The Data Driven Manager

1

Fundamentals of Sampling

Learning Objectives

- Characterize different types of sampling
- Use a sample to describe a population
- Maximize the probability that samples are an accurate representation of the population
- Create a vector of random numbers

Learning Objectives

- Describe the concept of sampling error
- Explore the concept of random sampling distributions in RStudio
- Describe the Central Limit Theorem
- Estimate probability using the Random Sampling Distribution of the mean

Acquiring Data

How to Acquire Data

- One of the primary reasons for collecting data is to be able make statements and draw conclusions about a particular population (or populations).
- Researchers always use samples to do this. Dealing with an entire population is an onerous ordeal.
- So, what about **populations** and **samples...**?

Population vs Sample

Population: Group of all items possessing a common characteristic of interest to a researcher

Population vs Sample

Population: Group of **all** items possessing a common characteristic of interest to a researcher

Sample: A representative **portion** of a population that is used to reach conclusions about the Population it represents

Populations & Samples

Population (**Target Population**)

- The entire group of objects, all with one characteristic of interest in common, and about which we want to make decisions
- Infinite, or finite but relatively huge

Populations & Samples

Research Population

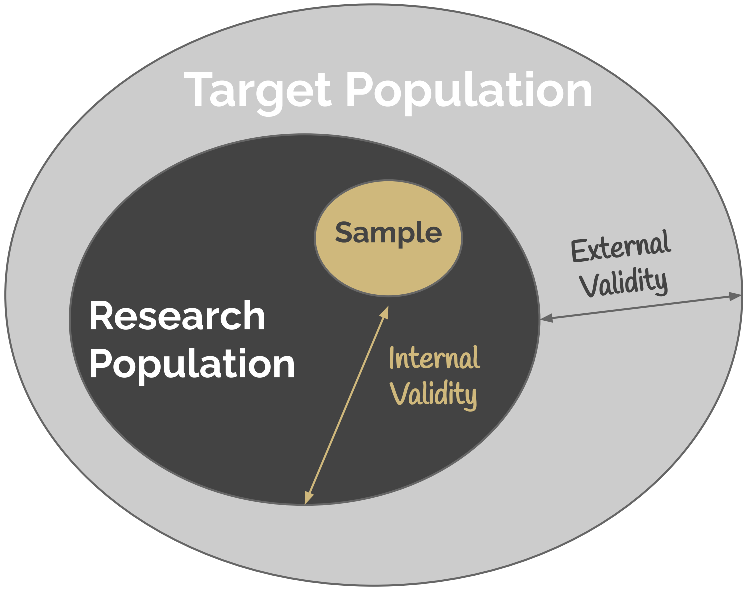
- That portion of the Target Population available for sampling

Populations & Samples

Sample

- A subgroup of the population of interest, usually selected randomly.
- Random sampling is a prerequisite to using any type of inferential statistics!

Population Definitions



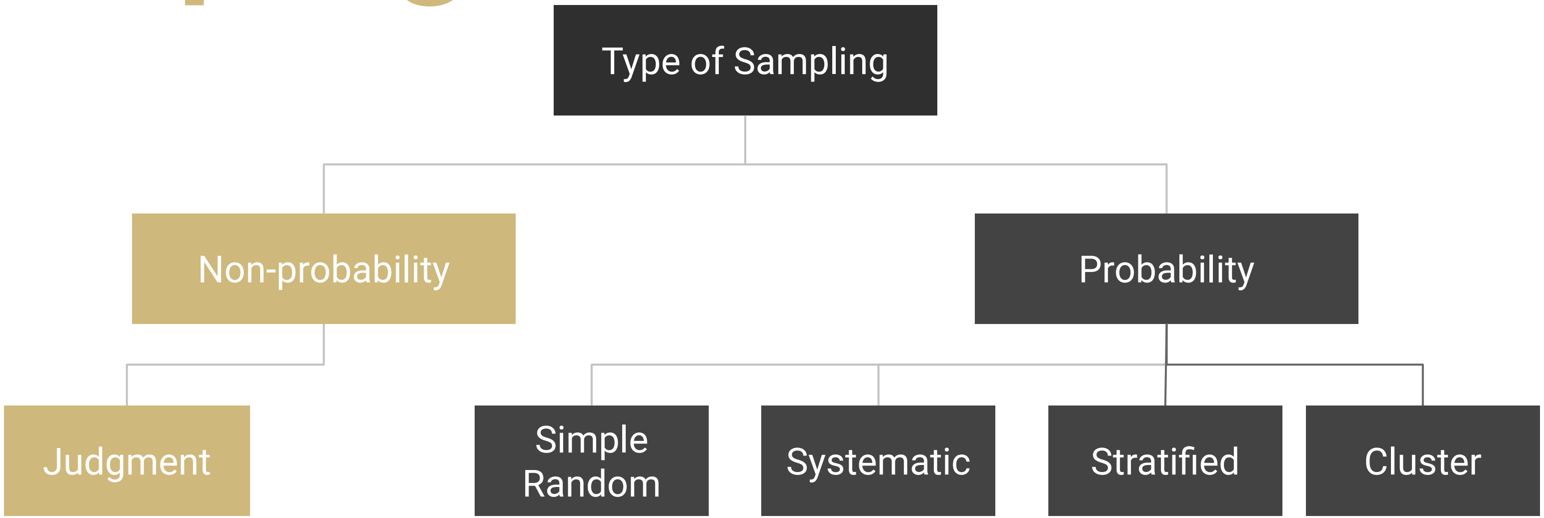
Why study Samples to understand Populations?

- Easier and more practical than studying the whole population
- Costs less
- Takes less time
- Sometimes testing involves risk
- Sometimes testing requires the destruction of the item being studied (the whole population would be destroyed)
- Not necessary (You don't give the doctor all of your blood for a blood test)

Types of Sampling

- Because sampling is so important we need understand the various types.
- There are many types of sampling, each have their **strengths** and their **weaknesses**.
- The violation of proper sampling (representation in particular) methods can seriously impact any research project, even render it useless.

Types of Sampling



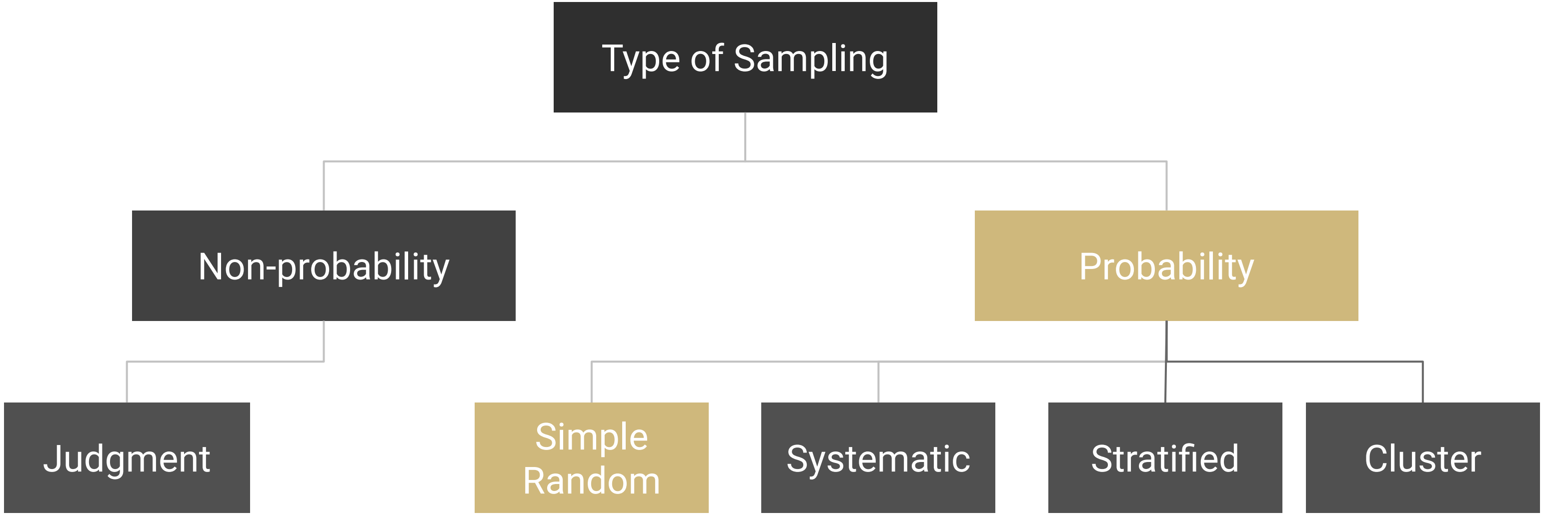
Nonrandom or Judgment Sampling

- Specimens or items are selected using personal judgment, reasoning, opinion, or convenience

Random or Probability Sampling

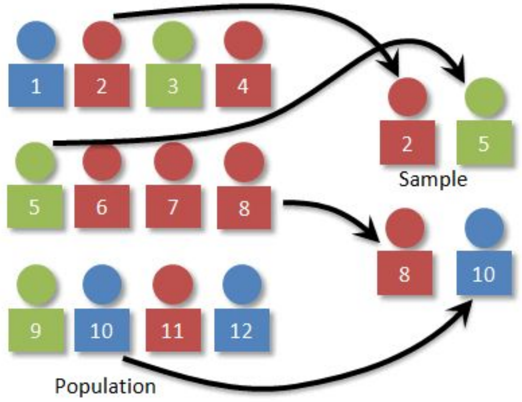
- All specimens or items have a probability of being included in the sample

Types of Sampling



Simple Random Sampling

- Each sample of size n has an equal chance of being selected



Simple Random Sampling

- Selecting 1 subject does not affect selecting others (helps to ensure independence)
- May use random numbers to identify which individual items are sampled (hence it is called simple random sampling)
- Foundation for statistical inference

Random Number Generation

https://www.random.org/sequences/

Generating Random Numbers and Sequences in R

Review file '**Random Numbers.R**'

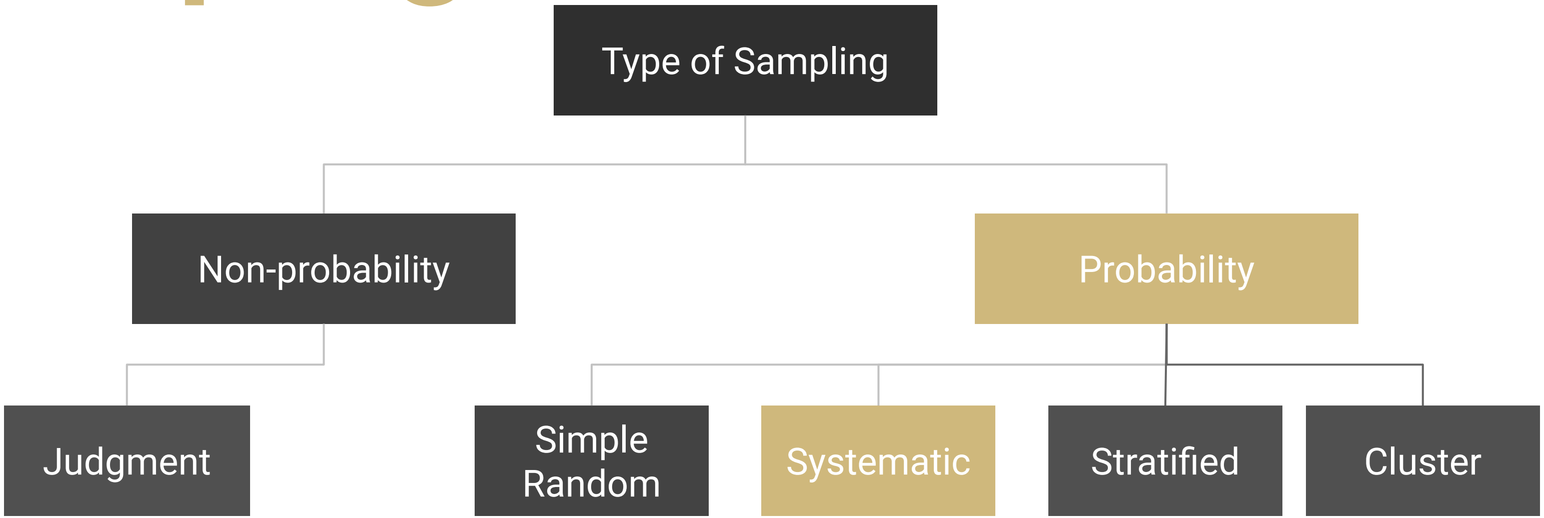
```
# Create a random sequence using numbers 1-10
# Create a sequence
> x<-seq(from = 1, to = 10, by = 1)

# Sample from the sequence without replacement
> sample(x = x, size = 10, replace = FALSE)
```

Generating Random Numbers and Sequences in R

Review file '**Random Numbers.R**'

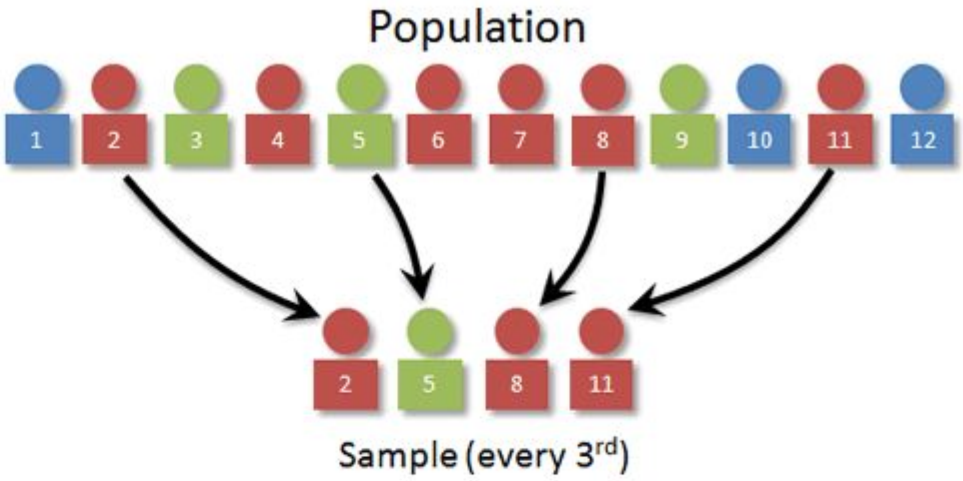
```
> install.packages('random')
> require(random)
> randomSequence(min=X, max=Y, col=Z)

# Create a sequence 1-10 in 2 columns
> randomSequence(min=1, max=10, col=2)
```

Types of Sampling



Systematic Random Sampling

- Specimens or items are selected at an interval



Systematic Random Sampling

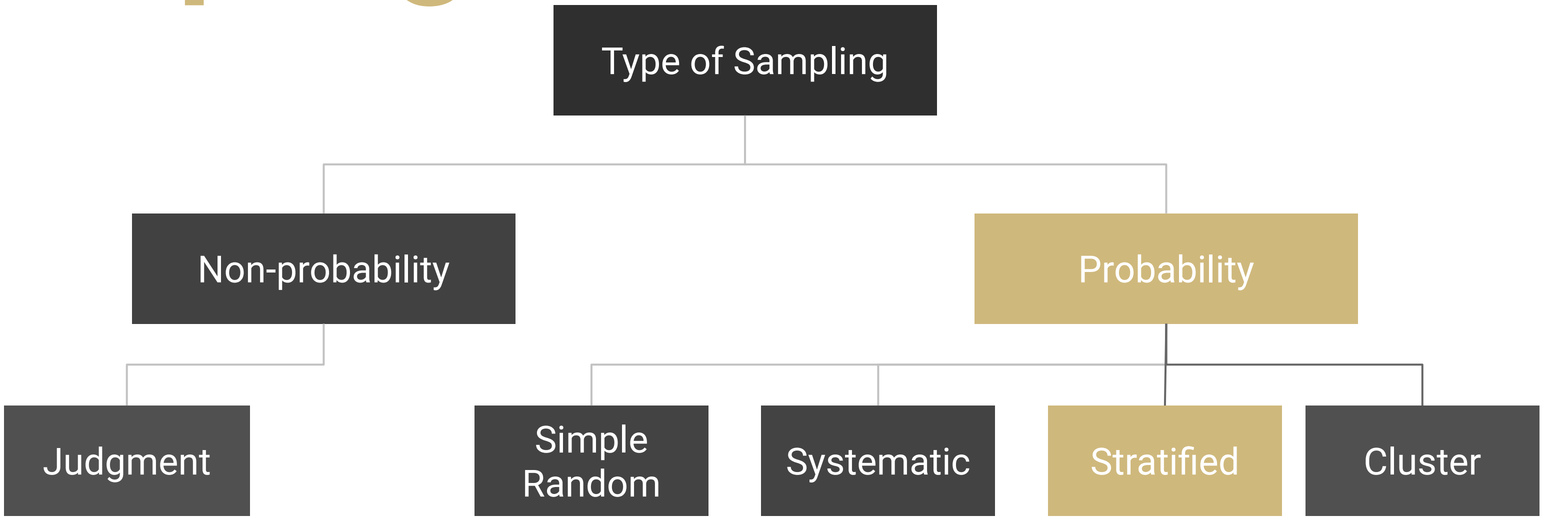
Advantages and Disadvantages

- Shortcoming: Problem of introducing bias into the sampling process

  e.g., Sample trash of 100 households every Monday (Watch out for "strata")

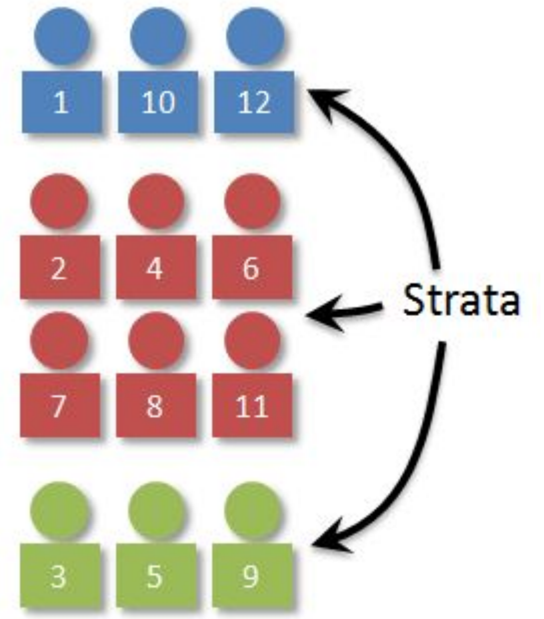
- Advantage: Requires less time and sometimes results in lower cost

Types of Sampling



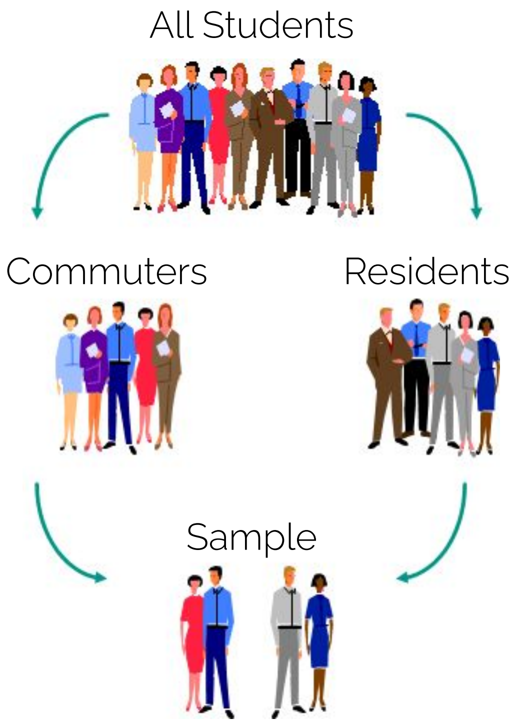
Stratified Random Sampling

- Specimens or items are divided into homogenous subsets, or strata

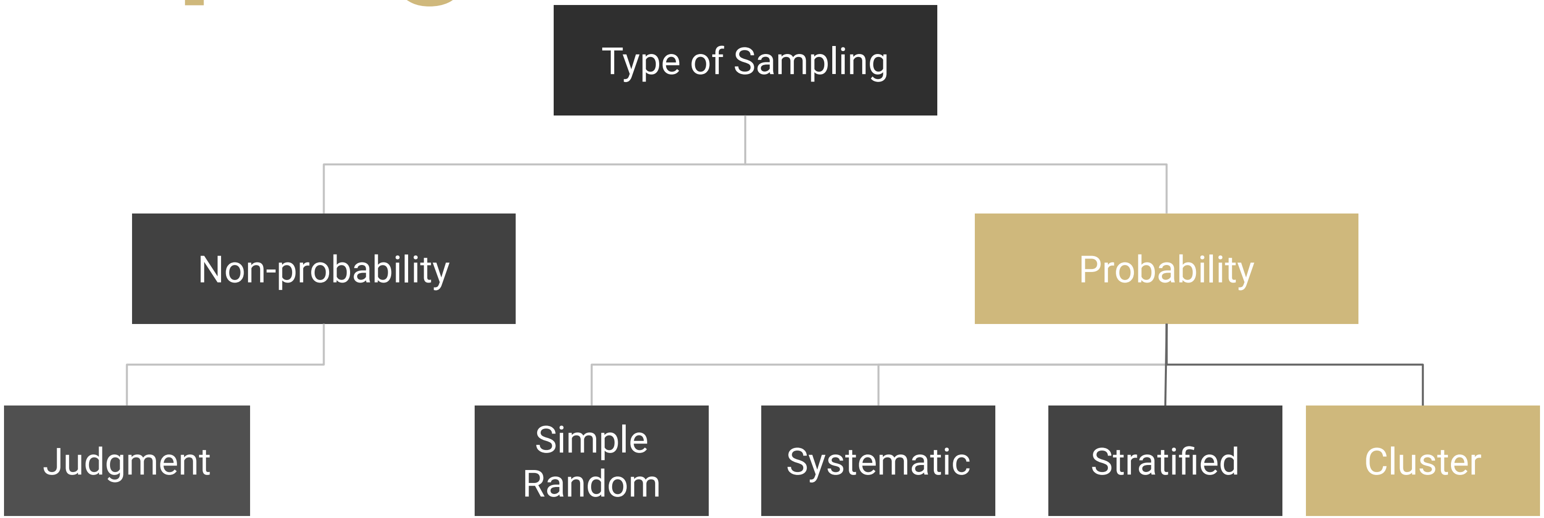


Stratified Random Sampling

- Divide population into homogeneous subgroups called Strata
  - Mutually exclusive
  - Exhaustive
- Select proportionate simple random samples from the subgroups (strata)
- More accurately reflects the characteristics of the population if it has multiple strata

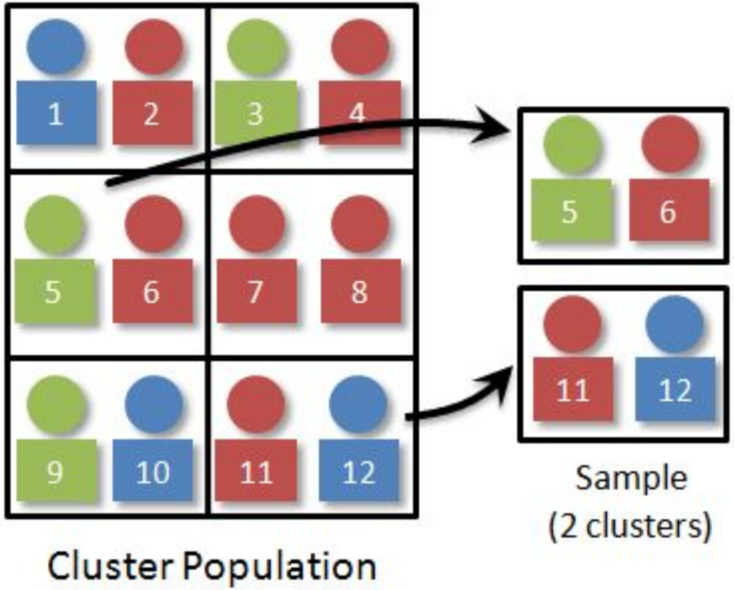


Types of Sampling



Cluster Sampling

- Specimens or items are divided into groups that are homogenous between each other, but heterogeneous within



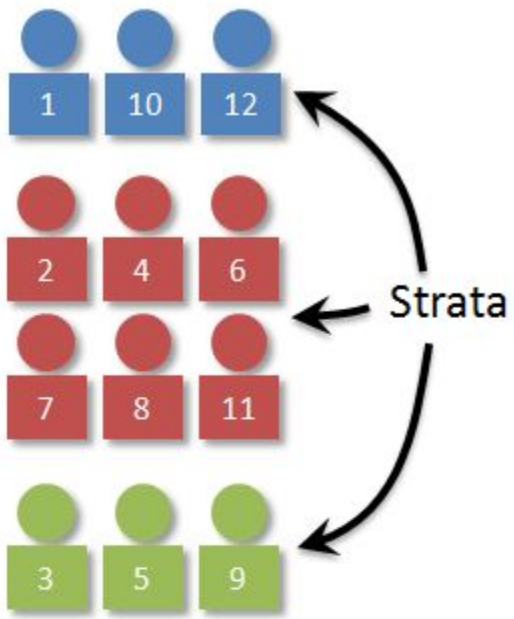
Cluster Sampling

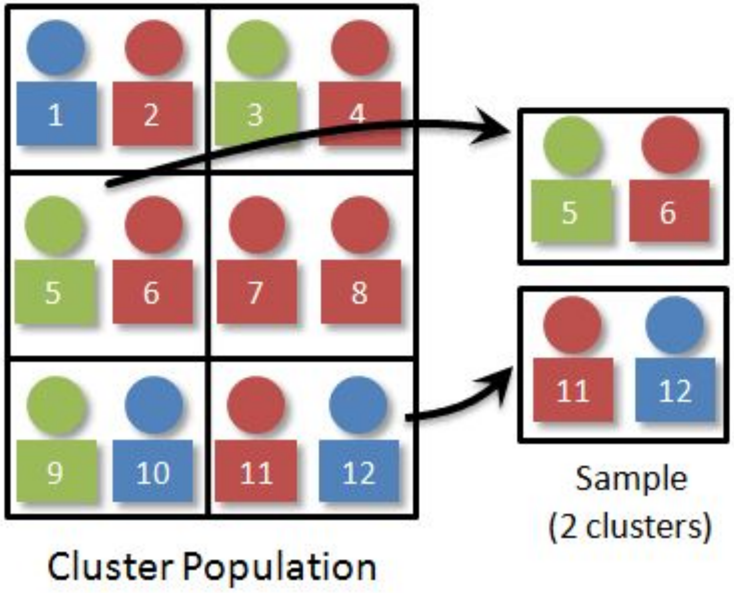
- Divide population into groups, or clusters which are similar, but with large within cluster variability

- Select clusters randomly (as it represents the diversity of the population)

- Survey all elements in the selected clusters

Stratify or Cluster ?

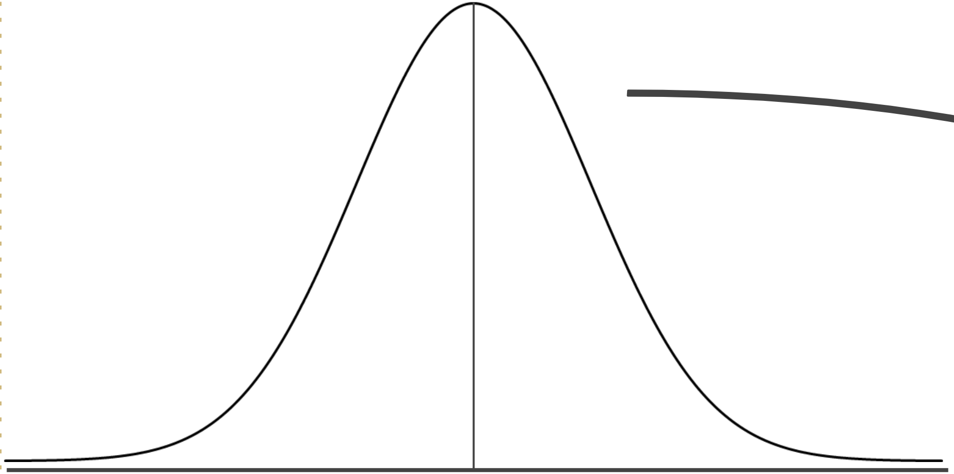




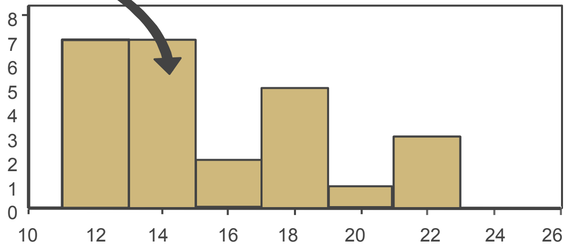
Sampling Error

Sampling Distributions and Estimation

Are you ***really*** interested in the sample?



Population



Sample

Parameters and Statistics

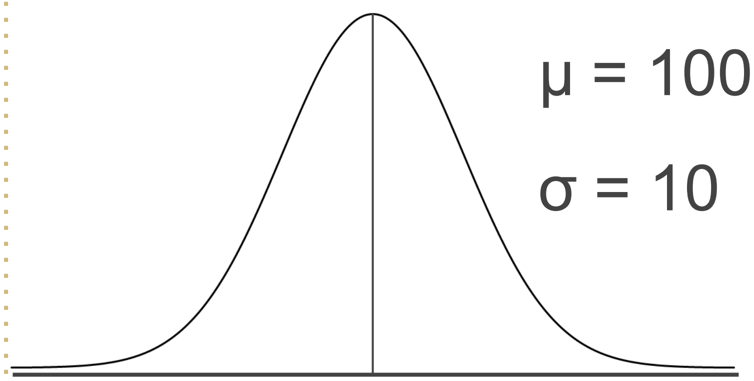
	Sample	Population
Definitions	Subgroup or portion of the population chosen for evaluation or study	Collection of all items produced or considered
Characteristics	Statistics	Parameters
Size	n	N
Mean	$\bar{X}$	μ
Median	$\tilde{X}$	M
Standard Deviation	s	σ
Variance	s^2	σ^2
Skewness	g_3	γ_3
Kurtosis	g_4	γ_4
Proportion	p	π
Rate	$\bar{c}$	λ

Creating Random Numbers in R

In R / Rstudio:

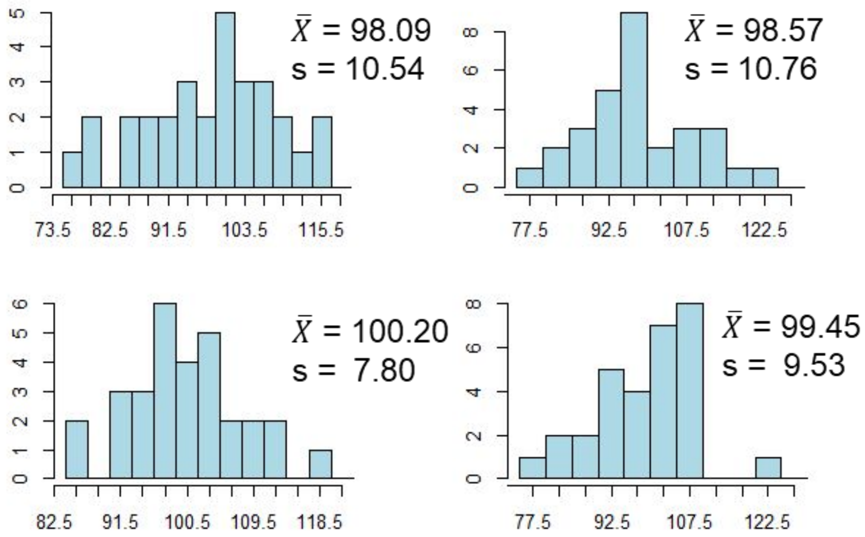
```
rnorm( )
rexp( )
rpois( )
rbinom( )
```

Sampling Distributions and Statistical Inference



$\mu = 100$

$\sigma = 10$



Sampling Error Defined

- An observed difference between a true parameter value and its associated sample descriptive statistic is caused by sampling error.

Sampling Error

- Repeated samples may not be **identical**
- Descriptive statistics calculated from repeated sampling (with replacement) will not be exactly the same, even though the population is unchanged.

Sampling Error

- This is an **expected** phenomenon since we are not measuring all of the subjects or units for the entire population.

- Statistical methods allow us to account for sampling error, and make appropriate decisions.

Sampling Error

- In spite of the presence of sampling error, random sampling allow us to use sample statistics as point estimators of population parameters; however

- Even when unbiased, sample statistics will probably not exactly equal their associated true population parameters.

Sampling Error & Probability

- Sampling Error is quantifiable using **Random Sampling Distributions** (RSDs).

- These distributions, like all probability distributions, are based on the principles of classical probability.

Random Sampling Distributions

Random Sampling Distributions

A random sampling distribution is a theoretical probability distribution that represents **all** of the possible sample statistics of a given size that could be obtained from a population of interest.

Random Sampling Distributions

A random sampling distribution is a theoretical **population** distribution and foundational for understanding statistical inference.

Random Sampling Distributions

- Draw all possible random samples of size n from a given research population
- Calculate descriptive statistics for each of the samples
- Construct a distribution for each of the sampled descriptive statistics

Random Sampling Distributions

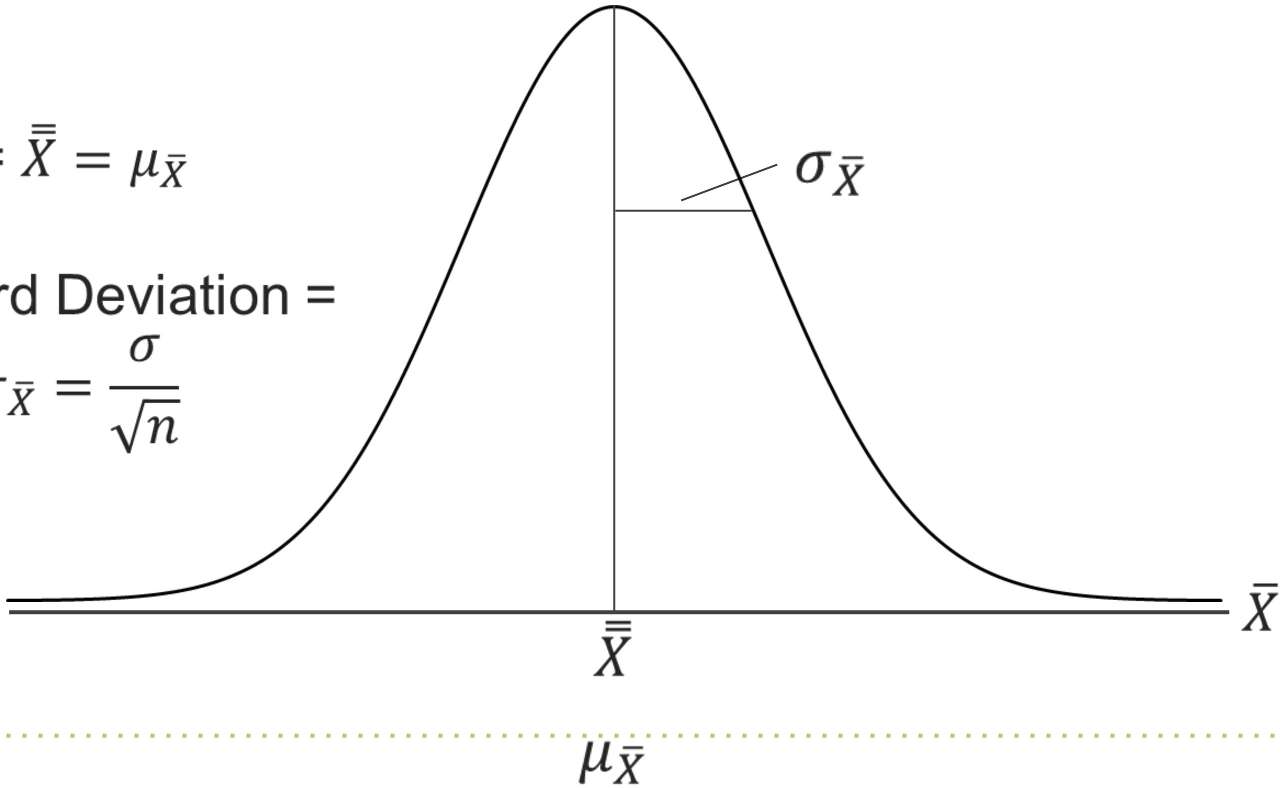
- Each of the resultant distributions constitutes the random sampling distribution of the statistics.

RSD of the Sample Averages

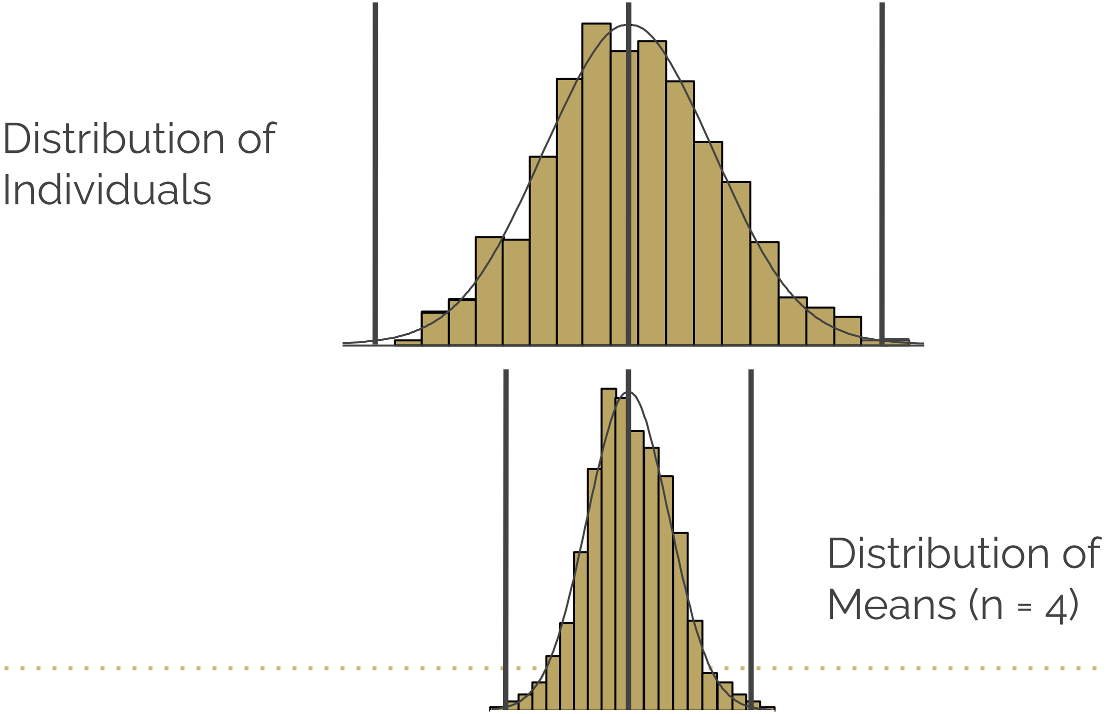
Mean = $\bar{\bar{X}} = \mu_{\bar{X}}$

Standard Deviation =

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$



RSD of the Sample Averages
(from a Normally Distributed Population)



Sampling Distributions

- A **Sample Statistic** is a random variable
    - Sample Mean
    - Sample Proportion
    - Standard Deviation
    - Variance
    - Skewness
    - Kurtosis, etc.

Examples of Sampling Distributions

Population	Sample	Sample Statistic	Sampling Dist.
Water in a river	10-gallon containers of water	Mean number of parts of mercury per million parts of water	Sampling distribution of the mean
All professional basketball teams	Groups of 5 players	Median height	Sampling distribution of the median
All parts in a manufacturing process	50 parts	Proportion defective	Sampling distribution of the proportion

RSD Visualization

https://onlinestatbook.com/stat_sim/sampling_dist/index.html

RSD of the Mean

Developing Sampling Distributions

Suppose we have a small and finite population of only **N = 5** individuals, with the following **ages** (our random variable):

18, 22, 24, 30, 35

Developing Sampling Distributions

We want to estimate the population mean age using a sample of size n=2.

To develop a random sampling distribution of the mean, we can follow these steps:

Developing Sampling Distributions

- Enumerate all possible samples of size n=2 that can be drawn from the population. In this case, there are a total of 25 possible samples.
- Calculate the mean age for each sample.

All Possible Samples of Size n = 2

25 Samples

1st Obs	2nd Observation				
	1	2	3	4	5
1	18,18	22,18	24,18	30,18	35,18
2	18,22	22,22	24,22	30,22	35,22
3	18,24	22,24	24,24	30,24	35,24
4	18,30	22,30	24,30	30,30	35,30
5	18,35	22,35	24,35	30,35	35,35

25 Sample Means

1st Obs	2nd Observation				
	1	2	3	4	5
1	18.0	20.0	21.0	24.0	26.5
2	20.0	22.0	23.0	26.0	28.5
3	21.0	23.0	24.0	27.0	29.5
4	24.0	26.0	27.0	30.0	32.5
5	26.5	28.5	29.5	32.5	35.0

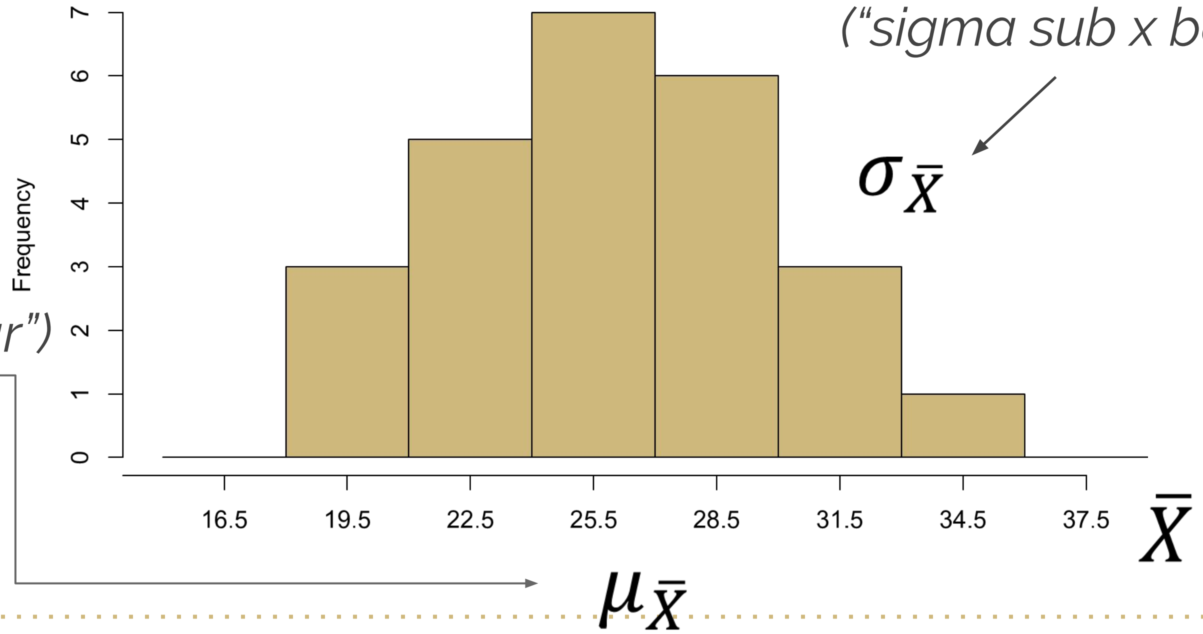
Developing Sampling Distributions

- Plot a histogram of the sample means to visualize the distribution.

Distribution of the Sample Means

Standard deviation of the mean
("sigma sub x bar")

Mean of the sampling distribution
("mu sub x bar")



Developing Sampling Distributions

- Calculate the mean and standard deviation of the sample means.

Formulae

$$\mu_{\bar{X}} = \frac{\sum \bar{X}}{N} = \frac{18 + 20 + \cdots 35}{25} = 25.8$$

$$\sigma_{\bar{X}} = \sqrt{\frac{\sum(\bar{X} - \mu_{\bar{X}})^2}{N}} = \frac{(18 - 25.8)^2 + (20 - 25.8)^2 + \cdots}{25} = 4.252$$

Standard Error of the Mean (or of the Estimate)

- Standard deviation of all possible sample means, $\sigma_{\bar{X}}$
  - Measures scatter in all sample means, $\bar{X}$
- Smaller in value than population standard deviation
- Formula

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$

Central Limit Theorem

Sampling from Normal Populations

- The mean of the Sampling distribution of the means equals the population mean

$$\mu_{\bar{X}} = \mu$$

- The standard deviation (standard error) of the sampling distribution of means equals the population standard deviation divided by the square root of the sample size

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Sampling from Normal Populations

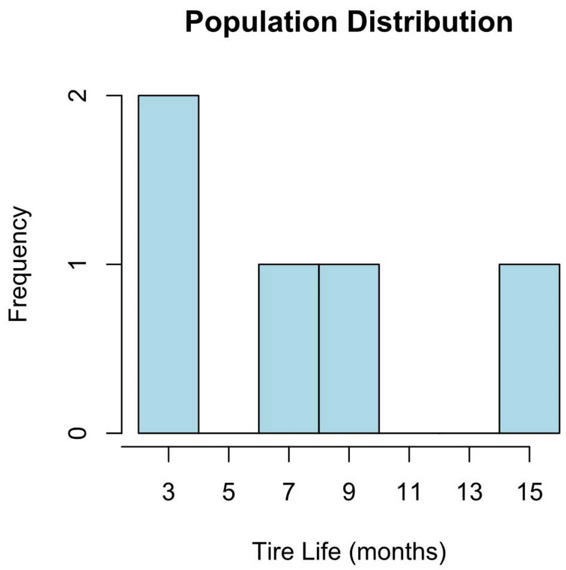
- Sampling distribution of means is normally distributed if the population is normal, and/or becomes increasingly normal as sample size increases if the population is not normally distributed.

- This is due to what is called the **Central Limit Theorem**

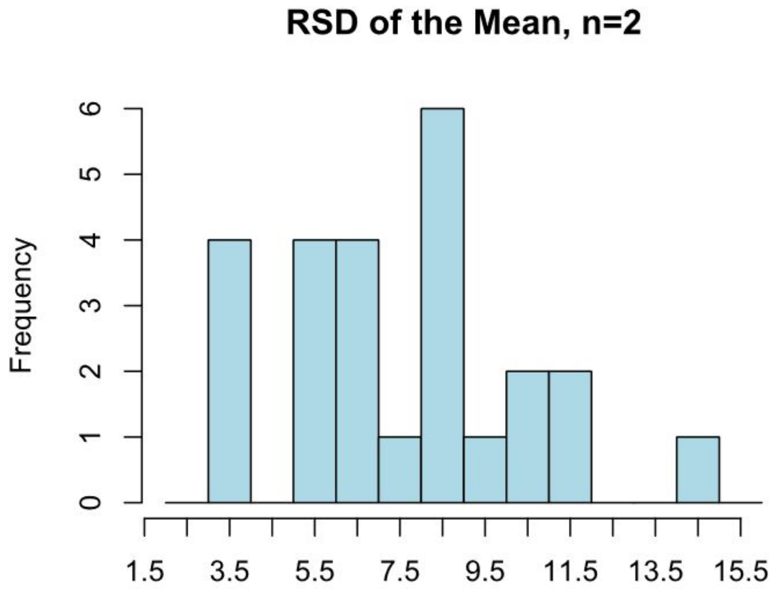
Sampling from Non-Normal Populations

Life of 5 motorcycle tires

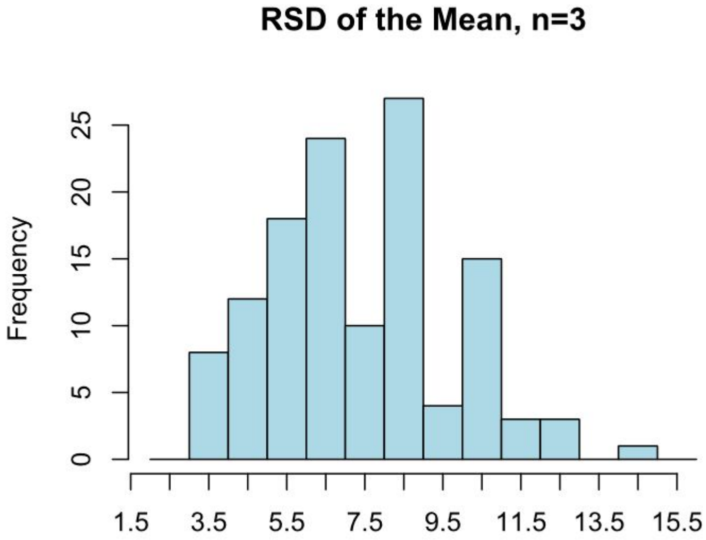
- Raw Data: 3, 3, 7, 9, 14
- N = 5 Mean = 7.2

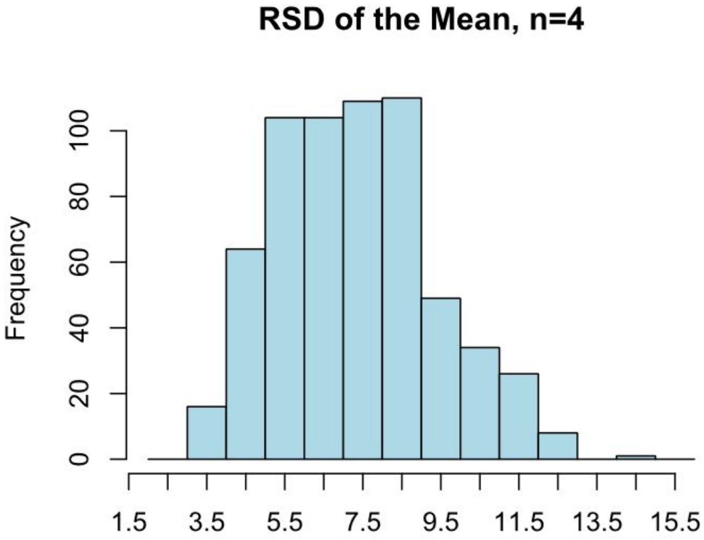


Sampling from Non-Normal Populations

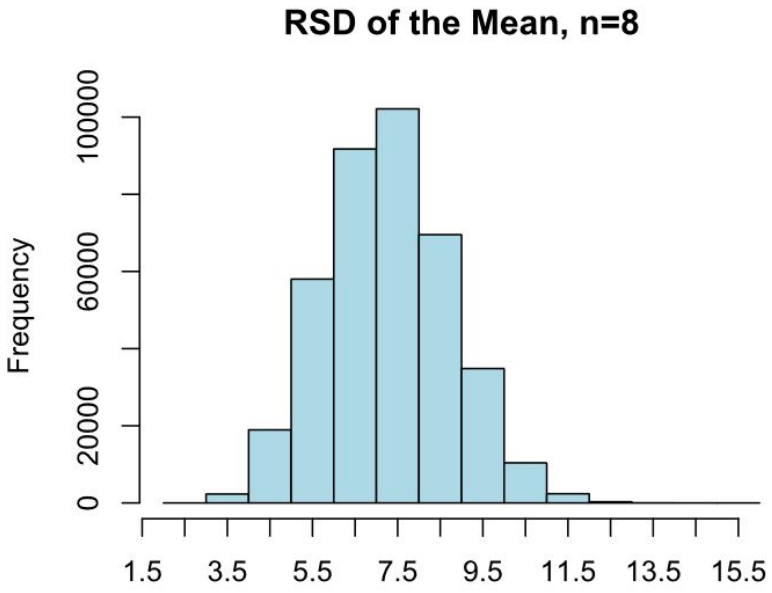


Effect of Sample Size (*n*) on Shape of Sampling Distribution





Effect of Sample Size (*n*) on Shape of Sampling Distribution



As sample size increases, the shape of the distribution becomes more normal.

Central Limit Theorem

- The mean of the RSD of the Mean = Population Mean
- The standard deviation of the RSD of the mean = Population St Dev divided by the square root of n, **regardless of sample size and type of distribution**.

Central Limit Theorem

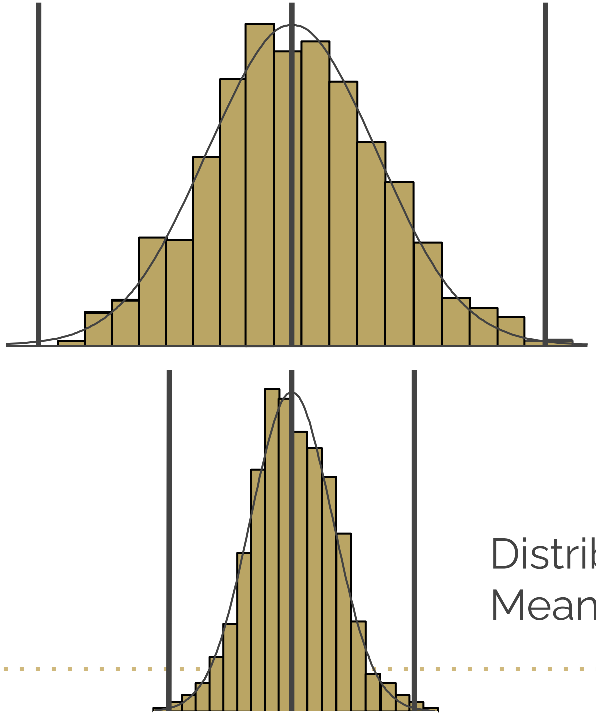
- As the sample size (n) increases, the RSD of the means will approach normality, regardless of the shape of the process distribution.

Central Limit Theorem

- This applies even without our knowing anything about the shape of that population other than what we can gather from the sample (in most cases).

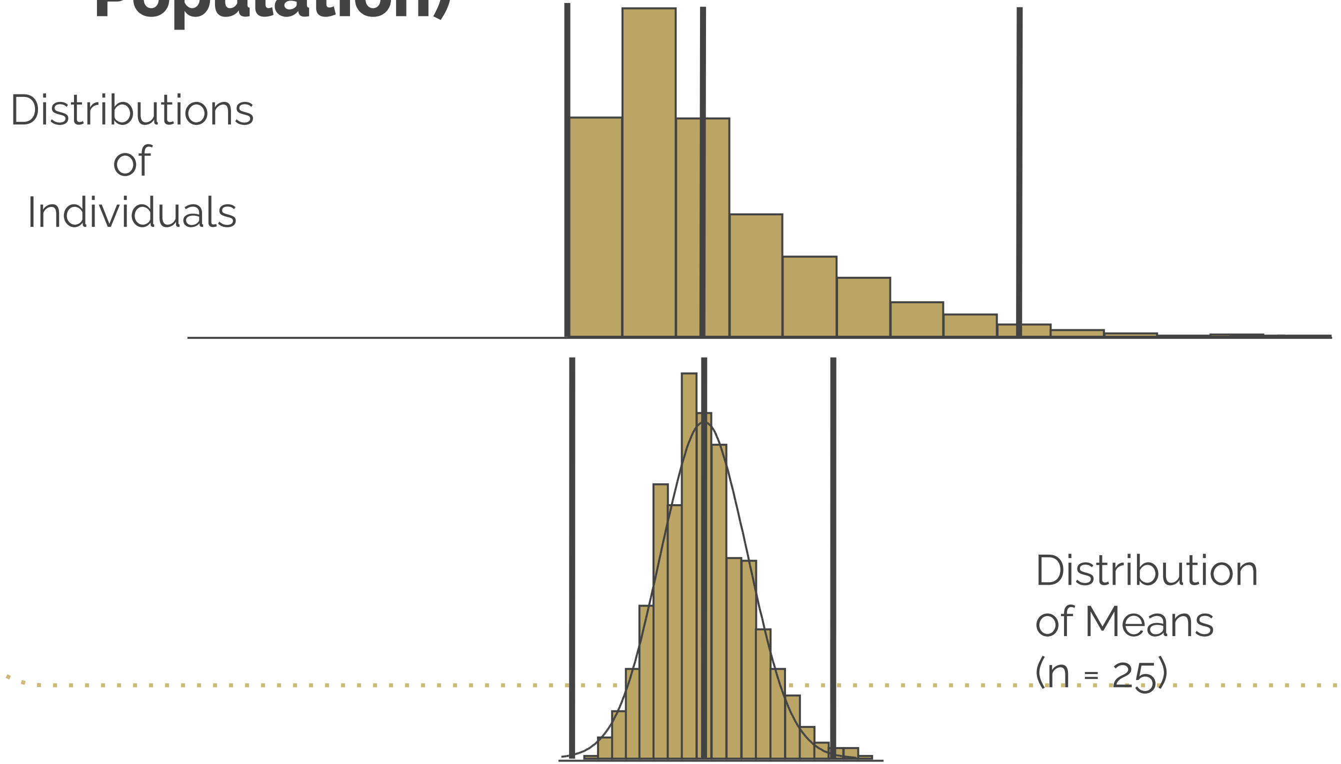
RSD of the Sample Averages
(from a Normally Distributed Population)

Distribution of Individuals



Distribution of Means (n = 4)

RSD of the Means
(from an Exponentially Distributed Population)



RSD Visualization

https://onlinestatbook.com/stat_sim/sampling_dist/index.html

Central Limit Theorem

- For large sample sizes ($n \geq 30$), regardless of the original population distribution, the normal distribution is a good approximation to the sampling distribution of X when **σ** is known.

Central Limit Theorem

The Central Limit Theorem ***Does Not Apply to All*** *Random Sampling Distributions – just the random sampling distribution of the mean*

Probability Problems using the RSD of the Mean

Estimating Probability Using the RSD of the Mean

Note that when we use the Standard Error of the Estimate to find areas on the RSD of the means, the z-score employed becomes:

$$Z_{\bar{X}} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

RSD of the Mean – Example 1

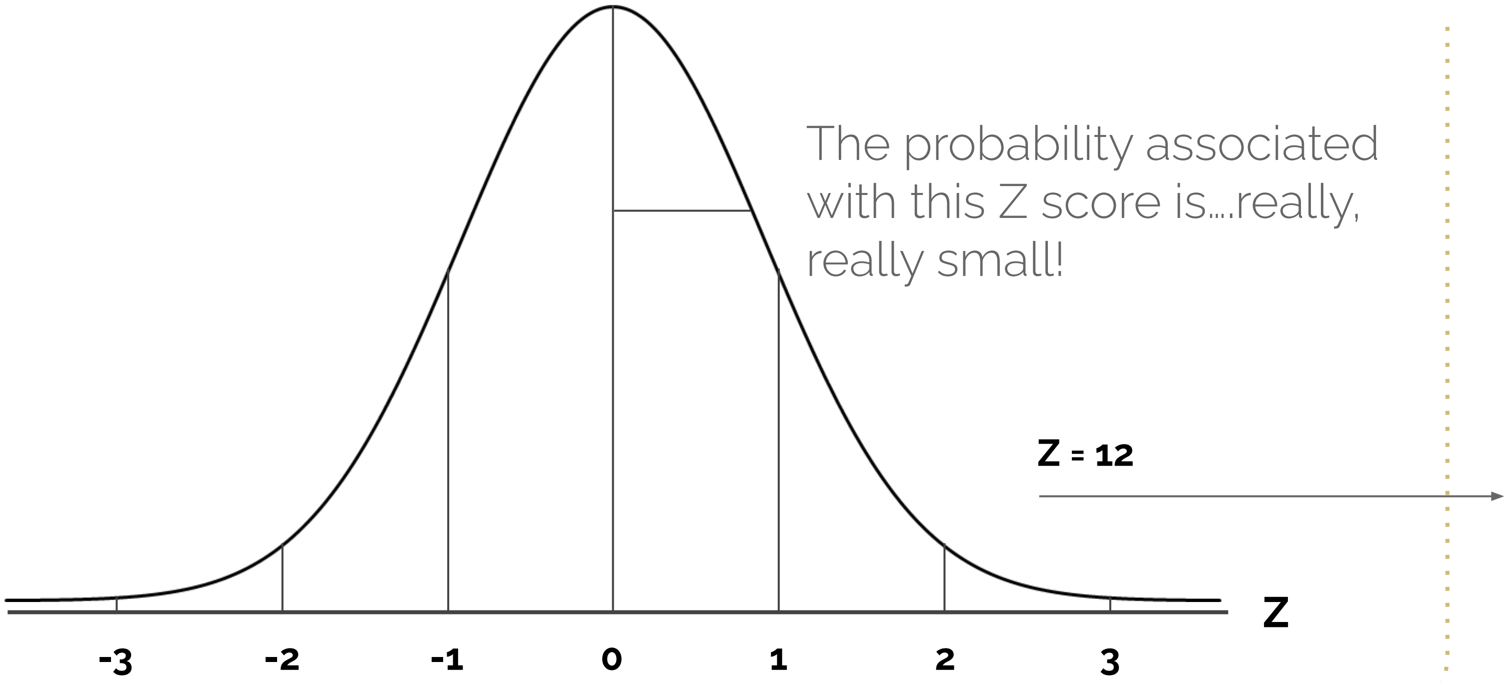
- A process has historically manufactured parts at a mean, μ, of 1.325, with a standard deviation, σ, of 0.045.
- Drawing a random sample of 25 units, what is the probability of finding an $\bar{X}$ of 1.433 or more for the sample if no change has occurred in the mean or dispersion of the process?

RSD of the Mean – Example 1

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{0.045}{\sqrt{25}} = 0.009$$

$$Z_{\bar{X}} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{1.433 - 1.325}{0.009} = 12$$

RSD of the Mean – Example 1



The probability associated with this Z score is….really, really small!

RSD of the Mean – Example 2

- A process has historically manufactured parts at a mean, μ, of 50, with a standard deviation, σ, of 14.4.
- Drawing a random sample of 16 units, what is the probability of finding an $\bar{X}$ of 55 or more for the sample if no change has occurred in the mean or dispersion of the process?

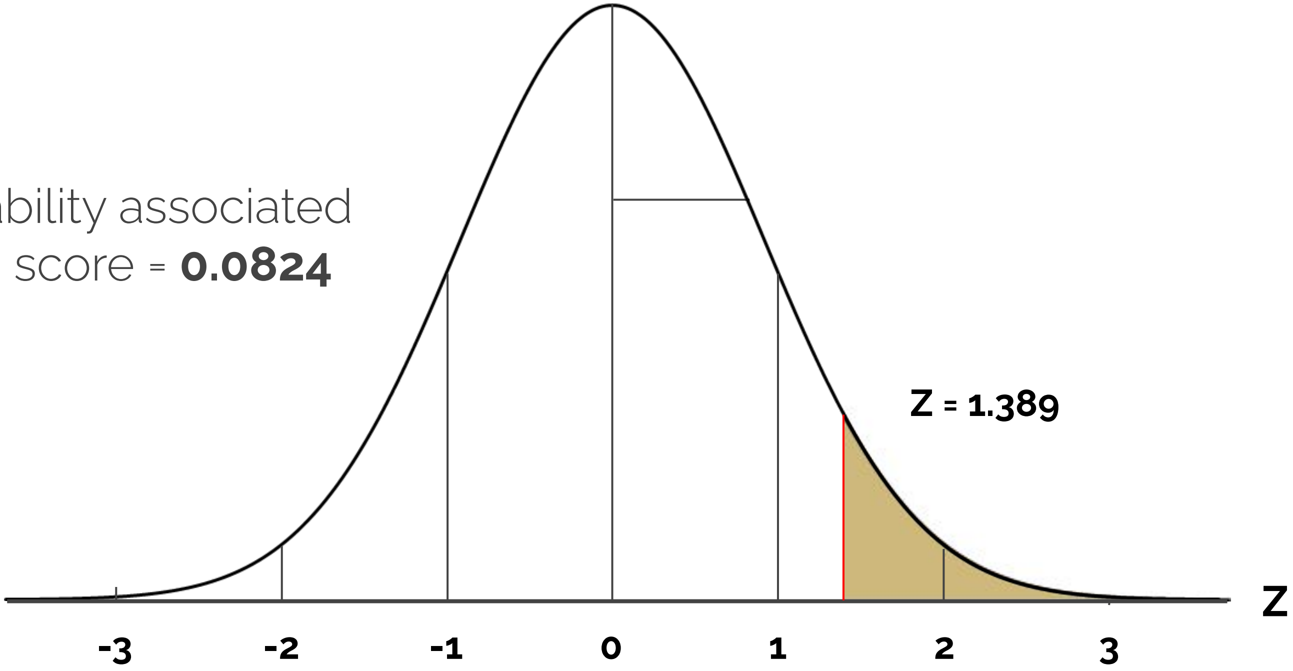
RSD of the Mean – Example 2

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{14.4}{\sqrt{16}} = 3.6$$

$$Z_{\bar{X}} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{55 - 50}{3.6} = 1.389$$

RSD of the Mean – Example 2

The probability associated with this Z score = **0.0824**



RSD of the Mean - Example 3

Suppose:

Individual savings accounts at a bank

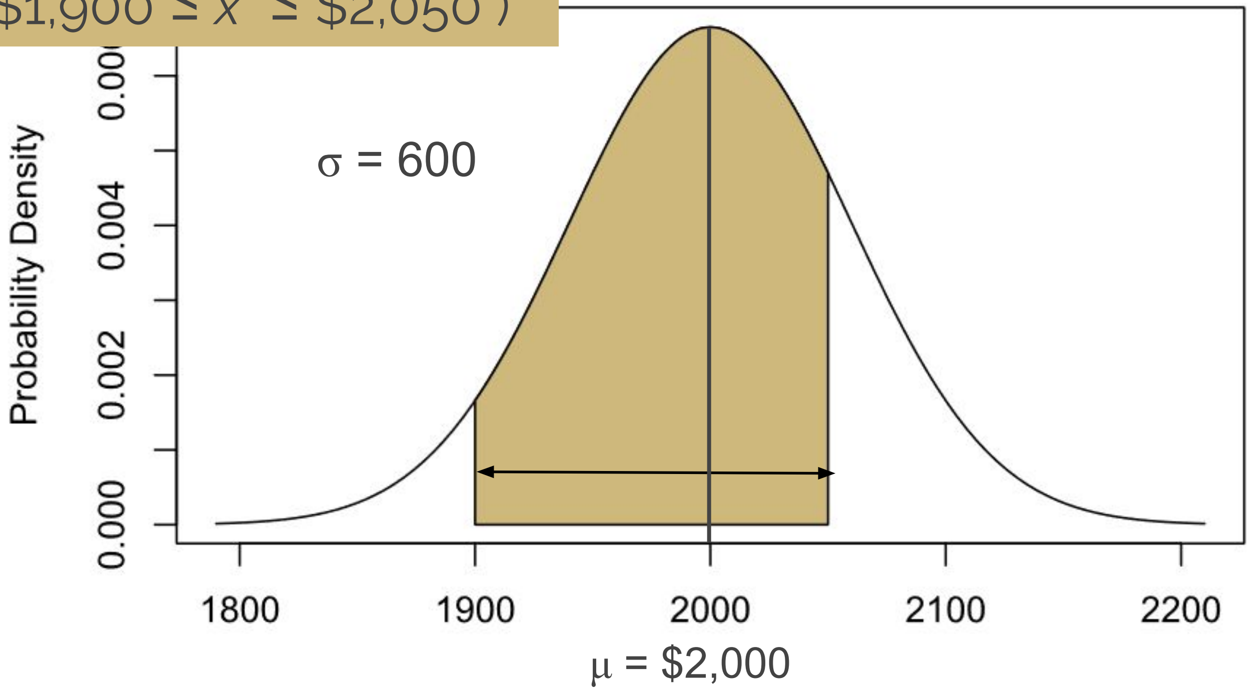
1. Average $2,000, with
2. Standard deviation $600, and
3. Normally Distributed

If 100 accounts are randomly selected, what is the probability that the **average** balance is between $1,900 and $2,050?

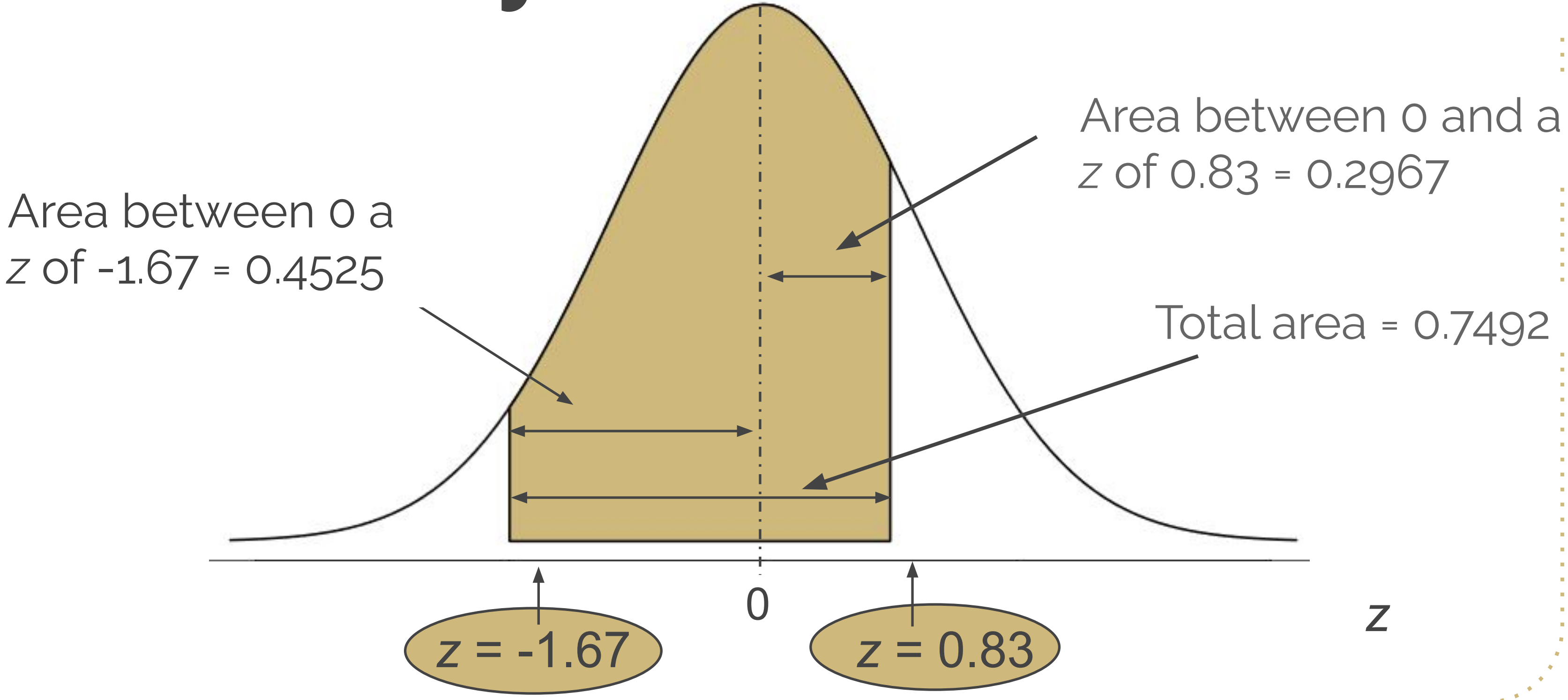
Graphical Representation of the Bank Problem

Shaded Area = Probability

P($1,900 ≤ $\bar{x}$ ≤ $2,050)



Finding the Area (Probability)



How To Say It

1. Say the meaning of the "basic" symbol. Use the word "population" for Greek letters and "sample" for English letters.
2. Then say the words "of the distribution of the sample"
3. Then say the meaning of subscript symbol using the plural form.

$\mu_{\bar{X}}$

The population mean of the distribution of the sample means

$\sigma^2_{s^2}$

The population variance of the distribution of the sample variances

RSD Examples

Statistic	RSD	Standard Error
$\bar{X}$	RSD of the mean	of the mean
$\tilde{X}$	RSD of the median	of the median
p	RSD of the proportion	of the proportion
R	RSD of the range	of the range