Batch: E2
Roll No.: 16010123325
Experiment No. 3

Title: Modeling real world data using suitable probability distributions

Aim: Exploring the Binomial, Poisson, and Normal Distributions in R

Course Outcome:

CO2

Books/ Journals/ Websites referred:

1. The Comprehensive R Archive Network
2. Posit

Resources used: Teachings in lab,

Theory:

Binomial Distribution

Definition

The binomial distribution is a discrete probability distribution that models the number of successes in a fixed number of independent trials, each with the same probability of success. It is commonly used in scenarios where outcomes are binary (success/failure, yes/no, etc.).

The probability mass function (PMF) for the binomial distribution is given by:

$$f(x) = \mathrm{P}[X = x] = \binom{n}{x} p^x (1-p)^{n-x}$$

Where:

- **n**: Number of trials
- **x**: Number of successes
- **p**: Probability of success in a single trial

dbinom

The function dbinom returns the value of the probability mass function (pmf) of the binomial distribution given a certain random variable x, number of trials (size) and probability of success on each trial (prob).

The syntax for using dbinom is as follows:

dbinom(x, size, prob)

Put simply, dbinom finds the probability of getting a certain number of successes (x) in a certain number of trials (size) where the probability of success on each trial is fixed (prob).

The following examples illustrate how to solve some probability questions using dbinom.

Example 1: Alice makes 60% of her free-throw attempts. If she shoots 12 free throws, what is the probability that she makes exactly 10?

```
> #find the probability of 10 successes during 12 trials where the probability of
> #success on each trial is 0.6
> dbinom(x=10, size=12, prob=.6)
[1] 0.06385228
```

The probability that he makes exactly 10 shots is 0.0639.

Example 2: Bob flips a fair coin 20 times. What is the probability that the coin lands on heads exactly 7 times?

```
> #find the probability of 7 successes during 20 trials where the probability of
> #success on each trial is 0.5
> dbinom(x=7, size=20, prob=.5)
[1] 0.07392883
```

The probability that the coin lands on heads exactly 7 times is 0.0739.

pbinom

The function pbinom returns the value of the cumulative density function (cdf) of the binomial distribution given a certain random variable q, number of trials (size) and probability of success on each trial (prob). The syntax for using pbinom is as follows:

pbinom(q, size, prob)

The following examples illustrate how to solve some probability questions using pbinom.

Example 1: Suppose Kishor scores a strike on 30% of his attempts when he bowls. If he bowls 10 times, what is the probability that he scores 4 or fewer strikes?

```
> #find the probability of 4 or fewer successes during 10 trials where the
> #probability of success on each trial is 0.3
> pbinom(4, size=10, prob=.3)
[1] 0.8497317
```

The probability that he scores 4 or fewer strikes is 0.8497.

Put simply, pbinom returns the area to the left of a given value q in the binomial distribution. If you're interested in the area to the right of a given value q, you can simply add the argument lower.tail = FALSE

pbinom(q, size, prob, lower.tail = FALSE)

Example 2: Ashok flips a fair coin 5 times. What is the probability that the coin lands on heads more than 2 times?

```
> #find the probability of more than 2 successes during 5 trials where the
> #probability of success on each trial is 0.5
> pbinom(2, size=5, prob=.5, lower.tail=FALSE)
[1] 0.5
```

The probability that the coin lands on heads more than 2 times is 0.5.

qbinom

The function qbinom returns the quantile (or the smallest integer) for which the cumulative density function (cdf) of the binomial distribution reaches or exceeds a given probability p. It effectively works as the inverse of pbinom, helping to find the threshold number of successes q for a specified cumulative probability p, given the number of trials (size) and the probability of success on each trial (prob).

The syntax for using qbinom is as follows:

qbinom(q, size, prob)

The following code illustrates a few examples of qbinom in action:

Example 1: Suppose Kishor scores a strike on 30% of his attempts when he bowls. If he bowls 10 times, what is the maximum number of strikes he can score given the cumulative probability is 0.8497?

```
> qbinom(0.8497, size = 10, prob = 0.3)
[1] 4
```

Example 2: Ashok flips a fair coin 5 times. What is the minimum number of heads he can observe to ensure the cumulative probability is greater than 0.5?

```
> qbinom(0.5, size = 5, prob = 0.5)
[1] 2
```

You can use qbinom to find out the percentile of the binomial distribution.

The following code illustrates a few examples:

#find the 10th percentile of a binomial distribution with 10 trials and prob

#of success on each trial = 0.4

```
> qbinom(.10, size=10, prob=.4)
[1] 2
```

#find the 40th percentile of a binomial distribution with 30 trials and prob

#of success on each trial = 0.25

```
> qbinom(.40, size=30, prob=.25)
[1] 7
```

rbinom

The function rbinom generates a vector of binomial distributed random variables given a vector length n, number of trials (size) and probability of success on each trial (prob).

The syntax for using rbinom is as follows:

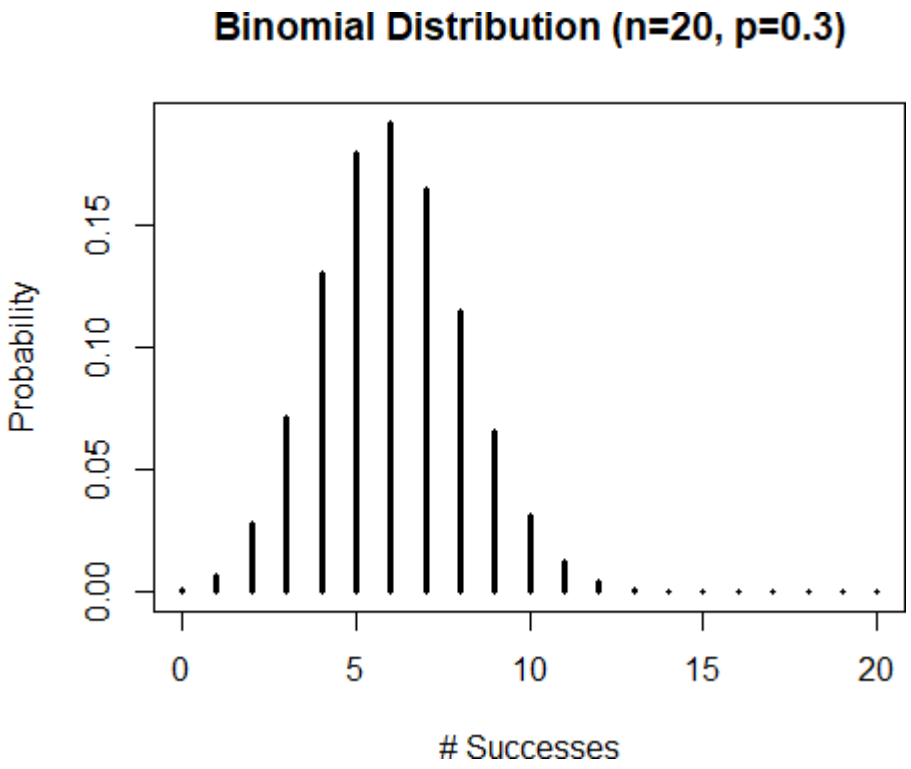
rbinom(n, size, prob)

The following code illustrates a few examples of rbinom in action:

```
> #generate a vector that shows the number of successes of 10 binomial experiments with
> #100 trials where the probability of success on each trial is 0.3.
> results <- rbinom(10, size=100, prob=.3)
>
> print(results)
 [1] 37 35 33 34 44 26 28 27 27 20
```

Visualization

```
> success <- 0:20

  plot(success,dbinom(success,size=20,prob=.3),
       type='h',
       main='Binomial Distribution (n=20, p=0.3)',
       ylab='Probability',
       xlab ='# Successes',
       lwd=3)
```



Poisson Distribution

Definition

The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space, provided these events occur with a known constant rate and independently of the time since the last event.

The probability mass function (PMF) for the Poisson distribution is given by:

$$f(x, \lambda) = P(X = x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}$$

Where:

- **x**: Number of events
- **λ (lambda)**: Average number of events per interval (rate parameter)
- **e**: Euler's number ≈2.718

dpois

The dpois function finds the probability that a certain number of successes occur based on an average rate of success, using the following syntax:

dpois(x, lambda)

where:

x: number of successes

lambda: average rate of success

Example : It is known that a certain website makes 10 sales per hour. In a given hour, what is the probability that the site makes exactly 8 sales?

```
> dpois(x=8, lambda=10)
[1] 0.112599
```

The probability that the site makes exactly 8 sales is 0.112599

ppois

The ppois function finds the probability that a certain number of successes or less occur based on an average rate of success, using the following syntax:

ppois(q, lambda)

where:

q: number of successes

lambda: average rate of success

Example : It is known that a certain website makes 10 sales per hour. In a given hour, what is the probability that the site makes 8 sales or less?

```
> ppois(q=8, lambda=10)
[1] 0.3328197
```

qpois

The qpois function finds the number of successes that corresponds to a certain percentile based on an average rate of success, using the following syntax:

qpois(p, lambda)

where:

p: percentile

lambda: average rate of success

Example : It is known that a certain website makes 10 sales per hour. How many sales would the site need to make to be at the 90th percentile for sales in an hour?

```
> qpois(p=.90, lambda=10)
[1] 14
```

rpois

The rpois function generates a list of random variables that follow a Poisson distribution with a certain average rate of success, using the following syntax:

rpois(n, lambda)

where:

n: number of random variables to generate

lambda: average rate of success

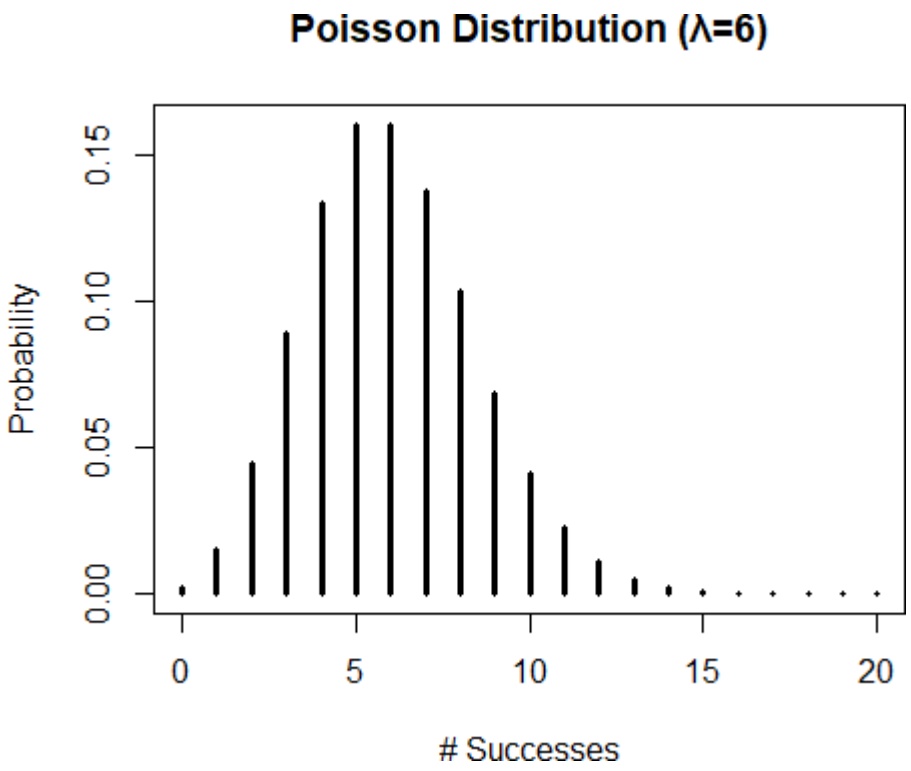
Here's an example of when you might use this function in practice:

Generate a list of 15 sales per hour that follow a Poisson distribution with an average rate of sales being equal to 10.

```
> rpois(n=15, lambda=10)
 [1]  8  7 10  6  7 10 12 14  8 10  8 14  6 10 13
```

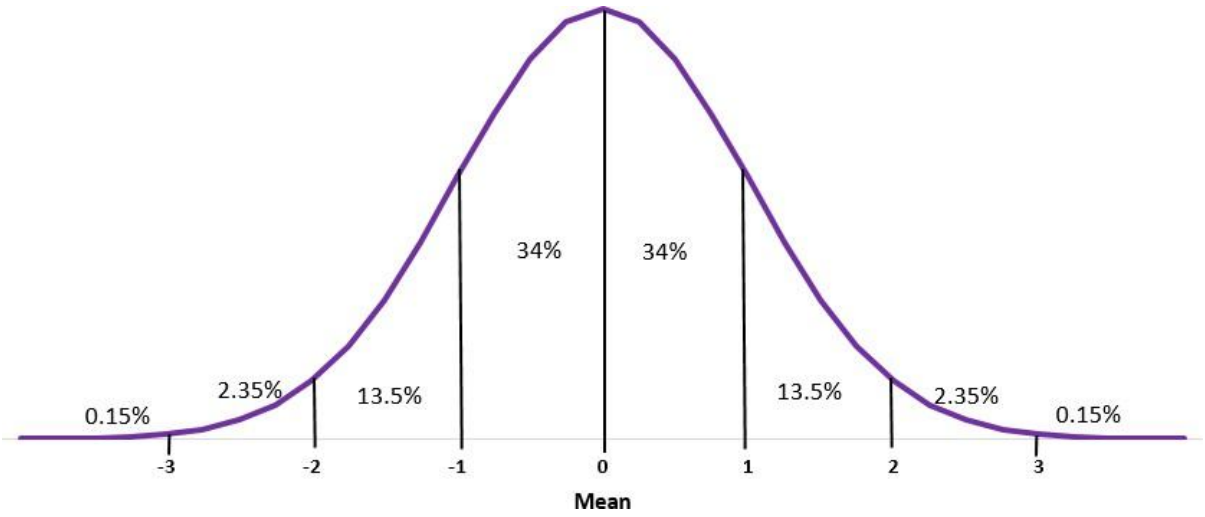
Visualization

```
> # Set parameters
  lambda <- 6  # Mean of the Poisson distribution
  success <- 0:20

  # Plot the Poisson distribution
  plot(success, dpois(success, lambda),
       type='h',
       main='Poisson Distribution (λ=6)',
       ylab='Probability',
       xlab='# Successes',
       lwd=3)
```



Normal distribution

The normal distribution is the most common probability distribution in statistics.



Normal distributions have the following features:

- Bell shape
- Symmetrical
- Mean and median are equal; both are located at the center of the distribution
- About 68% of data falls within one standard deviation of the mean
- About 95% of data falls within two standard deviations of the mean
- About 99.7% of data falls within three standard deviations of the mean

dnorm

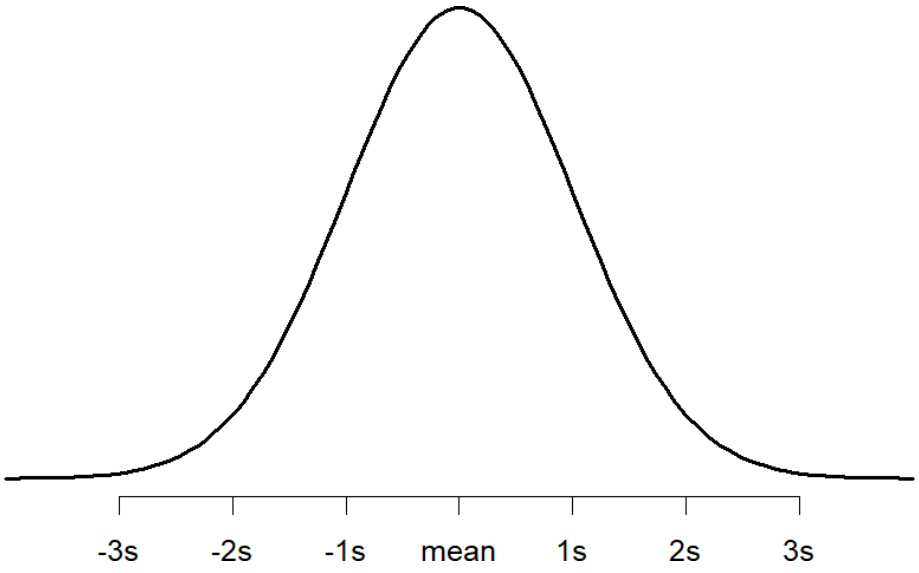
The function dnorm returns the value of the probability density function (pdf) of the normal distribution given a certain random variable x, a population mean μ and population standard deviation σ. The syntax for using dnorm is as follows:

dnorm(x, mean, sd)

```
> #find the value of the standard normal distribution pdf at x=0
> dnorm(x=0, mean=0, sd=1)
[1] 0.3989423
> #by default, R uses mean=0 and sd=1
> dnorm(x=0)
[1] 0.3989423
> #find the value of the normal distribution pdf at x=10 with mean=20 and sd=5
> dnorm(x=10, mean=20, sd=5)
[1] 0.01079819
```

Visualization

```
> #Create a sequence of 100 equally spaced numbers between -4 and 4
  x <- seq(-4, 4, length=100)

  #create a vector of values that shows the height of the probability distribution
  #for each value in x
  y <- dnorm(x)

  #plot x and y as a scatterplot with connected lines (type = "l") and add
  #an x-axis with custom labels
  plot(x,y, type = "l", lwd = 2, axes = FALSE, xlab = "", ylab = "")
  axis(1, at = -3:3, labels = c("-3s", "-2s", "-1s", "mean", "1s", "2s", "3s"))
```



pnorm

The function pnorm returns the value of the cumulative density function (cdf) of the normal distribution given a certain random variable q, a population mean μ and population standard deviation σ. The syntax for using pnorm is as follows:

pnorm(q, mean, sd)

Put simply, pnorm returns the area to the left of a given value x in the normal distribution. If you're interested in the area to the right of a given value q, you can simply add the argument lower.tail = FALSE

pnorm(q, mean, sd, lower.tail = FALSE)

Example 1: Suppose the height of males at a certain school is normally distributed with a mean of μ=70 inches and a standard deviation of σ = 2 inches. Approximately what percentage of males at this school are taller than 74 inches?

```
> #find percentage of males that are taller than 74 inches in a population with
> #mean = 70 and sd = 2
> pnorm(74, mean=70, sd=2, lower.tail=FALSE)
[1] 0.02275013
```

Example 2: Suppose the weight of a certain species of otters is normally distributed with a mean of μ=30 lbs and a standard deviation of σ = 5 lbs. Approximately what percentage of this species of otters weigh less than 22 lbs?

```
> #find percentage of otters that weight less than 22 lbs in a population with
> #mean = 30 and sd = 5
> pnorm(22, mean=30, sd=5)
[1] 0.05479929
```

Example 3: Suppose the height of plants in a certain region is normally distributed with a mean of μ=13 inches and a standard deviation of σ = 2 inches. Approximately what percentage of plants in this region are between 10 and 14 inches tall?

```
> #find percentage of plants that are less than 14 inches tall, then subtract the
> #percentage of plants that are less than 10 inches tall, based on a population
> #with mean = 13 and sd = 2
> pnorm(14, mean=13, sd=2) - pnorm(10, mean=13, sd=2)
[1] 0.6246553
```

qnorm

The function qnorm returns the value of the inverse cumulative density function (cdf) of the normal distribution given a certain random variable p, a population mean μ and population standard deviation σ. The syntax for using qnorm is as follows:

qnorm(p, mean, sd)

Put simply, you can use qnorm to find out what the Z-score is of the p-th quantile of the normal distribution.

```
> #find the z-score of the 99th quantile of the standard normal distribution
> qnorm(.99, mean=0, sd=1)
[1] 2.326348
```

```
> #by default, R uses mean=0 and sd=1
> qnorm(.99)
[1] 2.326348
> #find the Z-score of the 95th quantile of the standard normal distribution
> qnorm(.95)
[1] 1.644854
> #find the Z-score of the 10th quantile of the standard normal distribution
> qnorm(.10)
[1] -1.281552
```

rnorm

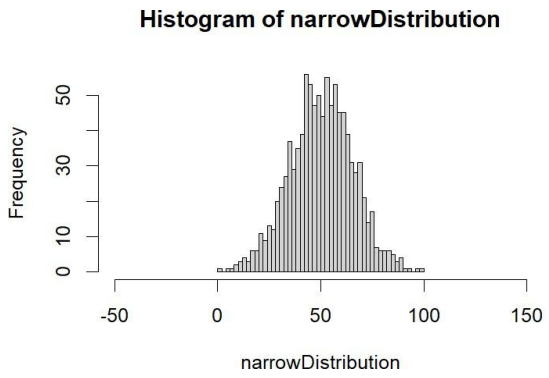
The function **rnorm** generates a vector of normally distributed random variables given a vector length *n*, a population mean μ and population standard deviation σ. The syntax for using rnorm is as follows:

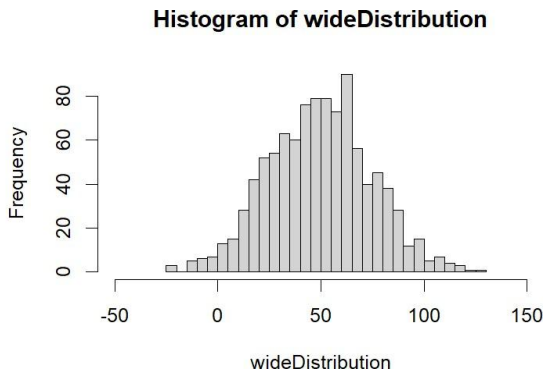
rnorm(n, mean, sd)

```
> #generate a vector of 5 normally distributed random variables with mean=10 and sd=2
> five <- rnorm(5, mean = 10, sd = 2)
> five
[1]  9.105876  6.522804 10.357730 13.794931  5.456149
```

Visualization

```
> #generate a vector of 1000 normally distributed random variables with mean=50 and sd=5
> narrowDistribution <- rnorm(1000, mean = 50, sd = 15)
> #generate a vector of 1000 normally distributed random variables with mean=50 and sd=25
> wideDistribution <- rnorm(1000, mean = 50, sd = 25)
>
> #generate two histograms to view these two distributions side by side, specify
> #50 bars in histogram and x-axis limits of -50 to 150
> par(mfrow=c(1, 2)) #one row, two columns
> hist(narrowDistribution, breaks=50, xlim=c(-50, 150))
> hist(wideDistribution, breaks=50, xlim=c(-50, 150))
```





Notice how the wide distribution is much more spread out compared to the narrow distribution. This is because we specified the standard deviation in the wide distribution to be 25 compared to just 15 in the narrow distribution.

Students have to experiment with different values of parameters - p, n, lambda, mu, sigma - to understand how the binomial, poisson and normal distributions behave.

Practise:

1) Binomial Distribution:

dbinom

```
> dbinom(x=10, size=12, prob=0.6)
[1] 0.06385228
```

```
> dbinom(x=7, size=30, prob=.5)
[1] 0.001895986
```

pbinom

```
> pbinom(4,size=10, prob=.3)
[1] 0.8497317
```

```
> pbinom(2, size=5, prob=0.5, lower.tail = FALSE)
[1] 0.5
```

qbinom

```
> qbinom(0.8497, size=10, prob=0.3)
[1] 4
```

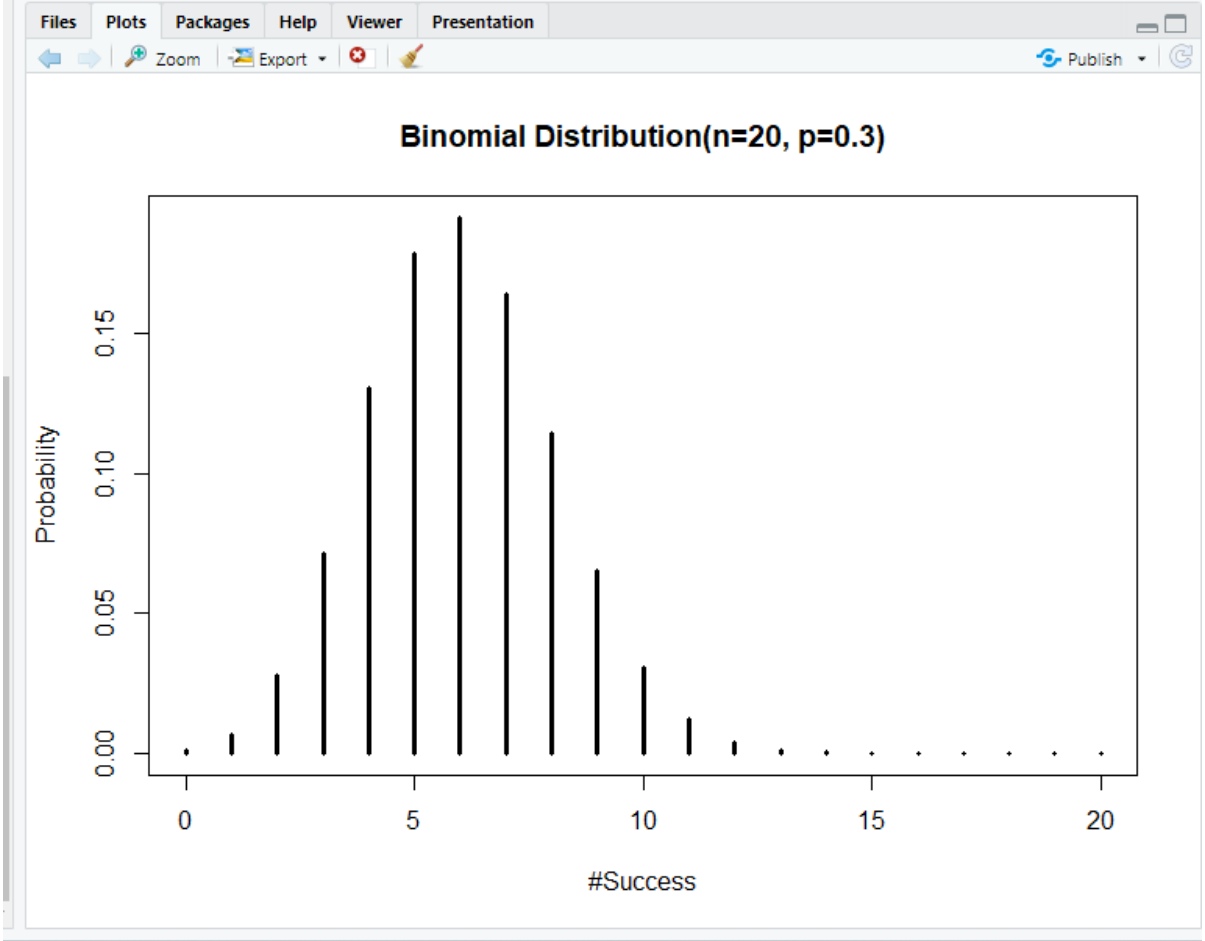
```
> qbinom(0.5, size=5, prob=0.5)
[1] 2
```

```
> qbinom(.10, size=10, prob=0.4)
[1] 2
```

```
[1] 2
> qbinom(.40, size=30, prob=.25)
[1] 7
```

rbinom

```
> rbinom(10, size=100, prob=0.3)
 [1] 34 29 34 24 25 29 30 31 29 34

> success<-0:20
> plot(success, dbinom(success, size=20, prob=0.3),
+      type='h',
+      main='Binomial Distribution(n=20, p=0.3)',
+      ylab='Probability',
+      xlab = '#Success',
+      lwd=3)
>
```



Poisson Distribution

dpois

```
> dpois(x=8, lambda=10)
[1] 0.112599
```

ppois

```
> ppois(q=8, lambda=10)
[1] 0.3328197
>
```

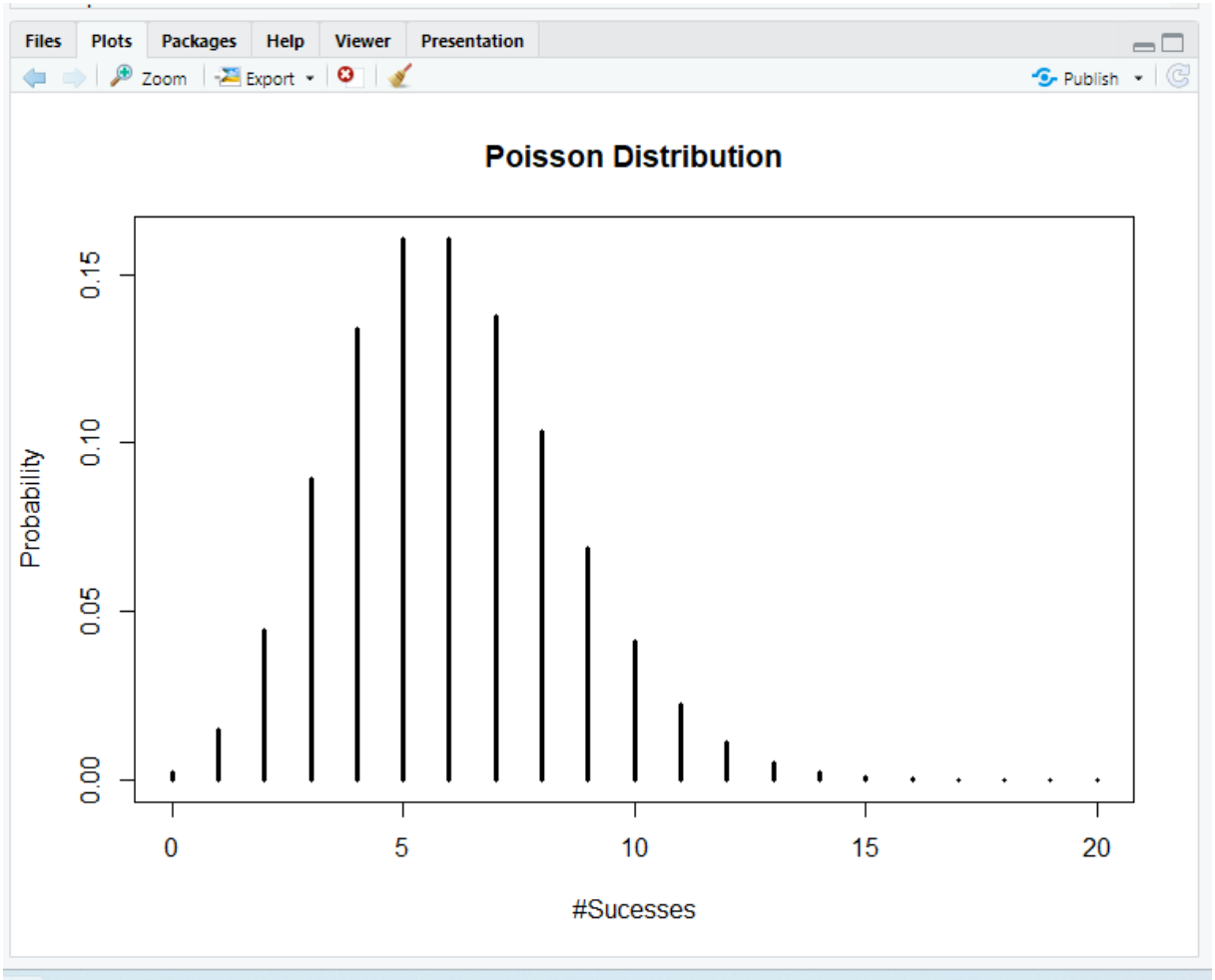
qpois

```
> qpois(p=.90,lambda=10)
[1] 14
>
```

rpois

```
> rpois(n=15, lambda=10)
 [1] 14 11  7  8 10 10 12  5  8  5 11 11  9  9 11
>
```

```
> lambda <- 6
> success <- 0:20
> plot(success, dpois(success, lambda),
+      type='h',
+      main='Poisson Distribution',
+      ylab='Probability',
+      xlab= '#Sucesses',
+      lwd=3)
>
```

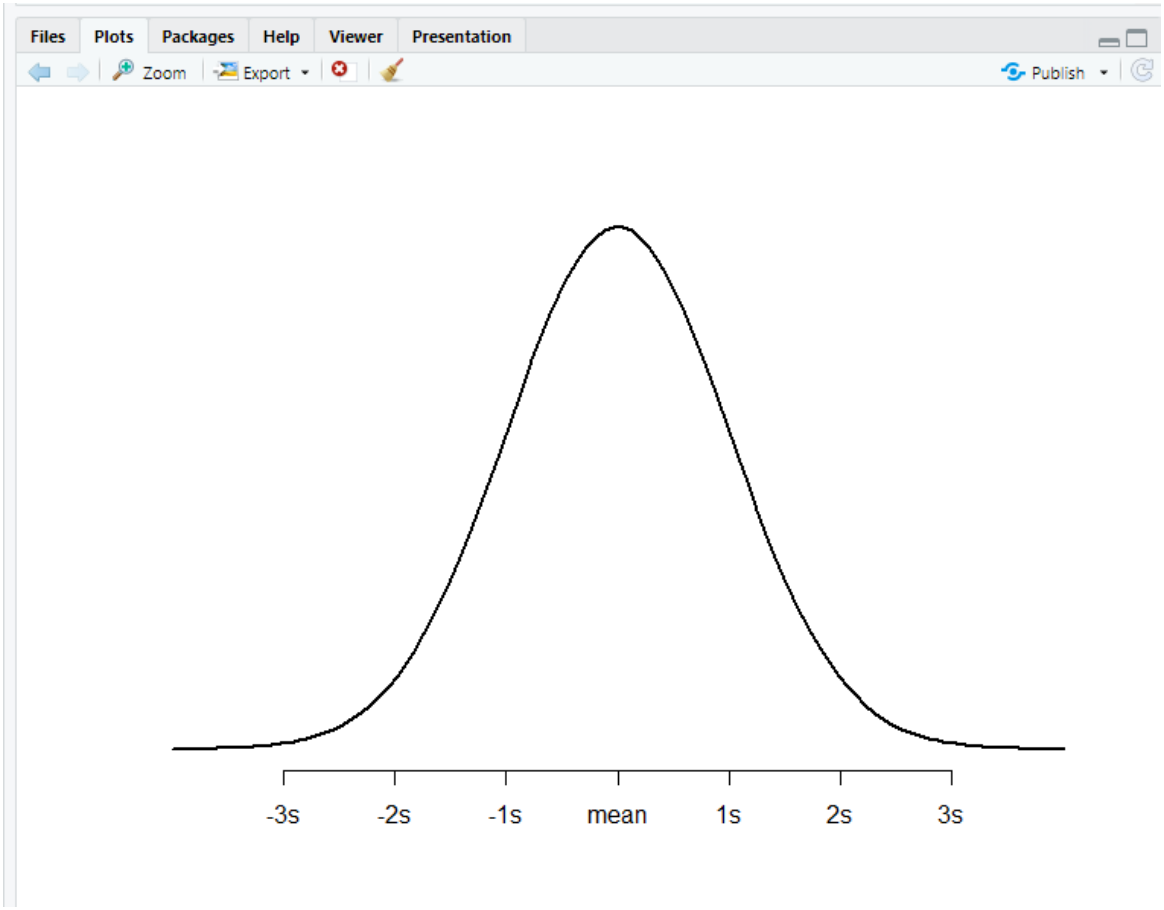


Normal distribution

dnorm

```
> dnorm(x=0, mean=0, sd=1)
[1] 0.3989423
> dnorm(x=0)
[1] 0.3989423
> dnorm(x=10, mean=20, sd=5)
[1] 0.01079819
```

```
> x<- seq(-4,4,length=100)
> y<- dnorm(x)
> plot(x,y, type="l", lwd=2, axes=FALSE, xlab="", ylab="")
> axis(1, at= -3:3, labels= c("-3s", "-2s", "-1s", "mean", "1s", "2s", "3s"))
>
```



pnorm

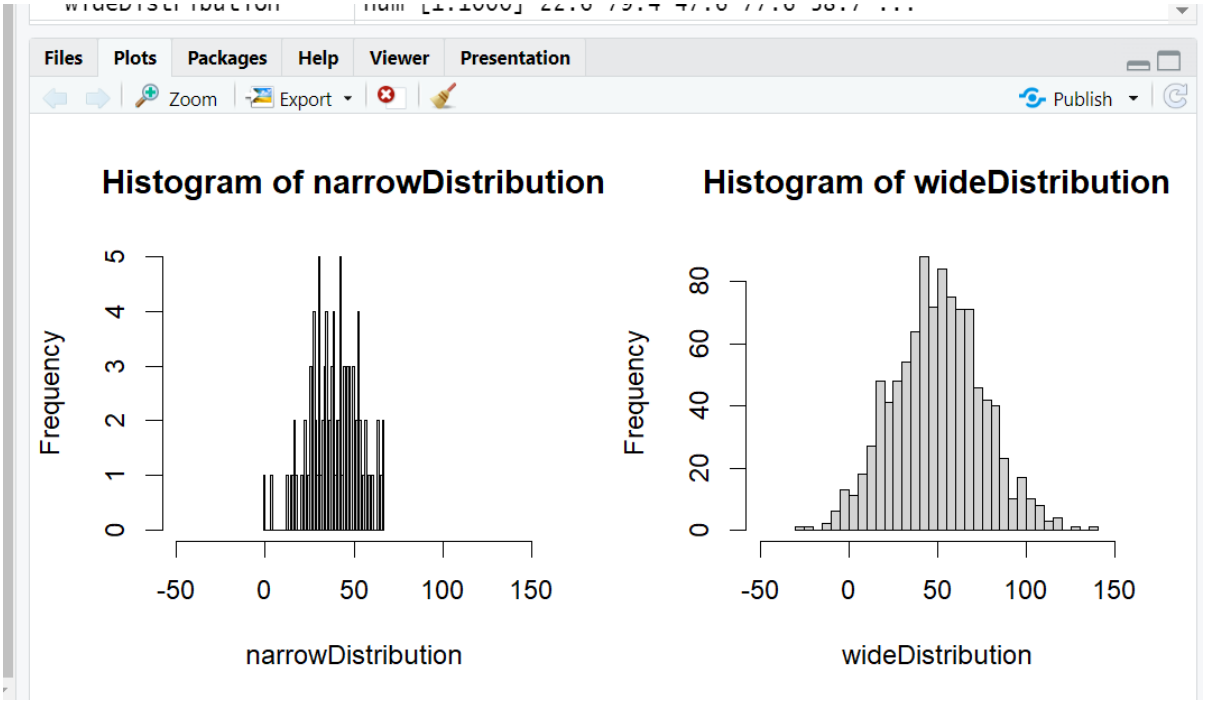
```
> pnorm(74, mean=70, sd=2, lower.tail = FALSE)
[1] 0.02275013
> pnorm(22, mean=30, sd=5)
[1] 0.05479929

> pnorm(14, mean=13, sd=2) - pnorm(10, mean=13, sd=2)
[1] 0.6246553
>
```

qnorm

```
> qnorm(.99, mean=0, sd=1)
[1] 2.326348
> qnorm(.99)
[1] 2.326348
> qnorm(.95)
[1] 1.644854
> qnorm(-.10)
[1] NaN
Warning message:
In qnorm(-0.1) : NaNs produced
> qnorm(.10)
[1] -1.281552
```

rnorm

```
> five<- rnorm(5, mean=10, sd=2)
> five
[1] 11.342442 11.049817  8.806532
[4]  9.987455  8.374685
> narrowDistribution <- rnorm(100, mean=40, sd = 15)
> wideDistribution <- rnorm(1000, mean = 50, sd = 25)
> par(mfrow=c(1,2))
> hist(narrowDistribution, breaks = 50, xlim = c(-50, 150))
> hist(wideDistribution, breaks = 50, xlim = c(-50, 150))
>
```



The wide distribution is much more spread out compared to the narrow distribution. This is because we specified the standard deviation in the wide distribution to be 25 compared to just 15 in the narrow distribution.

Conclusion: I learnt about the various functions for exploring the Binomial, Poisson, and Normal Distributions in R.

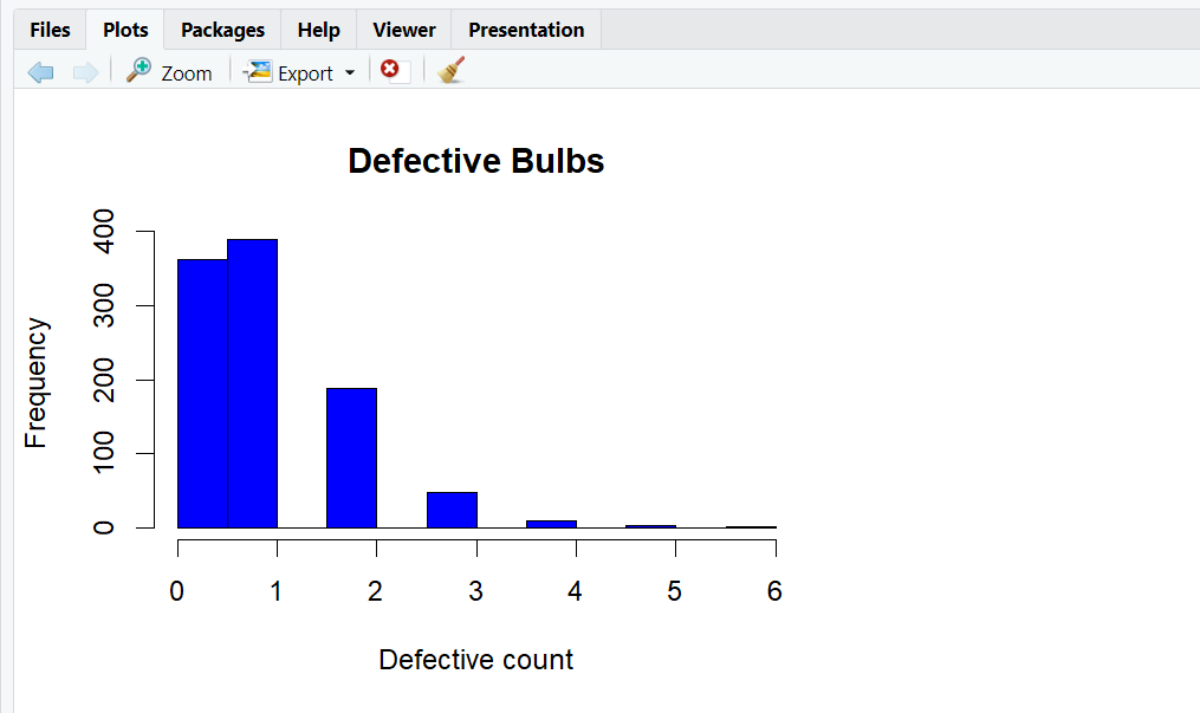
Post Lab questions

Solve the following using R:

1A. Imagine a factory producing light bulbs, where each bulb has a 90% probability of being non-defective. If 10 bulbs are selected at random, use the binomial distribution to calculate the probability of finding exactly 8 non-defective bulbs. Visualize the distribution of defective bulbs in repeated samples

Ans. Binomial Distribution (Finding Probability)

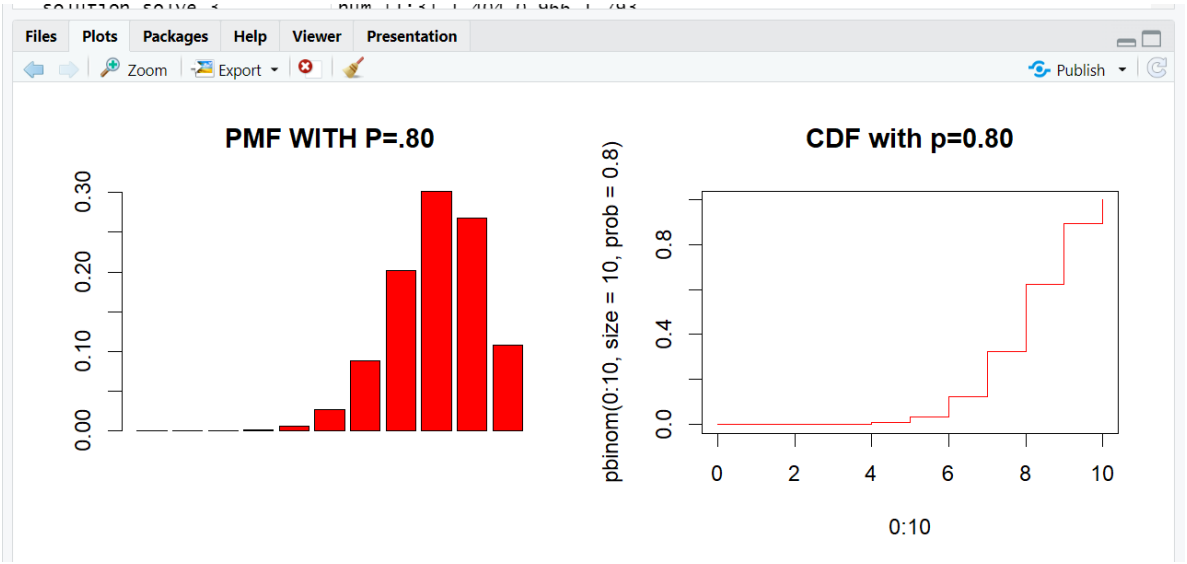
```
Console   Terminal   Background Jobs
R 4.4.2 · ~/
> dbinom(8,size=10, prob=0.90)
[1] 0.1937102
> hist(rbinom(1000, size=10, prob=0.10), col="blue", main="Defective Bulbs", xlab="Defective count")
>
```



1B. If the probability of a bulb being non-defective decreases to 80%, how does this affect the shape of the PMF and CDF? Visualize it in R and interpret the changes.

Ans. Changing Probability to 80% (Impact on PMF and CDF)

```
Console  Terminal  Background Jobs
R 4.4.2 · ~/
> dbinom(8, size=10, prob = 0.80)
[1] 0.3019899
> barplot(dbinom(0:10, size=10, prob=0.80), col="red", main="PMF WITH P=.80")
> plot(0:10, pbinom(0:10, size = 10, prob = 0.80), type="s", col="red", main="CDF with p=0.80")
>
```

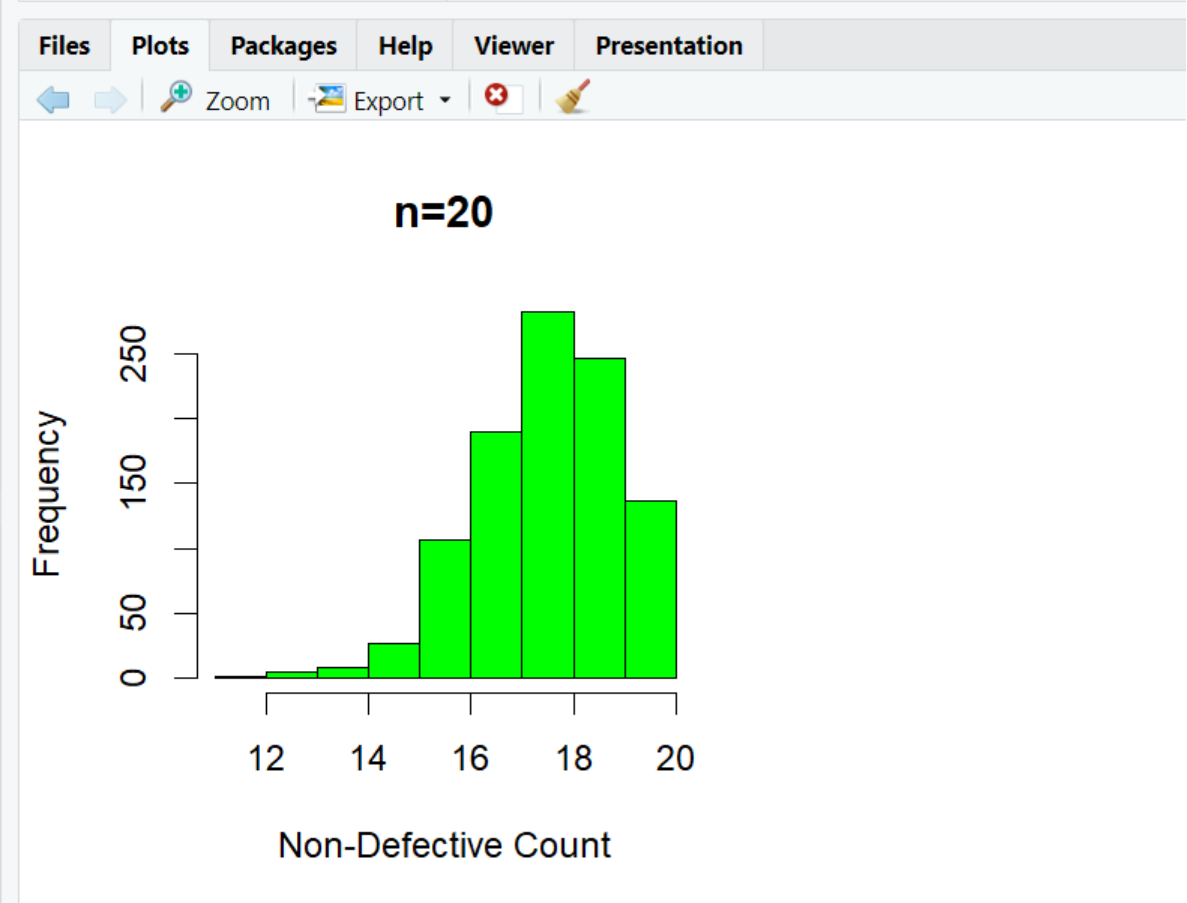


1C. In the context of quality control, how would increasing the sample size n to 20 impact the expected value and the spread of the distribution?

Ans. Increasing Sample Size to 20 (Impact on Expected Value and Spread)

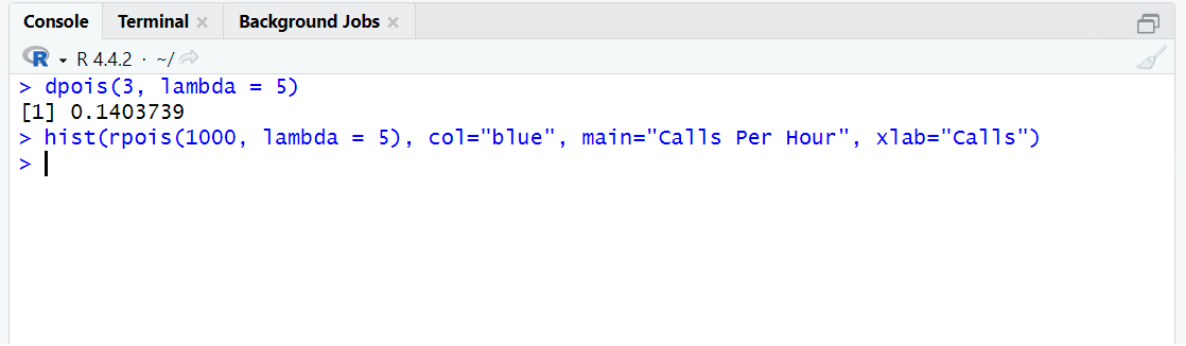
```
Console  Terminal  Background Jobs
R 4.4.2 · ~/
> expected_value <- 20*0.90
> std_dev <- sqrt(20 * 0.90 * 0.10)
> print(expected_value)
[1] 18
> print(std_dev)
[1] 1.341641
```

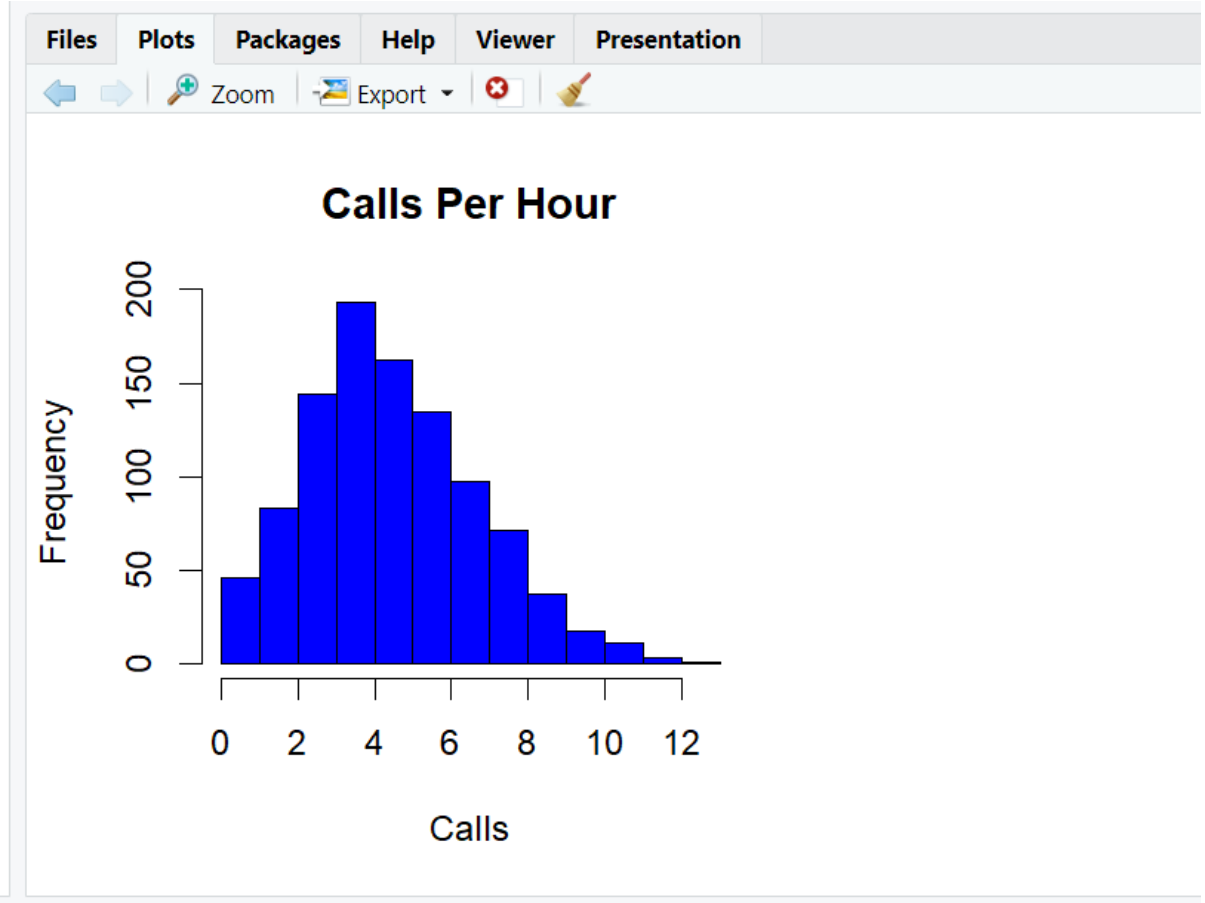
```
>
> hist(rbinom(1000, size = 20, prob = 0.90), col="green", main="n=20", xlab="Non-Defective Count")
```



2A. Consider a call center that receives an average of 5 calls per hour. Use the Poisson distribution to determine the probability of receiving exactly 3 calls in an hour. Visualize how the number of calls varies over multiple hours.

Ans. Poisson Distribution:

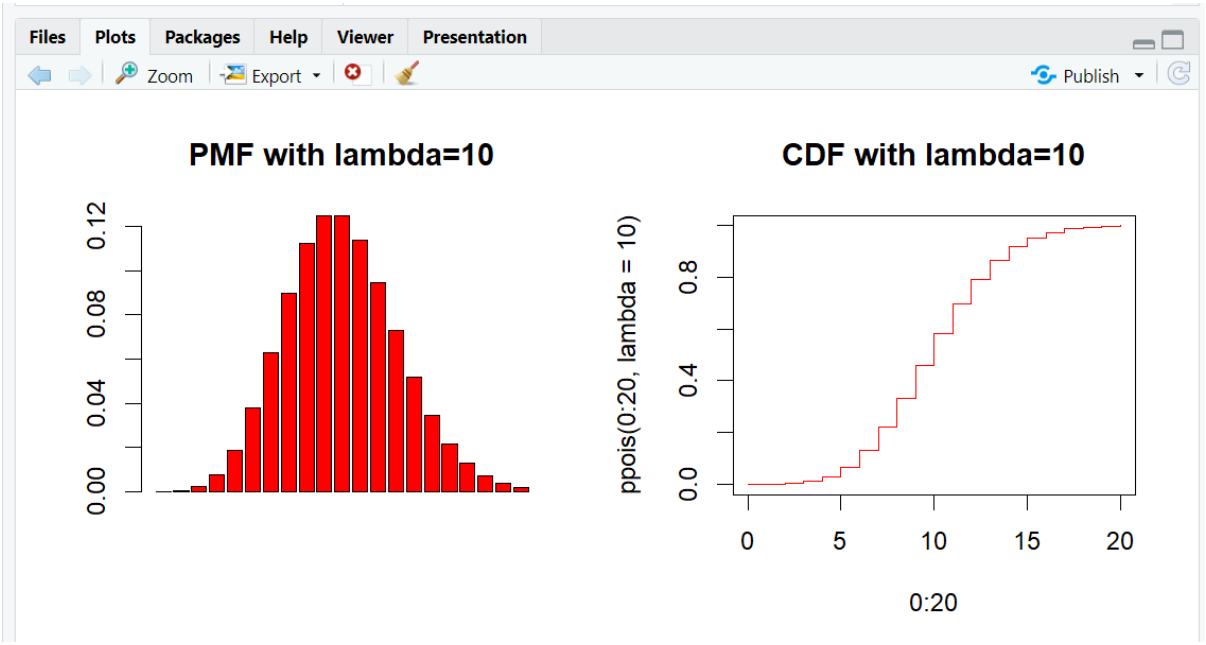




2B If the call rate increases to 10 calls per hour, how does the PMF and CDF change? Visualize it in R and explain.

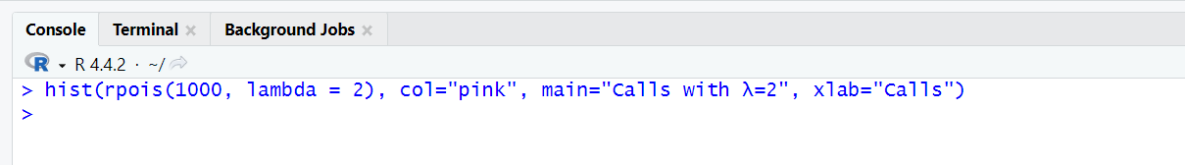
Ans. Increased Call Rate to 10 (PMF and CDF Change)

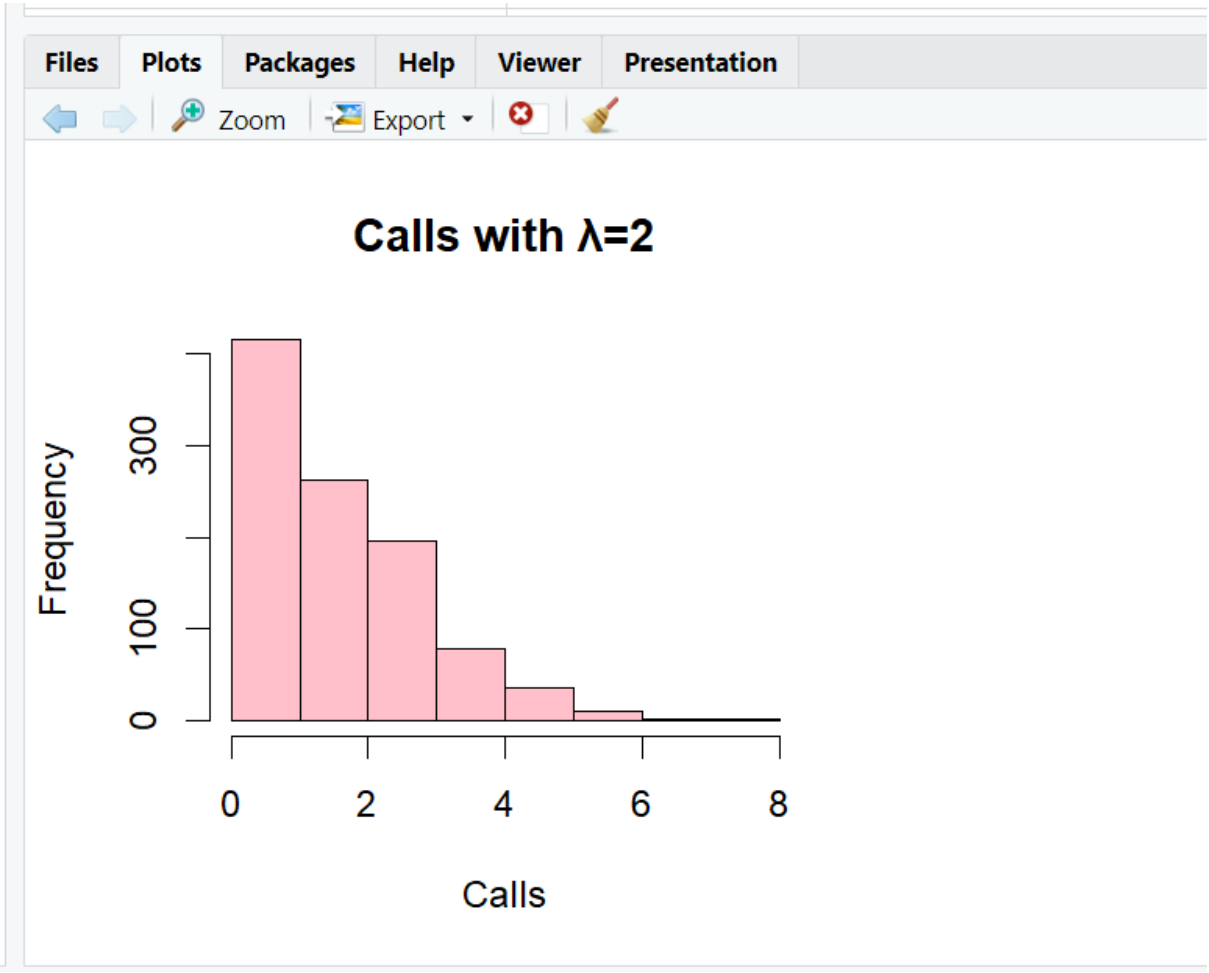
```
Console   Terminal   Background Jobs
R 4.4.2 · ~/
> dpois(3, lambda = 10)
[1] 0.007566655
> barplot(dpois(0:20, lambda=10), col="red", main="PMF with lambda=10")
> plot(0:20, ppois(0:20, lambda=10), type="s", col="red", main="CDF with lambda=10")
>
```



2C: What happens to the spread of the distribution as λ becomes smaller (e.g. λ=2)? Interpret this in the context of low call volumes.

Ans. Impact of Lower Call Rate (λ=2)

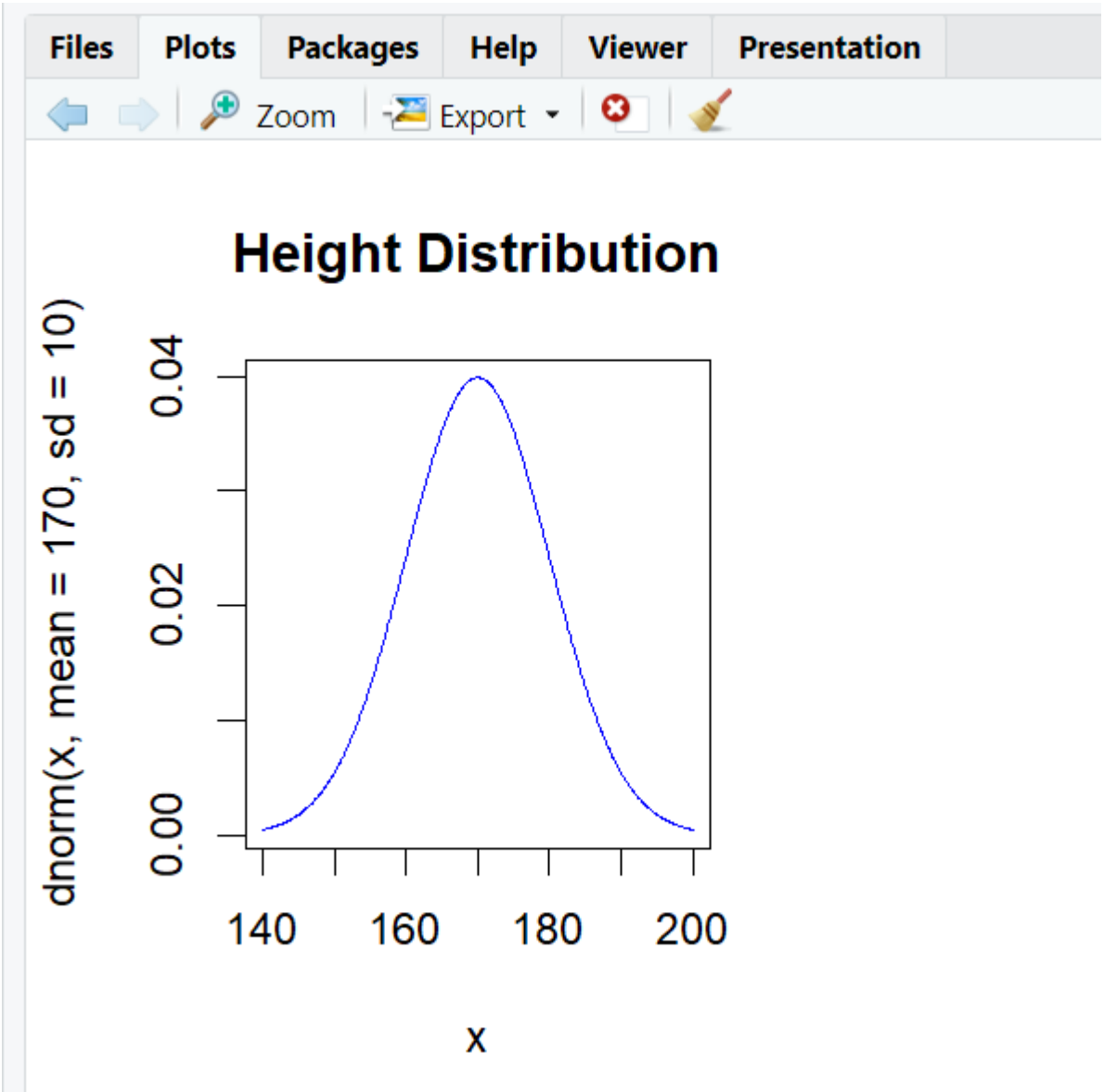




3A: The heights of individuals often follow a normal distribution. If the average height of adults in a region is 170 cm with a standard deviation of 10 cm, use the normal distribution to calculate the probability of height ranging between 160 to 170 cm. Visualize the distribution.

Ans. Normal Distribution (Height Range Calculation)

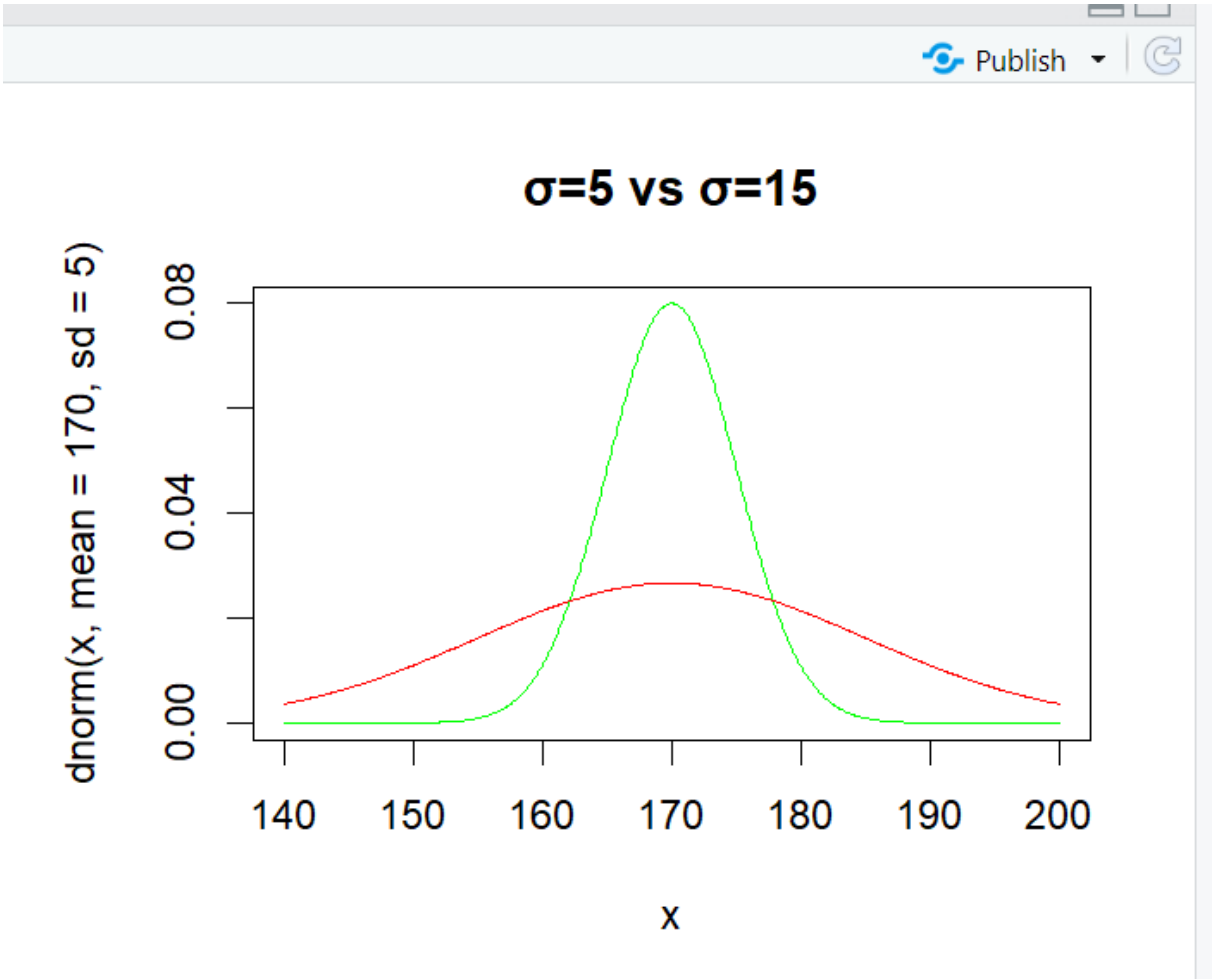
```
Console   Terminal   Background Jobs
R 4.4.2 · ~/
> pnorm(170, mean = 170, sd = 10) - pnorm(160, mean = 170, sd = 10)
[1] 0.3413447
> x <- seq(140, 200, by=0.1)
> plot(x, dnorm(x, mean=170, sd=10), type="l", col="blue", main="Height Distributio
n")
>
```



3B: How does changing the standard deviation affect the shape of the bell curve? Visualize this in R for σ=5 and σ=15.

Ans. Effect of Changing Standard Deviation (σ=5, σ=15)

```
Console   Terminal ×   Background Jobs ×
R 4.4.2 · ~/
> x <- seq(140, 200, by=0.1)
> plot(x, dnorm(x, mean=170, sd=5), type="l", col="green", main="σ=5 vs σ=15")
> lines(x, dnorm(x, mean=170, sd=15), col="red")
>
```



3C: What happens to the PDF if the mean shifts to 180 cm? How does this relate to real-world population differences?

Ans. Effect of Changing Mean to 180 cm

```
Console   Terminal   Background Jobs
R 4.4.2 · ~/
> x <- seq(140, 200, by=0.1)
> plot(x, dnorm(x, mean=170, sd=10), type="l", col="blue", main="Mean Shift")
> lines(x, dnorm(x, mean=180, sd=10), col="green")
>
```