the probability will be between zero
00:08:16	and one, as we mentioned before, for any probability. So in this unit,
00:08:22	you have seen that statistics is everywhere, and every day, we are expected to make
00:08:29	different kinds of statistical estimations and judgments. As part of this, we're bombarded with different kinds
00:08:36	of statistical claims from parties who want our vote or want our time or want our money.
00:08:43	In this course, we will learn how to evaluate these claims, and there are some key statistical terms,
00:08:48	which we need to understand to be able to develop our skills in understanding statistics.
00:08:56	These first terms we need to understand are population versus sample, randomness,
00:09:01	descriptive statistics, distributions, probabilities, and correlation.
00:09:07	I hope you've enjoyed this unit, and we'll now move into unit two,
00:09:10	where we'll be looking at numbers in everyday life.

Week 1 Unit 2

00:00:05	Hello, my name's Stuart Clarke and welcome to the second unit
00:00:08	in week one, where we'll be looking at numbers in everyday life.
00:00:13	In this unit, we look at everyday examples, where we have to evaluate
00:00:18	and make judgments using numbers and statistics. In many cases, we hardly think that we are doing this type
00:00:26	of evaluation, but, of course, we are. There are countless examples from everyday life,
00:00:31	but I've chosen three from the day I was building this unit.
00:00:36	For each of the claims made, we would like you to think about how you would evaluate
00:00:41	and test them. After each one, we're going to compare notes
00:00:46	with a few things I thought about. Let's look at some of the statistical claims
00:00:54	from my day today. Advertising, nine out of 10 cat owners say
00:00:59	their cats prefer x brand of cat food. In the newspaper, Bill Gates is saying
00:01:04	that poverty is decreasing. He couldn't be more wrong, of course,
00:01:08	so The Guardian says. And in shopping, does it make sense
00:01:12	for us to actually buy the Buy-One-Get-One- Free offer? Let's test these claims one by one.
00:01:21	Let's look at the advertising. Nine out of 10 cat owners prefer brand x.
00:01:27	So we need to consider what the sample size is. How could they test that the cats prefer the cat food
00:01:33	to other food? What were the testing procedures,
00:01:36	and how did they test it against each brand? Did the cats simply enjoy the food
00:01:41	because it was new and different? Will they get bored, of course, in the future?
00:01:47	In the newspaper, Bill Gates is saying poverty is decreasing.
00:01:50	He couldn't be more wrong, says The Guardian. How are we defining poverty?
00:01:56	What is the baseline? Do these facts make actual sense?
00:02:01	What's the reasoning behind it, and is there some bias? Do we have the knowledge
00:02:07	to legitimately make all these claims? And in shopping, which bananas should I buy?
00:02:14	A bag for a pound or individual ones for 25p each? Do I buy large ones or small ones
00:02:19	or organic ones? How do we compare all this?
00:02:22	Are the sizes the same? Is the quality and ripeness comparable between them?
00:02:29	If there is a Buy-One-Get-One-Free offer that could save me some money,
00:02:34	and how do I evaluate all of this? So to summarize this.
00:02:38	Numbers are in our everyday lives, and these claims are part of our lives,
00:02:43	and we have to interact with them. Selling goods and services often use tactics
00:02:49	which are not necessarily untruthful but make it difficult for us to check the claims.
00:02:54	And we are constantly trying to evaluate these claims to understand whether they are biased or truthful.
00:03:02	In the next unit, we'll be looking at the use and abuse of numbers.

Week 1 Unit 3

00:00:05	Hello and welcome to the third unit in week one, where we will be looking at
00:00:09	the use and abuse of numbers. Although numbers don't lie,
00:00:16	they can be used to mislead with half-truths. And this is known as "abuse of statistics".
00:00:23	You might think this misuse is confined to politicians, but a 2009 study show that nearly 34% of scientists
00:00:30	admitted to questionable research practices, including such things as modifying results
00:00:35	to improve outcomes, subjective data interpretation,
00:00:39	withholding analytical details and taking observations out of gut feelings.
00:00:46	To be able to interpret data, it's important that you are familiar with the basics
00:00:51	of statistical misuse. In this presentation, we will review
00:00:55	some of the most common forms of misuse. Often when a company promotes a product,
00:01:02	they will undertake studies to "prove" the product's effectiveness.
00:01:06	Let's say they conducted 40 studies with a confidence level of 95%,
00:01:12	so you might think that the analysis would be fairly robust. But they could produce one study showing it was beneficial,
00:01:21	one showing it was harmful and 38 they were inconclusive.
00:01:26	Then that 38 equates to the 95%. So, the company could be very selective
00:01:31	and cherry-pick the results. This is the type of misuse
00:01:34	that we sometimes see with tobacco companies, and some pharmaceutical companies arguably,
00:01:39	promoting miracle pills. The manner in which questions are phrased
00:01:44	can also have a massive impact on the way an audience answers the questions.
00:01:50	So specific wording patterns have a persuasive effect on people,
00:01:54	and influence respondents and way they answer in a predictable manner.
00:01:59	For example: Do you believe that you should be taxed
00:02:03	so other people don't have to work? Or do you think that the government
00:02:08	should help those people who can't find work? The question should be posed in a neutral way.
00:02:14	For example, what is your point regarding unemployment assistance?
00:02:19	This will ensure that the person being polled has no way of guessing from the wording
00:02:23	what the questioner might want to hear. Other examples:
00:02:29	Do you support the UK attempting to bring freedom and democracy
00:02:33	to other countries in the world? Or do you support the unprovoked military action by the UK?
00:02:41	Another unfair method of polling is to precede the question with a conditional statement
00:02:47	or a statement of fact. For example, "Given the rising costs to the middle class,
00:02:52	do you support government assistance programs?" Overgeneralization is a logical fallacy
00:02:59	that occurs when a conclusion about a group is drawn from an unrepresentative sample,
00:03:05	especially a sample that is too small or too narrow. For example,
00:03:09	there are common overgeneralizations regarding certain religions or races.
00:03:14	All Jews are rich. All Blacks are good athletes.
00:03:16	The French are rude. All Mexicans are lazy.
00:03:19	All Arabs are terrorists, and so on, and so on. These are basically stereotypes of certain groups of people.
00:03:26	These blanket statements maybe they are true for particular individuals,

00:03:31	but to assign that statement to the whole group is clearly wrong.
00:03:38	Gathering good experimental data for statistical analysis is difficult.
00:03:43	Sampling bias is a bias in which a sample is collected in such a way that some members of the intended population
00:03:50	are less likely to be included than others. Samples are systematically different from the population.
00:03:58	Researchers try to combat the effect of bias by using double-blind randomized comparative experiments.
00:04:05	This is reflected in a field of study known as in statistics as the design of experiments.
00:04:10	In 2007, Colgate was ordered by the UK Advertising Standards Authority
00:04:15	to abandon a claim that more than 80% of dentists recommend Colgate.
00:04:22	The claim, which was based on studies of dentists and hygienists carried out by the manufacturer,
00:04:28	was found to be misrepresentative as it allowed the participants to select
00:04:33	one or more toothpaste brands. The ASA stated that
00:04:37	this would "be understood by readers to mean that 80% of the dentists recommend Colgate
00:04:44	over and above other brands, and the remaining 20% would recommend different brands."
00:04:50	Because we understood that another competitor's brand was recommended almost as much as the Colgate brand
00:04:57	by the dentists surveyed, we concluded that the claim misleadingly
00:05:02	implied 80% of dentists recommend Colgate in preference to all other brands.
00:05:08	On election night of the 1948 presidential election, the Chicago Tribune printed the headline
00:05:15	"DEWEY DEFEATS TRUMAN." Truman won!
00:05:19	In the morning, the grinning president-elect, Harry S. Truman,
00:05:23	was photographed holding a newspaper bearing this headline. The reason the Tribune was mistaken
00:05:29	was due to the results of a biased phone survey. The sample of telephone users was not representative
00:05:37	of the general population. Telephones were not yet widespread,
00:05:42	and those who had them tended to be prosperous with stable addresses.
00:05:49	If you want to know how one million people feel about a topic, it's impractical to ask all of them.
00:05:56	Therefore, you choose to get a random sample of say 1000 people,
00:06:00	and you can be fairly certain that the results given by the sample
00:06:04	are representative of what the larger group would have said if they had all been asked.
00:06:11	This confidence can be quantified and is the "plus or minus" figure
00:06:16	often quoted for statistical surveys. For example, a survey might have an estimated error
00:06:22	of plus or minus 5% at 95% confidence. The smaller the estimated error,
00:06:28	the larger the required sample, at a given confidence level.
00:06:32	We'll look at this in more detail later in this course. Many people might assume that
00:06:37	if the confidence figure is omitted, then there is a 100% certainty
00:06:43	that the true result is within the estimated error. Of course, this is not mathematically correct.
00:06:50	Also the randomness of the sample is very important, because non-random sampling
00:06:56	makes the estimated error unreliable. Again, for example, many opinion polls
00:07:02	are conducted by phone, so the sample could be distorted in a number of ways:
00:07:07	It could be excluding people who don't have phones. It could be favoring people who have more than one phone.
00:07:15	Favoring people who are willing to participate in a phone survey over those who won't.

00:07:21	Sometimes only one margin of error is reported for a survey. However, if results are reported for a subgroup,
00:07:28	then a larger margin of error will apply, and this may not be made clear.
00:07:35	For example, a survey of 1000 people may contain 50 people from a certain ethnic group.
00:07:41	The results focusing specifically on that group will be much less reliable
00:07:46	than results for the full population. Therefore, the margin of error for that group
00:07:50	is much higher than for the total 1000 people. If a statistical test shows a correlation between X and Y,
00:08:00	there could be six possibilities: X causes Y.
00:08:03	Y causes X. X and Y both partly cause each other.
00:08:07	X and Y are both caused by a third factor. Y is caused by Z which is correlated to X.
00:08:14	Or the observed correlation was due to chance. An example of false causality
00:08:20	was widely reported a few years ago. There was concern that there was a link
00:08:25	between electromagnetic fields from power lines and cancer. However, there is now an alternative theory for this.
00:08:33	If there is a perception that a geographical location is dangerous,
00:08:38	even if it really isn't, then the property values in that area will decrease.
00:08:44	Secondly, low-income families will then move into that area. On top of that unfortunately, low-income families
00:08:51	are more likely to get cancer than high-income families.
00:08:56	This may be due to poorer diets or having less access to medical care.
00:09:00	Fourthly, and therefore, rates of cancer will increase, even though the electromagnetic fields
00:09:06	are not dangerous at all. In well-designed studies,
00:09:09	the effect of false causality can be eliminated by assigning some people into a "treatment group"
00:09:16	and some people into a randomized "control group". The "treatment group" is given the treatment,
00:09:23	and the "control group" is not given the treatment. You will look at correlation and causation
00:09:29	in more detail later in this course. Statistical significance is concerned
00:09:36	with whether a research result is due to chance or sampling variability.
00:09:41	Statistical significance refers to the unlikelihood that mean differences observed in a sample
00:09:47	have occurred due to sampling error. We'll be looking at the Null Hypothesis
00:09:51	in more detail later in this course. Given a large enough sample,
00:09:57	despite seemingly insignificant population differences, one might still find statistical significance.
00:10:04	Practical significance is concerned with whether the result is large enough to be of value
00:10:09	in the real world. A general problem with traditional statistics is that
00:10:14	if you take large enough samples, almost any difference or any correlation
00:10:20	will be significant. And due to this problem,
00:10:23	published statistics should include some information surrounding the practical significance
00:10:29	of the findings. If a survey on participation in sports
00:10:34	by boys and girls at school found that 60% of boys and 58% of girls
00:10:39	participated in outdoor sports, then there is a 2% difference between the boys and girls.
00:10:45	However, how much significance has this 2% difference statistically as well as practically?
00:10:52	The statistical significance of this 2% depends upon the size of data used
00:10:57	in determining the percentage of boys and girls. If a sufficiently large sample size is used
00:11:04	then the difference is statistically significant, and if a very small sample size is used,

00:11:09 then the difference is statistically insignificant. The sample size is an important factor
00:11:15 in determining the statistical significance of a computed figure.
00:11:21 The practical significance of this 2% will determine if the decision needs to be made
00:11:27 and action made as a result of this. If students could be promoted to participate in sports
00:11:32 to encourage more gender parity and so on, in outdoor sports.
00:11:35 Therefore, in this case the 2% difference though relatively small,
00:11:39 may be practically significant or not. You need to know how to detect faulty data
00:11:45 when reading statistically significant study results. And keep the following warning signs in mind:
00:11:52 Check the sample size used to obtain the study results. If the study is based on a very large sample size,
00:11:59 relationships found to be statistically significant may not have much practical significance.
00:12:05 Of course, the media tends to report only the eye-catching results.
00:12:11 Therefore, be skeptical of reports where many tests were conducted,
00:12:15 but where the results of only a small number of those studies are then presented as significant.

00:12:24 Data dredging, sometimes called "data fishing", "data snooping" and "p-hacking"
00:12:29 is the misuse of data analysis to find patterns in data that can be presented
00:12:34 as statistically significant, when in fact there is no real underlying effect.
00:12:41 The confidence interval to establish a relationship between two parameters is often chosen to be 95%.
00:12:48 This means that there is a 95% chance that the relationship observed is not due to random chance.
00:12:56 Therefore, there is a 5% chance of finding a correlation
00:12:58 between any two sets of completely random variables. Often extremely large datasets with many variables
00:13:07 will be analyzed, so spurious but apparently statistically significant results
00:13:13 can often be found. "P-hacking" is when a data scientist analyses
00:13:18 and presents the data in a way that supports already preconceived answers.
00:13:22 They know that by selectively munging, binning, constraining, cleansing, and sub-segmenting data,
00:13:30 they can get it to tell almost any story or validate almost any "fact".
00:13:35 One out of every 20 significant results might be random if you rely solely on statistical analysis.

00:13:42 This means that if you measure enough variables, eventually it will appear that some of them correlate.
00:13:50 Therefore, studies can be manipulated with enough data
00:13:53 to prove a correlation that does not exist, or that is not significant enough to prove causation.
00:14:01 Beware of correlation hunting. "P-hacking" is a reference to the p-value,
00:14:06 which is a measure of statistical significance. The p-value is the level of marginal significance

00:14:12 within a statistical hypothesis test. It represents the probability of the occurrence
00:14:18 of a given event. A small p-value of less than 0.05
00:14:24 means that there's stronger evidence in favor of the alternative hypothesis.
00:14:28 The hacking of a p-value can sometimes inadvertently happen through a statistical practice
00:14:34 known as overfitting, where an analytical model is excessively complex,
00:14:41 where there are too many explanatory variables in the model, relative to the number of observation points.

00:14:51	To summarize, we expect statistics should make data easier
00:14:55	for us to understand. Unfortunately, it's easy for statistics to be used
00:15:00	in a misleading way, to trick the casual observer
00:15:03	into believing something other than what the data shows. This misuse of statistics occurs
00:15:10	when a statistical argument asserts a falsehood. In some cases, the misuse may be accidental.
00:15:17	In others, it is purposeful and is designed to trick us into believing a lie.
00:15:24	This presentation will hopefully help you recognize the common forms of misuse.
00:15:30	In the next unit, we'll look at bias.

Week 1 Unit 4

00:00:05 Hello and welcome to the fourth unit in week one, where we're going to look at bias.
00:00:12 Bias refers to the tendency of a measurement process to over
00:00:17 or under-estimate the value of a population parameter.
00:00:22 In survey sampling, for example, bias would be the tendency
00:00:25 of a sample statistic to systematically over or under-estimate
00:00:30 a population parameter. There are many different types of bias.
00:00:35 It can be introduced at the stage of data collection, with surveys and questionnaires.
00:00:41 It can be bound up with the approach we take to analyze the data.
00:00:46 There are also a whole set of cognitive biases which influence how we as humans
00:00:52 interpret the data. Throughout this course, we will introduce you
00:00:58 to ways in which bad practice, deliberate or not, can lead
00:01:02 to a particular unjustified result. Many of these relate to the various types of bias
00:01:08 that can be introduced into your statistical analysis.
00:01:13 In this unit, we will pull these all together so that you can become more aware
00:01:17 of the opportunities for bias. The topic is a huge one,
00:01:21 but here we will give you a taster of the big areas of potential bias.
00:01:27 In their book, Common Errors in Statistics (and How to Avoid Them), Good and Hardin say
00:01:33 "With careful and prolonged planning, we may reduce or eliminate
00:01:38 many potential sources of bias, but seldom will we be able
00:01:42 to eliminate all of them. Accept bias as inevitable
00:01:47 and then endeavor to recognize and report all exceptions
00:01:50 that do slip thought the cracks." We will look at some
00:01:54 of the different types of bias, and these will include
00:01:57 selection or sampling bias, self-selection bias,
00:02:03 confirmation bias, and overfitting.
00:02:07 It's very important that samples are collected in an unbiased way
00:02:12 so that we can ensure there is no built-in bias.
00:02:16 The goal is to make sure that every member of the population that we are surveying
00:02:23 has an equal chance of actually being selected. So think about the potential problems
00:02:29 with the following examples, and who might be excluded.
00:02:34 Surveying on attitudes to Brexit on Facebook will obviously favor younger audiences.
00:02:40 Face-to-face surveying on attitudes to equality at a football match
00:02:44 might include self-selection. And telephone surveying
00:02:48 on holiday preferences might include some bias.
00:02:52 So there are various techniques that can be used to mitigate the risk of biased samples,
00:02:57 and these include randomizing or stratifying the data.
00:03:02 If you use data taken from a voluntary response sample,
00:03:07 for example, where the participants have actually volunteered to take part,
00:03:12 it becomes really difficult to avoid bias, where the self-selected group
00:03:16 contains more participants with a particular set of beliefs
00:03:21 about your study. For example, if you do a telephone survey
00:03:27 based on an area code or postal code, you're going to miss those increasing numbers
00:03:33 who only have mobile phones or those who choose

00:03:36 to restrict calls to their phones. What could be distinctive
00:03:41 about that subset of that population that could be really important
00:03:46 and you're missing? Confirmation bias is one
00:03:51 of a range of cognitive biases which affect how we read and interpret
00:03:57 the insights we think we have found. Cognitive bias means that it's inbuilt
00:04:02 into us as humans and how we actually think. Confirmation bias reflects our tendency
00:04:10 to pick out those parts of the data and, of course, the information within the data
00:04:15 in a way to support our previously held beliefs.
00:04:19 Confirmation bias therefore prevents us from being objective with the data
00:04:25 we have collected and how we actually interpret it. For example, we might have a tendency
00:04:31 to identify and remember events where a belief about another group,
00:04:36 for example, gender or ethnic, sports fans or work teams, and so on,
00:04:41 actually confirms our beliefs, and we don't notice evidence to the contrary.
00:04:48 In predictive modeling, one of the most important topics
00:04:52 is underfitting and overfitting the predictive model.
00:04:57 These are important concepts because they actually explain the state
00:05:02 of a model based on its performance. The term overfitting refers to a model
00:05:08 that has a close fit to the data with which it was trained,
00:05:12 but it doesn't generalize, meaning that, when we actually come to use it
00:05:17 and forecast values other than those we use for training,
00:05:21 it actually predicts these with very high error. Conversely, underfitting means that the model
00:05:28 doesn't fit even the training data very well. Overinterpreting poorly fitted
00:05:34 or overfitted model results can mean that you actually develop conclusions
00:05:39 that are not necessarily justified by the data. And this can be avoided by using
00:05:45 other sources of data to validate your conclusions.
00:05:51 So to summarize. It's almost impossible to avoid bias
00:05:55 in its various forms, but an awareness of bias
00:05:59 can help mitigate its worst effects. There are various forms of bias,
00:06:03 for example, technical, cognitive, and others, which impact what data to collect
00:06:09 and how it should be interpreted. In this unit, we've examined
00:06:14 four key examples of bias, including sampling, self-selection,
00:06:19 confirmation, and overfitting. There are many other types of bias
00:06:24 that you could read about. In the next unit, we're going to be looking
00:06:28 at different kinds of analytic approaches.

Week 1 Unit 5

00:00:05	Hello and welcome back to the fifth and final unit in week one, where we'll be looking
00:00:11	at different kinds of analytic approaches. The type of analytical approach you take really depends
00:00:20	on the type of data that you've collected and kind of the question that you're trying to answer.
00:00:27	There are two types of data: qualitative and quantitative. Qualitative data consist of words or narratives.
00:00:36	The analysis of this type of data includes highlighting keywords and the identification of themes.
00:00:45	For example, data captured from a focus group to understand the participants' perceptions:
00:00:54	The data could be in freeform, so it's a narrative, and you use qualitative techniques
00:01:00	to identify content and identify themes. The data deals with descriptions
00:01:05	and can be observed, but not measured. Some examples include colors, textures, smells, tastes, and so on.
00:01:17	Quantitative data are numerical and the analysis will involve statistical techniques.
00:01:23	For example, if you analyze a satisfaction survey where participants rated their experience
00:01:31	on a scale of one to 10, the data is numeric in form, and it can be measured.
00:01:38	You use statistical techniques to draw conclusions about participants' satisfaction,
00:01:44	and let's go through some examples. These can include length, height, area, speed, and time.
00:01:53	There are two common types of analysis that are referred to as descriptive and inferential.
00:02:01	Descriptive analysis informs you about the basic qualities of the data.
00:02:07	It includes basic descriptive statistics, such as the range, minimum value, maximum value, and frequency.
00:02:17	It also includes measures of central tendency, such as mean, median, mode, and standard deviation.
00:02:27	It tells you what the data look like, and it helps you to simplify and to summarize data,
00:02:34	as well as describe and visualize that data. Inferential analysis uses statistical techniques to analyze
00:02:41	whether a pattern in the data is due to chance or due to the intervention that is being observed.
00:02:49	And what the strength of that relationship is can also been in that.
00:02:53	So the first step is to understand the data distribution. Is it normal or non-normal?
00:03:00	Next, if the data are normally distributed, you will generally choose from a range of parametric tests.
00:03:08	However, if the data are non-normally distributed, you will choose from the set of nonparametric tests.
00:03:17	You will analyze samples of data and generalize the results to the whole population.
00:03:26	You will undertake hypothesis testing using statistical testing
00:03:30	and possibly make predictions using, for example, regression techniques or others.
00:03:37	You'll learn more about these later on. For inferential statistics,
00:03:42	you need to understand the data distribution. Is it normal or non-normal?
00:03:48	On the left side, you will see a normal distribution. It looks like the bell curve.
00:03:54	The majority of the data is clustered around one number or value in the middle.
00:04:00	Usually, when the data are normal, you will choose from statistical tests
00:04:06	called parametric tests. On the right, you will see a non-normal distribution.
00:04:13	There are several ways a distribution can be non-normal, and you'll learn a little bit more

00:04:18 about these later in the course. This can happen with a small sample size

00:04:24 or an unusual set of responses, for example. Usually, if the data is non-normal, you will choose

00:04:32 from statistical tests called nonparametric tests. You will learn more about these later on in the course.

00:04:43 A statistical hypothesis is an assumption about a population parameter that may or may not be true.

00:04:51 Statisticians use a formal procedure called hypothesis testing to accept or reject these hypotheses.

00:05:00 You will learn about this in more detail later in this course.

00:05:04 There are a wide range of statistical tests that you can use to test a hypothesis.

00:05:10 Some of the common ones are listed here. You will see some of the common tests for correlation,

00:05:19 the comparison of means of variables, and for regression, which analyzes how the change

00:05:26 in one variable predicts the change in another. The tests presented in this slide are called

00:05:34 parametric tests and are based on certain assumptions. For example, when running tests of hypothesis for means

00:05:44 of continuous outcomes, all parametric tests assume that the outcome is approximately

00:05:51 normally distributed in the population. Please note that this does not mean that the data

00:05:59 in the observed sample follows a normal distribution, but, rather, that the outcome follows a normal distribution,

00:06:09 and that is within the full population, which is not necessarily observed in the outcome.

00:06:14 For many outcomes, you may be comfortable with the normality assumption: that is, most of the observations

00:06:21 are in the center of the distribution, while fewer are at either extreme.

00:06:27 Also, many statistical tests are robust, which means that they maintain their statistical properties,

00:06:35 even when assumptions are not entirely met. When the sample size is small,

00:06:44 and the distribution of the outcome is not known and cannot be assumed to be approximately

00:06:51 normally distributed, then alternative tests called nonparametric tests are appropriate.

00:06:59 Nonparametric or distribution-free tests mean the test doesn't assume the data comes from

00:07:06 a particular distribution, like the normal distribution we had mentioned before.

00:07:13 It can sometimes be difficult to assess whether a continuous outcome follows a normal distribution

00:07:22 and whether a parametric or a nonparametric test is the appropriate one.

00:07:28 The most practical approach to assessing normality involves analyzing the distribution of the outcome

00:07:36 in the sample using a histogram. Nonparametric tests are sometimes called

00:07:43 distribution-free tests because they're based on fewer assumptions.

00:07:49 They do not assume that the outcome is approximately normally distributed.

00:07:56 However, parametric tests involve specific probability distributions,

00:08:00 for example, the normal distribution, and the tests involve estimation of the key parameters

00:08:07 of that distribution: for example, the mean or difference in the means from the sample data.

00:08:15 There are also several statistical tests that can be used to assess whether data

00:08:21 are likely to be coming from a normal distribution, and each test is essentially a goodness-of-fit test

00:08:27 and compares the observed data to quantiles of the normal distribution

00:08:32 for other specified distributions. To summarize, descriptive analysis informs you

00:08:41	about the basic qualities of the data. Inferential analysis uses statistical tests to analyze
00:08:50	whether a pattern in the data is due to chance or due to the intervention that is being observed
00:08:58	and what the strength of that relationship is. In this course, you'll learn about some of these descriptive
00:09:07	and inferential statistical techniques, and how these techniques can be applied and misused.
00:09:17	With this, I'd like to close the first week. I hope you enjoyed the first units of this course
00:09:22	and we're happy to get in touch with you in our discussion forum if you have
00:09:28	any content-related questions. Now, we wish you all the best for the upcoming weeks.
00:09:35	We hope you enjoy the course, and you enjoy doing the assignments
00:09:40	going forward, week by week. Bye for now.

www.sap.com/contactsap

THE BEST RUN