MAST7210 — Foundations of Data Science

Computing Session 6 — Random Variables and Probability distributions

Within this session we are going to learn how to draw random samples from a variety of probability distributions in R.

Task 1 – Introductory Probability

Watch the Introductory Probability video on moodle.

Task 2 – Axioms of Probability

Watch the Axioms of Probability video on moodle.

Task 3 – Discrete Random Variables and Probability Distributions

Watch the video on Random Variables – Discrete on moodle.

Task 4 – Binomial Distribution

The binomial probability distribution is one of the most widely used discrete probability distributions. It is applied to find the probability that an outcome will occur x times in n performances of an experiment. For example, if 75% of students at a University use Instagram, we may want to find the probability that in a random sample of five students at this University that exactly three use Instagram.

A binomial experiment must satisfy the following conditions:

1. There are n identical trials
2. Each trial has only two possible outcomes.
3. The probabilities of the two outcomes remain constant.
4. The trials are independent.

For a binomial experiment, the probability of exactly x successes in n trials is given by the formula:

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

where n denotes the number of trials and p denotes the probability of success.

The mean and the standard deviation of a binomial distribution are, respectively:

$$\mu = np \qquad \text{and} \qquad \sigma = \sqrt{np(1-p)}.$$

Example 1

Suppose there are twelve multiple choice questions in an English class quiz. Each question has five possible answers, and only one of them is correct. Find the probability of having four or less correct answers if a student attempts to answer every question at random.

Since only one out of five possible answers is correct, the probability of answering a question correctly by random is 1/5 = 0.2. We can find the probability of having exactly 4 correct answers by random attempts as follows:

```
> dbinom(4, size=12, prob=0.2)
```

To find the probability of having four or less correct answers by random attempts, we apply the function `dbinom` with x = 0, …, 4.

```
> dbinom(0, size=12, prob=0.2) +
+ dbinom(1, size=12, prob=0.2) +
+ dbinom(2, size=12, prob=0.2) +
+ dbinom(3, size=12, prob=0.2) +
+ dbinom(4, size=12, prob=0.2)
```

Alternatively, we can use the cumulative probability function for a binomial distribution `pbinom`, which calculates the probability that a random variable is less than or equal to a given value:

```
> pbinom(4, size=12, prob=0.2)
```

We can also use R to produce random samples from a binomial distribution. Here is the command to produce a sample of 20 binomial random variables from the binomial distribution with n = 12 and p = 0.2:

```
> rbinom(20, 12, 0.2)
```

Challenge 1

According to a survey, 30% of college students said that they spend too much time on Facebook. Suppose this result holds true for the current population of all college students. A random sample of six college students is selected.

Find the probability that exactly three of these six college students will say they spend too much time on Facebook.

Find the probability that at most two of these six college students will say that they spend too much time on Facebook.

Find the probability that at least three of these six college students will say that they spend too much time on Facebook.

Task 5 – Poisson Distribution

The Poisson probability distribution, named after the French mathematician, Simeon-Denis Poisson, is another important probability distribution of a discrete random variable that has a large number of applications. Suppose a washing machine in a launderette breaks down an average of three times a month, we may want to find the probability of exactly two break downs in the next month. This is an example of a Poisson probability distribution problem. Each breakdown is called an occurrence.

The following three conditions must have satisfied to apply the Poisson probability distribution:

1. x is a discrete random variable
2. The occurrences are random
3. The occurrences are independent

According to the **Poisson probability distribution**, the probability of x occurrences in an interval is

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where λ is the mean number of occurrences in that interval.

For the Poisson probability distribution, the mean and variance are both equal to λ, thus the standard deviation is $\sqrt{\lambda}$.

Example 2

On average, two new accounts are opened per day at a Barclays branch. Find the probability that on a given day the number of new accounts opened at this bank will be:

(a) exactly 6

```
> dpois(6,2)
```

(b) at most 3

```
> dpois(0,2)+dpois(1,2)+dpois(2,2)+dpois(3,2)
```

or

```
> ppois(3,2)
```

(c) at least 7

```
> 1-ppois(6,2)
```

Challenge 2

An insurance salesperson sells an average of 1.4 policies per day. Find the probability that this salesperson will sell no insurance policy on a certain day.

What is the probability that the salesperson will sell at least 1 policy on a certain day?

What is the probability that the salesperson will sell at least 2 policies on a 2 day period?

Task 6 – Continuous Random Variables and Probability Distributions

Watch the video on Random Variables – Continuous on moodle.

Task 7 – Normal Distribution

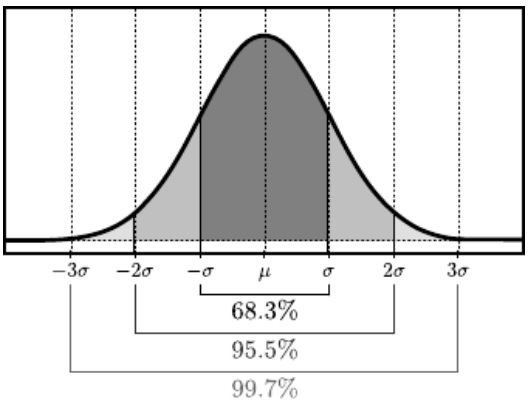
The normal distribution is one of the many probability distributions that a continuous random variable can possess. The normal distribution is the most important and most widely used of all probability distributions as a large number of phenomena in the real world are approximately normally distributed.

A normal probability distribution is a bell-shaped curve and its mean is denoted by μ and its standard deviation by σ. Note that not all bell-shaped curves represent a normal distribution, only a specific kind of bell-shaped curve represents a normal curve.

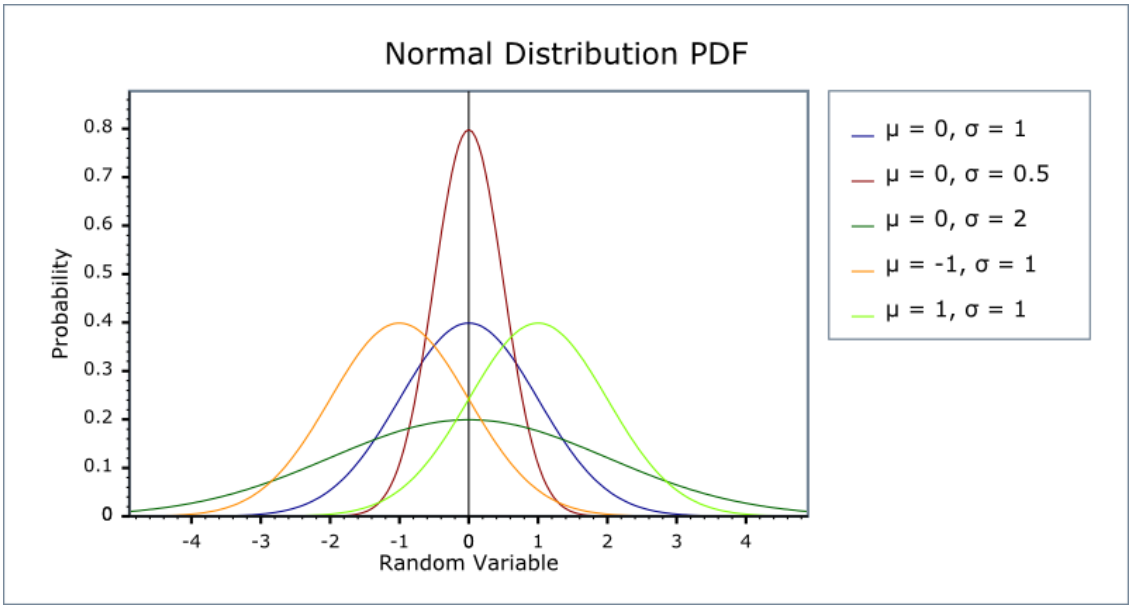
The bell-shaped curve is such that:

1. The total area under the curve is 1.
2. The curve is symmetric about the mean.
3. The two tails of the curve extend indefinitely.

Although the normal curve never meets the horizontal axis, beyond the points represented by $\mu - 3\sigma$ and $\mu + 3\sigma$ it becomes so close to the axis that the area under the curve beyond those points in both directions is very small. The actual area in each tail is 0.0013.



The mean, μ and standard deviation, σ are the parameters of the normal distribution. The value of μ determines the centre of the curve and the σ gives normal curves with different height/spread.



In R we can generate random samples from a normal distribution. Let us draw a sample of size 100 from a normal distribution with mean 2 and standard deviation 5.

```
> norm <- rnorm(100, 2, 5)
```

Let us display the first 10 observations:

```
> norm[1:10]
```

Let us calculate the mean and standard deviation of the sample:

```
> mean(norm)
```

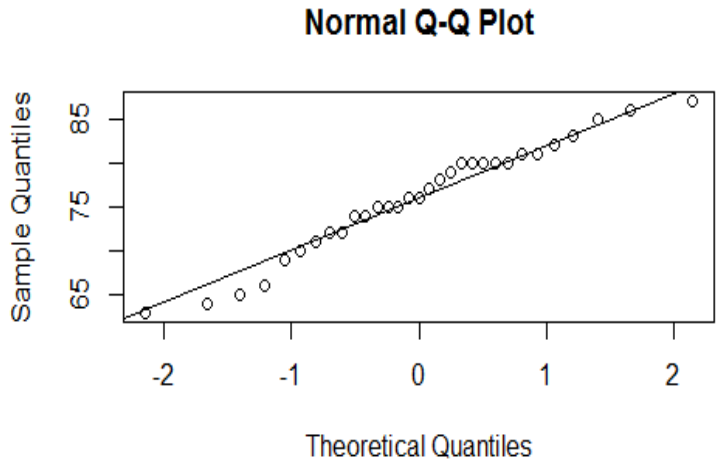
```
> sd(norm)
```

Task 8 – Assessing Normality

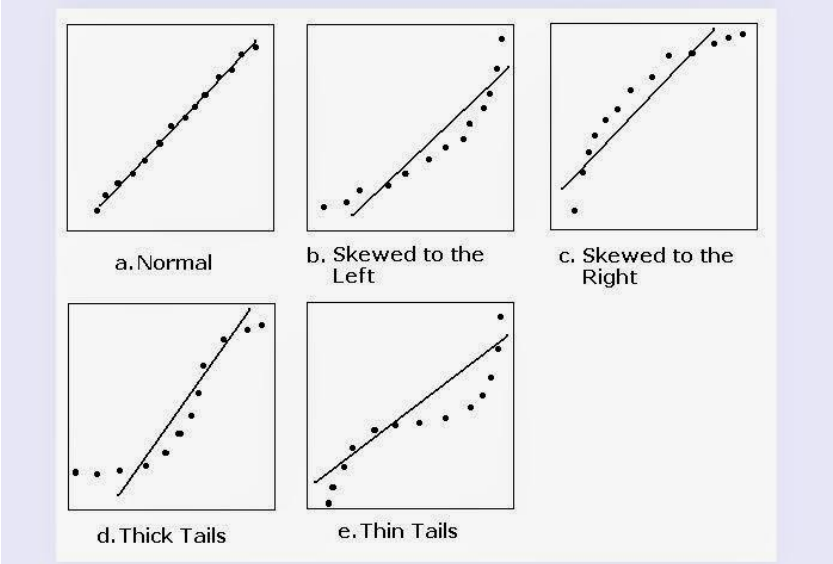
Many of the methods used in statistics require that the sampled data come from a normal distribution. While it is impossible to determine if this holds true without taking a complete census, there are statistical tools that can be used to determine if this is a reasonable assumption. One of the simplest tools is called a **normal quantile plot,** or **QQplot**. The idea of the plot is to compare the values in a data set with corresponding values one would predict for a standard normal distribution. If the data are in complete agreement with the normal distribution, the points should lie on the line displayed on the graph.

Consider the trees data set which is built into R. The data set provides measurements of the girth, height and volume of timber in 31 felled black cherry trees.

```
> qqnorm(trees$Height)
> qqline(trees$Height)
```



We can see from this plot that the points seem to fall about a straight line and so it appears to be a fairly safe assumption the data come from a population which is normally distributed.



Challenge 3

A data set consisting of weights pre and post-treatment for patients suffering from anorexia are contained in the R library MASS.

To load these data use the commands:

```
> library(MASS)
```

```
> attach(anorexia)
```

Plot QQplots for both the Prewt and Postwt data. What do you conclude?