Module 4

Probability Distributions: Random, but Predictable

James Baglin

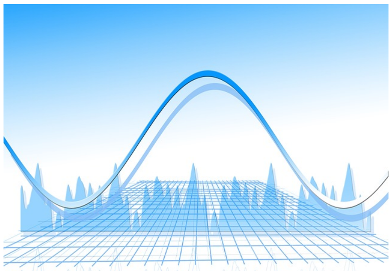
Last updated: 13 July, 2020

Overview

Despite the omnipresence of variability, many variables follow predicable patterns. That's not to say we can reliably predict an individual observation with great certainty, but over the course of many repeated observations of a variable, we can predict many informative outcomes. This module introduces two discrete probability distributions and one continuous probability distribution which are know to model the behaviour of many random processes.

##Summary

Learning Objectives



The learning objectives associated with this module are:

- Define and distinguish between random variables, discrete random variables, and continuous random variables.
- Define the properties of the Binomial and Poisson distributions.
- Correctly apply and work with the Binomial and Poisson distributions to solve Binomial and Poisson-based problems.
- Define the properties of the normal distribution and identify where it can be applied.
- Work with the normal distribution to solve normal-based problems.
- Define the standard normal z-distribution.
- Standardise random variables to standard normal variables, and vice-versa.
- Compare empirical distributions to theoretical probability distributions

Module Video

Probability Distributions

This module will introduce you to the first set of fundamental probability distributions used by statisticians to model random processes. Probability distributions are based on the central concept of statistics that, while an individual random event is almost impossible to predict, the behaviour of random processes in the long run can be very well understood. This module introduces discrete and continuous probability distributions. These distributions are used to model quantitative variables that can only take on discrete values (1, 2, 3…) and continuous values (1.45, 5.43, 2.39). These probability distributions can be used to model many random variables including guessing on a multiple choice exam, the number of heads flipped from five tosses of a coin, the number of goals scored in a football match, the infection rate of a disease or the number of people lining up in a queue at your local cafe. Let's start with the binomial distribution.

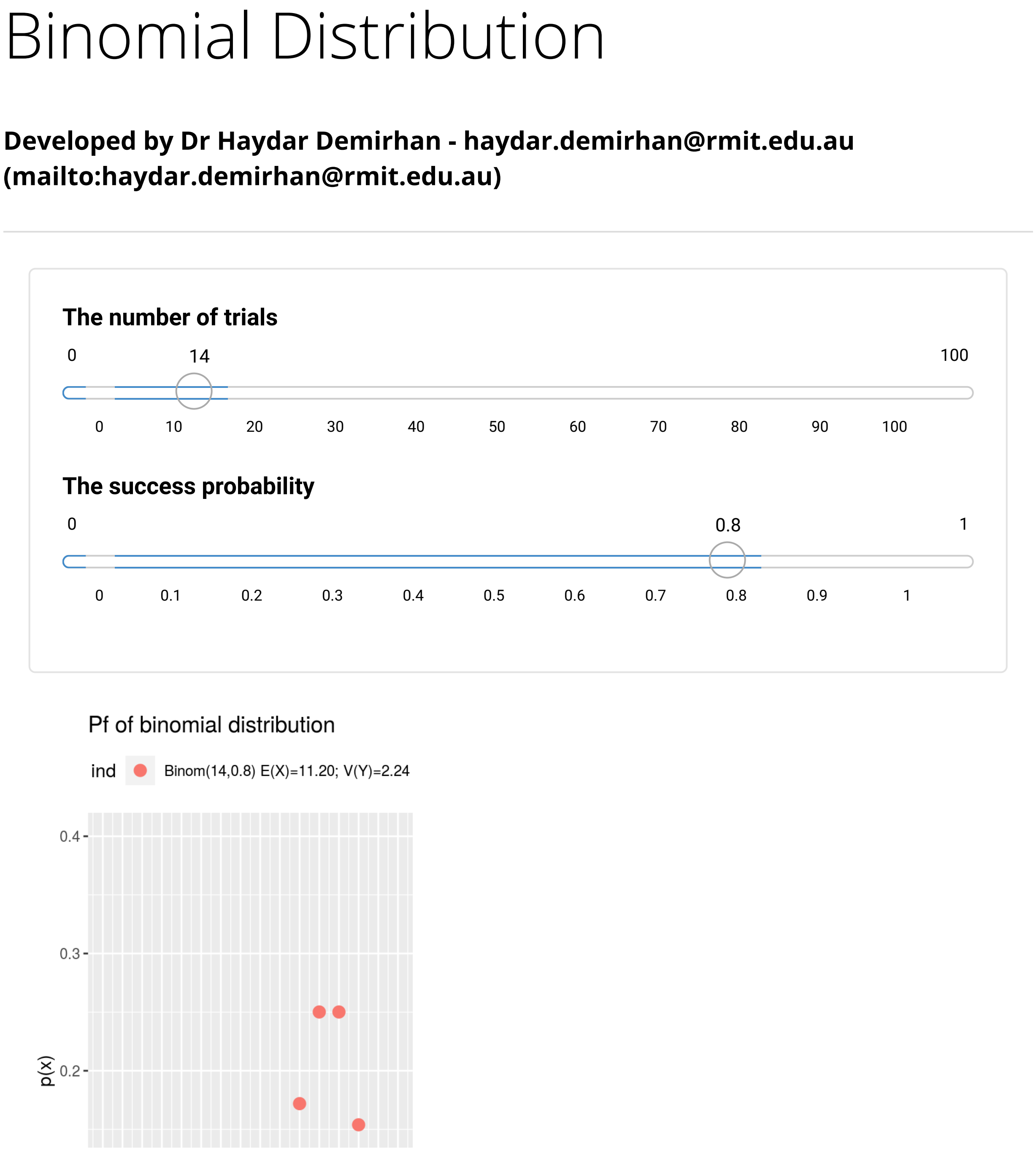
Binomial Distribution

The binomial distribution is used to model the number of successes/failures in n independent trials where the probability of success at each trial is fixed as p. The probability of failure is $1 - p$. For example, let's say a cancer vaccine is effective 85% of the time, p = .85. If we randomly select 12 vaccinated people from the population, n = 12, and expose them to the virus that causes the cancer, what is the probability that the vaccine will be successful for all 12 people? Before we answer this question, let's take a look at some theory.

The binomial distribution has the following mathematical form:

$$Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

where k = successes, n = no. of trials and p = probability of success. This formula is known as a **probability mass function** or PMF. You can visualise and interact with this distribution using Dr Haydar Demirhan's Binomial Shiny App.



Binomial Distribution

Developed by Dr Haydar Demirhan - haydar.demirhan@rmit.edu.au (mailto:haydar.demirhan@rmit.edu.au)

The number of trials: 14

The success probability: 0.8

Pf of binomial distribution

The mean, or expected value, $E(x)$, variance and standard deviation for a binomial distribution are as follows:

$$\text{Mean} = E(x) = np$$

$$\text{Variance} = np(1 - p)$$

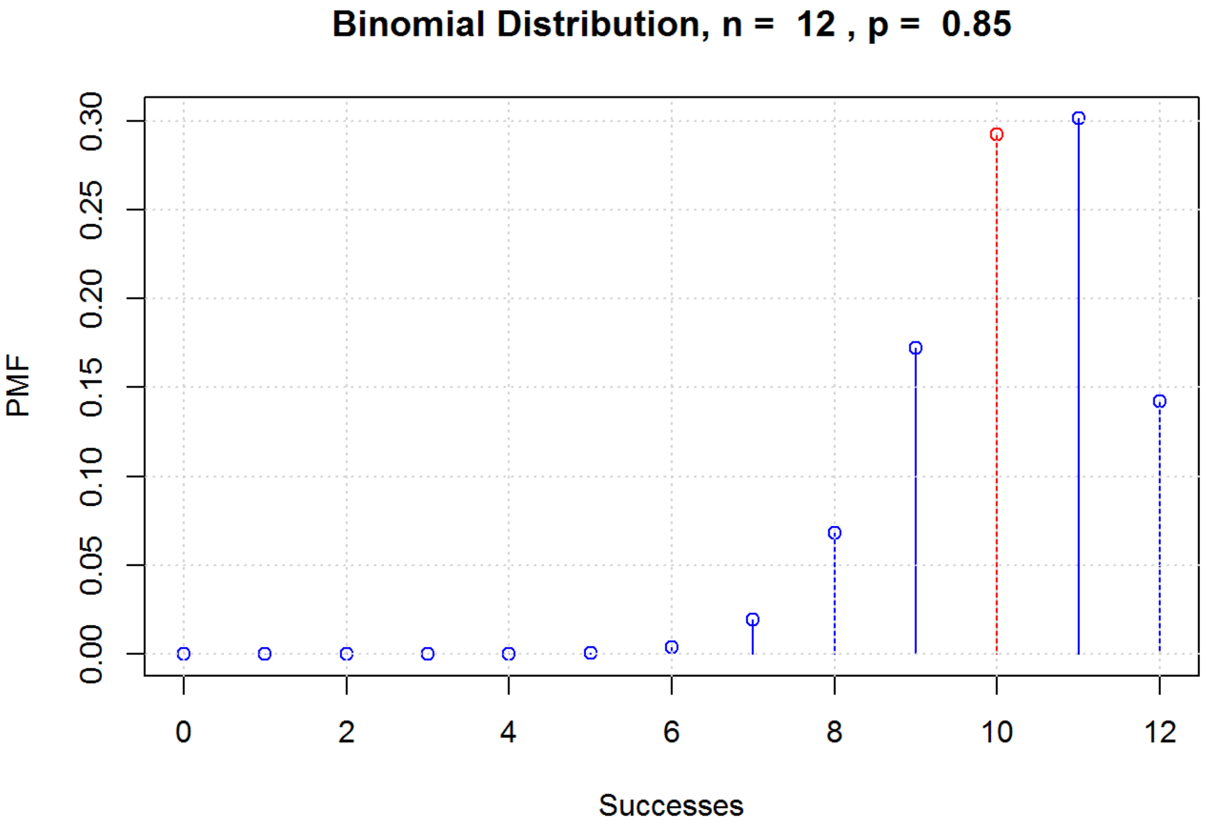
$$\text{Standard deviation} = \sqrt{np(1 - p)}$$

However, there is no need to remember these formulae. We will learn to use R to do the hard work for us. We will demonstrate R's binomial function to solve the following questions related to the vaccine scenario. We will assume p = .85 and n = 12.

$Pr(X = k)$

1.1. What is the probability that the vaccine will work for 10 people?

The key to solving these types of problems is to understand the question. Let's write it out. The question is asking $Pr(X = 10)$, assuming 12 trials and the vaccine to be effective 85% of the time. X is a short way to refer to the number of successes. We can visualise this in the following plot (the code will be introduced later). The height of each point line refers to the binomial probability of observing 0 to 12 successes in 12 trials assuming, p = 0.85. Each probability was calculated using the binomial PMF formula above. We can see 0 - 7 success all have probabilities below 0.05, while 10 and 11 successes have probabilities approximately 0.30. This makes sense as the expected, or mean value = np = 12*.85 = 10.2. If we added all the probabilities for each point line together, they would sum to 1. This is known as the total probability law. The $Pr(X = 10)$ has been coloured as a red line in the plot. We can quickly see $Pr(X = 10)$ will be about .30.



Now we could use the formula given above to calculate the exact probability...

$$Pr(X = 10) = \binom{12}{10}.85^{10}(1 - .85)^{12-10}$$

However, using an R function will do this a lot quicker and far more accurately. In R, we use the `dbinom(x, size, prob)` function. This function has three arguments:

- **x** : The value for k, or number of successes
- **size** : The number of trials for each experiment
- **prob** : The probability of success (p)

For Question 1. We type the following command into R:

```
dbinom(x = 10, size = 12, prob = 0.85)
```

```
## [1] 0.2923585
```

The answer is found to be $Pr(X = 10) = 0.29$. Given that the vaccine is 85% effective, it would be relatively common to find 10 successes in 12 randomly sampled people. In other words, there is a 29% chance that the vaccine will work for exactly 10/12 people.

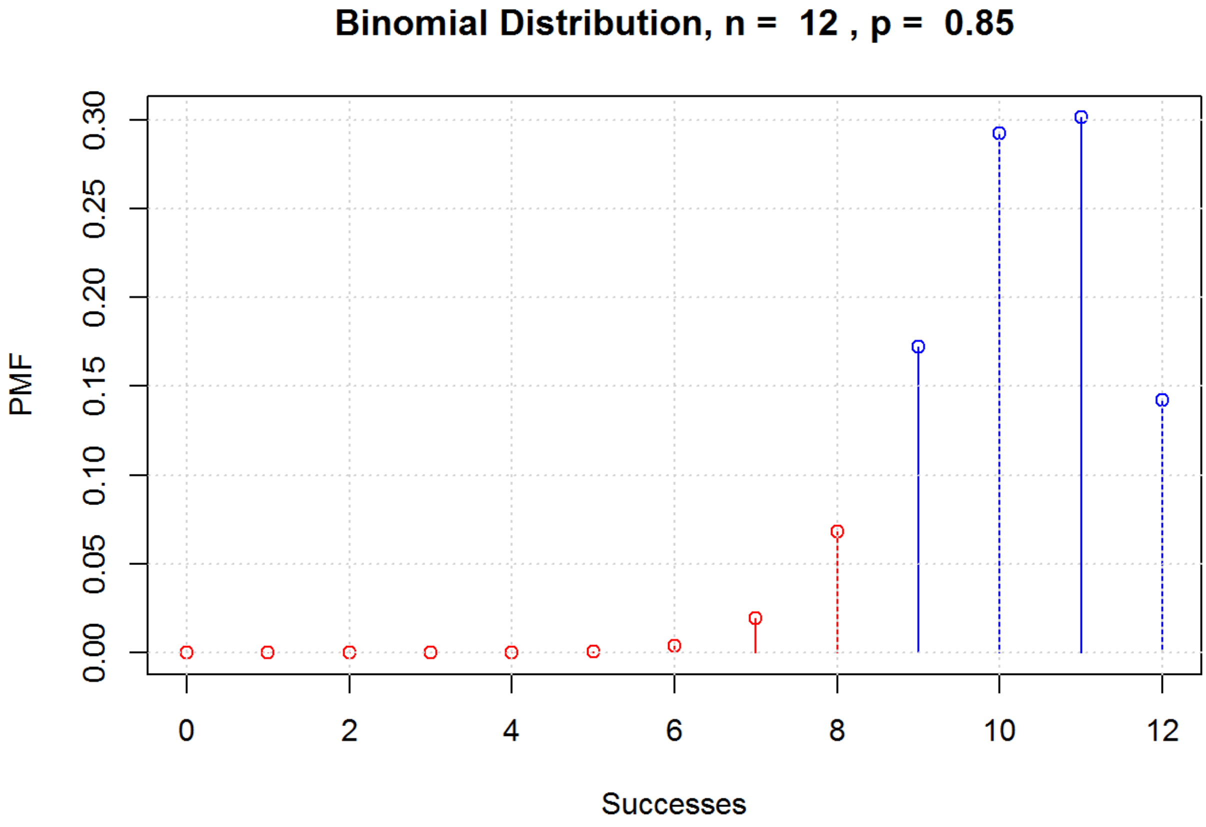
If you're interested in reproducing the plot above, you can use the following R code. Its a little lengthy, but it should mostly make sense.

```
# Set binomial parameters.

n <- 12
p = .85

# Define PMF to highlight - Pr(X < x), Pr(X > x), or Pr(a < x < b)
# Leave blank "" for no highlights
x <- 10
a <- ""
b <- ""



# Set sequence of x values to plot
Successes <- seq(0,n)

# Calculate PMF
PMF <- dbinom(x = Successes, size = n, prob = p)

# Define points to highlight in plot

highlight <- ifelse(Successes <= b &
                      Successes >= a |
                      Successes == x, "red", "blue")

# Plot PMF
plot(Successes, PMF, type = "p",
     main = paste("Binomial Distribution, n = ",n,", p = ",p), col = highlight)
lines(Successes,PMF, type = "h", col = highlight)
grid()
```

$Pr(X \leq k)$

1.2. What is the probability that the vaccine will work for 8 or less people?

This is a slightly different question. We are asked to find $Pr(X = 0) + Pr(X = 1) + \ldots + Pr(X = 8)$, or simply $Pr(X \leq 8)$. This is known as a cumulative probability. Visually, this question looks like the following plot. The red shaded point lines refer to $Pr(X \leq 8)$. If we add the probabilities of each of these point lines, we can solve the problem.

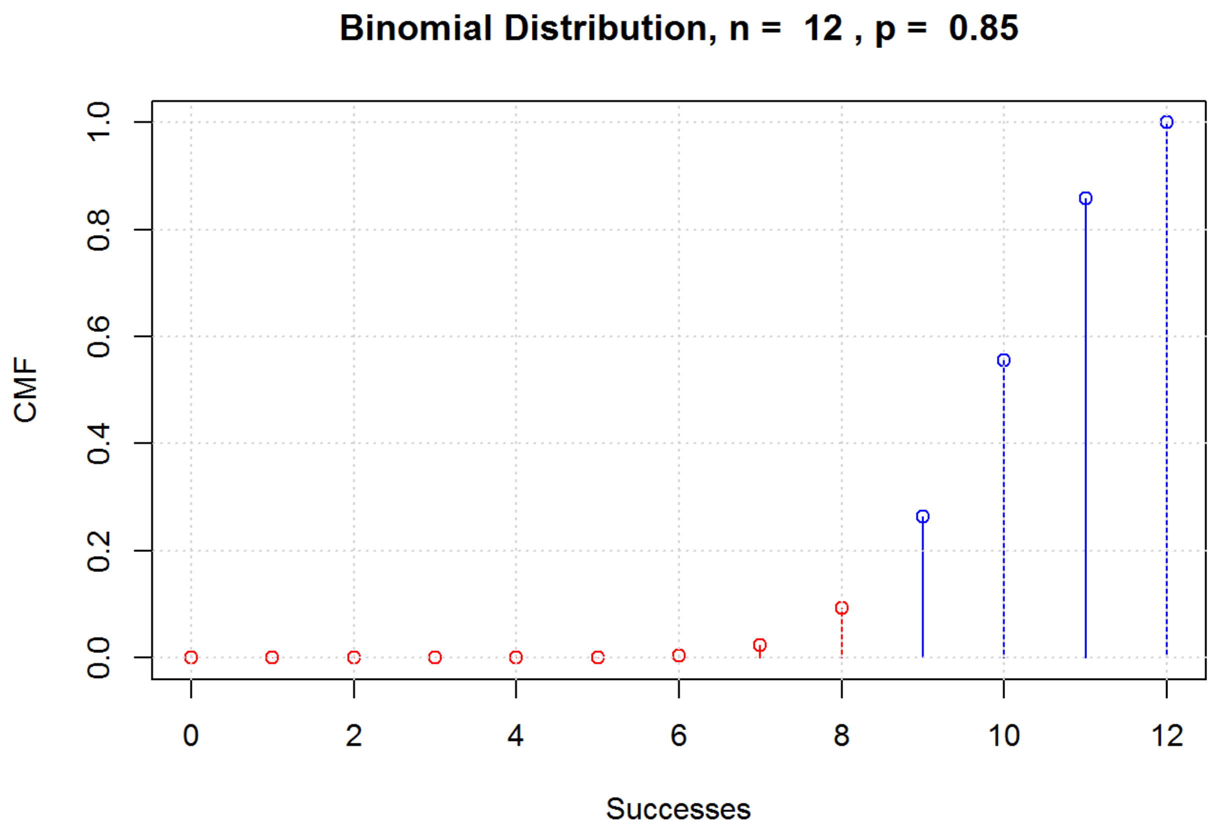


Using the `dbinom()` function:

```
dbinom(x = 0:8, size = 12, prob = .85) %>% sum()
```

```
## [1] 0.09220633
```

Alternatively, we can visualise the **cumulative masss function** (CMF), which plots, $Pr(X \leq x)$. For example: $Pr(X \leq 1) = Pr(X = 0) + Pr(X = 1)$, or $Pr(X \leq 8) = Pr(X = 0) + Pr(X = 1) + \ldots + Pr(X = 8)$ The CMF for the binomial distribution with $n = 12, p = 0.85$ is visualised below. As this is a CMF, it is bound between a probability of 0 and 1. To find $Pr(X \leq 8)$, we look for 8 on the x axis. We can’t see the exact value, but we know it’s going be be less than 0.1.



Easy to do in R using a slightly different `pbinom()` function:

```
pbinom(q = 8, size = 12, prob = 0.85, lower.tail = TRUE)
```

```
## [1] 0.09220633
```

which we recall is exactly the same as:

```
dbinom(x = 0:8, size = 12, prob = .85) %>% sum()
```

```
## [1] 0.09220633
```

We use `q` instead of `x` and we add `lower.tail = TRUE` to ensure we calculate $Pr(X \leq x)$ and not $Pr(X > x)$, which would be specified as `lower.tail = FALSE`. The answer is .092. Once again, it would be unusual to observe the vaccine only working for 8 or less people out of 12, given $p = 0.85$.

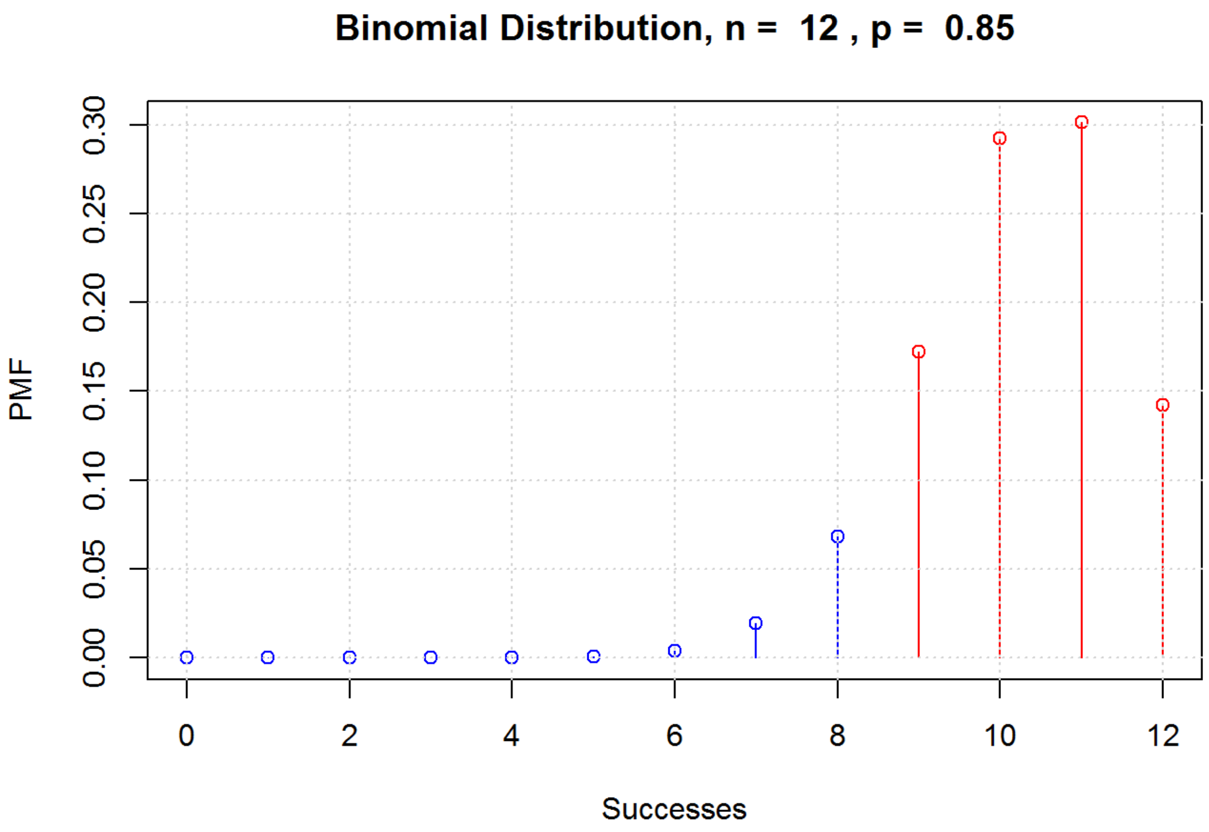
Here's the code for the CMF plot above.

```r
# Set binomial parameters.

n <- 12
p = .85

# Define CMF to highlight - Pr(X < x), Pr(X > x), or Pr(a < x < b)
# Leave blank "" for no highlights
x <- ""
a <- 0
b <- 8

# Set sequence of x values to plot
Successes <- seq(0,n)

# Calculate CMF
CMF <- pbinom(q = Successes, size = n, prob = p)

# Define points to highlight in plot

highlight <- ifelse(Successes <= b &
                      Successes >= a |
                      Successes == x, "red", "blue")


# Plot CMF
plot(Successes, CMF, type = "p",
     main = paste("Binomial Distribution, n = ",n,", p = ",p), col = highlight)
lines(Successes,CMF, type = "h", col = highlight)
grid()
```

$Pr(X > k)$

1.3. What is the probability that the vaccine will work in more than 8 random people?

Once again, a slightly different question. This time we need to find, $Pr(X > 8)$. According to the total probability rule, $Pr(0 \leq X \leq 12) = 1$. The `pbinom()` function in R, by default, gives us $Pr(X \leq 8)$, therefore, $Pr(X > 8) = 1 - Pr(X \leq 8)$. In other words, $Pr(X > 8) = Pr(X = 9) + Pr(X = 10) + Pr(X = 11) + Pr(X = 12)$. Let's visualise this question. The $Pr(X > 8)$ is shaded in red.



```
pbinom(q = 8, size = 12, prob = 0.85, lower.tail = FALSE)
```

```
## [1] 0.9077937
```

or…

```
1-pbinom(q = 8, size = 12, prob = 0.85, lower.tail = TRUE)
```

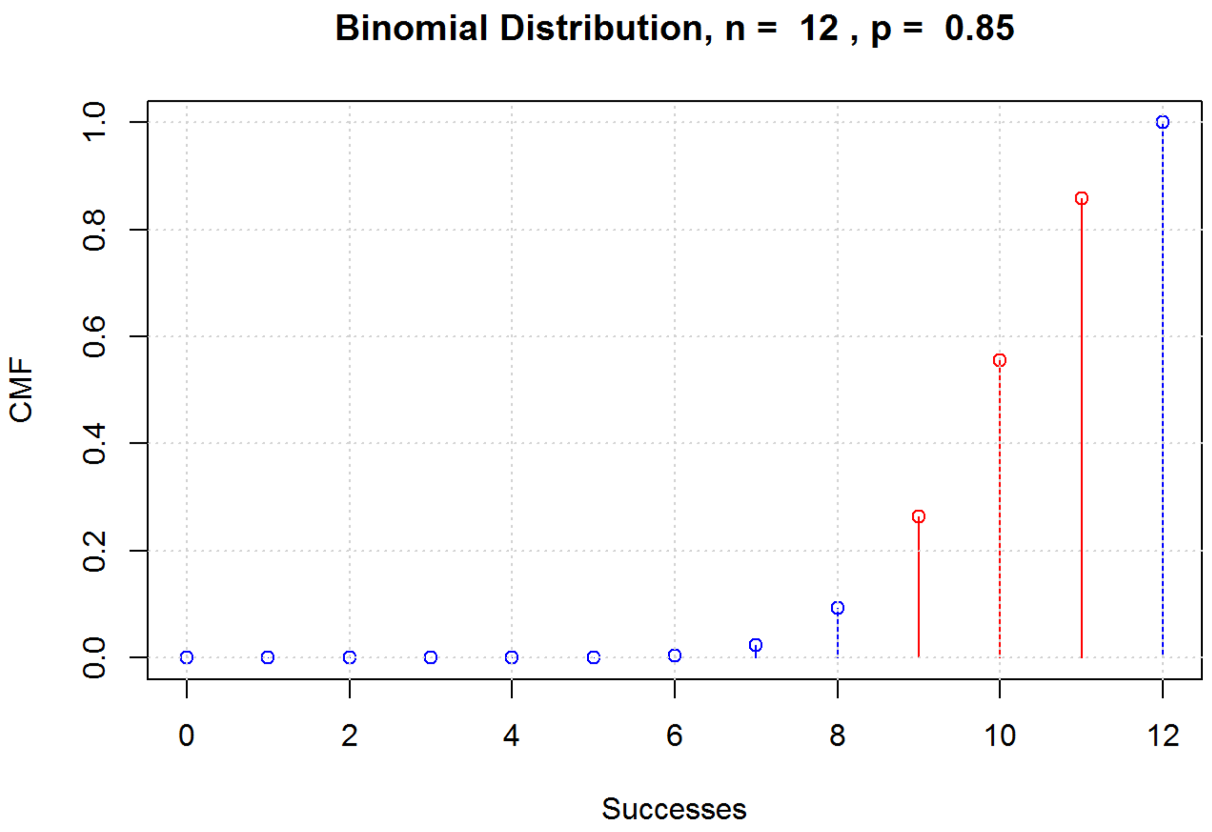
```
## [1] 0.9077937
```

Note the use of `lower.tail = FALSE` to calculate $Pr(X > 8)$. The answer should be .908. Now adding $Pr(X \leq 8) + Pr(X > 8) = .092 + .908 = 1$! This confirms the total probability rule.

$Pr(a \leq k \leq b)$

1.4. What is the probability that the vaccine will work for between 9 to 11 people?

Now this is a little tricky. We are asked to find $Pr(9 \leq X \leq 11)$. Let's visualise it.



We could just add, $Pr(X=9)+Pr(X=10)+Pr(X=11)$. That's a perfectly acceptable answer.

```
dbinom(x = 9:11, size = 12, prob = .85) %>% sum()
```

```
## [1] 0.7655519
```

We can also use some tricky subtraction:

```
pbinom(11, 12, 0.85, lower.tail = TRUE) - pbinom(8, 12, 0.85, lower.tail
   = TRUE)
```

```
## [1] 0.7655519
```

This formula takes the cumulative probability, $Pr(X \leq 11)$ and subtracts $Pr(X \leq 8)$. The probability left over includes $Pr(9 \leq X \leq 11)$.

$E(x)$ & $Var(x)$

1.5. What is the expected value and standard deviation of the binomial distribution with $n=12, p=0.85$?

Looking back at the formula, we can write some quick R code to answer this question. First we create two objects called `n` and `p`. We assign the parameters from the example to these objects.

```
n <- 12
p <- 0.85
```

Now we can call these objects into our formula.

```
# Mean or expected value
n*p
```

```
## [1] 10.2
```

```
# SD
sqrt(n*p*(1-p))
```

```
## [1] 1.236932
```

The advantage of this approach is if we need to change `n` and `p` to work on a different example. We simply change the object values (e.g. `n <- 50`, `p <- 0.5`) and re-run the formula. There is no need to rewrite the values into the formula. This can save a lot of time in the long run.

Poisson Distribution

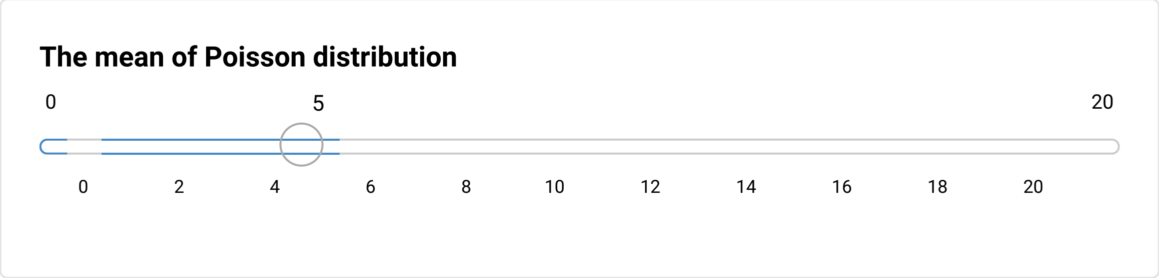
The Poisson distribution is used to model the occurrence of discrete events over a specific period of time, t. The Poisson distribution has one parameter, λ, which is the expected, $E(x)$, or mean, μ, number of events in a unit of time. For example, the expected daily number of patients for a doctor during winter might be $\lambda = 16$. λ can be adjusted to take into account different time periods using $\mu = \lambda t$. For example, the mean number of patients for the same doctor over a week in winter is $\mu = \lambda t = 16 * 7 = 112$. This assumes λ is time constant. The mean, μ, and variance, σ^2, of a Poisson random variable is simply λ. The Poisson distribution has the following probability mass function:

$$Pr(X = k) = \frac{e^{-\mu}\mu^k}{k!}$$

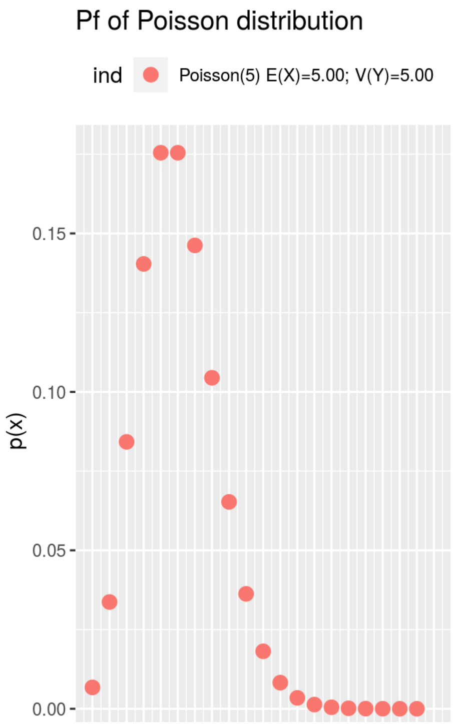
where, k = the number of occurrences, e = the exponential function, and $!$ is a factorial. You can visualise and interact with this distribution using Dr Haydar Demirhan's Poisson Shiny App.

Poisson Distribution

Developed by Dr Haydar Demirhan - haydar.demirhan@rmit.edu.au (mailto:haydar.demirhan@rmit.edu.au)



The mean of Poisson distribution



$E(x)$ & $Var(x)$

The mean, μ,or expected value, $E(X)$, variance, σ^2, and standard deviation, σ, for a Poisson distributed variable are as follows:

$$\mu = E(X) = \lambda$$

$$\sigma^2 = Var(X) = \lambda$$

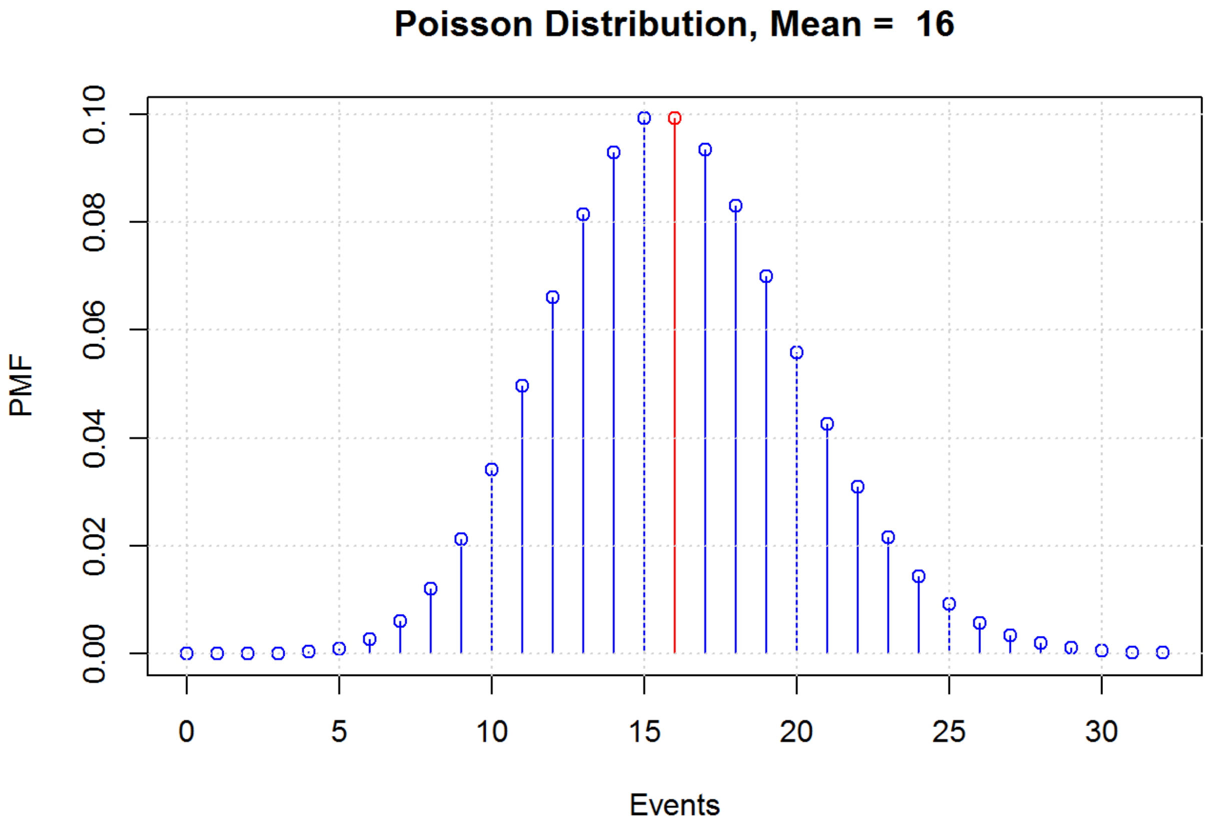
$$\sigma = \sqrt{\lambda}$$

We will work through some Poisson problems using R.

$Pr(X = x)$

2.1. What is the probability the doctor will see exactly 16 patients in a given day?

This is a simple one. We need to find $Pr(X = 16)$. The red area in the following plot highlights the probability visually. The plot represents a PMF of the Poisson distribution with $\lambda = 16$.



In R, use the `dpois(x, lambda)` function, which has the following options:

- **x**: This is the x value you want to look up under the Poisson distribution
- **lambda**: This the mean rate over time, μ

Therefore, we use the following formula:

```
dpois(x = 16, lambda = 16)
```

```
## [1] 0.09921753
```

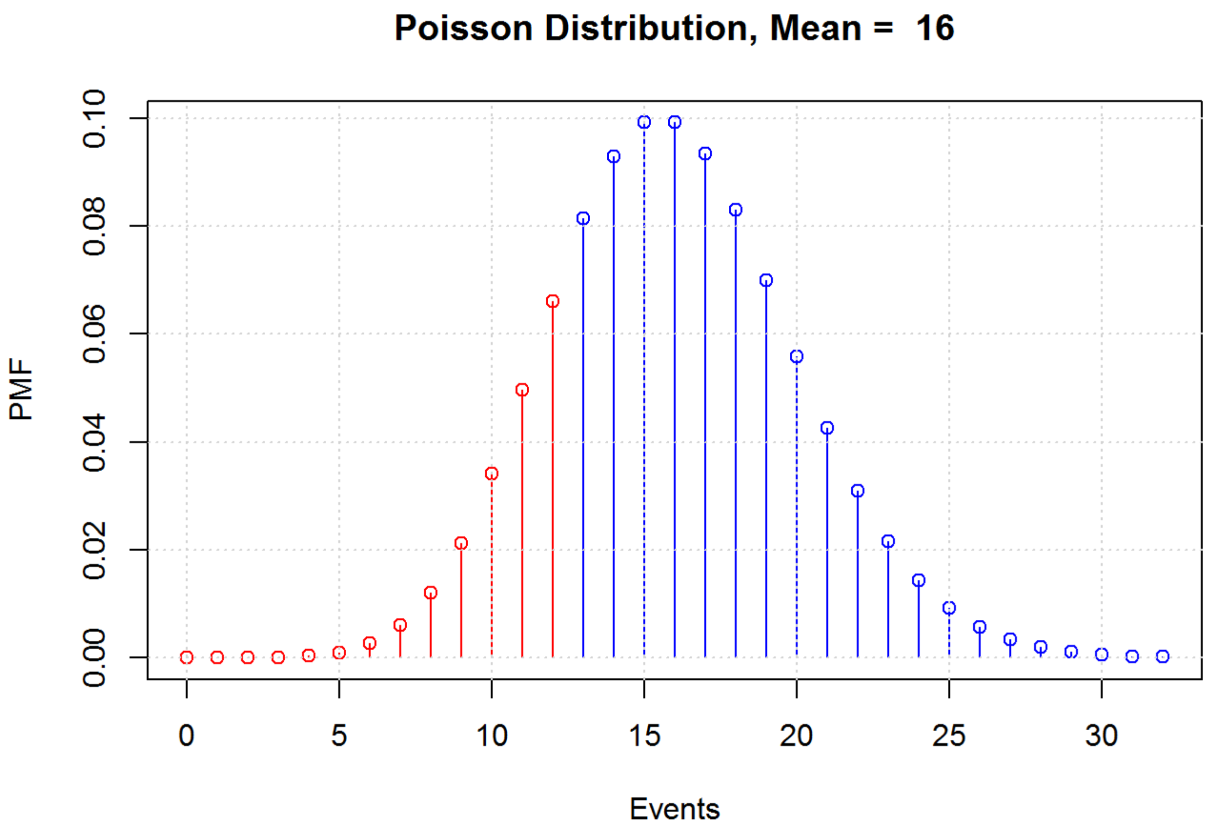
The function `dpois()` computes the probability of observing a particular number of occurrences for a Poisson distribution. The answer to question 1 is $Pr(X = 16) = .099$. We can use the following code to replicate the visualisation above:

```
# Set Poisson parameters.

lambda <- 16
time_multiplier <- 1
mu <- lambda*time_multiplier

# Define PMF to highlight - Pr(X < x), Pr(X > x), or Pr(a < x < b)
# Leave blank "" for no highlights
x <- 16
a <- ""
b <- ""

# Set sequence of x values to plot
Events <- seq(ifelse(sign(round(mu-sqrt(mu)*4,0))==-1,0,
                     round(mu-sqrt(mu)*4,0)),
              round(mu+sqrt(mu)*4,0))

# Calculate PMF
PMF <- dpois(x = Events, lambda = mu)

# Define points to highlight in plot

highlight <- ifelse(Events <= b &
                      Events >= a |
                      Events == x, "red", "blue")

# Plot PMF
plot(Events, PMF, type = "p",
     main = paste("Poisson Distribution, Mean = ",mu), col = highlight)
lines(Events, PMF, type = "h", col = highlight)
grid()
```

$Pr(X \leq x)$

2.2. What is the probability the doctor will see 12 or less patients in a given day?

This time we’re asked Pr(X ≤ 12). Visually…

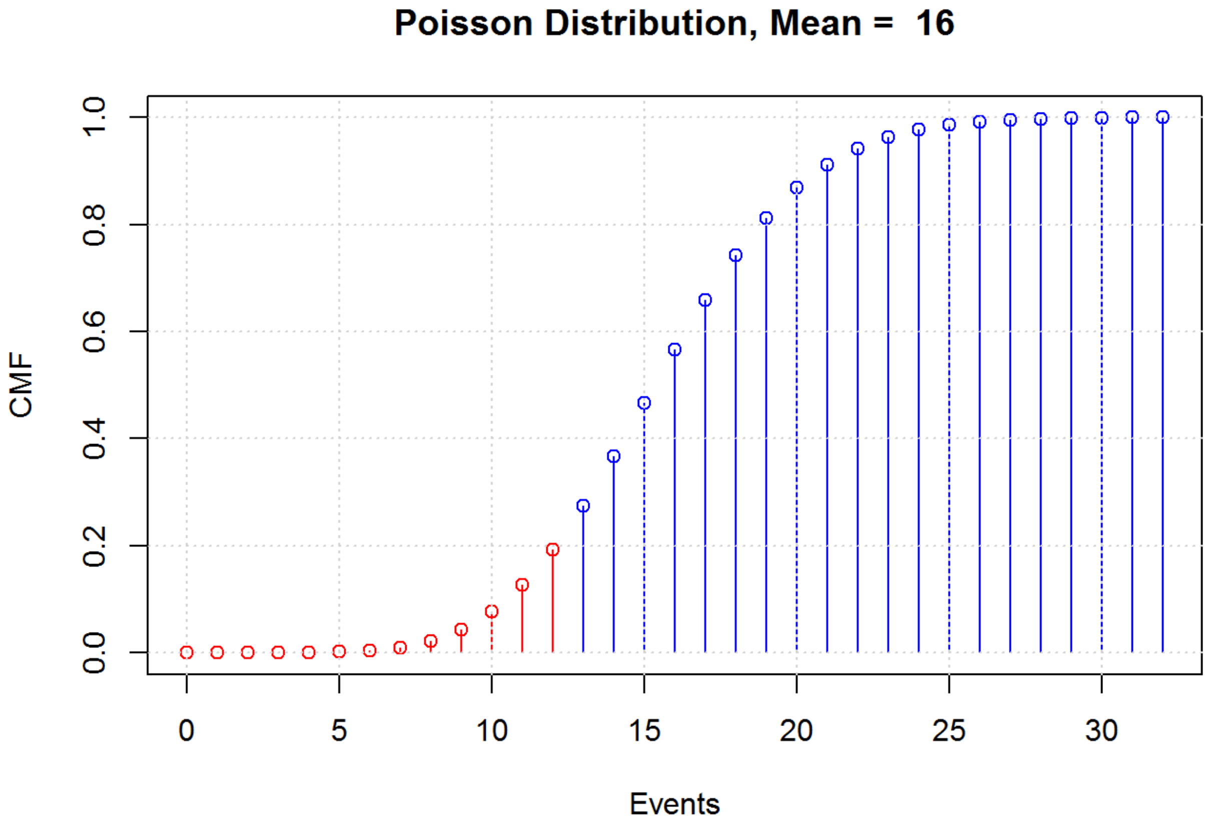


In R…

```
ppois(q = 12, lambda = 16)
```

```
## [1] 0.1931215
```

Notice that we need to use the function `ppois()`, which gives $Pr(X <= x)$, rather than the earlier function `dpois()`, to solve for the cumulative probability, CMF, of the doctor seeing 0 to 12 patients in a given day. Remember the difference between these two functions. Visually, the Poisson CMF looks like the following:

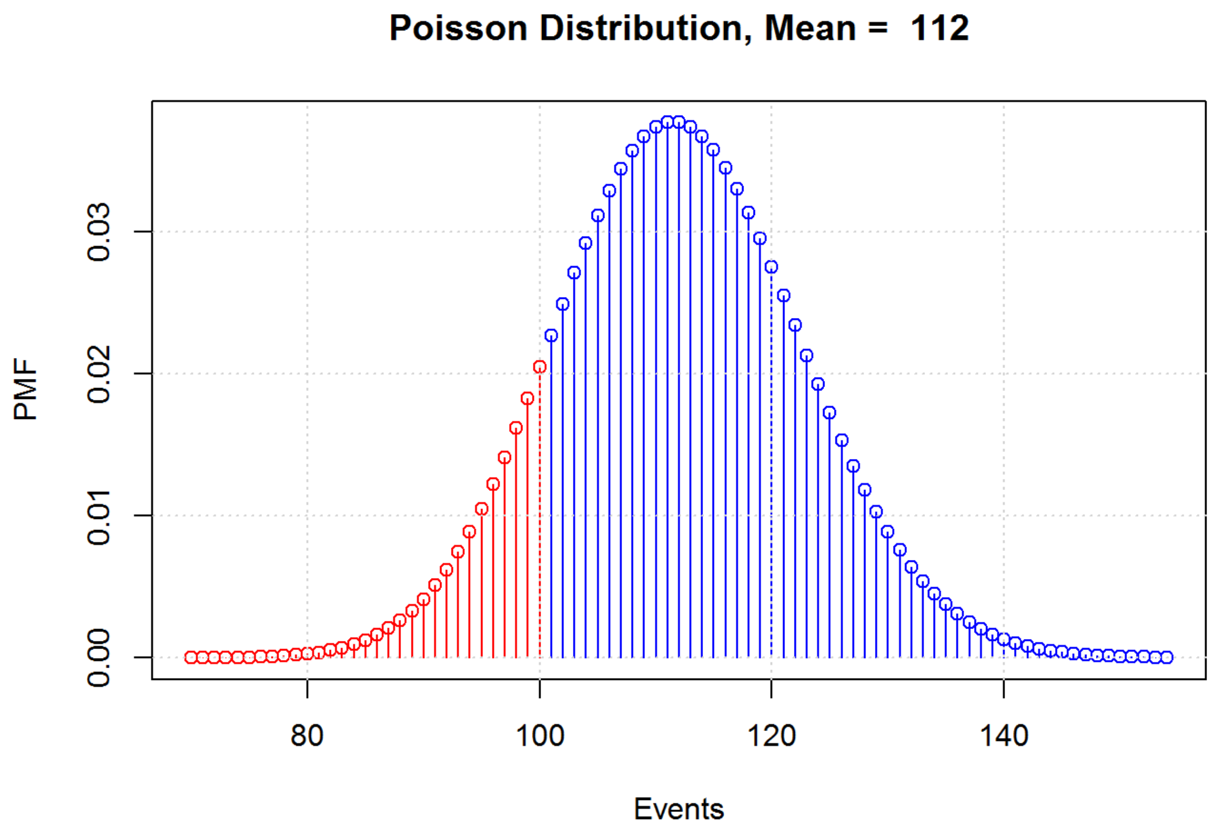


The code for generating the CMF plot above was as follows:

```
# Set Poisson parameters.

lambda <- 16
time_multiplier <- 1
mu <- lambda*time_multiplier

# Define PMF to highlight - Pr(X < x), Pr(X > x), or Pr(a < x < b)
# Leave blank "" for no highlights
x <- ""
a <- 0
b <- 12

# Set sequence of x values to plot
Events <- seq(ifelse(sign(round(mu-sqrt(mu)*4,0))==-1,0,
                     round(mu-sqrt(mu)*4,0)),
              round(mu+sqrt(mu)*4,0))

# Calculate CMF
CMF <- ppois(q = Events, lambda = mu)

# Define points to highlight in plot

highlight <- ifelse(Events <= b &
                      Events >= a |
                      Events == x, "red", "blue")

# Plot PMF
plot(Events, CMF, type = "p",
     main = paste("Poisson Distribution, Mean = ",mu), col = highlight)
lines(Events, CMF, type = "h", col = highlight)
grid()
```

2.3. What is the probability that the doctor will see less than or equal to 100 patients in a week?

We need to adjust the mean value to $\mu = 16 * 7 = 112$. The questions asks $Pr(X \leq 100)$. Visually...



Using R:

```
ppois(q = 100, lambda = 16*7)
```

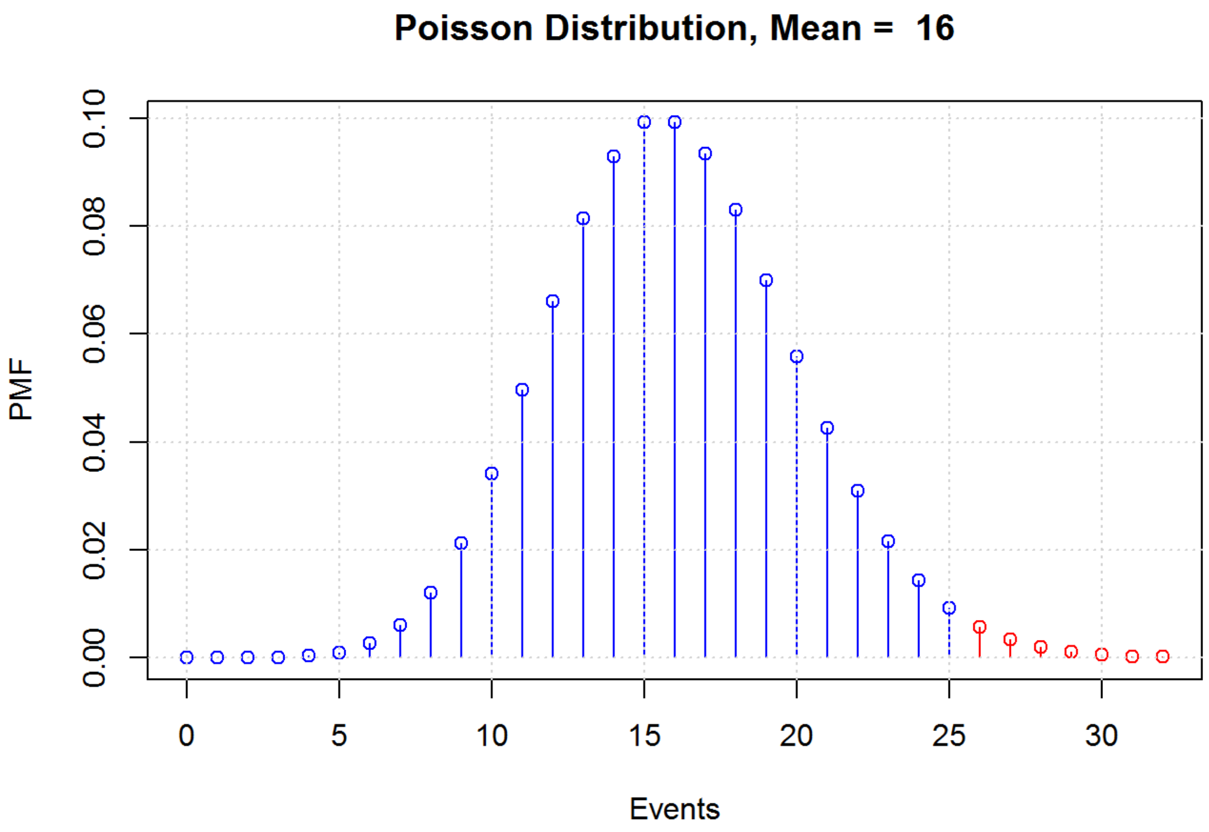
```
## [1] 0.1378483
```

The answer will be $Pr(X \leq 100) = .138$.

$Pr(X > x)$

2.4. What is the probability that the doctor will see more than 25 patients in a day?

This time we are asked $Pr(X > 25)$. Visually...



We can solve this by solving $Pr(X > 25) = 1 - Pr(X \leq 25)$.

```
1 - ppois(q = 25, lambda = 16)
```

```
## [1] 0.01311856
```

or…

```
ppois(q = 25, lambda = 16, lower.tail = FALSE)
```

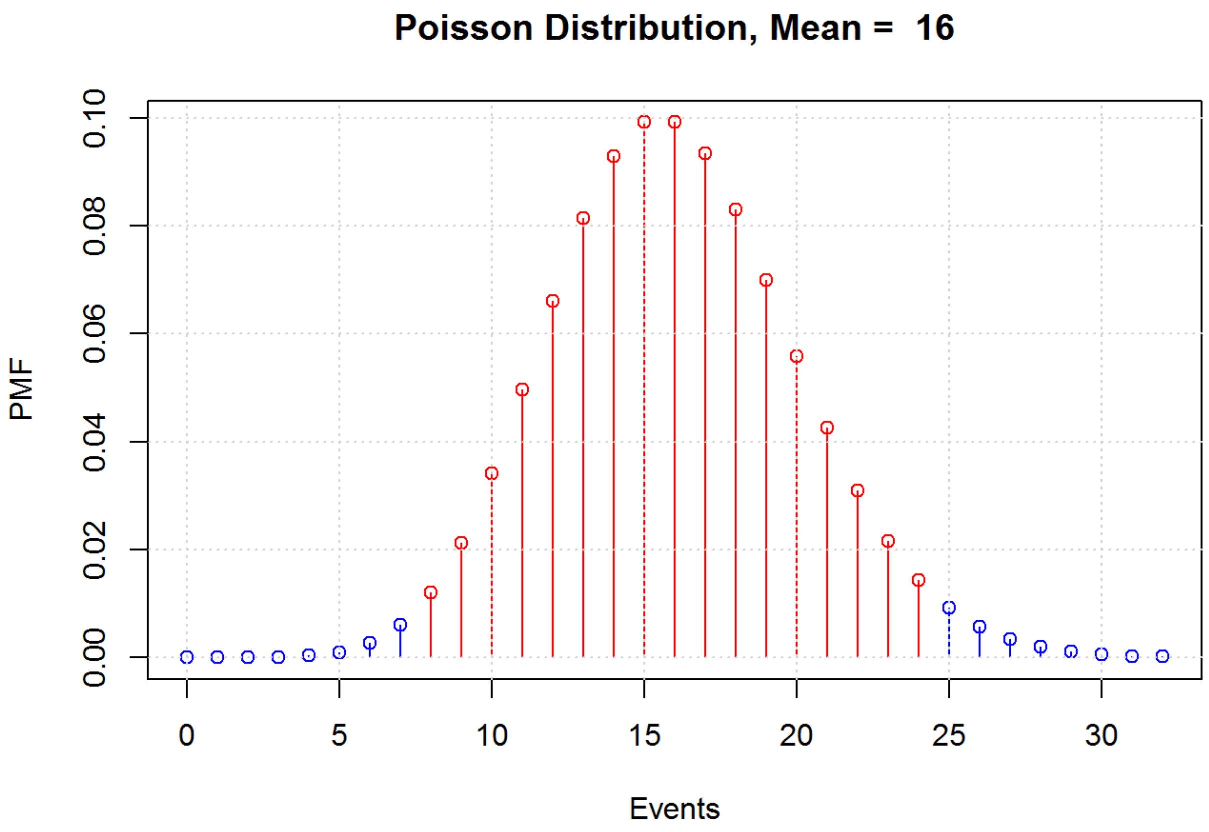
```
## [1] 0.01311856
```

The probability is very small at $Pr(X > 25) = .013$.

$Pr(a \leq X \leq b)$

2.5. What is the probability that the doctor will see between 8 to 24 patients in a given day?

This is a little tricky. We need to find $Pr(8 \leq X \leq 24)$. Visually…



We could just add $Pr(X = 8) + Pr(X = 9) + \ldots + Pr(X = 24)$. However, this is a little slow. Try...

```
ppois(q = 24, lambda = 16) - ppois(q = 7, lambda = 16)
```

```
## [1] 0.9676847
```

Much quicker! The answer is found to be $Pr(8 \leq X \leq 24) = 0.968$. The reason we put $x = 7$ into the second part of the formula is that we want to include $Pr(X = 8)$ in the left over cumulative probability.

Normal Distribution

The continuous, normal, or Gaussian, distribution is ubiquitous in the field of statistics. Many random variables exhibit a normal distribution shape or, at least, do so approximately. Continuous variables that are known to have a normal distribution make determining the probability of certain events easy to calculate. Probabilities for normal distributions can be readily obtained from tables in the back of most statistics textbooks or functions built into spreadsheets and statistical packages. We will focus on using technology to calculate these probabilities.

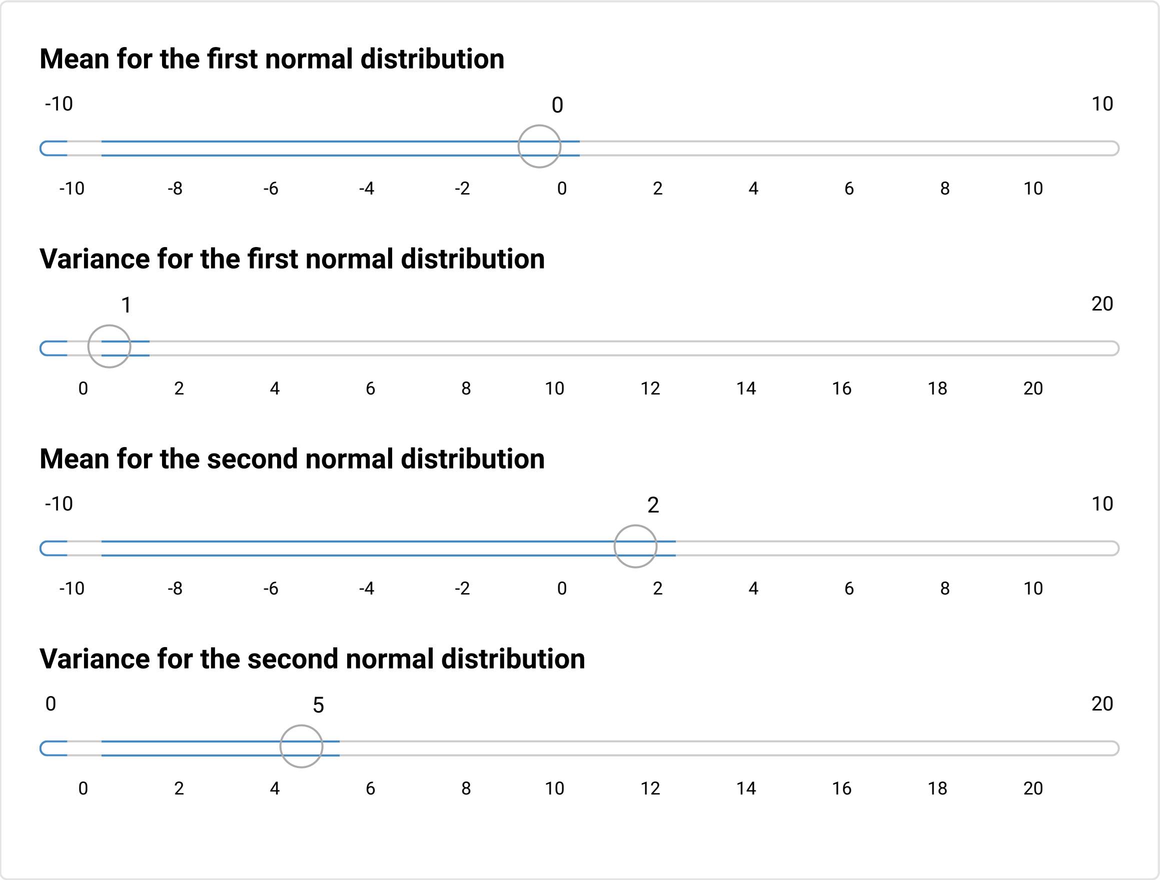
To illustrate the application of the normal distribution, we will consider looking at an example involving IQ or “intelligence” scores. IQ scores are believed to have a normal distribution in the population. Most people score close to the average, while few people score really low (e.g. those with learning disabilities) or really high (e.g. geniuses). The normal distribution has two parameters, a mean, μ, and a standard deviation, σ. For IQ we would denote the theoretical normal distribution as follows:

$$IQ \sim N(\mu, \sigma) \sim N(100, 15)$$

You can visualise and interact with two examples of normal distributions using Dr Haydar Demirhan’s Normal Distribution Shiny App.

Normal Distribution

Developed by Dr Haydar Demirhan - haydar.demirhan@rmit.edu.au (mailto:haydar.demirhan@rmit.edu.au)



Mean for the first normal distribution

-10 0 10

Variance for the first normal distribution

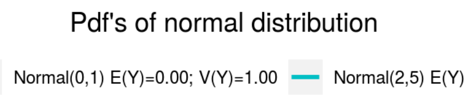
1 20

Mean for the second normal distribution

-10 2 10

Variance for the second normal distribution

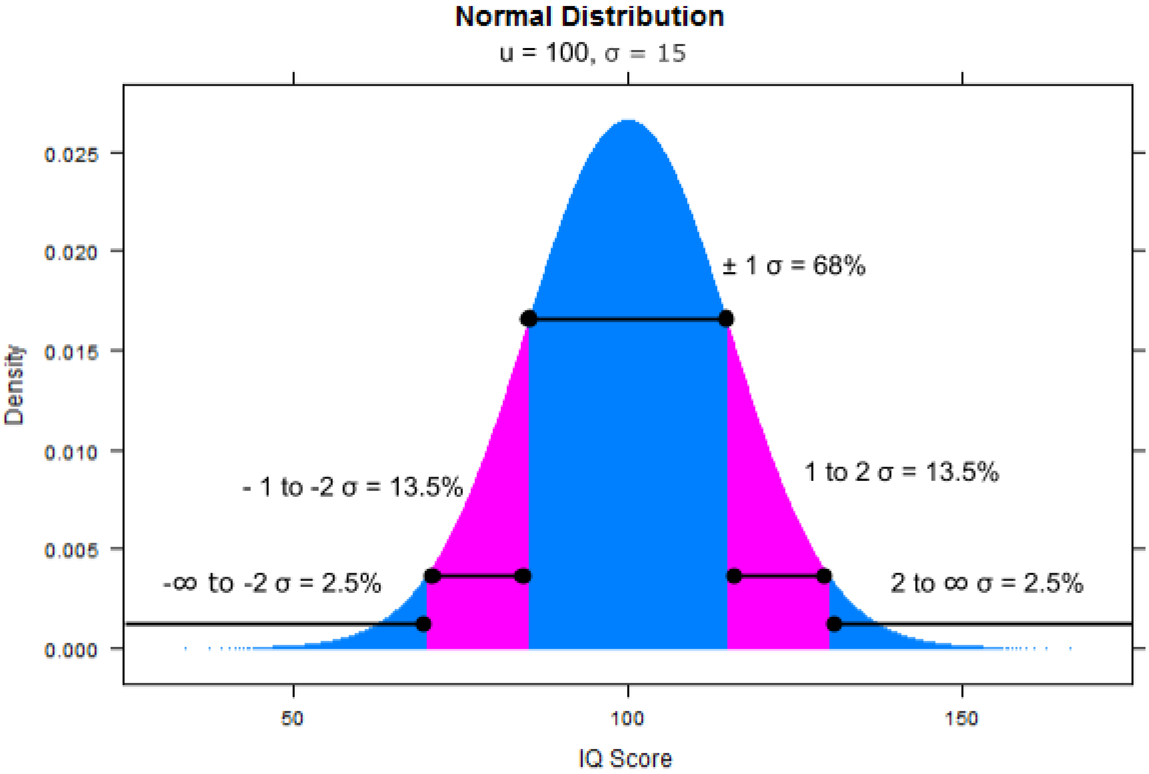
0 5 20



Pdf's of normal distribution

The following info-graphic shows the theoretical normal distribution of IQ scores. The shaded areas refer to 1 standard deviation, σ. Normal distributions have very specific properties. As you can see from the info graphic, 68% of a normal distribution falls within 1 standard deviation of the mean, $85 < x < 115$. From 1 to 2 standard deviations, $115 < x < 130$, 13.5% of values will fall. As the normal distribution is perfectly symmetric, we can also see that 13.5% of values will fall between -1 and -2 standard deviations, $70 < x < 85$. We also have 2.5% of values falling beyond 2 and -2 standard deviations, $x < 70$ and $x > 130$.

Note that for continuous distributions, there is no distinction between $Pr(X < x)$ and Pr $(X \leq x)$. These two statements are the same because for a continuous distribution the exact probability $Pr(X = x)$ always equals 0. This course will always use the $<$ sign in place of $\leq$ because $<$ is quicker to type!

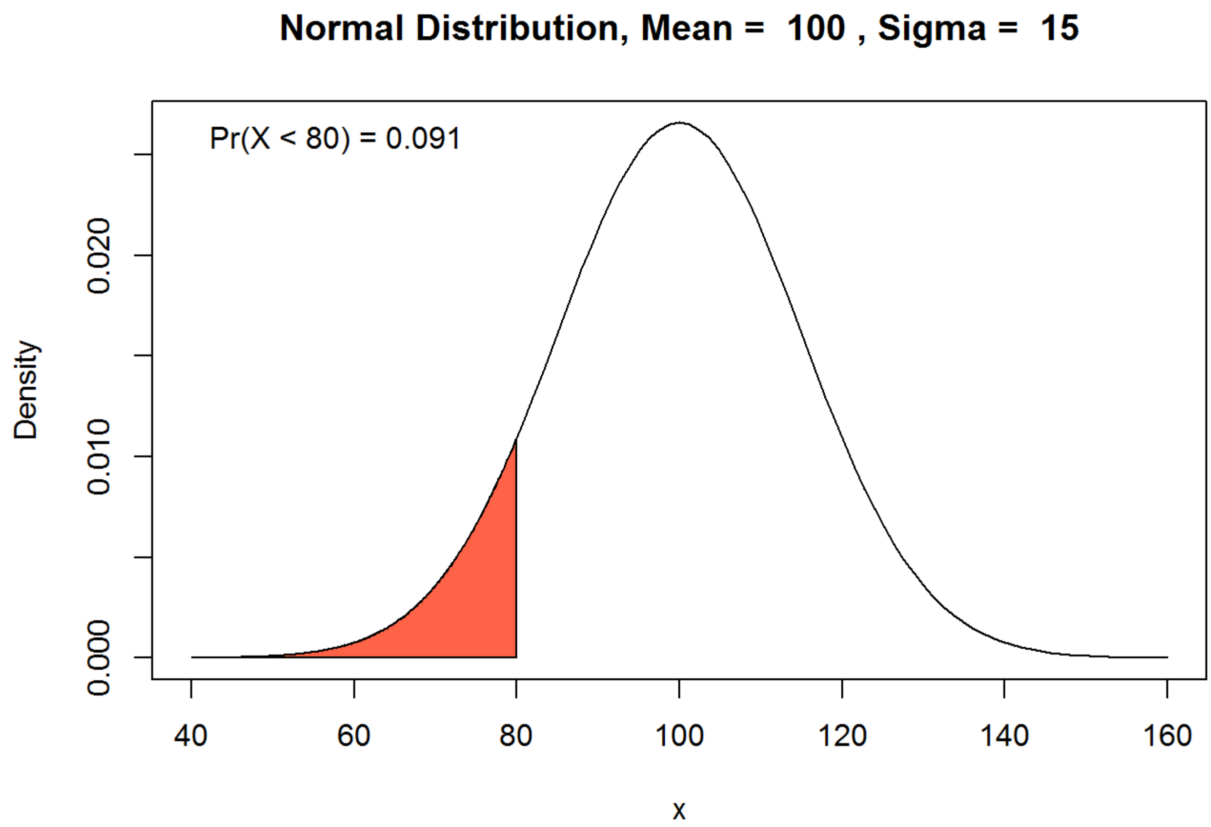


So assuming IQ scores are theoretically normally distributed in the population, what types of questions can we answer? Let's work through some examples to get an idea. Along the way, we will look at using simple functions in R to quickly and effectively answer these questions.

$Pr(X < x)$

1. What is the probability that a random person from the population will have an IQ score less than 80?

The question is asking $Pr(X < 80)$. We use X to represent a measurement for a random variable. If we plot the theoretical normal distribution and shade the area red that represents this probability, it would look like the following:



So what's the best way to find this probability? Use the built in functions of R. Specifically, we will use the function, `pnorm(q, mean , sd , lower.tail = TRUE)`. This function has the following arguments:

- **q**: This is the value you want to look up under the normal distribution.
- **mean**: The population mean.
- **sd**: The population standard deviation (don't confuse this with the variance).
- **lower.tail**: If TRUE = $Pr(X < x)$, the cumulative probability up to x . If false, $Pr(X > x)$ will be returned. `TRUE` is used by default, so it's often ignored.

Therefore, in R, we use the formula:

```
pnorm(q = 80, mean = 100, sd = 15)
```

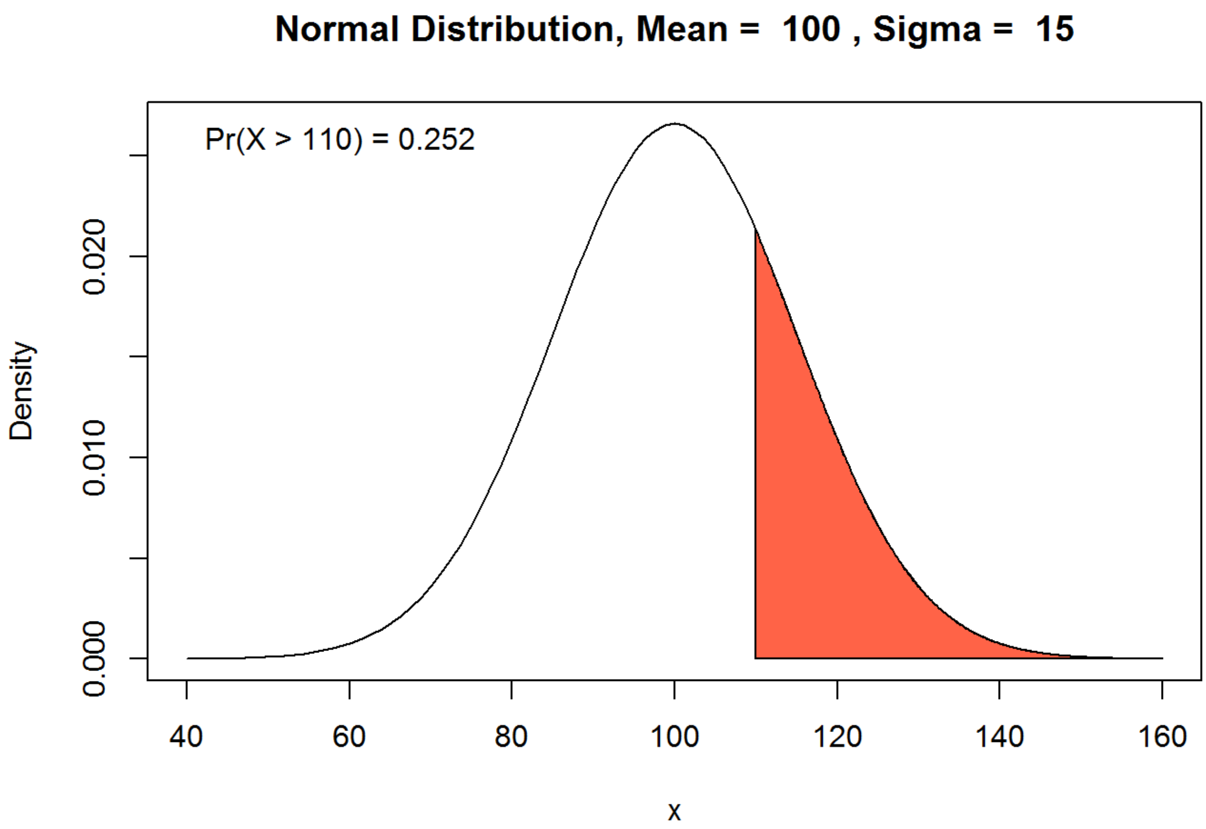
```
## [1] 0.09121122
```

We find $Pr(X < 80) = 0.091$. This means we have a 9.1% chance of randomly selecting a person with an IQ below 80. To reproduce the plot see the code at the end of this section (It's rather long).

$Pr(X > x)$

2. What is the probability that you will randomly select a person from the population with an IQ score above 110?

This time we need to find Pr(X > 110). Visually...



One way to solve this is to use the rule that $Pr(X > x) = 1 - Pr(X < x)$. In R, we would write the formula as either:

```
1 - pnorm(q = 110, mean = 100, sd = 15)
```

```
## [1] 0.2524925
```

or...

```
pnorm(q = 110, mean = 100, sd = 15, lower.tail = FALSE)
```

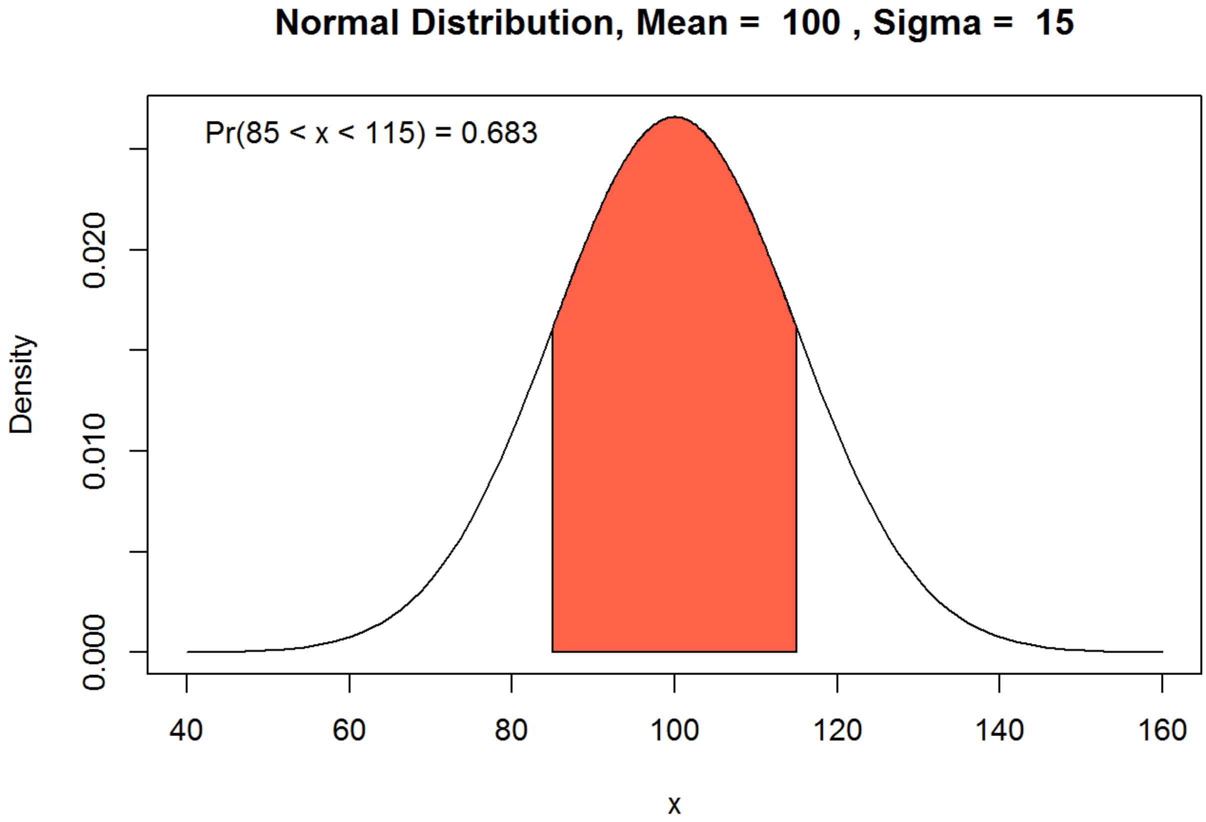
```
## [1] 0.2524925
```

Note the use of `lower.tail = FALSE` in the second formula. This reports $Pr(X > 110)$. The answer was found to be $Pr(X > 110) = .252$.

$Pr(a < x < b)$

3. What is the probability that a randomly selected person from the population will have an IQ score within one standard deviation from the mean?

The population standard deviation is 15 points, therefore, we need to find $Pr(85 < X < 115)$. Visually...



This is getting a little tricky. We can solve this by first calculating $Pr(X < 115)$. We then calculate $Pr(X < 85)$. If we subtract $Pr(X < 85)$ from $Pr(X < 115)$, we are left with $Pr(85 < X < 115)$. This gives us the following useful formula:

$$Pr(a < x < b) = Pr(X < b) - Pr(X < a)$$

In R:

```
pnorm(q = 115, mean = 100, sd = 15) - pnorm(q = 85, mean = 100, sd = 15)
```

```
## [1] 0.6826895
```

The answer is $Pr(85 < X < 115) = .683$. Now look back at the info-graphic. Therefore, there is a 68.3% chance that a person's IQ score will be within one standard deviation from the mean. Simple! Well, sort of...

Find X given percentile

4. A random person is selected from the population and their IQ is found to be in the 90th percentile. What was their IQ?

This is a different question again. We are given a percentile and asked to find the person's actual IQ. Scoring in the 90th percentile means that a person scored equal to or better than 90% of other people in the population. Therefore, we need to solve for x, in $Pr(X < x) = .9$. We need to use a different formula in R. The formula in question is `qnorm(p, mean, sd, lower.tail = TRUE/FALSE)`, which has four arguments.

- **p**: This is the cumulative probability Pr(X < x).
- **mean**: The population mean.
- **sd**: This is the population standard deviation (don't confuse this with the variance).
- **lower.tail**: If `TRUE` = $Pr(X < x)$, the cumulative probability up to x . If `FALSE` , $Pr(X > x)$, `TRUE` is set by default if left blank.

To answer the question, we complete the formula as follows:

```
qnorm(p = 0.9, mean = 100, sd = 15)
```

```
## [1] 119.2233
```

Rounding the answer to a whole number, we find $Pr(X < 119) = .9$. Therefore, if a person scored in the 90th percentile, their IQ was 119.

Find b given a and percentile

5. You are told that 95% of the population's IQ scores fall between a score of 71 and an unsolved upper value. What is the upper value?

The question asks $Pr(71 < X < b) = .95$. This one is also quite tricky. Let's have a think... We can find $Pr(X > 71) = .973$, which tells us how much probability is above a score of 71.

```
1 - pnorm(q = 71, mean = 100, sd = 15)
```

```
## [1] 0.9734024
```

or...

```
pnorm(q = 71, mean = 100, sd = 15, lower.tail = FALSE)
```

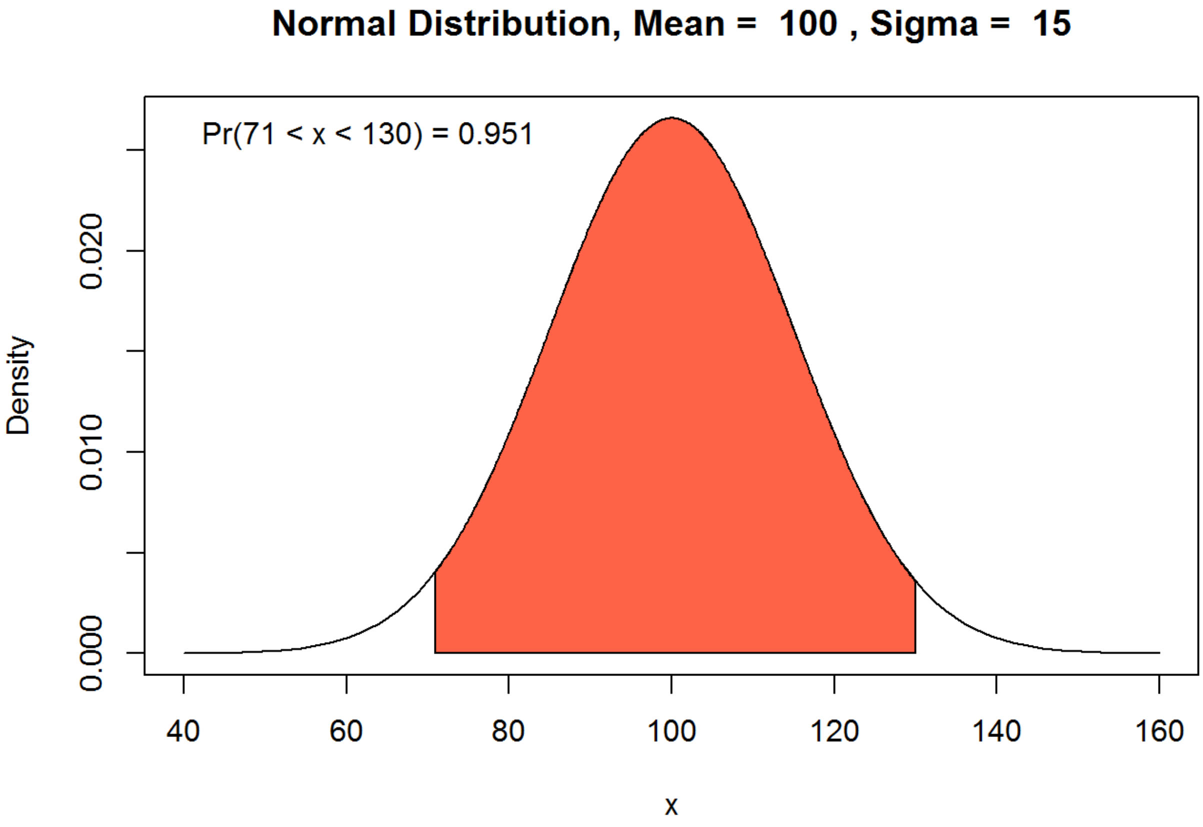
```
## [1] 0.9734024
```

This means that there is $.973 - .95 = .023$ probability left above the upper value x, or $Pr(X > x) = .023$. This x value is the answer to our question. To find this x value, we use:

```
qnorm(p = 1 - .023, mean = 100, sd = 15)
```

```
## [1] 129.9309
```

Note how we have used $1 - .023$, which corresponds to the 97.7th percentile. The answer to the above formula is $Pr(71 < X < 130) = .95$. Visually, the red shaded area corresponds to .95 probability.



We can check our answer using the formula:

```
pnorm(q = 130, mean = 100, sd = 15) - pnorm(q = 71, mean = 100, sd = 15)
```

```
## [1] 0.9506523
```

The answer is confirmed to be .95! Phew, that one was getting tricky.

The Standard Normal Z-Distribution

The standard normal z-distribution is sometimes used in statistics as it allows for probabilities to be looked up using standard tables available in textbooks. We need to know a little bit about the z-distribution, particularly the concept of standardisation, as it appears in a number of other modules. The standard normal distribution has the following properties:

$$z \sim N(0,1)$$

where z refers to a standard normal variable. Let's look at an example. Thinking back to IQ scores, we can convert IQ scores to a standard normal variable using the equation:

$$z = \frac{x - \mu}{\sigma}$$

The z-score divides the difference between an x value and the mean by a population standard deviation. Therefore, z scores are standard deviations. So, suppose we are asked to find $Pr(X > 110)$ using the standard normal distribution. First, we convert the IQ variable, x, to z, i.e. a z-score or standard deviation.

$$z = \frac{x - \mu}{\sigma} = \frac{110 - 100}{15} = \frac{10}{15} = .667$$

Therefore, an IQ score of 110 sits .667 standard deviations above the mean of 100. Simple!

Now we can find $Pr(Z > .667)$ using the `pnorm()` function in R.

```
1 - pnorm(q = .667, mean = 0, sd = 1)
```

```
## [1] 0.2523861
```

or...

```
pnorm(q = .667, mean = 0, sd = 1, lower.tail = FALSE)
```

```
## [1] 0.2523861
```

We find $Pr(X > 110) = 0.252$. You will find the answer to be exactly the same as when we used:

```
1 - pnorm(q = 110, mean = 100, sd = 15)
```

```
## [1] 0.2524925
```

or

```
pnorm(q = 110, mean = 100, sd = 15, lower.tail = FALSE)
```

```
## [1] 0.2524925
```

You will need to use standardisation techniques throughout the course. This section has aimed to give you a heads up. There is nothing special about the z-distribution. It's just a way to standardise a variable so that common probability tables can used. To convert a z-

score back to its original x value use the following equation:

$$x = \mu + \sigma z$$

E.g…

$$x = 100 + 15(.667) = 110$$

Normal Distribution Visualisation Code

This is the code used to create the visualisations of the normal distribution.

```
# Set Normal distribution parameters. Use SD, not Var!

mu <- 100
sd <- 15

# Set values to highlight
# x: Pr(x < x) or Pr(X > x)
# a and b: Pr(a < x < b)

x <- 80
a <- ""
b <- ""

# Highlight area under curve - Pr(X < x) = "less",
# Pr(X > x) = "greater", Pr(a < x < b) = "between"

area <- "less"

# Define area to highlight

if (area == "less") {
  auc <- c(0,dnorm(seq(mu-sd*4,x,sd*0.01), mean = mu, sd = sd),0)
  x_values <- c(mu-sd*4,seq(mu-sd*4,x,sd*0.01),x)
  } else {
  if (area == "greater") {
    auc <- c(0,dnorm(seq(x,mu+sd*4,sd*0.01), mean = mu, sd = sd),0)
    x_values <- c(x,seq(x,mu+sd*4,sd*0.01),mu+sd*4)
  } else {
    if (area == "between") {
      auc <- c(0,dnorm(seq(a,b,sd*0.01), mean = mu, sd = sd),0)
      x_values <- c(a,seq(a,b,sd*0.01),b)
    }
  }
}

# Create probability statement

if (area == "less") {
  prob_statement <- paste("Pr(X < ",x,") = ", round(pnorm(x,mu,sd),3), se
  p = "")
  } else {
    if (area == "greater") {
      prob_statement <- paste("Pr(X > ",x,") = ",
                              round(pnorm(x,mu,sd,lower.tail = FALSE),3),
  sep = "")
      } else {
        if (area == "between") {
```

```
     prob_statement <- paste("Pr(",a," < x < ",b,") = ",
                             round(pnorm(b,mu,sd) - pnorm(a,mu,sd),3
 ), sep = "")}
    else {
     prob_statement <- ""
    }
   }
 }



# Plot density
curve(expr = dnorm(x,mu,sd),
   xlim = c(mu-sd*4,mu+sd*4),
   main = paste("Normal Distribution, Mean = ",mu,", Sigma = ",sd),
   ylab = "Density")
if (area != "") {
 polygon(x = x_values, y = auc, col = "tomato")
 text(x = mu-sd*4, y = dnorm(mu-sd/4,mu,sd), labels = prob_statement, po
 s = 4)
}
```

References