Chapter 4 - Distributions of Random Variables

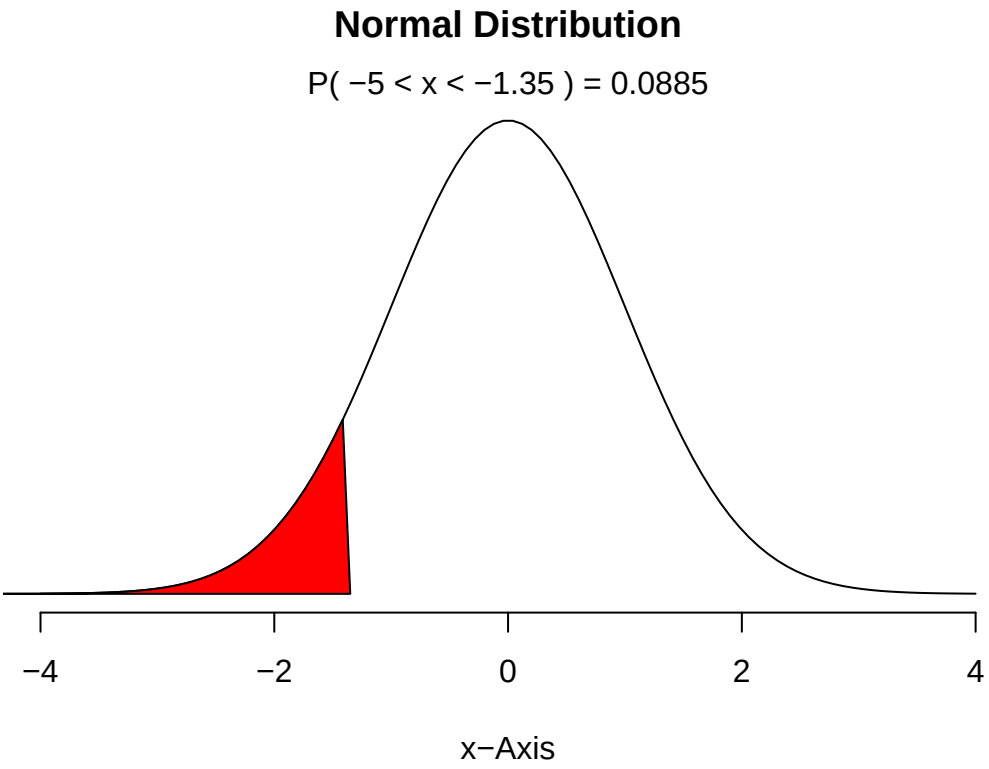
Area under the curve, Part I. (4.1, p. 142) What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph.

(a) $Z < -1.35$

```
auc_1<-pnorm(-1.35)*100
print(paste(auc_1,"%"))
```

```
## [1] "8.8507991437402 %"
```

```
DATA606::normalPlot(bounds = c(-5,-1.35))
```



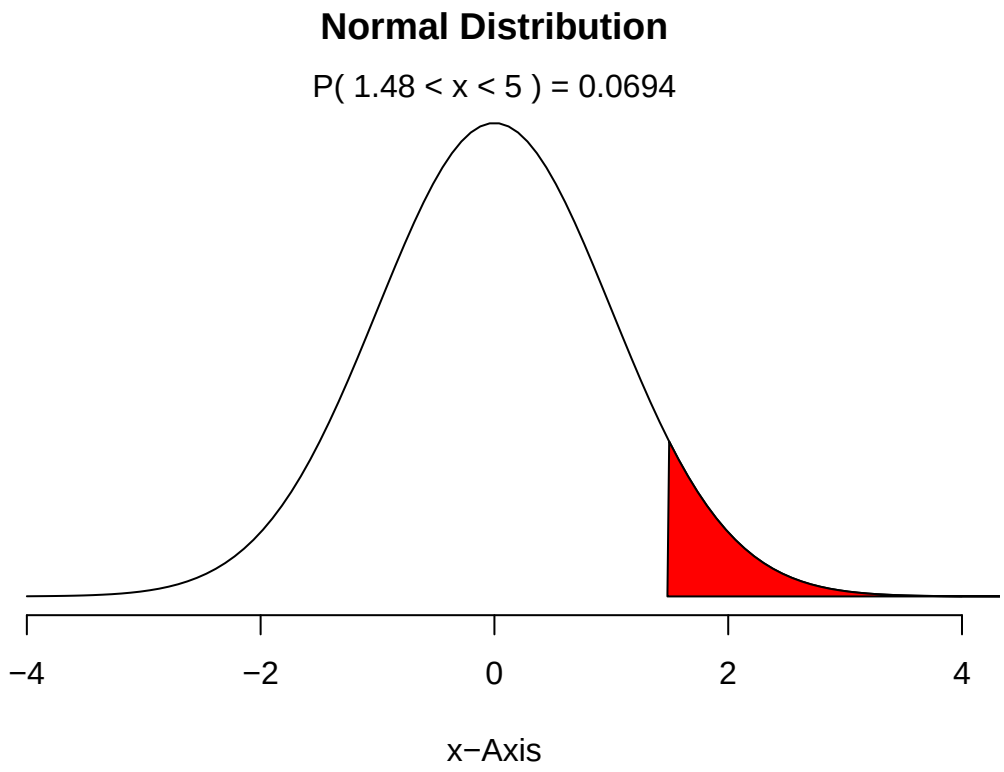
This is the amount under the curve to the left of Z-score = -1.35

(b) $Z > 1.48$

```
auc_2<- (1-pnorm(1.48))*100
print(paste(auc_2,"%"))
```

```
## [1] "6.94366233333318 %"
```

```
DATA606::normalPlot(bounds = c(1.48,5))
```



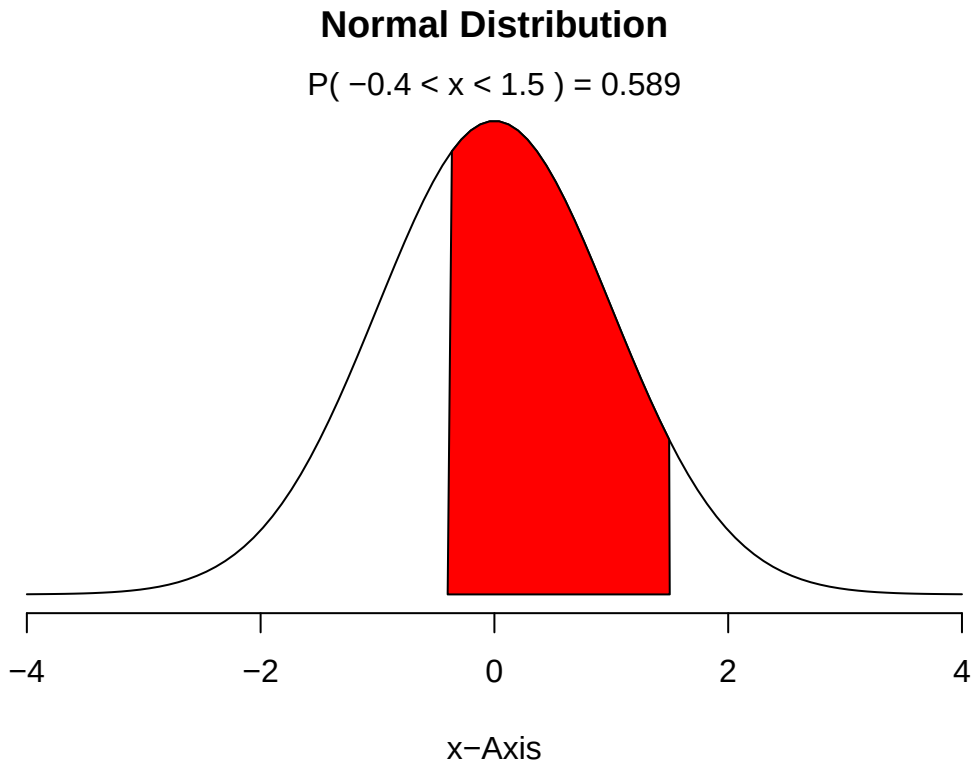
(c) $-0.4 < Z < 1.5$

This is the difference between everything right of Z-score = -0.4 and everything left of Z-score = 1.5

```
auc_31<-(1-pnorm(-0.4))*100 # right of -0.4
auc_32<-pnorm(1.5)*100 #left of 1.5
print(paste((auc_32-auc_31),"%"))
```

```
## [1] "27.7771057120818 %"
```

```
DATA606::normalPlot(bounds=c(-0.4,1.5))
```



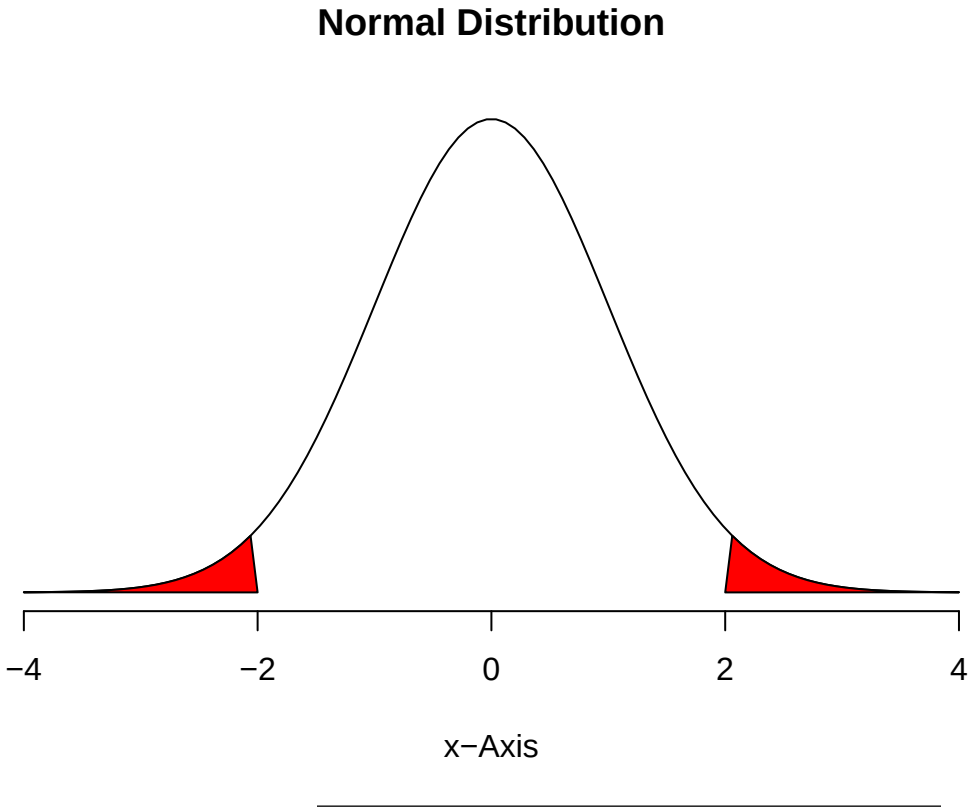
(d) $|Z| > 2$

This is the difference between Z-score < -2 and < 2. This is outside the classic 2 std swing and should capture about 5% of the data.

```
auc_41<-(pnorm(-2))*100 # right of -2
auc_42<-(1-pnorm(2))*100 #left of 2
print(paste((auc_41+auc_42),"%"))
```

```
## [1] "4.55002638963584 %"
```

```
DATA606::normalPlot(bounds = c(-2,2), tails = T)
```


Normal Distribution
−4
−2
0
2
4
x−Axis

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.

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

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

The Z-score are the amount of distance from the mean each time relates to.

For Leo: $Z = \frac{4948-4313}{583} = 1.089$

Mary: $Z = \frac{5513-5261}{807} = 0.3122$

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

In this case, Mary is much closer to the mean finish for her subgroup than Leo. So Leo seems to have had a better finish in his subgroup compared to Mary. Leo was +1 std from the mean finisher in his group, where as Mary was only +0.3 away from the mean finisher in her group. In this regard, Leo has a more rare finish time and Mary.

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

```
1- pnorm(4948, mean = 4313, sd = 583)
```

```
## [1] 0.1380342
```

86.1% percentile so 13.8 percent are faster (in group) than Leo. Nice, Leo

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

```
1- pnorm(0.3122)
```

```
## [1] 0.3774443
```

37.7% of the finishers were faster than Mary (in 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 it isn't normal of nearly there, we would need to find another way to determine mean and std so as to reconcile how far observations are from that mean. And how likely it is that the variance is explainable.

Heights of female college students Below are heights of 25 female college students.

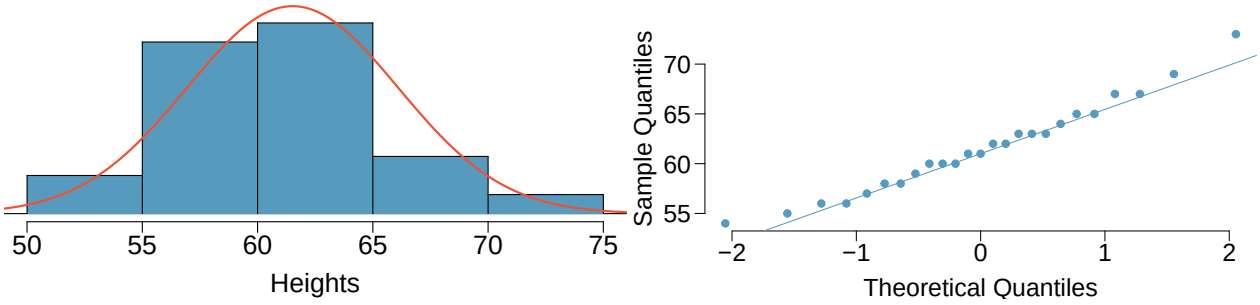
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
54,	55,	56,	56,	57,	58,	58,	59,	60,	60,	60,	61,	61,	62,	62,	63,	63,	63,	64,	65,	65,	67,	67,	69,	73

(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.

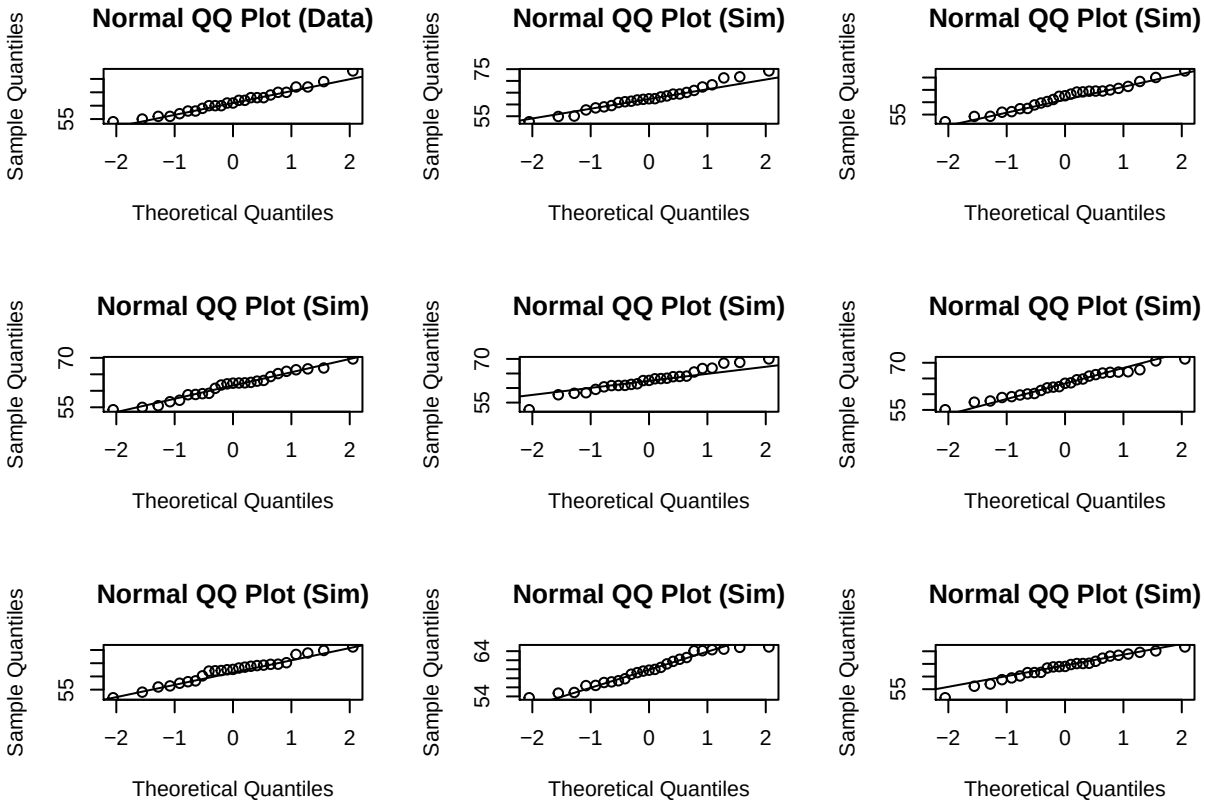
One sd above the mean would be 66.1 and there are only 4 more observations beyond that. The same is true for one sd below the mean. 4/25 is 16%, so the total amount beyond $\pm$ 4.58 is 32%. Thus, between one sd on either side of the mean, we have 68% of the data. That a good sign for calling this data normal. Similarly, at two sd beyond the mean we have only 1/25 observations outside of that range, which is 4%. Thus, this following the 68-95 rule.

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

The qq plot looks pretty good and the data seem to follow the 68-95 rule. Also the histogram appears to be normal enough with boxes both outside of the curve and below the same curve. Looking at the qqnormsim we see that a random sample of normal data looks very much in line with the qq plot we are given.




```
# Use the DATA606::qqnormsim function
DATA606::qqnormsim(heights)
```


Normal QQ Plot (Data)
Normal QQ Plot (Sim)
Normal QQ Plot (Sim)
Normal QQ Plot (Sim)
Normal QQ Plot (Sim)
Normal QQ Plot (Sim)
Normal QQ Plot (Sim)
Normal QQ Plot (Sim)
Normal QQ Plot (Sim)
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2
55
75
70
64
54

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?

Geometric probability. $p * (1-p)^{n-1}$

$0.02 * (1-0.02)^9 = 0.0166$ or 1.66%

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

If we take geometric probability to look for the first success in 101 trials then we can get the probability of 0 defects in 100 trials. $0.02 * (1-0.02)^{100} = 0.00266$ or 0.266%

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

The mean for geometric probability is $\mu = \frac{1}{p}$ so $\mu = \frac{1}{0.02} = 50$. The standard deviation is $\sigma = \sqrt{(\frac{1-p}{p^2})}$, which is $\sigma = \sqrt{(\frac{1-0.02}{0.02^2})} = 49.49$. Given these equations, the standard deviation with always be very close in value to the mean. That kind of makes me think that the mean and sd are not so meaningful. A plot of those equations is below, very confused as to how to interpret this.

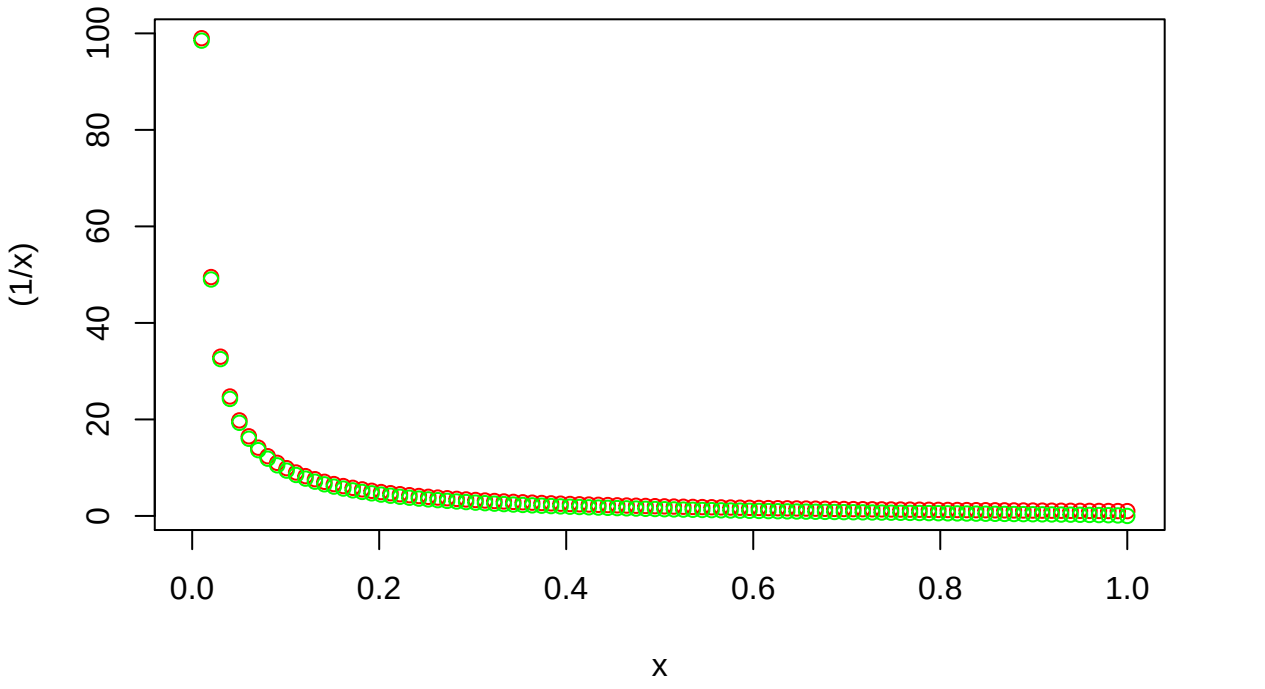
(d) Another machine that also produces transistors has a 5% defective rate where each transistor is produced independent of the others. On average how many transistors would you expect to be produced with this machine before the first with a defect? What is the standard deviation?

Same as above, different numbers. $\mu = \frac{1}{p} = 20$ transistors before a defect. $\sigma = \sqrt{(\frac{1-0.05}{0.05^2})} = 19.43$

(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?

It is... geometric. y ~ 1/x

```
x<- seq(0,1,length.out = 100)
plot(x, (1/x), col="red")
points(x, sqrt((1-x)/x^2), col="green")
```



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.

$\binom{3}{2} * (0.51)^2 * (0.49)^1 = 3 * 0.2601 * 0.49 = 0.382$

(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.

BBG: $0.52^2 * 0.49$ + BGB: $0.52^2 * 0.49$ + GBB: $0.52^2 * 0.49 = 3 * 0.52^2 * 0.49 = 0.382$

(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).

Writing out all 56 combinations of getting exactly 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?

Negative binomial. $\binom{n-1}{k-1} * (p)^k * (1-p)^{n-k} = 84 * (.15)^3 * (.85)^7 = 0.09$ or 9%

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

It is the same probability of completing any successful serve, i.e. 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 events are independent. for part (b) we are talking about a singular, independent event. We are given the probability for this. In part (a) we are to find the probability of successful serves in a given number of the attempts. This requires considering all the possible ways to have three successful serves that terminate on the 10th attempt.