Chapter 4 - Distributions of Random Variables

Donald Butler

09/26/2021

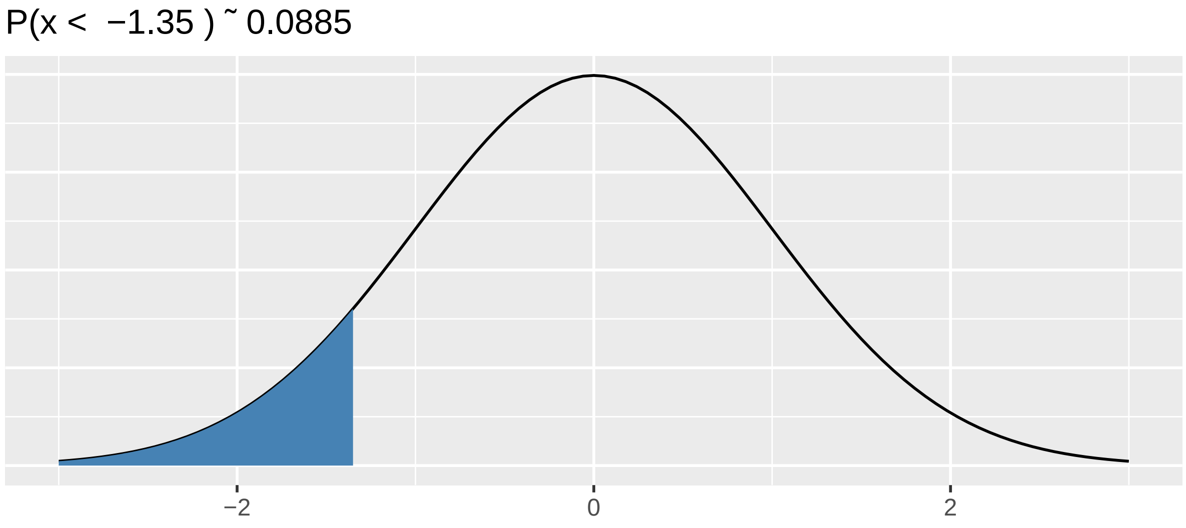
```
library(tidyverse)
library(DATA606)
```

```
##
## 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.
```

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.

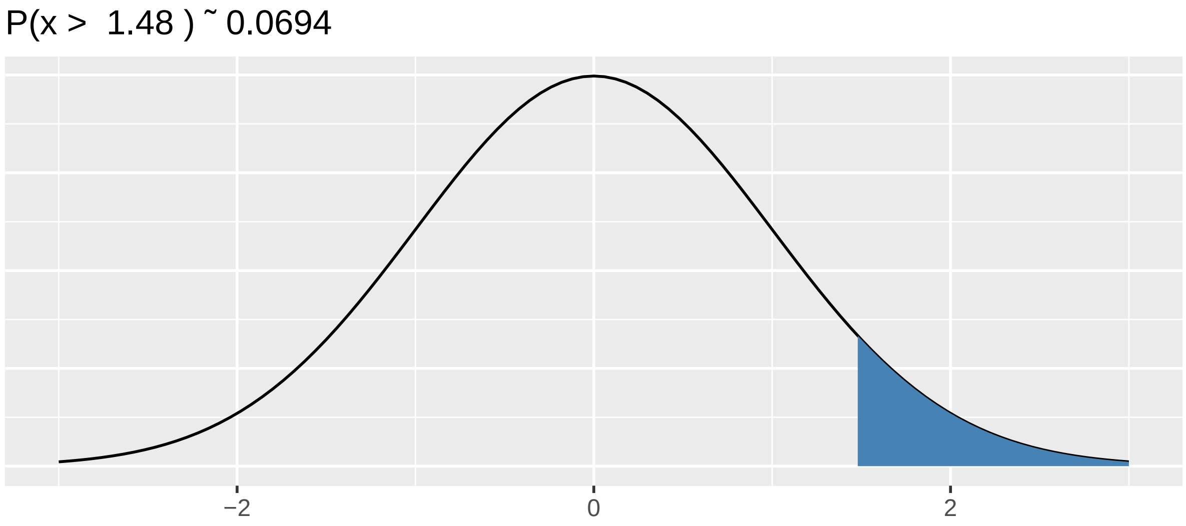
(a) $Z < -1.35$

```
DATA606::normal_plot(cv = -1.35, tails = 'less')
```



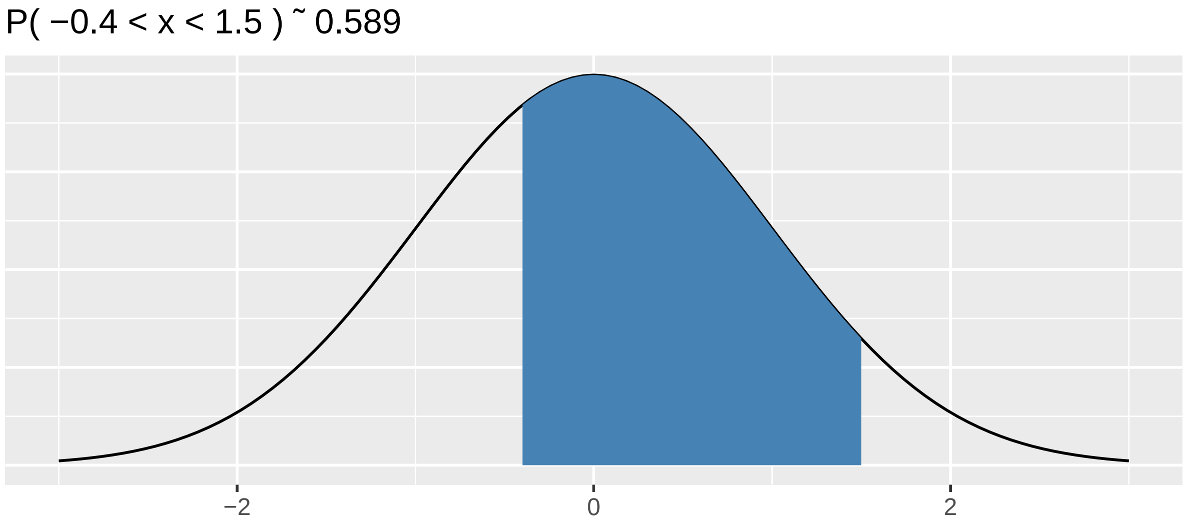
(b) $Z > 1.48$

```r
DATA606::normal_plot(cv = 1.48, tails = 'greater')
```



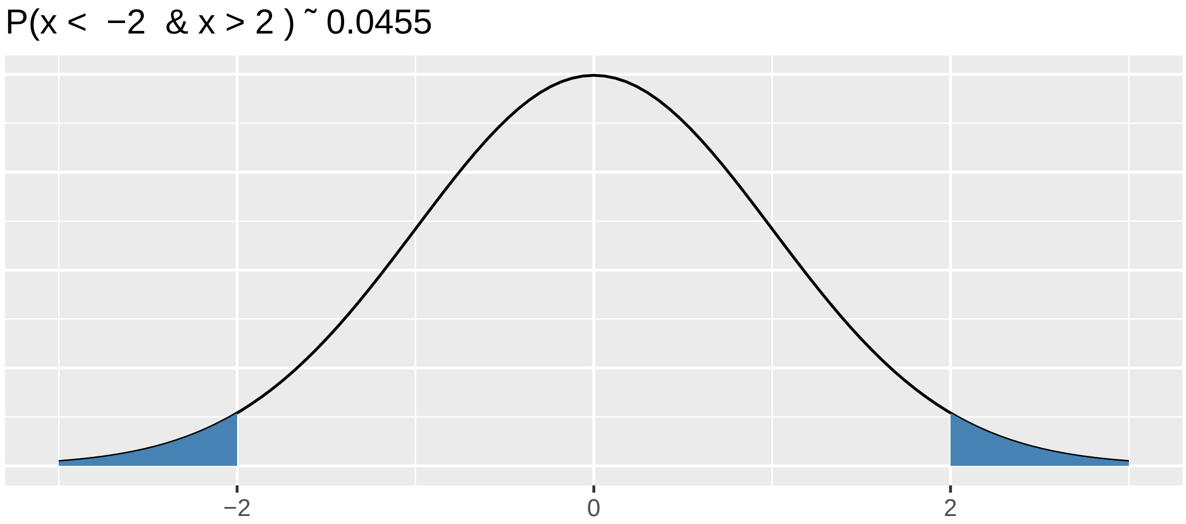
(c) $-0.4 < Z < 1.5$

```r
DATA606::normal_plot(cv = c(-.4,1.5), tails = 'no')
```



(d) $|Z| > 2$

```r
DATA606::normal_plot(cv = 2, tails = 'two.sided')
```


P(x < −2 & x > 2) ˜ 0.0455
−2
0
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.

For *Men, Ages 30 - 34*, N($\mu = 4313$, $\sigma = 583$).
For *Women, Ages 25 - 29*, N($\mu = 5261$, $\sigma = 807$).

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

```
finishLeo <- 4948
meanLeo <- 4313
sdLeo <- 583
(ZLeo <- (finishLeo - meanLeo)/sdLeo)
```

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

```
finishMary <- 5513
meanMary <- 5261
sdMary <- 807
(ZMary <- (finishMary - meanMary)/sdMary)
```

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

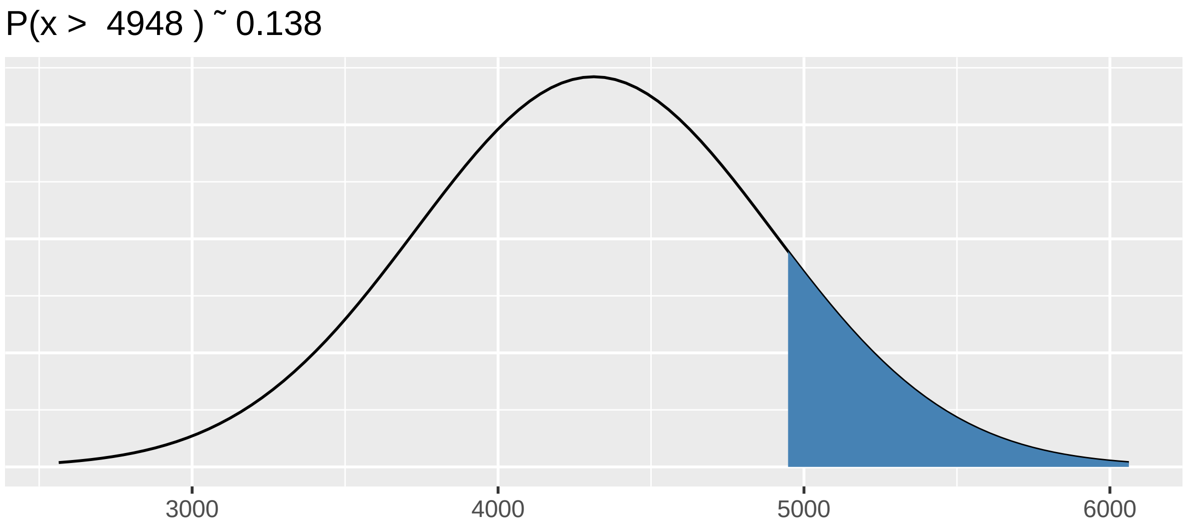
The Z scores represent the number of standard deviations from the mean that they finished. Leo finished 1.09 standard deviations slower than the mean of his group, while Mary finished 0.31 standard deviations slower than her group.

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

Mary ranked better than Leo because her Z score was lower.

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

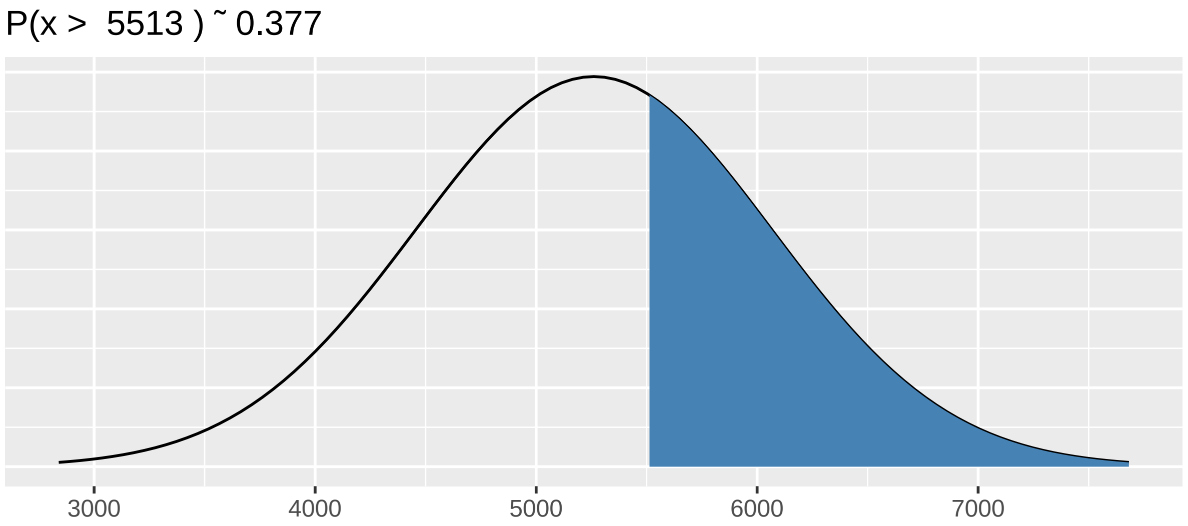
```
DATA606::normal_plot(mean = meanLeo, sd = sdLeo, cv = finishLeo, tails = 'greater' )
```



Leo finished faster than 13.8% of the men in his age group.

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

```
DATA606::normal_plot(mean = meanMary, sd = sdMary, cv = finishMary, tails = 'greater' )
```



Mary finished faster than 37.7% of the women in her age group.

(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 are not nearly normal, than the Z score calculation wouldn't be valid which would change the answers provided in the prior parts of the question.

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.

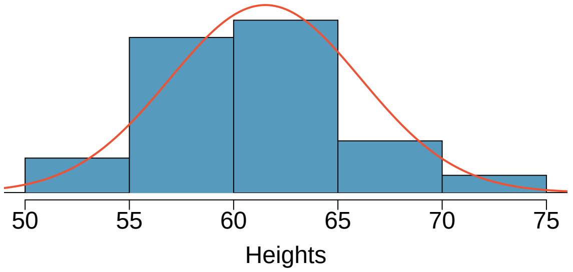
The range of heights within 1σ of the mean is (56.94,66.1). There are 17 observations within 1σ of the mean, which represents 68% of the observations.

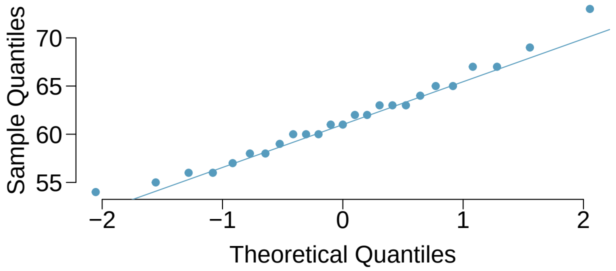
The range of heights within 2σ of the mean is (52.36,70.68). There are 24 observations within 2σ of the mean, which represents 96% of the observations.

The range of heights within 3σ of the mean is (47.78,75.26). All of the observations are within 3σ of the mean.

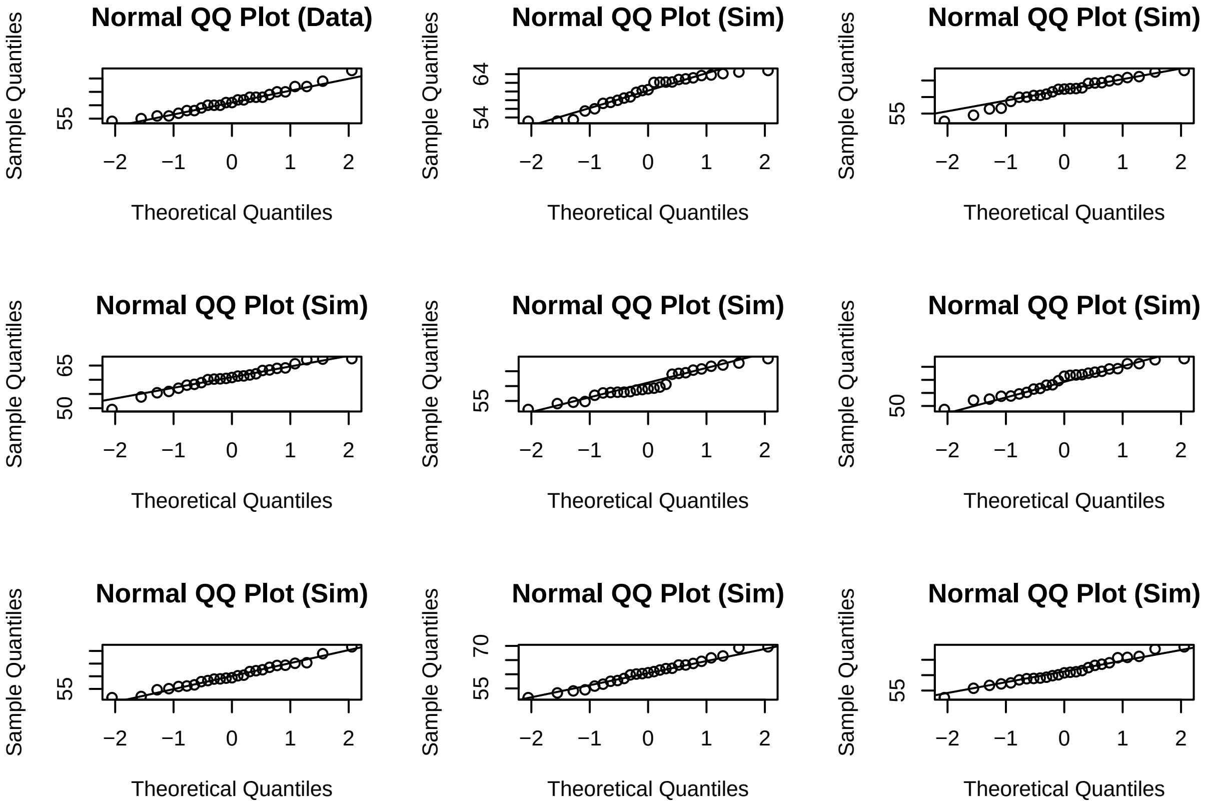
The height observations follow the 68-95-99.7 rule.

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





```
DATA606::qqnormsim(heights)
```



The plots above indicate that the height data is normal.

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?

```
(0.98)^9 * (0.02)
```

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

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

```
(.098)^100
```

```
## [1] 1.326196e-101
```

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

```
p = .02
(mu = 1/p)
```

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

```
(sd = sqrt((1-p)/p^2))
```

```
## [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?

```
p2 = .05
(mu2 = 1/p)
```

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

```
(sd2 = sqrt((1-p)/p^2))
```

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

(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?

Increasing the probability of an event will reduce 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.

```
choose(3,2) * .51^2 * .49
```

```
## [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.

There are 3 ways to have 2 boys; (bbg, bgb, gbb).

```
p_bbg = .51 * .51 * .49
p_bgb = .51 * .49 * .51
p_gbb = .49 * .51 * .51

(p_bbg + p_bgb + p_gbb)
```

```
## [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).

To follow the approach in part (b), you would need to list the 56 possible ways to have 3 boys in 8 children.

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?

```
choose(9,2) * .15^2 * .85^7 * .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 serves are independent so the 10th serve has a 15% probability of being successful regardless of the previous 9 attempts.

(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?

Each serve is independent of previous serves. Part (a) includes the probibilities of the first 9 serves, while part (b) does not.