Statistical Sampling

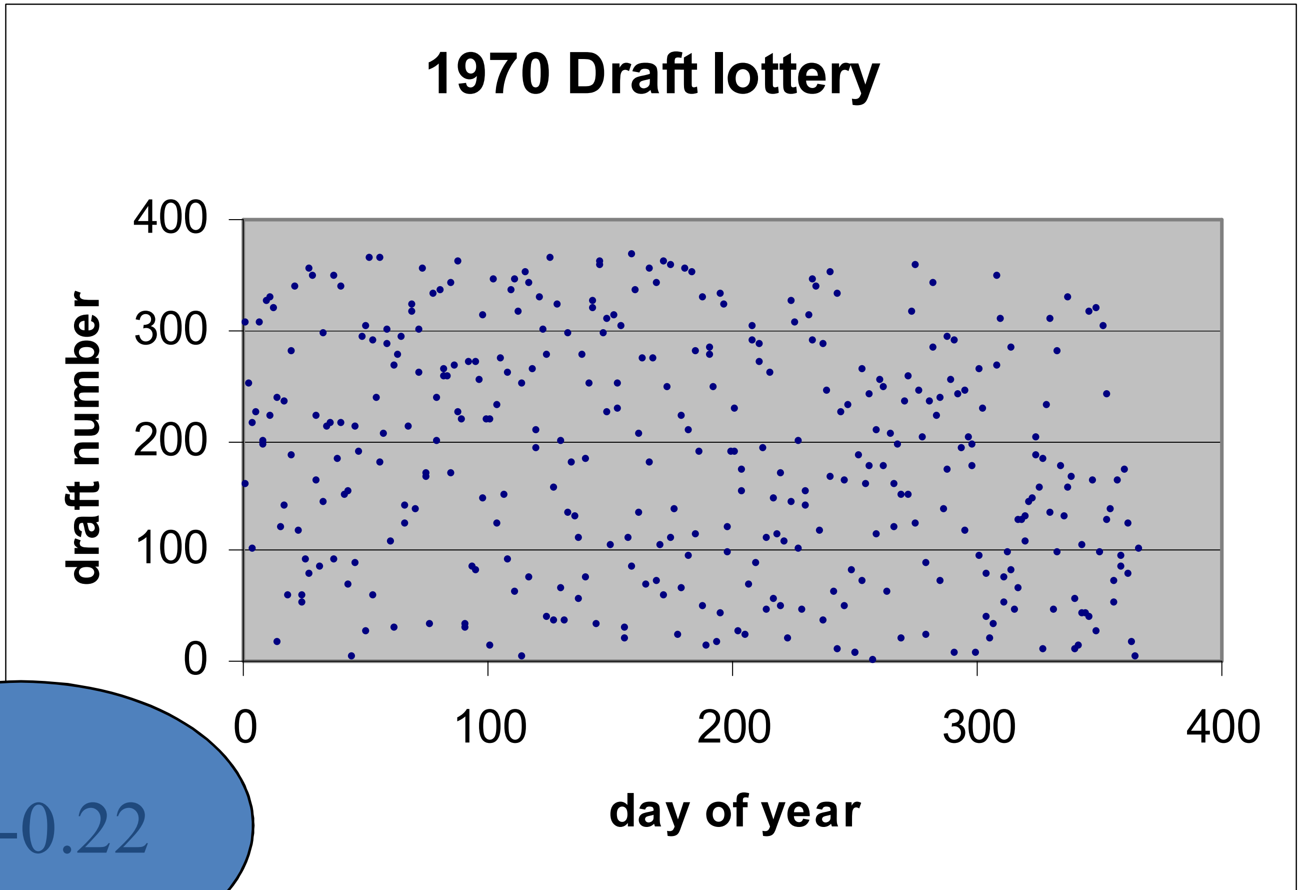
Live Session Unit 3

Future Plan

- May 23 : Live session 03- Sampling Distribution
- May 30 : Live session 04- Stratified Design
- Jun 06 : Live session 05- Sample size calculations
- Jun 13 : Live session 06 – Mid term review
- Jun 20 : Mid term

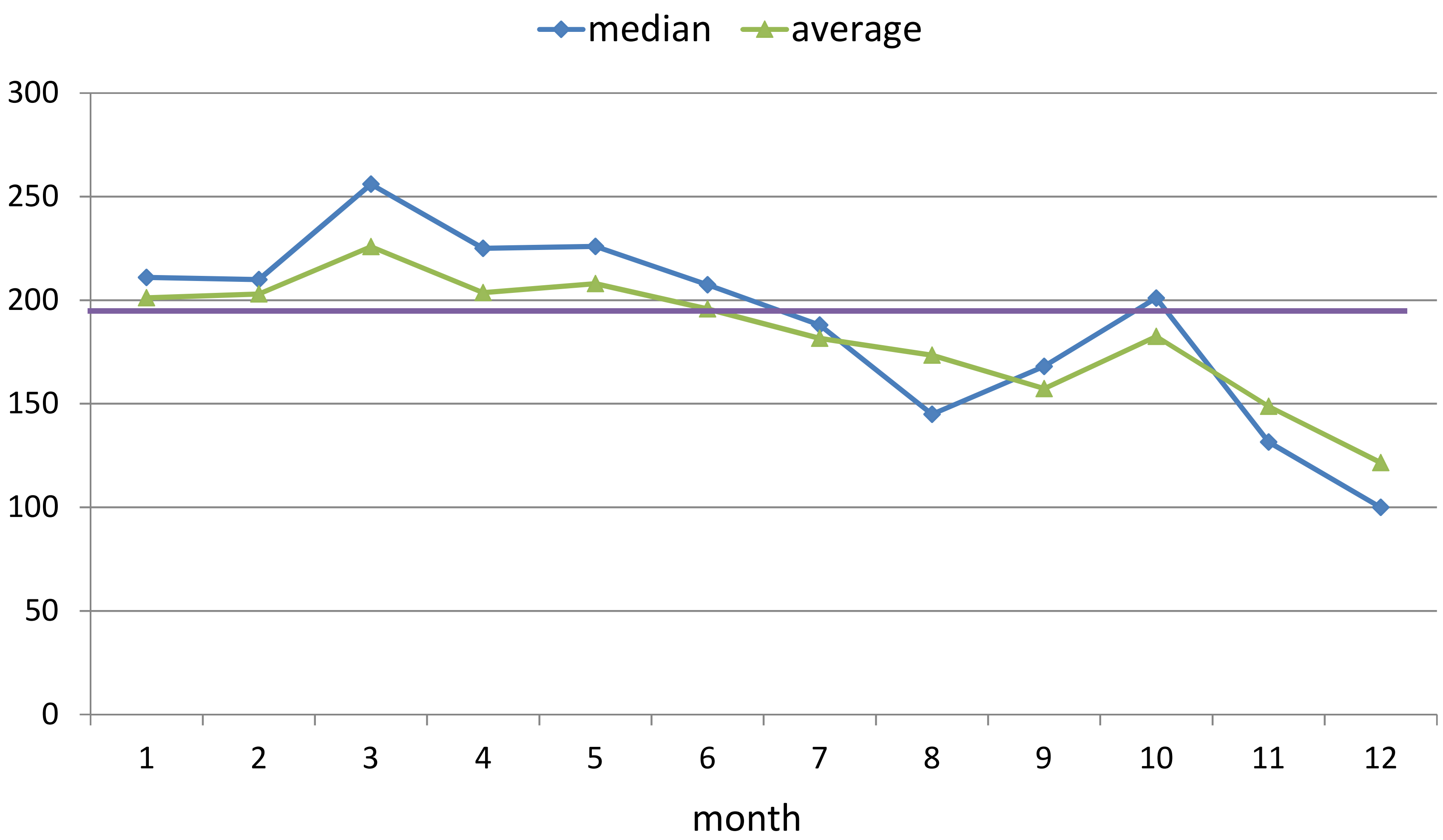
Lottery numbers used in 1970 & 1971 drafts

What is the correlation?

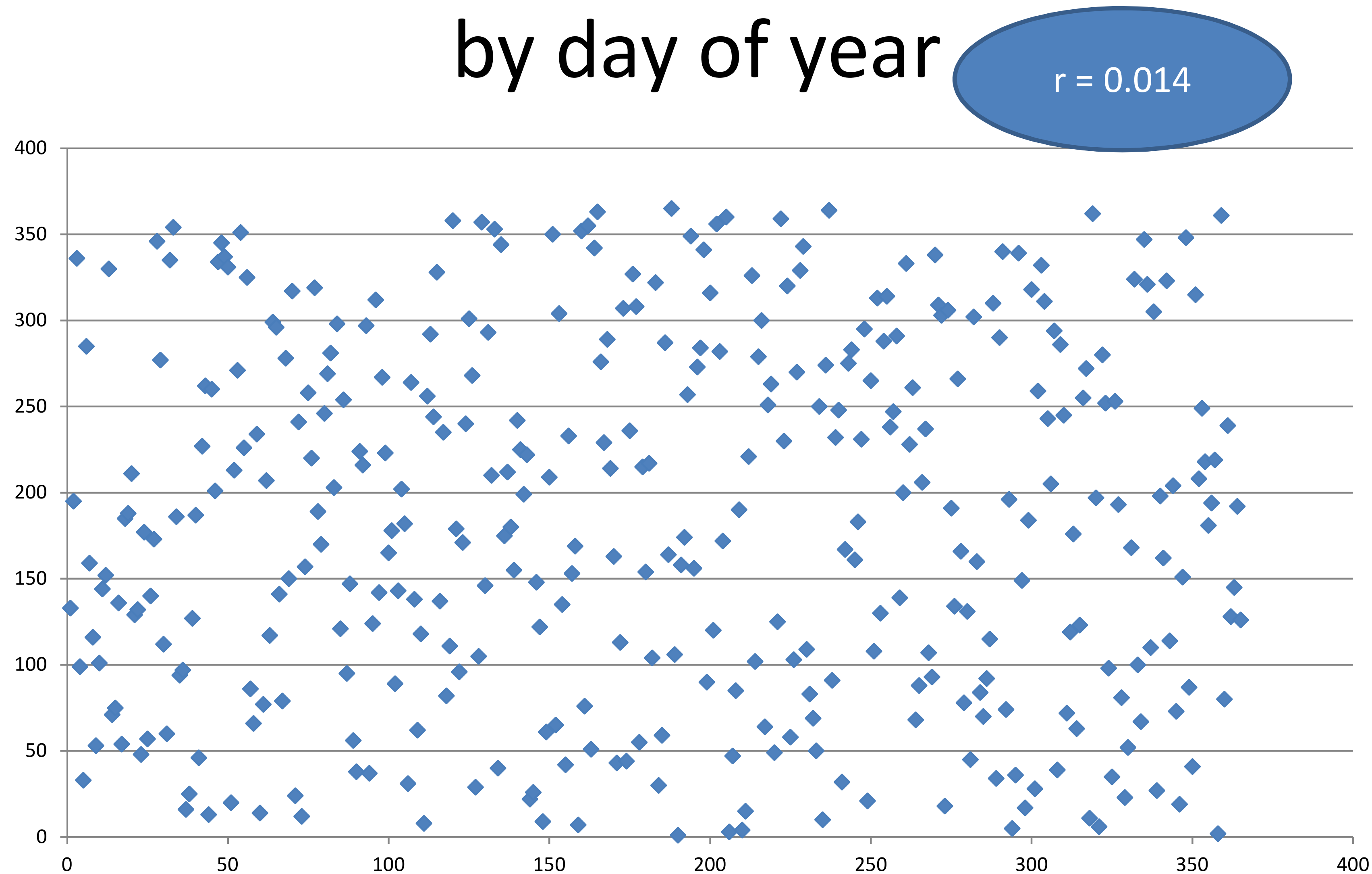


r = -0.22

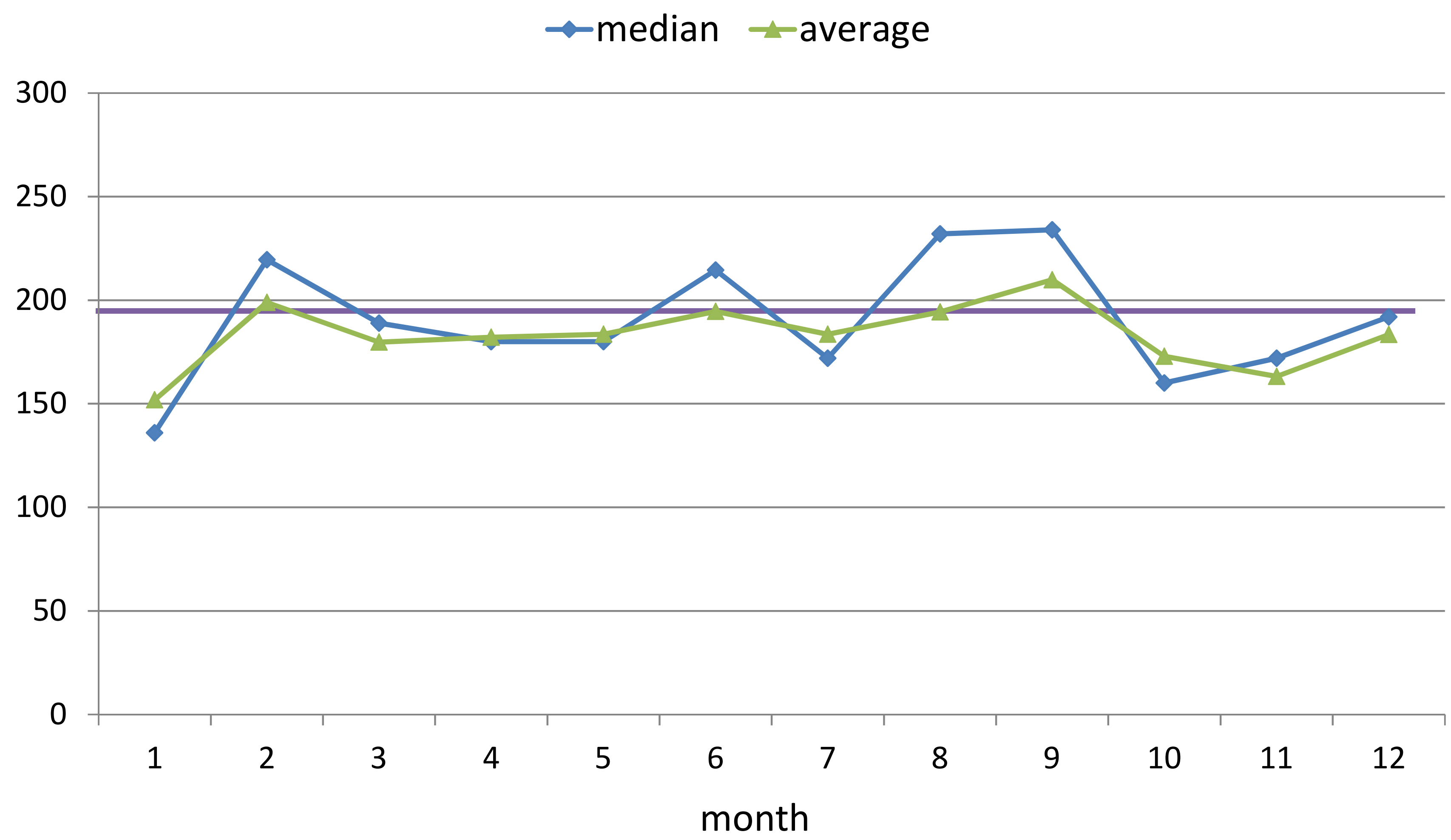
Lottery numbers used in 1970



Lottery numbers used in 1971 by day of year



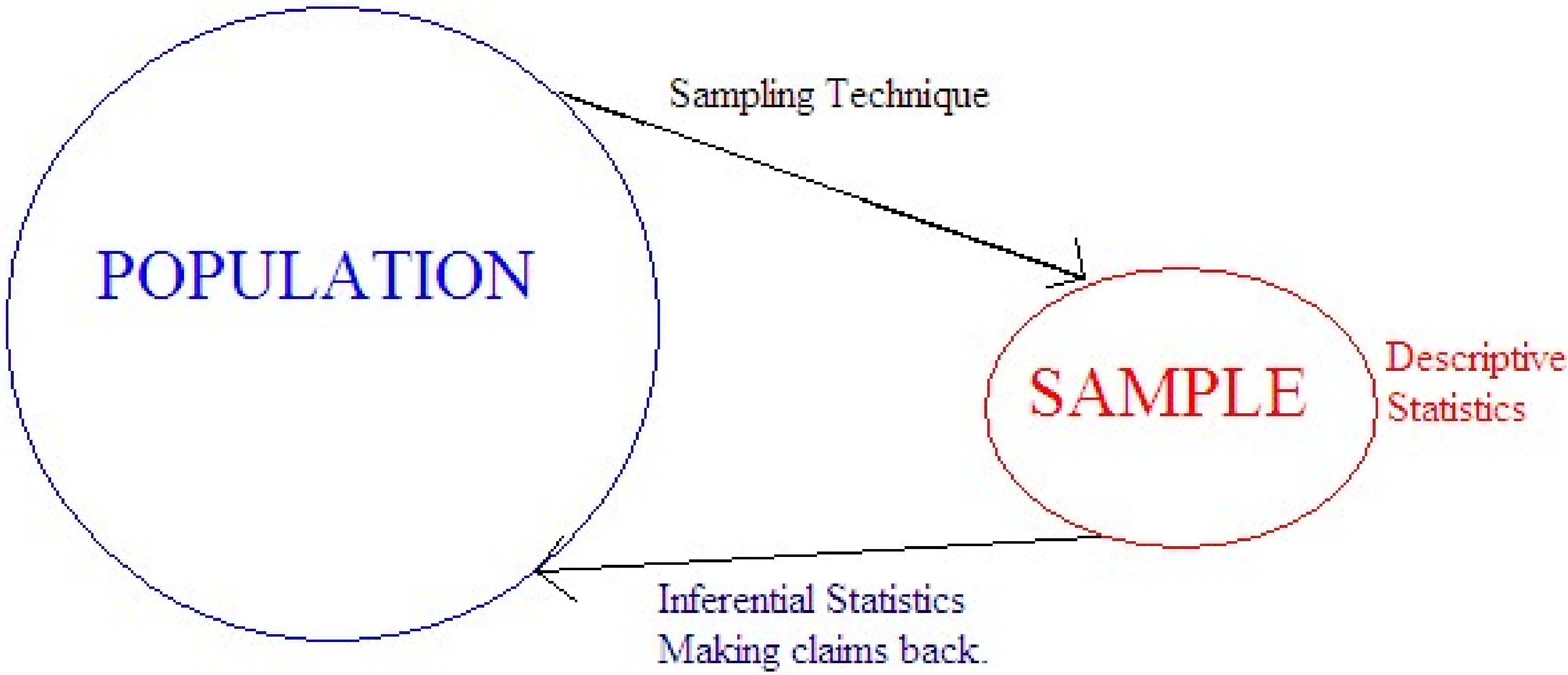
Lottery numbers used in 1971



Checking data for randomness

- No formula for checks that guarantees data are random
- A few basic checks for randomness
  - Construct a scatter plot and calculate correlation
  - Examine distributional properties (Lab 02)
    - Plot the means within the categories to see if they are approximately equal (assuming they should be).
    - Plot the medians within the categories of interest to see if they are approximately equal (assuming they should be)
      - Keep in mind that a median is robust to unusually large or unusually small observations while mean is not

Parameters and Statistics



Parameters and Statistics

A *statistic* is a characteristic or measure which uses the data values from a *sample.*

A *parameter* is a characteristic or measure which uses all of the data values from a specific *population*.

Parameters and Statistics (Mean)

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots\dots\dots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Statistic

$$\mu = \frac{x_1 + x_2 + x_3 + \dots\dots\dots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Parameter

Example

If we wanted to know the mean height of all SMU students, maybe we would find 100 SMU students and measure their heights.

These 100 students would be our sample.

The mean height of all SMU students would be a **parameter**.

The mean height of the 100 SMU students we measured would be a **statistic**.

Sampling Distribution

- Remember:
  - A *parameter* is a numerical index that describes some feature of a population (universe)
  - A *statistic* is a numerical index that describes some feature of the sample. It is a function of the data collected in a sample.
- The sampling distribution tells us how the statistic is related to the parameter.

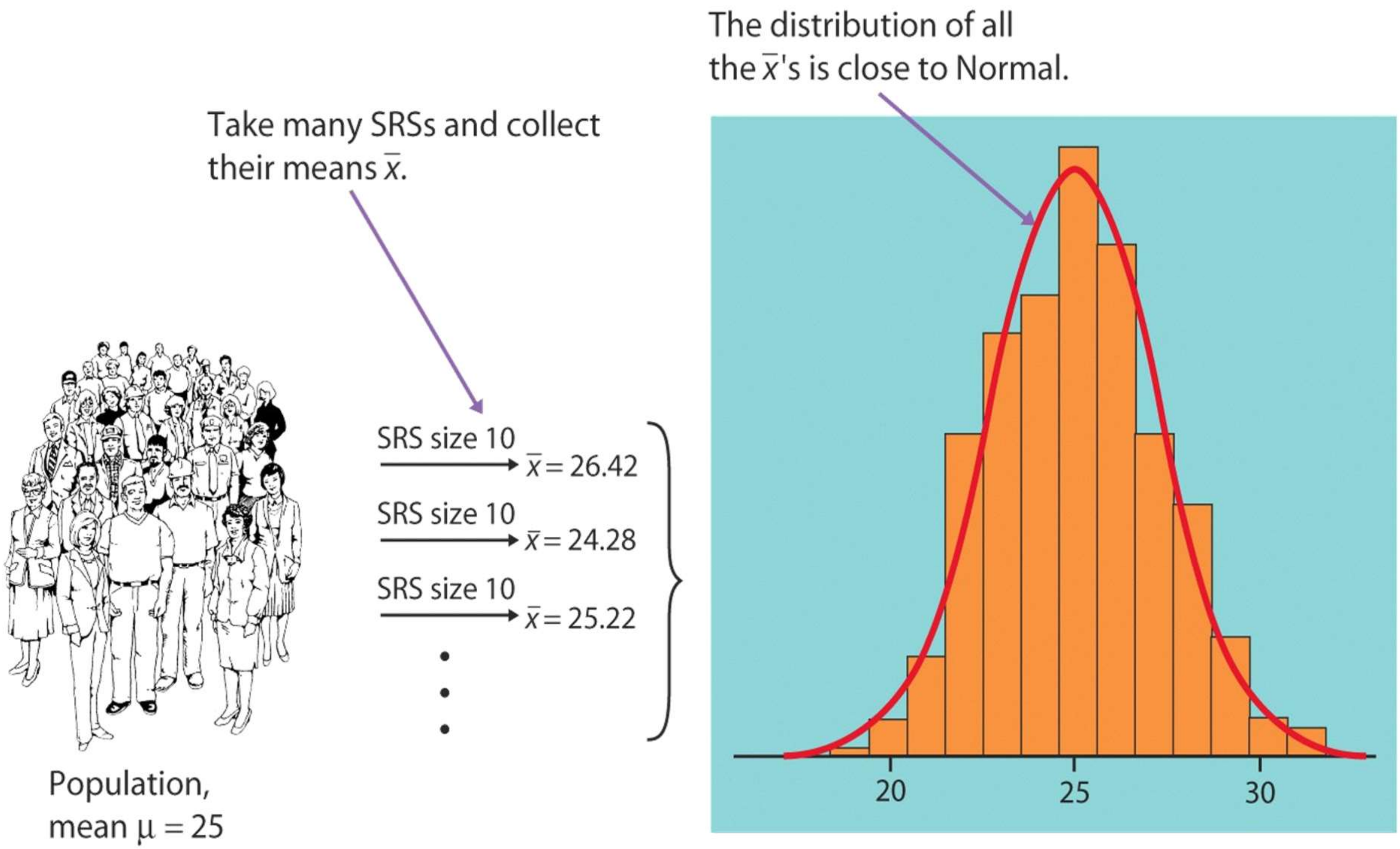
Sampling Distributions

Let's say we're interested in the mean of a population.

- We take a sample and find $\bar{x}$ and s
- What do this tell us about the population mean? Can we make a guess based on this?
- We need to know how sample means are distributed (in relation to the population distribution) to answer this.
- The distribution of $\bar{x}$ (all the sample means) is a <u>sampling distribution</u>

Sampling Distribution

- what would happen in many samples?



Notations

$\mu_{\bar{x}}$ The mean of the sample means.

$\sigma_{\bar{x}}$ The standard deviation of the sample means. (Standard error)

Sampling Distributions

This is a helpful site in understanding sampling distributions: http://onlinestatbook.com/stat_sim/sampling_dist/index.html

Original distribution			Sample Size	Distribution of Sample Means		
x Distribution	μ_x	σ_x	n	$\bar{x}$ Distribution	$\mu_{\bar{x}}$	$\sigma_{\bar{x}}$

Examples

X **Distribution**	μ_X	σ_X	**n**	$\bar{X}$ **Distribution**	$\mu_{\bar{X}}$	$\sigma_{\bar{X}}$
Normal	16	5	25	Normal	16	1
Normal	16	5	16	Normal	16	1.25
Normal	16	5	10	Normal	16	1.58
Uniform	16	9.52	25	Approx. Normal	16	1.90
Skewed	8.08	6.22	25	Approx. Normal	8.08	1.24

Mean

How are μ_X and $\mu_{\bar{X}}$ related?

Does it depend on n?

$$\boldsymbol{\mu_{\bar{X}} = \mu_X}$$

It doesn't matter what n is.

Standard Deviation

How are σ_x and $\sigma_{\bar{x}}$ related?

Does it depend on n?

$$\boldsymbol{\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}}$$

Their relationship does depend on n.

Central Limit Theorem

What we have just shown is the Central Limit Theorem!

Central Limit Theorem (CLT)

Say we take a **SRS** of size n from any population with mean μ and standard deviation σ.

If n is large enough, the sampling distribution of the sample mean is approximately normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

This is why the normal distribution is so important.

Central Limit Theorem

Original Population: x		Population of means, $\bar{x}$, of **all** samples
Any Shape		Approx. Normal Distribution
Mean: μ	Sample of size n $\rightarrow$	Mean: μ
Standard Deviation: σ		Standard Deviation: $\frac{\sigma}{\sqrt{n}}$

Example

Sample Size and the Standard Deviation

- The larger the sample size, the smaller the standard deviation of the $\bar{x}$

- As n increases, the standard deviation of the mean decreases

$$\boldsymbol{\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}}$$

Example

Population standard deviation ($\boldsymbol{\sigma_x}$) = 100

For n = 10 , $\boldsymbol{\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}}} = \boldsymbol{\frac{100}{\sqrt{10}} = 31.62}$

For n = 100 , $\boldsymbol{\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}}} = \boldsymbol{\frac{100}{\sqrt{100}} = 10.00}$

For n = 1000 , $\boldsymbol{\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}}} = \boldsymbol{\frac{100}{\sqrt{1000}} = 3.16}$

Example population of students

Student	Hrs completed at SMU	Born in US? (Y/N)
Ali	27	Y
Bucky	55	N
Judith	24	Y
Hal	100	N
Roy	18	Y
Gideon	60	N
John	0	N
Yusun	21	N
	mu = 38.125	Prop of Y = 0.375
	S = 31.77347	

SRSWOR of Size 2

- To find the sampling distribution directly, I must find every possible sample of size 2, all of which are equally likely:
- Ali and Bucky
- Ali and Judith
- Ali and Hal
- Ali and Roy
- Ali and Gideon
- Ali and John
- Ali and Yusun
- Bucky and Judith
- Bucky and Hal
- Etc.

Student
Ali
Bucky
Judith
Hal
Roy
Gideon
John
Yusun

HOW MANY?

Exercise

- How many SRSWOR of size 2 are there from a population of size N = 8 are possible?

a) 64

b) 56

c) 28

d) 23

$$\binom{N}{n} = \frac{N!}{(N-n)!\,n!}$$

$$\binom{8}{2} = \frac{8!}{(6)!2!} = \frac{56}{2} = 28$$

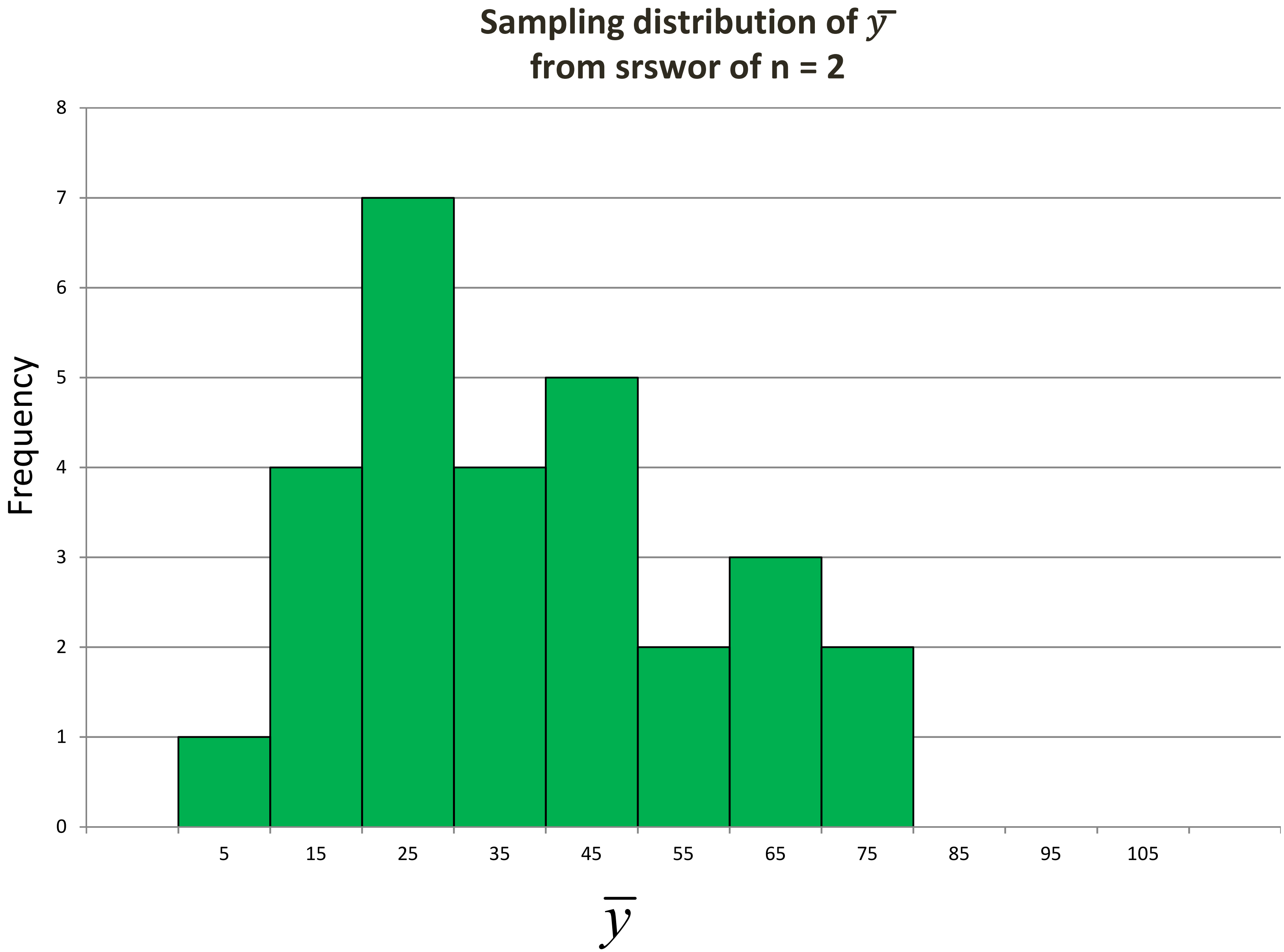
Sampling Distribution

			All possible sample means for sample of size 2, each having probability 1/(N choose n)						
	Ali	Bucky	Judith	Hal	Roy	Gideon	John	Yusun	Population
Ali		41	25.5	63.5	22.5	43.5	13.5	24	27
Bucky			39.5	77.5	36.5	57.5	27.5	38	55
Judith				62	21	42	12	22.5	24
Hal					59	80	50	60.5	100
Roy						39	9	19.5	18
Gideon							30	40.5	60
John								10.5	0
Yusun									21
Population	27	55	24	100	18	60	0	21	

$\bar{y}$	9	10.5	12	13.5	19.5	21	22.5	…	80
Prob	1/28	1/28	1/28	1/28	1/28	1/28	2/28	…	1/28

$\bar{y}$
9
10.5
12
13.5
19.5
21
22.5
22.5
24
25.5
27.5
30
36.5
38
39
39.5
40.5
41
42
43.5
50
57.5
59
60.5
62
63.5
77.5
80

Sampling Distribution



MEAN OF A DISCRETE RANDOM VARIABLE

Suppose that X is a discrete random variable whose distribution is

Value of X	x_1	x_2	x_3	$\cdots$	x_k
Probability	p_1	p_2	p_3	$\cdots$	p_k

To find the **mean** of X, multiply each possible value by its probability, then add all the products:

$$\begin{aligned}\mu_X &= x_1p_1 + x_2p_2 + \cdots + x_kp_k \\ &= \sum x_ip_i\end{aligned}$$

VARIANCE OF A DISCRETE RANDOM VARIABLE

Suppose that X is a discrete random variable whose distribution is

Value of X	x_1	x_2	x_3	$\cdots$	x_k
Probability	p_1	p_2	p_3	$\cdots$	p_k

and that μ is the mean of X. The **variance** of X

$$\begin{aligned}\sigma_X^2 &= (x_1 - \mu_X)^2 p_1 + (x_2 - \mu_X)^2 p_2 + \cdots + (x_k - \mu_X)^2 p_k \\ &= \sum (x_i - \mu_X)^2 p_i\end{aligned}$$

The **standard deviation** σ_X of X is the square root of the variance.

Find mean and standard deviation for our example

$\bar{y}_i$	9	10.5	12	13.5	19.5	21	22.5	…	80
p_i	1/28	1/28	1/28	1/28	1/28	1/28	2/28	…	1/28

$$\mu_{\bar{y}} = \text{mean of } \bar{y} = \sum \bar{y}_i p_i = \sum \left[9 * \left(\frac{1}{28}\right) + 10.5 * \left(\frac{1}{28}\right) + \cdots + 80 * \left(\frac{1}{28}\right)\right] = 38.125$$

$$\sigma_{\bar{y}} = \text{sd of } \bar{y} = \sqrt{\sum (\bar{y}_i - \mu_{\bar{y}})^2 p_i}$$

$$= \sqrt{\sum \left[(9 - 38.125)^2 * \left(\frac{1}{28}\right) + \cdots + (80 - 38.125)^2 * \left(\frac{1}{28}\right)\right]} = 19.4572$$

But that's a lot of work!

CLT for sample mean for SRSWOR

- The sampling distribution of the sample mean $\bar{y}$ from the SRSWOR of size n from population of size N has a mean

$$\mu_{\bar{y}} = \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

- The sampling distribution of the sample mean $\bar{y}$ from the SRSWOR has a standard deviation

$$\sigma_{\bar{y}} = \sqrt{\frac{S^2}{n}\left(1-\frac{n}{N}\right)},$$

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

Population
27
55
24
100
18
60
0
21

- The sampling distribution is approximately normal for large samples AND large populations.

$\bar{Y} =$	38.125
$\sqrt{\frac{S^2}{n}\left(1-\frac{n}{N}\right)} =$	19.4572

Note this is the same as we calculated

Example population of students

Student	Hrs completed at SMU	Born in US? (Y/N)
Ali	27	Y
Bucky	55	N
Judith	24	Y
Hal	100	N
Roy	18	Y
Gideon	60	N
John	0	N
Yusun	21	N
	mu = 38.125	Prop of Y = 0.375
	S = 31.77347	

Values of $\hat{p}$	a	b	c
Prob of each value	d	e	f

Estimate of the proportion of students who are native born, from a SRSWOR of size 2

			All possible sample means for sample of size 2, each having probability 1/(N choose n)						
	Ali	Judith	Roy	Bucky	Hal	Gideon	John	Yusun	Population
Ali		1	1	0.5	0.5	0.5	0.5	0.5	1
Judith			1	0.5	0.5	0.5	0.5	0.5	1
Roy				0.5	0.5	0.5	0.5	0.5	1
Bucky					0	0	0	0	0
Hal						0	0	0	0
Gideon							0	0	0
John								0	0
Yusun									0
Population	1	1	1	0	0	0	0	0	

Values of $\hat{p}$	**0**	**0.5**	**1**
Prob of each value	10/28	15/28	3/28

Similar CLT for sample proportion for SRSWOR

- The sampling distribution of the sample proportion $\hat{p}$ from the srswor of size n from population of size N has a mean

$$\mu_{\hat{p}} = p = \frac{1}{N}\sum_{i=1}^{N} y_i \quad \text{when } y_i \text{ is } 0/1$$

- The sampling distribution of the sample proportion $\hat{p}$ from the SRSWOR has a mean of p and a standard deviation

$$\sigma_{\hat{p}} = \sqrt{\frac{S^2}{n}\left(1 - \frac{n}{N}\right)},$$

where $S^2 = \frac{Np(1-p)}{N-1}$

- The sampling distribution is approximately normal for large samples AND large populations.

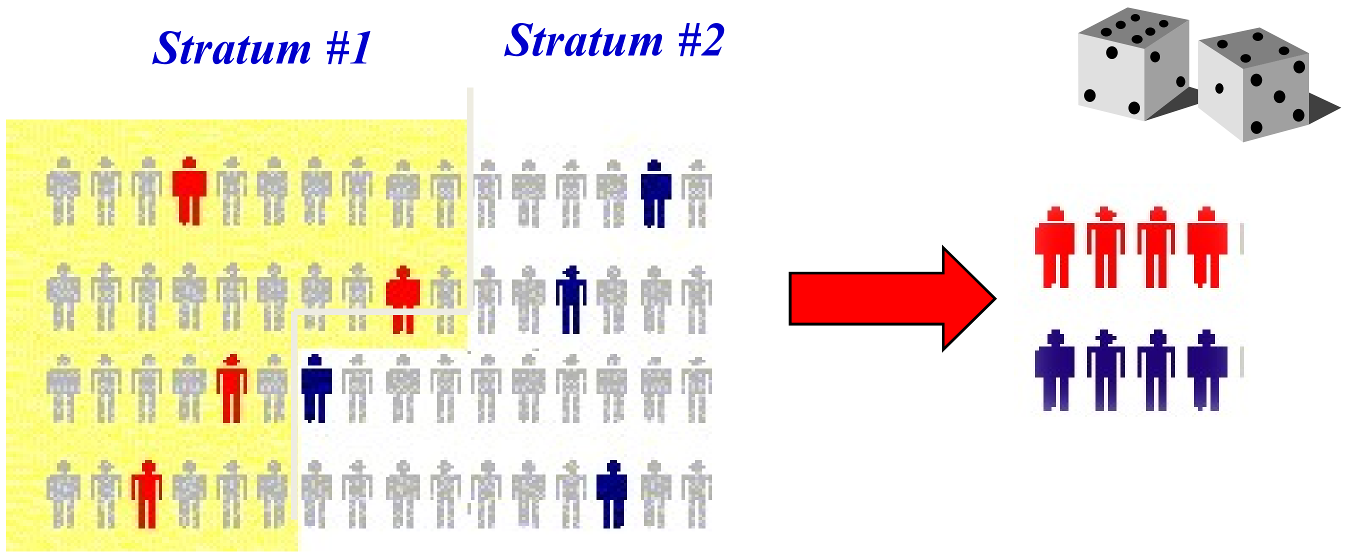
Stratified Random Sample

Guarantees Adequate Representation from Each of Several Groups

- Divide a Population into Mutually Exclusive and Exhaustive Strata (Groups)
  - Mutually Exclusive: Each Item Belongs to One (Exhaustive) and Only One (Mutually Exclusive) Stratum
- Take a Simple Random Sample from Each Stratum, Proportional to the Representation in the Population

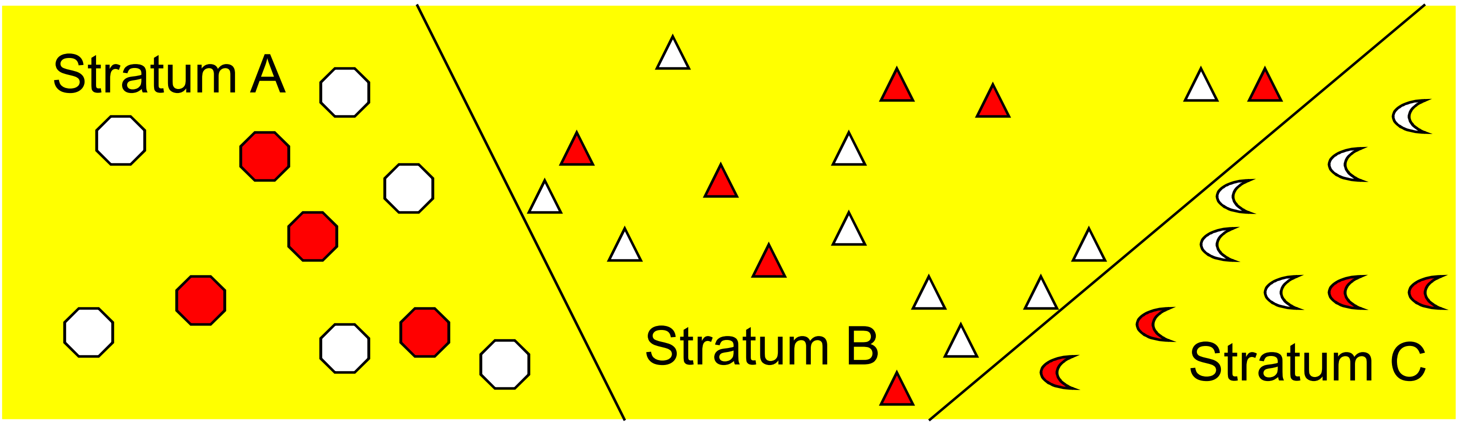
Stratified Samples

- Population divided into two or more groups according to some common characteristic
- Simple random sample selected from each group
- The two or more samples are combined into one



Stratified Sampling Details

- Units within stratum are similar
- Units in stratum A are different from units in stratum B and stratum C
- Use similarity within each stratum to obtain more precise information about population



Note: Symbols Similar in Each Stratum
Different Colors Represent Different Responses

Stratified Design

- Divide the sampling frame into groups based on stratifying variable(s)
- Select a simple random sample from each stratum
- Put the samples together to estimate

$$\bar{y}_{str} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) \bar{y}_h$$

and

$$\hat{p}_{str} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) \hat{p}_h$$

Stratified Design 01

- Same data with one additional variable

Student	Hrs completed at SMU	Gender
Ali	27	M
Bucky	55	M
Judith	24	F
Hal	100	F
Roy	18	M
Gideon	60	M
John	0	M
Yusun	21	M

- This time we choose a sample by selecting 1 male (at random) and 1 female (at random)

- Because there are 2 females and 6 males, we have 12 possible samples

- The male "represents" 5 other unseen males; the female 1 unseen female. Thus we estimate by $\bar{y}_{str} = \left(\frac{2}{8}\right)\bar{y}_f + \left(\frac{6}{8}\right)\bar{y}_m$

Sampling distribution of $\bar{y}_{str}$

- Does this work better for estimating mean hrs complete?
- To see, we will enumerate the sampling distribution

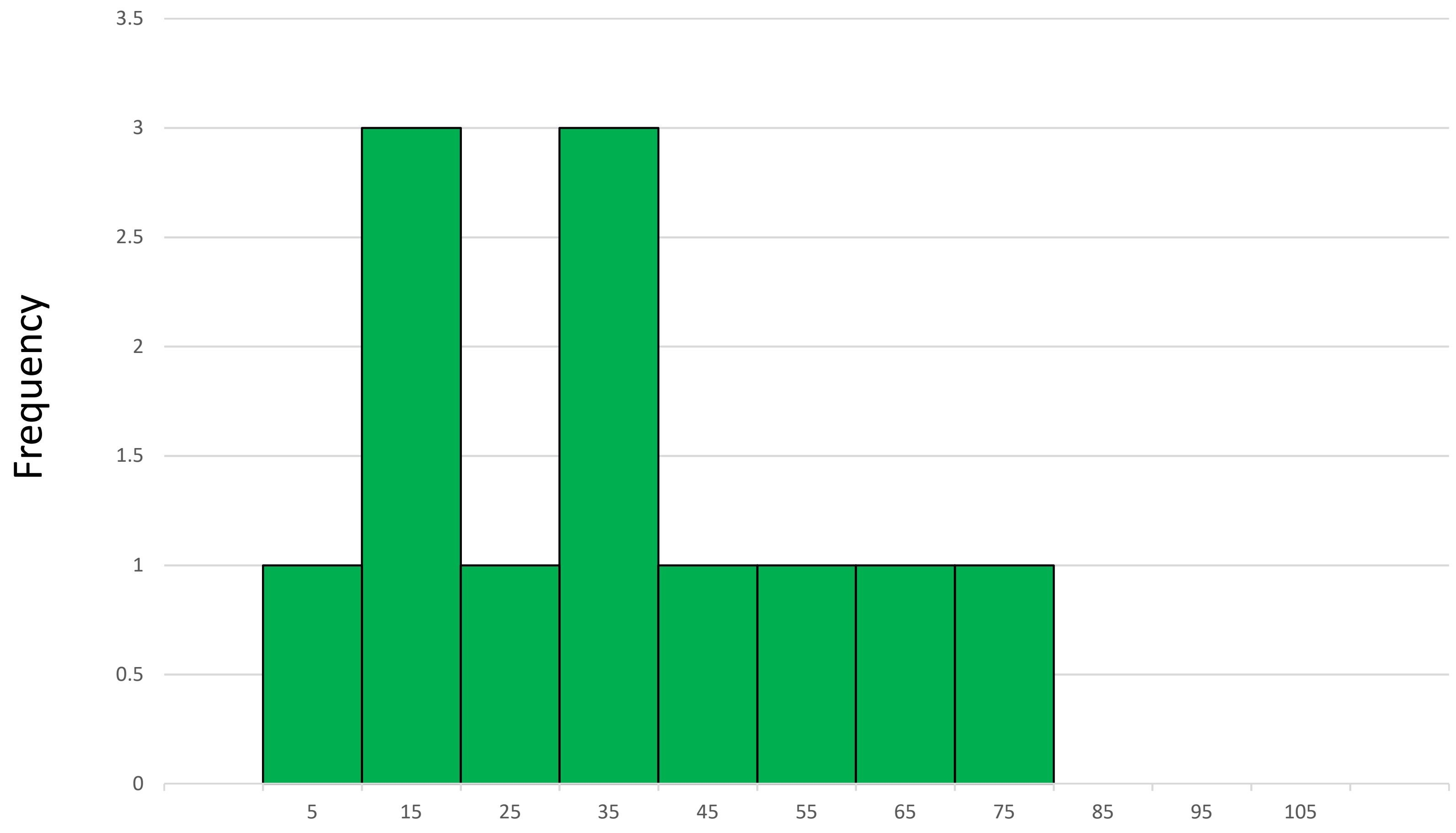
	Judith	Hal	Population
Ali	26.25	45.25	27
Bucky	47.25	66.25	55
Roy	19.5	38.5	18
Gideon	51	70	60
John	6	25	0
Yusun	21.75		21
Population	24	100	

$\frac{6}{8}*27 + \frac{2}{8}*24$

$\bar{y}_{new,i}$	6	19.5	21.75	…	70
p_i	1/12	1/12	1/12	…	1/12

Sampling distribution

Sampling distribution of $\bar{y}_{str}$ from new design with sample of size 2



$\bar{y}_{str}$

Mean and sd of new design's sampling distribution

$\bar{y}_{new,i}$	6	19.5	21.75	…	70
p_i	1/12	1/12	1/12	…	1/12

$$\mu_{\bar{y}_{str}} = \text{mean of } \bar{y}_{str} = \sum \bar{y}_i p_i = \sum \left[6 * \left(\frac{1}{12}\right) + \cdots + 70 * \left(\frac{1}{12}\right) \right] = 38.125$$

Yea! Still unbiased!

$$\sigma_{\bar{y}_{str}} = \text{sd of } \bar{y}_{str} = \sqrt{\sum (\bar{y}_i - \mu_{\bar{y}_{str}})^2 p_i}$$

$$= \sqrt{\sum \left[(6 - 38.125)^2 * \left(\frac{1}{12}\right) + \cdots + (70 - 38.125)^2 * \left(\frac{1}{12}\right) \right]} = 18.426$$

$< \sigma_{\bar{y}}$= 19.4572

So this design is slightly better than the SRSWOR because the variability is reduced

Stratified Design 02

- Another variable for stratification. Does it work better?

Student	Hrs completed at SMU	Gender	Grad/ Undergrad
Ali	27	M	G
Bucky	55	M	U
Judith	24	F	G
Hal	100	F	U
Roy	18	M	G
Gideon	60	M	U
John	0	M	G
Yusun	21	M	G

Step 1. Stratify by graduate status; select 1 student from each stratum
Step 2. Enumerate all possible samples (hint: there are 15)
Step 3. Calculate $\bar{y}_{str}$ for every sample
Step 4. Compare it to the other two designs we have examined.

Sampling distribution of $\bar{y}_{str}$

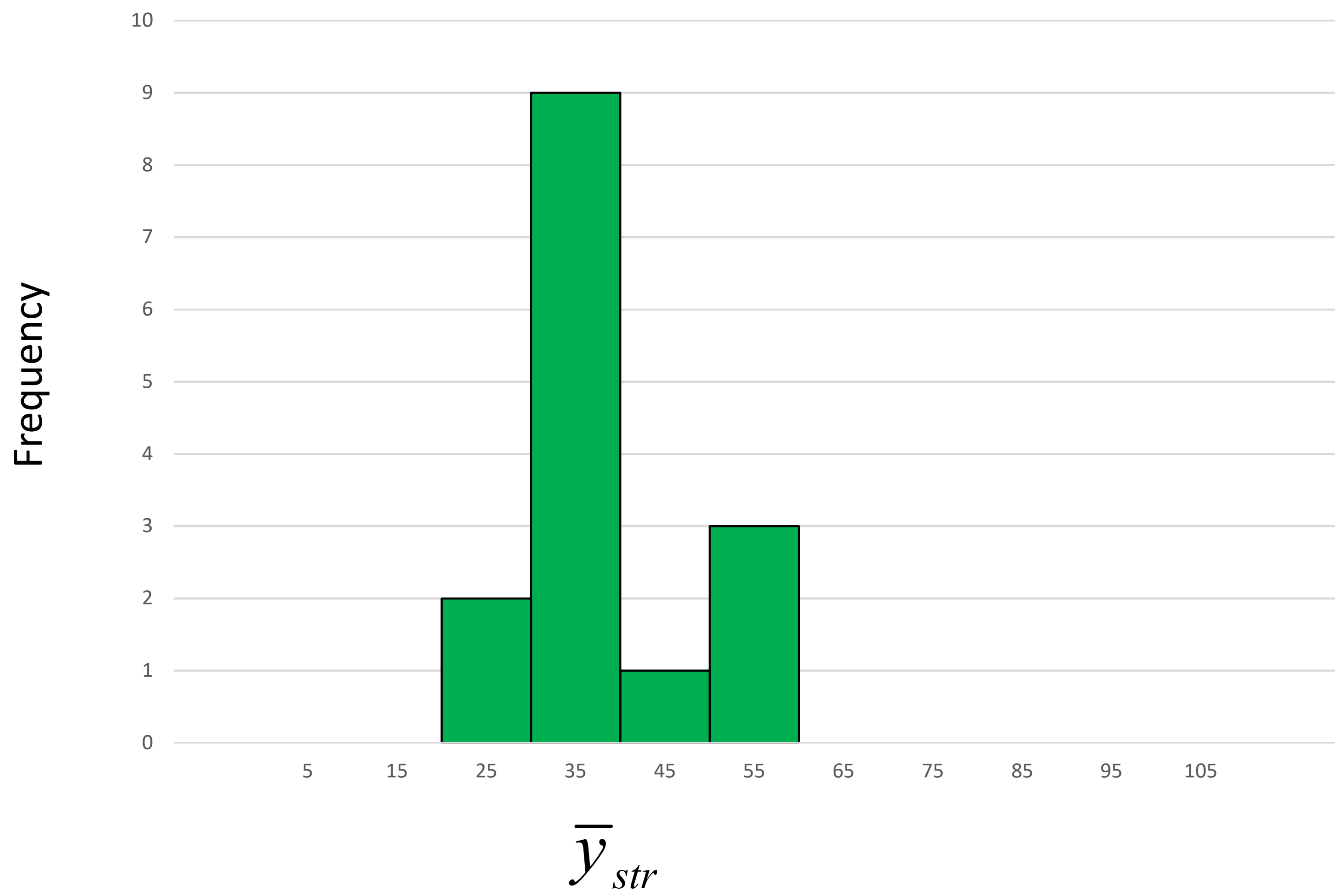
	Bucky	Hal	Gideon	Population
Ali	37.5	54.375	39.375	27
Judith	35.625	52.5	37.5	24
Roy	31.875	48.75	33.75	18
John	20.625	37.5	22.5	0
Yusun	33.75		35.625	21
Population	55	100	60	

$\frac{5}{8}$*27 + $\frac{3}{8}$*55

$\bar{y}_{str,i}$	**20.625**	**22.5**	**31.875**	**…**	**54.375**
p_i	1/15	1/15	1/15	…	1/15

Sampling distribution

Sampling distribution of $\bar{y}_{str}$ from stratified (G/U) with sample of size 2



Summary of sampling distribution

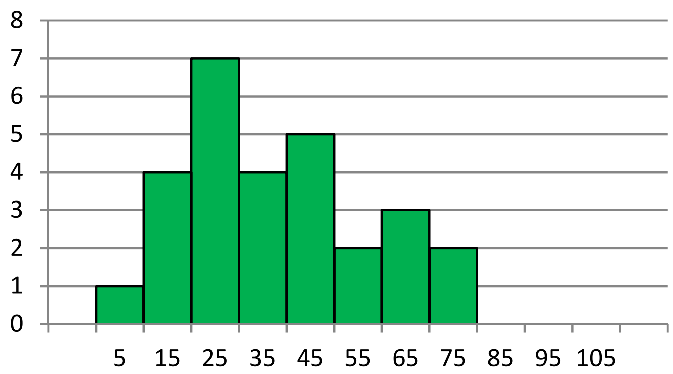
$\bar{y}_{str,i}$	**20.625**	**22.5**	**31.875**	**…**	**54.375**
p_i	1/15	1/15	1/15	…	1/15

$$\mu_{\bar{y}_{str2}} = \sum \left[20.625 * \left(\frac{1}{15} \right) + \cdots + 54.375 * \left(\frac{1}{15} \right) \right] = 38.125$$

$$\sigma_{\bar{y}_{str2}} = \sqrt{\sum \left[(20.625 - 38.125)^2 * \left(\frac{1}{15} \right) + \cdots + (54.375 - 38.125)^2 * \left(\frac{1}{15} \right) \right]} = 9.60$$

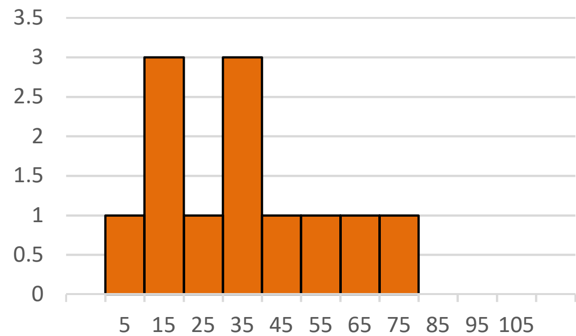
Comparing our three designs

SRSWOR



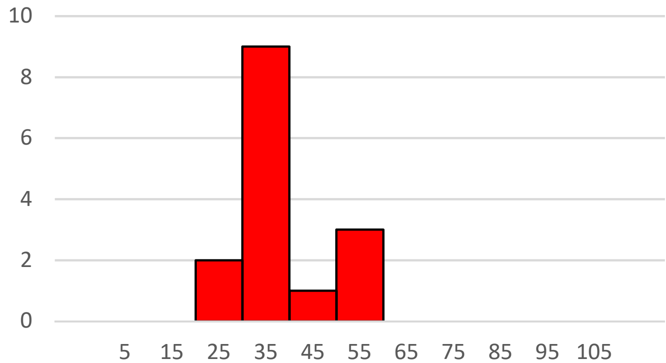
Mean = 38.125
SD = 19.45

Stratified M/F



Mean = 38.125
SD = 18.42

Stratified G/U



Mean = 38.125
SD = 9.60

Choice of stratifying variable

- Different methods of stratifying yield different results.
- The goal of stratification is to ensure representation from all types of units in the population.
- Stratification reduces the chance of “bad luck samples” of all similar units that SRS can produce.
- Make the strata as different from each other as possible.
- Because G/U status separated the students by # of hours completed better than gender did, it was a more effective stratifying variable.

Summary

- The sampling distribution of a statistic consists of its possible values, along with the probability that each occurs.
- There is a CLT that describes the approximate sampling distribution for estimators of means from sample designs from finite populations.
- The sampling distribution for a sample mean from a SRS is similar to that for a sample mean from an infinite populations except that the variance of the estimator is smaller by a multiplicative factor:

$$fpc = 1 - \frac{n}{N}$$

- The sampling distribution of a statistic from a small population can be obtained by complete enumeration of the samples and the statistics they produce.
- A stratified design can produce an estimator of means with smaller variance than a SRS.