DATA 606 - Chapter 3 Homework

Adam Douglas

10/1/2018

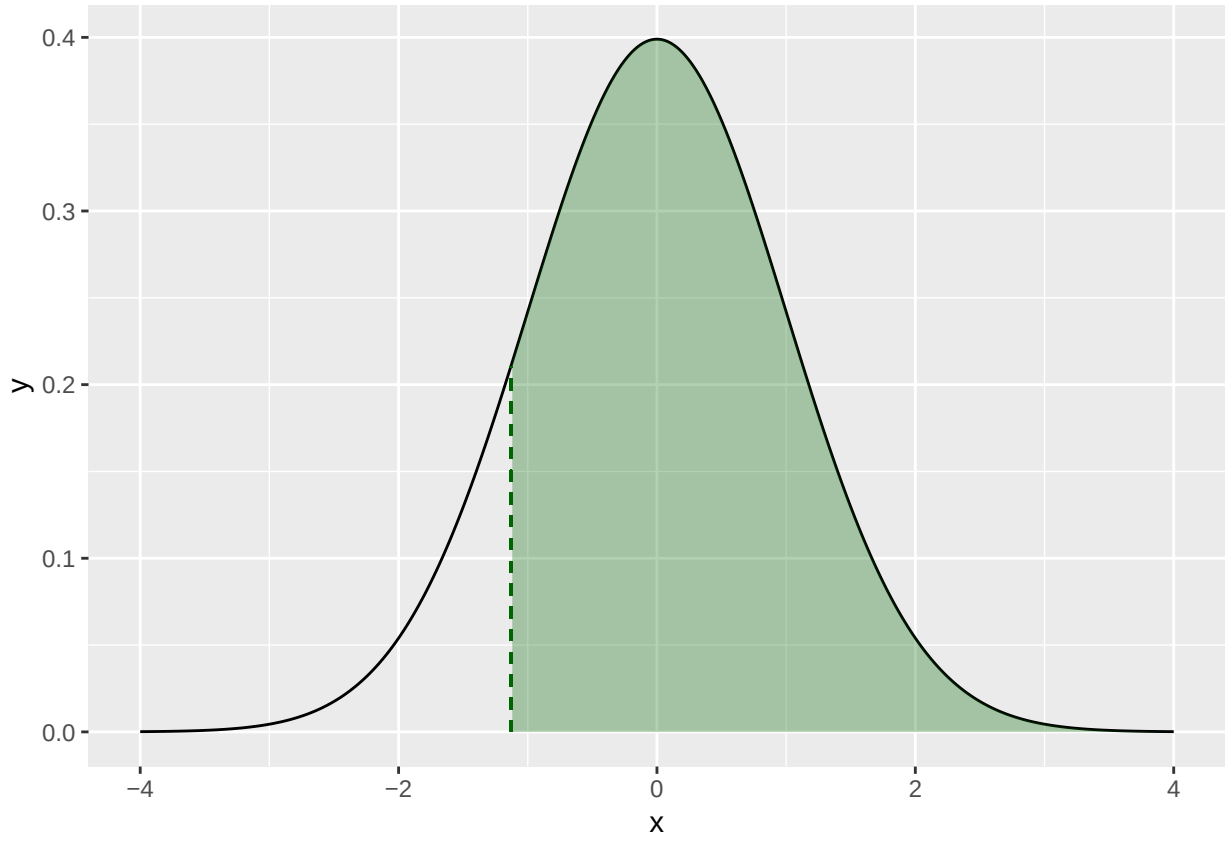
3.2

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

$Z > -1.13$ $Z < 0.18$ $Z > 8$ $|Z| < 0.5$

```
x = seq(-4,4,0.01)
y <- dnorm(x, 0, 1)
stdNorm <- data.frame(x,y)
```

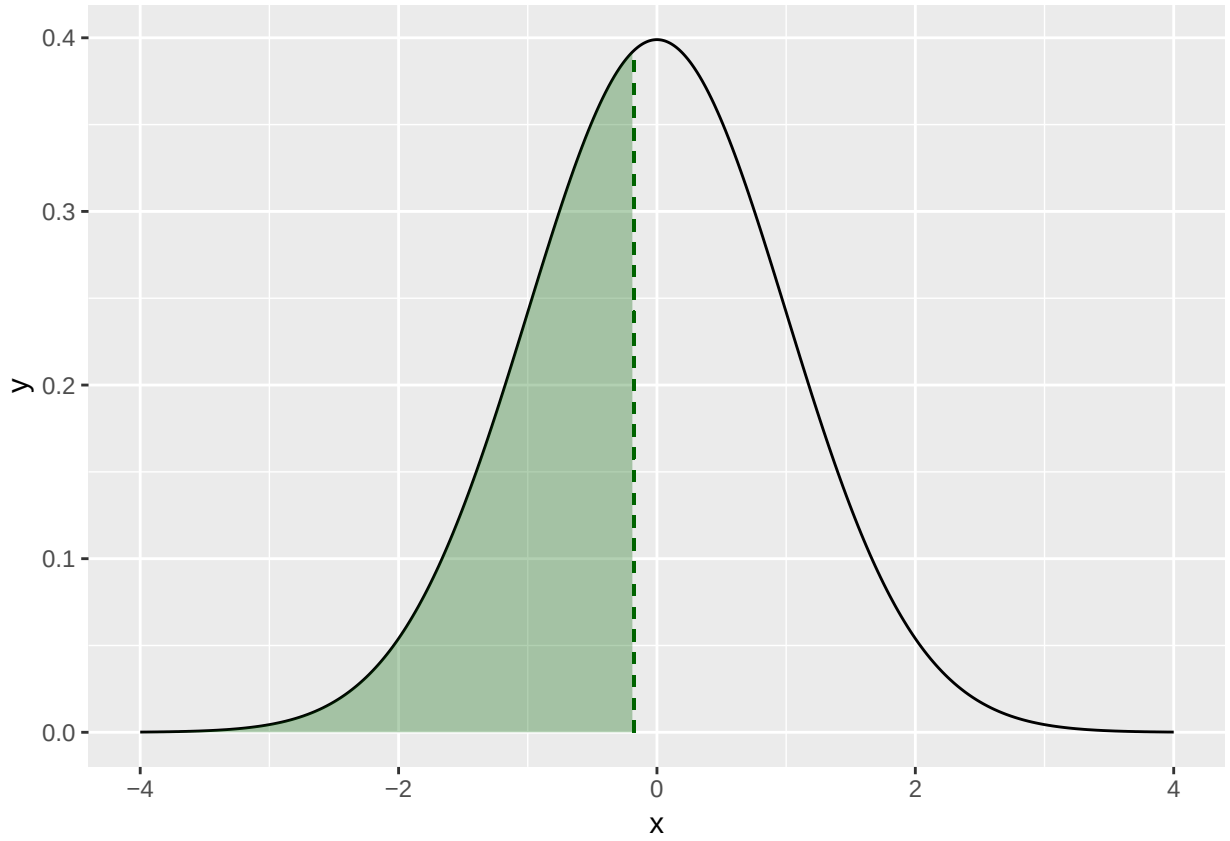
```
# Plot
stdNorm %>% ggplot(aes(x,y)) + geom_line() +
  geom_segment(aes(-1.13,0,xend=-1.13,yend=dnorm(-1.13,0,1)),
               linetype=2,col="darkgreen") +
  geom_ribbon(data=subset(stdNorm,x > -1.13),aes(x=x,ymax=y),
              ymin=0,fill="darkgreen", alpha=0.3)
```




```
# Answer
a <- round(100 * pnorm(-1.13,0,1,lower.tail=FALSE),2)
```

(a) 87.08% of the standard normal distribution falls < -1.13.

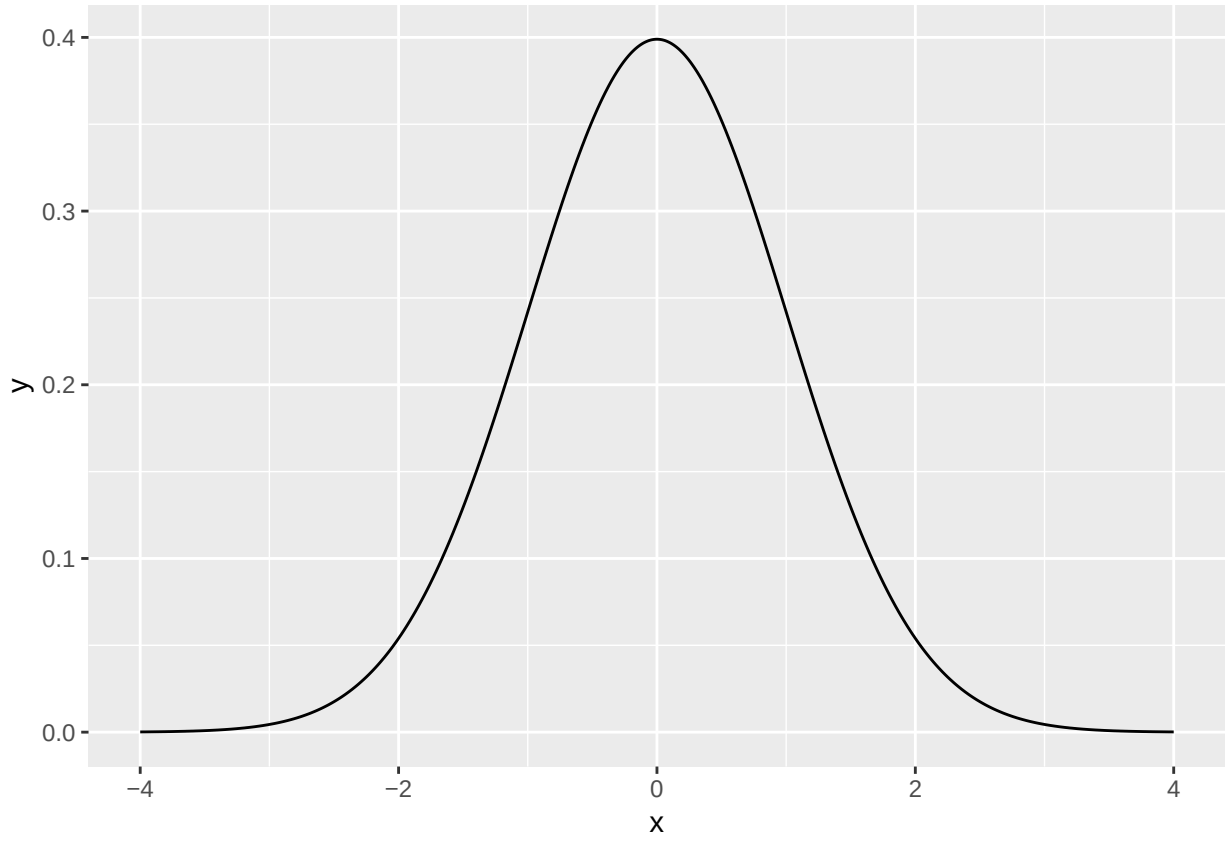
```r
# Plot
stdNorm %>% ggplot(aes(x,y)) + geom_line() +
  geom_segment(aes(-0.18,0,xend=-0.18,yend=dnorm(-0.18,0,1)),
               linetype=2,col="darkgreen") +
  geom_ribbon(data=subset(stdNorm,x < -0.18),aes(x=x,ymax=y),
              ymin=0,fill="darkgreen", alpha=0.3)
```




```r
# Answer
b <- round(100 * pnorm(-0.18,0,1,lower.tail=TRUE),2)
```

(b) 42.86% of the standard normal distribution lies < -0.18.

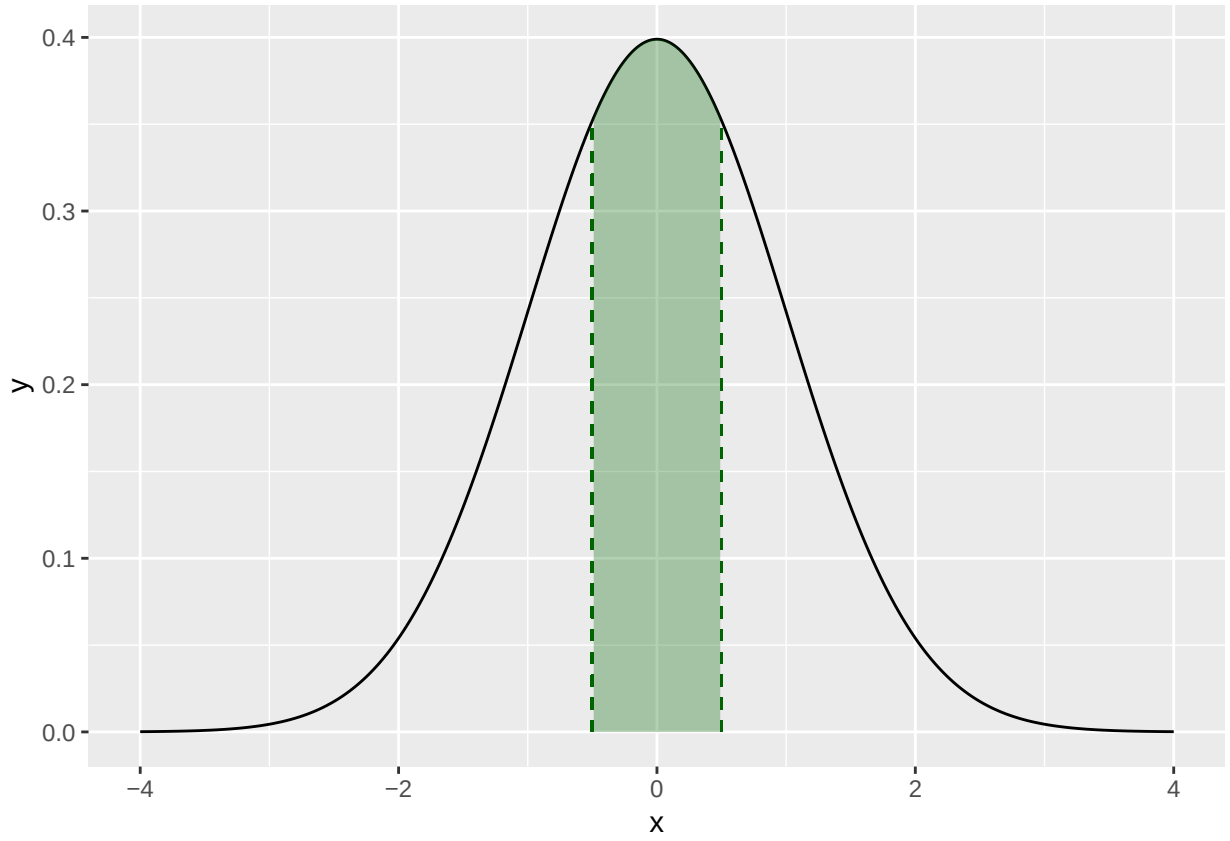
```r
# Plot
stdNorm %>% ggplot(aes(x,y)) + geom_line()
```




```
# Answer
c <- 100 * pnorm(8,0,1,lower.tail=FALSE)
```

(c) A very, very small percentage, 0.00000000000006220961%, lies to the right of 8.

```
# Plot
stdNorm %>% ggplot(aes(x,y)) + geom_line() +
  geom_segment(aes(-0.5,0,xend=-0.5,yend=dnorm(-0.5,0,1)),
               linetype=2,col="darkgreen") +
  geom_segment(aes(0.5,0,xend=0.5,yend=dnorm(0.5,0,1)),
               linetype=2,col="darkgreen") +
  geom_ribbon(data=subset(stdNorm,x > -0.5 & x < 0.5),aes(x=x,ymax=y),
              ymin=0,fill="darkgreen", alpha=0.3)
```




```
# Answer
d <- round(100 * pnorm(-0.5,0,1,lower.tail=TRUE) - pnorm(-0.5,0,1,lower.tail=TRUE),2)
```

(d) Approximately 30.55% of the standard normal distribution has an absolute value < 5.

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.

The distribution for the Men, Ages 30-34 group is $N(4313, 583)$. The distribution for the Women, Ages 25-29 group is $N(5261, 807)$.

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

```r
# Manual method
zLeo <- (4948 - 4313)/583

# Function
zMary <- scale(5513,5261,807)
```

Leo's Z-score is 1.09 and Mary's is 0.31.

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

While both athletes were above (slower than) the mean Mary did better, as she was closer to the mean than Leo was. This is because her Z-score was lower (less number of standard deviations).

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

This part is asking for the percentile of Leo's finish time. Since a higher number is a slower time, we need to reverse the typical calculation. In this case he finished in the 14 percentile.

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

Mary finished in the 38 percentile.

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

Yes. We couldn't do a Z-score comparison if the finish times were not nearly normal. We'd have to calculate, most likely, from all participant times.

3.18 - Heights of female college students.

Below are heights of 25 female college students.

```r
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)
```

(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.

```r
# The range of 1 standard deviation is:
lower <- 61.52 - 4.58
upper <- 61.52 + 4.58

# We should then see about 68% of the 25 entries (~17) within 1 standard deviation
length(heights[which(heights < upper & heights > lower)])
```

```
## [1] 17
```

```r
# Which we see
```

```r
# The range of 2 standard deviations is:
lower <- 61.52 - (2 * 4.58)
upper <- 61.52 + (2 * 4.58)

# We should then see about 95% of the 25 entries (~24) within 2 standard deviations
length(heights[which(heights < upper & heights > lower)])
```

```
## [1] 24
```

```r
# Which we also see
```

```r
# The range of 3 standard deviations is:
lower <- 61.52 - (3 * 4.58)
upper <- 61.52 + (3 * 4.58)

# We should then see about 99.7% of the 25 entries (~24-25) within 3 standard deviations
length(heights[which(heights < upper & heights > lower)])
```

```
## [1] 25
```

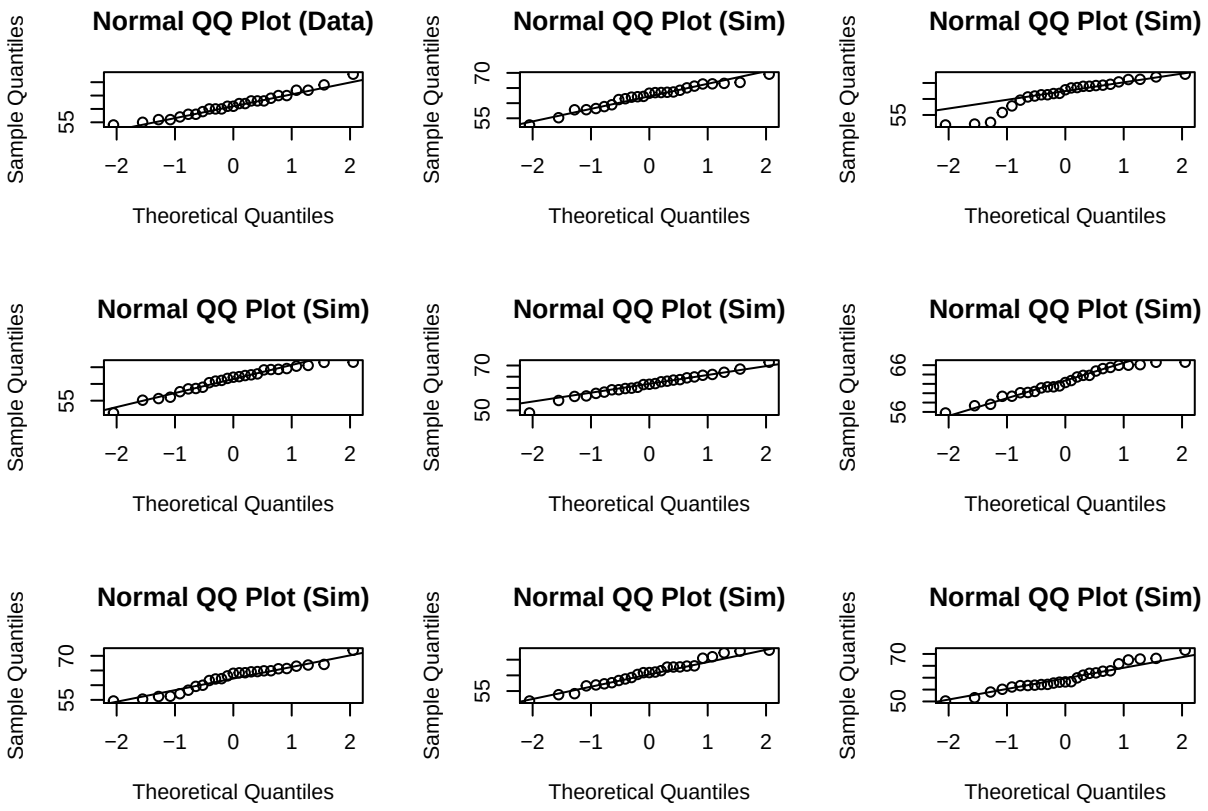
```r
# Which we more or less see
```

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

The data does appear to be nearly normal. The Q-Q plot shows that the data is close to the theoretical straight line (with perhaps a small skew). The histogram with the normal curve overlaid also shows that these data may be nearly normal.

We can also sample normal data with the same mean and standard deviation to compare:

```r
qqnormsim(heights)
```



This shows us that the heights have some similarity to the simulations of the normal distribution.

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?

This is an application of the geometric distribution. That defines the probability of the first success (here, a faulty transistor) on the 10th trial as:

$$(1-p)^9 p$$

```
# Manual calculation
(0.98^9)*0.02
```

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

```
# R function
dgeom(9,0.02)
```

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

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

This is quite simply $(1-p)^n$ or 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?

This is the same as the mean of the geometric distribution, or $\frac{1}{p}$, which is 50. That makes intuitive sense too, as 0.02 is 1 in 50.

(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'd expect the first defect to show up around the 20th transitor. The standard deviation would be $\sqrt{\frac{1-p}{p^2}}$, or 19.49.

(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?

As probability increases, the mean decreases, as does the standard deviation.

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.

The probability of having 2 boys and 1 girl is 0.382.

(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 potential ways to have 2 boys and 1 girl out of 3 children: (B,B,G), (B,G,B), and (G,B,B). Each of these scenarios has $(0.51)^2 * 0.49 = 0.1274$. We then add those scenario probabilities: $0.1274+0.1274+0.1274 = 0.3822$. This matches what we got above.

(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 method used in (b) above would be considerably more tedious because we'd have to count all the possible combinations of 3 boys amongst 8 kids. Then, we'd have to add all those probabilities.

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 is an application of the negative binomial distribution. We want the 3rd success on the 10th trial, which is $\binom{9}{2}(0.15^3)(0.85^7)$ or 0.039

(b) Suppose she has made two successful serves in nine attempts. What is the probability that her 10th serve will be successful?

The same as her first, 0.15. Because each serve is an independent event the probability is the same each time, regardless of what may have happened on previous serves.

(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 event is independent, so in (a) we're adding probabilities of various disjoint independent events. In (b) it is simply the one independent event, the 10th serve.
