Week 2 - Before the lecture

Here is a list of items to read and be familiar before the lecture:

- Read through the following sections of Crawley's book: 2,3,4 (up to and *not* including Bootstrap) and 5 (up to and including Student's t Distribution).
- The material in chapter 5 is most pertinent to this week's lecture and lab material

The concepts and definitions below **are not core** to the module however they can be helpful to be familiar with as the module progresses. You may want to read them either before or after the lecture.

1 Random Variables

In real world problems we encounter quantities that do not have a fixed value. An example of this is the number of ice-creams a shop can make each day. In probability, these quantities are called *random variables*. The numerical values of random variables are unknown before the experiment takes place. In other words we don't know how many ice-creams will actually be made tomorrow.

- Let S be the **sample space**, which is the set of all possible outcomes.
- A **Random Variable** is a function that assigns a real value to each outcome in S. A **random variable** can be discrete or continuous.
- for example X is the number of ice creams made by a shop in a day and the sample space could be $0 \leq X \leq 200$, if the ice-cream shop can make maximum 200 ice-creams per day,
- Random variables are useful to compute probability of events happening, such as what is the probability that $X = 99$? or that X is greater than or less than a number or in a specific range?

An example of this is in Figure 1 which shows us the density of ice creams made each day, based on data from 100 days. The x axis represents the number of ice creams made and the y axis is the probability of that number of ice creams being made. Most days somewhere between 80 and 120 ice creams were made.

1.1 Cumulative Distribution Function (CDF)

- If X is a Random Variable then F on $(-\infty, \infty)$ by $F(x) = P(X \leq x)$, F is called the **Cumulative Distribution Function (CDF)** of X
- $0 \leq F(x) \leq 1$ for all x
- F is a non decreasing function of x

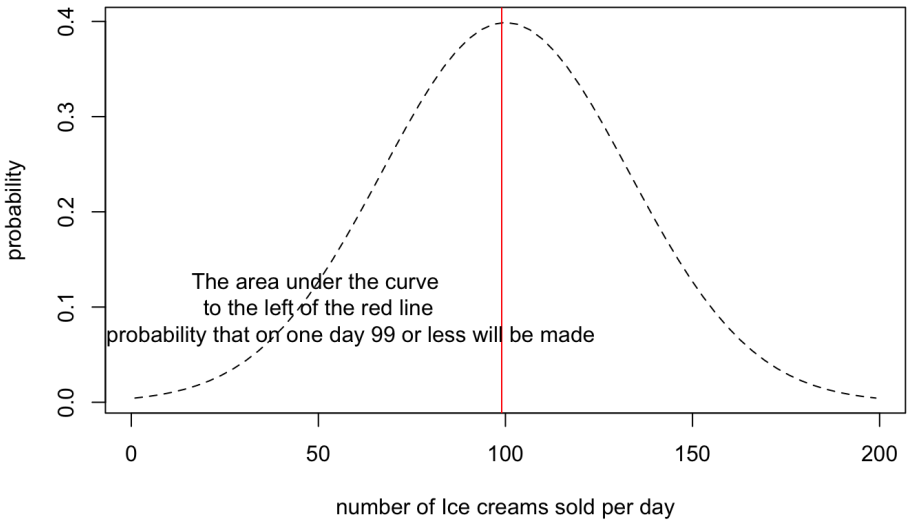


Figure 1: Distribution of number of ice-creams made each day based on data from 100s of days - Density function. The area under the curve to the left of the red line is the probability that on a given day 99 or less ice creams will be made

- $lim_{x\to\infty}F(x) = 1$, $lim_{x\to-\infty}F(x) = 0$.

The CDF for our ice cream example helps us obtain the probability that $X \leq 99$, i.e. the probability that on one day 99 or less ice creams will be made.

2 Random Variable distributions

There are two types of random variable distributions we will look at:

- **Discrete distribution** - the random variable can only take only a specific set of values (for example a dice rolled once)
- **Continuous distribution** - can take any continuous value (for example the number of ice creams or the height of a 12 year old child)

Firstly lets look at some definitions for **Discrete Random Variables**.

2.1 Discrete Random Variables

The **Probability Mass Function** PMF of a **discrete random variable** X whose set of possible values is $\{x_1, x_2, \ldots\}$ is a function from $\mathbb{R}$ to $\mathbb{R}$ such that:

1. $f(x) = 0$ if $x \notin \{x_1, x_2, \ldots\}$
2. $f(x_i) = P(X = x_i)$ and hence $f(x_i) \geqslant 0$ for $i = 1, 2, 3 \ldots$

3. $\Sigma_{i=1}^{\infty} f(x_i) = 1$

Intuitively the **PMF gives a probability to each possible outcome**, for example in the die example the probability of each possible outcome is $1/6$.

An example: PMF and CDF for rolling a single fair die

- The **sample space** for a single die is 1,2,3,4,5,6
- The **PMF** for the die is $Pr(X = k) = \frac{1}{6}$ for $k = 1, 2 \ldots, 6$
- The **CDF** for the die is:
  - $Pr(X \leq 1) = 1/6$
  - $Pr(X \leq 2) = 2/6$
  - $Pr(X \leq 3) = 3/6$
  - $Pr(X \leq 4) = 4/6$
  - $Pr(X \leq 5) = 5/6$
  - $Pr(X \leq 6) = 6/6 = 1$

Discrete random variable common distributions include: Binomial Distribution, Poisson Distribution, Bernoulli distribution and many more

2.2 Binomial Distribution

The **Binomial Distribution** is defined by 2 parameters, $B(n, p), n =$ the no of trials; $p =$ the probability of success each time. An example of a binomial distribution is fair coin toss $(p = 0.5)$ where p is the probability of Heads. The probability mass function for *heads* for $n = 100$ coin tosses is in Figure 2. Figure 3 shows the density function when the coin is tossed 10 times and Figure 4 shows the same when the coin is tossed 1000 times.

The larger the sample (the higher the number of coin tosses) the closer this probability mass distribution function is to having all its mass closest to 0.5.

Binomial Distribution in R

In R this family of functions is associated with the `binom` functions and the parameters are n number of trials, p probability of success at each trial.

- the PMF is obtained using `dbinom(v, n, p)` where v is the specific outcome (i.e. the number of successes)
- the CDF is the probability of up to k successes in n trials with a probability of success p is `pbinom(k,n,p)`

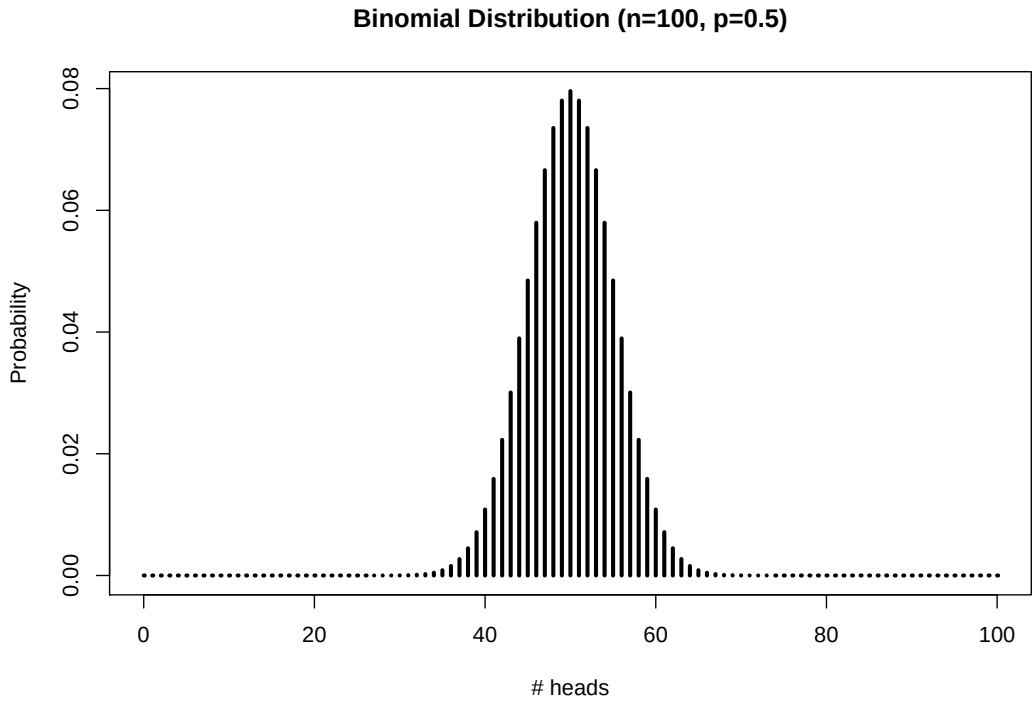


Figure 2: Binomial Distribution n=100 and p=0.5

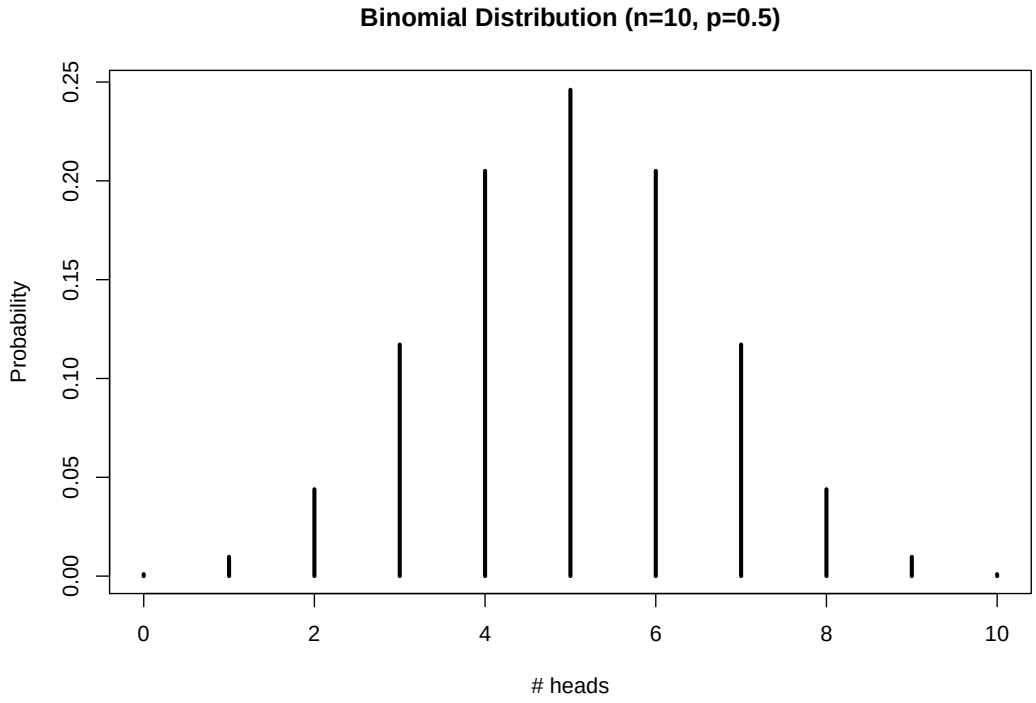


Figure 3: Binomial Distribution n=10 and p=0.5

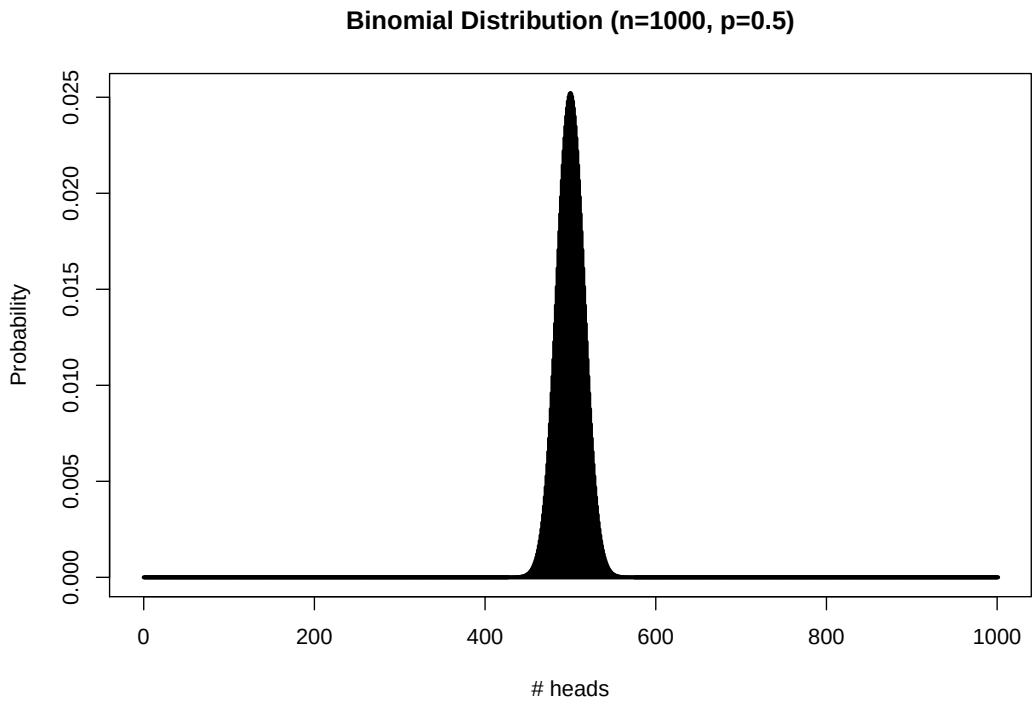


Figure 4: Binomial Distribution n=1000 and p=0.5

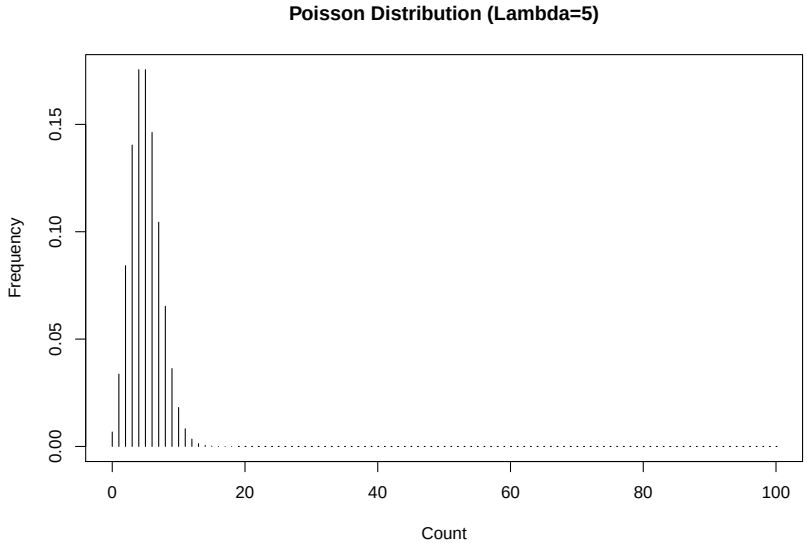


Figure 5: Poisson distribution with $\lambda = 5$

2.3 Poisson Distribution

The **Poisson** Distribution is used for description of count data. This can be the count of the number of times something happened, but not how many times it did not (e.g. bomb hits). It is a one parameter distribution , defined by λ which is the mean and the variance. The variance is the same as the mean for this type of distributionn.

The Poisson distribution's λ is the parameter for both the mean and the variance.

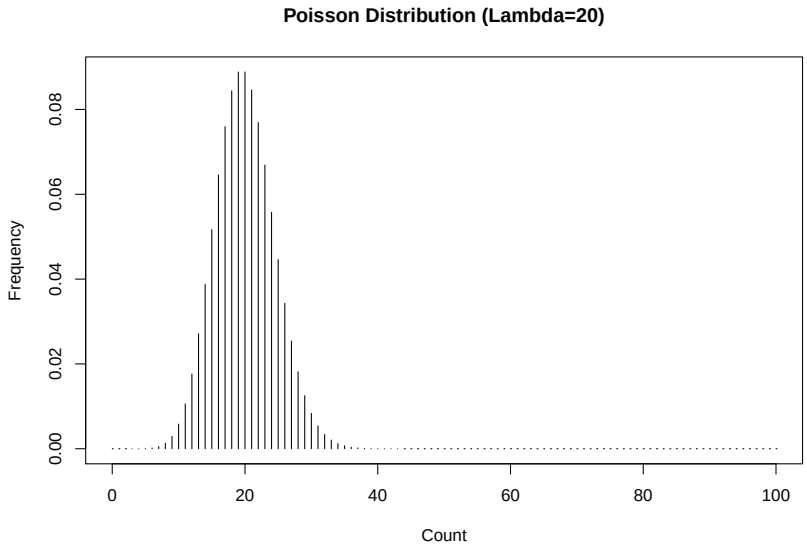


Figure 6: Poisson distribution with $\lambda = 20$

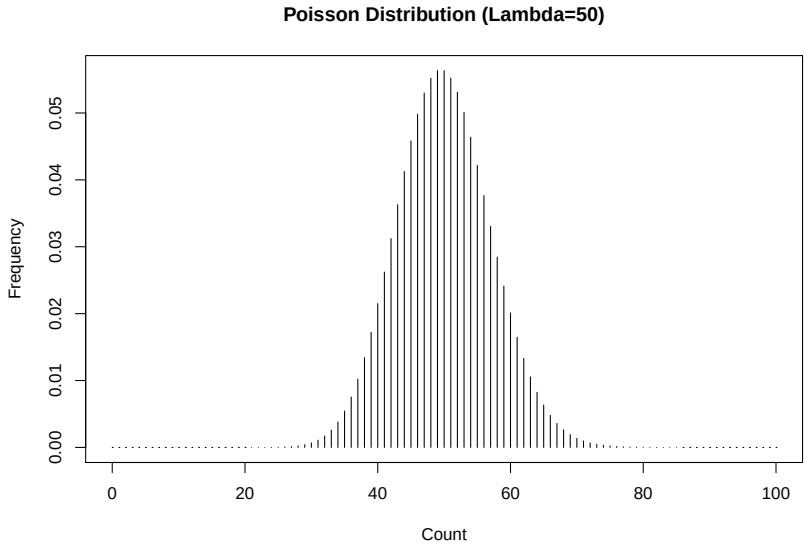


Figure 7: Poisson distribution with $\lambda = 50$

3 Density Function for a Continuous Random Variable

The **Probability Density Function - PDF** of a continuous random variable X is a function $f : \mathbb{R} \to (0, \infty)$, such that:

$$P(X \in A) = \int_a f(x)dx$$

1. $F(t) = \int_{-\infty}^{t} f(x)dx$, so $f(x) = F'(x)$
2. $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P(a < X < b) = \int_a^b f(x)dx$.

The PDF can help us find the probability that X is between two values, for example if X is the height of 12 year old child, this will tell us the probability that a 12 year old child's height will be between 145cm and 155cm.

Continuous random variable common distributions include: Normal Distribution, Uniform Distribution, Exponential distribution and many more

3.1 Normal Distribution

- The **Normal Distribution** has a central place in statistical analysis
- If repeat samples are taken from a population, and their averages calculated, then these averages will be normally distributed - this is referred to as the **Central Limit Theory**
- The **Normal Distribution** $N(\mu, \sigma^2)$
- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad x \in \mathbb{R}$
- When the mean is 0 and the s.d. is 1, this is known as the **Standard Normal Distribution** X.

3.2 Student's t-distribution

The central limit theory depends on a large sample size and on knowing the standard deviation of the population. In such cases it is possible to estimate the standard deviation from the sample. **Student's t-distribution** is used instead of the normal. t is more spread out then z (the normal distribution) because the use of s (the estimated s.d. from the sample) introduces more uncertainty resulting in "fatter tails" in the density function. See Figure 8 where the density functions for Z (normal), t with a very small sample and t with a large sample are compared.

For more see page 82 Crawley

- In situations when the limits of the **central limit theory** are **not** met, such as:
  - when we have a small sample (less than 30)

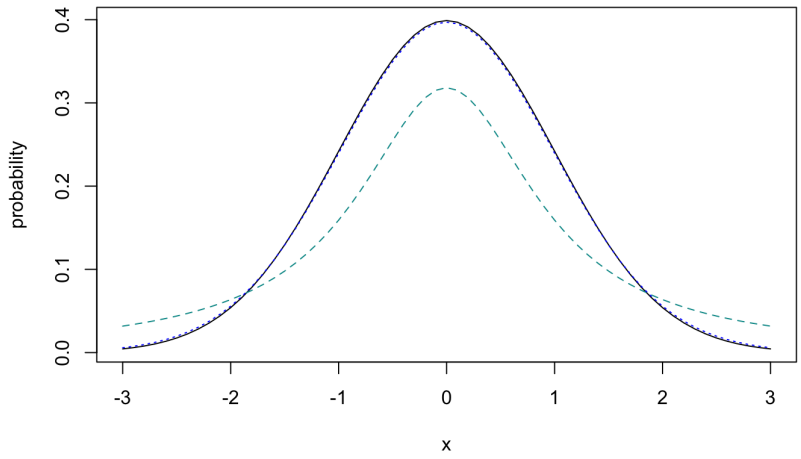


Figure 8: Normal distribution is in black, the t distribution with a small sample is in green and the dotted blue line is t with a large sample

  - and the s.d. σ needs to be estimated from the sample too

- In such situations the sample σ can be estimated from the data using

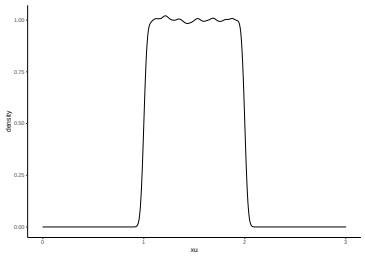
$$s = \sqrt{\frac{1}{n-1}\Sigma_{i=1}^{n}(x_i - \bar{x})^2}$$

Degrees of Freedom is the sample size n, minus the number of parameters , p estimated from the data. If we estimate the mean and the sd then df=n-1.

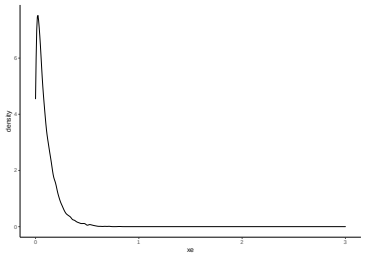
For more see p53 Crawley

3.3 Other distributions

- **Uniform** distribution is defined as $U(a, b)$ where $f(x) = \frac{1}{b-a}$ if $a < x < b$ and zero otherwise. All points between a and b are equally likely.



- **Exponential** distribution is defined as $f(x) = \lambda e^{-\lambda t}$, $0 < x < \infty$. Exponential distributions are *memoryless*.



If you want to extend your knowledge on the topic of Distributions see Chapter 6 of *Probability and Statistics for Data Science : Math + R + Data* see the module BBL page for a link in the reading list.

3.4 Probability Functions applications

- Probability functions can model measurements or real world data
- A smaller set of these are used to model the behaviours of **sample statistics**
- The Normal distribution belongs to both categories
- Student's t, χ^2 and F distributions are important for inference and we will be using them later in the module.

3.5 Probability functions in R

There are four basic probability functions for each probability distribution in R.

R's probability functions begins with one of four prefixes: d, p, q, or r followed by a root name that identifies the probability distribution. For the normal distribution the root name is norm. The meaning of these prefixes is as follows.

- d is for "density" and the corresponding function returns the value from the probability density function (continuous) or probability mass function (discrete).
- p is for "probability" and the corresponding function returns a value from the cumulative distribution function.
- q is for "quantile" and the corresponding function returns a value from the inverse cumulative distribution function.
- r is for "random and the corresponding function returns a value drawn randomly from the given distribution.

To use these prefixes you will need to also specify the distribution to use. For example `norm`, `binom` or `t`.

Some specific examples:

- `dnorm(2)` will return the value of the PDF for a given value of x =2 - See Figure 9

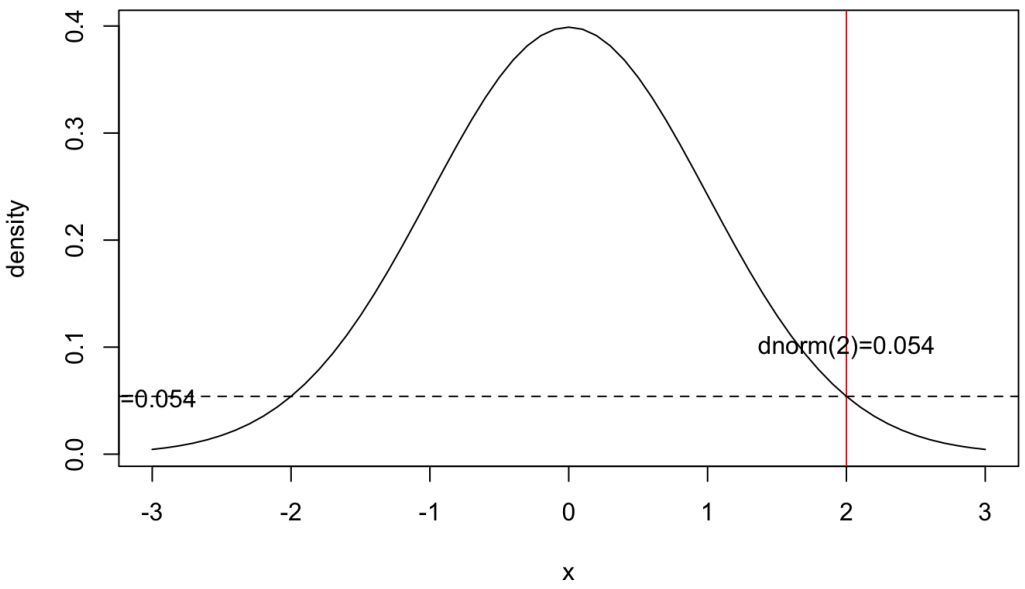


Figure 9: The PDF for this distribution - `dnorm`. In this case the function `dnorm(2)` returns the value of y (the density) in this curve when $x = 2$

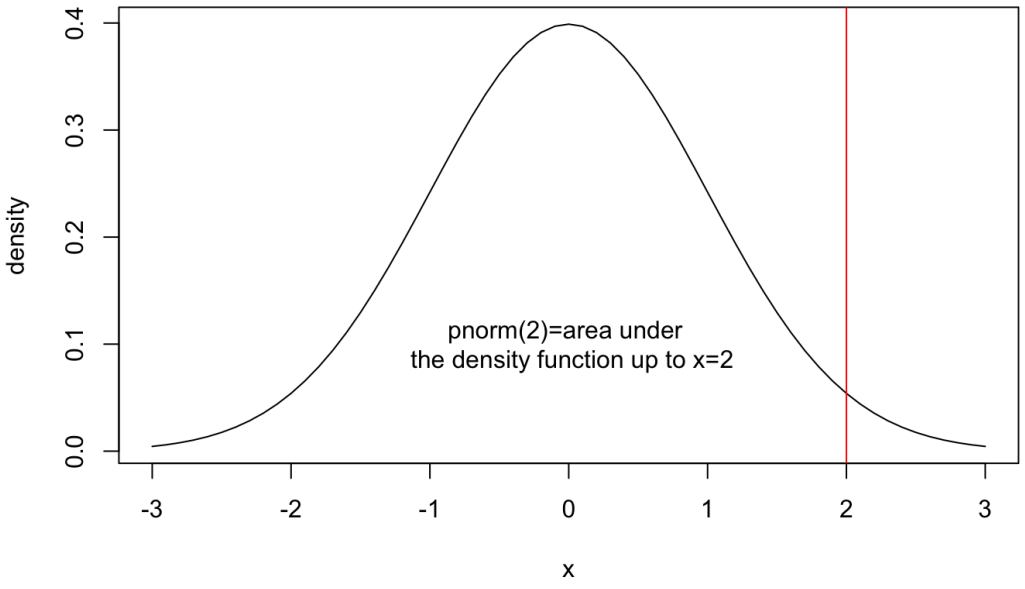


Figure 10: The CDF for this distribution `pnorm`. In this case the function `pnorm` returns the area under the curve (density) up to x=2

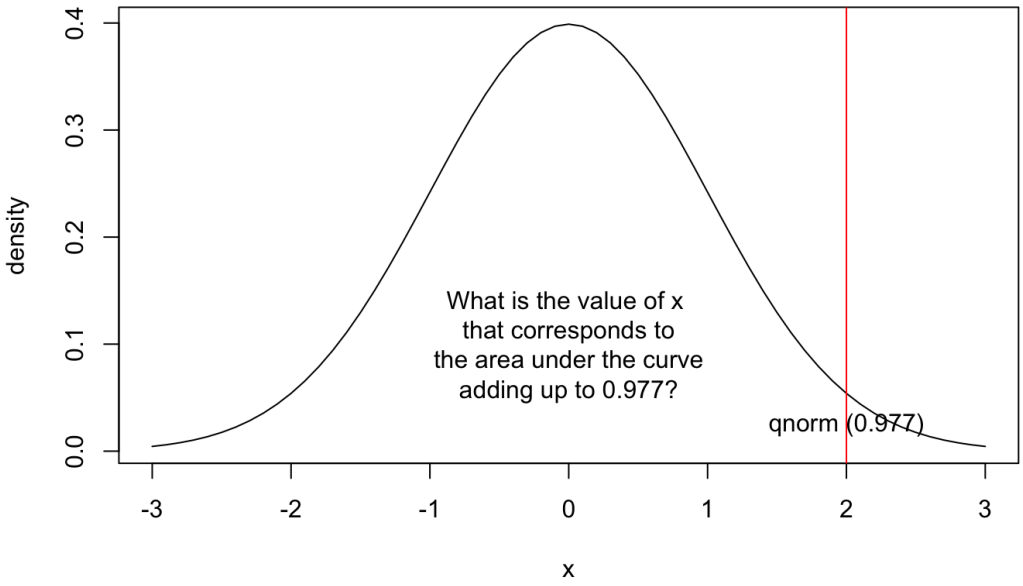


Figure 11: The value of X that corresponds to the CDF `qnorm`. What is the value of x that corresponds to the area under the curve adding up to 0.977?

- `pnorm(2)` will return the cumulative distribution function when x=2, which is the probability that x is ≤ 2 in a normal distribution. See Figure 10 and `pnorm(2)`=0.977.
- `rnorm(20)` will return a random sample of 20 taken from a normal distribution with mean =0 and sd=1.
- `qnorm(0.0975)` answers the question what is the Z score for a given significance level. We will use this in the coming weeks.