STAT 5P81 Final Project

Sampling Methods on NCAA March Madness Tournament Adjusted Offensive Rating in 2021

Christopher Salim (7358955)

Abstract

In the following study, three different sampling methods were used and compared to determine which method provided the best estimator for the population mean, or in this context of study, the mean adjusted offensive rating for teams playing in 2021 Division I college basketball season in United States. The population consisted of 68 teams with their respective seed placement and adjusted offensive rating. For stratified sampling, the dataset was split into 3 strata: seed 1 – 5, seed 6 – 10, and seed 11 – 16. The estimated population means for simple random sampling, stratified random sampling with proportional allocation, and stratified random sampling with Neyman allocation were $\bar{y}_{SRSWOR} = 110.0441$, $\bar{y}_{st} = 109.6529$, and $\bar{y}_{stNeyman} = 110.285$, respectively. Sample mean from simple random sampling had the smallest difference to the population mean. However, Neyman allocation had the smallest variance and standard error. It was found that $var(\bar{y}_{stNeyman}) < var(\bar{y}_{st}) < var(\bar{y}_{SRSWOR})$, or $0.2952667 < 0.31754 < 0.61841$ and $SE(\bar{y}_{stNeyman}) < SE(\bar{y}_{st}) < SE(\bar{y}_{SRSWOR})$ or $0.54338 < 0.563507 < 0.78639$. Therefore, Neyman allocation has the highest level of precision for the estimator of the population mean. It could also be concluded that teams in top seeds had larger adjusted offensive rating mean compared to teams in lower seeds, which conformed to our hypothesis that efficiency is correlated to the seed in NCAA March Madness tournament.

Key Words

Neyman allocation, NCAA March Madness, proportional allocation, simple random sampling, stratified random sampling.

Background and Objective

Sports analytics has become a key component in sports these days. By analyzing sports data, including player and team performance, coaches and management teams are able to plan accordingly and efficiently (Mondello & Kamke, 2014). In basketball, the main objective is to outscore the opposing team in a game. The easier method to measure this is typically by looking at how many points a team can score per match. However, there is an issue to this approach, mainly because points alone do not reflect whether the team is winning. For example, in a NBA game back in 2019, Atlanta Hawks scored 161 points despite they ended up losing the game. This led analysts to shift attention towards other parameters as well. In this project, adjusted offensive rating will be used as a variable of interest. Adjusted offensive rating is an estimate of the offensive efficiency by calculating the points scored per 100 possessions a team would have against the average league defense. It is believed that efficiency gives a better estimate of how well a team can attack the opposing team. To add, during the data collection, most of the times only sample data are available to use for further analysis, therefore the objectives of this project are to determine which sampling method works best to capture the adjusted

offensive ratings in the population and conclude whether the estimated mean is larger for teams placed in the top seeds.

Data Description

The dataset is the match record data from the Division I college basketball season in United States. The dataset was collected from Kaggle (https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset?resource=download). Since basketball is constantly evolving, the dataset from 2021 season, the most recent year in the data, was selected to make sure the dataset is still relevant to today's basketball. Another column of interest, besides adjusted offensive rating, is seed. It is a column that shows where a team is placed in the NCAA March Madness Tournament, an annual mid-March single elimination tournament to determine who will be that year's National Champion. In total, there are 16 seeds and 68 teams qualified every year. The full data table can be found in Appendix A. Figure 1 illustrates the distribution of the population adjusted offensive rating. By quick inspection, it was concluded that the data followed a normal distribution.

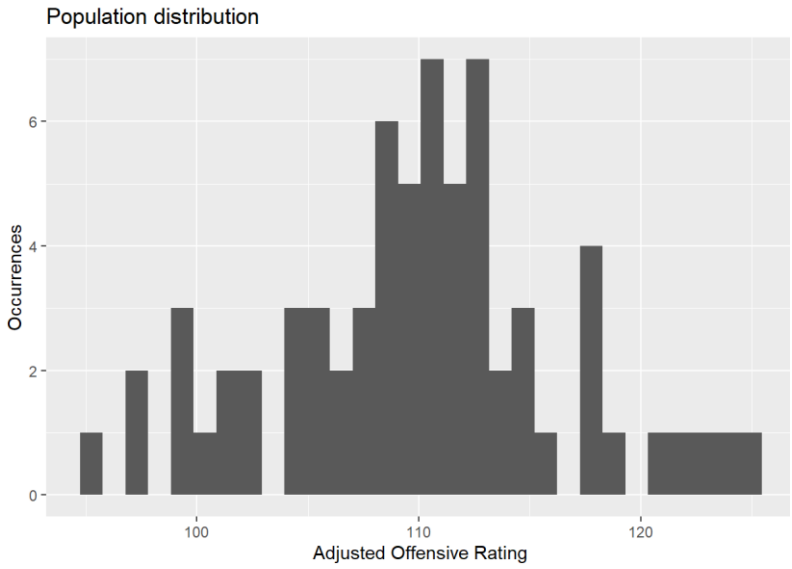

Figure 1. Histogram of the population adjusted offensive rating.

Population Parameters

Some formulas were used to find the population parameters:

$$\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}$$

$$S^2 = \frac{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}{N - 1}$$

$$SE = \frac{S}{\sqrt{N}}$$

$$Margin\ of\ error = e = SE \times Z_{\frac{\alpha}{2}}$$

where $\bar{Y}$ is population mean, S^2 is population variance, S is standard deviation, SE is standard error, and $Z_{\frac{\alpha}{2}}$ is the $\frac{\alpha}{2}$ level's Z-score, in this case α is equal to 0.05.

From R calculations, the following population parameters were found:

$\bar{Y} = 109.9647$

$S^2 = 42.05187$

$S = 6.484741$

$e = 1.541297$

The code in R programming language is provided in the Appendix.

Sampling Methods

Simple Random Sampling

Simple random sampling means every data point in the data set has equal probabilities of being selected (Cochran, 1977). It is also found that sampling without replacement is more efficient than with replacement, so simple random sampling without replacement was implemented here.

Variables

First, the variables needed to calculate the sample size were prepared.

$N = 68$

$\alpha = 0.05$

$S = 6.484741$

$e = 1.541297$

$Z_{\frac{\alpha}{2}} = 1.96$

Sample Size

Since the pre-specified variance was unknown, pre-specified estimation error was assumed instead.

$$n = \frac{\left(\frac{Z_{\frac{\alpha}{2}} \cdot S}{e}\right)^2}{1 + \frac{1}{N}\left(\frac{Z_{\frac{\alpha}{2}} \cdot S}{e}\right)^2} = 34$$

Moving forward, the total sample size would be 34.

Sampled Data from Simple Random Sampling

After obtaining the sample size, a simple random sample of $n = 34$ was selected using R to generate 34 random numbers between 1-68, without replacement. Those random numbers had to be between 1 and 68 because those random numbers would be the indexes for the selected samples from the population. In this case, the population size is 68, so the maximum random number we could get to select the data point is 68. Table 1 shows the selected data from the process. The first column refers to the selected indexes.

Table 1. Sampled data from simple random sampling.

index	team	adjusted_off	seed
9	West Virginia	115.9	3
55	Colgate	110.8	14
54	Liberty	109.1	13
39	Maryland	110	10
63	Drexel	107.9	16
19	Creighton	114.4	5
24	San Diego St.	111.1	6
15	Purdue	112.4	4
61	Grand Canyon	102.7	15
42	Wichita St.	109.6	11
7	Houston	120.6	2
10	Texas	113.8	3
34	Missouri	111.3	9
62	Cleveland St.	99.9	15
13	Florida St.	117.8	4

31	LSU	119.3	8
8	Alabama	111	2
46	Drake	111.8	11
58	Morehead St.	99.3	14
38	Virginia Tech	109.3	10
65	Hartford	97.3	16
27	Clemson	105.9	7
22	BYU	112.8	6
18	Tennessee	108.8	5
29	Oklahoma	111	8
44	UCLA	112.3	11
21	Texas Tech	112.3	6
28	Oregon	113.1	7
26	Connecticut	114.6	7
37	Rutgers	106.7	10
33	St. Bonaventure	111.1	9
50	Winthrop	104.2	12
40	VCU	104.6	10
30	North Carolina	108.8	8

Normality Testing

To check whether the sample still followed the population distribution, a Shapiro-Wilk's normality test was conducted for the sample dataframe as seen in Figure 2. Our p-value was 0.2804, which is > 0.05 and implies that the distribution of the data is not significantly different from the normal distribution.

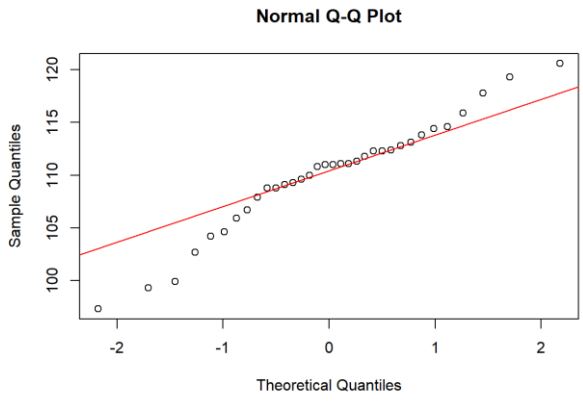

Figure 2. QQ plot to test for normality.

Population Estimates for Simple Random Sampling

Finally, to measure the performance of this sampling method, the population estimates were calculated. In the end, the standard error for the mean estimator was evaluated and the standard error for each method will be compared.

$\bar{y}_{SRSWOR} = 110.0441$, $s^2 = 27.98496$, $s = 5.290082$, $N = 68$, $n = 34$

$$var(\bar{y}_{SRSWOR}) = \frac{N-n}{Nn} S^2 = 0.6184099$$

$$SE_{SRSWOR} = \sqrt{var(\bar{y}_{SRSWOR})} = 0.7863904$$

Stratified Random Sampling (Proportional Allocation)

Strata Definition

Stratified random sampling with proportional allocation is a sampling technique that separates the data into different strata and allocates sample size for each stratum based on the amount of data in the stratum relative to the population size. In order to make sure the comparison between sampling methods was consistent, stratified random sampling also used $n = 34$ for the total sample size. However, the difference is here we split the data into 3 strata:

Stratum 1: Seed 1 – 5

Stratum 2: Seed 6 – 10

Stratum 3: Seed 11-16

Sample Size

After splitting the data, the following values for the population size for each stratum were found:

$N_1 = N_2 = 20$, $N_3 = 28$

Then, to conduct a proportional allocation, the weight for each stratum was calculated by the following formula:

$$w_1 = \frac{N_1}{N} = 0.2941176$$

$$w_2 = \frac{N_2}{N} = 0.2941176$$

$$w_3 = \frac{N_3}{N} = 0.4117647$$

From the weights, the sample size for each stratum could be obtained:

$n_1 = w_1 n = 10$

$n_2 = w_2 n = 10$

$n_3 = w_3 n = 14$

which made sense, because $n_1 + n_2 + n_3 = n$ in this case.

Sampled Data from Stratified Random Sampling (Proportional Allocation)

Simple random sampling was then implemented within each of the strata to choose the sampling units. In each case, R was used to generate n_i random numbers between $1 - N_i$ without replacement. Sampled data points for stratum 1, 2, and 3 can be seen below:

Table 2. Sampled data from stratified random sampling, proportional allocation.

index	Team	adjusted_off	seed	stratum
4	Gonzaga	125.4	1	1
18	Tennessee	108.8	5	1
19	Creighton	114.4	5	1
11	Kansas	108.5	3	1
15	Purdue	112.4	4	1
1	Michigan	118.1	1	1
10	Texas	113.8	3	1
12	Arkansas	110.4	3	1
17	Villanova	117.9	5	1
14	Virginia	114.6	4	1
14	Missouri	111.3	9	2
13	St. Bonaventure	111.1	9	2
1	Texas Tech	112.3	6	2
9	Oklahoma	111	8	2
17	Rutgers	106.7	10	2
8	Oregon	113.1	7	2
5	Florida	110.6	7	2
10	North Carolina	108.8	8	2
16	Georgia Tech	113	9	2
18	Virginia Tech	109.3	10	2
16	Eastern Washington	107.7	14	3
23	Drexel	107.9	16	3
15	Colgate	110.8	14	3

14	Liberty	109.1	13	3
24	Mount St. Mary's	95.7	16	3
2	Wichita St.	109.6	11	3
4	UCLA	112.3	11	3
3	Syracuse	112.8	11	3
25	Hartford	97.3	16	3
1	Michigan St.	105.8	11	3
21	Grand Canyon	102.7	15	3
10	Winthrop	104.2	12	3
17	Abilene Christian	101.8	14	3
28	Appalachian St.	99	16	3

Population Estimates for Stratified Random Sampling (Proportional Allocation)

Variables

$\bar{Y}_1 = 115.955, \bar{Y}_2 = 110.865, \bar{Y}_3 = 105.0429$

$S_1^2 = 26.62576, S_2^2 = 10.34871, S_3^2 = 26.02921$

$N_1 = N_2 = 20, N_3 = 28$

$n_1 = 10, n_2 = 10, n_3 = 14$

$N = 68, n = 34$

$\bar{y}_1 = 114.43, \bar{y}_2 = 110.72, \bar{y}_3 = 105.4786$

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{3} N_i \bar{y}_i = 109.6529$$

$$var(\bar{y}_{st}) = \frac{N-n}{Nn}\sum_{i=1}^{3} w_i S_i^2 = 0.3175405$$

$$SE_{st} = \sqrt{var(\bar{y}_{st})} = 0.5635073$$

Stratified Random Sampling (Neyman Allocation)

Strata Definition

A stratified random sample of $n = 34$ was selected again, so as to provide an easy comparison between sampling methods. Neyman allocation is slightly different from proportional allocation because this also considers the variance of each stratum and the cost of sampling.

Stratum 1: Seed 1 – 5

Stratum 2: Seed 6 – 10

Stratum 3: Seed 11-16

Variables

The sample sizes were determined by using the following variables and calculations for Neyman allocation:

$N_1 = N_2 = 20, N_3 = 28$

$N = 68, n = 34$

$S_1 = 5.160016, S_2 = 3.216941, S_3 = 5.101883$

Sample Size

$$n_1 = \frac{n \cdot N_1 \cdot S_1}{\sum_{i=1}^{3} N_i \cdot S_i} = 11.30446 \approx 11$$

$$n_2 = \frac{n \cdot N_2 \cdot S_2}{\sum_{i=1}^{3} N_i \cdot S_i} = 7.047608 \approx 7$$

$$n_3 = \frac{n \cdot N_3 \cdot S_3}{\sum_{i=1}^{3} N_i \cdot S_i} = 15.64794 \approx 16$$

Upon closer inspection, those values made sense because $n_1 + n_2 + n_3 = n$.

Sampled Data from Stratified Random Sampling (Neyman Allocation)

Simple random sampling was then implemented within each stratum. Similar to proportional allocation, R was used to generate n_i random numbers between $1 - N_i$ without replacement. However, this time we used different values for n_i. Sampled data points for stratum 1, 2, and 3 can be seen below:

Table 3. Sampled data from stratified random sampling, Neyman allocation.

index	team	adjusted_off	seed	stratum
7	Houston	120.6	2	1
16	Oklahoma St.	109.3	4	1
15	Purdue	112.4	4	1
10	Texas	113.8	3	1
5	Iowa	123.5	2	1
3	Illinois	117.7	1	1
19	Creighton	114.4	5	1
14	Virginia	114.6	4	1
20	Colorado	113.7	5	1
18	Tennessee	108.8	5	1

6	Ohio St.	122.1	2	1
18	Virginia Tech	109.3	10	2
3	USC	112	6	2
10	North Carolina	108.8	8	2
20	VCU	104.6	10	2
14	Missouri	111.3	9	2
11	LSU	119.3	8	2
9	Oklahoma	111	8	2
15	Colgate	110.8	14	3
21	Grand Canyon	102.7	15	3
13	UNC Greensboro	102.3	13	3
10	Winthrop	104.2	12	3
14	Liberty	109.1	13	3
2	Wichita St.	109.6	11	3
4	UCLA	112.3	11	3
20	Oral Roberts	107	15	3
3	Syracuse	112.8	11	3
17	Abilene Christian	101.8	14	3
16	Eastern Washington	107.7	14	3
27	Texas Southern	97.3	16	3
12	North Texas	104.4	13	3
11	Ohio	111.9	13	3
9	UC Santa Barbara	107.9	12	3
24	Mount St. Mary's	95.7	16	3

Population Estimates for Stratified Random Sampling (Neyman Allocation)

$\bar{Y}_1 = 115.955, \bar{Y}_2 = 110.865, \bar{Y}_3 = 105.0429$

$S_1^2 = 26.62576, S_2^2 = 10.34871, S_3^2 = 26.02921$

$N_1 = N_2 = 20, N_3 = 28$

$N = 68, n = 34$

$n_1 = 11, n_2 = 7, n_3 = 16$

$\bar{y}_1 = 115.536, \bar{y}_2 = 110.9, \bar{y}_3 = 106.094$

$$\bar{y}_{stNeyman} = \frac{1}{N}\sum_{i=1}^{3} N_i \bar{y}_i = 110.285$$

$$var(\bar{y}_{stNeyman}) = \frac{N-n}{Nn}\sum_{i=1}^{3} w_i S_i^2 = 0.2952667$$

$SE_{stNeyman} = \sqrt{var(\bar{y}_{st})} = 0.5433845$

Precision of Estimators and Conclusion

After applying three different sampling methods to the data and calculating the estimators for each, we could now compare the precision of estimators. When using simple random sampling, the estimated population mean is $\bar{y}_{SRSWOR} = 110.0441$. For stratified sampling with proportional allocation, the estimated population mean is $\bar{y}_{st} = 109.6529$, whereas stratified sampling with Neyman allocation estimated the population mean to be $\bar{y}_{stNeyman} = 110.285$. Compared to the population mean, which is 109.9647, sample mean from simple random sampling had the smallest difference = 0.07941176. Stratified sampling with proportional allocation came next (0.3117647) and stratified sampling with Neyman allocation had the largest difference from the population mean (0.3198864), although the last two were not that far from each other.

However, looking at the other estimators, simple random sampling had the largest variance compared to the other methods, and Neyman allocation had the smallest variance. It was found that $var(\bar{y}_{stNeyman}) < var(\bar{y}_{st}) < var(\bar{y}_{SRSWOR})$, or $0.2952667 < 0.31754 < 0.61841$. Similarly, simple random sampling had the largest standard error, compared to the other methods, and Neyman allocation had the smallest standard error. The result was the following: $SE(\bar{y}_{stNeyman}) < SE(\bar{y}_{st}) < SE(\bar{y}_{SRSWOR})$ or $0.54338 < 0.563507 < 0.78639$. Although simple random sampling provided the closest estimation to the population mean, its variance was 109% larger than the variance from stratified random sampling with Neyman allocation, so we had to be careful when taking this result into consideration.

To conclude, to have a more efficient estimator, stratified random sampling with Neyman allocation is the most appropriate method to use here since it also has the highest level of precision for the estimator of the population mean. Lastly, if we consider the means within stratum under Neyman allocation, $\bar{y}_1 = 115.536$, $\bar{y}_2 = 110.9$, $\bar{y}_3 = 106.094$, which made sense because naturally teams with higher offensive efficiency will score better in the league, thus finishing in the top seeds (seed 1 – 5), whereas teams that are not as efficient will lose more games and be placed in lower seeds in the NCAA March Madness tournament.

References

Cochran, W. G. (1977). *Sampling Techniques.* John Wiley & Sons.

Mondello, M., & Kamke, C. (2014). The Introduction and Application of Sports Analytics in Professional Sport Organizations . *Journal of Applied Sport Management*.

Appendix

Appendix A. Full Data Table

TEAM	ADJOE	SEED
Michigan	118.1	1
Baylor	123.2	1
Illinois	117.7	1
Gonzaga	125.4	1
Iowa	123.5	2
Ohio St.	122.1	2
Houston	120.6	2
Alabama	111	2
West Virginia	115.9	3
Texas	113.8	3
Kansas	108.5	3
Arkansas	110.4	3
Florida St.	117.8	4
Virginia	114.6	4
Purdue	112.4	4
Oklahoma St.	109.3	4
Villanova	117.9	5
Tennessee	108.8	5
Creighton	114.4	5
Colorado	113.7	5
Texas Tech	112.3	6
BYU	112.8	6
USC	112	6
San Diego St.	111.1	6
Florida	110.6	7
Connecticut	114.6	7

Clemson	105.9	7
Oregon	113.1	7
Oklahoma	111	8
North Carolina	108.8	8
LSU	119.3	8
Loyola Chicago	108.5	8
St. Bonaventure	111.1	9
Missouri	111.3	9
Wisconsin	111.3	9
Georgia Tech	113	9
Rutgers	106.7	10
Virginia Tech	109.3	10
Maryland	110	10
VCU	104.6	10
Michigan St.	105.8	11
Wichita St.	109.6	11
Syracuse	112.8	11
UCLA	112.3	11
Utah St.	105.2	11
Drake	111.8	11
Georgetown	108.5	12
Oregon St.	108.5	12
UC Santa Barbara	107.9	12
Winthrop	104.2	12
Ohio	111.9	13
North Texas	104.4	13
UNC Greensboro	102.3	13
Liberty	109.1	13
Colgate	110.8	14
Eastern Washington	107.7	14
Abilene Christian	101.8	14

Morehead St.	99.3	14
Iona	101.1	15
Oral Roberts	107	15
Grand Canyon	102.7	15
Cleveland St.	99.9	15
Drexel	107.9	16
Mount St. Mary's	95.7	16
Hartford	97.3	16
Norfolk St.	99.4	16
Texas Southern	97.3	16
Appalachian St.	99	16
Howard	95.8	NA
South Carolina St.	86.5	NA
Idaho	91.7	NA
Maine	85.2	NA
Fordham	87.5	NA
Denver	93.9	NA
Iowa St.	100.4	NA
Mississippi Valley St.	80	NA
San Diego	97.7	NA
Northern Illinois	94.7	NA
Delaware St.	91.1	NA
Charleston Southern	92.1	NA
American	97.9	NA
Lehigh	92.8	NA
Binghamton	97.1	NA
Towson	99.1	NA
Alabama St.	86.1	NA
Robert Morris	98.1	NA
Boston College	103.3	NA
Tennessee St.	91.2	NA

Cal Poly	90.3	NA
Arkansas Pine Bluff	90.4	NA
Bucknell	99.9	NA
North Carolina Central	91.8	NA
Temple	101.9	NA
Holy Cross	99.4	NA
George Washington	101	NA
DePaul	98.9	NA
Saint Joseph's	101	NA
Western Michigan	95.6	NA
San Jose St.	96.9	NA
Central Connecticut	92.8	NA
USC Upstate	94.6	NA
Middle Tennessee	93.4	NA
Kennesaw St.	90.3	NA
Central Arkansas	97.3	NA
Texas A&M Corpus Chris	90.2	NA
Air Force	96.1	NA
Nebraska Omaha	93.1	NA
Washington	101.4	NA
Tennessee Tech	95.8	NA
Alabama A&M	85.4	NA
Cal St. Fullerton	102.6	NA
Loyola MD	97.7	NA
Long Beach St.	97.4	NA
Eastern Michigan	98.6	NA
Samford	96.4	NA
Alcorn St.	91.1	NA

South Carolina	103.4	NA
Portland	97.6	NA
Wake Forest	102	NA
Northern Arizona	96.7	NA
New Mexico	93.3	NA
St. Francis PA	94.7	NA
Rider	99.8	NA
Houston Baptist	91.2	NA
Canisius	98.5	NA
Delaware	98	NA
Albany	98.6	NA
UC San Diego	100.7	NA
UNC Wilmington	101.6	NA
William & Mary	94.8	NA
Boston University	101.4	NA
NJIT	94.1	NA
Manhattan	91.7	NA
Presbyterian	91.2	NA
Western Illinois	97	NA
Central Michigan	99.1	NA
Illinois St.	97.8	NA
Louisiana Monroe	93.6	NA
Nebraska	101.5	NA
Massachusetts	105.4	NA
IUPUI	96.6	NA
Texas A&M	102.1	NA
East Carolina	98.4	NA
Southern	91.3	NA
Sacramento St.	101.9	NA
Florida A&M	89.1	NA

Dixie St.	91.6	NA
Incarnate Word	94.9	NA
North Florida	100.5	NA
Fort Wayne	100.3	NA
Tennessee Martin	93.2	NA
Southern Miss	92.2	NA
Green Bay	102.2	NA
Southeastern Louisiana	91.3	NA
Lafayette	103.5	NA
Duquesne	102.1	NA
LIU Brooklyn	95.3	NA
Merrimack	92.8	NA
Sacred Heart	99.9	NA
Pacific	102.5	NA
College of Charleston	102.4	NA
St. Francis NY	101.1	NA
UT Rio Grande Valley	90.4	NA
Niagara	98.8	NA
South Florida	100.4	NA
Portland St.	92.4	NA
Cal St. Northridge	99.8	NA
Illinois Chicago	92.7	NA
Quinnipiac	91.9	NA
Coppin St.	91.9	NA
Stony Brook	94.3	NA
Northwestern	103.8	NA
High Point	96.3	NA
Fairleigh Dickinson	101.6	NA
Kentucky	108.5	NA
La Salle	101.9	NA

Charlotte	95.9	NA
Evansville	105.1	NA
Vanderbilt	108.6	NA
FIU	96.5	NA
SIU Edwardsville	93	NA
North Dakota	97.1	NA
Eastern Illinois	94.8	NA
Kansas St.	97.8	NA
California	102.7	NA
Vermont	105.3	NA
Florida Gulf Coast	92.1	NA
UC Davis	96.5	NA
New Hampshire	95.8	NA
Northeastern	98.1	NA
Elon	97.3	NA
UNC Asheville	99.6	NA
Tarleton St.	98.9	NA
Pittsburgh	107.9	NA
Milwaukee	101.3	NA
Tulane	98.7	NA
Ball St.	102.7	NA
McNeese St.	95.8	NA
Rhode Island	103.4	NA
Butler	102.3	NA
Northern Iowa	102	NA
New Orleans	97.3	NA
Miami FL	102	NA
Fairfield	95.9	NA
Valparaiso	96	NA
Lamar	93.5	NA
SMU	107.9	NA
Hawaii	100.7	NA

North Carolina A&T	92.7	NA
Northern Colorado	98.6	NA
Utah Valley	100.8	NA
UMass Lowell	98.7	NA
UCF	105.9	NA
Tulsa	101.2	NA
Jacksonville	91.8	NA
Arkansas St.	100	NA
UMKC	95.6	NA
Penn St.	111.2	NA
Stetson	101.6	NA
Hampton	96.1	NA
Arizona St.	106.6	NA
San Francisco	105.1	NA
Notre Dame	113.4	NA
Gardner Webb	104.6	NA
Little Rock	96.4	NA
Southeast Missouri St.	94.9	NA
Western Carolina	99.4	NA
Troy	92.7	NA
Northwestern St.	96.7	NA
Siena	99.3	NA
Jackson St.	88.2	NA
Monmouth	98.6	NA
New Mexico St.	104.7	NA
Santa Clara	100.8	NA
Marist	92.9	NA
Army	99.2	NA
Detroit	108.3	NA
Cincinnati	102.5	NA
Miami OH	105.3	NA

Seattle	97.9	NA
UTEP	102.7	NA
Fresno St.	100.5	NA
Grambling St.	92.4	NA
Pepperdine	107.6	NA
Utah	110.8	NA
TCU	104.3	NA
Southern Illinois	99.3	NA
Indiana	108.3	NA
UNLV	103.5	NA
Longwood	97.9	NA
Bradley	102.7	NA
Oakland	102.8	NA
Chicago St.	85	NA
North Carolina St.	110.6	NA
Montana St.	98.9	NA
Hofstra	106.3	NA
Florida Atlantic	102.2	NA
Cal Baptist	104.8	NA
Duke	115.1	NA
North Alabama	94.7	NA
Idaho St.	92.3	NA
East Tennessee St.	102.7	NA
VMI	106.2	NA
The Citadel	102.8	NA
Providence	107.7	NA
Murray St.	103.9	NA
Georgia Southern	94.8	NA
UT Arlington	97.6	NA
Marquette	107.4	NA
Auburn	110.5	NA
Louisville	106.6	NA

Bellarmine	107.5	NA
James Madison	104.4	NA
Wagner	101.3	NA
Richmond	109.7	NA
Davidson	112.9	NA
Xavier	109.8	NA
George Mason	101.4	NA
Loyola Marymount	107	NA
Northern Kentucky	102.8	NA
Saint Peter's	90.2	NA
Bowling Green	104.5	NA
Wyoming	108.4	NA
South Dakota	106.7	NA
Georgia	108.2	NA
Seton Hall	109.5	NA
Austin Peay	104.3	NA
Stanford	103.8	NA
Washington St.	101.1	NA
Minnesota	107.4	NA
Saint Louis	110.1	NA
UMBC	100	NA
UC Riverside	102.5	NA
Morgan St.	99.9	NA
Dayton	107.5	NA
Saint Mary's	99.2	NA
Indiana St.	101.5	NA
Cal St. Bakersfield	102.3	NA
UTSA	106.2	NA
Lipscomb	100.2	NA
Radford	97.9	NA
Youngstown St.	101.7	NA
North Dakota St.	101.9	NA

Montana	98.4	NA
Rice	103.3	NA
Mississippi St.	105.4	NA
Navy	103.6	NA
Bryant	103.3	NA
Marshall	108.8	NA
Old Dominion	100.8	NA
Akron	106.6	NA
Kent St.	104	NA
Wofford	104.8	NA
Nevada	108	NA
Mississippi	106.4	NA
St. John's	110.7	NA
Stephen F. Austin	102.9	NA
Prairie View A&M	94.9	NA
Georgia St.	105.3	NA
Coastal Carolina	99.1	NA
South Dakota St.	108.6	NA
Memphis	103.2	NA
Buffalo	106.7	NA
Furman	108.1	NA
Campbell	102.8	NA
South Alabama	101.9	NA
Weber St.	105.7	NA
Missouri St.	105.3	NA
Arizona	113.1	NA
Louisiana Lafayette	101.2	NA
Mercer	105.2	NA
Wright St.	107	NA
Colorado St.	106.4	NA
Texas St.	99.6	NA

Nicholls St.	98	NA
Boise St.	107.7	NA
Chattanooga	102.9	NA
UC Irvine	99	NA
Jacksonville St.	104.5	NA
Sam Houston St.	100.4	NA
Southern Utah	106.6	NA
Western Kentucky	104.6	NA
Louisiana Tech	102.7	NA
Toledo	113.3	NA
UAB	102.5	NA
Eastern Kentucky	101.5	NA
Belmont	108.5	NA

Appendix B. R Code

```
rm(list = ls())

library(tidyverse)

library(ggplot2)

data <- read.csv("dataset.csv") %>%

  filter(!is.na(SEED)) %>%

  select(TEAM, ADJOE, SEED) %>%

  rename(team = "TEAM", adjusted_off = "ADJOE", seed = "SEED")

N <- nrow(data)

pop_mean <- mean(data$adjusted_off)

pop_var <- var(data$adjusted_off)

pop_sd <- sqrt(pop_var)

se <- pop_sd/sqrt(N)

alpha <- 0.05
```

```
z <- qnorm(alpha/2, lower.tail = FALSE)

margin_of_error <- se * z


histogram <- data %>%

  ggplot(aes(x = adjusted_off)) +

  geom_histogram(bins = 20)

histogram


# determining sample size using pre-specified estimation error

n <- (z * pop_sd/margin_of_error)^2/

  (1 + (z * pop_sd/margin_of_error)^2/N)


# SRS without replacement

set.seed(10)

sample_idx_srs <- sample(1:N, n, replace=FALSE)

sample_srs <- data[sample_idx_srs, ]

# normality test

qqnorm(sample_srs$adjusted_off)

qqline(sample_srs$adjusted_off, col="red")

shapiro.test(sample_srs$adjusted_off)


# we know that our sample data is not significantly different

# from normal distribution

mean_srs <- mean(sample_srs$adjusted_off)

var_srs <- var(sample_srs$adjusted_off)
```

```
sd_srs <- sqrt(var_srs)

var_ybar_srs <- (N - n) * pop_var/(n * N)

se_ybar_srs <- sqrt(var_ybar_srs)

histogram_srs <- sample_srs %>%

  ggplot(aes(x = adjusted_off)) +

  geom_histogram(bins = 20)

histogram_srs


# stratified random sampling with allocation

# stratum 1: seed 1-5, stratum 2: seed 6-10

# stratum 3: seed 11-16

data_strat_1 <- data[data$seed >= 1 & data$seed <= 5, ]

data_strat_2 <- data[data$seed >= 6 & data$seed <= 10, ]

data_strat_3 <- data[data$seed >= 11 & data$seed <= 16, ]

rownames(data_strat_1) <- 1:nrow(data_strat_1)

rownames(data_strat_2) <- 1:nrow(data_strat_2)

rownames(data_strat_3) <- 1:nrow(data_strat_3)


pop_mean_strat_1 <- mean(data_strat_1$adjusted_off)

pop_mean_strat_2 <- mean(data_strat_2$adjusted_off)

pop_mean_strat_3 <- mean(data_strat_3$adjusted_off)


pop_var_strat_1 <- var(data_strat_1$adjusted_off)

pop_var_strat_2 <- var(data_strat_2$adjusted_off)

pop_var_strat_3 <- var(data_strat_3$adjusted_off)
```

```
pop_sd_strat_1 <- sqrt(pop_var_strat_1)

pop_sd_strat_2 <- sqrt(pop_var_strat_2)

pop_sd_strat_3 <- sqrt(pop_var_strat_3)


N1 <- nrow(data_strat_1)

N2 <- nrow(data_strat_2)

N3 <- nrow(data_strat_3)

n1 <- N1/N * n

n2 <- N2/N * n

n3 <- N3/N * n


sample_idx_prop_1 <- sample(1:N1, n1, replace=FALSE)

sample_idx_prop_2 <- sample(1:N2, n2, replace=FALSE)

sample_idx_prop_3 <- sample(1:N3, n3, replace=FALSE)

sample_prop_1 <- data_strat_1[sample_idx_prop_1, ]

sample_prop_2 <- data_strat_2[sample_idx_prop_2, ]

sample_prop_3 <- data_strat_3[sample_idx_prop_3, ]


mean_prop_1 <- mean(sample_prop_1$adjusted_off)

mean_prop_2 <- mean(sample_prop_2$adjusted_off)

mean_prop_3 <- mean(sample_prop_3$adjusted_off)


yst_prop <- 1/N * (mean_prop_1 * N1 + mean_prop_2 * N2 + mean_prop_3 * N3)
```

```
var_yst_prop <- (N - n)*(N1/N * pop_var_strat_1 + N2/N * pop_var_strat_2 + N3/N *
pop_var_strat_3)/(N*n)

se_yst_prop <- sqrt(var_yst_prop)


# neyman allocation

n1_neyman <- round((n * N1 * pop_sd_strat_1)/(N1 * pop_sd_strat_1 + N2 * pop_sd_strat_2 + N3 *
pop_sd_strat_3))

n2_neyman <- round((n * N2 * pop_sd_strat_2)/(N1 * pop_sd_strat_1 + N2 * pop_sd_strat_2 + N3 *
pop_sd_strat_3))

n3_neyman <- round((n * N3 * pop_sd_strat_3)/(N1 * pop_sd_strat_1 + N2 * pop_sd_strat_2 + N3 *
pop_sd_strat_3))


sample_idx_neyman_1 <- sample(1:N1, n1_neyman, replace=FALSE)

sample_idx_neyman_2 <- sample(1:N2, n2_neyman, replace=FALSE)

sample_idx_neyman_3 <- sample(1:N3, n3_neyman, replace=FALSE)

sample_neyman_1 <- data_strat_1[sample_idx_neyman_1, ]

sample_neyman_2 <- data_strat_2[sample_idx_neyman_2, ]

sample_neyman_3 <- data_strat_3[sample_idx_neyman_3, ]


mean_neyman_1 <- mean(sample_neyman_1$adjusted_off)

mean_neyman_2 <- mean(sample_neyman_2$adjusted_off)

mean_neyman_3 <- mean(sample_neyman_3$adjusted_off)


yst_neyman <- 1/N * (mean_neyman_1 * N1 + mean_neyman_2 * N2 + mean_neyman_3 * N3)

var_yst_neyman <- (N1/N * pop_sd_strat_1 + N2/N * pop_sd_strat_2 + N3/N * pop_sd_strat_3)^2/n
- (N1/N * pop_var_strat_1 + N2/N * pop_var_strat_2 + N3/N * pop_var_strat_3)/N
```

```
se_yst_neyman <- sqrt(var_yst_neyman)
```