Chapter 4 - Distributions of Random Variables

Shana Green

DATA 606 - Homework 4

Due Date: 9/20/2020

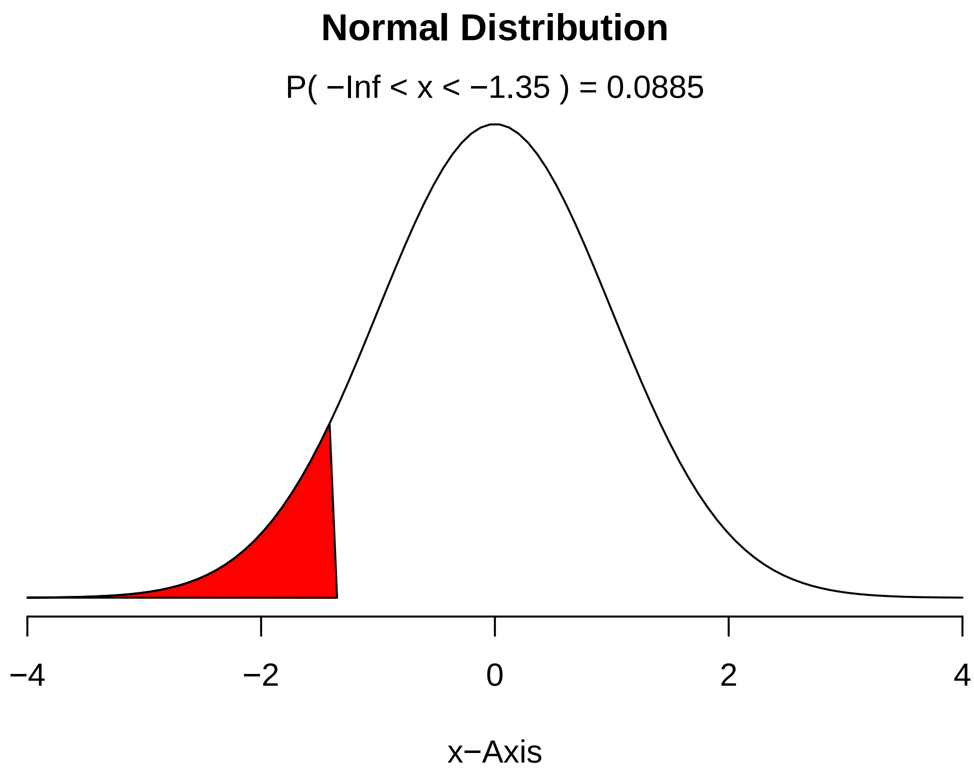
Area under the curve, Part I. (4.1, p. 142) What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph.

(a) $Z < -1.35$

```
mu <- 0
sd <- 1
Z <- -1.35

#find x
x <- (Z * sd) + mu
1 - pnorm(x, mean = 0, sd = 1)
```

```
## [1] 0.911492
```

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

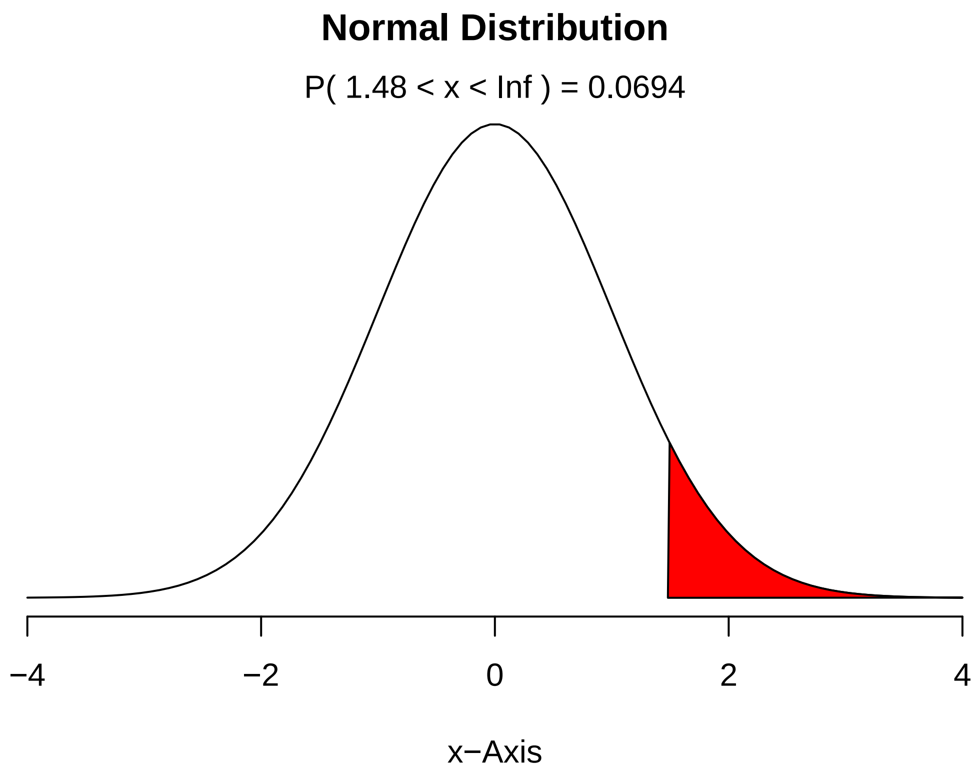


(b) $Z > 1.48$

```
mu <- 0
sd <- 1
Z <- 1.48

#find x
x <- (Z * sd) + mu
1 - pnorm(x, mean = 0, sd = 1)
```

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

```
normalPlot(mean = 0, sd = 1, bounds = c(x, Inf), tails = FALSE)
```

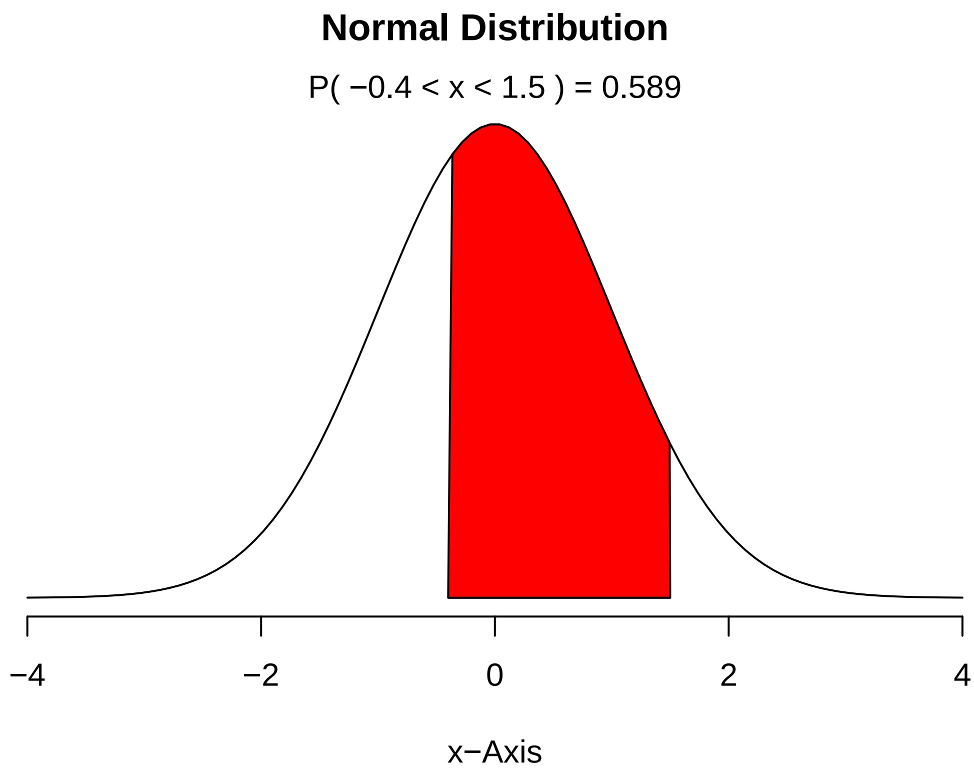


(c) $-0.4 < Z < 1.5$

```
mu <- 0
sd <- 1
Z1 <- -0.4
Z2 <- 1.5

#find x1 and x2
x1 <- (Z1 * sd) + mu
x2 <- (Z2 * sd) + mu

p1 <- 1 - pnorm(x1, mean = 0, sd = 1)
p2 <- 1 - pnorm(x2, mean = 0, sd = 1)

p2 - p1
```

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

```
normalPlot(mean = 0, sd = 1, bounds = c(x1, x2), tails = FALSE)
```

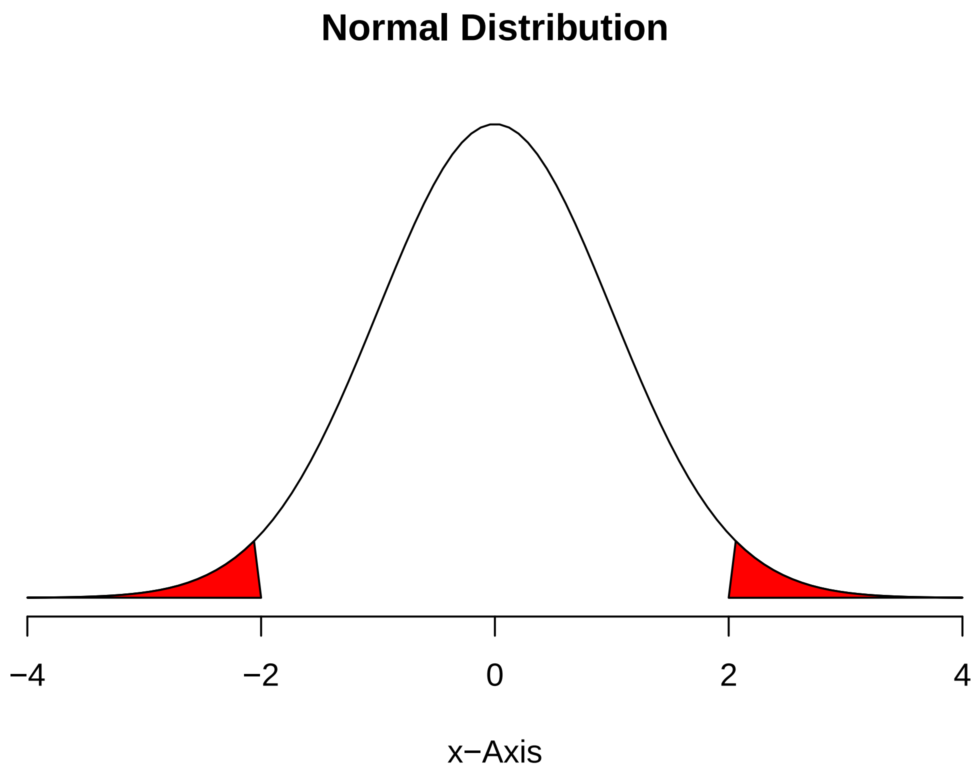


(d) $|Z| > 2$

```
mu <- 0
sd <- 1

# Since this is an absolute value, we need two pnorm values for Z
pnorm(-2, mean = 0, sd = 1) + 1 - pnorm(2, mean = 0, sd = 1)
```

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

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


Normal Distribution
−4
−2
0
2
4
x−Axis

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.

N(mu = 4313, s^2 = 583) -> Men, Ages 30 - 34 N(mu = 5261, s^2 = 807) -> Women, Ages 25 - 29

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

```
Leo <- (4948-4313)/583
Mary <- (5513-5261)/807

Leo
```

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

```
Mary
```

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

Leo is 1.09 sds slower than the men's mean for his specific group. Mary is 0.31 sds slower than the women's mean for her specific group.

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

```
# Leo performed better than 13.8% of his group
1 - pnorm(4948, mean = 4313, sd = 583)
```

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

```
# Mary performed 37.8% of her group
1 - pnorm(5513, mean = 5261, sd = 807)
```

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

Although Leo had a faster time, his z-score is lower than the average. Mary, on the other hand, her scores were closer to the mean. Both Mary and Leo performed below average.

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

```r
x <- 1 - pnorm(4948, mean = 4313, sd = 583)
x * 100
```

```
## [1] 13.80342
```

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

```r
y <- 1 - pnorm(5513, mean = 5261, sd = 807)
y * 100
```

```
## [1] 37.74186
```

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

If the distribution of the finishing times are not nearly normal then Z-scores and percentage will change but not much as the sample size is large enough.

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.

```
female_students <- 25
mu <- 61.52
sd <- 4.58
sd_2 <- 2 * sd
sd_3 <- 3 * sd

x <- mu+sd
x_2<- mu-sd
x
```

```
## [1] 66.1
```

```
x_2
```

```
## [1] 56.94
```

```
y <- mu + sd_2
y_2 <- mu - sd_2

y
```

```
## [1] 70.68
```

```
y_2
```

```
## [1] 52.36
```

```
z <- mu + sd_3
z_2 <- mu - sd_3

z
```

```
## [1] 75.26
```

```
z_2
```

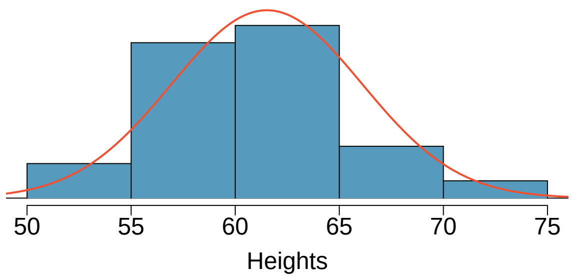
```
## [1] 47.78
```

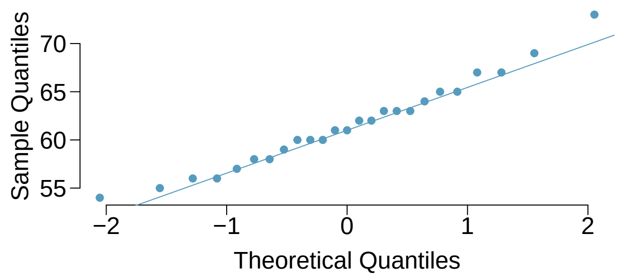
According to 68-95-99.7% rule, 68% data should be within one standard deviation of mean. That is 68% of data should be between 56.94 and 66.1. In the given data, 19 data values are within 1 standard deviation of mean. Since there are total 25 data values so for the given data, (19/25)*100=76% data lies within 1 standard deviation of mean.

According to 68-95-99.7% rule, 95% data should be within 2 standard deviations of mean. That is 95% data should be between 52.36 and 70.68. In the given data, 24 data values are within 2 standard deviations of mean. Since there are total 25 data values so for the given data, (24/25)*100=96% data lies within 2 standard deviations of mean.

According to 68-95-99.7% rule, 99.7% data should be within 3 standard deviations of mean. That is 99.7% data should be between 47.78 and 75.26. In the given data, 25 data values are within 3 standard deviations of mean. Since there are total 25 data values so for the given data, (25/25)*100=100% data lies within 3 standard deviations of mean.

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






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

Both graphs shows that distribution of height is approximately normal. The second graph shows a linear relationship between the theoretical and sample quantities.

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?

```
p <- ((1-.02)^(10-1))*.02
p
```

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

The probability that the 10th transistor produced is the first with a defect is 0.01667496.

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

```
p <- ((1-.02)^(100))
p
```

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

The probability that the machine produces no defective transistors in a batch of 100 is 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?

```
avg = 1/.02
avg
```

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

```
sd <- sqrt((1 - .02)/(.02 * .02))
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?

```
avg_2 <- 1/.05
avg_2
```

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

```
sd <- sqrt((1 - .05)/(.05 * .05))
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?

If we increase the probability of an event, then the event is more likely to happen and it would reduce both the wait times 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.

```
p <- dbinom(2,3, .51)
p
```

```
## [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.

```
# Six possibilities
p <- (0.51 * 0.51 * 0.49) + (0.51 * 0.49 * 0.51) + (0.49 * 0.51 * 0.51)
p
```

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

Both answers 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).

```
p <- dbinom(3, 8, 0.51)
p
```

```
## [1] 0.2098355
```

The method under section b would be too tedious since a total of 56 combinations would be required. 8!/((8-3)!*3!) = 56

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?

```
serve <- 0.15
n <- 10
k <- 3
p <- choose(n - 1, k - 1) * (1 - serve)^(n - k) * serve^k
p
```

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

The probability that on the 10th try, she will make her 3rd successful serve is 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 would remain at 15%; the prior numbers are irrelevant.

(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?

There is a discrepancy because we know that she already has two successful attempts. In answer (a), there is unconditional probability. Whereas, in answer (b), there is conditional probability.