Week3hw

Charley Ferrari

Wednesday, September 23, 2015

Question 3.2: Area under the curve, Part II

What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph.

Note: Since the plot goes to 4, I'm using 4 in the normalPlot bounds for these diagrams. Technically this would be ∞ or $-\infty$.

a. $Z > -1.13$

```
library(IS606)
```

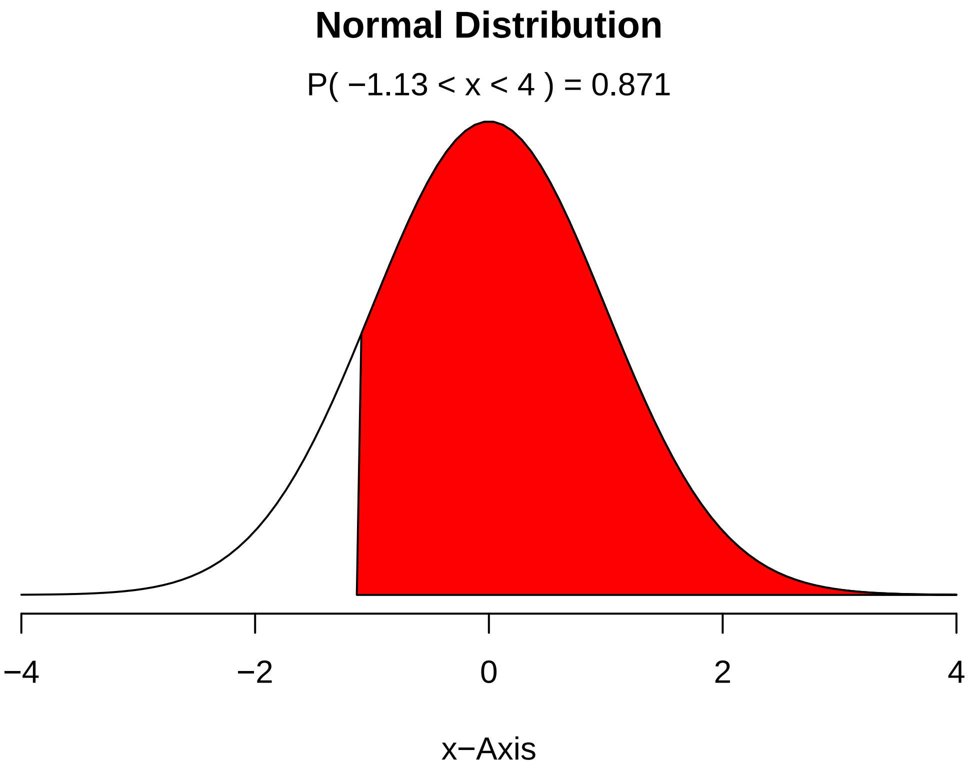
```
##
## Welcome to CUNY IS606 Statistics and Probability for Data Analytics
## This package is designed to support this course. The text book used
## is OpenIntro Statistics, 3rd Edition. You can read this by typing
## vignette('os3') or visit www.OpenIntro.org.
##
## The getLabs() function will return a list of the labs available.
##
## The demo(package='IS606') will list the demos that are available.
```

```
##
## Attaching package: 'IS606'
##
## The following object is masked from 'package:utils':
##
##     demo
```

```
1-pnorm(-1.13)
```

```
## [1] 0.8707619
```

```
normalPlot(bounds = c(-1.13,4))
```

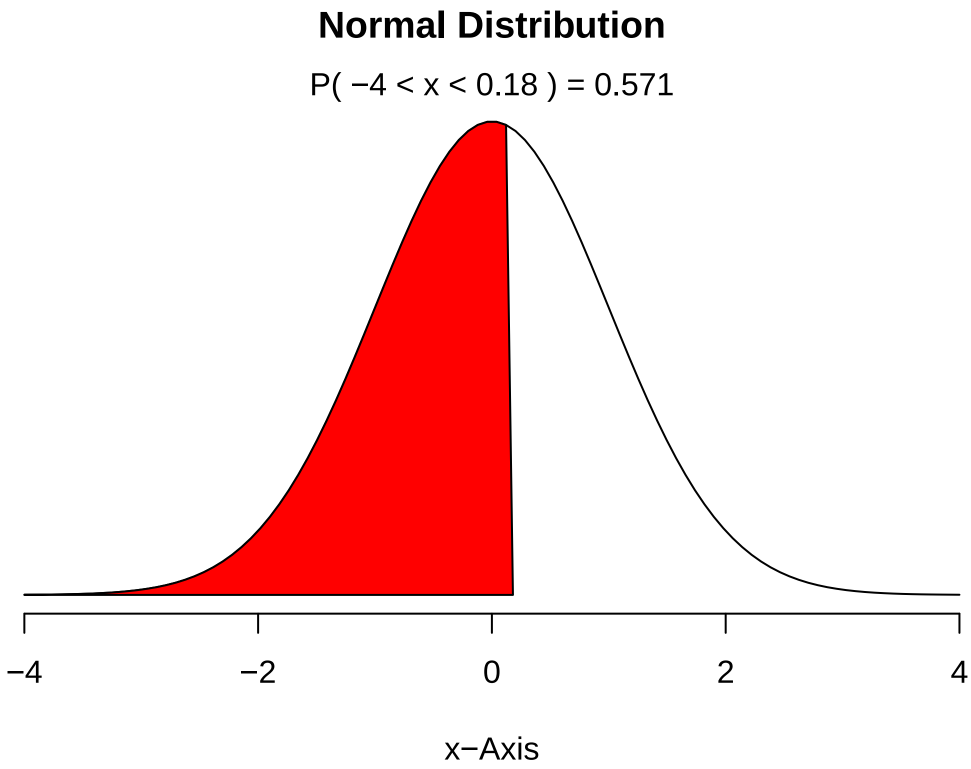


b. $Z < 0.18$

```
pnorm(0.18)
```

```
## [1] 0.5714237
```

```
normalPlot(bounds = c(-4, 0.18))
```

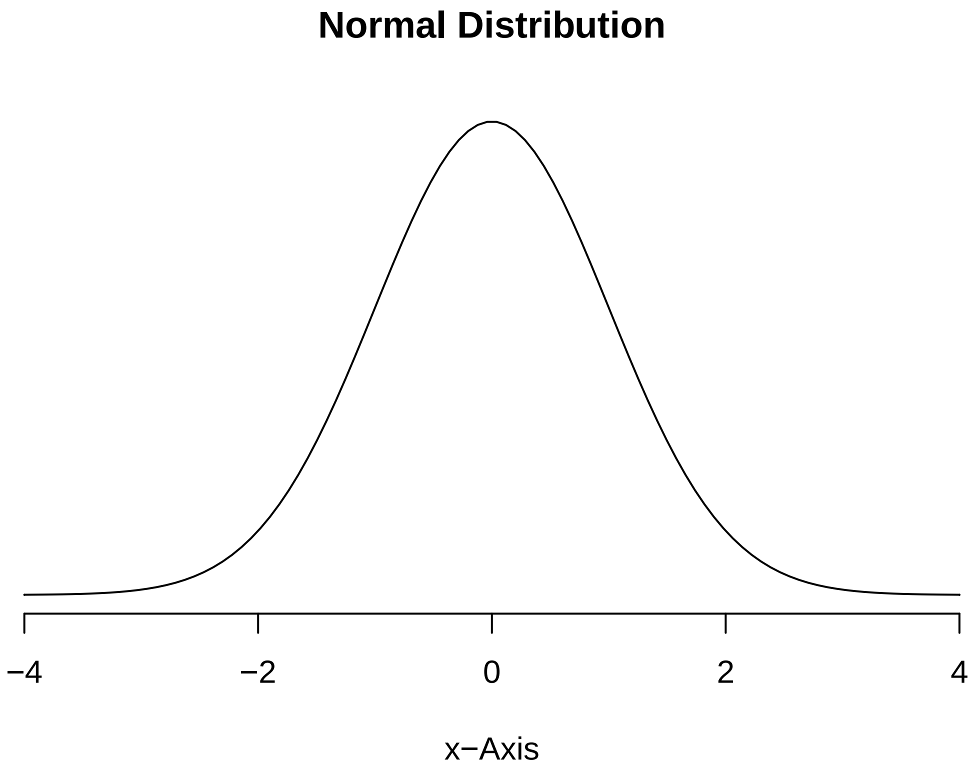


c. $Z > 8$

```
pnorm(8)
```

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

```
normalPlot(bounds = c(-4,4), tails=TRUE)
```



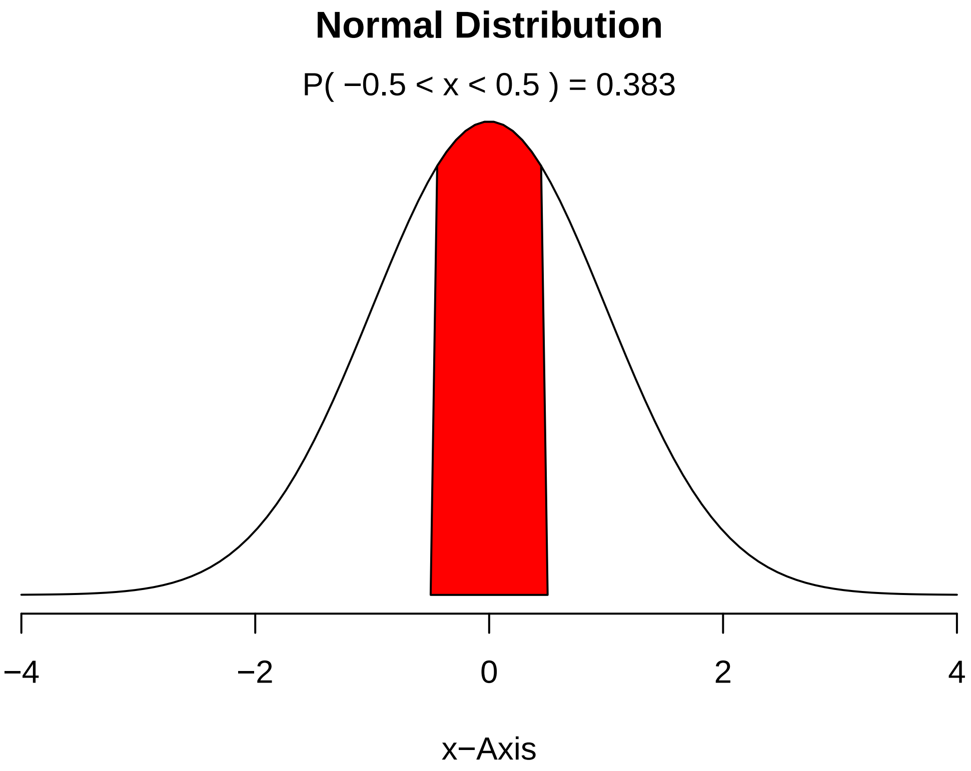
This is the probability of an observation being 8 standard deviations away from the mean, which in a normal distribution is incredibly rare (even though the normal curve assymptotically approaches 0.)

d. $|Z| < 0.5$

```
pnorm(0.5) - pnorm(-0.5)
```

```
## [1] 0.3829249
```

```
normalPlot(bounds=c(-0.5, 0.5))
```



3.3: GRE Scores, Part I

Sophia who took the GRE scored 160 on the Verbal Reasoning section and 157 on the Quantitative Reasoning section. The mean score for Verbal Reasoning section for all test takers was 151 with a standard deviation of 7, and the mean score for the Quantitative Reasoning was 153 with a standard deviation of 7.67. Suppose that both distribution are nearly normal.

a. Write down the short-hand for these two normal distributions:

Verbal Reasoning ~ $N(\mu = 151, \sigma = 7)$ Quantitative Reasoning ~ $N(\mu = 153, \sigma = 7.67)$

b. What is Sophia's Z-score on the Verbal Reasoning section?

$$Z = \frac{x - \mu}{\sigma} = \frac{160 - 151}{7}$$

```
zv <- (160-151)/7
zv
```

```
## [1] 1.285714
```

On the Quantitative Reasoning section?

```
zq <- (157-153)/7.67
zq
```

```
## [1] 0.5215124
```

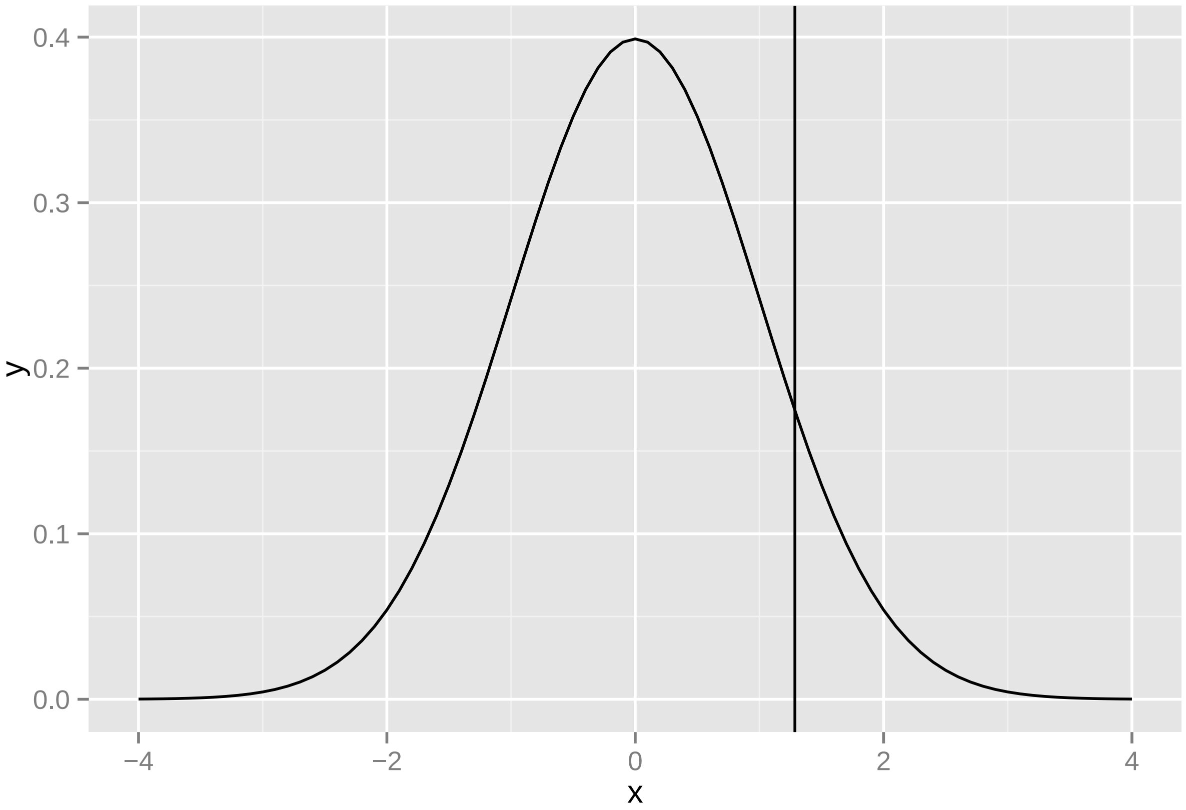
Draw a standard normal distribution curve and mark these two Z-Scores

For the Verbal Reasoning z-score:

```
library(ggplot2)
library(dplyr)
```

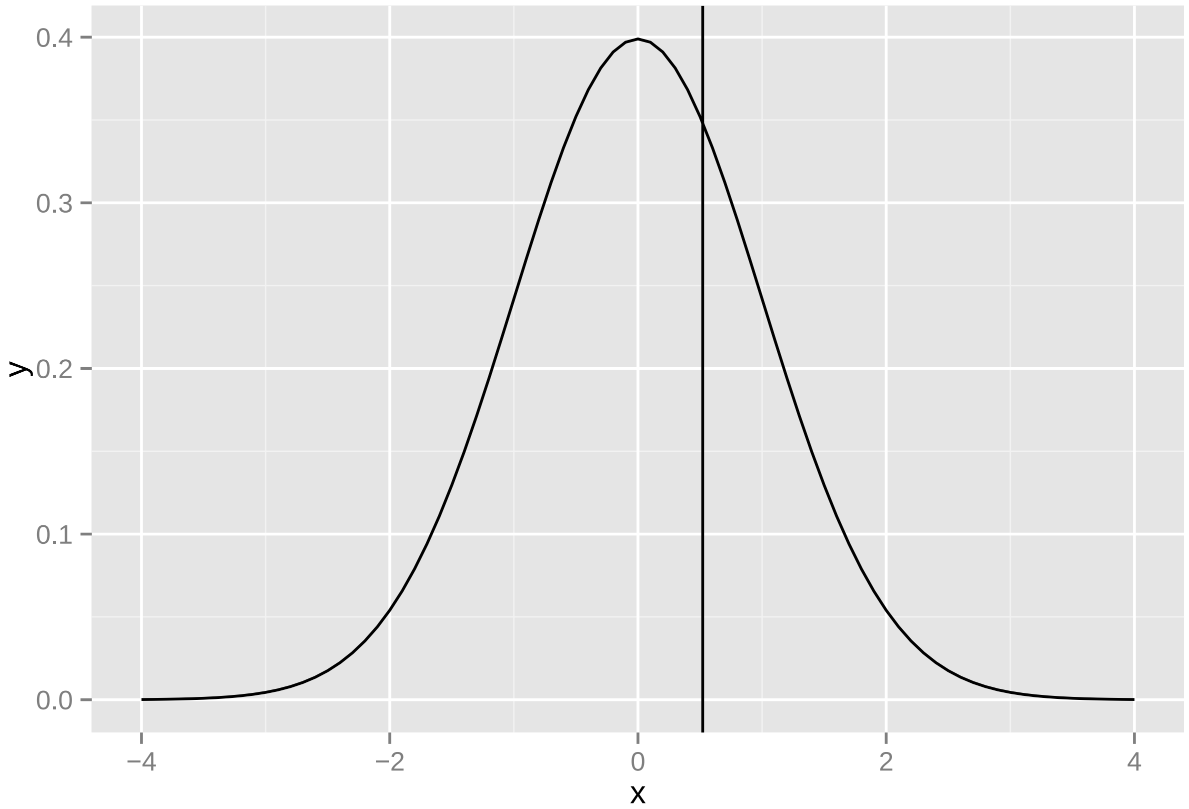
```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
x = seq(-4,4,by=0.1)

normaldf <- data.frame(x = x, y = dnorm(x))

ggplot(normaldf,aes(x=x,y=y)) + geom_line() + geom_vline(xintercept=zv)
```



For the Quantitative Reasoning z-score:

```r
ggplot(normaldf, aes(x=x, y=y)) + geom_line() + geom_vline(xintercept=zq)
```



c. What do these Z-scores tell you?

These Z-scores tell me how many standard deviations away from the mean Sophia's scores are for the verbal and quantitative scores. Plotting them along with the normal distribution give a visual feel for how Sophia performed relative to the general population. Using the normal distribution, one can find out the percentile Sophia was in.

d. Relative to others, which section did she do better on? Relative to others, she did better on the verbal reasoning section. This is because her Z-score in this section is higher than her Z-score in the quantitative reasoning section.

e. Find her percentile scores for the two exams

```
verbalquant <- pnorm(zv)

quantitativequant <- pnorm(zq)

verbalquant
```

```
## [1] 0.9007286
```

```
quantitativequant
```

```
## [1] 0.6989951
```

FYI, pnorm(zv) = pnorm(rawverbalscore, mean=151, sd=7)

f. What percentage of test takers did better than her on the Verbal Reasoning section?

```
1 - verbalquant
```

```
## [1] 0.0992714
```

On the Quantitative Reasoning section?

```
1 - quantitativequant
```

```
## [1] 0.3010049
```

g. Explain why simply comparing her raw scores from the two sections would lead to the incorrect conclusion that she did better on the Quantitative Reasoning section.

The raw scores seem fairly close on their own. If the two scores are out of the same total score, one might think that she did a bit worse on the quantitative reasoning section. However, the percentiles show that she's in the 70th percentile for the quantitative section while she's in the 90th percentile for the verbal section.

h. If the distributions of the scores on these exams are not nearly normal, would your answers to parts b - f change? Explain your reasoning.

My answers would definitely change, because the Z scores are calculated using the normal distribution. The answers depend on the area under the normal curve, and if there were a different distribution, percentiles would need to be calculated based on those distributions. If scores are uniformly distributed, for example, the percentile would be equal to the perntage of the score out of the highest possible score.

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 teh heights approximately follow the 65-95-99.7% rule

I'll test this out by taking the length of the heights vector where the heights fall in one SD, two SD, etc and dividing it by the total number of observations.

```
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)

heightsmean <- 61.52

heightssd <- 4.58

length(heights[heights<heightsmean+heightssd & heights>heightsmean-heightssd])/length(heights)
```

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

```
length(heights[heights<heightsmean+2*heightssd & heights>heightsmean-2*heightssd])/length(heights)
```

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

```
length(heights[heights<heightsmean+3*heightssd & heights>heightsmean-3*heightssd])/length(heights)
```

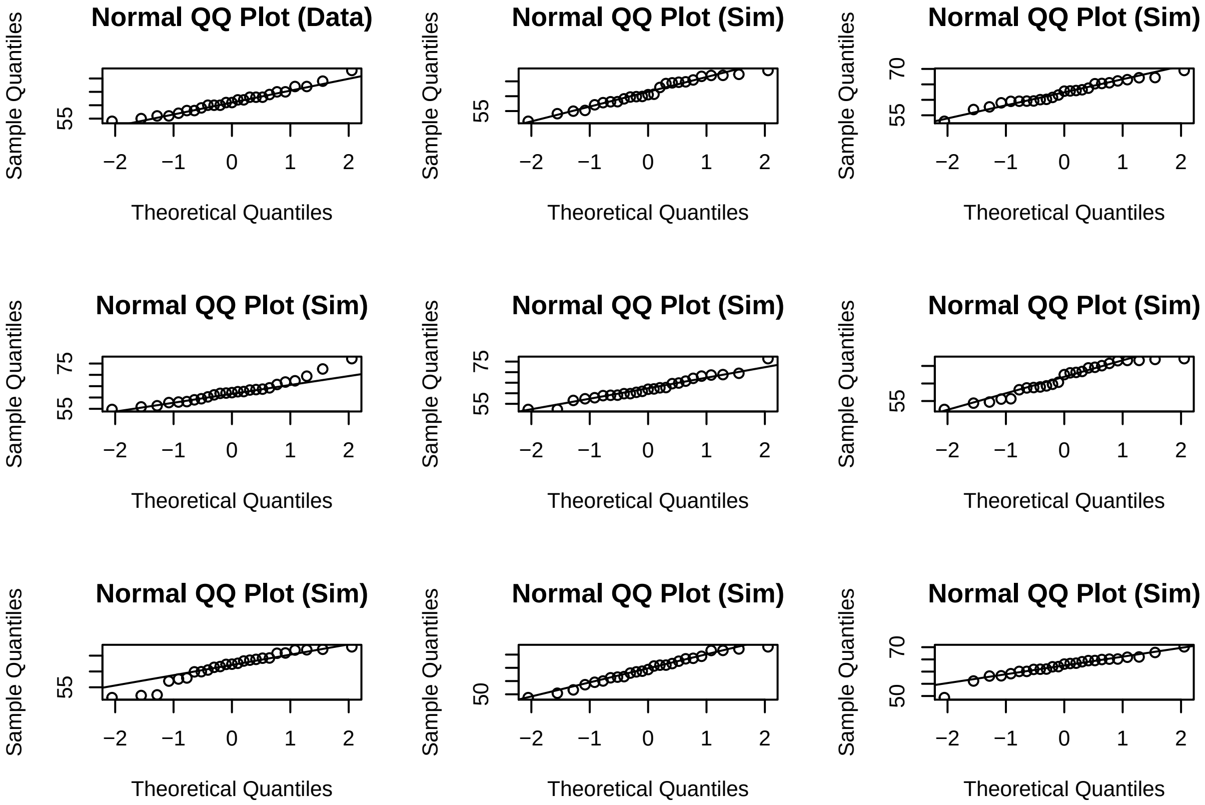
```
## [1] 1
```

This data does seem to match the 65-95-99.7% rule approximately.

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

Lets use the qqnormsim function to compare the data to simulations of normally distributed data with the same mean and standard deviation

```
qqnormsim(heights)
```



It does appear to be normal (even more normal than many of the simulations!)

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?

Geometric Distribution:

In this case, we're counting "success", as finding the defective part. So, with p being the defective rate, we end up with a probability of $(1-p)^{n-1}p$

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

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

This also makes intuitive sense. The success rate above is 98%. If the first defective transistor is the 10th one, that means 9 successful transistors have been made. So by just multiplying the probabilities together, you end up with theanswer.

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

Once again, the success rate is 98%, so you just multiply that 100 times to get the total probability of getting 100 successes.

```
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?

Since this is a geometric distribution, $\mu = \frac{1}{p}$, and $\sigma = \sqrt{\frac{1-p}{p^2}}$

```
ptransistor <- 0.02

mutransistor <- 1/ptransistor

sdtransistor <- sqrt((1-ptransistor)/(ptransistor^2))

mutransistor
```

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

```
sdtransistor
```

```
## [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?

Here, we're just going to replace the 2% above with 5%, so p = 0.05

```
ptransistor <- 0.05

mutransistor <- 1/ptransistor

sdtransistor <- sqrt((1-ptransistor)/(ptransistor^2))

mutransistor
```

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

```
sdtransistor
```

```
## [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?

Increasing the probability of an event (in this case a transistor failure) decreases the mean number of days before the first time that event occurs. This makes intuitive sense, if the probability of an event occurring is increased, it's going to happen on average sooner. The fact that the standard deviation also increases makes intuitive sense. If something has a low chance of occurring, there are more chances that it will occur on different days. If the probability of something occurring is low, the probability of it occurring on any particular day is low. Because these events are independent, there is no such concept as "this event is due to occur". In the above examples, if we went 150 days without the event occurring, the probability of the event occurring on the 151st day is still 2% and 5%.

3.38: Male Children

While it is often assumed that he probabilities of having a boy or 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

Here, we're using p = 0.51, p being the probability of a child being a boy. The formula for the binomial model is:

$$\binom{n}{k} p^k (1-p)^{n-k}$$

For this example, n = 3, and k = 2

```
n <- 3

k <- 2

p <- 0.51

numscenarios <- factorial(n) / (factorial(k)*factorial(n - k))

pboy <- numscenarios*(p^k)*((1-p)^(n-k))

pboy
```

```
## [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.

The possible orderings for 3 children are:

GGG BGG GBG **BBG** GGB **BGB GBB** BBB

With the scenarios where there are two boys highlighted. We can calculate the probability of each of these events occurring by using the fact that the probability of having a girl is 0.49 and the probability of having a boy is 0.51. For each of these, we'll multiply the probabilities over three children.

In the case of there being two boys and a girl, the probability will always be the same:

```r
0.51*0.51*0.49
```

```
## [1] 0.127449
```

So, to calculate the probability, we just take the three cases with two boys, and multiply that by the probability of having three boys:

```r
0.51*0.51*0.49*3
```

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

```r
pboy
```

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

They 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.

This would become tedious mainly because of the work involved in listing out the total number of outcomes, and then counting the outcomes we want. For 8 children, this very quickly becomes hard to do. The number of scenarios is displayed below:

```r
factorial(8) / (factorial(3)*factorial(8 - 3))
```

```
## [1] 56
```

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.

a. What is the probability that on the 10th try she will make her 3rd successful serve?

This can be solved using the negative binomial distribution. p in this case will be 0.15. We're trying to calculate the probability of observing the kth (3rd) success on her nth (10th) trial. The formula that we're using is below:

$$\binom{n-1}{k-1} p^k (1-p)^{n-k}$$

```r
p <- 0.15

n <- 10

k <- 3

nchoosek <- factorial(n-1)/(factorial(k-1)*(factorial(n-k)))

pserve1 <- nchoosek*(p^k)*((1-p)^(n-k))

pserve1
```

```
## [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?

For this we'll use the binomial distribution again:

```r
n <- 9

k <- 2

p <- 0.15

numscenarios <- factorial(n) / (factorial(k)*factorial(n - k))

pserve2 <- numscenarios*(p^k)*((1-p)^(n-k))

pserve2
```

```
## [1] 0.2596674
```

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 reason for this discrepancy is that in the first case, we're calculating the probability that the 10th serve is a hit, while two of the former nine are hits. In the second case, we're only calculating the probability that two of the former nine are hits. We can arrive at our first probability by multiplying the second by the probability of that 10th serve being a hit: 0.15:

```r
pserve2*0.15
```

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

```
pserve1
```

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

Now they match.