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

See Fish Article:

Scientists Conduct First-Ever Fish Census

Posted on: Thursday, 23 October 2003, 06:00 CDT ,
http://www.redorbit.com/news/general/35372/scientists_conduct_firstever_fish_census/

An unprecedented census of life in the world's oceans is discovering three new fish species a week on average and predicts as many as 5,000 unknown fish species may be lurking undetected, according to the first interim report.

By the time they're done in 2010, scientists say they may find more than 2 million different species of marine life.

Three hundred scientists from 53 countries participating in the $1 billion study reported their first findings Thursday, three years into the decade-long project. So far, the Census of Marine Life comprised 15,304 species of fish and 194,696 to 214,696 species of animals and plants, estimated to be roughly 10 percent of the world's total.

The census is adding about 150 to 200 species of fish and 1,700 species of animals and plants each year.

The scientists said they believe the oceans that extend across 70 percent of Earth's surface hold about 20,000 species of fish and up to 1.98 million species of animals and plants. Many of those could be basic and small life forms, such as worms and jellyfish.

"We've tended to be interested in the things that we eat," said Jesse Ausubel, an environmental scientist at The Rockefeller University in New York City. He helps run the census for the Alfred P. Sloan Foundation, which provided $20 million in funding.

"We've tended not to be interested in the things that pass through our nets or don't taste good," Ausubel said. "But the small critters are tremendously important in the ecosystem ... and in an evolutionary sense, the small things came first. They're ancient, and they're survivors."

Scientists hope to gain a better understanding of life in the mostly unexplored seas, learning about evolution and climate, pole to pole. Environmentalists hope to use it to counter overfishing and pollution that has depleted the ocean's resources. Industry hopes it will lead to more efficient fishing and shipping, new pharmaceuticals and industrial compounds.

Time 2000 - 2010

"Our goal by 2010 is to know as much about life in the oceans as we know about life on land now," said Ronald O'Dor, a marine biologist at Dalhousie University in Canada and the project's chief scientist.

"No one would claim that we know everything about life on land," he said. "There are probably still a few hundred thousand beetles in tropical forests that haven't been described. But we'd like to aim for parity."

The project grew from scientists' concerns after a 1995 report by the National Academy of Sciences found that human population growth was fast changing the diversity of life in the oceans, possibly irreversibly.

They wanted to learn what still was there.

The census started organization six years ago, partly through the efforts of J. Frederick Grassle, director of Rutgers University's Institute of Marine & Coastal Sciences. Actual work began in 2000. It has cost $70 million so far and the price tag eventually is expected to reach $1 billion, paid by participating governments.

Their work may never really be finished. - impossible to count all.

"We know we won't have counted every animal," said Grassle, who chairs the census's scientific steering committee. "The limit on the knowable is in major part the resources that can go into the problem."

Statistical Inference:
The process of using data obtained from a sample to make estimates or test hypotheses about characteristics of the population (like mean)
Infer:
Conclude from evidence

Sampled Population:
Population from which the sample is drawn
Target Population:
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, they are the same
  - If your goal is to take a sample of all movie goers and you sample only matinee movie goers:
    - **Sampled Population** = matinee movie goers
      and
    - **Target Population** = all movie goers
      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
- How a sample can help to decide whether the process is working properly:
  - Processes not working properly (like machine filling too much) will produce sample statistics that are not close (statistically significant) to the population parameter
  - Processes working properly (like machine filling just right) will produce sample statistics that are close (statistically insignificant) to the population parameter

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

Random Sample:

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**
  - Names of classmates in a hat, mix up names, select until sample size, "n" is reached
  - Using Excel's RAND() function to select a sample from a population
  - There are other methods in Excel also

See → Excel workbook

2. **Random Sample:**

- These must hold true:
  - Each one of the sampled elements is independent (each has no affect on others) of the other elements in the population
  - Each one of the sampled elements follows the same probability distribution as the elements in the population
- 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**

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 $\mathbf{X_{bar}}$ as the point estimator of the population mean **mu**.
- We refer to **s** as the point estimator of the population mean **sigma**.
- We refer to $\mathbf{P_{bar}}$ as the point estimator of the population mean **p**.

Sampling Error:

Does $\mathbf{X_{bar}}$ always equal **Mu**?
Rarely!

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

Example

$\bar{X}$ = weight of cereal box = 14.14 oz

μ = 14 oz.

$\bar{X} - \mu$ = 14.14 oz − 14 oz = .14 oz

4 **Sampling Error** → $\bar{x} - \mu$, $\hat{p} - p$, $s - \sigma$

Why is there sampling Error?

Because the sample only uses some of values.

Example:

Pop	sample 1
12	16
13	11
11	10
10	12
12	
15	$\bar{X}$ = 12.25
16	

μ = 12.71429

{Sample Error} = $\bar{X} - \mu$ = 12.25 − 12.71 = −0.46

Q1: Is this sampling error acceptable?

Q2: Is our estimate of 12.25 good?

Q3: How close is the estimate of 12.25 to our pop. mean?

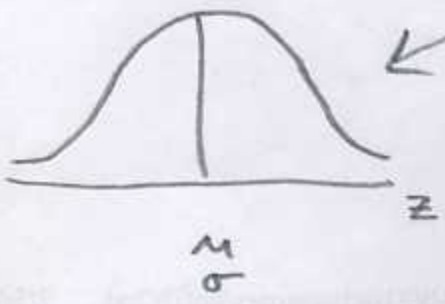
Q4: What sampling methods must we use to get good estimates?

Q5: How large might the sample error be? what risk of being wrong are we willing to take?

In chapter 5-6 we talked about Probability Distributions concerning X, and standard Normal curve

We would like to begin to talk about $\bar{X}$ & the standard Normal curve (SNC).

If we can talk about $\bar{X}$ & the SNC, we will be able to test things (take samples) ourselves and compare our $\bar{X}$ (from our sample) to the SNC in order to make reasonable conclusions about the population.



X is what we have been talking about

NOT $\bar{X}$!!

we must investigate

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}

Step 1

Sample 1
Sample 2
Sample 3
.
.
.
Sample n

Step 2

Sample 1	n = 36	Sample 1	$X_{bar\ 1}$
Sample 2	n = 36	Sample 2	$X_{bar\ 2}$
Sample 3	n = 36	Sample 3	$X_{bar\ 3}$
.	.	.	.
.	.	.	.
.	.	.	.
Sample n	n = 36	Sample n	$X_{bar\ n}$

Step 3

Sample 1	n = 36	Sample 1	$X_{bar\ 1}$
Sample 2	n = 36	Sample 2	$X_{bar\ 2}$
Sample 3	n = 36	Sample 3	$X_{bar\ 3}$
.	.	.	.
.	.	.	.
.	.	.	.
Sample n	n = 36	Sample n	$X_{bar\ n}$

Plot All X_{bar}

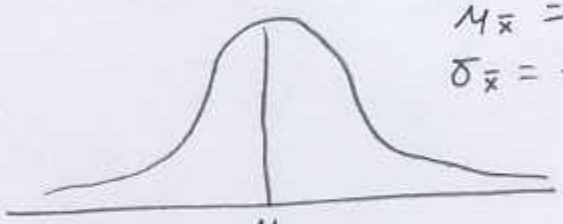
Step 4

We will Discover

① $E(\bar{x}) = \mu_{\bar{x}} = \mu$

② 

population spread is greater

③ 

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

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

⑤ $$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Mean Of The Distribution Of The Sample Mean

$$\mu_{\bar{X}} = \frac{\text{Sum of all sample means}}{\text{Total number of samples}}$$

- If we are able to select all possible samples of a particular size from a given population, then the mean of the Sampling Distribution Of The Sample Mean will exactly equal the population mean:

$\mu_{\bar{X}} = \mu =$ Mean of the Distribution of the Sample Means

- Even if we do not select all possible samples, they will be approximately equal:

$\mu_{\bar{X}} \approx \mu \approx$ Mean of the Distribution of the Sample Means

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

Standard Deviation Of The Sampling Distribution Of The Sample Mean

(Standard Error Of The Mean)

- There is less dispersion in the sampling distribution of the sample mean than in the population (each value is an average!!)

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \text{ SD of the Sampling Distribution of the Sample Mean}$$

- σ = population standard deviation
- n = sample size
- When we increase "n" the standard deviation of the sample will decrease

12 Standard Deviation of $\bar{x}$ (Finite Population)

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

$$\left\{\begin{array}{c}\text{Finite population correction Factor}\end{array}\right\} = \sqrt{\frac{N-n}{N-1}}$$

use when $\frac{n}{N} \leq .05$

sample size is less than or equal to 5% of population size.

* Text assumes:

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

unless specifically stated

Why:

usually populations are very big and samples small, so the correction factor has little affect.

Z-Value

- To determine the probability a sample mean falls within a particular region, use:

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

Sampling error

Standard error of sampling distribution of the sample mean

standard Deviation of Sampling Distribution of $\overline{X}$

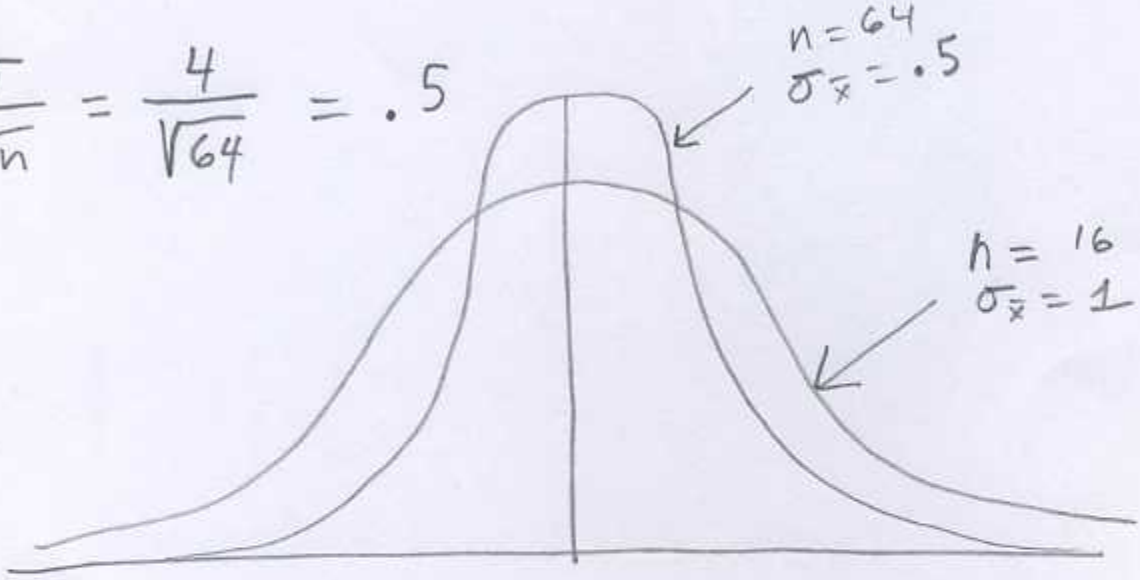
We are interested in the distribution Xbar, the sample mean, instead of X

13 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 simple 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 (NORMSDIST) or Normal Curve (NORMDIST) 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

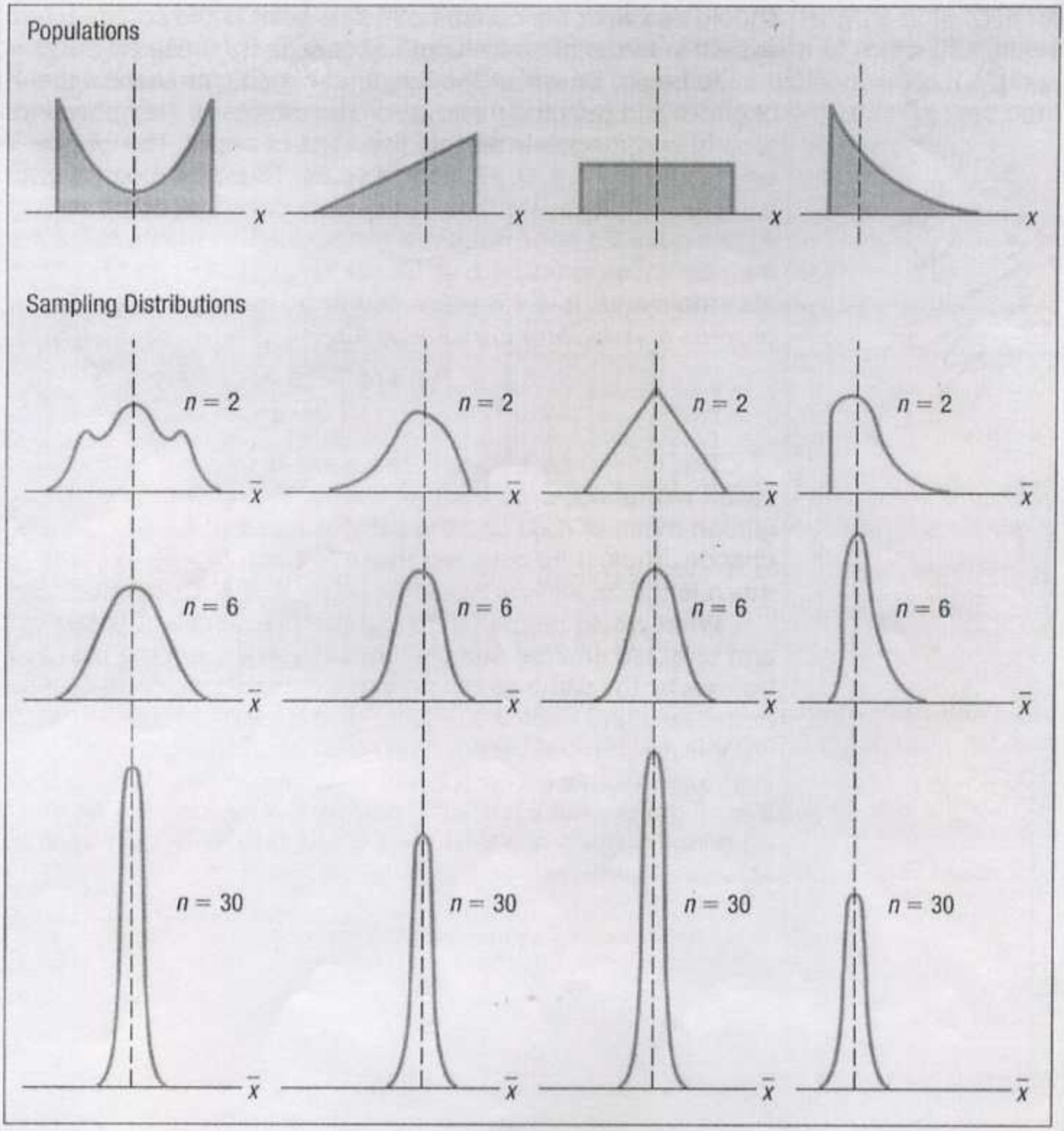


CHART 8–2 Results of the Central Limit Theorem for Several Populations

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?

(4) conclude continued...

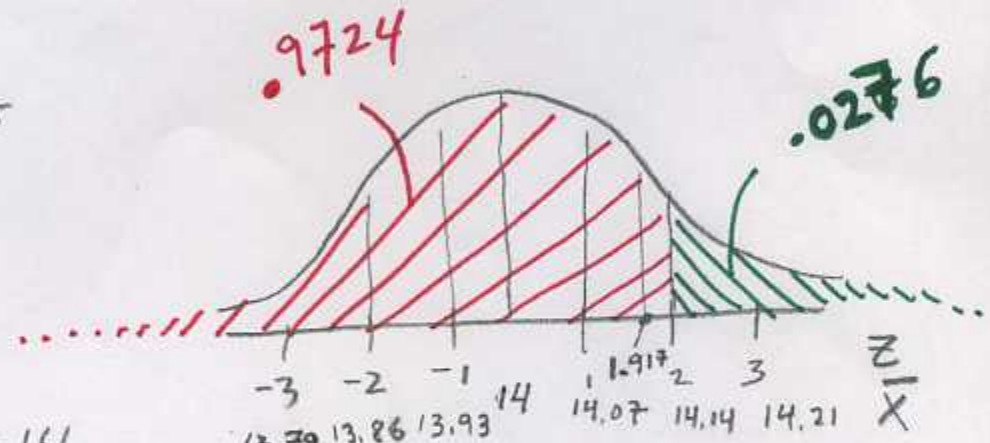
Because it is unlikely that the sample error is due to chance, the 14.14 probably represents a machine that is filling too much.

shut down and FIX

(1) variables

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

(2) Draw Picture



(3) Calculate Z

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

$$Z_{14.14} = 1.917$$

$M = M_{\bar{X}} = 14$ oz

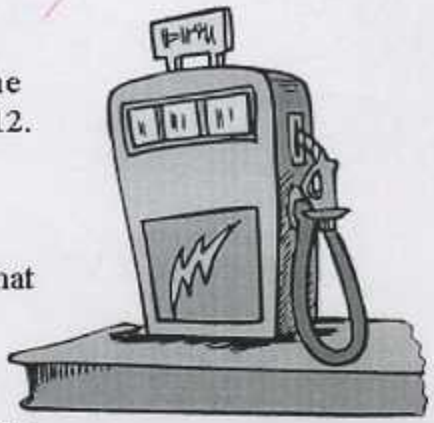
$\frac{\sigma}{\sqrt{n}} = \frac{.4}{\sqrt{30}} = .07303$

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

(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 sample of 14.14 & had the sample error ($14.14 - 14 = .14$) occur by chance ....↗

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 = $\sigma = \$.98$

$n = 35$

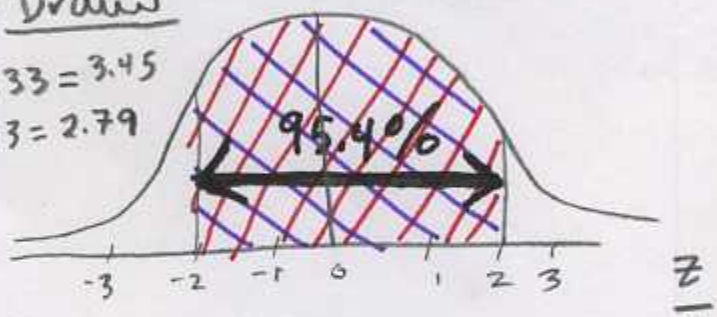
{distance on either side of μ} = .33

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

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

② Draw

$\pm 2 = 95.4\%$



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

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

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

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:

(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

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 error of the proportion"

(17) # Standard Deviation of $\bar{p}$ (p. 25)

Finite population

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

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

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

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

Example;

If $p = .55$, $n = 30$ and you want to find probability of finding $\bar{p}$ within a margin of error of .05:

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

$$\sigma_{\bar{p}} = \sqrt{\frac{.55*.45}{30}} = .09083$$

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

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

The Crossett Trucking Company

The Crossett Trucking Company claims that the mean weight of their delivery trucks when they are fully loaded is 6000 lbs. And the standard deviation is 150 lbs. Assume that the population follows the normal distribution. 40 trucks are randomly selected and weighed.

Within what limits will 95% of the sample means occur?

This problem is similar to chapter 8. In chapter 7, we know what the population mean, μ, is, and we can say things like "we are 95% sure that $\bar{x}$ occurs between 6047 lbs. and 5954 lbs." In chapter 8, we do not know what the population mean, μ, is and we say things like "we are 95% sure that μ occurs between two calculated values."

(1) $\mu = 6000$ lbs.
$\sigma = 150$ lbs.
$n = 40$

{standard error "standard deviation for Distribution of Sample Mean"} $= \frac{\sigma}{\sqrt{n}} = \frac{150 \text{ lbs.}}{\sqrt{40}} = 23.71708$

(2) In this problem we are not given $\bar{x}$ and asked to find the probability, we are given the probability and asked to find the $\bar{x}$. → Remember we can solve for $\bar{x}$ from our z formula →

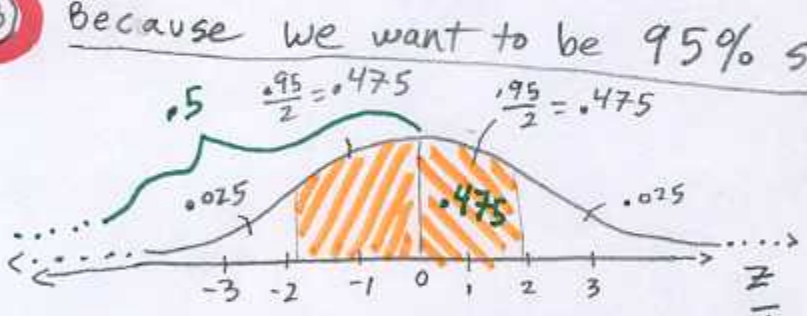
$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad (1)$$

$$z\frac{\sigma}{\sqrt{n}} = \bar{x} - \mu \quad (2)$$

$$\mu + z\frac{\sigma}{\sqrt{n}} = \bar{x} \quad (3)$$

$$\bar{x} = \mu + z\frac{\sigma}{\sqrt{n}}$$

(3) Because we want to be 95% sure, we need to divide 95% / 2



$\mu = \mu_{\bar{x}} = 6000$ lbs

{SD of sample mean} $= \frac{\sigma}{\sqrt{n}} = 23.71708$

(4) Find z
① Book Table Method, look .475 in Table
or
② = NORMSINV(.475 + .5) ≈ 1.96

(5) Find $\bar{x}$ (one on upperside and one on lower side)

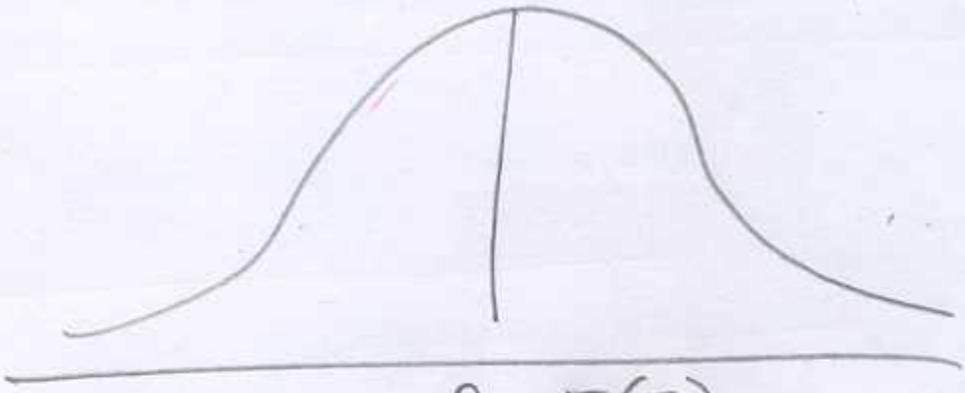
$$\bar{x} = \mu \pm z\frac{\sigma}{\sqrt{n}}$$

$6000 \pm 1.959964 * 23.71708$

$\bar{x}$ Rounded to the pound: 6046 and 5954

Answer:

It is reasonable to assume that the sample means for truck weight will occur between the limits 5954 lbs. and 6046 lbs. 95% of the time. However we do run of 5% risk that they will not occur between our limits.



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

$\sigma_{\bar{p}} = \sqrt{\frac{.3 * .7}{100}} = .045825757$ ✓

$5 \leq n * p = 100 * .3 = 30$ ✓

$5 \leq n * (1-p) = 100 * .7 = 70$ ✓