Chapter 4 - Distributions of Random Variables

Leticia Salazar

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) $Z < -1.35$

```
#$Z < -1.35$ = 0.089
pnorm(-1.35)
```

```
## [1] 0.08850799
```

```
#DATA606::normalPlot(mean = 0, sd = 1, bounds = c(-5, -1.35))
```

(b) $Z > 1.48$

```
#$Z > 1.48$ = 0.069
1-pnorm(1.48)
```

```
## [1] 0.06943662
```

```
#DATA606::normalPlot(mean = 0, sd = 1, bounds = c(1.48, 5))
```

(c) $-0.4 < Z < 1.5$

```
#$-0.4 < Z < 1.5$ = 0.589
pnorm(1.5)- pnorm(-0.4)
```

```
## [1] 0.5886145
```

```
#DATA606::normalPlot(mean = 0, sd = 1, bounds = c(-0.4, 1.5))
```

(d) $|Z| > 2$

```
#$|Z| > 2$ = 0.046
pnorm(-2) + (1-pnorm(2))
```

```
## [1] 0.04550026
```

```r
#DATA606::normalPlot(mean = 0, sd = 1, bounds = c(-5, 2))
```

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 *Men, Ages 30 - 34* group has a mean of 4313 seconds with a standard deviation of 583 seconds.
- The finishing times of the *Women, Ages 25 - 29* 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.

```r
leo_time <- 4948
mary_time <- 5513

#Men's Mean and SD
m_mean <- 4313
m_sd <- 583

#Women's Mean and SD
w_mean <- 5261
w_sd <- 807
```

(b) What are the Z-scores for Leo's and Mary's finishing times? What do these Z-scores tell you?

- *Z-score for Leo is 1.089 and for Mary is 0.312*
- *Both z-scores tell you distance in standard deviations of both participants in their respective groups.*

```r
#z-score = (value - mean) / sd

#Z-Score for Leo
z_score_Leo <- (leo_time - m_mean) / m_sd

#Z-Score for Mary
z_score_Mary <- (mary_time - w_mean) / w_sd

print(c(z_score_Leo, z_score_Mary))
```

```
## [1] 1.0891938 0.3122677
```

(c) Did Leo or Mary rank better in their respective groups? Explain your reasoning.

- _ For the men's Leo is 1.089 standard deviations above the mean; he was slower than average in his group. Meanwhile, Mary's z-score is 0.312 standard deviations below the mean; she is ranked faster than average in her group. Therefore, based on these details, Mary ranked better than Leo in her group._

(d) What percent of the triathletes did Leo finish faster than in his group?

- *Leo finished 13.8% faster than his group.*

```
1-pnorm(z_score_Leo)
```

```
## [1] 0.1380342
```

(e) What percent of the triathletes did Mary finish faster than in her group?

- *Mary finished 37.7% faster than his group.*

```
1-pnorm(z_score_Mary)
```

```
## [1] 0.3774186
```

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

- *If the distributions of finishing times were not nearly normal then there would be a change in the skeweness of the distribution and affect the z-scores. With z-scores changing, so will the percentage and ranking for both Leo and Mary within their groups. I will be assuming their ranking would be further from the mean and standard deviations from average and there will be outliers more visible in those groups as well.*

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

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

(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.

- *The 68-95-99.7% rules does apply with this data set since the percentages below very similar.*

```
heights <- c(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)

#Mean of Heights
H_mean <- mean(heights)
H_sd <- sd(heights)

#Z-Score Heights
z_score_H <- (heights - H_mean) / H_sd
z_score_H
```

```
##  [1] -1.6406080 -1.4224420 -1.2042761 -1.2042761 -0.9861101 -0.7679442
##  [7] -0.7679442 -0.5497782 -0.3316122 -0.3316122 -0.3316122 -0.1134463
## [13] -0.1134463  0.1047197  0.1047197  0.3228856  0.3228856  0.3228856
## [19]  0.5410516  0.7592175  0.7592175  1.1955494  1.1955494  1.6318813
## [25]  2.5045451
```

```
#Percent to compare with 68% (±1 SD)
pnorm(H_mean + H_sd, mean = H_mean, sd = H_sd) - pnorm(H_mean - H_sd, mean = H_mean, sd = H_sd)
```

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

```
#Percent to compare with 95% (±2 SD)
pnorm(H_mean + (2 * H_sd), mean = H_mean, sd = H_sd) - pnorm(H_mean -  (2 * H_sd), mean = H_mean, sd = 
```

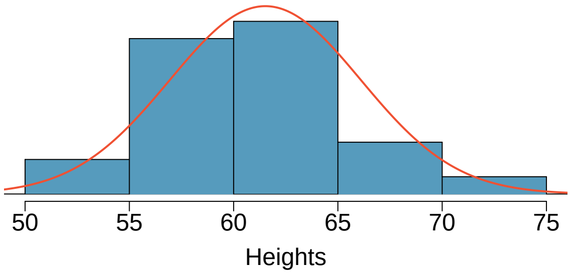
```
## [1] 0.9544997
```

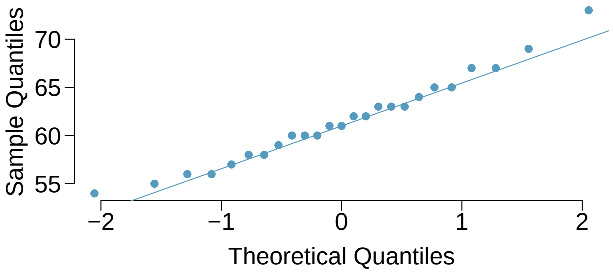
```
#Percent to compare with 99.7% (±3 SD)
pnorm(H_mean +  (3 * H_sd), mean = H_mean, sd = H_sd) - pnorm(H_mean -  (3 * H_sd), mean = H_mean, sd =
```

```
## [1] 0.9973002
```

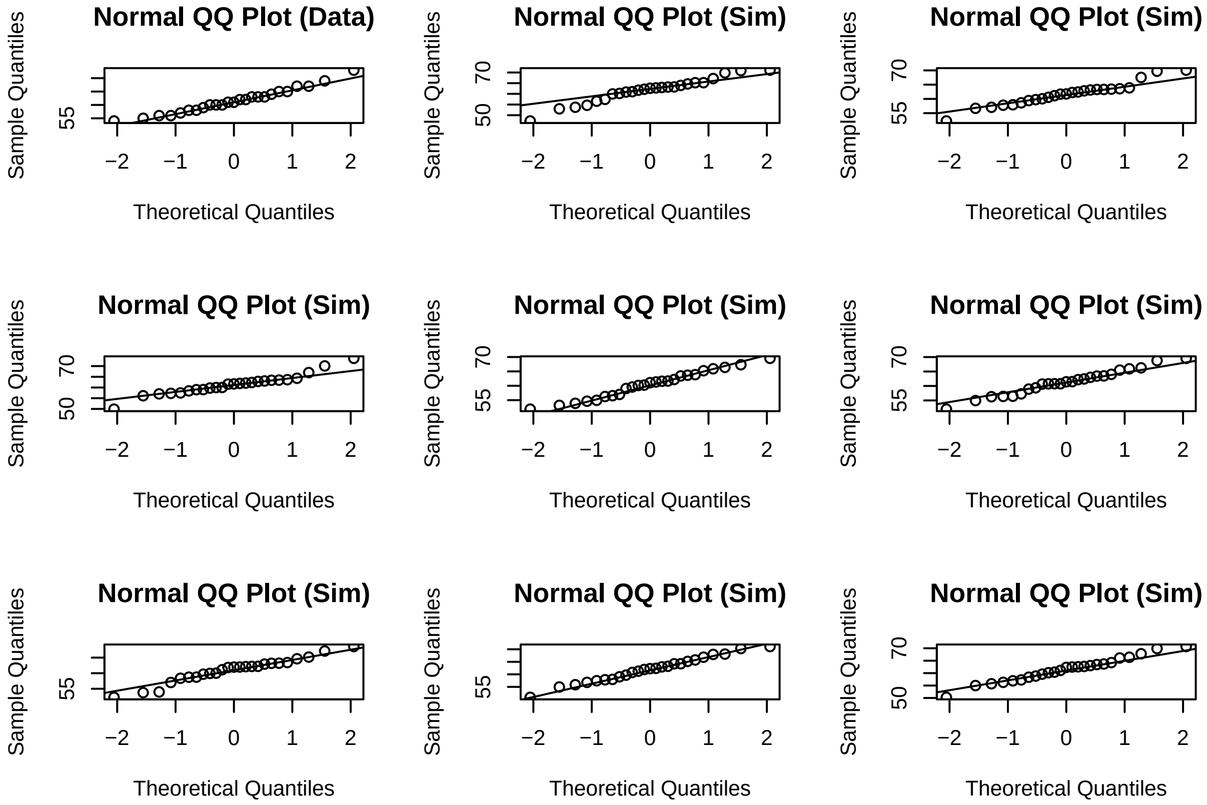
(b) Do these data appear to follow a normal distribution? Explain your reasoning using the graphs provided below.

- *Based from the plots below, the data does appear to be normally distributed. The plots in the Q-Q plot are very close to the line as well as the histogram showing a unimodal distribution with the mean close to the center.*





```
# Use the DATA606::qqnormsim function
DATA606::qqnormsim(heights)
```



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?

- *There is a 1.67% chance that the 10th transitor will the first to be defect.*

```
#Geometric distribution in R
dgeom(10 - 1, 0.02)
```

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

(b) What is the probability that the machine produces no defective transistors in a batch of 100?

- *There is a 13.26% probability that the machine produces no defective transitors in a batch of 100.*

```
#Binomal distribution in R
dbinom(0, 100, 0.02)
```

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

(c) On average, how many transistors would you expect to be produced before the first with a defect? What is the standard deviation?

- *On average, 49 transistors are expected to be produced before the first (50th) is defective. The standard deviation is 49.5*

```
#Mean of expected transistors expected to be produced before first defect
 p <- 0.02
mean <- 1 / p
mean
```

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

```
#Standard deviation of expected transistors expected to be produced before first defect
p <- 0.02
sd <- sqrt((1 - p) / p^2)
sd
```

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

(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?

- *With a 5% defective rate, there is an expected 19 transistors to be produced before the first (20th) is defective. The standard deviation in this case is 19.5.*

```r
#Mean of expected transistors expected to be produced before first defect
 p <- 0.05
mean <- 1 / p
mean
```

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

```r
#Standard deviation of expected transistors expected to be produced before first defect
p <- 0.05
sd <- sqrt((1 - p) / p^2)
sd
```

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

(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?

- *Based on answers to parts (c) and (d), the increasing of probability leads to a decrease in the mean and standard deviation.*

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.

- *Probability that two of the three children will be boys is 0.3823 or 38.23%*

```
#Binomial Distribution
dbinom(2, 3, 0.51)
```

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

(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.

- *Probability is 38.23% matching the answer to part (a).*
- 1 | G B B | 0.49 * 0.51 * 0.51
- 2 | B B G | 0.51 * 0.51 * 0.49
- 3 | B G B | 0.51 * 0.49 * 0.51

```
#Scenarios for 2 boys
P1 <- 0.49 * 0.51 * 0.51
P2 <- 0.51 * 0.51 * 0.49
P3 <- 0.51 * 0.49 * 0.51

P <- P1 + P2 + P3
P
```

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

(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).

- *The approach from part(b) is more tediuous because we would have to write out each individual scenario. When having larger data, this becomes a very complicated process that wastes time. As opposed to the approach from part(a) it'll make the process much smoother with the same outcome as part(b).*

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?

```
#Negative Binomial Distribution
dnbinom(7, 3, 0.15)
```

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

(b) Suppose she has made two successful serves in nine attempts. What is the probability that her 10th serve will be successful?

- *The probability that her 10th serve will be successful is 15% since her success does not depend on the previous attempt.*

(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?

- *The probabilities in both parts(a) and (b) are different because there are two different questions being asked. In part(a) we are being asked for the probability of the event in a particular order, where as in part(b) there's a sense of independence between the events giving the volleyball player the same percentage with each serve.*