# Week3HW: OpenIntro Statistics

*Walt Wells, Fall 2016*

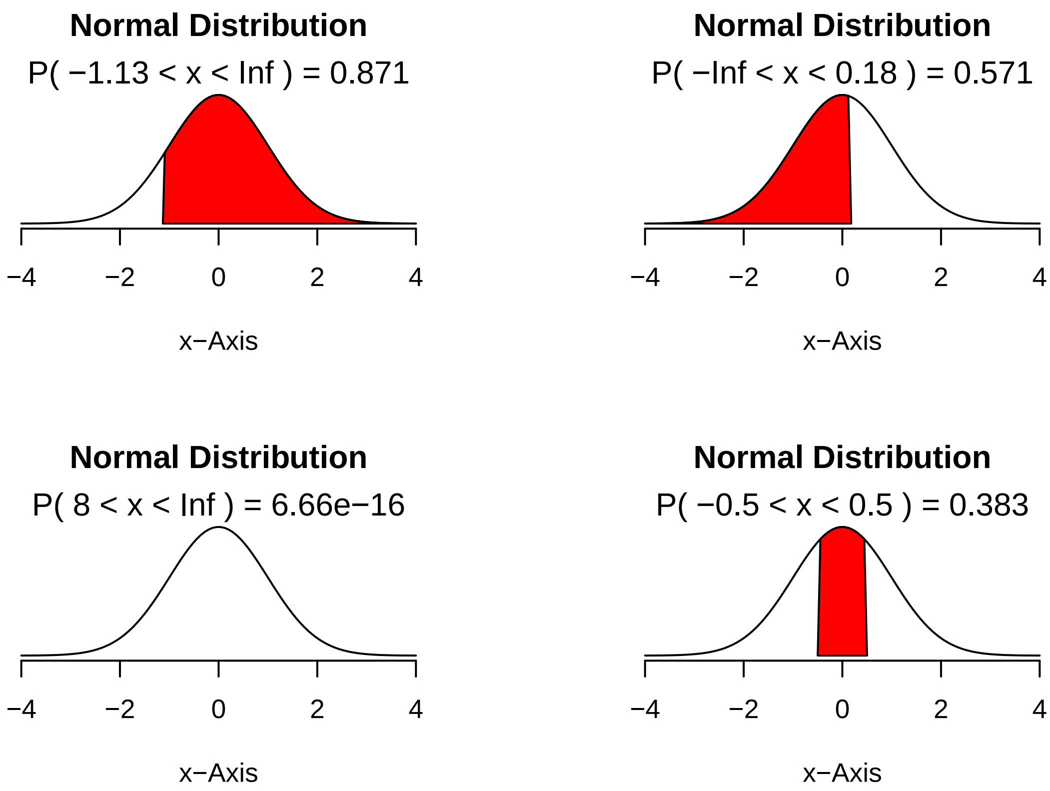
## Graded

### 3.2 Area under the curve, Part II

What percent of a standard normal distribution is found in each region? Be sure to draw a graph.

(a) Z > -1.13
(b) Z < 0.18
(c) Z > 8
(d) |Z| < 0.5

```
# need to review and calc d
par(mfrow=c(2,2))
normalPlot(bounds=c(-1.13, Inf))
normalPlot(bounds=c(-Inf, 0.18))
normalPlot(bounds=c(8, Inf))
normalPlot(bounds=c(-0.5, 0.5))
```




```
#a
round(pnorm(-1.13, lower.tail=FALSE), 3)
```

```
## [1] 0.871
```

```r
#b
round(pnorm(0.18), 3)
```

```
## [1] 0.571
```

```r
#c
pnorm(8, lower.tail=FALSE)
```

```
## [1] 6.220961e-16
```

```r
#d
round(pnorm(.5) - pnorm(-.5), 3)
```

```
## [1] 0.383
```

**3.4 Triathlon times, Part I.**

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.
**Men's: N(4313, 583); Women's: N(5261, 807)**
(b) What are the Z-scores for Leo's and Mary's finishing times? What do these Z-scores tell you?

```r
leoz <- round((4948 - 4313)/583, 2)
maryz <- round((5513 - 5261)/807, 2)
```

**Leo's Z score is 1.09 and Mary's Z score is 0.31. Since higher race time numbers indicate a slower race, in this instance a higher Z score means slower, relative to peers.**

(c) Did Leo or Mary rank better in their respective groups? Explain your reasoning.
**Since her Z score was 0.31 compared to Leo's 1.09, we can determine she was faster. Said another way, Leo was 1.09 standard deviations slower than the mean in his peer group, while Mary was only 0.31 slower than the mean in hers.**

(d) What percent of the triathletes did Leo finish faster than in his group? **Leo was faster than only 13.8% of his group**

(e) What percent of the triathletes did Mary finish faster than in her group? **Mary was faster than only 37.8% 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. **We would not have been able to standardize their scores if the distribution were not nearly normal, so the approaches we took in parts b-e would all need to be reconsidered.**

**3.18 Heights of female college students.**

(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. **Below we calculate the percentage of the distribution that fall within 1, 2, and 3 standard deviations of the mean. Our results (.68, .96, and 1) follow the rule closely.**

(b) Do these data appear to follow a normal distribution? Explain your reasoning using the graphs provided below. **The data does seem to follow the normal distribution. The QQ plot and histogram both indicate normality even with a small sample. We test this assumption by using the qqnormsim function compare a QQ plot of our data to 8 additional QQ plots of selected from the normal distribution with a similar sample size. Our data is actually appears more normal than some of the samples!**

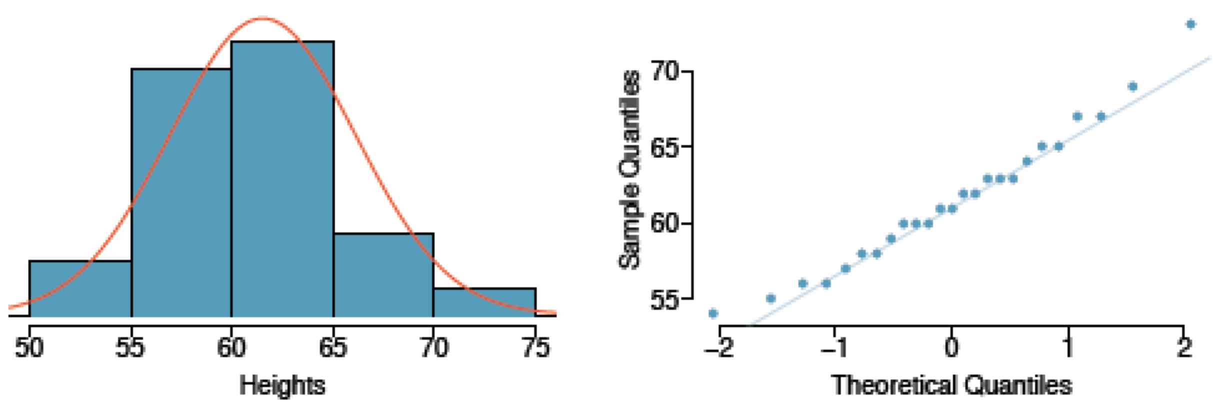



Figure 1: 3.18

```
fheight <- 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)

mu <- mean(fheight)
sd <- sd(fheight)

# test 68
length(fheight[fheight < mu + sd & fheight > mu - sd]) / length(fheight)
```

```
## [1] 0.68
```

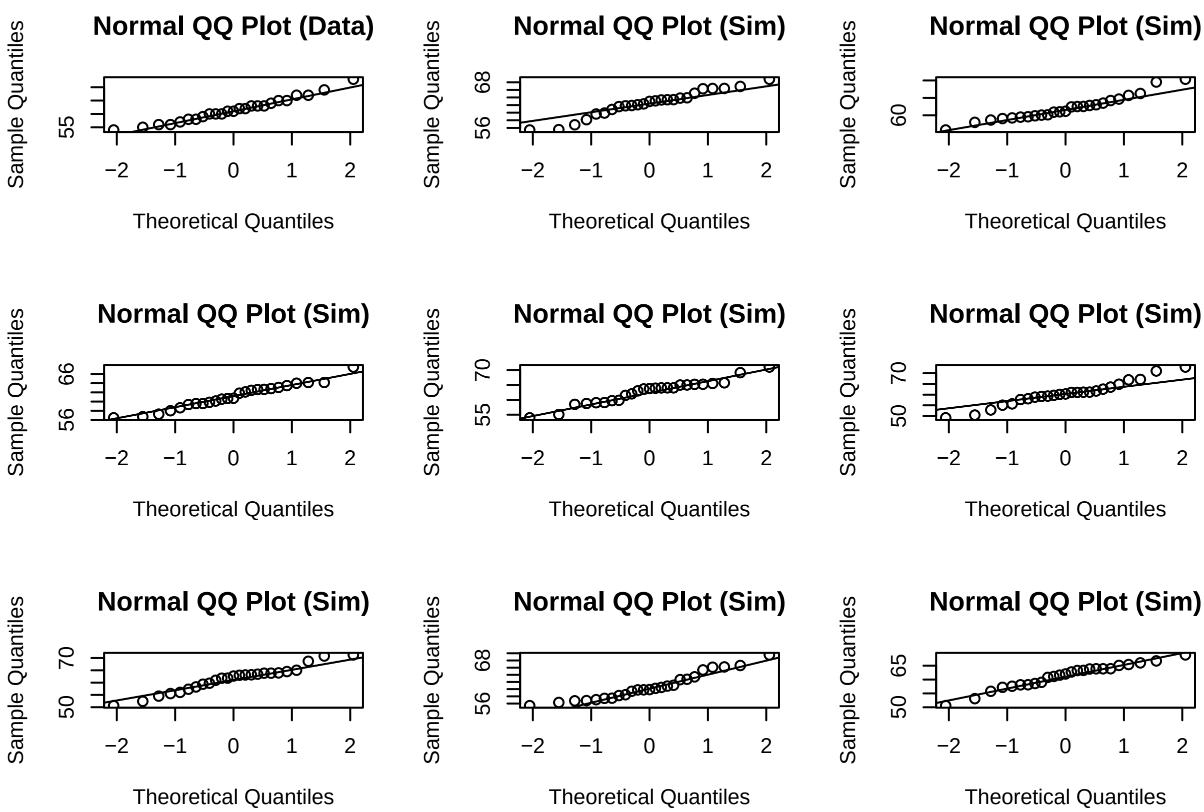
```r
# test 95
length(fheight[fheight < mu + 2*sd & fheight > mu - 2*sd]) / length(fheight)
```

```
## [1] 0.96
```

```r
# test 99.7
length(fheight[fheight < mu + 3*sd & fheight > mu - 3*sd]) / length(fheight)
```

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

```r
qqnormsim(fheight)
```




**3.22 Defective rate.**

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?

```r
#note - r's _geom package does not seem to use n-1, so built own function
geom <- function(n, p){
    p * (1-p) ^ (n-1)
}
p <- .02
geom(10, p)
```

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

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

```r
(1-p)^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?

```r
# first defect at:
1/p
```

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

```r
# 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?

```r
# new p:
p <- .05
# first defect at:
1/p
```

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

```r
# sd:
sqrt((1-p)/p^2)
```

```
## [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? **In the geometric distribution, increasing the probability of an event will lower both the mean and the standard deviation regarding how many trials until an occurence.**

### 3.38 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.

```r
dbinom(2, 3, prob=0.51)
```

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

```r
#confirm
bindist <- function(n, k, p) {
    choose(n, k) * p^k * (1-p) ^ (n-k)
}
bindist(3, 2, 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. **We have a match!**

```r
# B B G
s1 <- .51 * .51 * .49
# B G B
s2 <- .51 * .49 * .51
# G B B
s3 <- .49 * .51 * .51

s1 + s2 + s3
```

```
## [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). **Drawing out the sample space would require 56 samples. That's not a particularly fun exercise.**

**3.42 Serving in volleyball.**

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.

```r
# let's use the negative binomial distribution.   note, similar to
# the geometric distribution, dnbinom does not use n-1 and k-1, so
# we'll define our own function to ensure accuracy
nb <- function(n, k, p) {
    choose(n-1, k-1) * p^k * ((1-p)^(n-k))
}
```

(a) What is the probability that on the 10th try she will make her 3rd successful serve?
**The probability is 0.039**

(b) Suppose she has made two successful serves in nine attempts. What is the probability that her 10th serve will be successful? **Since we are supposing that her serves are independant, we believe that her 10th serve will simply be the original probability of .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? **This is consistent with the principles of the negative binomial distribution. Sine we're calculating at n-1 and k-1, even applying something like the probability of 1 in 1 (.15) times the probability of 2 in 9 (0.0577) would yield different results (0.0087) than calculating 3 in 10 (0.039).**