



Sampling Distribution of the sample mean

Mathematics Education Section

CDI, EMB

Email: stchan@emb.gov.hk



Key Points

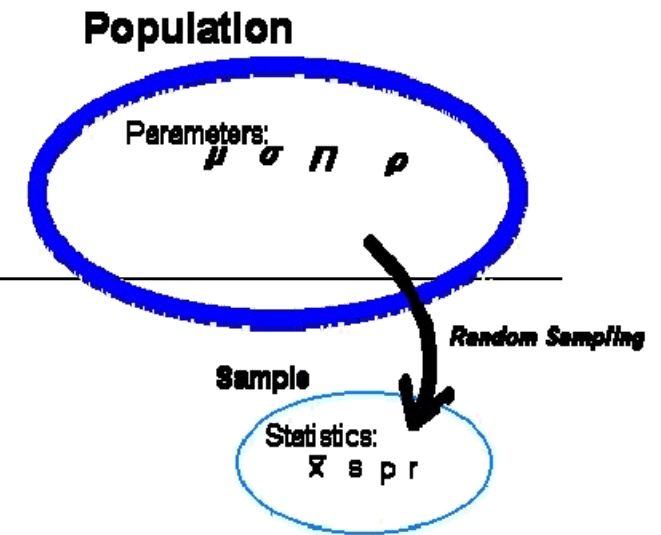
- *Sampling Distribution of the sample mean*
- *The Central Limit Theorem*
- *Point Estimates*



Reference

- MOORE, DAVID S., STATISTICS: CONCEPTS AND CONTROVERSIES, 1991, NEW YORK, W.H. FREEMAN EMPHASIZES IDEAS, RATHER THAN MECHANICS.
- FREEDMAN, DAVID, PISANI, ROBERT, AND PURVES, ROGER, STATISTICS, 1991. NEW YORK, W.W. NORTON.
- A course in Business Statistics
http://wps.prenhall.com/bp_shannon_coursebus_3/0,6134,221218-,00.html
- Introduction to the Practice of Statistics
<http://bcs.whfreeman.com/ips5e/default.asp>
- Statistics CAI
<http://newton.ma.polyu.edu.hk/Stat/Laboratory>

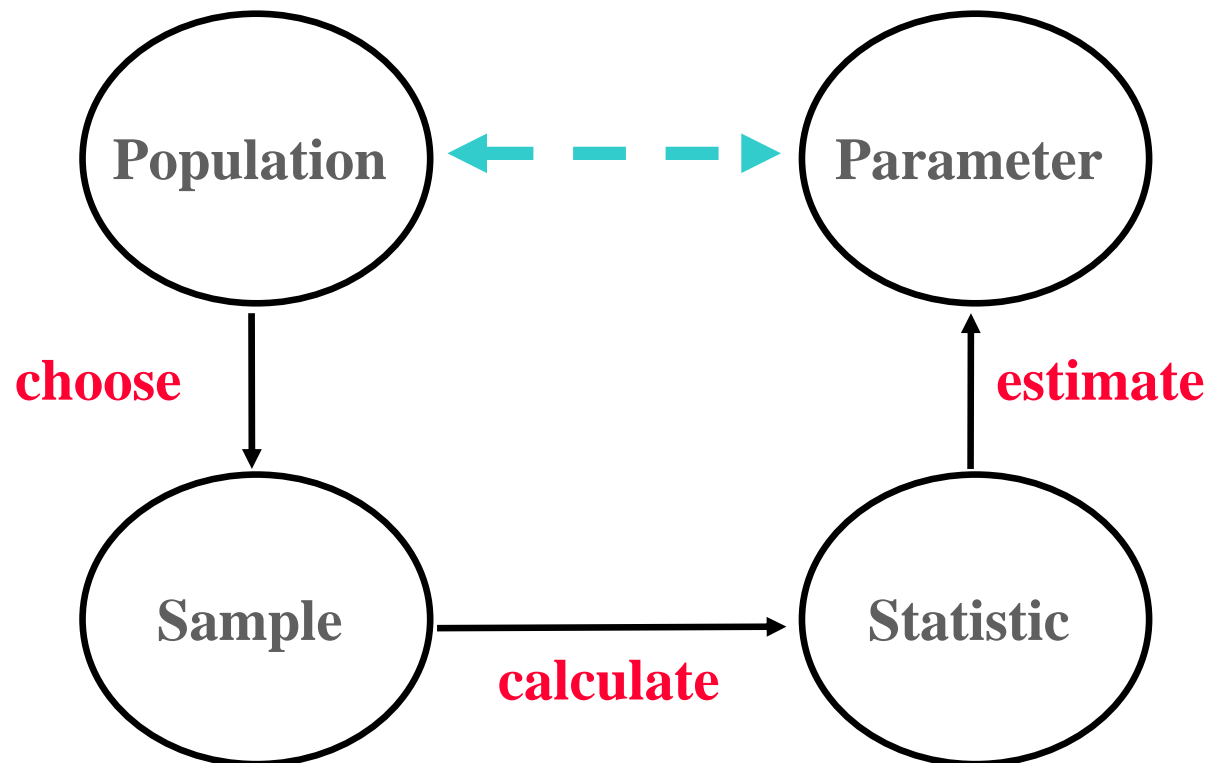
Why sample the population?



1. **Expense.** A census may not be cost effective.
2. **Speed of Response.** There may not be enough time to obtain more than a sample.
3. **Accuracy.** A carefully obtained sample may be more accurate than a census.
4. **Destructive Sampling.** In destructive testing of products a sample has to suffice.
5. **The large (infinite) population.** Sometimes a census is impossible.

Parameter and Statistic

- A number that describes a population is called a **parameter**
- A number that describes a sample is a **statistic**
- If we take a sample and calculate a statistic, we often use that statistic to infer something about the population from which the sample was drawn





Population

- Imagine that our population consists of only three numbers:
 - the number 1, the number 2 and the number 3
- Our plan is to draw an infinite number of random samples of size $n = 2$
- Our population is small, we are sampling with replacement



Population

	X
	1.00
	2.00
	3.00
Mean	2.00
σ^2	0.67

1

2

3

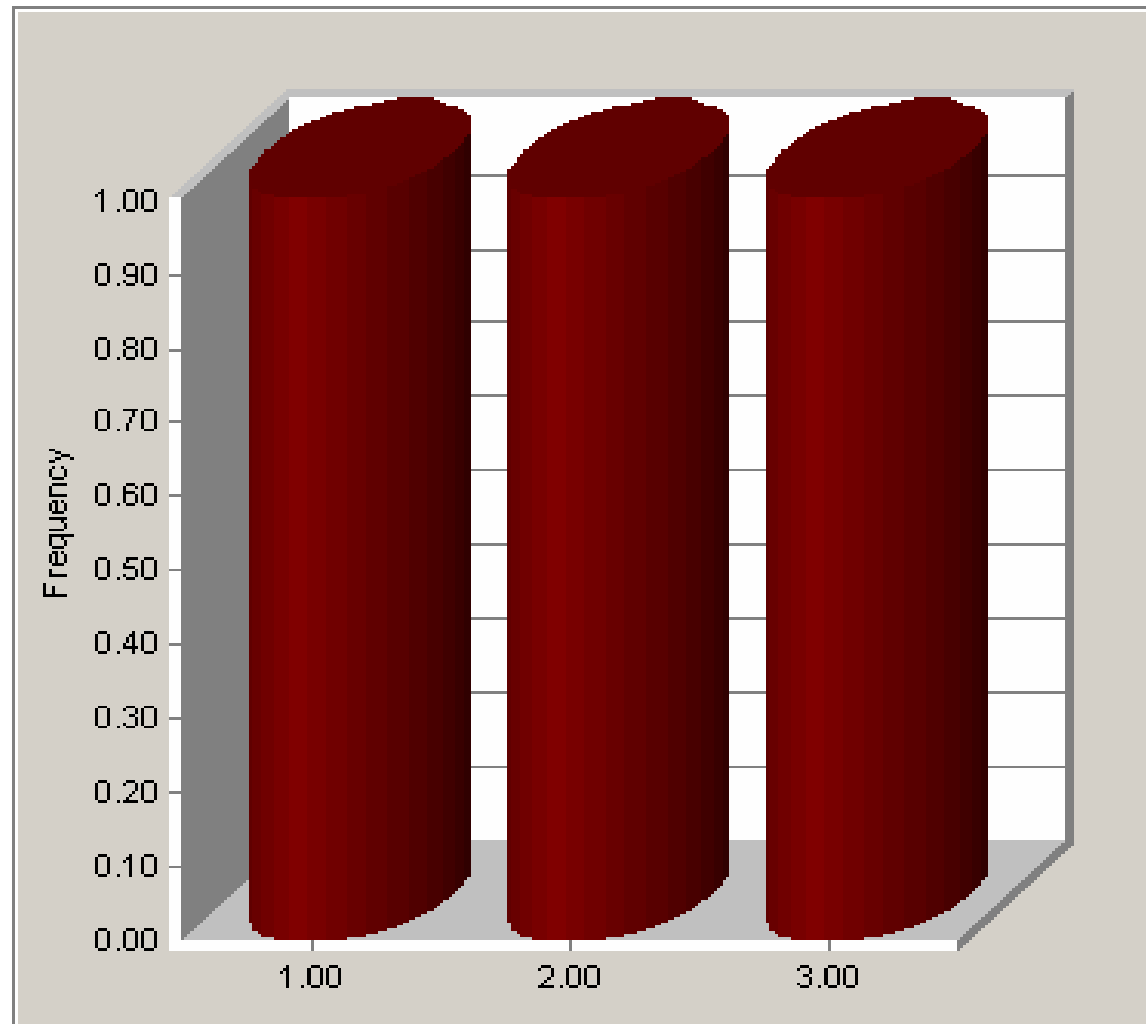
$$\sigma^2 = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3} = \frac{2}{3}$$



Frequency Table

<i>X</i>	<i>Frequency</i>
<i>1</i>	1
<i>2</i>	1
<i>3</i>	1

The Distribution of the Population

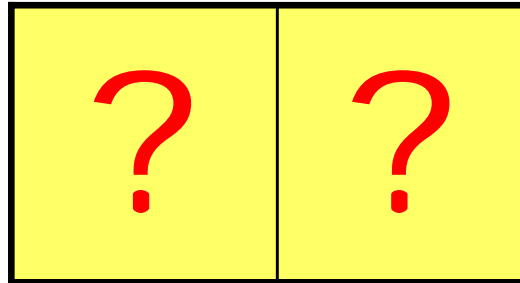




Theoretical Sampling Distribution of the Sample Mean



Sample size $n = 2$





All Possible Samples

<i>Sample₁</i>	1	1
<i>Sample₂</i>	1	2
<i>Sample₃</i>	1	3
<i>Sample₄</i>	2	1
<i>Sample₅</i>	2	2
<i>Sample₆</i>	2	3
<i>Sample₇</i>	3	1
<i>Sample₈</i>	3	2
<i>Sample₉</i>	3	3



Compute
the Sample
Means

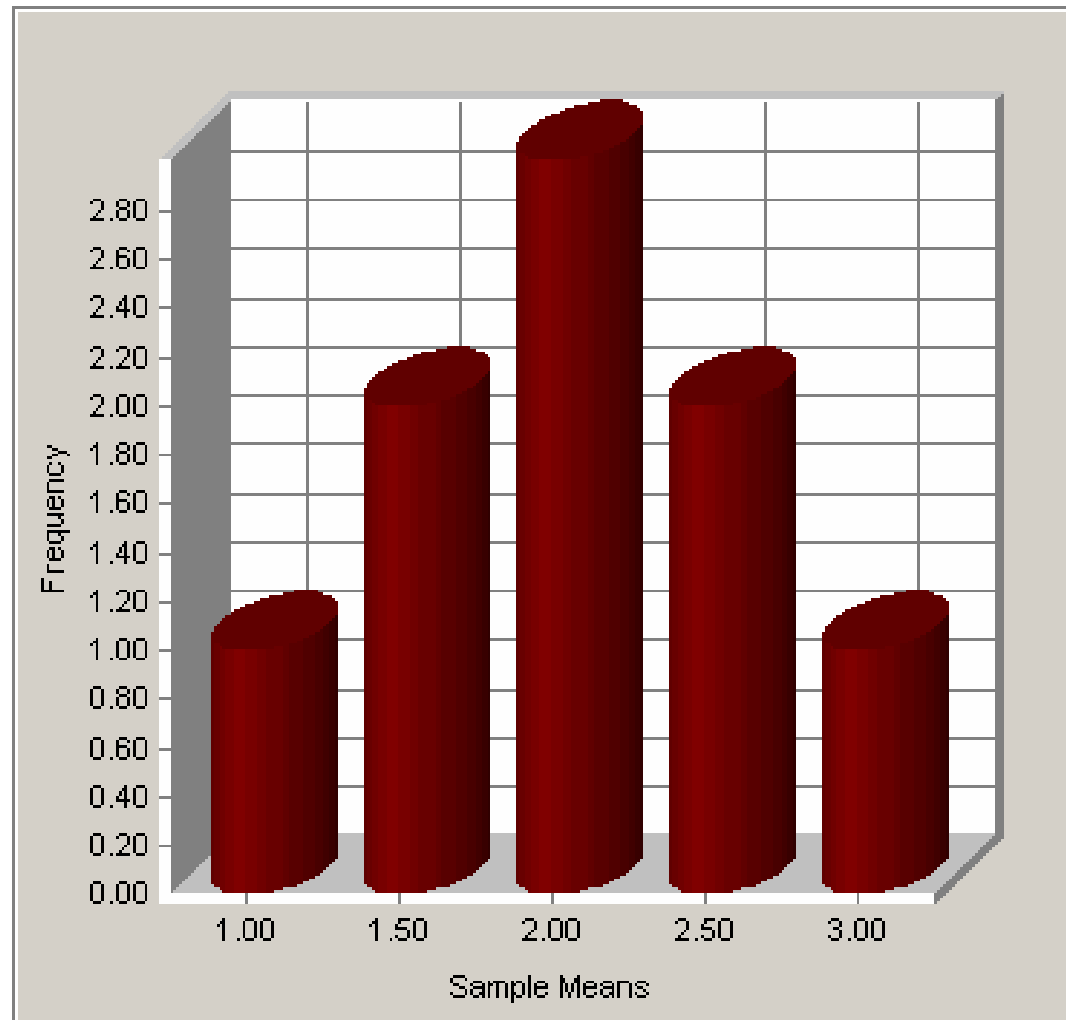
	X_1	X_2	Sample Mean
Sample 1	1	1	$(1 + 1) / 2 = 1$
Sample 2	1	2	$(1 + 2) / 2 = 1.5$
Sample 3	1	3	$(1 + 3) / 2 = 2$
Sample 4	2	1	$(2 + 1) / 2 = 1.5$
Sample 5	2	2	$(2 + 2) / 2 = 2$
Sample 6	2	3	$(2 + 3) / 2 = 2.5$
Sample 7	3	1	$(3 + 1) / 2 = 2$
Sample 8	3	2	$(3 + 2) / 2 = 2.5$
Sample 9	3	3	$(3 + 3) / 2 = 3$



Frequency Table

<i>M</i>	<i>Frequency</i>
<i>1</i>	1
<i>1.5</i>	2
<i>2</i>	3
<i>2.5</i>	2
<i>3</i>	1

The Distribution of the Sample Mean \bar{X}





The mean of the
sampling distribution
of the sample mean

$$\mu_{\bar{X}} = 2$$

	X_1	X_2	Sample Mean
Sample 1	1	1	1
Sample 2	1	2	1.5
Sample 3	1	3	2
Sample 4	2	1	1.5
Sample 5	2	2	2
Sample 6	2	3	2.5
Sample 7	3	1	2
Sample 8	3	2	2.5
Sample 9	3	3	3
Mean			2



The Mean of the Population

X

1.00

2.00

3.00


Mean 2.00



The variance of the
sampling distribution
of the sample mean

$$\sigma_{\bar{X}}^2 = 0.33$$

	X_1	X_2	Sample Mean
Sample 1	1	1	1
Sample 2	1	2	1.5
Sample 3	1	3	2
Sample 4	2	1	1.5
Sample 5	2	2	2
Sample 6	2	3	2.5
Sample 7	3	1	2
Sample 8	3	2	2.5
Sample 9	3	3	3
Variance			0.33




The variance of the sampling distribution of the sample mean

$$\begin{aligned}\sigma_{\bar{X}}^2 &= \frac{(1-2)^2 + (1.5-2)^2 + (2-2)^2 + (1.5-2)^2 + (2-2)^2 + (2.5-2)^2 + (2-2)^2 + (2.5-2)^2 + (3-2)^2}{9} \\ &= \frac{1 + 0.25 + 0 + 0.25 + 0 + 0.25 + 0 + 0.25 + 1}{9} \\ &= \frac{3}{9} \\ &= \frac{1}{3}\end{aligned}$$

The Variance of the Population

The population's distribution of individual scores has far **more variability** than does the distribution of sample means.

<hr/>	
X	
<hr/>	
	1.00
	2.00
	3.00
<hr/>	
σ^2	0.67
<hr/>	




The variance of the sampling distribution of the sample mean

$$\sigma_{\frac{x}{n}}^2 = \frac{\sigma^2}{n}$$

n = sample size

$$0.67 / 2 = 0.33$$



standard error of the sampling distribution of the sample mean

$$\sigma_{\bar{X}}$$

standard error of the sampling distribution

of the sample mean = $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$



The sample mean
= the population mean ?

Rolling a fair dice

- Cast a fair dice and take X to be the uppermost number, the population mean is $\mu = 3.5$, and that the population median is also $m = 3.5$.
- Take a sample of **four** throws, the mean may be far from 3.5



The results of 5 such samples of 4 throws

	X_1	X_2	X_3	X_4	\overline{X}
Sample 1	6	2	5	6	4.75
Sample 2	2	3	1	6	3
Sample 3	1	1	4	6	3
Sample 4	6	2	2	1	2.75
Sample 5	1	5	1	3	2.75

The sample size $n = 4$



Rolling a fair dice

Quick-Calc Random number generator

- <http://www.graphpad.com/quickcalcs/randomN1.cfm>

Introduction to the Practice of Statistics

- <http://bcs.whfreeman.com/ips5e/default.asp>

Rice Virtual Lab in Statistics

- <http://www.ruf.rice.edu/~lane/rvls.html>



The Central Limit Theorem:

- ◆ The mean of the sampling distribution of **ALL** the sample means is equal to the true population mean.

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

- ◆ The standard deviation of a sampling distribution of means is called **STANDARD ERROR** of the mean

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

The variability of a sample mean decreases as the sample size increases



Central Limit Theorem (CLT)

- ◆ If the population distribution is normal, so is the sampling distribution of \bar{X}
- ◆ For **ANY** population (regardless of its shape) the distribution of sample means will approach a normal distribution as ***n*** approaches ***infinity***

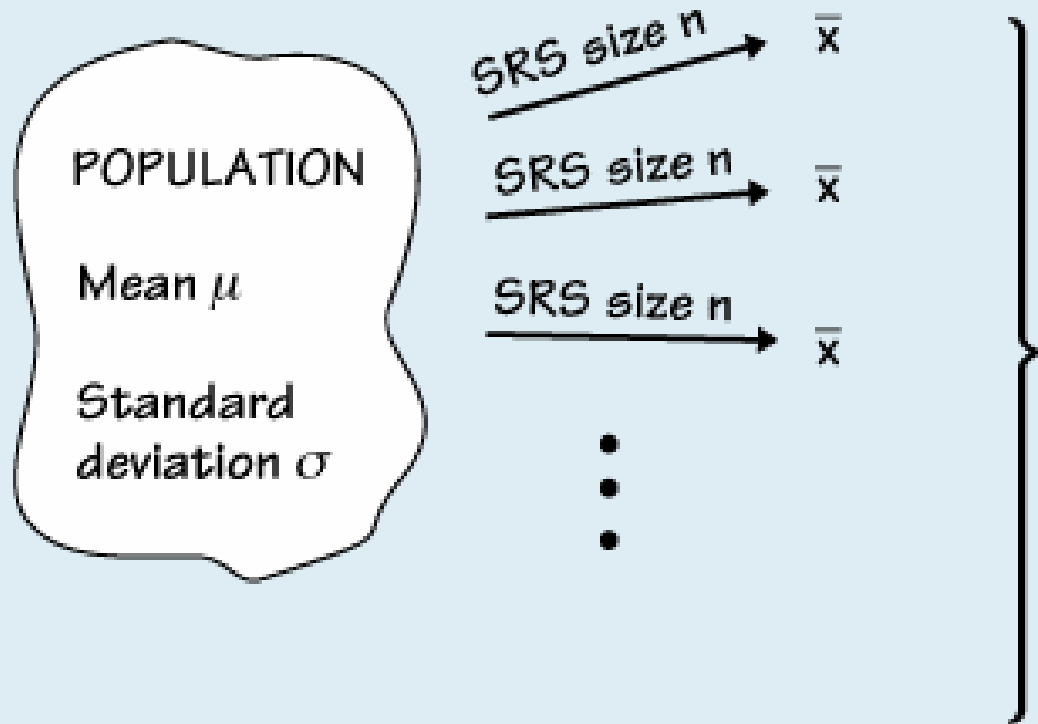


Central Limit Theorem (CLT)

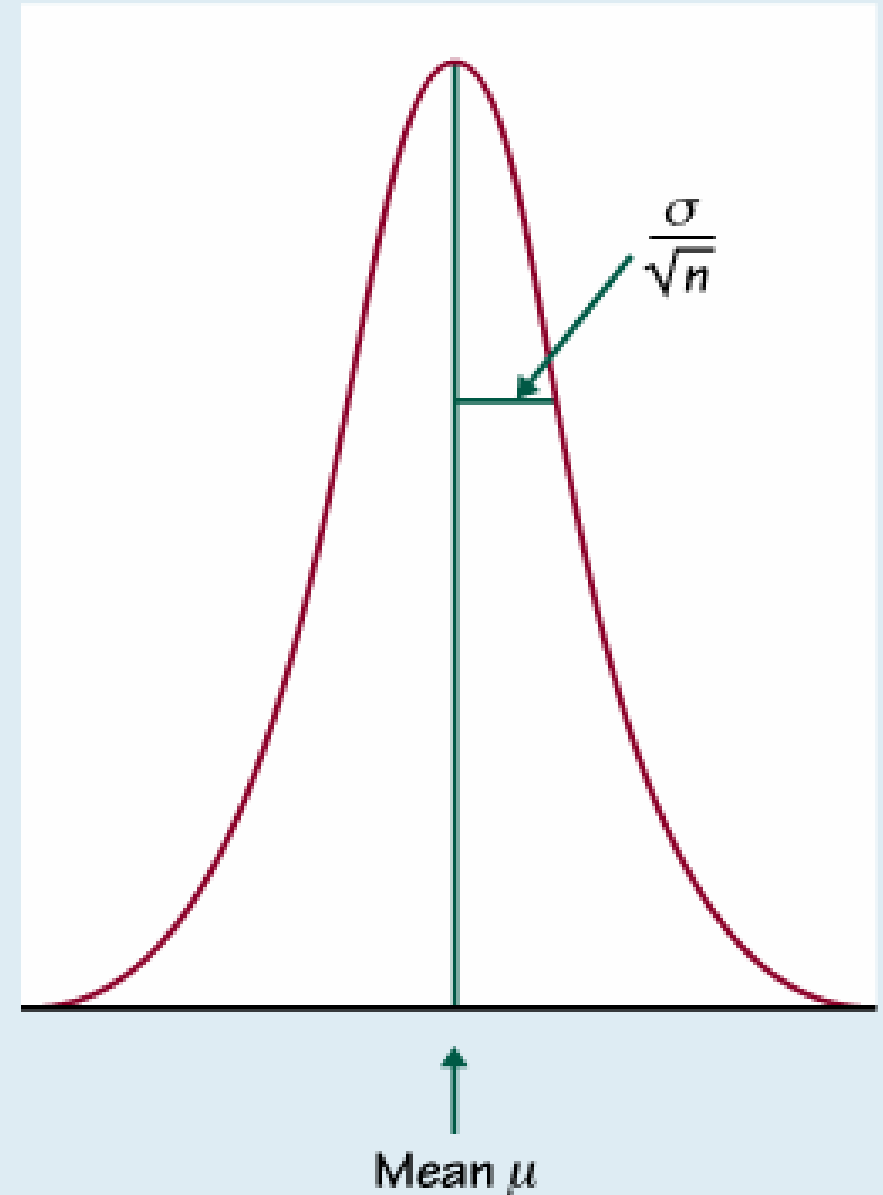
How large is a “large sample”?

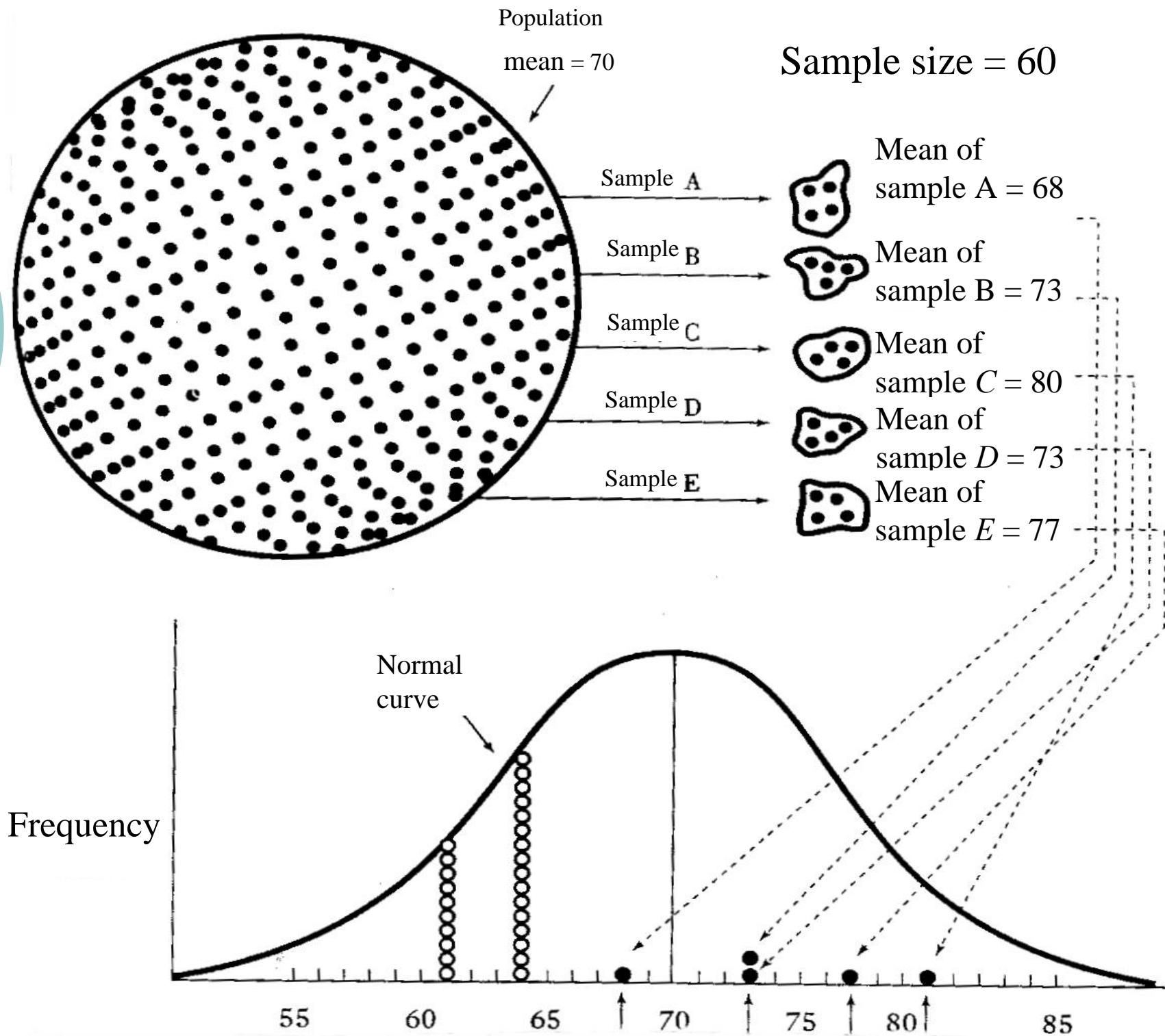
It depends upon the form of the distribution from which the samples were taken. If the population distribution deviates greatly from normality larger samples will be needed to approximate normality.

The Central Limit Theorem



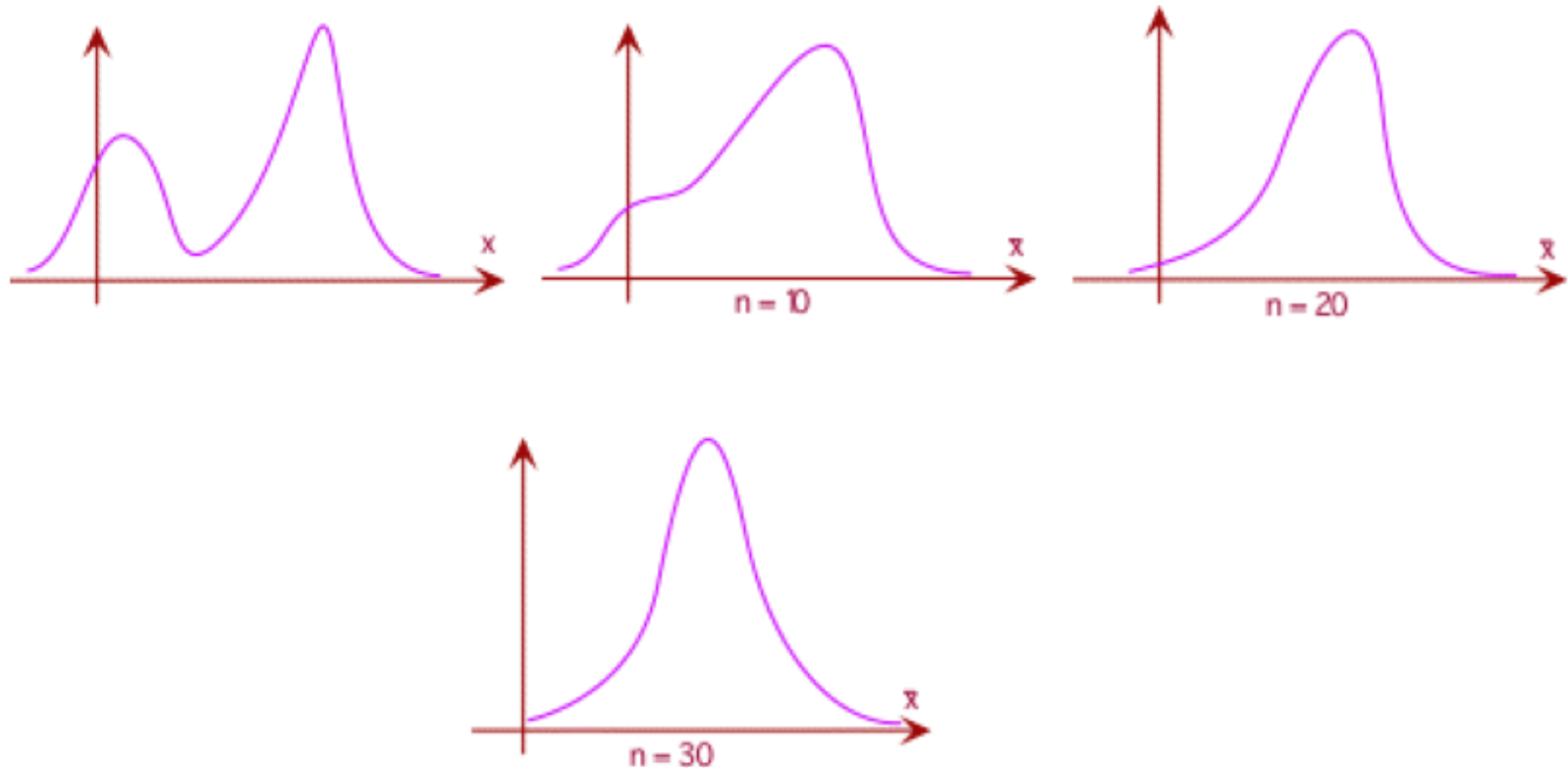
Sampling distribution of \bar{x}





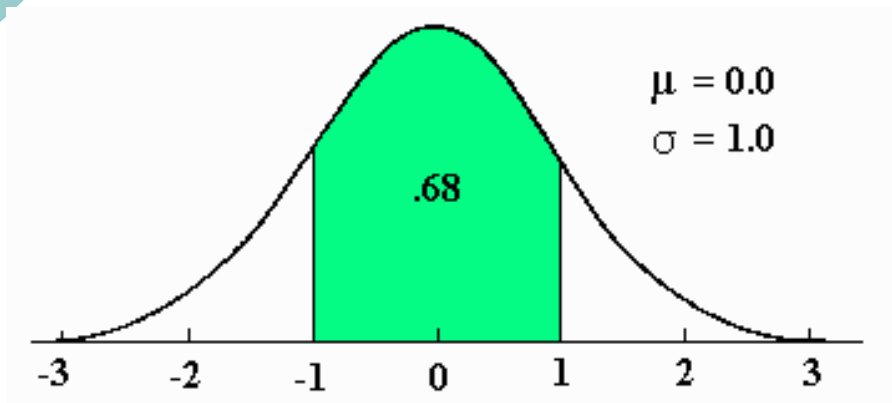
Central Limit Theorem (CLT)

The following illustration shows how the sample size effects the shape of the sampling distribution.



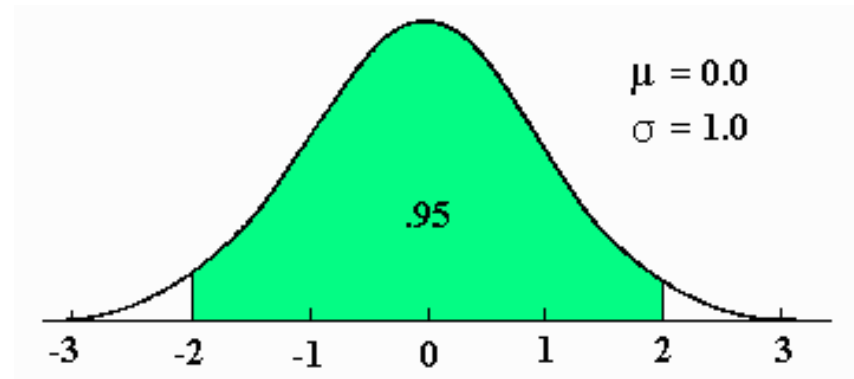
- [http://www.statisticalengineering.com/central_limit_theorem_\(summary\).htm](http://www.statisticalengineering.com/central_limit_theorem_(summary).htm)

The Central Limit Theorem



$$\bar{X} : N(\mu, \frac{\sigma}{\sqrt{n}})$$

68% probability that our
will be in this region \bar{X}



95% probability that our
will be in this region \bar{X}



Using the Central Limit Theorem

- A lightbulb manufacturer claims that the lifespan of its lightbulbs has a mean of 54 months and a standard deviation of 6 months.
- Your consumer advocacy group tests 50 of them. Assuming the manufacturer's claims are true, what is the probability that it finds a mean lifetime of less than 52 months?



Solution

- By Central Limit Theorem, \bar{X} is approximately normal and $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
standard deviation of \bar{X} is $\frac{\sigma}{\sqrt{n}}$

$\therefore \mu_{\bar{X}} = 54$ and a standard error of $\sigma_{\bar{X}} = \frac{6}{\sqrt{50}} \approx 0.85$ months



Solution

- To find $P(\bar{X} \leq 52)$, we need to convert to z-scores:

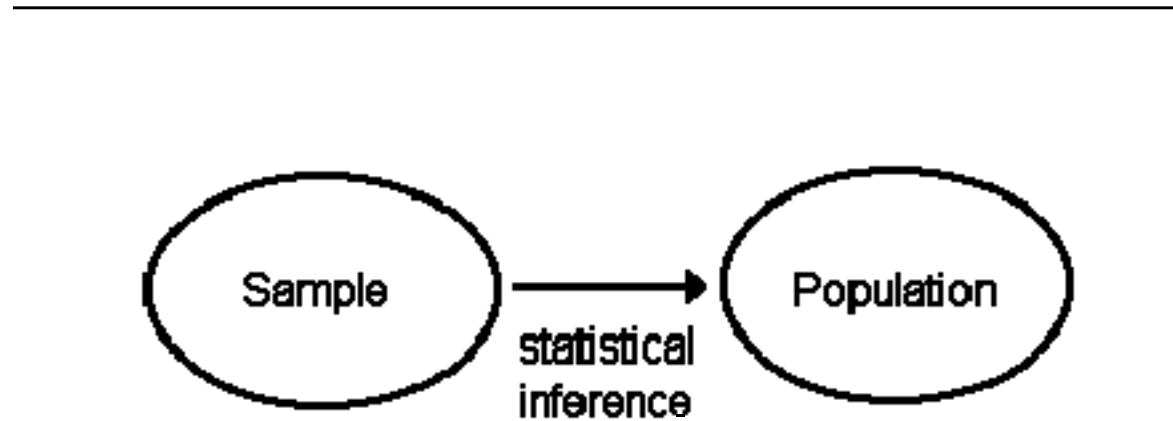
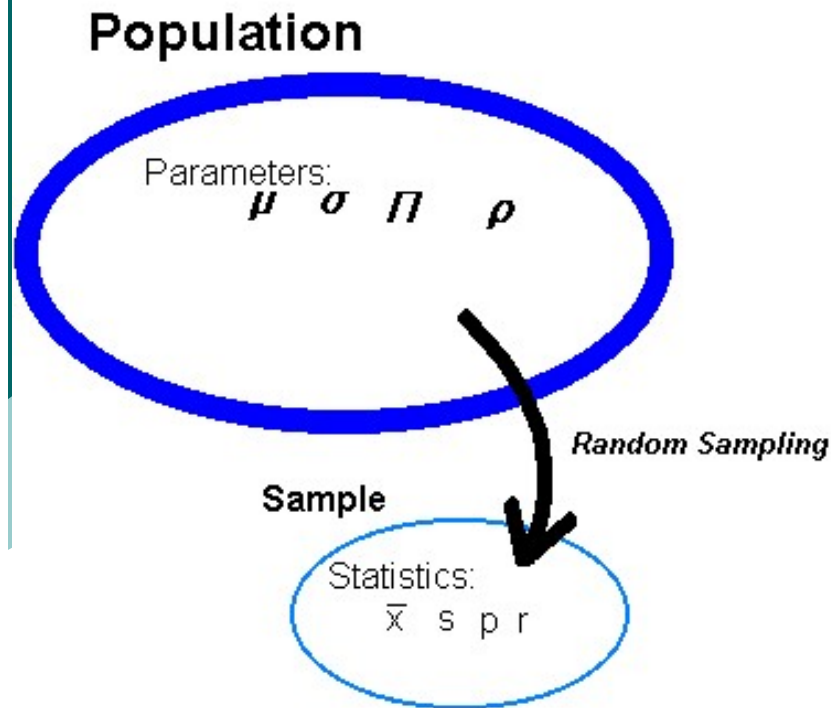
$$Z = \frac{\bar{x} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{52 - 54}{0.85} \approx -2.35$$

- Hence, we need to use the table to find $P(\bar{Z} \leq -2.35)$

$$0.5 - P(0 \leq Z \leq 2.35) = 0.5 - 0.4906 = 0.0094$$

- Hence, the probability of this happening is 0.00094.
We have 99.06% certain that this will not happen
(if the manufacturer's claim is correct!).

Statistical Inference



- ◆ Statistical inference is a formal process that uses information from a sample to draw conclusions about a population.
- ◆ It also provides a statement of how much confidence can be placed in the conclusion.

Statistical Inference

○ *Estimation*

- ◆ Estimating unknown value of a population parameter

○ *Hypothesis testing*

- ◆ Making decisions about the value of a parameter by testing a pre-conceived hypothesis

- Both types of inference are based on the **sampling distribution of statistics**
- Both report probabilities that state *what would happen if we used the inference method many times*

Statistical Inference

- *When do we use estimation?*

When we want to estimate the unknown population parameters and we do not have any previous knowledge about the population.



Point Estimate

- ◆ is a single number (**our best guess**), calculated from available sample data, that is used to estimate the value of an unknown population parameter.

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$



Questions:

- An unfair coin has a 75% chance of landing heads-up.

Let $X = 1$ if it lands heads-up,
and $X = 0$ if it lands tails-up.

- Find the sampling distribution of the mean \bar{X} for samples of size 3.



Solution:

- The experiment consists of tossing a coin 3 times and measuring the sample mean \bar{X}

	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
Probability	27/64	9/64	9/64	3/64	9/64	3/64	3/64	1/64
\bar{X}	1	2/3	2/3	1/3	2/3	1/3	1/3	0

The possible values of \bar{X} are 0, 1/3, 2/3 and 1.



The distribution of the sample mean is a binomial distribution

\bar{X}	0	1/3	2/3	1
$P(\bar{X} = \bar{x})$	1/64	9/64	27/64	27/64

Is the Sample Mean an Unbiased estimator of the Population Mean?

$$E[\text{sample mean}] = \text{population mean} ?$$



s^2 is an unbiased estimator of σ^2

$$\begin{aligned} E(s^2) &= E\left(\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2\right) \\ &= \frac{1}{n-1} E\left(\sum_{i=1}^n x_i^2 - 2\bar{x} \cdot \sum_{i=1}^n x_i + n\bar{x}^2\right) \\ &= \frac{1}{n} E\left(\sum_{i=1}^n x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right) \\ &= \frac{1}{n-1} E\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right) \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n E(x_i^2) - nE(\bar{x}^2) \right] \\ &= \frac{1}{n-1} \{n[\text{Var}(x_i) + E^2(x_i)] - n[\text{Var}(\bar{x}) + E^2(\bar{x})]\} \\ &= \frac{1}{n-1} \left[n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) \right] \\ &= \frac{1}{n-1} (n\sigma^2 - \sigma^2) \\ &= \sigma^2. \end{aligned}$$



s^2 is an unbiased estimator of σ^2

$$E[\textit{statistic}] = \textit{parameter}$$

$$E(s^2) = \sigma^2$$

How to do?

Refer to the example in Applied
Mathematics (Advanced Level) Syllabus



Example

- Let X_1 , X_2 and X_3 be random samples taken from a population with mean μ and σ^2

$$T_1 = \frac{X_1 + X_2 + X_3}{3}$$

$$T_2 = \frac{X_1 + 2X_2}{3}$$

$$T_3 = \frac{X_1 + 2X_2 + 3X_3}{3}$$

are estimators of μ . Which one are unbiased estimates of μ ?



Solution:

$$\begin{aligned}E[T_1] &= \frac{1}{3}\{E[X_1] + E[X_2] + E[X_3]\} \\&= \frac{1}{3}(\mu + \mu + \mu) \\&= \mu\end{aligned}$$

$$\begin{aligned}E[T_3] &= \frac{1}{3}E[X_1] + \frac{2}{3}E[X_2] + E[X_3] \\&= \frac{1}{3}\mu + \frac{2}{3}\mu + \mu \\&= 2\mu\end{aligned}$$

$$\begin{aligned}E[T_2] &= \frac{1}{3}E[X_1] + \frac{2}{3}E[X_2] \\&= \frac{1}{3}\mu + \frac{2}{3}\mu \\&= \mu\end{aligned}$$

$\therefore T_1$ and T_2 are unbiased estimate of μ