



SAMPLING METHODS ON NCAA MARCH MADNESS TOURNAMENT ADJUSTED OFFENSIVE RATING DATA IN 2021



Christopher Salim

Presentation Outline



1

Background and objective

2

Dataset

3

Results for each sampling method

4

Conclusion



BACKGROUND AND OBJECTIVE

Why sampling in sports analytics?

Why analyze basketball offensive efficiency, to be more specific?

Answer:

Atlanta Hawks scored 161 points but still lost the game. What does it tell us?

We want to sample adjusted offensive ratings (points per 100 possessions) in relation to seed placements in NCAA March Madness Tournament 2021



Objectives

1. Determine which sampling method works best to capture the adjusted offensive ratings in the population.
2. Conclude whether the estimated mean is larger for teams placed in the top seeds.

Sampling methods used:

- Simple random sampling
- Stratified sampling (proportional allocation)
- Stratified sampling (Neyman allocation)



DATASET

Data Description

Source:

<https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset?resource=download>

Only picked the dataset from 2021 season.

Initially we had 347 rows and 22 columns, but after data cleaning (only considering teams that qualified for the NCAA March Madness), we got 68 rows left.

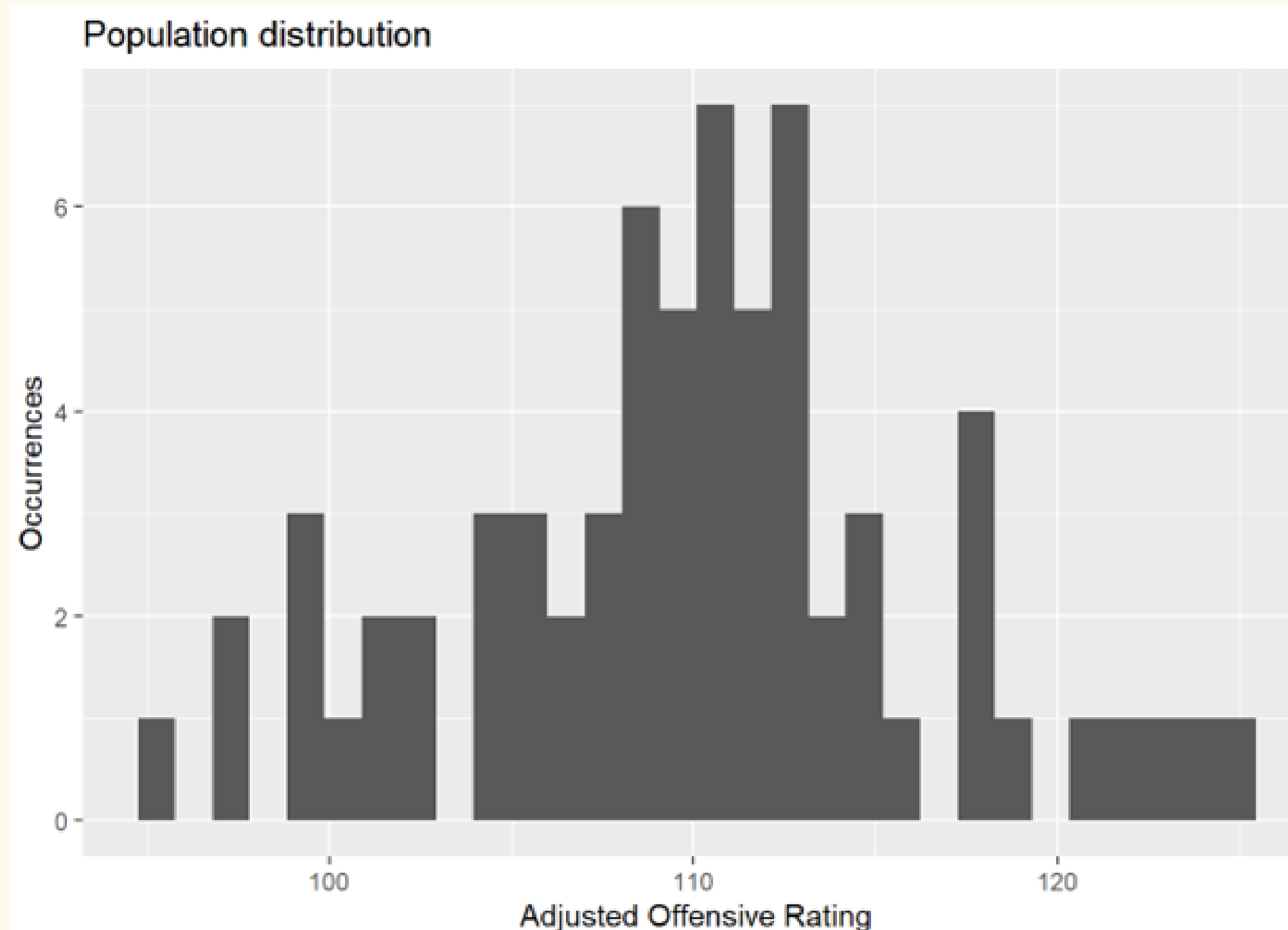
3 variables of interest (team name, adjusted offensive rating, and seed placement)

Data Overview

First 5 rows of the data

| team | adjusted_off | seed |
|----------|--------------|------|
| Michigan | 118.1 | 1 |
| Baylor | 123.2 | 1 |
| Illinois | 117.7 | 1 |
| Gonzaga | 125.4 | 1 |
| Iowa | 123.5 | 2 |

Data Overview (cont.)



Some teams were more efficient than others, but overall the dataset is normally distributed.

Population Parameters

Important formulas:

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

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

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

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

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

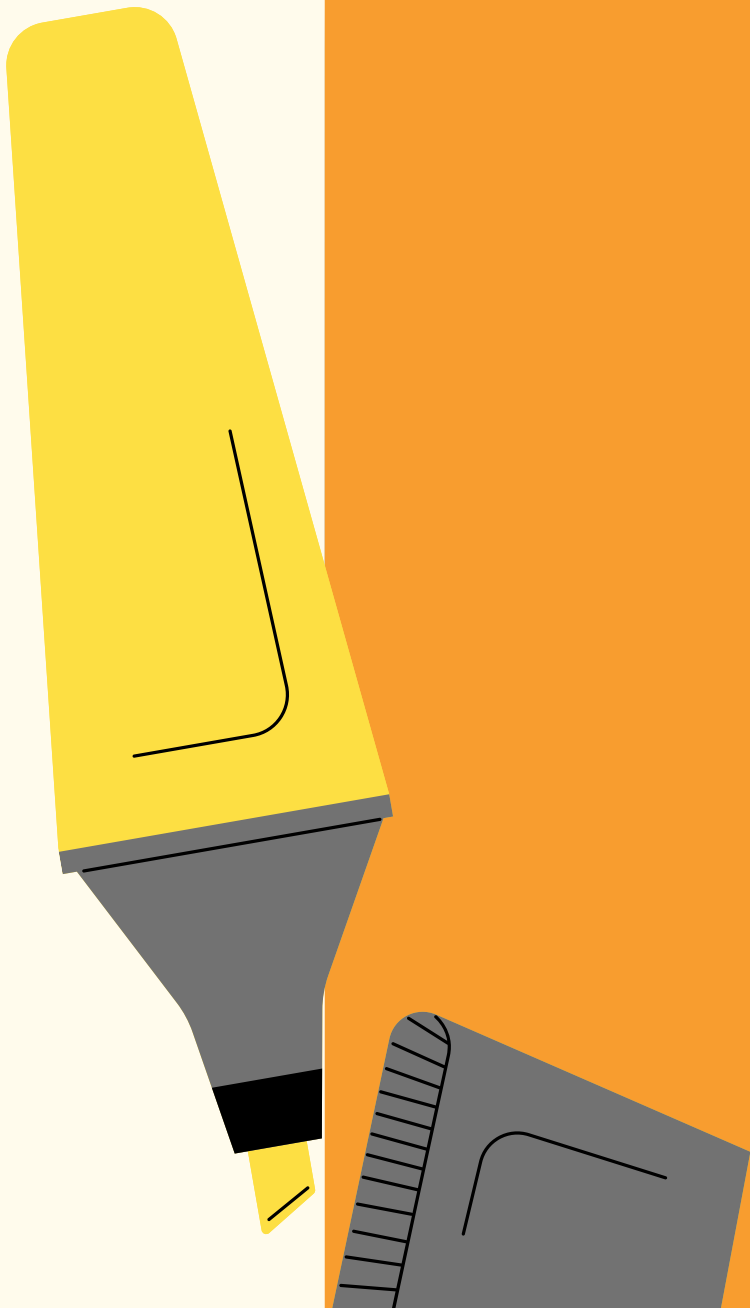
Population Parameters (cont.)

$$\bar{Y} = 109.9647$$

$$S^2 = 42.05187$$

$$S = 6.484741$$

$$e = 1.541297$$



SAMPLING METHODS

Simple random sampling (SRS)

All teams in all seed have the same probability to be picked (**1/68**).

SRS without replacement's efficiency > SRS with replacement's efficiency

We would use SRS without replacement here.

But what should our sample size be?

Sample Size Determination

Given the pre-specified estimation error:

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

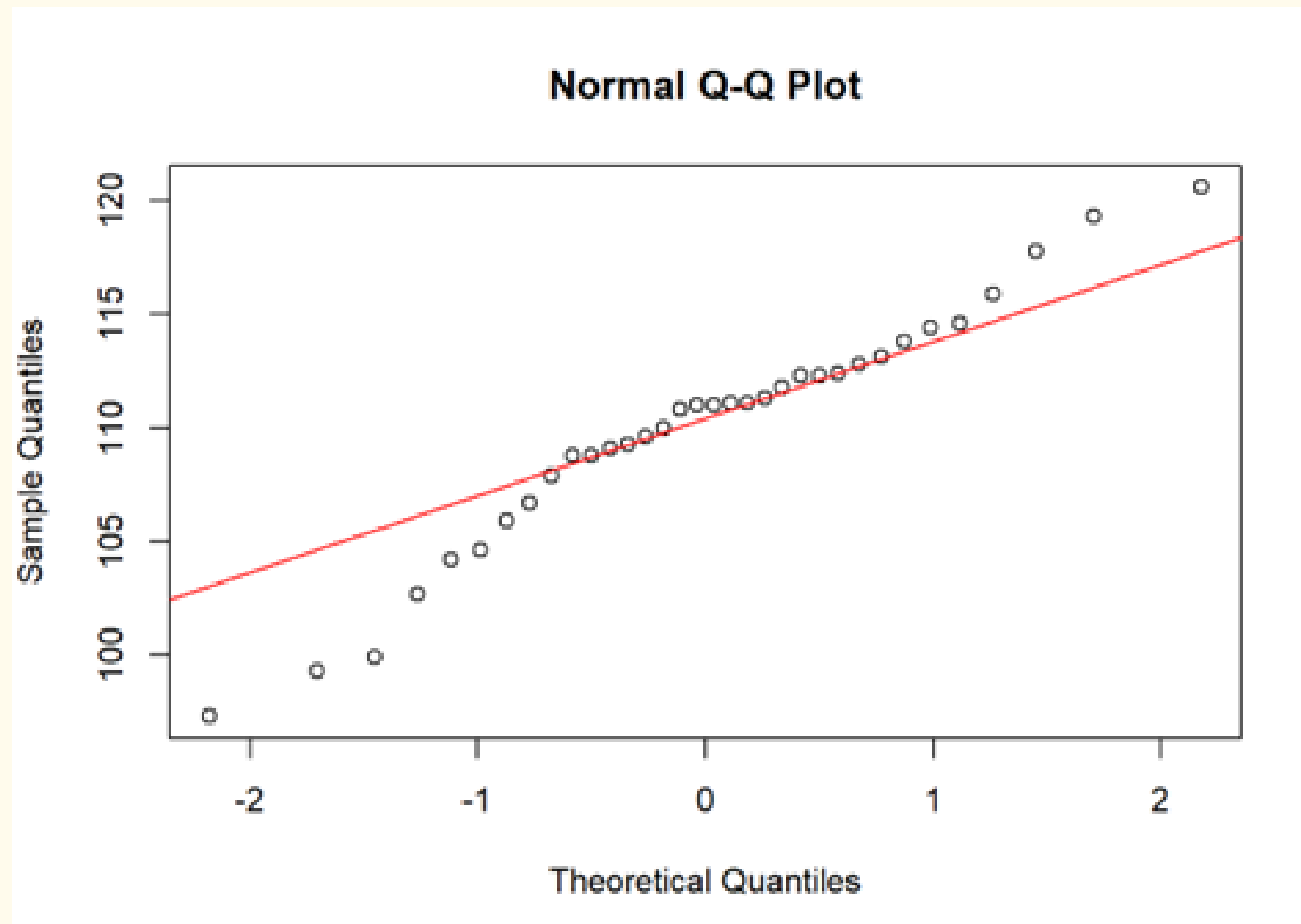
Simple random sampling (SRS) (cont.)

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.

Is our sample dataset representative and normally distributed?

Simple random sampling (SRS) (cont.)

QQ plot was generated and Shapiro-Wilk normality test was conducted to answer the previous question.



p-value = 0.2804 > 0.05, this sample dataset is normally distributed.

Population estimates for SRS

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

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

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

Stratified random sampling (proportional allocation)

The dataset was divided into 3 strata:

Stratum 1: Seed 1 – 5

Stratum 2: Seed 6 – 10

Stratum 3: Seed 11-16

This method allocates sample size for each stratum based on the amount of data points in the stratum relative to the population size

Stratified random sampling (proportional allocation) (cont.)

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

Sample size for each stratum:

Weight for each stratum:

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

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

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

$$n_1 = w_1 n = 10$$

$$n_2 = w_2 n = 10$$

$$n_3 = w_3 n = 14$$

Stratified random sampling (proportional allocation) (cont.)

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

Population estimates for stratified random sampling (proportional allocation)

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

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

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

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

$$N = 68, n = 34$$

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

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

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

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

Stratified random sampling (Neyman allocation)

Key difference from proportional allocation:

This also considers the variance of each stratum and the cost of sampling.

Sample sizes for Neyman allocation

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

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

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

Stratified random sampling (Neyman allocation) (cont.)

Similar to proportional allocation, R was used to generate n_i random numbers between 1- N_i without replacement.

Population estimates for stratified random sampling (Neyman allocation)

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

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

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

$$N = 68, n = 34$$

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

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

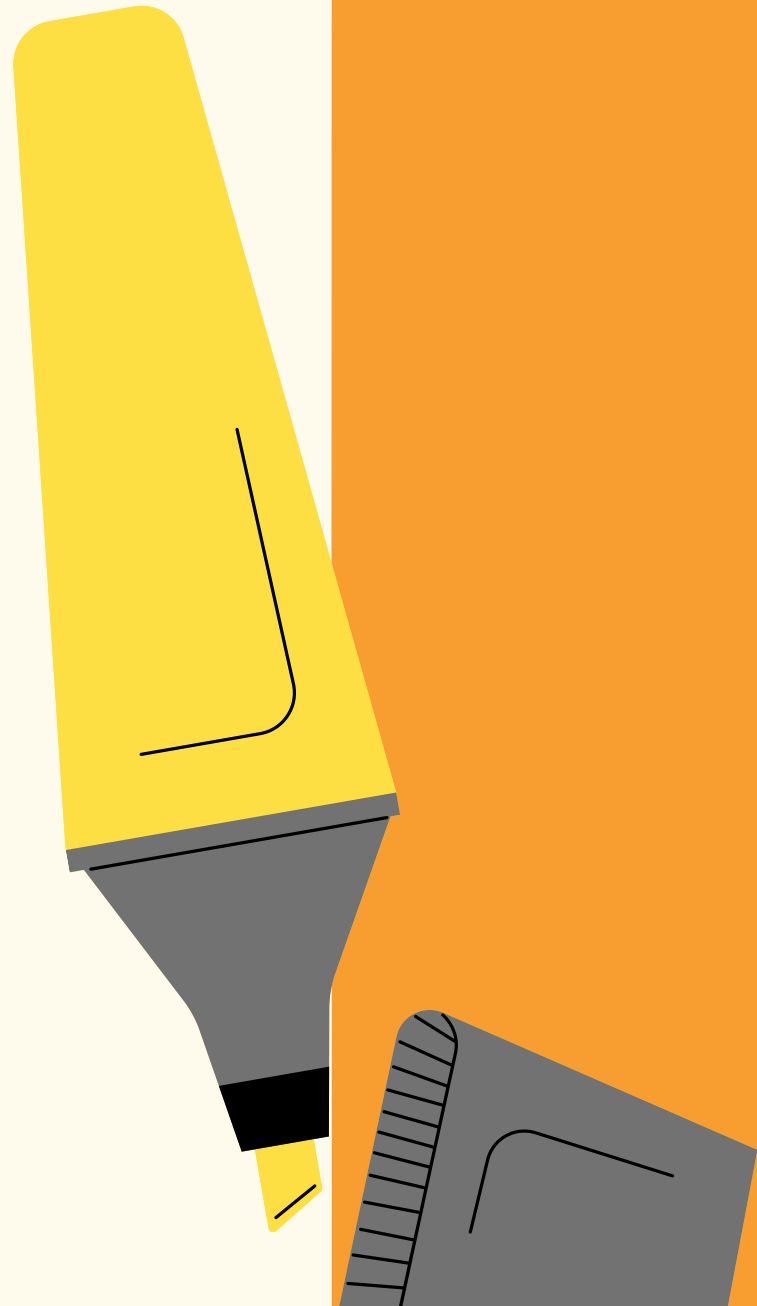
Population estimates for stratified random sampling (Neyman allocation) (cont.)

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

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

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

CONCLUSION



$$\text{var}(\bar{y}_{stNeyman}) < \text{var}(\bar{y}_{st}) < \text{var}(\bar{y}_{SRSWOR})$$

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

If we consider the means estimation between strata:

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

Teams at higher seeds have more offensive efficiency than teams at lower seeds.



Thank you
Any questions?



| | |
|--|----------------------------------------------------------|
| | <p>Christopher Salim christopher.salim@brocku.ca</p> |
| | |