|  |  |  |  |
| --- | --- | --- | --- |
| Date: | **01-08-2020** | Name: | **Varun G Shetty** |
| Course: | **Coursera** | USN: | **4AL17EC093** |
| Topic: | Basic Statestics | Semester & Section: | **6th & ‘B’** |
| GitHub Repository: | **Varunshetty4** |  |  |

**Report-**

R Language:

Probability: Probability is calculated as the relative frequency of an ouctome. Independent intersecting events are two events that do not influence each other and can occur similtaneously. An example might be the outcome of rolling two dices.

Disjoint exhaustive events are mutually exclusive, so only one of the events can happen at a time. Let's use these probability values to find some conditional probabilities. To do this you need to use the formula P(A|B)=P(A+B)/P(B)P(A|B)=P(A+B)/P(B).

The probability of one event is independent of another if the probability of the first event occuring is unaffected by whether or not the other event occurs, i.e. if P(A|B) = P (A).

Bayes law is based on the idea that prior knowledge about A (e.g. whether or not a plant is a tree), provides information about B (e.g. whether or not a plant lives indoors).

The formula for Bayes law is: P(A|B)=(P(B|A)∗P(A))/P(B)

Probability mass and density Function:  Probability mass functions relate to the probability distributions discrete variables, while probability density functions relate to probability distributions of continuous variables.

For continuous variables, the values of a variable are associated with a probability density. To get a probability, you will need to consider an interval under the curve of the probability density function. Probabilities here are thus considered surface areas.

 we will simulate some random normally distributed data using the rnorm() function. dnorm gives the density, pnorm gives the distribution function, qnorm gives the quantile function, and rnorm generates random deviates.

**Usage**

dnorm(x, mean = 0, sd = 1, log = FALSE)

pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)

qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)

rnorm(n, mean = 0, sd = 1)

**Arguments**

**x, q -** vector of quantiles.

**P -** vector of probabilities.

**n -** number of observations. If length(n) > 1, the length is taken to be the number required.

**Mean -** vector of means.

**Sd -** vector of standard deviations.

**log, log.p -** logical; if TRUE, probabilities p are given as log(p).

**lower.tail -** logical; if TRUE (default), probabilities are \(P[X \le x]\) otherwise, \(P[X > x]\).

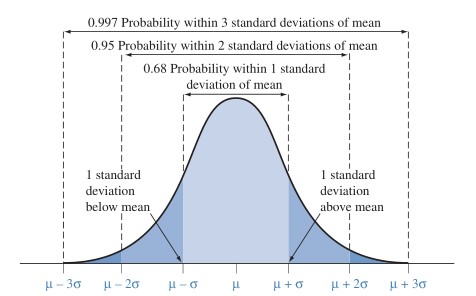
We could sum individual probabilities in order to get a cumulative probability of a given value. However, in some cases, the function cumsum() may come in handy. What cumsum() does is that returns a vector whose elements are the cumulative sums of the elements of the arguments. For instance, if we would have a vector which contains the elements: c(1, 2, 3), cumsum() would return c(1, 3, 6)

Summary Statestics: One of the first things that you would like to know about a probability distribution are some summary statistics that capture the essence of the distribution. One example of such a summary statistics is the mean. The mean of a probability distribution is calculated by taking the weighted average of all possible values that a random variable can take.

 The variance is often taken as a measure of spread of a distribution. It is the squared deviation of an observation from its mean. If you want to calculate it on the basis of a probability distribution, it is the sum of the squared difference between the individual observation and their mean multiplied by their probabilities. See the following formula: var(X)=∑(xi−x¯)2∗Pi(xi)var(X)=∑(xi−x¯)2∗Pi(xi).

If we want to turn that variance into the standard deviation, all we need to do is to take its square root.

Normal Distribution: The normal distribution, als known as the Gaussian distribution, is the probability distribution that is encountered most frequently. It is characterized by a nice bell curve. A normal distribution is centered at its mean called μμ. Its spread is defined by the standard deviation. The image below gives an idea how the probability density function and the standard deviation of a normal distribution are related:



We can do probability in R using the pnorm() function. This function calculates the cumultative probability. We can use it the following way: pnorm(30, mean = 25, sd = 5). If you wanted to calculate the probability of a woman having a hair length larger or equal to 30 centimers, you can set the lower.tail argument to FALSE. For instance, pnorm(30, mean = 25, sd = 5, lower.tail = FALSE).

Sometimes we have a probability that we want to associate with a value. This is basically the opposite situation as the situation described in the previous question. Say we want the value of a woman's hair length that corresponds with the 0.2 quantile (=20th percentile).

This value is the 0.2 quantile (=20th percentile) and divides the curve in an area that contains the lower 20% of the scores and an area that the rest of the scores. If our variable is normally distributed, in R we can use the function qnorm() to do so. We can specify the probability as the first parameter, then specify the mean and then specify the standard deviation, for example, qnorm(0.2, mean = 25, sd = 5).

Standard normal deviation / Z-score: A special form of the normal probability distribution is the standard normal distribution, also known as the z - distribution. A z distribution has a mean of 0 and a standard deviation of 1. Often you can transform variables to z values.

You can transform the values of a variable to z-scores by subtracting the mean, and dividing this by the standard deviation. If you perform this transformation on the values of a data set, your transformed data set will ave a mean of 0 and a standard deviation of 1. The formula to transform a value to a z score is the following:

Zi=xi−x¯/sx