openSAP
Introduction to Statistics for Data Science

Week 5 Unit 1

00:00:05	Hello, and welcome back to week five of the openSAP course, Introduction to Statistics for Data Science.
00:00:14	In this unit, we are going to look at properties of distributions.
00:00:20	A probability distribution is a mathematical function that provides the probabilities of occurrence
00:00:27	of different possible outcomes in an experiment. So, for an example, if the random variable x
00:00:35	is used to denote the outcome of a coin toss, which is the experiment,
00:00:40	then the probability distribution of x would take the value 0.5 for heads,
00:00:46	and 0.5 for tails, if the coin is fair. A probability distribution is the set
00:00:51	of all possible outcomes of the random phenomenon being observed.
00:00:59	Probability distributions are generally divided into two classes: discrete probability distributions,
00:01:06	where the set of possible outcomes is discrete, so for example, rolling a dice or tossing a coin.
00:01:12	And continuous probability distributions, where the set of possible outcomes
00:01:18	can take on values in a continuous range, for example, temperature over a day,
00:01:24	and for example, one of these would be the normal distribution,
00:01:27	which is a commonly encountered continuous distribution. Discrete probability functions
00:01:34	are also known as probability mass functions. These can assume a discrete number of values,
00:01:43	so for example, you can have only heads or tails in a coin toss.
00:01:48	Similarly, if you're counting the number of cars sold per day by a car sales company,
00:01:55	you can count 10 or 11 cars but nothing in between. For discrete probability distribution functions,
00:02:03	each possible value has a non-zero likelihood. The probabilities for all possible values must sum to one.
00:02:12	For example, the likelihood of rolling a specific number on a dice is one divided by six.
00:02:18	The total probability for all six values equals one. When you roll a die, you will definitely obtain
00:02:25	one of the six possible values. There are a variety of discrete probability distributions
00:02:31	that you can use to model different types of data. The correct discrete distribution
00:02:38	depends on the properties of your data. So for example, use the binomial distribution
00:02:45	to model binary data, for example, a coin toss;
00:02:49	the Poisson distribution to model count data, for example, the count of cars sold per day;
00:02:56	the uniform distribution to model multiple events with the same probability,
00:03:01	for example, rolling a dice. In the diagram, the probability mass function
00:03:07	specifies the probability distribution for the sum of counts from two dice.
00:03:15	For example, the figure shows that the probability of throwing an 11 is two divided by 36 or 1/18.
00:03:25	This computation of probabilities of events is, for example, the probability of throwing the dice
00:03:33	with a combined value greater than nine. This means you add together the probability

00:03:39	for dice combinations of 10, 11, and 12, this is equal to 1/12 plus an 1/18 plus a 1/36.
00:03:51	Suppose you flip a coin two times. This simple statistical experiment
00:03:56	can have four possible outcomes, as shown on the slide. Let the random variable x
00:04:03	represent the number of heads that result from this. The random variable can only take on the values
00:04:12	zero, one, or two. So it's a discrete random variable.
00:04:17	The probability distribution for this statistical experiment is shown in the table.
00:04:23	You can see that the table represents a discrete probability distribution
00:04:28	because it relates each value of a discrete random variable
00:04:34	with its probability of occurrence. Continuous probability functions
00:04:41	are also known as probability density functions. Sometimes they are referred to as a density function or a pdf.
00:04:52	Probabilities for continuous distributions are measured over ranges of values,
00:04:58	rather than single points. Therefore, a probability indicates the likelihood
00:05:03	that the value will fall within an interval. In a continuous distribution,
00:05:09	the variable can assume an infinite number of values between any two values.
00:05:15	For example, continuous variables are often measurements on a scale,
00:05:20	such as temperature, height, and weight. Specific values in continuous distributions
00:05:28	can have a zero probability, unlike discrete probability distributions.
00:05:35	On the probability plot, the entire area under the distribution curve equals one.
00:05:42	The proportion of the area under the curve that falls within a range of values
00:05:48	along the x-axis represents the likelihood that a value will fall within that range.
00:05:55	Each continuous probability distribution has parameters that define its shape.
00:06:02	When you specify these parameters, they establish the shape of the distribution
00:06:07	and all of its probabilities. The parameters represent essential properties
00:06:14	of the distribution, such as the central tendency and the variability.
00:06:20	The most commonly encountered continuous distribution is the normal distribution,
00:06:26	which is also known as the Gaussian distribution or the bell curve.
00:06:31	This symmetric distribution fits a wide variety of phenomena,
00:06:35	such as human height and IQ scores. It's defined by two parameters:
00:06:42	the mean and the standard deviation. The Weibull distribution
00:06:48	and the lognormal distribution are other commonly encountered continuous distributions.
00:06:54	Both of these distributions can fit skewed data. The diagram shows the probability density function
00:07:02	of the normal distribution. The probabilities of intervals of values
00:07:08	correspond to the area under the curve. For example, consider the probability density function
00:07:18	shown in the graph. Suppose you wanted to know the probability
00:07:23	that the random variable x was less than or equal to a. The probability that x is less than or equal to a
00:07:32	is equal to the area under the curve bounded by a and minus infinity,
00:07:38	as indicated by the shaded area. In general, for a continuous probability distribution,
00:07:45	the density function has the following properties: Since the continuous random variable
00:07:51	is defined over a continuous range of values, called the domain of the variable,
00:07:58	the graph of the density function will also be continuous over that range.
00:08:05	The area bounded by the curve of the density function and the x-axis is equal to one,
00:08:12	when computed over the domain of the variable. The shaded area in the graph

00:08:19	represents the probability that the random variable x is less than or equal to a.
00:08:26	This is a cumulative probability. However, the probability that x is exactly equal
00:08:34	to a would be zero. A continuous random variable can take on
00:08:38	an infinite number of values. The probability that it will equal a specific value,
00:08:45	such as a, is always zero. The probability that a random variable
00:08:53	assumes a value between a and b is equal to the area under the density function
00:09:00	bounded by a and b. Assume that the distribution of IQ scores
00:09:06	in a school is defined as a normal distribution with a mean of 100 and a standard deviation of
00:09:13	You want to determine the likelihood that an IQ score will be between 120 and 140.
00:09:21	The probability plot is a symmetric distribution with the most frequent values occurring around
00:09:29	which is the mean. The probabilities reduce as you move away
00:09:35	from the mean in both directions. The shaded area for the range of IQ scores
00:09:41	between 120 and 140 contains nearly 14% of the total area under the curve.
00:09:48	Therefore, the likelihood that an IQ score falls within this range is 0.14.
00:09:56	There are three main differences between a continuous
00:10:00	and a discrete probability distribution. Firstly, the probability that a continuous variable
00:10:08	will take a specific value is equal to zero. For example, the likelihood of measuring a temperature
00:10:16	that is exactly 25 degrees Celsius is zero. This is because the temperature can be an infinite number
00:10:24	of other temperatures that are infinitesimally higher or lower than 25.
00:10:30	So, statisticians say that an individual value has an infinitesimally small probability
00:10:37	that is equivalent to zero. Secondly, because of this,
00:10:42	continuous probability distributions are not displayed in a tabular form.
00:10:48	And thirdly, a graph with specified parameters, for example, the mean and the standard deviation,
00:10:55	are used to describe continuous distributions. The graph is called the probability density function.
00:11:03	So to summarize: A probability distribution
00:11:08	is a mathematical function that provides the probabilities of occurrence
00:11:14	of different possible outcomes in an experiment. A discrete random variable
00:11:20	can take only a finite number of different values like zero, one, two, three, four,
00:11:26	whereas a continuous random variable is a variable that can take an infinite number
00:11:31	of possible values. Discrete probability functions
00:11:36	are also known as probability mass functions, and they can assume a discrete number of values.
00:11:47	Continuous probability functions are also known as probability density functions,
00:11:54	and the probabilities are measured over ranges of values, rather than single points.
00:12:01	In the next unit, we will consider the normal distribution in more detail.

Week 5 Unit 2

00:00:05	Hello, and welcome back to week five, unit two of the openSAP course,
00:00:10	Introduction to Statistics for Data Science. In this unit, we'll look at the normal distribution
00:00:17	in more detail. In many cases, data tends to a central value,
00:00:23	with no bias left or right. This is called a normal distribution.
00:00:28	The normal distribution is often called a bell curve because it looks like a bell,
00:00:33	or referred to as the Gaussian or Gauss-Laplace distribution. It's a very common continuous probability distribution.
00:00:44	In the slide, you can see an example where the yellow histogram shows some data
00:00:50	that follows the normal distribution, not perfectly, but closely.
00:00:56	Normal distributions are important in statistics and often used in the natural and social sciences
00:01:03	to represent real values, and real-valued random variances,
00:01:08	whose distributions are not known. It is a theoretical distribution
00:01:15	with the mean, median, and mode positioned at the same point,
00:01:21	which is the exact center of the distribution. It's a unimodal frequency distribution curve,
00:01:28	a bell shape with a single peak in the center. This means most of the values are clustered in the center,
00:01:38	around the mean or median. It's symmetrical about the mean,
00:01:43	with half of the distribution on each side of the mean. The total area under the normal distribution
00:01:53	is equal to 100%. It's asymptotic, meaning the two tails of the curve
00:02:00	fall and extend indefinitely in both directions, but never touching the x-axis. Thus, it has infinite range.
00:02:10	The location of a normal distribution is determined by the mean and the spread.
00:02:15	And the spread is determined by the standard deviation. Distance away from the mean is measured
00:02:22	in standard deviations, also known as z-scores. You've learned that the standard deviation
00:02:32	is a measure of how spread out numbers are. The standard deviation enables you
00:02:40	to say that any value is likely to be within one standard deviation, so 68 out of 100;
00:02:46	very likely to be within two standard deviations, which would be 95 out of 100;
00:02:52	or almost certainly within three standard deviations, representing 997 out of 1,000.
00:03:00	The number of standard deviations from the mean is also called the standard score, sigma, or z- score.
00:03:10	To convert a value to a standard score, the z- score, subtract the mean, divide by the standard deviation.
00:03:19	This is called standardizing, and the formula is shown on the slide.
00:03:26	Here is an example using the standard, normal distribution. In a recent data science test, you did really well,
00:03:35	and scored 1.5 standard deviations above the average. How many students scored lower than you?
00:03:46	From the graph you can see, that between zero and 1.5 standard deviations,
00:03:51	the percentage population is 19.1, plus 15, plus which equals 43.3%.
00:03:59	Less than zero is 50%, the left half of the curve. Therefore, in theory, the total less than yours
00:04:06	is 50% plus 43.3, which is 93.3%. That's a very good result.
00:04:14	The empirical rule states that for a normal distribution, nearly all of the data will fall

00:04:19	within three standard deviations of the mean. The rule is also called the 68-95-99.7 rule,
00:04:28	or the three sigma rule. The empirical rule is often used in statistics
00:04:33	for forecasting, especially when obtaining the right data is difficult
00:04:38	or impossible to get. The rule can give you a rough estimate
00:04:43	of what your data collection might look like if you were able to survey the entire population.
00:04:52	This rule applies, generally, to a random variable, x, following the shape of a normal distribution.
00:04:59	The rule doesn't apply to distributions that are not normally distributed,
00:05:05	but you can apply it to other kinds of distributions using Chebyshev's theorem.
00:05:13	The z-score can be used to indicate if a measurement is deemed to be an outlier.
00:05:21	Observations with z-scores greater than three in absolute values are considered outliers.
00:05:27	For some highly skewed data sets, observations with z-scores greater than two
00:05:33	in absolute values may also be outliers. However, the presence of one or more outliers
00:05:39	in a data set can inflate the computed values of the standard deviation.
00:05:47	However, it is unlikely than an error observation would have a z-score larger than absolute three.
00:05:54	In a previous lesson, you were introduced to box plots. In contrast to z-scores, the values of the core tiles
00:06:02	used to calculate the intervals for a box plot are not affected by the presence of outliers.
00:06:12	In an experiment, suppose that a sample is obtained containing a large number of observations,
00:06:19	where each observation is randomly generated in a way that does not depend on the values
00:06:25	of the other observations. And that the arithmetic average of the observed values
00:06:32	is calculated. If this procedure is performed many times,
00:06:37	the central limit theorem says that the distribution of the average,
00:06:41	will be closely approximated by a normal distribution. The central limit theorem
00:06:48	establishes that when independent random variables are added,
00:06:53	their properly normalized sum tends towards a normal distribution,
00:06:58	even if the original variables themselves are not normally distributed.
00:07:05	The theorem is a key – that is, central – concept, because it applies that probabilistic
00:07:11	and statistical methods that work for normal distributions can be applicable to many problems
00:07:20	involving other types of distributions. A simple example of this is that if you flip a coin
00:07:28	many times, the probability of getting a given number of heads
00:07:33	in a series of flips will approach a normal curve, with a mean equal to half the total number of flips
00:07:40	in each series. In summary, the normal distribution is a very commonly
00:07:48	encountered continuous probability distribution. The characteristics of the normal distribution are:
00:07:57	mean equals median equals mode. Symmetry about the center.
00:08:02	50% of values less than the mean and 50% greater than the mean.
00:08:08	When we calculate the standard deviation, we find that generally, 68% of values
00:08:14	are within one standard deviation. 95% of values are within two standard deviations
00:08:22	of the mean. And finally, 99.7% of the values
00:08:29	are within three standard deviations of the mean. The empirical rule states that for a normal distribution,
00:08:36	nearly all of the data will fall within three standard deviations of the mean.
00:08:43	The central limit theorem establishes that when independent random variables are added,

00:08:51 their properly normalized sum tends towards a normal distribution.
00:08:55 Even if the original variables themselves are not normally distributed.
00:09:03 In the next unit, we'll consider kurtosis and skewness.

Week 5 Unit 3

00:00:05	Hi, and welcome back to week five, unit three of this openSAP course,
00:00:11	Introduction to Statistics for Data Science. In this unit, we're going to look at kurtosis and skewness.
00:00:20	Kurtosis is a measure of the tailedness of the probability distribution.
00:00:26	It's a descriptor of the shape of a probability distribution.
00:00:33	For any univariate normal distribution, it has a value of three.
00:00:38	It's common to compare the kurtosis of other distributions to the value for a normal distribution.
00:00:47	Data sets with high kurtosis tend to have heavy tails or outliers.
00:00:52	This means that there are more cases far from the mean than is found in a normal distribution.
00:00:59	Distributions with kurtosis greater than three are said to be leptokurtic.
00:01:05	Data sets with low kurtosis tend to have light tails or lack of outliers.
00:01:12	This means that there are fewer cases in the tails than would be expected in a normal distribution.
00:01:19	Distributions with kurtosis less than three are said to be platykurtic.
00:01:24	In terms of shape, a leptokurtic distribution has fatter tails.
00:01:29	Examples of leptokurtic distributions include the student's t-distribution, the Rayleigh distribution,
00:01:36	the Laplace distribution, exponential distributions, Poisson distributions, and the logistic distribution.
00:01:45	In terms of shape, a platykurtic distribution has thinner tails.
00:01:51	Examples of platykurtic distributions include the continuous
00:01:55	and discrete uniform distributions and the raised cosine distribution.
00:02:01	The most platykurtic distribution of all is the Bernoulli distribution.
00:02:09	Distributions with zero excess kurtosis are called mesokurtic.
00:02:16	The most prominent example of a mesokurtic distribution is the normal distribution family
00:02:22	regardless of the values of its parameters. Excess kurtosis is a measure of how the distribution's tails
00:02:33	compare to the normal distribution. It's usually defined as kurtosis minus three.
00:02:39	Excess kurtosis for the normal distribution is zero, three minus three.
00:02:45	Negative excess equals higher tails than the normal distribution.
00:02:50	Positive excess equals heavier tails than the normal distribution.
00:02:55	This graph here shows a variety of distributions. Note how the tails are fatter
00:03:01	or thinner than the normal, shown in black. Kurtosis has real life applications,
00:03:08	especially in the world of economics. Fund managers usually focus on risks
00:03:14	and returns and this can be indicated by kurtosis. A leptokurtic return means that risks
00:03:21	are coming from outlier events. This would be a stock for investors
00:03:26	willing to take extreme risks. For example, in real estate with a high kurtosis
00:03:33	and high-yield U.S. bonds, these are high-risk investments. Investment-grade U.S. bonds
00:03:40	and small capitalized U.S. stocks would be considered much safer investments
00:03:45	with lower kurtosis. Skewness is a measure of the asymmetry
00:03:52	of a probability distribution. A distribution is symmetric if it looks the same to the left
00:03:58	and right of the center point. We briefly looked at skewness in a previous lesson.
00:04:04	In a normal distribution, the graph appears as a classical, symmetric bell-shaped curve.

00:04:11	The mean and the mode are equal. This is shown by the green solid curve in the diagram.
00:04:18	The tails on either side of the curve are exact mirror images of each other.
00:04:24	If the distribution is symmetric, then the mean is equal to the median,
00:04:29	and the distribution has zero skewness. If the distribution is both symmetric
00:04:37	and unimodal, then the mean equals the median equals the mode.
00:04:43	When a distribution is skewed to the left, shown by the red dashed curve here,
00:04:50	the tail on the curve's left-hand side is longer than the tail on the right-hand side,
00:04:57	and the mean is less than the mode. This is also called negative skewness.
00:05:03	When a distribution is skewed to the right, shown by the blue dotted curve,
00:05:09	the tail on the curve's right-hand side is longer than the tail on the left-hand side,
00:05:15	and the mean is greater than the mode. This is also called positive skewness.
00:05:23	The skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero.
00:05:32	Negative values for the skewness indicate data that are skewed left,
00:05:37	and positive values for the skewness indicate data that are skewed right.
00:05:45	For distributions that are unimodal or with tails of similar weight
00:05:50	or for discrete distributions with areas to the left and the right of the median that are similar,
00:05:58	then the relationship between the mean and the median is very insightful.
00:06:04	This is shown in the histograms in the diagram. If most of the data are on the left side of the histogram,
00:06:12	but a few larger values are on the right, the data are said to be skewed to the right.
00:06:18	It's a positive skew. Histogram A shows an example of data
00:06:23	that are skewed to the right. The few larger values bring the mean upwards,
00:06:31	but they don't really affect the median, so when data are skewed right,
00:06:36	the mean is often larger than the median. An example of right-skewed data
00:06:42	would be for football team salaries, where star players make a lot more than their teammates.
00:06:50	Other common examples for right, positive, skewness include people's incomes, mileage on used cars,
00:06:59	reaction times in a psychology experiment, house prices, the number of accident claims by an insurance customer,
00:07:07	and the number of children in a family. If most of the data are on the right
00:07:13	with a few smaller values showing up on the left of the histogram, the data are skewed to the left,
00:07:21	which is a negative skew. Histogram B in the figure shows an example
00:07:27	of data that are skewed to the left. The few smaller values bring the mean down,
00:07:34	and, again, the median is minimally affected, if at all. When data are skewed left,
00:07:41	the mean is often smaller than the median. An example of skewed-left data is the amount of time
00:07:48	students use to take an exam. Some students leave early, more of them stay later,
00:07:55	and many stay until the end. There are fewer real-world examples
00:08:00	of left, negative, skewness. However, age at death is negatively skewed
00:08:07	in developed countries. Skewed data often occur due to lower
00:08:12	or upper bounds on the data – that is, data that have a lower bound are often skewed right
00:08:20	while data that have an upper bound are often skewed left. Skewness can also result from startup effects.
00:08:28	So for example, in reliability applications, some processes may have a large number of initial failures

00:08:36 that could cause left skewness. On the other hand, a reliability process

00:08:43 could have a long startup period, where failures are rare, resulting in right-skewed data.

00:08:50 If the data are symmetric, they have about the same shape on either side of the middle.

00:08:56 In other words, if you fold the histogram in half, it looks about the same on both sides.

00:09:03 Histogram C in the figure shows an example of symmetric data.

00:09:09 With symmetric data, the mean and the median are close together. However, there is a very interesting quote regarding

00:09:19 the mean and the median that you should remember. Paul von Hippel wrote: "Many textbooks, teach a rule

00:09:28 of thumb stating that the mean is right of the median under right skew and left of the median under left skew."

00:09:37 This rule fails with surprising frequency. It can fail in multimodal distributions

00:09:43 or in distributions where one tail is long, but the other is heavy.

00:09:49 Most commonly, though, the rule fails in discrete distributions where the areas to the left

00:09:56 and right of the median are not equal. One of the assumptions for many parametric tests

00:10:04 to be reliable is that the data is approximately normally distributed.

00:10:10 Significant skewness and kurtosis clearly indicate that data are not normal.

00:10:18 You need to fix the positive skewness. Often the skewed variables are transformed by a function

00:10:25 that has a disproportionate effect on the tails of the distribution.

00:10:29 Ideally, for most modeling algorithms, the desired outcome of skew correction is a new version

00:10:37 of the variable that is normally distributed. For positive skew, the most common corrections

00:10:44 are the log transform, the multiplicative inverse, the square root transform.

00:10:53 And these work by reducing larger values more than the smaller values or, in the case of the inverse,

00:11:01 increasing the smaller values. Of these, the log transform is the most often used

00:11:07 transformation to correct for positive skew. To fix negative skewness,

00:11:14 this is obviously less common than positive skew, but has the same problems with bias that positive skew has.

00:11:23 For negative skew, the most common corrections are a power transform, so like the square, the cube,

00:11:31 or raising the variable to a higher power; a log transform, as described previously.

00:11:38 However, the log transform is undefined for negative values, so you must first ensure the values

00:11:44 are positive before applying the log, so it's fairly complicated.

00:11:50 In summary: Kurtosis is a measure of the tailedness of the probability distribution.

00:11:57 Data sets with high kurtosis tend to have heavy tails or outliers. That's leptokurtic.

00:12:04 Data sets with low kurtosis tend to have light tails or lack of outliers.

00:12:09 That's platykurtic. Distributions with zero excess kurtosis

00:12:15 are called mesokurtic, like the normal distribution family. Skewness is a measure of the asymmetry

00:12:22 of a probability distribution. A distribution is symmetric if it looks

00:12:27 the same to the left and right of the center point. If most of the data is on the left side of the histogram,

00:12:35 but a few larger values are on the right, the data is said to be skewed to the right

00:12:42 with positive skew. If most of the data is on the right

00:12:46 with a few smaller values showing up on the left of the histogram,

00:12:51	the data is skewed to the left with negative skew. If the distribution is symmetric, then the mean is equal
00:13:01	to the median, and the distribution has zero skewness. If the distribution is both symmetric
00:13:10	and unimodal, then the mean equals the median equals the mode.
00:13:17	In the next unit, we'll be considering how to use the normal distribution to calculate probability.

Week 5 Unit 4

00:00:06	Hello, and welcome back to week five, unit four of the openSAP course
00:00:11	Introduction to Statistics for Data Science. In this unit, we'll look at using the normal distribution
00:00:18	to calculate probability. The normal distribution refers to a family
00:00:25	of continuous probability distributions. The graph of the normal distribution depends on two factors:
00:00:34	the mean and the standard deviation. The mean of the distribution
00:00:40	determines the location of the center of the graph, and the standard deviation
00:00:46	determines the height and width of the graph. All normal distributions look like a symmetric,
00:00:53	bell-shaped curve. The graphs show that when the standard deviation is small,
00:01:00	the curve is tall and narrow, and when the standard deviation is big,
00:01:05	the curve is short and wide. The area under the normal distribution
00:01:13	can be used to calculate probabilities for a normally distributed random variable.
00:01:20	This means that the probability that a normal random variable X
00:01:27	equals any particular value is zero. The probability that X is greater than a
00:01:33	is equal to the area underneath the normal curve between a and plus infinity,
00:01:40	the non-shaded area in the diagram. The probability that X is less than a
00:01:46	is equal to the area under the normal curve between a and minus infinity.
00:01:52	That's the shaded area in the diagram. The total area under the curve is equal to one.
00:02:02	Every normal distribution, regardless of its mean or standard deviation,
00:02:07	conforms to the following rule: About 68% of the area under the curve
00:02:12	falls within one standard deviation of the mean. About 95% of the area under the curve
00:02:19	falls within two standard deviations of the mean, and about 99.7% of the area under the curve
00:02:26	falls within three standard deviations of the mean. This is known as the empirical rule
00:02:33	or the 68-95-99.7 rule. Therefore, given a normal distribution,
00:02:40	most outcomes will be within three standard deviations of the mean.
00:02:45	Question: 95% of students at school are between 1.2 and 1.8 meters tall.
00:02:52	Assuming this data is normally distributed, calculate the mean and standard deviation.
00:02:59	Well, the solution? The mean equals 1.2 plus 1.8 divided by 2,
00:03:04	and so, that equals 1.5 meters. 95% is two standard deviations either side of the mean,
00:03:10	a total of four standard deviations. Therefore, one standard deviation equals
00:03:17	1.8 meters minus 1.2 meters divided by 4, which equals 0.15 meters.
00:03:24	This can then be visualized on a normal curve, as you can see on the slide.
00:03:32	How can you use this theory in practice? To find the probability associated
00:03:37	with a normal random variable, use a graphing calculator,
00:03:41	an online normal distribution calculator, or a normal distribution table.
00:03:46	There are lots of normal distribution calculators available, and there are some links given,
00:03:53	so you can choose one for yourself. You are now going to see some simple example calculations.
00:04:03	Question: On average, a light bulb lasts 300 days with a standard deviation of 50 days.
00:04:10	Assuming that bulb life is normally distributed, what is the probability
00:04:15	that the light bulb will last at most 365 days? The solution?
00:04:20	Given a mean score of 300 days and a standard deviation of 50 days,
00:04:26	you need to find the cumulative probability that bulb life is less than or equal to 365 days.

00:04:34	The value of the normal random variable is 365 days. The mean is equal to 300 days.
00:04:41	The standard deviation is equal to 50 days. You enter these values
00:04:47	into the normal distribution calculator and compute the cumulative probability.
00:04:55	The answer: The probability that X is less than 365 days equals 0.9032.
00:05:02	There is a 90% chance that a light bulb will burn out within 365 days.
00:05:09	Question: Scores on an IQ test are normally distributed. If the test has a mean of 110
00:05:16	and a standard deviation of 20, what's the probability that a person who takes the test
00:05:22	will score between 90 and 120? Well, the solution:
00:05:28	Here, you want to know the probability that the test score falls between 90 and 120.
00:05:35	To do this, you use the following simple formula. The probability that X is between 90 and 120
00:05:45	equals the probability that X is less than 120 minus the probability that X is less than 90.
00:05:53	You can use the normal distribution calculator to compute both probabilities
00:05:59	on the right side of the equation. To compute the probability X is less than 120,
00:06:06	you enter the following inputs into the calculator: The value of the normal random variable is 120,
00:06:15	the mean is 110, and the standard deviation is You find that the probability that X is less than
00:06:23	is 0.6915. To compute probability that X is less than 90,
00:06:29	you enter the following inputs into the calculator. The value of the normal random variable is 90,
00:06:37	the mean is 110, and the standard deviation is You find that probability of X is less than 90
00:06:46	is 0.1587. We use these findings
00:06:51	to compute the final answer as follows: The probability that X is between 90 and 120
00:06:59	is equal to 0.6915 minus 0.1587, which equals 0.5328.
00:07:07	Therefore, about 53% probability that the test scores will fall between 90 and 120.
00:07:15	Alternatively, you can use the calculator and enter the lower and upper values,
00:07:20	and it will compute everything for you directly. This is shown in the picture on the slide.
00:07:27	Question: A student achieves a score of 900 in an exam. The mean test score is 825 with a standard deviation of 100.
00:07:37	Assuming that test scores are normally distributed, what proportion of students
00:07:44	achieved a higher score than 900? Well, the solution,
00:07:48	as part of this solution to this problem, you assume that test scores are normally distributed.
00:07:55	In this way, you use the normal distribution to model the distribution of test scores in the real world.
00:08:03	You can use the normal distribution calculator to compute the probability that X is greater than 900
00:08:10	equals to 0.2266. You enter the parameters,
00:08:16	and the calculator computes that you would expect 22.66% of the students
00:08:24	to achieve a higher score than 900. To summarize: The normal distribution refers to a family
00:08:32	of continuous probability distributions. The area under a normal distribution curve
00:08:39	can be used to calculate probabilities for a normally distributed random variable.
00:08:47	There are lots of normal distribution calculators available. Given the mean and the standard deviation,
00:08:54	the calculator can be used to calculate the area under the normal curve, the probability:
00:09:01	less than a value, greater than a value, between values, outside two values.
00:09:09	In the next unit, we will consider hypothesis testing in more detail.

Week 5 Unit 5

00:00:05	Hi, and welcome back to week five, unit five of the openSAP course Introduction to Statistics
00:00:13	for Data Science. In this unit, we're going to look at hypothesis testing
00:00:18	in a little bit more detail. A hypothesis is a proposed explanation for a phenomenon.
00:00:27	It's made on the basis of limited evidence as a starting point for further investigation.
00:00:33	A statistical hypothesis is an assumption about a population parameter.
00:00:40	This assumption may or may not be true. Hypothesis testing refers to the formal procedure
00:00:47	used by statisticians to accept or reject statistical hypotheses.
00:00:55	There are two types of statistical hypothesis: The null hypothesis, which is denoted by Ho,
00:01:04	and is usually the hypothesis that the sample observations result purely from chance.
00:01:11	And the alternative hypothesis, which is denoted by H1 or Ha, and this is the hypothesis that the sample
00:01:20	observations are influenced by some non- random cause. The best way to determine whether a statistical hypothesis
00:01:30	is true would be to examine the entire population. Since that is often impractical, researchers typically
00:01:39	examine a random sample from the population. If sample data are not consistent
00:01:46	with the statistical hypothesis, the hypothesis is rejected. For example, suppose we wanted to determine
00:01:55	whether a coin was fair and balanced. A null hypothesis might be that half the flips
00:02:02	would result in heads and half in tails. The alternative hypothesis might be that the number
00:02:09	of heads and tails would be very different. Symbolically, these hypotheses would be expressed as
00:02:19	Ho probability of 0.5 and H1 probability not equal to 0.5.
00:02:26	Suppose we flipped the coin 50 times, resulting in 40 heads and 10 tails.
00:02:31	Given this result, we would be inclined to reject the null hypothesis.
00:02:37	We would conclude, based on the evidence, that the coin was probably not fair and balanced.
00:02:44	The hypothesis test can have one of two outcomes. You accept the null hypothesis,
00:02:51	or you reject the null hypothesis. However, some statisticians don't agree
00:02:57	with the idea of accepting the null hypothesis. They prefer to say you reject the null hypothesis,
00:03:05	or you fail to reject the null hypothesis. There is a difference between acceptance
00:03:11	and failure to reject. Acceptance implies that the null hypothesis is true.
00:03:18	The failure to reject implies the test is not sufficiently persuasive for us to prefer
00:03:25	the alternative hypothesis over the null hypothesis. Statisticians follow a formal process to determine
00:03:35	whether to reject a null hypothesis based on sample data.
00:03:40	This process, called hypothesis testing, consists of four steps:
00:03:45	Firstly, state the hypotheses. This involves stating the null and alternative hypotheses.
00:03:52	The hypotheses are stated in such a way that they are mutually exclusive.
00:03:58	That is, if one is true, the other must be false. Secondly, formulate an analysis plan.
00:04:05	The plan describes how to use sample data to evaluate the null hypothesis.
00:04:12	The evaluation often focuses around a single test statistic, which is the mean score, the proportion,
00:04:19	the z-score, for example. Thirdly, analyze the sample data.

00:04:24	Find the value of the test statistic described in the analysis plan.
00:04:28	And fourthly, interpret the results. Apply the decision rule described in the analysis plan.
00:04:34	If the value of the test statistic is unlikely, based on the null hypothesis, reject the null hypothesis.
00:04:44	The analysis plan includes decision rules for rejecting the null hypothesis.
00:04:49	Statisticians describe these decision rules in two ways: With reference to a P-value or with reference
00:04:57	to a region of acceptance. The strength of evidence in support of a null hypothesis
00:05:05	is measured by the P-value. Suppose the test statistic is equal to S.
00:05:12	The P-value is the probability of observing a test statistic as extreme as S,
00:05:18	assuming the null hypothesis is true. If the P-value is less than the significance level,
00:05:26	we reject the null hypothesis. The acceptance region or non-rejection region
00:05:35	is a range of values. If the test statistic falls within the acceptance region,
00:05:41	the null hypothesis is not rejected. The set of values outside the region of acceptance
00:05:48	is called the rejection region. If the test statistic falls within the rejection region,
00:05:55	the null hypothesis is rejected. The significance level, alpha, is the probability
00:06:04	of rejecting the null hypothesis when it's true. For example, a significance level of 0.05
00:06:11	indicates a 5% risk of concluding that a difference exists when there is no actual difference.
00:06:21	The significance level for a given hypothesis test is a value for which a P-value less than or equal
00:06:29	to alpha is considered statistically significant. Typical values for the significance level
00:06:36	are 0.1, 0.05, and 0.01. An alternative hypothesis may be one-sided or two-sided.
00:06:45	A one-sided hypothesis claims that a parameter is either larger or smaller than the value
00:06:51	given by the null hypothesis. A two-sided hypothesis claims that a parameter
00:06:57	is not equal to the value given by the null hypothesis. The direction doesn't matter.
00:07:04	For a one-tailed test, the region of rejection is on only one side of the sampling distribution.
00:07:11	For example, suppose the null hypothesis states that the mean is less than or equal to 100.
00:07:20	The alternative hypothesis would be that the mean is greater than 100.
00:07:25	The region of rejection would consist of a range of numbers located on the right side of the sampling distribution.
00:07:34	That is, a set of numbers greater than 100. For a two-tailed test, the region of rejection
00:07:42	is on both sides of the sampling distribution. For example, suppose the null hypothesis
00:07:49	states that the mean is equal to 100. The alternative hypothesis would be that
00:07:55	the mean is less than 100 or greater than 100. The region of rejection would consist of a range of numbers
00:08:04	located on both sides of the sampling distribution. That is the region of rejection would consist partly
00:08:12	of numbers that were less than 100 and partly of numbers that were greater than
00:08:20	Two types of errors can result from a hypothesis test. Type I errors:
00:08:27	These often occur when the researcher rejects a null hypothesis when it is true.
00:08:33	The probability of committing a type I error is the significance level, alpha (α).
00:08:39	Type II errors: These often occur when the researcher fails to reject a null hypothesis that is false.
00:08:48	The probability of committing a type II error is called beta (β).
00:08:53	The probability of not committing a type II error is called the power of the test.
00:09:00	The statistical power ranges from zero to one, and as the statistical power increases,

00:09:07	the probability of making a type II error, in other words, wrongly failing to reject the null, decreases.
00:09:15	For a type II error probability of beta, the corresponding statistical power is one minus beta.
00:09:24	So, to summarize: Hypothesis testing refers to the formal procedures used by statisticians
00:09:32	to accept or reject statistical hypotheses. There are two types of statistical hypotheses.
00:09:39	The null hypothesis Ho, is usually the hypothesis that the sample observations result purely from chance.
00:09:48	Secondly, the alternative hypothesis H1 or Ha, is the hypothesis that the sample observations
00:09:56	are influenced by some non-random cause. An analysis plan includes decision rules
00:10:04	for rejecting the null hypothesis. Statisticians describe these decision rules in two ways:
00:10:11	With reference to a P-value or with reference to a region of acceptance.
00:10:18	Two types of errors can result from a hypothesis test. A type I error occurs when the researcher
00:10:25	rejects a null hypothesis when it's true. A type II error occurs when the researcher fails
00:10:32	to reject a null hypothesis that is false. With this, I'd like to close the fifth week.
00:10:40	I hope you enjoyed these units of this course, and we are happy to get in touch with you
00:10:46	in our discussion forum if you've got any content-related questions.
00:10:52	Now, we wish you all the best for the weekly assignment, and we'll see you next week, where we'll be covering
00:11:00	how we can use statistics in the real world and use SAP solutions to support our statistical analysis.

www.sap.com/contactsap

© 2019 SAP SE or an SAP affiliate company. All rights reserved.
No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP SE or an SAP affiliate company.

The information contained herein may be changed without prior notice. Some software products marketed by SAP SE and its distributors contain proprietary software components of other software vendors. National product specifications may vary.

These materials are provided by SAP SE or an SAP affiliate company for informational purposes only, without representation or warranty of any kind, and SAP or its affiliated companies shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP or SAP affiliate company products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

In particular, SAP SE or its affiliated companies have no obligation to pursue any course of business outlined in this document or any related presentation, or to develop or release any functionality mentioned therein. This document, or any related presentation, and SAP SE's or its affiliated companies' strategy and possible future developments, products, and/or platform directions and functionality are all subject to change and may be changed by SAP SE or its affiliated companies at any time for any reason without notice. The information in this document is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. All forward-looking statements are subject to various risks and uncertainties that could cause actual results to differ materially from expectations. Readers are cautioned not to place undue reliance on these forward-looking statements, and they should not be relied upon in making purchasing decisions.

SAP and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP SE (or an SAP affiliate company) in Germany and other countries. All other product and service names mentioned are the trademarks of their respective companies. See www.sap.com/copyright for additional trademark information and notices.

THE BEST RUN SAP®