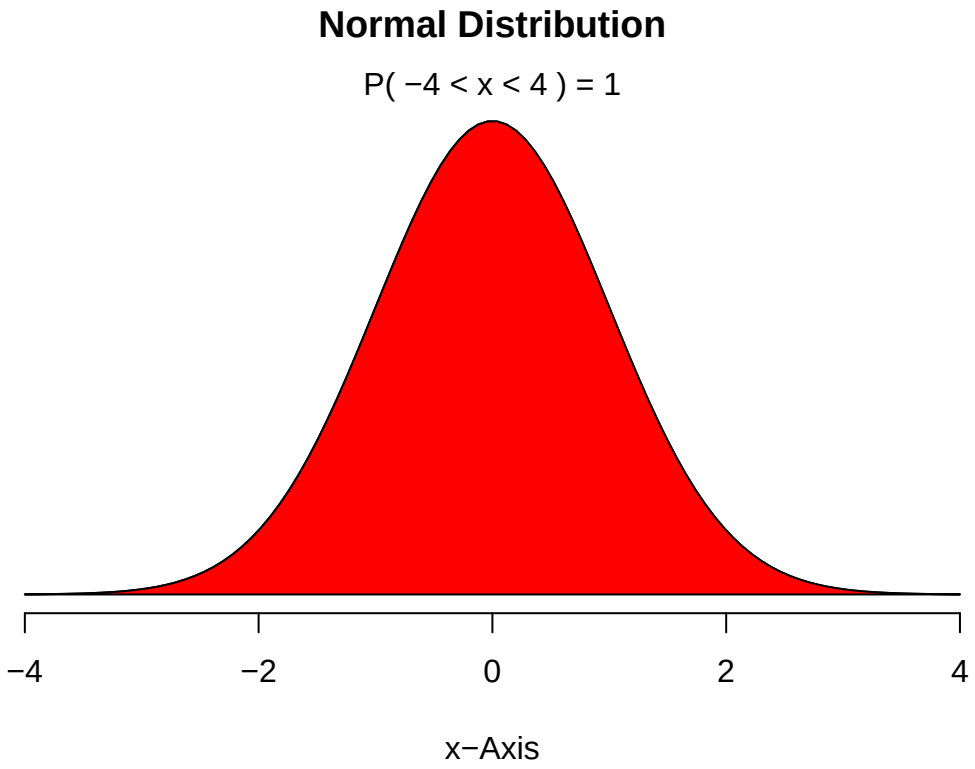
# Chapter 4 - Distributions of Random Variables

Joshua Registe

**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) $|Z| > 2$




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

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

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

```
## [1] 0.9772499
```

a) 8.89 %
b) 93.0 %
c) 58.9 %
d) 97.7 %

---

**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.

The short hand for these two normal distributions would be:

Leo - $N(\mu = 4313, \sigma = 583)$ Mary - $N(\mu = 5261, \sigma = 807)$

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

```
Leo_Z <- (4948-4313)/583
Mary_Z <- (5513-5261)/807

paste("Leo's Z score is",Leo_Z)
```

```
## [1] "Leo's Z score is 1.08919382504288"
```

```
paste("Mary's Z score is",Mary_Z)
```

```
## [1] "Mary's Z score is 0.312267657992565"
```

These z scores are indicating that Leo's finishing time was one standard deviation greater than his group's mean while Mary's finishing time was much closer at less than half a standard deviation greater.

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

Mary ranked better in her respective group because she was closer to the mean finishing time of her group while leo was further away in the positive direction based on the Z-scores.

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

```
paste0("Leo finished faster than ",(1-round(pnorm(Leo_Z),2))*100,"% of his group")
```

```
## [1] "Leo finished faster than 14% of his group"
```

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

```r
paste0("Mary finished faster than ",(1-round(pnorm(Mary_Z),2))*100,"% of her group")
```

```
## [1] "Mary finished faster than 38% of her group"
```

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

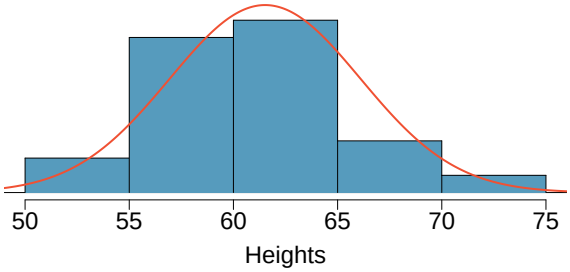
if the distributions of the finishing times are not normal, my answers might change since the z-scores are based on datasets with appoximate normal distributions. The data would have to be transformed prior to doing the analysis

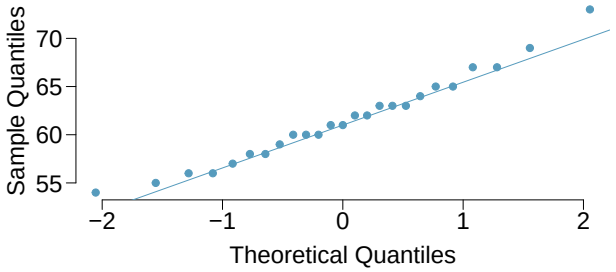
---

**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.




```
# Use the DATA606::qqnormsim function
quantile(heights,.68)
```

```
## 68%
##  63
```

```
qnorm(.5, 61.52,4.58)
```

```
## [1] 61.52
```

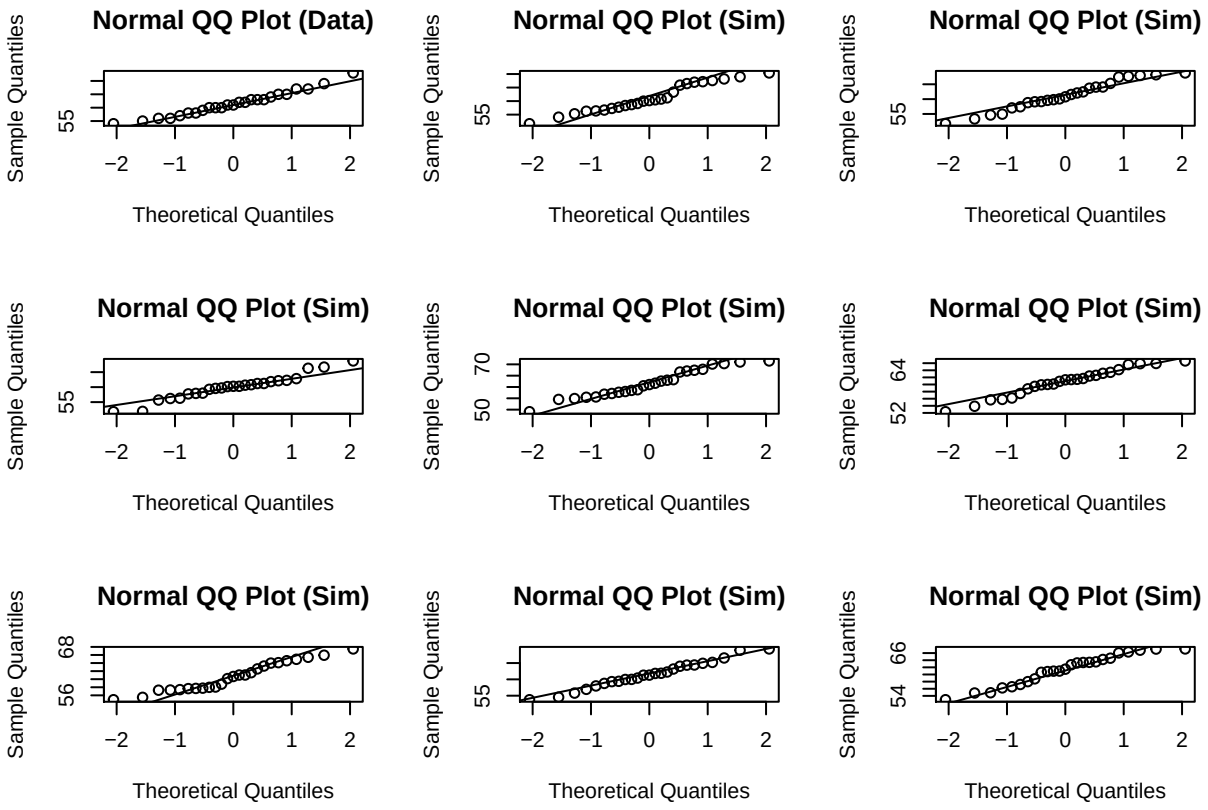
```
quantile(heights,.997)
```

```
##  99.7%
## 72.712
```

```
qnorm(.997, 61.52,4.58)
```

```
## [1] 74.10484
```

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




The dataset does appear to be normally distributed, most of the qqnorm simulation plots do follow the normal distibution z-line with a few of them falling outside the range especially in once you get 2-3 standard deviations away from the mean. this dataset is probably slightly skewed but normal distribution approximations seem appropriate to use.

---

**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?

This will be the probability that the first 9 are not defective and the probability that the 10th is defective:

```
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?

The probability will be 0.98 to the 10th. calculated as:

```
0.98^100
```

```
## [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, I would expect 50 transistors to be produced before the first

(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?

I would expect approximately 1/p to be produced before first failure, in this case 1/0.05 = 20. the variance 1-p/p^2 and the standard deviation is the square root of that. this is calculated as 19.49

```
p = .05

mean_p = 1/p
mean_p
```

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

```
var_p = (1-p)/p^2
stdev_p = sqrt(var_p)
stdev_p
```

```
## [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 the answers to c and d, increasing the probability of an event causes the mean and standard deviation of that expected event to happen in n tries deacreases.

---

**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.

```
binom_model<-function(k,n,p){
choose(n,k)*p^k*(1-p)^(n-k)
}

binom_model(2,3,.51)
```

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

as shown in the above function, the probability is 38.2%

(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.

bbg bgb gbb this is out of 9 orders total ggg gbg ggb bgg bbb

there are 3 different oderings that allow for 2 boys and 1 girls as shown above. Using the rule of disjoints this is 3*1/8 = 37.5%

p = 3!/(2!*(3-2)!)$(0.51)^{2(1-0.51)}$1

p = 3x2x1/[(2x1) * (1)] * $(0.51)^{2(1-0.51)}$1

p = 38.2%

(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 would be more tedious because as you increase the number of trials, the number of possible orderings needed to write out grows extensively and also leads to more room for human error. using the method from part (a) allows for a more efficient method to calculate this.

---

**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?

```
neg_binom_model<-function(k,n,p){
choose(n-1,k-1)*p^k*(1-p)^(n-k)
}

neg_binom_model(3,10,.15)
```

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

Based on the negative binomial function created above, the probability that the volleyball player will make 3 successful serves by her 10th trial is 3.89%

(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 the probability that any 1 of her serves are successful which is approximately 15%

(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 reasoning is because for question (a) we are observing the probability of 3 successes in 10 trials, while question (b) were observing the probability of 1 success in 1 trial.