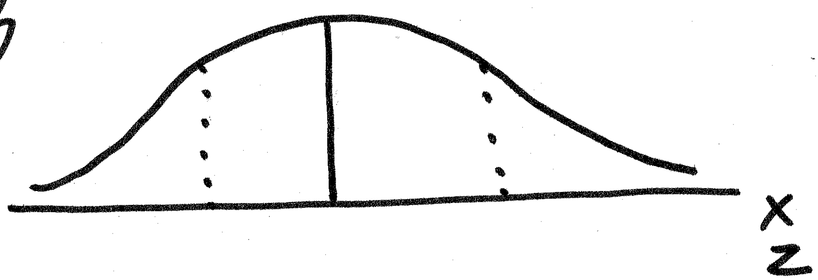
Chapter 7 BUSN 210 Statistics

1. Simple Random Samples
2. Point Estimate & Sampling Error
3. Sampling Distributions

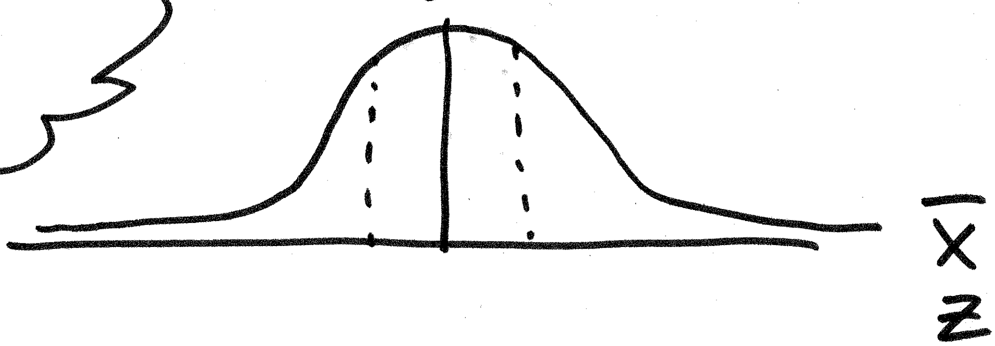
Chapter 6: Looked at X or P values & compared to Normal curve

X-values & Prob.



Chapter 7: Looked at X_{Bar} or P_{Bar} values & compare to Normal curve (sigma will be known)

X_{Bar} values & Prob.



Chapter 8: Look at X_{Bar} where sigma is not known & compare to a Normal curve (t-distribution)

Table of Contents

Sampling,
Sampling Distribution of Sample Means,
Central Limit Theorem

Element:
The entities on which data is collected (Primary Key)

Variable:
The characteristic of interest for the elements (Fields)

Observation:
The set of measurements obtained for a particular element

Data set:

Variable (Field) ==>

	Copmany	Total Sales	Earnings/ Share	Share Price	Mean Sales	Standard Deviation For Sales	
Element (Primary Key) ==>	Deere	$3B	$5.77	$71.00	$150.00	$75.00	<== Observation (Record)
	e Bay	$1.5B	$0.57	$43.00	$10.52	$9.00	
	ComCast	$2B	$0.43	$32.00	$68.95	$10.32	

Population:
All elements of interest

Sample:
A subset of the population

Why do we sample instead of look at whole population?

- We select sample to collect data to answer research question about a population
- Specific reasons:
  - The physical impossibility of checking all items in the population
    - Example:
      - Can't count all the fish in the ocean
  - The cost of studying all the items in a population
    - Example:
      - General Mills hires firm to test a new cereal:
        - Sample test: cost ≈ $40,000
        - Population test: cost ≈ $1,000,000,000
  - Contacting the whole population would often be time-consuming
    - Political polls can be completed in one or two days
    - Polling all the USA voters would take nearly 200 years!
  - The destructive nature of certain tests
    - Examples:
      - Test each bottle of wine?!!?
      - Testing all seeds from Burpee → there'd be none left
  - The sample results are usually adequate
    - Consumer price index constructed from a sample is an excellent estimate for a consumer price index that could be constructed from the population

<u>Statistical Inference:</u>

- The process of using data obtained from a sample to make estimates or test hypotheses about characteristics of the population (like mean).
- Draw reasonable conclusions about population from statistics

<u>Infer:</u>

Conclude from evidence

<u>Sampled Population:</u>

Population from which the sample is drawn

<u>Target Population:</u>

Population we want to make an inference about

- **Sampled Population** and **Target Population** are not always the same!
  - If you took a sample from a College Registration List, **Sampled Population** and **Target Population** are the same
  - If you take a sample from only matenee movie-goers and you want to make inferences about all movie goers your **Sampled Population** (matenee) is different than your **Target Population** (all movie goers): they are not the same.
  - Conclusion: When a sample is used to make inferences about the population, make sure that the sampled and target population are in close agreement. This is not a mathematical calculation, it is a judgment call.

Frame:

List of elements that sample will be selected from. It is not always possible to construct a **Frame**.

- **Frame** that CAN be constructed:
  - Take sample at Highline Community College to see how many people have iPods
  - Sampled Population = List of registered students
  - **Frame** = List of registered students
  - The sampled population has a finite number of elements
  - This is called "Sampling from a Finite Population". Use "Simple Random Sampling" method to select a sample
- **Frame** that CANNOT be constructed:
  - Population is too big (like counting all the fish in the sea) or not feasible (cots too much)
  - Take a sample of cereal box weights from a cereal box filling machine
    - Sampled Population = conceptual population of all boxes that could have been filled at that particular point in time. In this sense, the sampled population is considered infinite.
    - Frame = impossible to construct frame from infinite population because all the elements are not present
    - The sampled population has a conceptually infinite number of elements
    - This is called "Sampling from an infinite Population or Process". Use "Random Sampling" method to select a sample
      - "Random Sampling" is the same as "Simple Random Sampling", except for two assumptions have to hold true (more later)

Sampling from a Finite Population:

- Replacing each sampled element before selecting subsequent elements is called sampling with replacement
- Sampling without replacement is the procedure used most often
- In large sampling projects, computer-generated random numbers are often used to automate the sample selection process

Processes (Sampling from an infinite Population):

- Examples of processes:
  - Machine fills boxes of cereal
  - Machine fills bags of lettuce
  - Machines make bolts and screws for airplanes
  - Router makes boomerangs
  - Transactions occur at bank
  - Calls arrive at Highline help desk
  - Customers entering store
- All are viewed as coming from a process generating elements from a conceptually infinite population

Random Variable:

Numerical Description of the outcome of an experiment

- If we consider the process of selecting a "Random Sample" as an experiment, the X_{bar} is the numerical description of the outcome of the experiment. Thus X_{bar} is the random variable

Probability Sample

1. Each possible sample has a known probability of selection and a random process is used to select the elements for the sample. Good because we have techniques to help us understand if our sample is "good", reasonable – more later…
2. Examples:
    a. Simple Random Sample (Finite pop)
        i. A **simple random sample** of size ***n*** from a finite population of size ***N*** is a sample selected such that each possible sample of size ***n*** has the same probability of being selected
    b. Random Sample (Infinite population)
        i. A random sample of size n from an infinite population is a sample selected such that the following is true;
            1. Each element selected comes from same population
            2. Each element is selected independently
    c. Stratified Random Sampling (allows smaller sample size and lower cost)
        i. Population divided into mutually exclusive strata where elements in each strata are similar
        ii. Works best when the variation among the elements in each strata is relatively small.
        iii. Formulas available
    d. Cluster Sampling
        i. Elements in clusters are not a like – each cluster is like a min population
        ii. Helps to reduce cost.
    e. Systematic Sampling
        i. Like with invoices or other ordered populations.

Non-probability Sampling

1. Not good because we can't calculate how reasonable the sample results are
2. Convenience Sampling
3. Judgment Sampling

<u>Random Sample:</u>

1. **Simple Random Sample:**

- A sample selected so that each item or person in the population has the same chance of being included
- Used for Finite Populations
- How to select a sample:
    - **Select any n units in a random way**
        - Book method:
            - Step 1: Assign Random Number to each element in population
            - Step 2: Select "n" elements that have the "n" smallest (or largest) random numbers
                - Using Excel's RAND or RANDBETWEEN functions (each follows a Uniform Distribution between 0 and 1)
        - Names of classmates in a hat, mix up names, select until sample size, "n" is reached
        - There are other methods in Excel also

2. **Random Sample:**

- These must hold true:

    1) Each element selected comes from same population (Sampled and Targeted Populations are the same).

    2) Each element is selected independently – to prevent selection bias (prevent all from similar group, or similar attitudes, or choosing to get desired result)
- Used for Infinite Populations or Populations where it is not feasible to list all elements
- How to select a sample:
    - **Select any n units in a random way**
        - Examples:
            - Machines filling boxes or bags, choose sample from same point in time
            - People arriving at a restaurant, choose customer directly after customer who uses coupon (McDonald's did this to simulate a random selection.

Samples are only estimates:

- In point estimation we use the data from the sample to compute a value of a sample statistic that serves as an estimate of a population parameter.
- We refer to X_{bar} as the point estimator of the population mean **mew.**
- We refer to **s** as the point estimator of the population mean **sigma**.
- We refer to P_{bar} as the point estimator of the population mean **p**.

Sampling Error:

Does X_{bar} always equal **Mew**?
Rarely!

But if $\bar{X}$ is a point estimate for μ, what if they are different?

Example:

cereal Box filling machine

$\bar{X}$ = 14.14 oz. = sample of boxes

μ = 14 oz. = weight on box

$$\underbrace{\bar{X} - \mu = 14.14\text{ oz.} - 14\text{ oz.} = .14\text{ oz}}_{\text{sampling Error}}$$

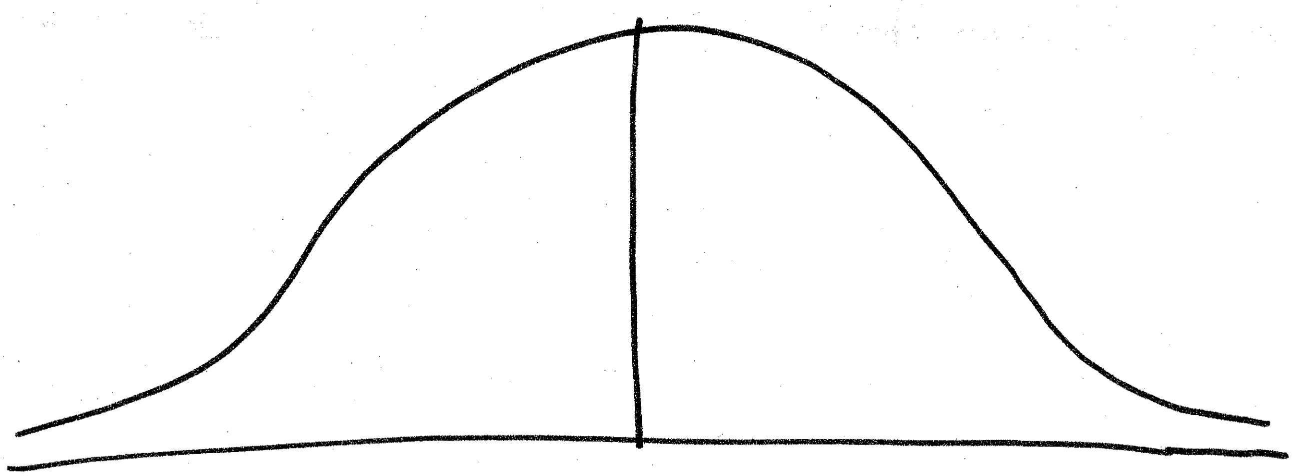
① Is this sampling error okay, and machine is filling properly?

② Is it not sampling error, and machine is off?

To answer this Question we need to learn about the :

$$\boxed{\text{Sampling Distribution of } \bar{X}}$$

$\bar{X}$ is Random variable



$\bar{X}$

$$\mu = \mu_{\bar{x}} = E(\bar{x})$$

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

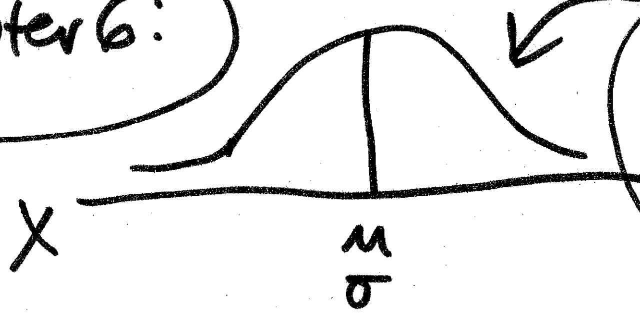
$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

In chapter 6 we talked about X values & Normal Probability Distributions.

we would like to begin to talk about $\bar{X}$ & the sampling Distribution of $\bar{X}$ (SDo$\bar{X}$)

If we can talk about $\bar{X}$ & the SDo$\bar{X}$, we will be able to take samples & campare $\bar{X}$ to SDo$\bar{X}$ in order to make reasonable conclusions about the population.

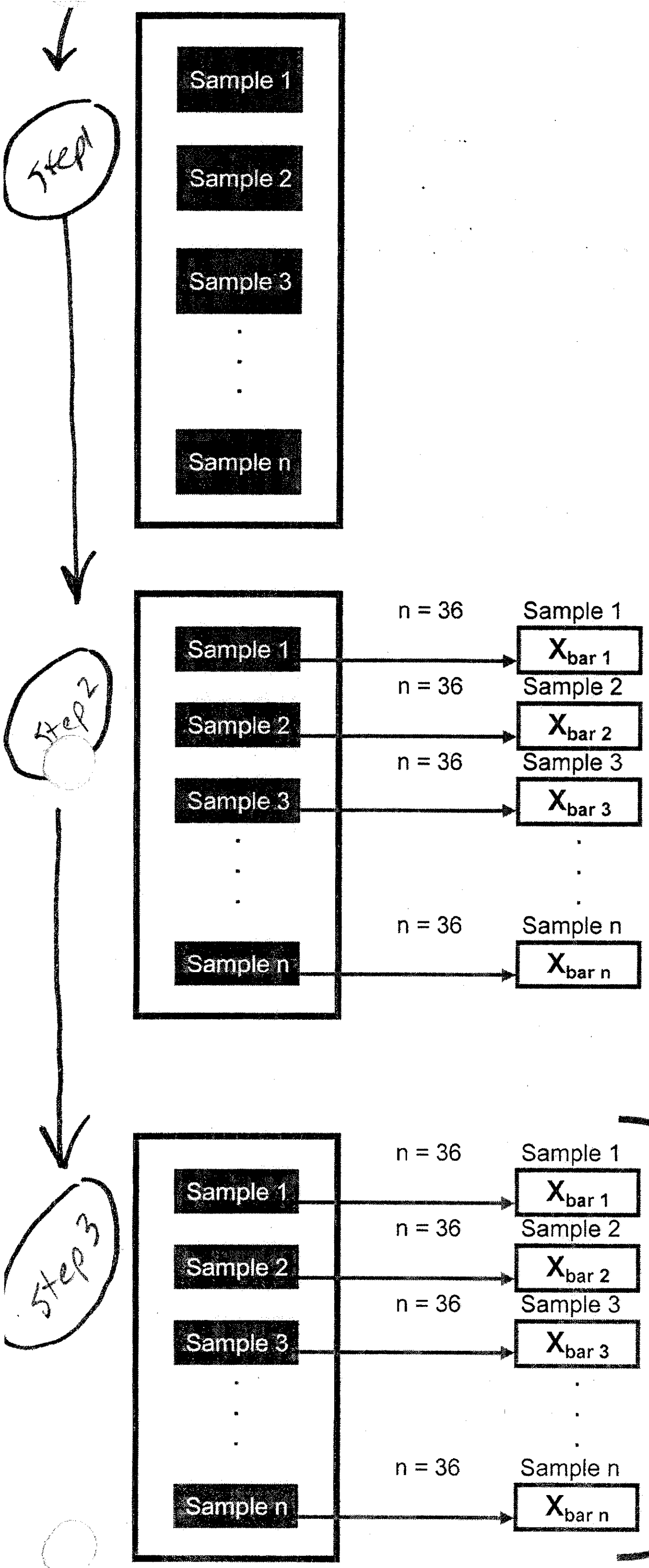
Chapter 6:



X is what we were talking about no $\bar{X}$

But if we begin to talk about $\bar{X}$, what about our sampling error, $\bar{X} - \mu$? Is it OK to take a sample and use the sample mean, $\bar{X}$, to say something about the population? For example, if we are manufacturer that fills cereal boxes and we take a sample of box weights and get $\bar{X} = 14.14$ oz. and the box is supposed to weigh 14 oz., is the filling machine putting too much into the box or is the sampling error ($\bar{X} - \mu$, $14.14 - 14 = .14$) acceptable? We must investigate further →

How to construct The Sampling Distribution Of The Sample Mean X_{bar}

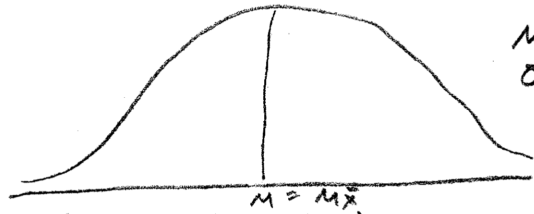


Plot All X_{bar}

step4 — We will Discover

① $E(\bar{x}) = \mu_{\bar{x}} = \mu$

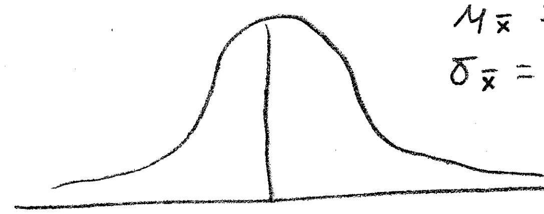
②



μ
σ

population spread is greater

③



$\mu_{\bar{x}} = \mu$
$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

$\mu_{\bar{x}}$
$\sigma_{\bar{x}}$

Sampling Distribution of $\bar{x}$ is less

④ Sampling Distribution of $\bar{X}$ is Normally Distributed (if n big)

⑤

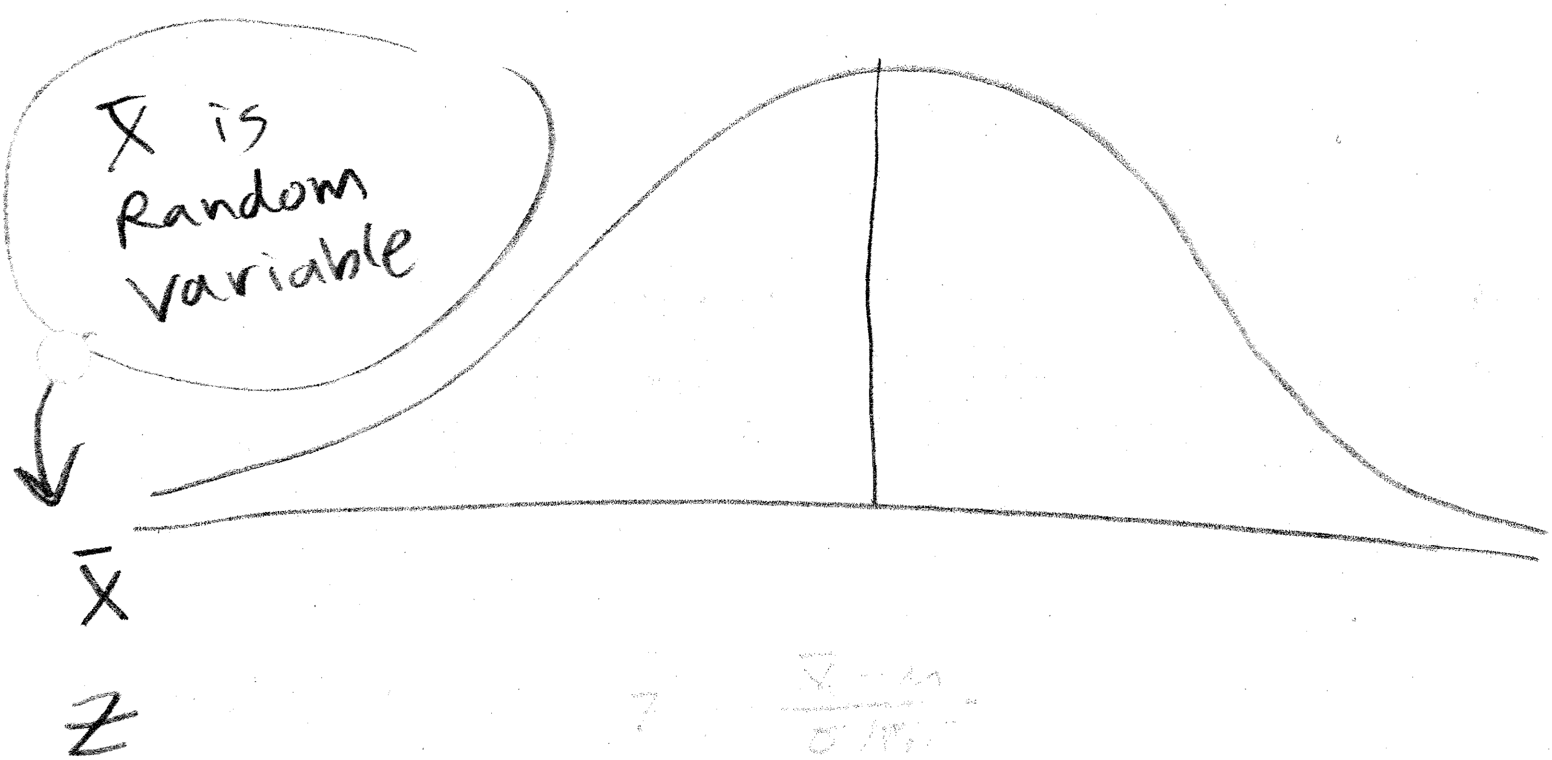
$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

we will look at Example in Excel

Sampling Distribution of $\bar{x}$ (SDO$\bar{x}$)

12.5

probability Distribution of all possible values of sample mean $\bar{x}$



$$\mu = \mu_{\bar{x}} = E(\bar{x})$$

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

we can now take a sample & compare it to SDO$\bar{x}$ & see if our sample seems reasonable or not.

Expected value of $\bar{X}$ of SDO$\bar{X}$

or

Mean of Sampling Distribution of $\bar{X}$

$$E(\bar{x}) = \mu_{\bar{x}} = \mu$$

$$\underbrace{\mu_{\bar{x}} = \frac{\text{Sum of all possible Sample Means}}{\text{Total number of Samples}}}$$

if we are able to select all possible samples of a particular size from a given population, then the Mean of Sampling Distribution of $\bar{x}$ is equal to population Mean.

we'll do an example in Excel.

Standard Deviation of Sampling Distribution of $\bar{X}$ or Standard Error

Infinite Population

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

Finite Population

Finite population correction factor

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}}$$

when $\frac{n}{N} \leq 0.05$

then simply:

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

text assumes: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ unless stated otherwise

why:
usually populations are very large & sample size very big, so correction factor close to 1 (No affect)

Z for Sampling Distribution of $\overline{X}$

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

Sampling error

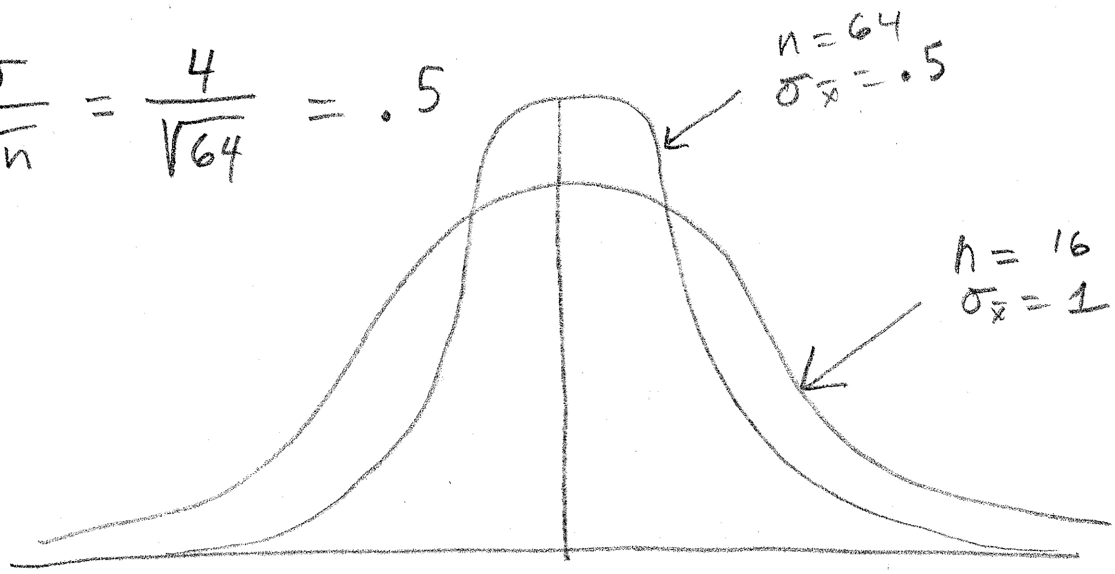
Standard Error

Relationship between sample size and the Sampling Distribution of $\bar{x}$ (sample mean)

* As sample size n increases, the Standard Error $\frac{\sigma}{\sqrt{n}}$ decreases

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{16}} = 1$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{64}} = .5$$



This means:

* The bigger the n, the higher the probability that the sample mean falls within a specified distance of the population mean

Leading up to the Central Limit Theorem:

- If all samples of a particular size are selected from any population, the sampling distribution of the sample mean X_{bar} is approximately a normal distribution. This approximation improves with larger samples → See Next Page

Central Limit Theorem:

- In selecting random samples of size n from a population, the sampling distribution of the sample mean X_{bar} can be approximated by a normal distribution as the sample size becomes large
    - If population distribution is symmetrical but not normal, the distribution will converge toward normal when n > 10
    - Skewed or thick-tailed distributions converge toward normal when n > 30
    - Heavily skew distributions converge n > 50

Use of Central Limit Theorem:

- We can reason about the Sampling Distribution of Xbar with absolutely no information about the shape of the original distribution from which the sample is taken
- This means that:
    - We can take one sample and compare it to the Standard Normal Curve (NORM.S.DIST) or Normal Curve (NORM.DIST) to see if our sample result is reasonable or not.
    - If it is reasonable, the process or claim is reasonable
    - If it is not reasonable, the process or claim is not reasonable

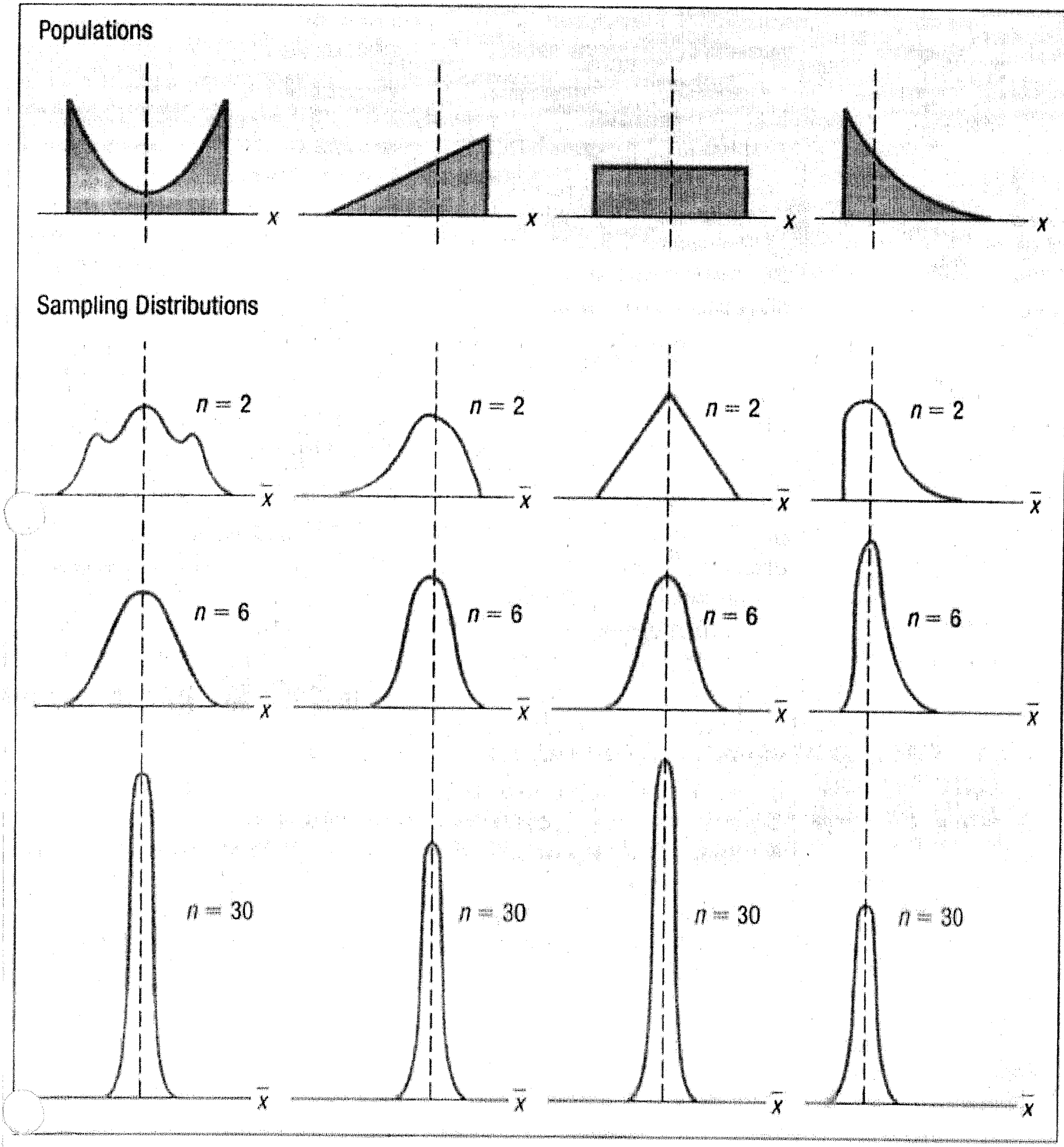


CHART 8–2 Results of the Central Limit Theorem for Several Populations

Business Decisions Example 1

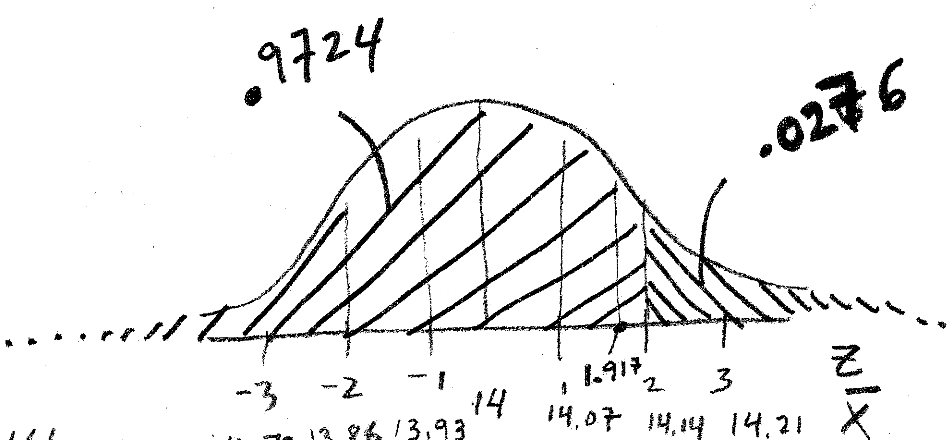
- History for a food manufacturer shows the weight for a Chocolate Covered Sugar Bombs (popular breakfast cereal) is:
  - $\mu = 14$ oz.
  - $\sigma = .4$ oz.
- If the morning shift sample shows:
  - $X_{bar} = 14.14$ oz.
  - $n = 30$
- Is this sampling error reasonable, or do we need to shut down the filling operations?

32

① variables

$\mu = 14$ oz.
$\sigma = .4$ oz
$\bar{X} = 14.14$ oz
$n = 30$

② Draw Picture



$\mu = \mu_{\bar{X}} = 14$ oz

$$\frac{\sigma}{\sqrt{n}} = \frac{.4}{\sqrt{30}} = .07303$$

Standard Deviation of sample Means "Standard Error" Because Distr. of $\bar{X}$ is less spread out

③ Calculate Z

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{30}}} = \frac{14.14 - 14}{.07303}$$

$$Z_{14.14} = 1.917$$

④ conclude

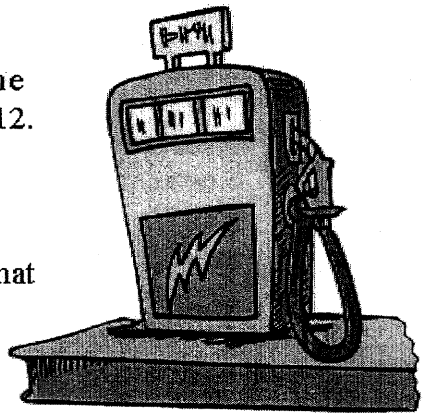
The probability associated with $\bar{X} = 14.14$ oz. or greater is .0276. This is low. It is unlikely that we could have taken a sampe of 14.14 & had the sample error $(14.14 - 14 = .14)$ occur by chance ....

④ conclude continued...

Because it is unlikely that the sample error is due to chance, the 14.14 probably represents a machine that is filling too much.

Shut down and Fix

Suppose the mean selling price of a gallon of gasoline in the United States is \$3.12. (μ) Further, assume the distribution is positively skewed, with a standard deviation of \$0.98 (σ). What is the probability of selecting a sample of 35 gasoline stations (n = 35) and finding the sample mean within \$.33?



① variables

$\mu = \$3.12$

sigma $= \sigma = \$.98$

$n = 35$

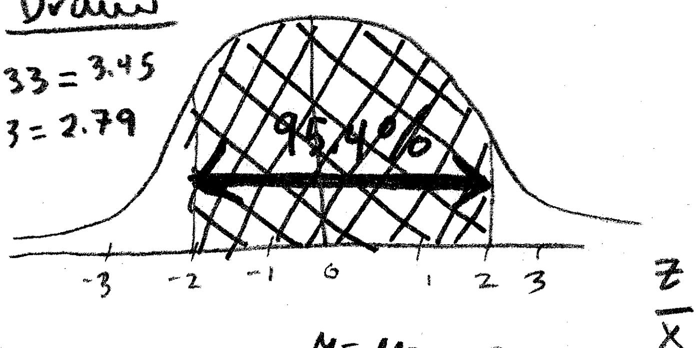
$\left\{ \begin{array}{l} \text{distance on either} \\ \text{side of } \mu \end{array} \right\} = .33$

est. $\bar{X}_1 = 3.12 + .33 = 3.45$

est. $\bar{X}_2 = 3.12 - .33 = 2.79$

② Draw

$\pm 2 = 95.4\%$



$\mu = \mu_{\bar{x}} = 3.12$

$\frac{\sigma}{\sqrt{n}} = .16565$

$\left\{ \text{Standard error} = \text{SD of Distribution of sample means} \right\} = \frac{\sigma}{\sqrt{n}} = \frac{.98}{\sqrt{35}} = .16565$

③ Calculate $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

$z_{3.45} = \frac{3.45 - 3.12}{.16565} = 1.99 \approx 2$

$z_{2.79} = \frac{2.79 - 3.12}{.16565} = -1.99 \approx -2$

④ The probability of selecting a sample of 35 gas stations & finding the sample mean within \$.33 of \$3.12 is .954.

↳ alternative ways of stating answer →

Alternative ways to state Answer:

21

1 "simple random sample of 35 gas stations has a .954 probability of providing a sample mean $\bar{x}$ that is within \$.33 of the population mean of \$3.12."

OR

2 ".046 probability that the sampling error will be more than $\pm$ \$.33."

The Sampling Distribution can be used to provide probability information about how close the sample mean is to the population mean μ

Sample Proportion

$\bar{p} = \frac{x}{n}$ = Sample proportion = Random Variable

x = the number of elements in the sample that possess the characteristic of interest

n = sample size

Note: x is a binomial variable
Bi = 2
Nomial = Nominal variable

Sampling Distribution of $\bar{p}$

① The sampling Distribution of $\bar{p}$ is the probability distribution of all possible values of the sample proportion $\bar{p}$.

② The sampling Distribution of $\bar{p}$ can be approximated by a Normal distribution whenever:

$$n * p \geq 5$$

and

$$n * (1 - p) \geq 5$$

Expected Value of $\bar{p}$

$$E(\bar{p}) = p$$

$E(\bar{p})$ = Expected value of $\bar{p}$ = unbiased estimator

p = population proportion

Standard Deviation of $\bar{p}$

Finite population

$$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}} * \sqrt{\frac{p*(1-p)}{n}}$$

$*$ if $\frac{n}{N} \le .05$ use : $\sigma_{\bar{p}} = \sqrt{\frac{p*(1-p)}{n}}$

Infinite pop. or process or Not feasible to List all Elements

$$\sigma_{p} = \sqrt{\frac{p*(1-p)}{n}}$$

Example:

If $p = .55$, $n = 30$

and you want to find probability of finding $\bar{p}$ within a margin of error of .05:

$n*p = .55 * 30 = 16.5$
$n*(1-p) = .45 * 30 = 13.5$

$$\sigma_p = \sqrt{\frac{.55*.45}{30}} = .09083$$

Probability that $\bar{p}$ will lie between .5 & .6 is:

=NORMDIST(.6, .55, .09083, 1) - NORMDIST(.50, .55, .09083, 1)
= .418011