Applying Probability to Georgia Tech Volleyball

ISYE 2027 Project

Mike Wiggins and Nathanael Williams

15

Purpose

During the Fall 2015 semester, we started by learning very basic set theory and then moved onto modeling distributions of data and calculating complex probabilities. This project applies the major concepts we have learned to a context that we found interesting, the Georgia Tech Women's Volleyball team. The game Georgia Tech played against Clemson University is the main focus of this study on basic set theory, discrete and continuous random variables, and limit theories.

Table of Contents

Section 1.1: Motivation

Why is this topic interesting at large?

As is the case with any sport, the coaches and analysts for volleyball rely heavily on statistics. The women's volleyball team at the Georgia Institute of Technology employs a full time video analyst for the purpose of collecting and evaluating information pertaining to certain aspects of the team and their games. At the time of the game from which we will be recording data, Georgia Tech was ranked significantly higher than their opponent (Clemson), so this analysis will be an interesting representation of how Georgia Tech performs when theoretically they are at their best or feel very confident.

Why is this topic interesting to us personally?

Growing up, we played volleyball at our churches and at various clubs in our hometowns. We are very familiar with how volleyball is played at a recreational level, but not as familiar with the collegiate rules and regulations. It will be very interesting to analyze a known topic in a completely different context than we are used to. While we are attending the Georgia Institute of Technology, we plan to participate in many intramural sports and attend as many varsity-sporting events as possible. This analysis will allow us to be more engaged participants and provide us with a new lens with which we can view our women's volleyball team's games and the recreational volleyball activities around campus.

Why is this context good for modeling probability distributions?

A great deal of the information gathered can be easily modeled by fundamental probability distributions. The motivation of this project is to combine some basic probability distributions with empirical data from a particular game in order to come up with a few interesting assertions. We are going to model several interesting aspects including how many serves Georgia Tech wins before they lose a serve, how many points Georgia Tech typically scores in the first twenty minutes, how many points they score in the first 10 serves, and whether or not the player who is serving has an impact on whether Georgia Tech is able to accomplish a kill.

Section 1.2: Context Description

We will discuss the applications of probability to a volleyball match. Here is some useful information about how the game is played, to help you better understand some of the concepts that we will discuss.

Basic setup of a volleyball match:

Volleyball is a sport that is played on a rectangular shaped court (See Figure 1). The sport is played using a volleyball (See Figure 2). In a particular volleyball match, there are two teams that are competing; each team stands on a different side of the net. During a match, each team has six players on the court; one of the players is allowed to be a libero, a strictly defensive player. The libero's jersey is a different color than the jerseys of the rest of the members of the team. (Moore, T., "The Volleyball Court ")

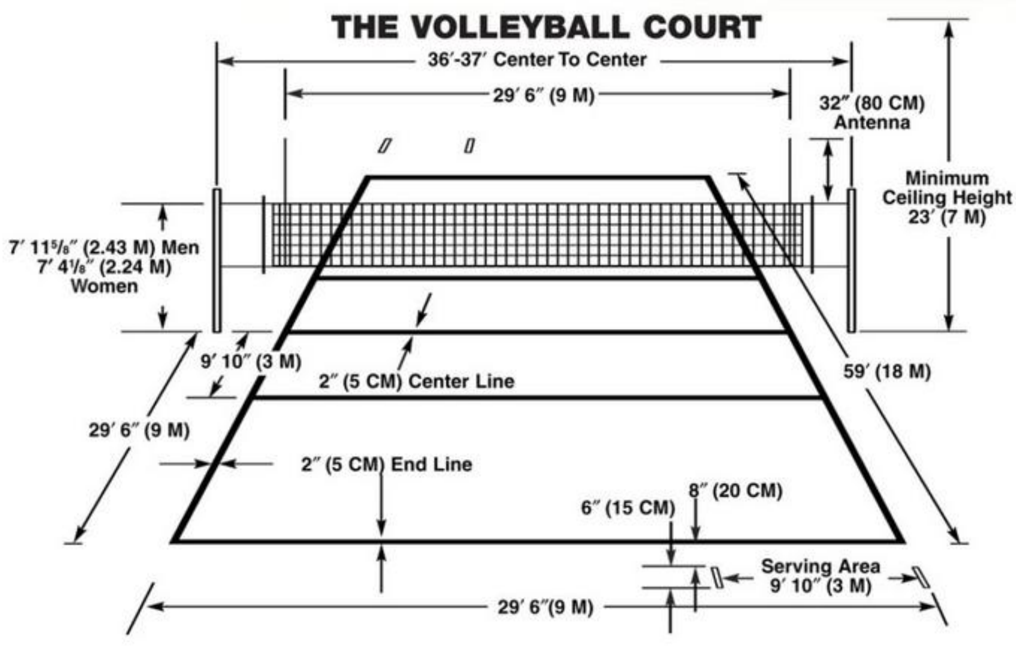


Figure 1: Diagram of a Volleyball Court

Figure 2: Picture of a Volleyball
(Nothstein, J., "The Volleyball")

Basic rules of the gameplay a volleyball match:

The court is split into two halves by the net; each team is responsible for protecting its own half of the court. The area inside the giant rectangle is considered to be in bounds. Volleyball is separated into a series of rallies; each rally is initiated when a server hits the ball, putting it into play. During a rally, each team is allowed to hit the ball no more than three times before sending it back over the net. During a rally, the same player is not allowed to make two consecutive hits between times that the ball is sent over the net. At the end of each rally, one of the teams is awarded a point based on the following conditions:

- A team wins the point if the ball hits the ground in play on the other team's side of the net.
- A team wins the point if the other team hits the ball out of play.
- A team wins the point if the other team illegally touches the ball. An illegal hit could be caused by the same player hitting the ball twice in a row or by a team hitting the ball too many times before sending it over the net.

The team that won the most recent point is responsible for serving to start the next rally. A kill happens when a player on one of the teams legally hits the ball over the net in play in such a way that the other team in unable to return it; this results in the team that hit the ball being rewarded the point.

Scoring in a volleyball match:

A volleyball match is separated into five sets. In each set, the teams play until one of the teams has scored 25 points in that set; however, a team must win a set by at least two points, so 25-24 is not a possible score to end a set. In a set, once the score gets to 24-24, the set ends as soon as one of the teams is ahead by two points. The fifth set is an exception to this rule: the teams play until one team has scored 15 points in the fifth set. The rule about winning a set by two points also applies to the fifth set. The first team to win three sets wins the match; if one of the teams has one three sets before the end of the fifth set, the remaining sets are not played.

Section 2.1: Experiments, Sample Spaces, Events

a. Georgia Tech Starting Strong

Immediate success after the match begins is sometimes referred to as “starting strong.” If Georgia Tech wins many serves soon after the match begins, they are in a good position to win the match.

Experiment 1: Monitor the first 20 serves by either team for the first set. Record the serve numbers of the serves that Georgia Tech won.

The sample space is the serve numbers of the serves that Georgia Tech won from the first 20 serves in the match, regardless of the serving team.

Sample Space = Power set of {1, 2, 3, … , 20}
The power set is this huge set of subsets: {{1}, {2}, {3}, … , {1, 2}, {1, 3}, {1, 4}, … , {1, 2, 3, 4, 5, 6, 7, 8, 9, 10 ,11, 12, 13, 14, 15, 16, 17, 18, 19, 20}}

This is an example of a discrete-finite sample space.

For the first set in the match, here is the set of data:
Serve numbers that Georgia Tech won: A = {2, 3, 4, 5, 6, 10, 11, 13, 14, 16, 19}
(McColskey, T., "Georgia Tech vs. Clemson Women’s Volleyball Game Film")

We can use the cardinality (number of elements in a set) of the set and values of the numbers to determine if Georgia Tech started off the match strong. We can use the size of the set to determine how many out of the first 20 serves Georgia Tech wins. Also, the values in the set tell us which particular serves Georgia Tech won. Since Georgia Tech won 5 of the first 6 serves and then only won 6 out of the next 14 serves, we evaluate that as a strong start that seems to fall back to average. We can also evaluate the beginning of a match as being a strong start across 20 serves if Georgia Tech wins 15 out of the first 20 serves.

b. Starting Off on a Bad Streak

For a team that is not able to start strong, there is a possibility that that team is starting off the match by losing many consecutive serves before winning any of the serves. This is referred to as a “bad streak.”

Experiment 2: Count the number of serves until Georgia Tech wins a point. Record the result.

The sample space includes all possible values for the number of serves in the match until Georgia Tech wins a point.

Sample Space = {1, 2, 3, …}

Since a match will end if the score is {25-0, 25-0, 25-0}, the maximum value for the sample space for one match is {75}. However, if we look at this example as the number of serves from the beginning of the Georgia Tech vs. Clemson match until Georgia Tech scores, the sample space is infinite (assuming Georgia Tech will continue to play matches until they score).

This is an example of a discrete-infinite sample space.

For the first set in the match, here is the set of data:
Serve number of the first serve that Georgia Tech won: B = {2}
(McColskey, T., "Georgia Tech vs. Clemson Women’s Volleyball Game Film")

Unless the value that is the result of this experiment is a really large number, a single iteration of this experiment is not particularly useful. We could apply this experiment to many different matches to see if Georgia Tech often starts off on a bad streak. If the result is a small number, then Georgia Tech scored soon after the match started. This result of this experiment does not necessarily tell us much information about how the match will end. If the result is 10 or larger, then Georgia Tech is off to a really bad start and will probably struggle in the match unless they make some major adjustments in their strategy. However, if the result is smaller than 4, then the match is just beginning and Georgia Tech still has a good chance of winning. This experiment is particularly useful in identifying catastrophically bad starts to matches.

c. Missing the Action

During a volleyball match, the time between consecutive scores can give the spectators an idea of how long the match will last. This will also give fans some idea how much of the action they will miss if they take a break to get something to eat from the concession stand.

Experiment 3: Monitor the time between consecutive scores. Record the result to the nearest second.

The sample space includes all possible times between two consecutive scores.

Sample Space = $\{t: t \in (0, \infty)\}$

Technically, there is a lower bound for this sample space: there is some amount of time that must elapse between serves: the referees and players need to reset their positions, the server needs to have the ball, etc.

This is an example of a continuous sample space

For the first set in the match, here is the set of data:
Time between the first two serves in seconds: C = {31}
(McColskey, T., "Georgia Tech vs. Clemson Women's Volleyball Game Film")

This is another experiment for which a single trial will not be very useful. Performing this experiment many times and then averaging the time between serves will provide much more useful information. The time between the first two serves in the match could be considerably longer or considerably shorter than the average time between serves. An average of many trials of this experiment can give fans an estimation of how many points will be scored during their trips to the restroom and/or concession stand. It will also give them some idea of how much time to allot if they plan to watch the entire game.

Section 2.2: Probabilities of Events

a. Likelihood of Early Success

To calculate the probability that Georgia Tech will win the exact serve numbers that they did in Set 1, we need to determine the probability of winning a particular point in the set. The probability that Georgia Tech loses a given point in the set is equal to 1 minus the probability that they win that point. Since Georgia Tech's volleyball team is better than Clemson's, we anticipate that the probability that Georgia Tech wins a particular point in a set is slightly better than 0.50. We also think that any particular result of the first 20 serves is extremely unlikely, but the probability of doing well is quite high.

Using the number of serves and the outcomes of the serves, we can estimate the probability that the first set of a volleyball match will turn out exactly the way that it did in this match. The probability that Georgia Tech wins a particular point can be estimated by dividing the total number of points that they won in the match by the total number of points that either team won in the match. [1]

$$p = \frac{75}{131} \approx 0.57$$

The probability that Georgia Tech loses a particular point in the match is 1 – p.

$$1 - p = 1 - 0.57 = 0.43$$

The probability that Georgia Tech wins the exact serve numbers that they did in Set 1 is found by multiplying the probability of each point turning out the way that it did.

$$P\{Set\ 1\} = (1-p) * p * p * \dots * (1-p) = p^{11}(1-p)^{9} = 1.03\times10^{-6}$$

Although the probability of this particular outcome of the set is very small, the probability that Georgia Tech wins at least 11 out of the first 20 serves in the first set is quite high.

We anticipated that Georgia Tech's probability of winning a certain point was slightly better than 0.50, and it was 0.57. We also thought that the exact outcome of the first 20 serves of the first set was extremely unlikely to happen.

[1] Data gathered from McColskey, T., "Georgia Tech vs. Clemson Women's Volleyball Game Film"

b. Likelihood of Extended Failure

To calculate the probability that the first serve that Georgia Tech wins in the second serve, we will use the probability of winning a particular point in the set that we estimated in part a. We will also use the probability that Georgia Tech loses a given point in the set that we estimated in part a. Since Georgia Tech wins more points than they lose, we anticipate that Georgia Tech will win about every other serve. We think that the probability that Georgia Tech loses the first serve, and then wins the second serve is about 0.25 (0.50 * 0.50).

Using the probabilities from part a, we can estimate the probability that Georgia Tech loses the first serve, and then wins the second serve.

$$p(1-p) = 0.57 * 0.43 = 0.24$$

We anticipated that Georgia Tech's probability of winning a certain point was about 0.25, and it turned out to be 0.24. Our estimation is a very reasonable guess for the probability of this particular outcome.

c. Likelihood of Missing Most of the Game

To calculate the probability that there will be 31 seconds between the first two serves, we need to determine some way to estimate the time between two serves in a match. We think that on average the time between consecutive serves is about 25 to 30 seconds. Therefore, we think that the probability of 31 seconds elapsing between the first two serves is very likely.

Using the number of serves in the first set and the time of the first set, we can estimate the time between consecutive serves. We will divide the length of the game by the number of serves to determine the average time between serves. This will not be a perfect estimate because there are timeouts in the first set that will make this estimate a little bit too large. [2]

$$t = \frac{1{,}119}{42} \approx 26.6$$

To account for the set breaks, we will round this down to 26 seconds between serves that do not involve timeouts.

From this, we will estimate that 50% of the time, the time between consecutive serves is between 24 and 28 seconds, and that about 20% of the time, the time between consecutive serves is between 29 and 31 seconds. We will estimate that the probability that the time between the first two serves is 31 seconds is about 0.06.

Our prediction that the time between rallies was between 25 and 30 seconds was not very precise. As a result, our prediction that 31 seconds between consecutive serves was likely was also wrong. The real probability of the time between consecutive serves (excluding those with timeout breaks) being 31 seconds is quite small.

[2] Data gathered from McColskey, T., "Georgia Tech vs. Clemson Women’s Volleyball Game Film"

Section 2.3: Probability Relationships

a. Losing Serve

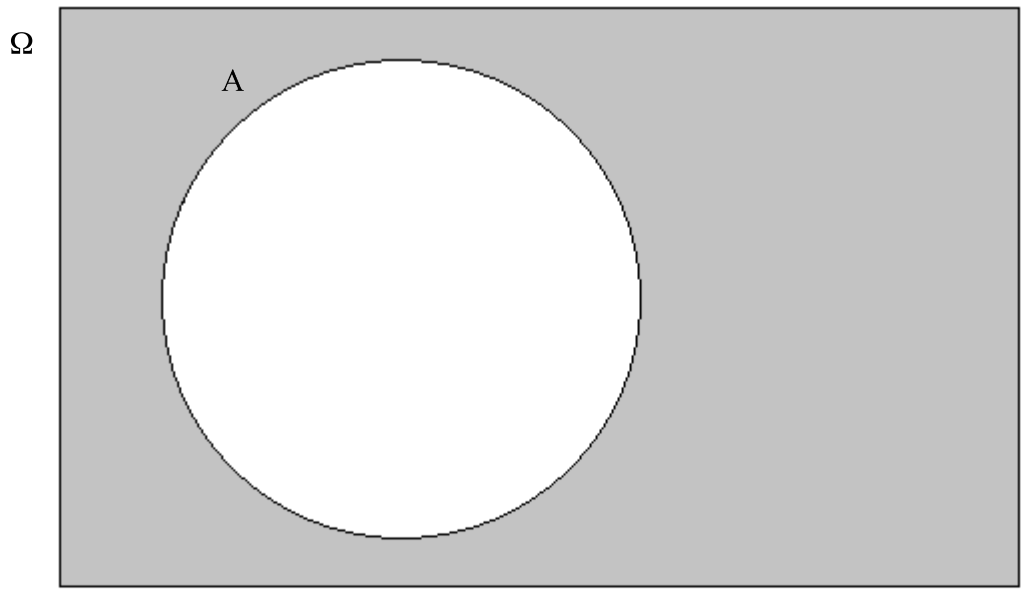


Figure 3: Complement of A

The symbol Ω is used to signify the universe, meaning all of the elements that exist. The complement of a set is defined as the set of all elements not in that set. To illustrate this concept, we will define set A as the serve numbers that Georgia Tech won in the first 20 serves of the match.

The universe is all of the elements that exist; in this case, Ω is all of the serve numbers 1 – 20.
$\Omega = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$
Serves 1 – 20 that Georgia Tech won in Set 1: $A = \{2, 3, 4, 5, 6, 10, 11, 13, 14, 16, 19\}$
(McColskey, T., "Georgia Tech vs. Clemson Women's Volleyball Game Film")

In Figure 3, Set A is represented by the white circle, and Ω is represented by the rectangle.

The complement of A includes all of the elements that are in Ω, but not in A. These are the serve numbers that Georgia Tech lost in the first 20 serves of the match. The notation for the complement of A is A^c.

$A^c = \{1, 7, 8, 9, 12, 15, 17, 18, 20\}$
The complement of A is the shaded region in Figure 3.

The sum of these two sets is the universe: $A + A^c = \Omega$. Set addition is combining all of the elements in the summed sets into a single set.

b. Winning Every Serve

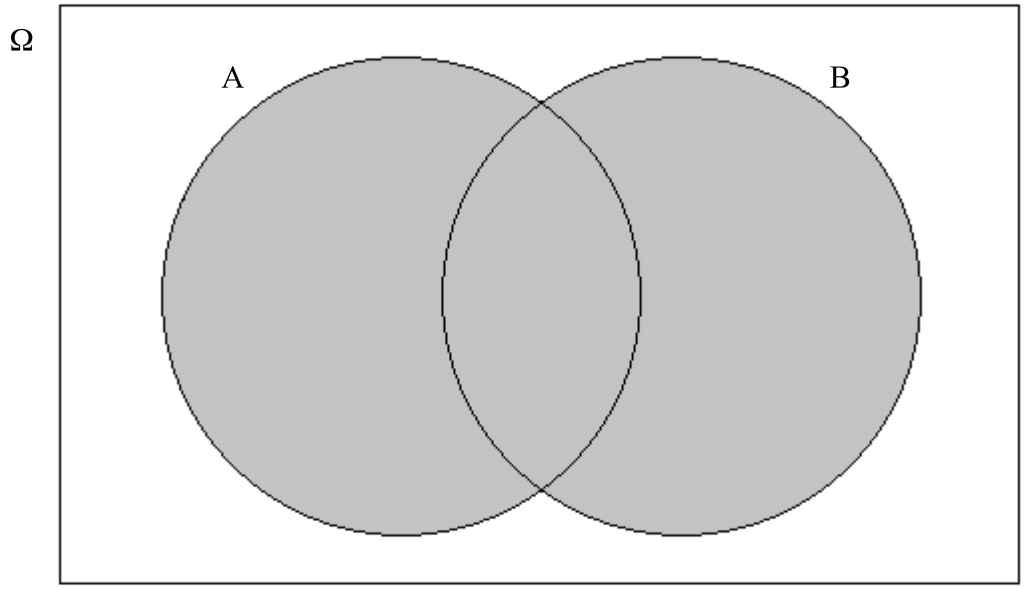


Figure 4: Union of A and B

The union of two sets is defined as the set of all elements that are in either set, or both. To illustrate this concept, we will define set A as the serve numbers that Georgia Tech won in the first 20 serves of Set 1 and B as the serve numbers that Georgia Tech won in the first 20 serves of Set 2.

Serves 1 – 20 that Georgia Tech won in Set 1: A = {2, 3, 4, 5, 6, 10, 11, 13, 14, 16, 19}
Serves 1 – 20 that Georgia Tech won in Set 2: B = {2, 3, 5, 7, 9, 12, 14, 17, 20}
(McColskey, T., "Georgia Tech vs. Clemson Women's Volleyball Game Film")

For this set of data, the union of A and B includes all of the serve numbers 1 – 20 that Georgia Tech won in either Set 1, Set 2, or both. The notation for the union of A and B is $A \cup B$.

$A \cup B$ = {2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 16, 17, 19, 20}
The union of A and B is the shaded region in Figure 4.

If the union of A and B had included all of the numbers 1 – 20, then Georgia Tech would have won every possible serve number 1 – 20 in at least one of the first two sets. If Georgia Tech is having success in a match, $A \cup B$ will contain most, if not all of the numbers 1 – 20. The numbers missing from A ∪ B are serves that Georgia Tech did not win in either set. Interpreting the missing numbers can tell us if Georgia Tech had certain parts of both sets in which they lost the same serve. In this match, Georgia Tech won most of the different serve numbers, and therefore seemed equally strong throughout the first 20 serves of both Set 1 and Set 2.

c. Winning the Same Serve

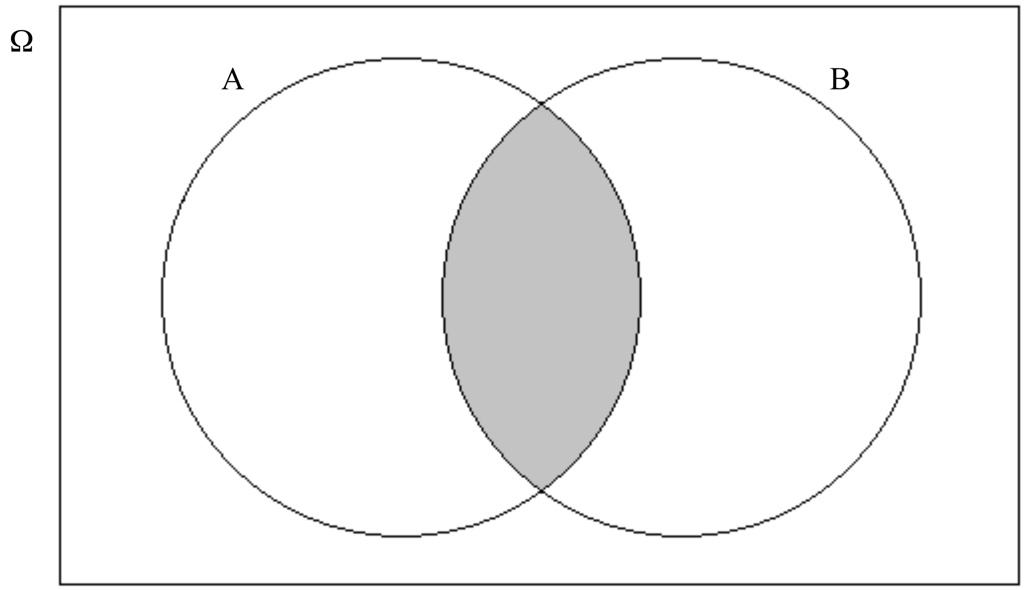


Figure 5: Intersection of A and B

The intersection of two sets is defined as the set of all elements that are in both sets. To illustrate this concept, we will use the same sets A and B that we defined in part b of Section 2.3.

Serves 1 – 20 that Georgia Tech won in Set 1: A = {2, 3, 4, 5, 6, 10, 11, 13, 14, 16, 19}
Serves 1 – 20 that Georgia Tech won in Set 2: B = {2, 3, 5, 7, 9, 12, 14, 17, 20}
(McColskey, T., "Georgia Tech vs. Clemson Women's Volleyball Game Film")

For this set of data, the intersection of A and B includes all of the serve numbers 1 – 20 that Georgia Tech won in both Set 1 and Set 2. The notation for the intersection of A and B is A ∩ B.

A ∩ B = {2, 3, 5, 14}
The intersection of A and B is the shaded region in Figure 5.

If the intersection of A and B included all of the numbers 1 – 20, then Georgia Tech would have won every possible serve number 1 – 20 in both of the first two sets. This would be a very dominant start to both sets. In this match, Georgia Tech did not win very many of the same serve number from the first 20 serves in the first two sets. This means that Georgia Tech was not particularly dominant early in both of the first two sets. The fact that the cardinality of $A \cup B$ is 16 and the cardinality of $A \cap B$ is 4 seems to indicate that the two teams are evenly matched.

Section 3.1: Discrete Random Variables

a. Winning Streak

Question: How many points did Georgia Tech score before Clemson scored a point?

The probability that Georgia Tech loses a particular serve is p; there are only two possible outcomes of each serve: success (Clemson scores) and failure (Georgia Tech scores). Although it can be argued that a particular team is capable of having momentum, or is more likely to score again given that they have scored several times already, we think that, for the most part, it is a reasonable assumption to say that each serve is independent of the others. We will let the random variable X_1 denote the number of serves until Clemson scores a point. This random variable maps a geometric distribution with a parameter of probability p. The expected value of X_1 will be the number of serves that Georgia Tech typically wins in a row. The winning streak will be one serve plus the expected value of X_1, since it needs to include the serve that Georgia Tech initially won when Clemson was serving.

Let X_1 = the number of serves by Georgia Tech until Clemson's first point
$X_1 \sim$ Geo(p)
p = the probability Clemson scores on any serve by Georgia Tech
E[X_1] = the expected number of serves by Georgia Tech until Clemson scores

Analogy of Geometric Distribution:

Analogy	Volleyball Match
Flip a coin	Service by Georgia Tech
Until first head	Until Clemson Scores
Result is head	Result is Clemson Scores
Result is tails	Result is Georgia Tech Scores
Probability of head	Probability Clemson Scores

Parameter p is the probability that Clemson will score on any of Georgia Tech's serves. For this match, we think that it is somewhat likely that Georgia Tech will score on any particular one of their serves; therefore, we expect p to be somewhere in the range of 0.3 to 0.5. We will estimate p by dividing the total number of serves by Georgia Tech in the match by the total number of serves by Georgia Tech that resulted in Clemson scoring a point. The expected value of X_1 is the expected number of serves until Clemson scores a point. Our intuition is that for this match, the expected value will be somewhere in the range of 3 to 5 serves.

Here is the data from the match:

Set #	GT Serves Lost	Number of GT Serves
1	11	25
2	15	24
3	12	25
Total	38	74

(McColskey, T., "Georgia Tech vs. Clemson Women's Volleyball Game Film")

Using the formulas for geometric distributions, we will calculate p and E[X_1] from the data:

$$p = \frac{GT\ Serves\ Lost}{Number\ of\ GT\ Serves} = \frac{38}{74} = 0.513$$

$$E[X_1] = \frac{1}{p} = \frac{1}{0.513} = 1.95$$

Our intuition about the value of p was fairly accurate. We predicted that Georgia Tech won most of their serves; however, we see now that according to our data Georgia Tech fails to score about half the time when they are serving. Our intuition about the expected number of serves was close, but a little bit high. According to our data, Georgia typically wins about two serves in a row when they are serving. This implies that Georgia Tech's expected winning streak is about 3 serves (including the serve they won in order to switch which team was serving).

The biggest gap in our model is that due to the psychology of sports, whether or not a team is winning has an effect on how they perform. If a team is losing, they could either be discouraged or highly motivated to perform better. If a team is winning, they could play less than their best due to over-confidence or play well due a realization that they need to work hard in order to maintain the lead. All of these scenarios illustrate how the probability of winning the next serve (p) is dependent on how the team has performed so far. Not only would varying the value of p have an effect on the expected value of X_1, but more importantly, it would make the use of a geometric distribution inaccurate because the events would no longer be independent.

b. Early Score Update

Question: How many out of the first ten serves did Georgia Tech win?

The probability that Georgia Tech wins a particular serve is p; there are only two possible outcomes of each serve: success (Georgia Tech scores) and failure (Clemson Scores). Using the same argument that we presented in part a, we will assume that each serve is an independent event. Let the random variable X_2 denote the number of serves out of the first ten that Georgia Tech wins. This random variable maps a binomial distribution with probability p and number of trials n = 10. The expected value of X_2 will be the number of serves that Georgia Tech typically wins in the first 10 serves.

Let X_2 = the number of points scored by Georgia Tech in the first 10 serves
$X_2 \sim \text{Bin}(n, p)$
p = the probability that Georgia Tech scores on any given serve
n = the number of serves (in this case, n = 10)
E[X_2] = the expected number of points for Georgia Tech in first 10 serves

Analogy of Binomial Distribution:

Analogy	Volleyball Match
Flip 10 coins	Volleyball is served 10 times
Result is heads	Result is Georgia Tech scores
Result is tails	Result is Clemson scores
Probability of head	Probability Georgia Tech scores

Parameter p is the probability that Georgia Tech will score on any given serve. In this match, we think that it is likely that Georgia Tech will score on any particular serve, so we expect that p will be somewhere in the range of 0.5 to 0.6. We will estimate p by dividing the total number of serves in the match by the total number of points won by Georgia Tech. The expected value of X_2 is the expected number of points that Georgia Tech wins in the first ten serves of the match. We expect that the number of points scored by Georgia Tech in the first ten serves to be either 5 or 6, so the expected value of X_2 should be between 5 and 6. The number of trials in the distribution dictates parameter n. In our case this value is 10.

Here is the data from the match:

Set #	Serves GT Won	Number of Serves
1	25	42
2	25	45
3	25	44
Total	75	131

(McColskey, T., "Georgia Tech vs. Clemson Women's Volleyball Game Film")

Using the formulas for binomial distributions, we will calculate p and E[X_2] from the data:

$$p = \frac{Serves\ GT\ Won}{Number\ of\ Serves} = \frac{75}{131} = 0.573$$

$$E[X_2] = np = 10 * 0.573 = 5.73$$

The probability that we win any particular serve was within the range we estimated. It makes sense that it is greater than 0.5 since we won the match. In a volleyball match, the winning team wins more sets than it loses; this usually translates into winning more points in the match. We correctly predicted the range in which expected value fell. It is interesting to note that in the match between Georgia Tech and Clemson, the score after the first ten serves was 6 – 4 (Georgia Tech's score listed first). This aligns surprisingly well with our calculated expected value.

The biggest gap in this distribution is once again the fact that the probability for winning each serve is not independent of the other serves. A psychologist would argue that the value for p fluctuates during a match based on which team is currently winning and how each team performs under their current conditions. If p is not the same for every serve, then each serve is not independent and the binomial distribution is inappropriate for modeling the data.

c. Scoring in the Long Run

Question: How many points did Georgia Tech score in the twenty minute interval beginning at the start of the match?

Points occur at random during the duration of a match. Let X_3 denote the random variable that counts the number of points in a length of T minutes of game time. Also suppose that the average number of points per minute is λ. The expected value of X_3 is the mean number of points scored in a length of game time. Necessary assumptions for modeling this data with a Poisson distribution include assuming that the points are scored at random times. As previously stated, the probability that we will score is dependent of a few factors such as which team is currently serving, which team is currently winning, etc. These factors are insignificant enough for us to ignore for now and assume that the time of each point is random. The random variable maps a Poisson distribution with an expected value of μ = λT, with T = 20.

Let X_3 = the number of points scored by Georgia Tech in the first 20 minutes
$X_3 \sim \text{Pois}(\mu)$
μ = λT
λ = the average number of points scored per minute
T = the length of interval (20 minutes)
$E[X_3]$ = the mean number of points scored in 20 minutes of game time

Parameter λ is the average number of points scored by Georgia Tech per minute. We will estimate this by dividing the total points scored by Georgia Tech by the total length of the match in minutes. We think that λ will be around one point per minute. Parameter μ is the mean number of points scored in a length of game time; it is calculated by multiplying λ by T. We think that for T = 20, μ will be around 20 points.

Using the formulas for a Poisson distribution, we will calculate λ, μ and $E[X_3]$ from the data:

Total length of game play = 1:19:51 = 79.85 minutes
Total points scored by Georgia Tech = 75

(McColskey, T., "Georgia Tech vs. Clemson Women's Volleyball Game Film")

$$\lambda = \frac{Total\ points\ scored\ by\ GT}{Length\ of\ the\ match\ in\ minutes} = \frac{75}{79.85} = 0.939\ points/min$$

$$\mu = \lambda T = 0.939 * 20 = 18.79\ points$$

We estimated that λ would be around 1 point per minute and μ would be around 20 points. Both of these were a little high. After the first 20 minutes of the Clemson match, Georgia Tech had scored 25 times. This value exceeds our calculated expected value considerably. The reason for this is that the length of the match includes the time that elapsed during the set breaks. The first set break began shortly before the 20 minute mark in the match; the proportion of the time of set breaks to the total duration of the match is not proportionally represented in the first 20 minutes of the match.

Because the time that points are scored in a volleyball match is not completely random, there exists a gap between our model and reality. The probability that a team scores is slightly dependent upon which team is serving, which team is winning, and how much time has already passed. Team members are typically more energetic and prone to scoring at the beginning and end of matches. For these reasons, the exact time of earning points is not entirely random.

Section 3.2: Continuous Random Variables

a. Rapid Fire Scoring

Question: What is the time between two consecutive scores?

To answer this question, we will define a random variable X_4 that is equal to the difference in the time between two consecutive scores. Our intuition is that this random variable maps a normal distribution. We think that most of the results of the time between scores will be close to the average time between scores. There will be some exceptions to this, but we think that those are rare and the normal distribution will be an appropriate model for this random variable. The parameters for a normal distribution are the mean and the variance. We think that the mean is around 30 seconds and the variance is between 9 and 25 seconds squared.

Let X_4 = the time between consecutive scores
$X_4 \sim N(\mu, \sigma^2)$
μ = the mean and σ^2 = the variance

Here is the data for the first set:

Serve #	Time of Score	Difference	Serve #	Time of Score	Difference
1	27:01	-------	22	35:07	0:19
2	27:28	0:27	23	35:29	0:22
3	27:55	0:27	24	35:48	0:19
4	28:15	0:20	25	36:09	0:21
5	28:37	0:22	26	36:34	0:25
6	29:04	0:27	27	38:23	1:49
7	29:24	0:20	28	38:52	0:29
8	29:42	0:18	29	39:16	0:24
9	30:03	0:21	30	39:48	0:32
10	30:31	0:28	31	40:05	0:17
11	30:48	0:17	32	40:27	0:22
12	31:24	0:36	33	40:46	0:19
13	31:50	0:26	34	41:09	0:23
14	32:14	0:24	35	41:43	0:34
15	32:36	0:22	36	42:02	0:19
16	32:59	0:23	37	42:25	0:23
17	33:19	0:20	38	42:44	0:19
18	33:38	0:19	39	43:01	0:17
19	33:57	0:19	40	44:43	1:42
20	34:21	0:24	41	45:04	0:21
21	34:48	0:27	42	45:29	0:25

(McColskey, T., "Georgia Tech vs. Clemson Women’s Volleyball Game Film")

Using the formulas for normal distributions, we will calculate μ and σ^2 from the data:

$$\mu = E[X_4] \approx average\ time\ difference = \frac{\sum t_i}{n} = \frac{\sum t_i}{41} = 27.0\ s$$

The variable t_i is the ith time difference between consecutive scores.

$$\sigma^2 = VAR[X_4] = E[(X_4 - E[X_4])^2] = \frac{\sum (t_i - \mu)^2}{n} = \frac{\sum (t_i - \mu)^2}{41} = 344\ s^2$$

The mean value for time between consecutive scores was close to our expected value for the mean: we slightly overestimated the lengths of the rallies. Our estimation for the variance for this model was not close at all: the consecutive scores with timeouts between them drastically skewed the variance value that we calculated. If we used all of the data except for the two outliers with timeout breaks between the scores, the variance is equal to 21.1 s^2. This is close to our expected variance value. Most of the values in the table do seem to be close to the mean, which is a characteristic of a normal distribution.

The biggest gap between our model and reality is the problem with timeouts between consecutive scores. Another problem with this model is the fact that a normal distribution extends forever in both directions, and our data does not correspond to that information. There is a certain amount of time between the end of a rally and the start of the next rally that sets an unfixed lower bound on the time between consecutive scores. The normal distribution can be an effective model to answer the question if solutions to these gaps are incorporated into the model.

b. Timing of Timeouts

Question: How much time elapses after 20 points are scored in a set until a timeout is granted?

To answer this question, we will define a random variable X_5 that is equal to the time between the 20th point in a set and the time at which the timeout was granted. Our intuition is that this random variable maps a continuous uniform distribution. We think that the time between the 20th point and the timeout is equally likely to happen soon as it is to happen later in the set. The parameters for a continuous uniform distribution are the starting and ending points among which the variable will occur. The starting point for this variable will be at exactly time equals 0 minutes. We think that the ending point will be around 17 minutes since the first 20 points in a set requires about 6 minutes of time, and we think that the set will last around 45 points. The ending time is not an exact time.

Let X_5 = the time between the 20th point in a set and a timeout
$X_5 \sim U(\alpha, \beta)$
α = the starting time of the interval and β = the ending time of the interval

Here is the data from the entire match:

Set #	Time of 20th Point	Time of Timeout	Time Difference
1	34:21	36:44	2:23
1	34:21	43:08	8:47
2	58:56	1:06:01	7:06
2	58:56	1:10:06	11:10
3	1:32:41	1:37:43	5:02
3	1:32:41	1:41:58	9:17

(McColskey, T., "Georgia Tech vs. Clemson Women's Volleyball Game Film")

The starting time is 0 minutes because the timeout could be called immediately after the 20th point is scored. From this data, we cannot identify an solid limits for the ending time for a continuous uniform variable. The ending point is definitely over 11 minutes and 10 seconds, so our guess of around 17 minutes is reasonable.

The biggest gap between this model and reality is the problem that arises when a set requires extra points. Since one of the two teams must win a set by at least 2 points, a set cannot end with the score of 25-24. Our model does not account for the extra time past this score during which a timeout could be called and granted. The variation in length of a set makes the ending time for the random variable very unpredictable. For example, a set could end 25-0 or it could end 25-23; our model will be most effective if the set length is known.

c. Rally Duration

Question: How much time elapses from the start of a rally until one of the teams wins the rally?

To answer this question, we will define a random variable that is equal to the time between the start and end of rally. Our intuition is that this random variable maps an exponential distribution. We think that during a rally, a score is going to happen soon no matter how long or how short the rally has already lasted. The parameter for an exponential distribution is equal to the reciprocal of the expected value. We think that a rally will last around 7 seconds, so our predicted parameter value is 1/7.

Let X_6 = the time between the start and end of a rally
$X_6 \sim \text{Exp}(\lambda)$
$\text{E}[X_6] = 1/\lambda$
$\text{VAR}[X_6] = 1/\lambda^2$

Here is the data for the first set.

Serve #	Duration	Serve #	Duration	Serve #	Duration	Serve #	Duration
1	11 s	12	18 s	23	4 s	34	7 s
2	7 s	13	10 s	24	0 s	35	17 s
3	10 s	14	5 s	25	5 s	36	0 s
4	7 s	15	9 s	26	8 s	37	5 s
5	10 s	16	6 s	27	4 s	38	1 s
6	10 s	17	5 s	28	5 s	39	3 s
7	6 s	18	5 s	29	6 s	40	7 s
8	2 s	19	6 s	30	16 s	41	4 s
9	4 s	20	5 s	31	11 s	42	9 s
10	13 s	21	6 s	32	6 s		
11	0 s	22	3 s	33	2 s		

(McColskey, T., "Georgia Tech vs. Clemson Women's Volleyball Game Film")

Using the formulas for exponential distributions, we will calculate $\text{E}[X_6]$, λ, and $\text{VAR}[X_6]$ from the data:

$$E[X_6] = average\ duration = \frac{\sum t_i}{n} = \frac{\sum t_i}{42} = 6.62\ s$$

The variable t_i is the duration of the ith serve.

$$\lambda = \frac{1}{E[X_6]} = \frac{1}{6.62} = 0.151$$

$$VAR[X_6] = \frac{1}{\lambda^2} = \frac{1}{0.151^2} = 43.8$$

Our guess of 7 seconds seems reasonable for this model. An important thing to note is that in an exponential model, the variance should be equal to one over the parameter value for λ. This does not seem to be close to accurate in our model.

The biggest gap between our model and reality is the fact that very few rallies lasted 1 or 2 seconds. The majority of the rallies lasted 5 to 7 seconds long; there were a few rallies that lasted 16 seconds. When we chose the exponential model, we were trying to accommodate the "rallies" that were service aces or service errors: these last no more than 1 second. There do not seem to be as many of these as we had anticipated when choosing our model. Another problem with this model is that it does not account for the rounding of times that we used in our calculations.

Section 3.3: Joint RVs

a. Georgia Tech Killing to Score

In volleyball, because the players rotate, every player gets a chance to serve. Our intuition is that when Georgia Tech serves and scores, the likelihood that they got a kill does not depend on who served to start the rally. In other words, given that Georgia Tech served, the probability that Georgia Tech got a kill to finish a point is independent of who served.

Here is the data from the volleyball match:

Jersey Number	Player Name	Points GT Won	Kills by GT	Kills / Points
2	London Ackermann	4	2	0.500
4	Sydney Wilson	0	0	N/A
5	Annika Van Gunst	9	6	0.667
7	Teegan Van Gunst	6	3	0.500
8	Rebecca Martin	5	2	0.400
11	Gabriela Stavnetchei	0	0	N/A
12	Wimberly Wilson	3	1	0.333
17	Anna Kavalchuk	0	0	N/A
18	Lauren Pitz	7	0	0.000
33	Ashley Askin	2	1	0.500
	Total	36	15	0.417

(McColskey, T., "Georgia Tech vs. Clemson Women's Volleyball Game Film")

Let X = the result of the serve when Georgia Tech wins the serve
$X = 1$ if Georgia Tech wins the point by a kill
$X = 2$ if Georgia Tech wins the point by any other means
Y = jersey number of server

	a		
b	1	2	P{Y = b}
2	2/36	2/36	4/36
4	0/36	0/36	0/36
5	6/36	3/36	9/36
7	3/36	3/36	6/36
8	2/36	3/36	5/36
11	0/36	0/36	0/36
12	1/36	2/36	3/36
17	0/36	0/36	0/36
18	0/36	7/36	7/36
33	1/36	1/36	2/36
P{X = a}	15/36	21/36	1

In this table, the values in the right column are the probabilities that each player serves the ball during the match. The bottom row includes the probabilities that Georgia Tech wins a point by a kill (column 2) or wins a point by some other means (column 3) when they are serving. These are the marginal probability distributions of variables X and Y. In the middle of the table are the joint probabilities between these two variables: these are the probabilities that X equals a particular value and Y equals a particular value. The X and Y values for the joint probabilities are the a and b values that correspond to that particular cell in the table.

We will check to see if these variables are independent during this match:

If X and Y are independent, then for every entry in the table,

$$P\{X = a; Y = b\} = P\{X = a\} * P\{Y = b\}$$

To prove that X and Y are not independent, we need to find only one counterexample.

$$P\{X = 1; Y = 2\} = P\{X = 1 \cap Y = 2\} = \frac{2}{36}$$
$$P\{X = 1\} * P\{Y = 2\} = \frac{15}{36} * \frac{4}{36} = \frac{5}{108}$$

Since these two probabilities are not equal, X and Y are not independent.

Just because our data does not imply that these variables are independent does not mean that they are not independent. Since this is a relatively small data set, it does not necessarily prove that the variables are dependent. This data looks as though it could have come from random selection of independent random variables. If we had collected data over the whole season and still did not have data that indicated the variables might be independent, then we might consider stating that the variables are dependent. For the serves in the match that Georgia Tech wins, the probability that Georgia Tech kills to score the point does not necessarily depend on who is serving.

b. The Best Server

In sports, because some players are better than others, the probability of success depends on the location of the players. Our intuition is that in volleyball, the result of a serve depends on who is serving: this means that Georgia Tech is more likely to score when certain players are serving than when other players are serving. This means that the probability that Georgia Tech wins when they are serving is dependent upon who is serving.

Here is the data from the volleyball match:

Jersey Number	Player Name	Points GT Won	Serves	GT Point %
2	London Ackermann	4	11	0.364
4	Sydney Wilson	0	1	0.000
5	Annika Van Gunst	9	16	0.563
7	Teegan Van Gunst	6	11	0.545
8	Rebecca Martin	5	11	0.455
11	Gabriela Stavnetchei	0	1	0.000
12	Wimberly Wilson	3	5	0.600
17	Anna Kavalchuk	0	0	N/A
18	Lauren Pitz	7	12	0.583
33	Ashley Askin	2	6	0.333
	Total	36	74	0.486

(McColskey, T., "Georgia Tech vs. Clemson Women's Volleyball Game Film")

X = the result of the serve
X = 1 if Georgia Tech wins the serve
X = 2 if Georgia Tech loses the serve (Clemson wins the serve)

Y = jersey number of server

	a		
b	1	2	P{Y = b}
2	4/74	7/74	11/74
4	0/74	1/74	1/74
5	9/74	7/74	16/74
7	6/74	5/74	11/74
8	5/74	6/74	11/74
11	0/74	1/74	1/74
12	3/74	2/74	5/74
17	0/74	0/74	0/74
18	7/74	5/74	12/74
33	2/74	4/74	6/74
P{X = a}	36/74	38/74	1

As was the case in part a of this section, the values in the right column are the probabilities that each player serves the ball during the match. The bottom row includes the probabilities that Georgia Tech wins a point (column 2) or loses a point (column 3) when they are serving. These are the marginal probability distributions of variables X and Y. In the middle of the table are the joint probabilities between these two variables; these are read the same way that the joint probabilities are read in part a.

We will check to see if these variables are independent during this match:

If X and Y are independent, then for every entry in the table,

$$P\{X = a; Y = b\} = P\{X = a\} * P\{Y = b\}$$

To prove that X and Y are not independent, we need to find only one counterexample.

$$P\{X = 1; Y = 2\} = P\{X = 1 \cap Y = 2\} = \frac{4}{74}$$
$$P\{X = 1\} * P\{Y = 2\} = \frac{36}{74} * \frac{11}{74} = \frac{99}{1369}$$

Since these two probabilities are not equal, X and Y are not independent.

Since this is a small data sample, we cannot conclude independence solely based on our data. If we collected this data for the entire season, we would have a better argument for stating that these variables are dependent. This data does seem to indicate that Georgia Tech had more success when Lauren Pitz and Annika Van Gunst were serving than when London Ackermann and Ashley Askin were serving. This data does seem to be useful in proving our intuition that the player who serves for Georgia Tech and the result of the serve are dependent upon each other.

Section 4.1: Law of Large Numbers

We want to find the average length of time of rallies for a volleyball match. For the law of large numbers to be applied here, we need to assume that the individual rallies map independent variables with the same mean and variance. We have already argued why these assumptions are valid (see Section 3.1 part a). The law of large numbers tells us that the larger our sample size, the more accurate the prediction of our mean will be. To calculate the mean and to illustrate the law of large numbers, we will display a graph of running averages of rally lengths for the first set. Each data point will represent the average length of rallies up to and including the current rally. For example, the first point is just the length of the first rally. The second point is the average length of the first two rallies. The third point is the average of the lengths of the first three rallies, and so on until the 42^{nd} rally.

Let X = the amount of time elapsed during one rally
$\bar{X}_n$ = the average of the time elapsed in the first n rallies

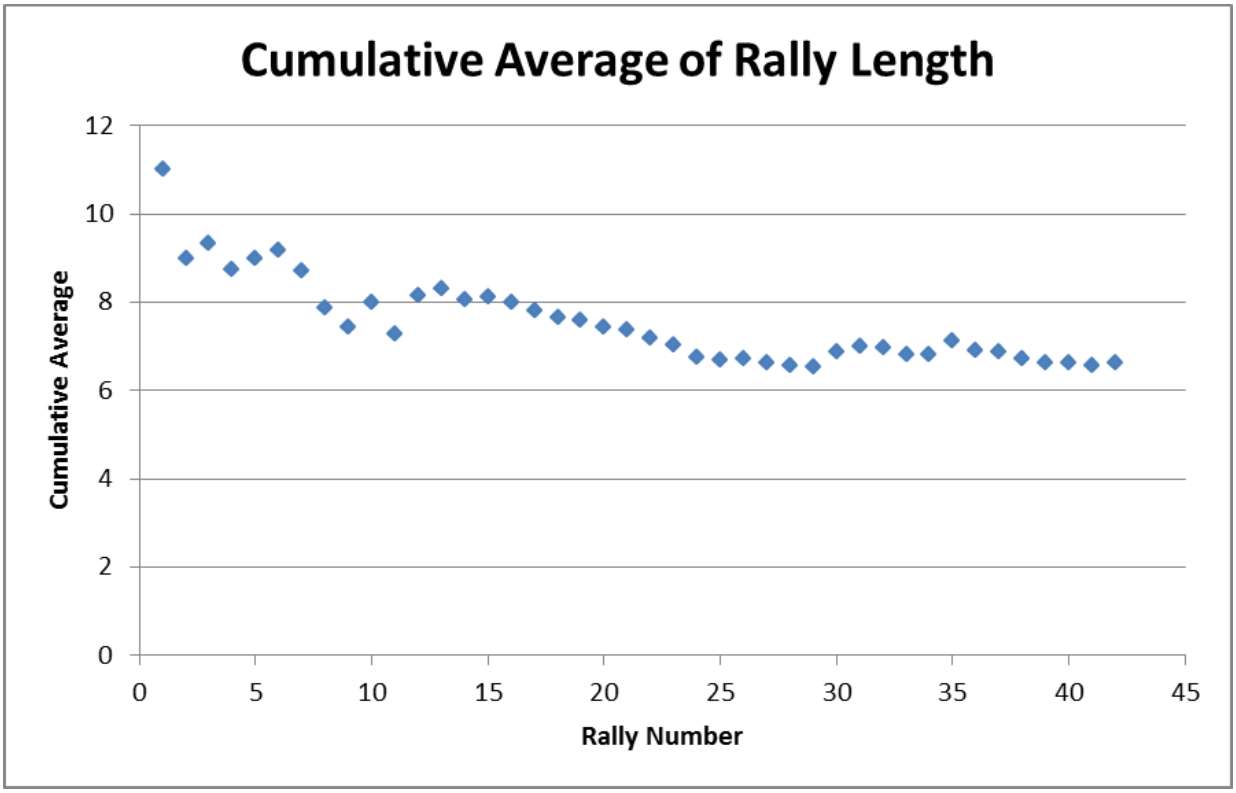
Sample of Data:

Serve #	Start Time	End Time	Length of Volley	Cumulative Avg
1	26:50	27:01	11 s	11.00 s
2	27:21	27:28	7 s	9.00 s
3	27:45	27:55	10- s	7.33 s
4	28:08	28:15	7 s	6.25 s
5	28:27	28:37	10 s	7.00 s
…	…	…	…	…
42	45:20	45:29	9 s	6.38 s

(McColskey, T., "Georgia Tech vs. Clemson Women's Volleyball Game Film")

The full set of Length of Volley times can be found in Section 2.3 part c.

Here is a graph depicting the cumulative average of rally lengths:



As illustrated above, the cumulative average approaches the theoretical average of all rallies as the number of rallies in our sample increases. The average length of all 42 rallies of this set is 6.38 seconds. A more formal way to state this observation is:

$$E[X] \approx E[\bar{X}_n] = 6.62 \, seconds$$

This equation holds true when n is sufficiently large. For instance, the average of only the first three rallies was not very accurate, but after 25 rallies the average was very close to the theoretical mean. As more and more values are averaged, the value obtained gets closer and closer to the expected value. This illustrates the law of large numbers.

Section 4.2: Central Limit Theorem

In the previous example we calculated the mean of the average of the length of rallies. Now, we want to model the lengths of rallies and be able to find the probability of a certain rally length occurring. To apply the Central Limit Theorem, we must assume that the individual rallies map independent and identically distributed random variables. We have already argued why this assumption is valid (see Section 3.1 part a). Next, we will need to calculate the standard deviation of our data set. Once we have the standard deviation, we will apply the Central Limit Theorem to determine the distribution of our sample mean.

Here are the values of X and X^2:

X	11 s	7 s	10 s	7 s	10 s	10 s	…	9 s
X^2	121 s^2	49 s^2	100 s^2	49 s^2	100 s^2	100 s^2	…	81 s^2

We will use the formulas variance and standard deviation to solve for σ.

$$VAR[X] = \frac{\sum(X-\bar{X})}{n-1} = \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1} = \frac{2460-1710}{41} = 18.29$$

$$\sigma = \sqrt{VAR[X]} = \sqrt{18.29} = 4.28$$

The Central Limit Theorem tells us that the sample mean will be approximately normally distributed for large sample sizes, regardless of the distribution from which we are sampling.

$$\frac{\bar{X}-E[\bar{X}]}{\sqrt{VAR[\bar{X}]}} \sim Z(0,1) \; which\ implies\ \bar{X} \sim N(E[\bar{X}], VAR[\bar{X}])$$

From this distribution, it is very easy to calculated many interesting probabilities. For instance, if we wanted to know the probability that a rally would last longer than 17 seconds we would solve as follows:

$$P(a\ rally\ last\ more\ than\ 17\ seconds) = P(\bar{X} > 17)$$
$$= P(\frac{\bar{X}-E[\bar{X}]}{\sqrt{VAR[\bar{X}]}} > \frac{17-E[\bar{X}]}{\sqrt{VAR[\bar{X}]}}) = P\left(Z > \frac{17-6.38}{4.28}\right) = P(Z > 2.48)$$
$$= 1 - P(Z < 2.48) = 1 - 0.9934 = 0.0066$$

Bibliography

General

James, N. D. (2007, August 21) *How to read a volleyball box score*. Retrieved November 22, 2015, from http://tennesseevolleyball.net/2007/08/how-to-read-a-volleyball-box-score/

Montgomery, Douglas C., & Runger, George C. (2011). *Applied Statistics and Probability for Engineers* (5th ed.). Hoboken, N.J.: J. Wiley & Sons.

Dekking, F. M., Kraaikamp, C., Lopuhaä, Meester, L. E.. (2007, February 14). *A Modern Introduction to Probability and Statistics: Understanding Why and How*. United States: Springer London.

Cited

McColskey, T. (Director). (2015). *Georgia Tech vs. Clemson Women's Volleyball Game Film* [Motion picture on DVD]. Georgia Institute of Technology.

McColskey, T. (2015, September 26). Georgia Tech vs Clemson (Sep 25, 2015). Retrieved November 22, 2015, from http://www.ramblinwreck.com/sports/w-volley/stats/2015-2016/13_gt.html

Moore, T. (2015, August 23). *The Volleyball Court* [Online Image]. Retrieved November 22, 2015 from http://conquestvb.com/wp-content/uploads/2015/08/volleyball-court-diagram-with-positions.jpg

Nothstein, J. (2014, March 5). *Volleyball* [Online Image]. Retrieved November 22, 2015 from http://www.cambridgesports.org/uploads/Volleyball.jpg

Data Summary

	Set 1										GT	Total	
18	Pitz	CU					0		2	Ackermann	4	11	
5	A. Van Gunst	K	GT	K	K	CU	5		4	S. Wilson	0	1	
33	Askin	GT	CU				7		5	A. Van Gunst	9	16	
2	Ackermann	GT	CU				9		7	T. Van Gunst	6	11	
8	Martin	CU					10		8	Martin	5	11	
7	T. Van Gunst	CU					11		11	Stavnetchei	0	1	
18	Pitz	GT	GT	CU			14		12	W. Wilson	3	5	
5	A. Van Gunst	K	K	CU			17		18	Pitz	7	12	
33	Askin	K	CU				19		33	Askin	2	6	
2	Ackermann	K	CU				21				36	74	
8	Martin	CU					22						
7	T. Van Gunst	GT	K				25						
	Set 2								2	Ackermann	Point %	Kills by GT	Kill/Point
5	A. Van Gunst	K	CU				2		4	S. Wilson	0.364	2	0.500
33	Askin	CU					3		5	A. Van Gunst	0.000	0	#DIV/0!
2	Ackermann	CU					4		7	T. Van Gunst	0.563	6	0.667
8	Martin	CU					5		8	Martin	0.545	3	0.500
7	T. Van Gunst	CU					6		11	Stavnetchei	0.455	2	0.400
18	Pitz	CU					7		12	W. Wilson	0.000	0	#DIV/0!
5	A. Van Gunst	CU					8		18	Pitz	0.600	1	0.333
11	Stavnetchei	CU					9		33	Askin	0.583	0	0.000
2	Ackermann	CU					10				0.333	1	0.500
8	Martin	K	GT	GT	CU		14					15	
7	T. Van Gunst	GT	K	CU			17						
4	S. Wilson	CU					18						
5	A. Van Gunst	GT	CU				20						
33	Askin	CU					21						
2	Ackermann	K	GT	CU			24						
							25						
	Set 3									Kill	K		
8	Martin	CU					0			GT point	GT		
7	T. Van Gunst	K	GT	CU			3			CU point	CU		
18	Pitz	GT	CU				5			The listed score is the score after			
5	A. Van Gunst	GT	CU				7			the server has finished serving .			
12	W. Wilson	GT	CU				9						
2	Ackermann	CU					10						
8	Martin	K	GT	CU			13						
7	T. Van Gunst	CU					14						
18	Pitz	GT	GT	GT	GT	CU	19						
5	A. Van Gunst	CU					20						
12	W. Wilson	K	GT	CU			23						
2	Ackermann	CU					24						
							25						

Acknowledgements

Dr. Alisha Waller:

Our professor this semester challenged us in essentially every way possible academically. She encouraged us to interrogate our textbook instead of just reading it, form *effective* study groups, and dedicate the time necessary to get the most out of her class. Dr. A was willing to meet with us numerous times outside of her regular office hours in order to explain confusing concepts and my partner and I both feel like we have achieved a much deeper understanding of basic probability and statistics than we would have under a less engaged teacher. I hope this project is a reflection of her effectiveness as a teacher.

Ingrid Cai:

Throughout this semester Ingrid and Mike met together several times in order to help each other. We studied for tests, assisted each other with understanding concepts, and compared class notes. The topics covered in this project are certainly understood by the authors a little better due to the collaborative efforts of Ingrid and Mike in encouraging and investing in one another.

Carolina Fernandez:

Carolina was a constant source of encouragement and inspiration for this project. Mike and Carolina were both new to Georgia Tech this semester. Their friendship that began in ISyE 2027 helped ease the adjustment for both into the rigorous academic environment that is Georgia Tech. Without her company, it is likely Mike would have taken the easy way out and transferred to UGA weeks ago.