Homework 4 : The normal distribution

Ramnivas Singh

02/28/2021

```
## Loading required package: shiny
```

```
## Loading required package: openintro
```

```
## Loading required package: airports
```

```
## Loading required package: cherryblossom
```

```
## Loading required package: usdata
```

```
## Loading required package: OIdata
```

```
## Loading required package: RCurl
```

```
## Loading required package: maps
```

```
## Loading required package: ggplot2
```

```
## Loading required package: markdown
```

```
##
## Welcome to CUNY DATA606 Statistics and Probability for Data Analytics
## This package is designed to support this course. The text book used
## is OpenIntro Statistics, 4th Edition. You can read this by typing
## vignette('os4') or visit www.OpenIntro.org.
##
## The getLabs() function will return a list of the labs available.
##
## The demo(package='DATA606') will list the demos that are available.
```

```
##
## Attaching package: 'DATA606'
```

```
## The following objects are masked from 'package:openintro':
##
##     calc_streak, present, qqnormsim
## The following object is masked from 'package:utils':
##
##     demo
```

```
##
## Attaching package: 'DT'
```

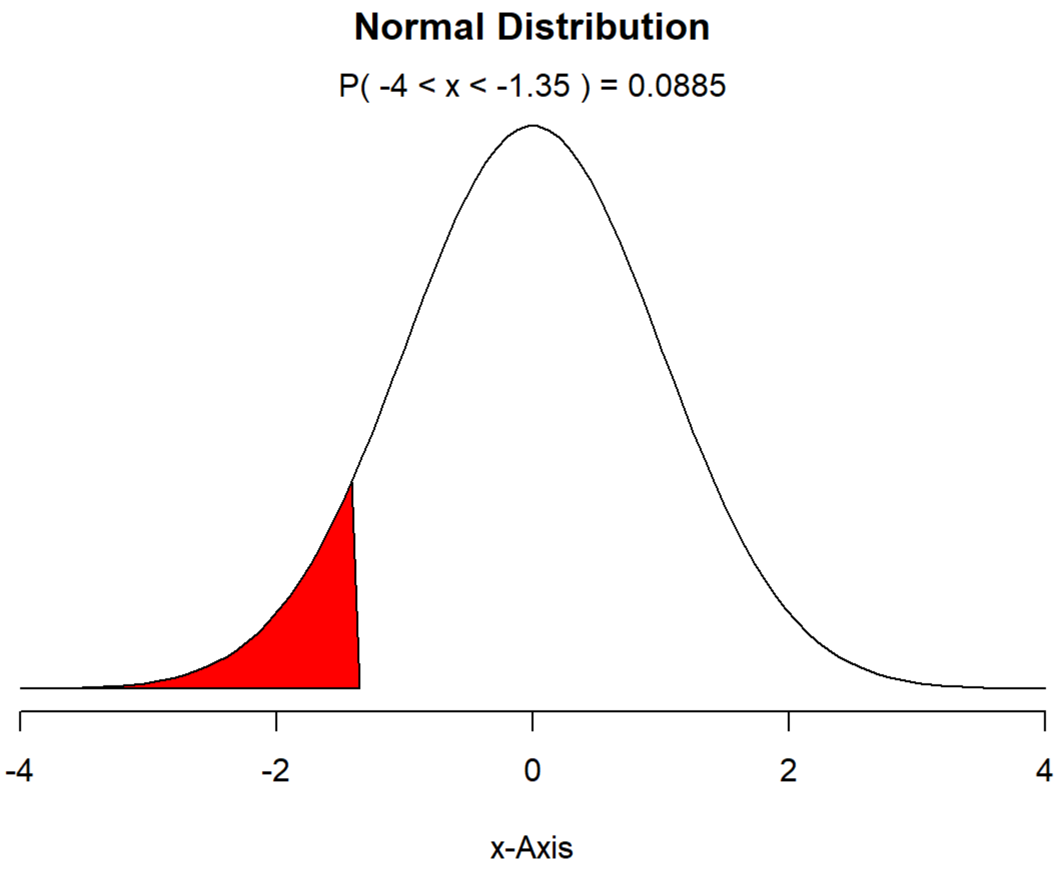
```
## The following objects are masked from 'package:shiny':
##
##     dataTableOutput, renderDataTable
```

Area under the curve, Part I. (4.1, p. 142) What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph.

a. $Z < -1.35$
b. $Z > 1.48$
c. $-0.4 < Z < 1.5$
d. $|Z| > 2$

Answers: (a) $Z < -1.35$

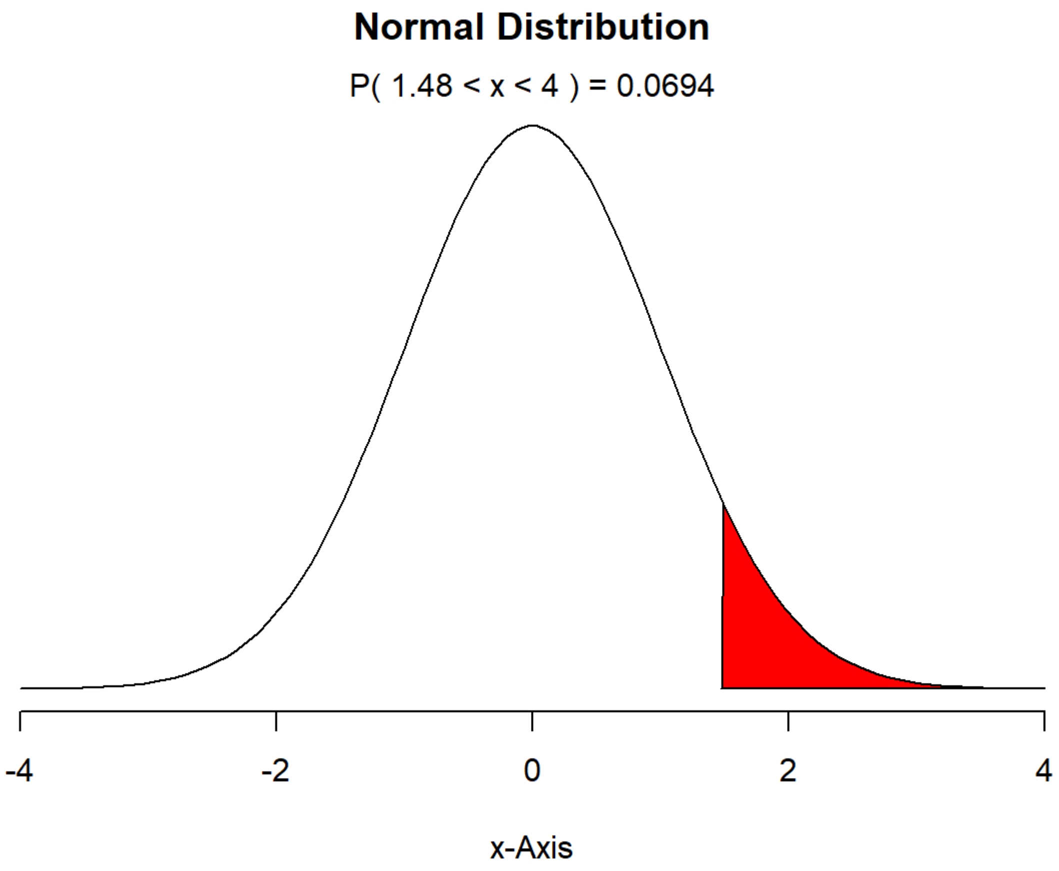
```
# use the DATA606::normalPlot function to create a plot or the pnorm function for the actual val
ues.
# ?normalPlot
normalPlot(mean = 0, bounds = c(-4, -1.35), sd = 1, tails = FALSE)
```



Percent of a standard normal distribution =8.85%

b. $Z > 1.48$

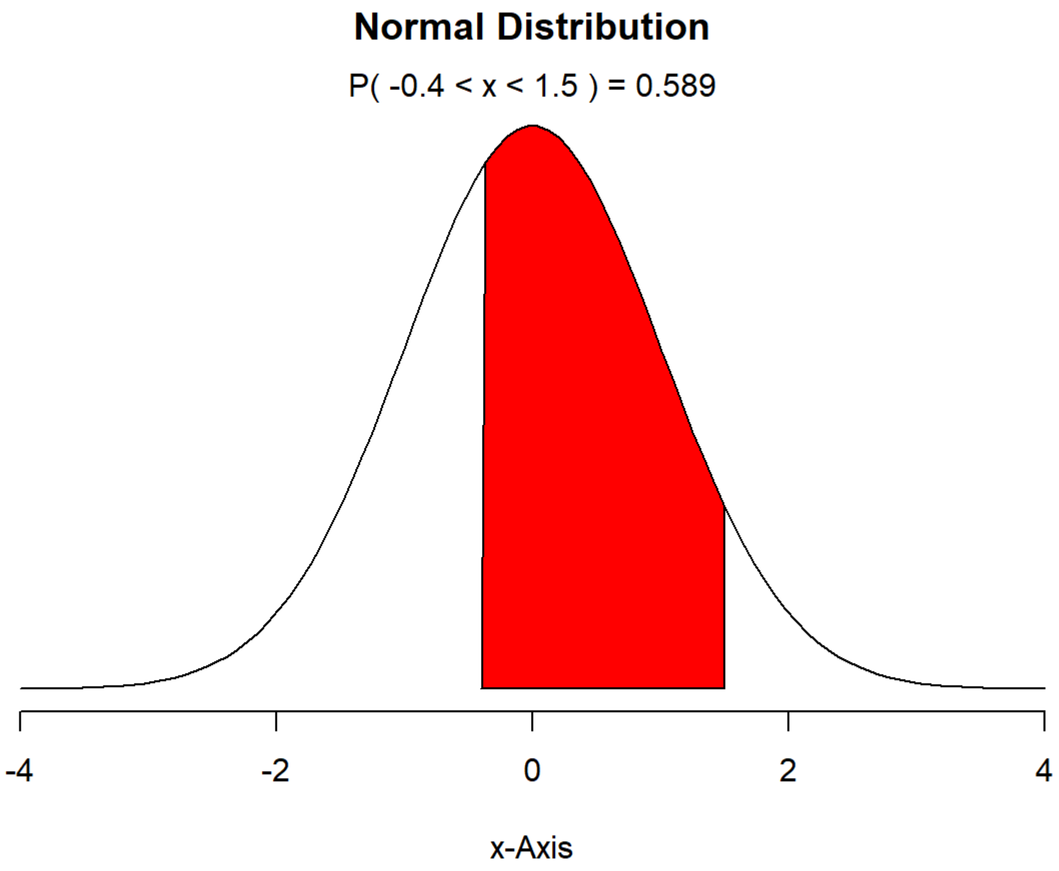
```
normalPlot(mean = 0, bounds = c(1.48, 4), sd = 1)
```



Percent of a standard normal distribution =6.94%

c. $-0.4 < Z < 1.5$

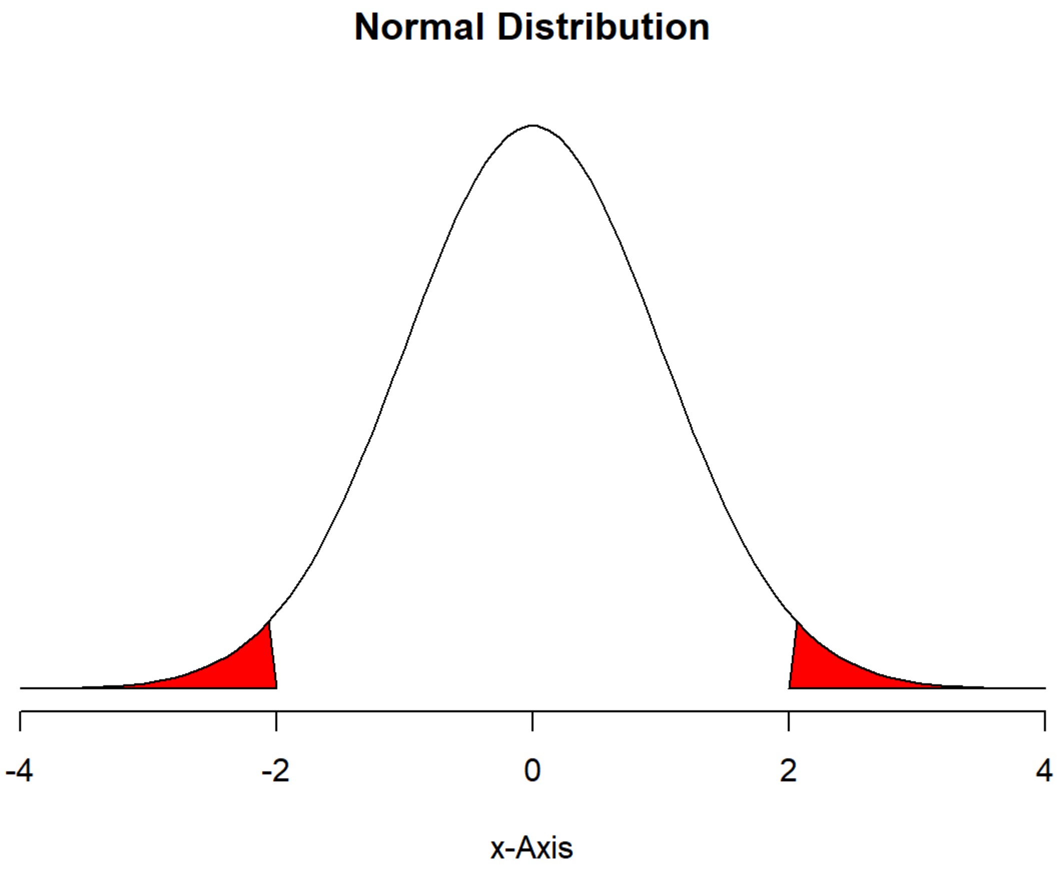
```
normalPlot(mean = 0, bounds = c(-0.4,1.5), sd = 1)
```



Percent of a standard normal distribution =58.9%

d. $|Z| > 2$

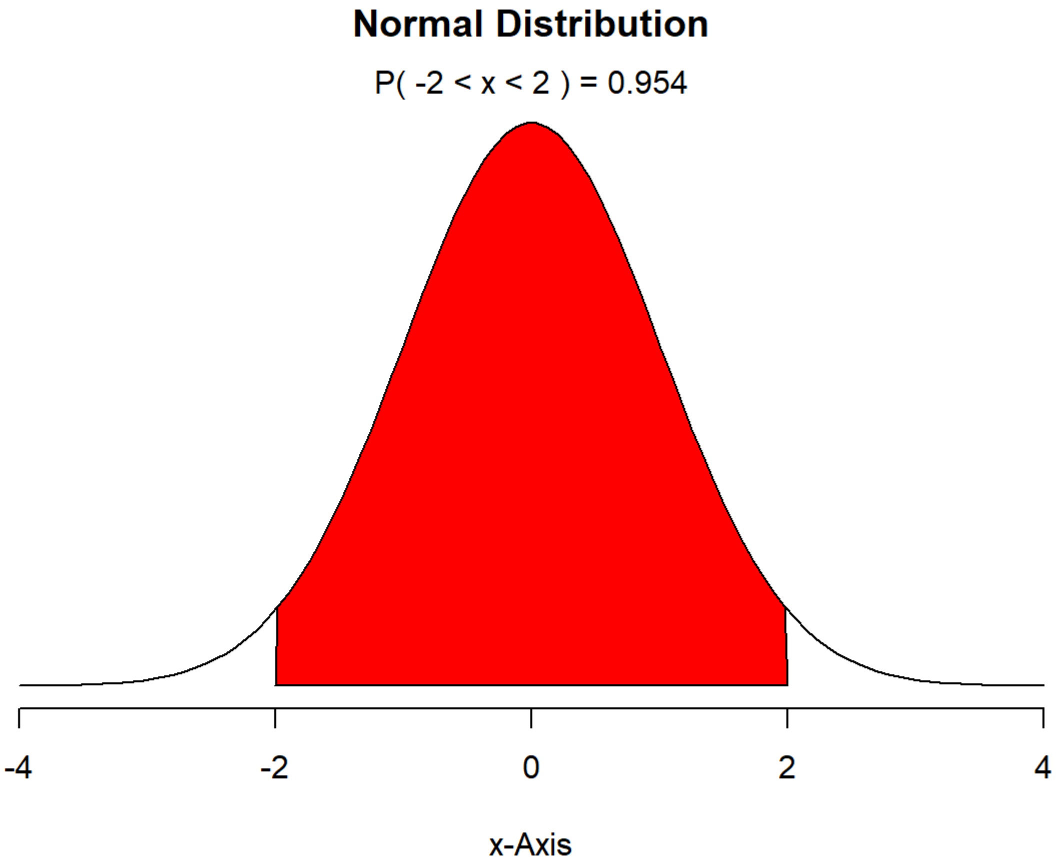
```
normalPlot(mean = 0, bounds = c(-2,2), sd = 1,tails = TRUE)
```



Lets

plot another area with between -2,2

```
normalPlot(mean = 0, bounds = c(-2,2), sd = 1,tails = FALSE)
```



```
1-.954
```

```
## [1] 0.046
```

Percent of a standard normal distribution = 4.6

Triathlon times, Part I (4.4, p. 142) In triathlons, it is common for racers to be placed into age and gender groups. Friends Leo and Mary both completed the Hermosa Beach Triathlon, where Leo competed in the *Men, Ages 30 - 34* group while Mary competed in the *Women, Ages 25 - 29* group. Leo completed the race in 1:22:28 (4948 seconds), while Mary completed the race in 1:31:53 (5513 seconds). Obviously Leo finished faster, but they are curious about how they did within their respective groups. Can you help them? Here is some information on the performance of their groups:

- The finishing times of the group has a mean of 4313 seconds with a standard deviation of 583 seconds.
- The finishing times of the group has a mean of 5261 seconds with a standard deviation of 807 seconds.
- The distributions of finishing times for both groups are approximately Normal.

Remember: a better performance corresponds to a faster finish.

a. Write down the short-hand for these two normal distributions.
b. What are the Z-scores for Leo's and Mary's finishing times? What do these Z-scores tell you?
c. Did Leo or Mary rank better in their respective groups? Explain your reasoning.
d. What percent of the triathletes did Leo finish faster than in his group?
e. What percent of the triathletes did Mary finish faster than in her group?

f. If the distributions of finishing times are not nearly normal, would your answers to parts (b) - (e) change? Explain your reasoning.

Answers:
(a)

```
#Men Ages 30-40:
men_mean=4313
men_sd=583

#Women Ages 30-34:
women_mean=5261
women_sd=807
```

b.

```
# Z_score = (x -μ)/sd
# Z_score for Leo
Z_score_Leo<-(4948-men_mean)/men_sd
print(Z_score_Leo)
```

```
## [1] 1.089194
```

```
# Z_score for Mary
Z_score_Mary<-(5513-women_mean)/women_sd
print(Z_score_Mary)
```

```
## [1] 0.3122677
```

What do these Z-scores tell you? (c)

```
With lower Z-Score, Mary did better than Leo.
```

(d)

```
pnorm(1.09,lower.tail=FALSE)
```

```
## [1] 0.1378566
```

13 %

(e)

```
pnorm(0.31,lower.tail=FALSE)
```

```
## [1] 0.3782805
```

37 %

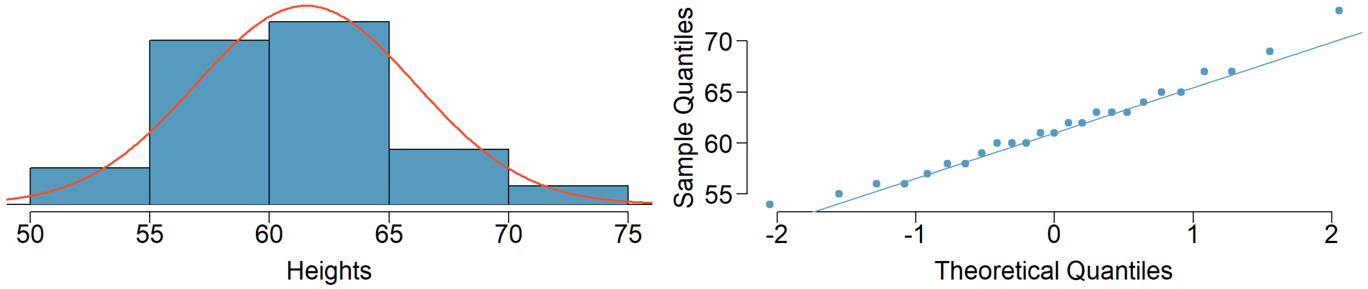
(f)
Yes, it would change. Any change distribution will have a change in mean and standard deviation, z score and other percentiles above will change as these parameters changes.

Heights of female college students Below are heights of 25 female college students.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
54,	55,	56,	56,	57,	58,	58,	59,	60,	60,	60,	61,	61,	62,	62,	63,	63,	63,	64,	65,	65,	67,	67,	69,	73

a. The mean height is 61.52 inches with a standard deviation of 4.58 inches. Use this information to determine if the heights approximately follow the 68-95-99.7% Rule.
b. Do these data appear to follow a normal distribution? Explain your reasoning using the graphs provided below.

#This finds the proportion of heights within 1, 2, and 3 stdevs, respectively



Answers:
(a)

```
sd<-sd(heights)
summary(heights)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   54.00   58.00   61.00   61.52   64.00   73.00
```

```
mean <- 61.52
sd <- 4.58
sds <- sd * c(1:3)
pnorm(mean + sds, mean = mean, sd = sd) - pnorm(mean - sds, mean = mean, sd = sd)
```

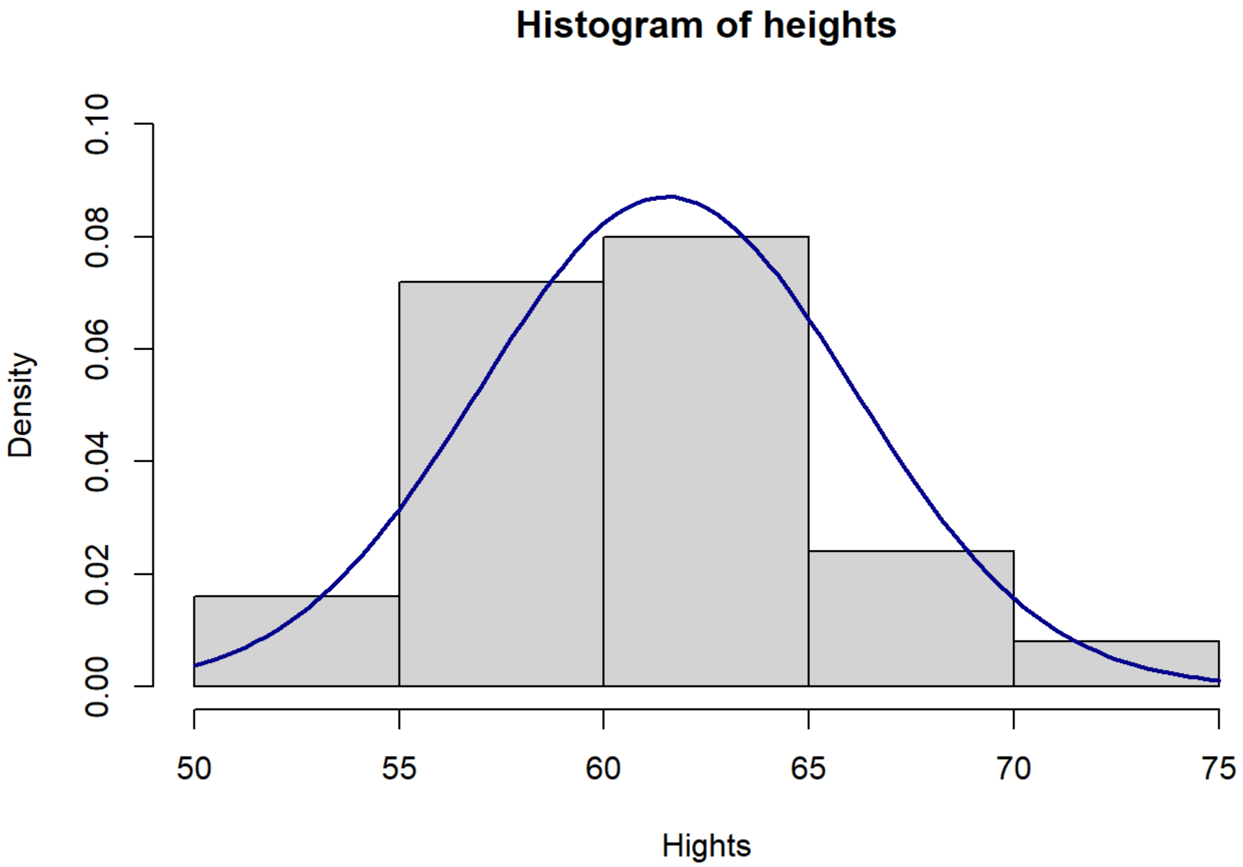
```
## [1] 0.6826895 0.9544997 0.9973002
```

Above values appears to follow 68-95-99.7% rule

```
# Use the DATA606::qqnormsim function
```

b.

```
hist(heights, prob=TRUE,
     xlab="Hights", ylim=c(0, .1))
curve(dnorm(x, mean=mean, sd=sd),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")
```



- The data does approximately follow a normal distribution.
- The histogram appears to have an approximate bell curve, and the dots mostly fall on the line in the qq-plot

Defective rate. (4.14, p. 148) A machine that produces a special type of transistor (a component of computers) has a 2% defective rate. The production is considered a random process where each transistor is independent of the others.

a. What is the probability that the 10th transistor produced is the first with a defect?
b. What is the probability that the machine produces no defective transistors in a batch of 100?
c. On average, how many transistors would you expect to be produced before the first with a defect? What is the standard deviation?
d. Another machine that also produces transistors has a 5% defective rate where each transistor is produced independent of the others. On average how many transistors would you expect to be produced with this machine before the first with a defect? What is the standard deviation?
e. Based on your answers to parts (c) and (d), how does increasing the probability of an event affect the mean and standard deviation of the wait time until success?

Answers:

a.

```
# p(success) as the probability, transistor has a 2% defective rate
p = 0.02
(1-p)^9 * p
```

```
## [1] 0.01667496
```

b.

```
p = 0.02
(1-p)^100
```

```
## [1] 0.1326196
```

c.

```
mean<-1/p
mean
```

```
## [1] 50
```

```
sd<-((1-p)/p^2)^0.5
sd
```

```
## [1] 49.49747
```

d.

```
p2 <- 0.05
mean<-1/p2
mean
```

```
## [1] 20
```

```
sd<-((1-p2)/p2^2)^0.5
sd
```

```
## [1] 19.49359
```

e.

Probability of success increases, wait time for success and the spread in the distribution the decreases

Male children. While it is often assumed that the probabilities of having a boy or a girl are the same, the actual probability of having a boy is slightly higher at 0.51. Suppose a couple plans to have 3 kids.

a. Use the binomial model to calculate the probability that two of them will be boys.
b. Write out all possible orderings of 3 children, 2 of whom are boys. Use these scenarios to calculate the same probability from part (a) but using the addition rule for disjoint outcomes. Confirm that your answers from parts (a) and (b) match.
c. If we wanted to calculate the probability that a couple who plans to have 8 kids will have 3 boys, briefly describe why the approach from part (b) would be more tedious than the approach from part (a).

Answers:
(a)

```
dbinom(2,3,0.51)
```

```
## [1] 0.382347
```

b.

```
# boyp * boyp * girlp +
# boyp * girlp * boyp +
# girlp * boyp * boyp
(0.51^2)*0.49*3
```

```
## [1] 0.382347
```

c. The approach in part b would be more tedious because it would involve writing out every combination of 3 boys among 8 kids

---

Serving in volleyball. (4.30, p. 162) A not-so-skilled volleyball player has a 15% chance of making the serve, which involves hitting the ball so it passes over the net on a trajectory such that it will land in the opposing team's court. Suppose that her serves are independent of each other.

a. What is the probability that on the 10th try she will make her 3rd successful serve?
b. Suppose she has made two successful serves in nine attempts. What is the probability that her 10th serve will be successful?
c. Even though parts (a) and (b) discuss the same scenario, the probabilities you calculated should be different. Can you explain the reason for this discrepancy?

Answers:
(a) 3rd successful serve on the 10th try matches with the negative binomial distribution:

```
dnbinom(7, 3, 0.15)
```

```
## [1] 0.03895012
```

(b) The probability that her 10th serve will be successful is 15%.
(c) Probability of a single event by itself would be different probability of the same event in a combination of

observation.