# Lecture 8: Sampling Methods

Donglei Du
(ddu@unb.edu)

Faculty of Business Administration, University of New Brunswick, NB Canada Fredericton
E3B 9Y2

# Table of contents


1. Sampling Methods
   - Why Sampling
   - Probability vs non-probability sampling methods
   - Sampling with replacement vs without replacement
   - Random Sampling Methods
2. Simple random sampling with and without replacement
   - Simple random sampling without replacement
   - Simple random sampling with replacement
3. Sampling error vs non-sampling error
4. Sampling distribution of sample statistic
   - Histogram of the sample mean under SRR
5. Distribution of the sample mean under SRR: The central limit theorem

# Layout


1. Sampling Methods
   - Why Sampling
   - Probability vs non-probability sampling methods
   - Sampling with replacement vs without replacement
   - Random Sampling Methods
2. Simple random sampling with and without replacement
   - Simple random sampling without replacement
   - Simple random sampling with replacement
3. Sampling error vs non-sampling error
4. Sampling distribution of sample statistic
   - Histogram of the sample mean under SRR
5. Distribution of the sample mean under SRR: The central limit theorem

# Why sampling?

- The physical impossibility of checking all items in the population, and, also, it would be too time-consuming
- The studying of all the items in a population would not be cost effective
- The sample results are usually adequate
- The destructive nature of certain tests

# Sampling Methods

- Probability Sampling: Each data unit in the population has a known likelihood of being included in the sample.
- Non-probability Sampling: Does not involve random selection; inclusion of an item is based on convenience

# Sampling Methods

- Sampling with replacement: Each data unit in the population is allowed to appear in the sample more than once.
- Sampling without replacement: Each data unit in the population is allowed to appear in the sample no more than once.

# Random Sampling Methods

- Most commonly used probability/random sampling techniques are
  - Simple random sampling
  - Stratified random sampling
  - Cluster random sampling

# Simple random sampling

- Each item(person) in the population has an equal chance of being included.

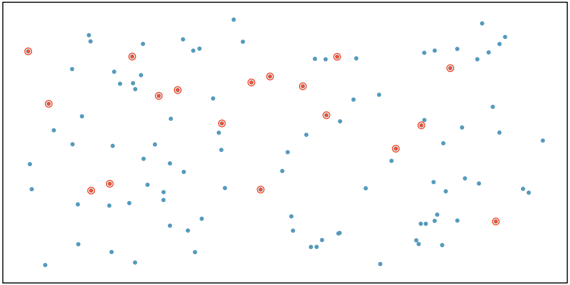


Figure: Credit: Open source textbook: OpenIntro Statistics, 2nd Edition, D. M. Diez, C. D. Barr, and M. Cetinkaya-Rundel (http://www.openintro.org/stat/textbook.php)

# Stratified random sampling

- A population is first divided into strata which are made up of similar observations. Take a simple random sample from each stratum.

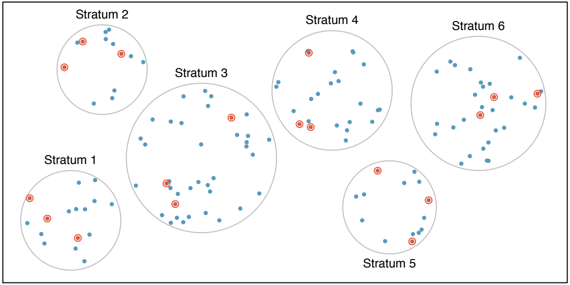



Figure: Credit: Open source textbook: OpenIntro Statistics, 2nd Edition, D. M. Diez, C. D. Barr, and M. Cetinkaya-Rundel (http://www.openintro.org/stat/textbook.php)

# Cluster random sampling

- A population is first divided into clusters which are usually not made up of homogeneous observations, and take a simple random sample from a random sample of clusters.

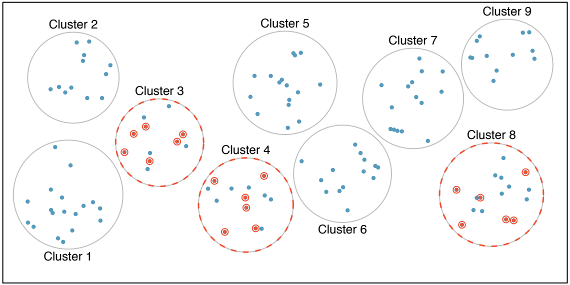



Figure: Credit: Open source textbook: OpenIntro Statistics, 2nd Edition, D. M. Diez, C. D. Barr, and M. Cetinkaya-Rundel (http://www.openintro.org/stat/textbook.php)

# Layout


1. Sampling Methods
   - Why Sampling
   - Probability vs non-probability sampling methods
   - Sampling with replacement vs without replacement
   - Random Sampling Methods
2. Simple random sampling with and without replacement
   - Simple random sampling without replacement
   - Simple random sampling with replacement
3. Sampling error vs non-sampling error
4. Sampling distribution of sample statistic
   - Histogram of the sample mean under SRR
5. Distribution of the sample mean under SRR: The central limit theorem

# Simple random sampling without replacement (SRN)

- Repeat the following process until the requested sample is obtained:
  - Randomly (with equal probability) select an item, record it, and discard it
  - Example: draw cards one by one from a deck without replacement.
- This technique is the simplest and most often used sampling technique in practice.

# R code

- Given a population of size $N$, choose a sample of size $n$ using SRN

```
> N<-5
> n<-2
> sample(1:N, n, replace=FALSE)
```

# Simple random sampling with replacement (SRR)

- Repeat the following process until the requested sample is obtained:
  - Randomly (with equal probability) select an item, record it, and replace it
  - Example: draw cards one by one from a deck with replacement.
- This is rarely used in practice, since there is no meaning to include the same item more than once.
- However, it is preferred from a theoretical point of view, since
  - It is easy to analyze mathematically.
  - Moreover, SRR is a very good approximation for SRN when N is large.

# R code

- Given a population $\{1, \ldots, N\}$ of size $N$, choose a sample of size $n$ using SRR

```
> N<-5
> n<-2
> sample(1:N, n, replace=TRUE)
```

# Layout


1. Sampling Methods
   - Why Sampling
   - Probability vs non-probability sampling methods
   - Sampling with replacement vs without replacement
   - Random Sampling Methods
2. Simple random sampling with and without replacement
   - Simple random sampling without replacement
   - Simple random sampling with replacement
3. Sampling error vs non-sampling error
4. Sampling distribution of sample statistic
   - Histogram of the sample mean under SRR
5. Distribution of the sample mean under SRR: The central limit theorem

# Sampling error vs non-sampling error

- **Sampling error**: the difference between a sample statistic and its corresponding population parameter. This error is inherent in
  - The sampling process (since sample is only part of the population)
  - The choice of statistics (since a statistics is computed based on the sample).
- **Non-sample Error**: This error has no relationship to the sampling technique or the estimator. The main reasons are human-related
  - data recording
  - non-response
  - sample selection

# Layout


1. Sampling Methods
   - Why Sampling
   - Probability vs non-probability sampling methods
   - Sampling with replacement vs without replacement
   - Random Sampling Methods
2. Simple random sampling with and without replacement
   - Simple random sampling without replacement
   - Simple random sampling with replacement
3. Sampling error vs non-sampling error
4. **Sampling distribution of sample statistic**
   - Histogram of the sample mean under SRR
5. Distribution of the sample mean under SRR: The central limit theorem

# Sampling distribution of sample statistic

- Sampling distribution of sample statistic: The probability distribution consisting of all possible sample statistics of a given sample size selected from a population using one probability sampling.
- Example: we can consider the sampling distribution of the sample mean, sample variance etc.

# An example of the sampling distribution of sample mean under SRR

- Consider a small population $\{1, 2, 3, 4, 5\}$ with size $N = 5$. Let us randomly choose a sample of size $n = 2$ via SRR.
- It is understood that sample is ordered. Then there are $N^n = 5^2 = 25$ possible samples; namely

| sample | $\bar{x}$ | sample | $\bar{x}$ | sample | $\bar{x}$ | sample | $\bar{x}$ | sample | $\bar{x}$ |
|---|---|---|---|---|---|---|---|---|---|
| (1,1) | 1 | (2,1) | 1.5 | (3,1) | 2 | (4,1) | 2.5 | (5,1) | 3 |
| (1,2) | 1.5 | (2,2) | 2 | (3,2) | 2.5 | (4,2) | 3 | (5,2) | 3.5 |
| (1,3) | 2 | (2,3) | 2.5 | (3,3) | 3 | (4,3) | 3.5 | (5,1) | 4 |
| (1,4) | 2.5 | (2,4) | 3 | (3,4) | 3.5 | (4,4) | 4 | (5,1) | 4.5 |
| (1,5) | 3 | (2,5) | 3.5 | (3,5) | 4 | (4,5) | 4.5 | (5,1) | 5 |

# An example of the sampling distribution of sample mean under SRR

- Let us find the sampling distribution of the sample mean:

| $\bar{X}$ | Probability |
|---|---|
| 1 | 1/25 |
| 1.5 | 2/25 |
| 2 | 3/25 |
| 2.5 | 4/25 |
| 3 | 5/25 |
| 3.5 | 4/25 |
| 4 | 3/25 |
| 4.5 | 2/25 |
| 5 | 1/25 |

# The mean and variance of the sample mean under SRR

- Let us find the mean and variance of the sampling distribution of the sample mean:

| $\bar{X}$ | $P(\bar{X})$ | $\bar{X}P(\bar{X})$ | $\bar{X}^2P(\bar{X})$ |
|---|---|---|---|
| 1 | 1/25 | 1/25 | 1/25 |
| 1.5 | 2/25 | 3/25 | 4.5/25 |
| 2 | 3/25 | 6/25 | 12/25 |
| 2.5 | 4/25 | 10/25 | 25/25 |
| 3 | 5/25 | 15/25 | 45/25 |
| 3.5 | 4/25 | 14/25 | 49/25 |
| 4 | 3/25 | 12/25 | 48/25 |
| 4.5 | 2/25 | 9/25 | 40.5/25 |
| 5 | 1/25 | 5/25 | 25/25 |
| | | 75/25=3 | 250/25=10 |

# The mean and variance of the sample mean under SRR

- So the mean and variance of the sample mean are given as

$$\begin{aligned} \bar{x} &= 3 \\ s^2 &= 10 - 3^2 = 1 \end{aligned}$$

- On the other hand, the population mean and variance are given as

$$\begin{aligned} \mu &= \frac{1 + 2 \ldots + 5}{5} = 3 \\ \sigma^2 &= \frac{55 - \frac{15^2}{5}}{5} = 2 \end{aligned}$$

# Relationship between sample and population mean and variance under SRR

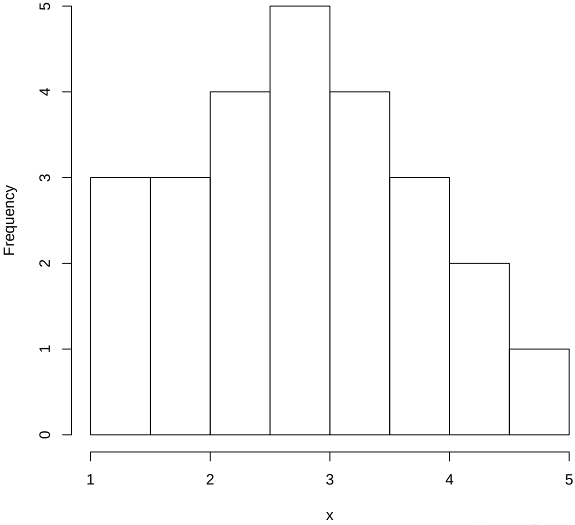
- So from this example

$$\begin{aligned} \bar{x} &= \mu = 3 \\ s^2 &= \frac{\sigma^2}{2} = \frac{2}{2} = 1 \end{aligned}$$

- The above relationship is true for any population of size $N$ and sample of size $n$

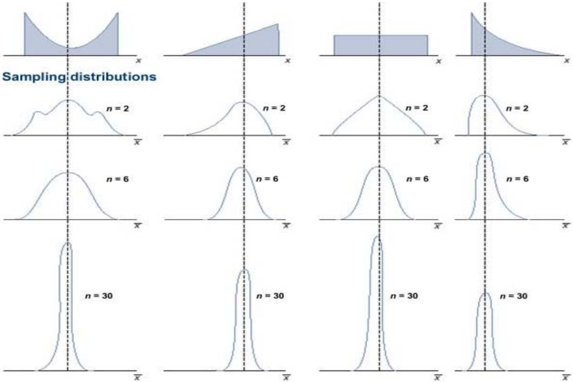
$$\begin{aligned} \bar{x} &= \mu \\ s^2 &= \frac{\sigma^2}{n} \end{aligned}$$

- Let us look the histogram of the sample mean in the above example.



# Distribution of the sample mean under SRR for various population

- Let us look the histogram of the sample mean for various population.



# Layout

1. Sampling Methods
   - Why Sampling
   - Probability vs non-probability sampling methods
   - Sampling with replacement vs without replacement
   - Random Sampling Methods
2. Simple random sampling with and without replacement
   - Simple random sampling without replacement
   - Simple random sampling with replacement
3. Sampling error vs non-sampling error
4. Sampling distribution of sample statistic
   - Histogram of the sample mean under SRR
5. Distribution of the sample mean under SRR: The central limit theorem

# Distribution of the sample mean under SRR: The central limit theorem

- **The central limit theorem**: The sampling distribution of the means of all possible samples of size $n$ generated from the population using SRR will be approximately normally distributed when $n$ goes to infinity.

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

- How large should $n$ be for the sampling mean distribution to be approximately normal?
  - In practice, $n \geq 30$
  - If $n$ large, and we do not know $\sigma$, then we can use sample standard deviation instead. Then Central Limit Theorem is still true!

# Distribution of the sample mean under SRR for small sample

- If $n$ small, and we do not know $\sigma$, but we know the population is normally distributed, then replacing the standard deviation with sample standard deviation results in the Student's $t$ distribution with degrees of freedom $df = n - 1$:

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t(n-1)$$

- Like $Z$, the $t$-distribution is continuous
- Takes values between $-\infty$ and $\infty$
- It is bell-shaped and symmetric about zero
- It is more spread out and flatter at the center than the $z$-distribution
- For larger and larger values of degrees of freedom, the $t$-distribution becomes closer and closer to the standard normal distribution

# Comparison of $t$ Distributions with Normal distribution