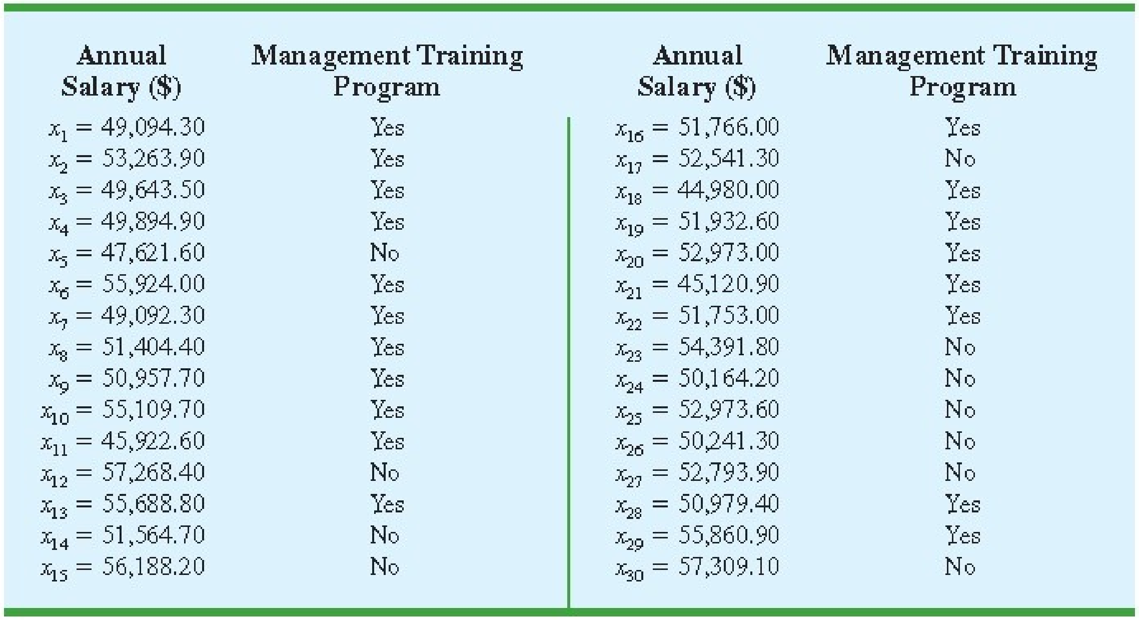
Sampling Distribution and Hypothesis Testing

Dr. P. Kalpana, M.E., PhD.
Faculty of Mechanical Engineering
IIITDM Kancheepuram

Point Estimation-The Electronics Associates Sampling Problem

- The director of personnel for Electronics Associates, Inc. (EAI), has been assigned the task of developing a profile of the company’s 2500 managers. The characteristics to be identified include the mean annual salary for the managers and the proportion of managers having completed the company’s management training program.

ANNUAL SALARYAND TRAINING PROGRAM STATUS FOR A SIMPLE RANDOM SAMPLE OF 30 EAI MANAGERS



Annual Salary ($)	Management Training Program	Annual Salary ($)	Management Training Program
x_1 = 49,094.30	Yes	x_{16} = 51,766.00	Yes
x_2 = 53,263.90	Yes	x_{17} = 52,541.30	No
x_3 = 49,643.50	Yes	x_{18} = 44,980.00	Yes
x_4 = 49,894.90	Yes	x_{19} = 51,932.60	Yes
x_5 = 47,621.60	No	x_{20} = 52,973.00	Yes
x_6 = 55,924.00	Yes	x_{21} = 45,120.90	Yes
x_7 = 49,092.30	Yes	x_{22} = 51,753.00	Yes
x_8 = 51,404.40	Yes	x_{23} = 54,391.80	No
x_9 = 50,957.70	Yes	x_{24} = 50,164.20	No
x_{10} = 55,109.70	Yes	x_{25} = 52,973.60	No
x_{11} = 45,922.60	Yes	x_{26} = 50,241.30	No
x_{12} = 57,268.40	No	x_{27} = 52,793.90	No
x_{13} = 55,688.80	Yes	x_{28} = 50,979.40	Yes
x_{14} = 51,564.70	No	x_{29} = 55,860.90	Yes
x_{15} = 56,188.20	No	x_{30} = 57,309.10	No

- To estimate the value of a population parameter, we compute a corresponding characteristic of the sample, referred to as a **sample statistic**

Point Estimation

- The sample mean is

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{1{,}554{,}420}{30} = \$51{,}814$$

- The sample standard deviation is

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{325{,}009{,}260}{29}} = \$3348$$

- To estimate p, the proportion of managers in the population who completed the management training program, sample proportion

- n=30, x=19

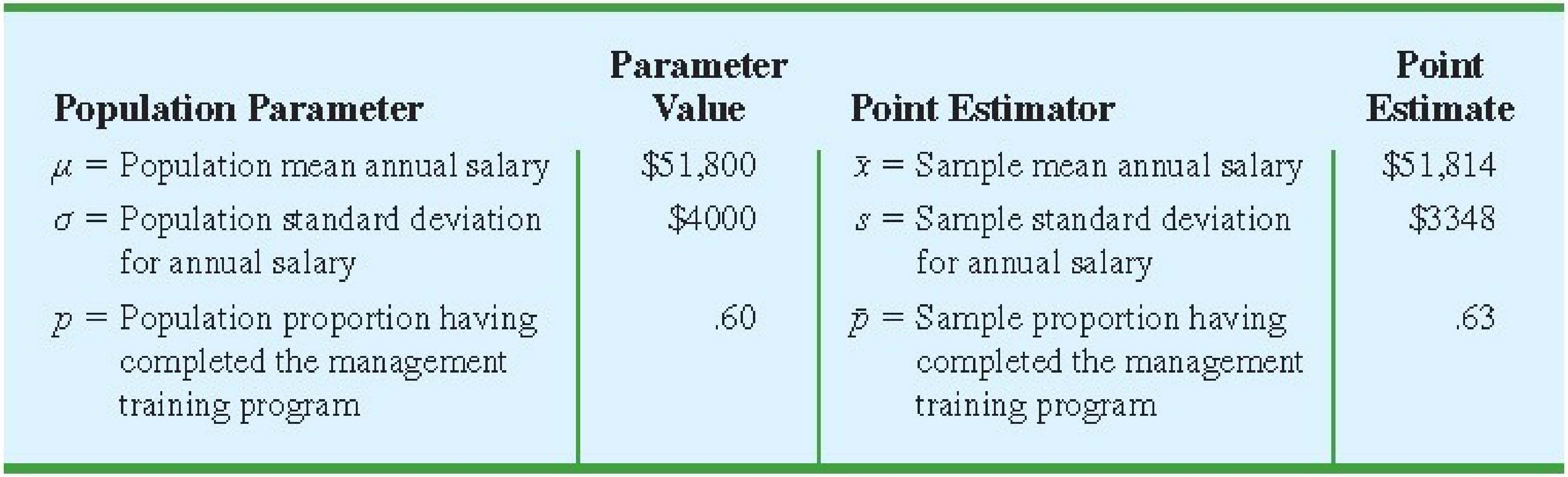
$$\bar{p} = \frac{x}{n} = \frac{19}{30} = .63$$

This statistical procedure called point estimation

Point Estimation

The sample mean $\bar{x}$ as the **point estimator** of the population mean μ
The sample standard deviation ***s*** as the point estimator of the population standard deviation σ
The sample proportion $\bar{p}$ as the point estimator of the population proportion ***p***
The numerical value obtained for $\bar{x}$, s,$\bar{p}$ or is called the point estimate

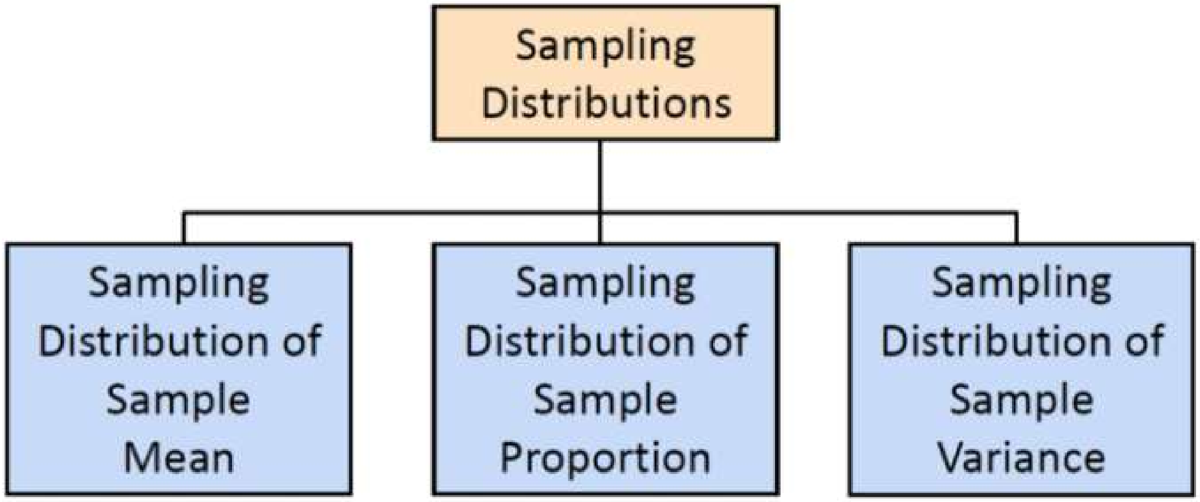
SUMMARY OF POINT ESTIMATES OBTAINED FROM A SIMPLE RANDOMSAMPLE OF 30 EAI MANAGERS



Population Parameter	**Parameter Value**	**Point Estimator**	**Point Estimate**
μ = Population mean annual salary	\$51,800	$\bar{x}$ = Sample mean annual salary	\$51,814
σ = Population standard deviation for annual salary	\$4000	s = Sample standard deviation for annual salary	\$3348
p = Population proportion having completed the management training program	.60	$\bar{p}$ = Sample proportion having completed the management training program	.63

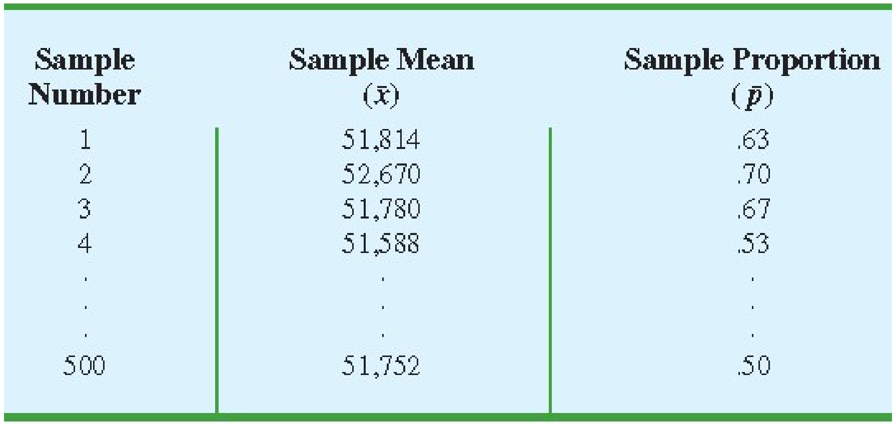
Introduction to Sampling Distributions

The sampling distribution sampling distribution is a distribution of all of the possible values of your statistic for a given size sample selected from the population


Sampling Distributions
Sampling Distribution of Sample Mean
Sampling Distribution of Sample Proportion
Sampling Distribution of Sample Variance

Sampling Distributions of Sample mean $\bar{x}$

VALUES OF AND FROM 500 SIMPLE RANDOM SAMPLES OF 30 EAI MANAGERS

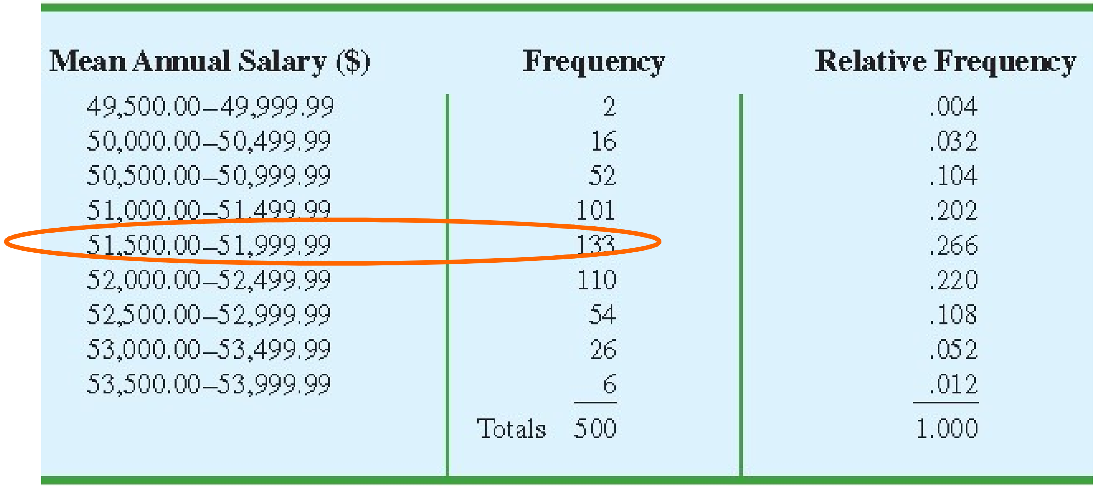


Sample Number	Sample Mean ($\bar{x}$)	Sample Proportion ($\bar{p}$)
1	51,814	.63
2	52,670	.70
3	51,780	.67
4	51,588	.53
.	.	.
.	.	.
.	.	.
500	51,752	.50

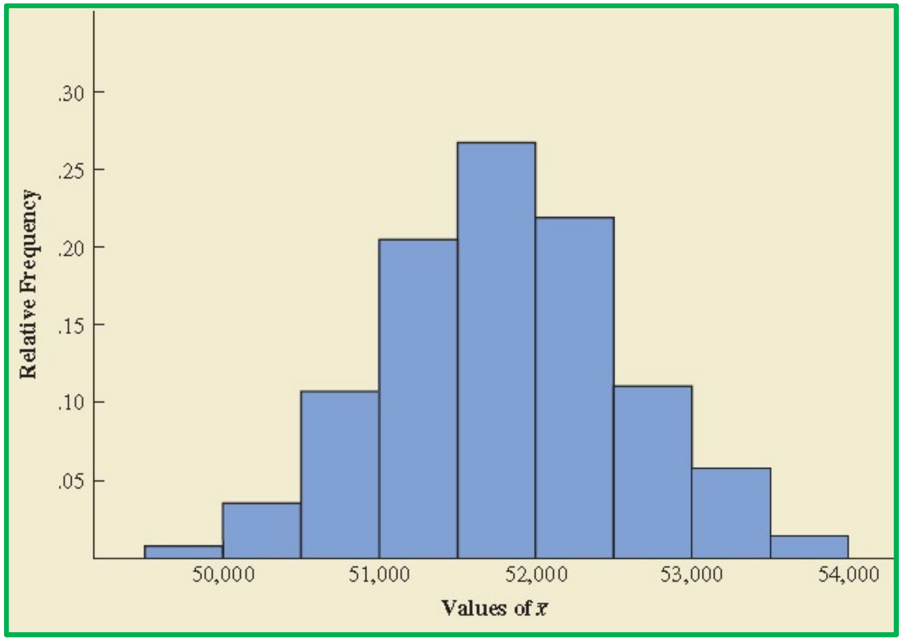
- Consider the process of selecting a simple random sample as an experiment, the sample mean $\bar{x}$ is the numerical description of the outcome of the experiment.
- Thus, the sample mean $\bar{x}$ is a random variable.
- It has a mean or expected value, a standard deviation, and a probability distribution.
- The probability distribution of is called the sampling distribution of $\bar{x}$
- It enable us to make probability statements about how close the sample mean $\bar{x}$ is to the population mean μ

Sampling Distributions of Sample mean $\bar{x}$

FREQUENCYAND RELATIVE FREQUENCY DISTRIBUTIONS OF FROM 500 SIMPLE RANDOM SAMPLES OF 30 EAI MANAGERS



Mean Annual Salary ($)	Frequency	Relative Frequency
49,500.00–49,999.99	2	.004
50,000.00–50,499.99	16	.032
50,500.00–50,999.99	52	.104
51,000.00–51,499.99	101	.202
51,500.00–51,999.99	133	.266
52,000.00–52,499.99	110	.220
52,500.00–52,999.99	54	.108
53,000.00–53,499.99	26	.052
53,500.00–53,999.99	6	.012
Totals	500	1.000



The largest concentration of the values and the mean of the 500 values is near the population mean μ \$51,800

Sampling distribution of $\bar{x}$

The sampling distribution of is the probability distribution of all possible values of the sample mean $\bar{x}$

Sampling Distributions of Sample mean $\bar{x}$

- Expected Value of $\bar{x}$

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

$E(\bar{x})$ =the expected value of $\bar{x}$

μ =the population mean

parameter$\bar{x}$ is an unbiased estimator of the population mean μ.

When the expected value of a point estimator equals the population

Sampling Distributions of Sample mean $\bar{x}$

- Standard Deviation of $\bar{x}$

Finite Population

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

Infinite Population

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

$\sigma_{\bar{x}}$ = the standard deviation of $\bar{x}$

σ = the standard deviation of the population

n = the sample size

N = the population size

- Finite correction factor is $\sqrt{(N-n)/(N-1)}$

The correction factor is required because the calculated value of SD of $\bar{x}$ is different from the value $\frac{\sigma}{\sqrt{n}}$ for the larger sample size

- If you are given N and n check if n is greater than 5% of N
- If n/N>5%, then use finite correction factor
- When n is small relative to N then fcf~1

Central Limit Theorem

- In selecting random samples of size *n* from a population, the sampling distribution of the sample mean$\bar{x}$ can be approximated by a *normal distribution* as the sample size becomes large.

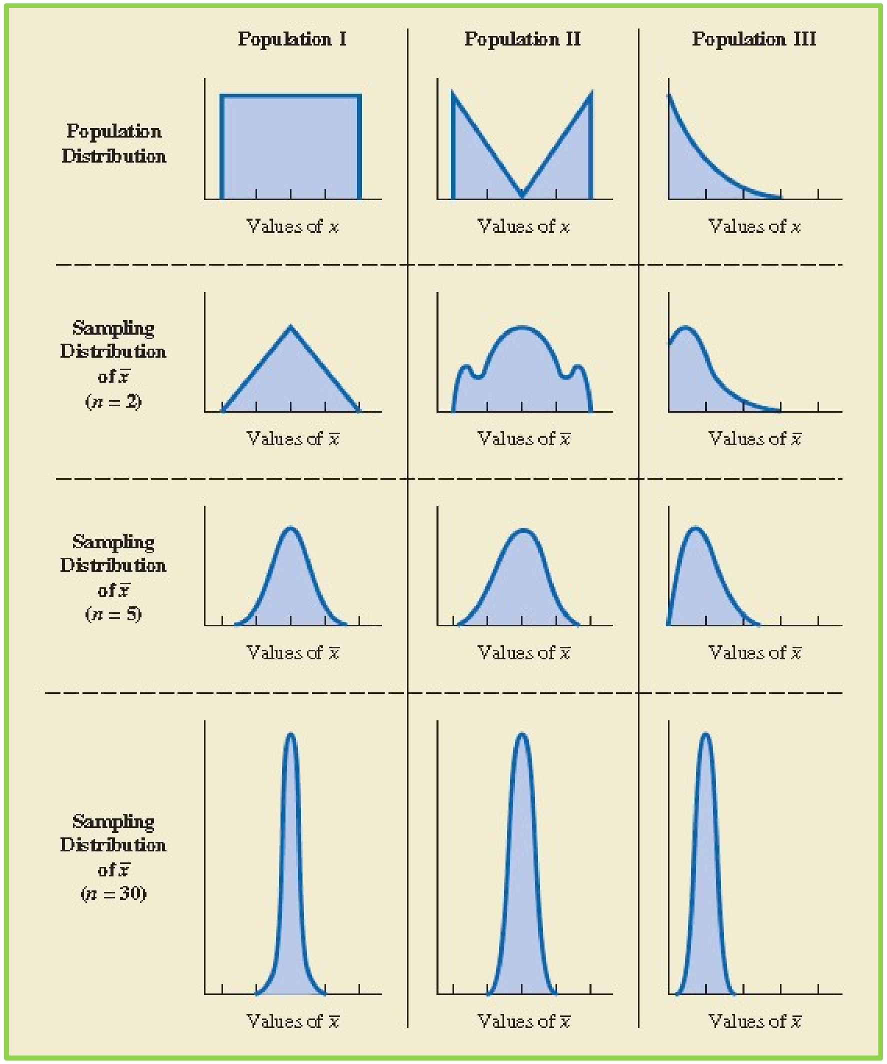
- If sample means are normally distributed, the z score formula applied to sample means would be

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

$\sigma_{\bar{x}}$ the standard error of the mean

Z Formula for sample means of a finite population

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N - n}{N - 1}}}$$



Sampling Distribution Of $\hat{p}$ ($\overline{p}$)

- The sampling distribution of is the probability distribution of all possible values of the sample proportion $\hat{p}$ $\hat{p} = \frac{x}{n}$
- *x* = number of items in a sample that have the characteristic
- *n* = number of items in the sample
- sample mean is the best choice of statistic for measurable items (weight, distance, distance, time and income etc)
- Sample proportion the best choice of statistic for countable items (how many number of ……? And questions involving Yes or No answers)
- Central limit theorem
  - Normal distribution approximates the shape of the distribution of sample proportions if *n.p*> 5 and *n.q*> 5(*p* is the population proportion and *q* = 1 - *p*)
- The standard deviation of sample proportions or standard error of the proportion $\sqrt{\frac{p \cdot q}{n}}$
- Z formula for sample proportions for n.p > 5 AND n.q > 5 $z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}}$

Sampling Distribution Of $\hat{p}$ ($\bar{p}$)

- Expected value of $\boldsymbol{\bar{p}}$

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

- Standard Deviation of $\boldsymbol{\bar{p}}$

Finite Population

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

If $n/N > .05$

Infinite Population

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

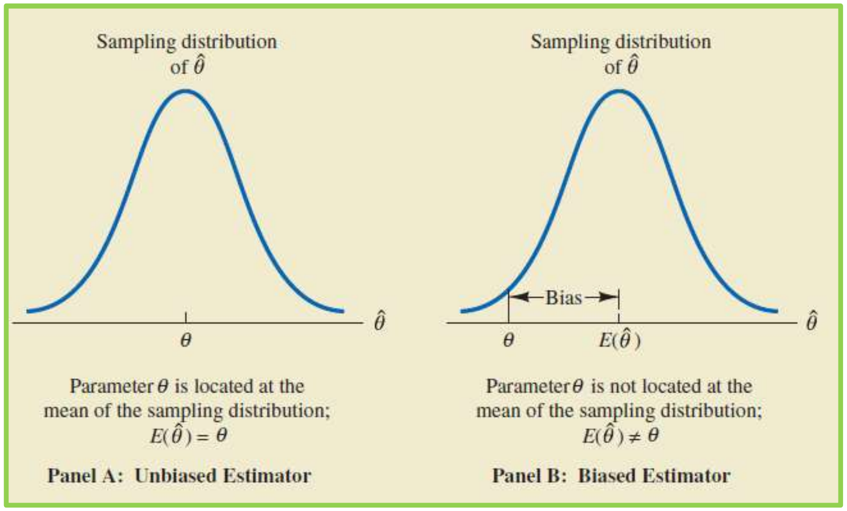
If $n/N \leq .05$

Properties of Point Estimators

- ➢sample statistics such as a sample mean $\bar{x}$, a sample standard deviation *s*, and a sample proportion$\bar{p}$ can be used as point estimators of their corresponding population parameters *μ*, *σ*, and *p.*
- ➢Three properties of good point estimators:
  - ➢Unbiased
  - ➢Efficiency
  - ➢Consistency

general notation

- θ the population parameter of interest
- $\hat{\theta}$ the sample statistic or point estimator of θ



Parameter θ is located at the mean of the sampling distribution; $E(\hat{\theta}) = \theta$

Panel A: Unbiased Estimator

Parameter θ is not located at the mean of the sampling distribution; $E(\hat{\theta}) \neq \theta$

Panel B: Biased Estimator

- Unbiased:
- If the expected value of the sample statistic is equal to the population parameter being estimated, the sample statistic is said to be an *unbiased estimator* of the population parameter

$$\boxed{E(\hat{\theta}) = \theta}$$

$$\boxed{E(\hat{\theta}) = \text{the expected value of the sample statistic } \hat{\theta}}$$

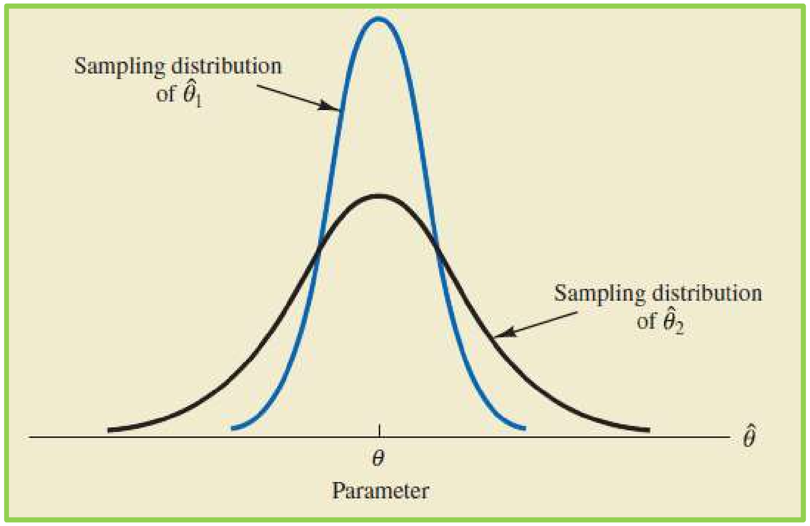
$$\boxed{\text{Bias}(\hat{\Theta}) = \text{E}(\hat{\Theta}) - \Theta}$$

Properties of Point Estimators

Efficiency

The point estimator with the smaller standard error is said to have greater **relative efficiency** than the other

SAMPLING DISTRIBUTIONS OF TWO UNBIASED POINT ESTIMATORS



$\widehat{\theta_1}$have a greater chance of being close to the parameter θ than do values of $\widehat{\theta_2}$.
It has less standard error of point estimator

$$\boxed{\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2)}$$

$$\boxed{\text{Relative Efficiency} = \frac{\mathrm{Var}(\hat{\theta}_2)}{\mathrm{Var}(\hat{\theta}_1)}}$$

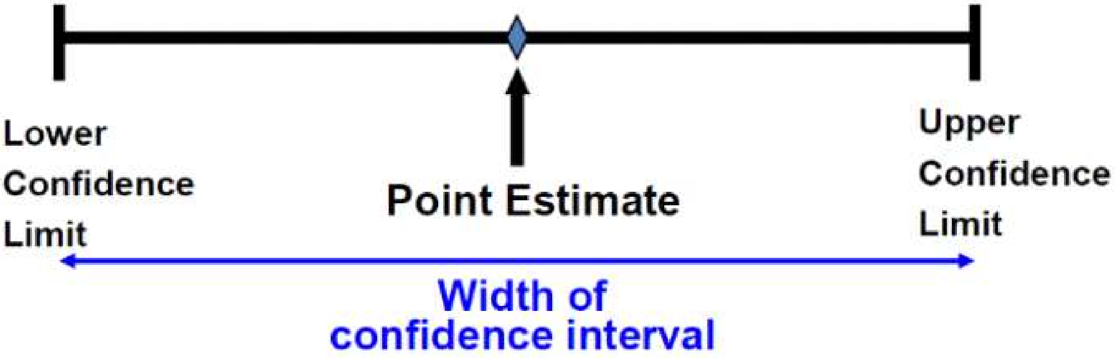
Consistency

a point estimator is consistent if the values of the point estimator tend to become closer to the population parameter as the sample size becomes larger.

Interval Estimation

- An **estimator** of a population parameter is a random variable that depends on the sample information;
- its value provides approximations of this unknown parameter.
- A specific value of that random variable is called an **estimate**.

- A point estimate is a single number, where as a Interval Estimate (confidence interval) provides additional information about variability.



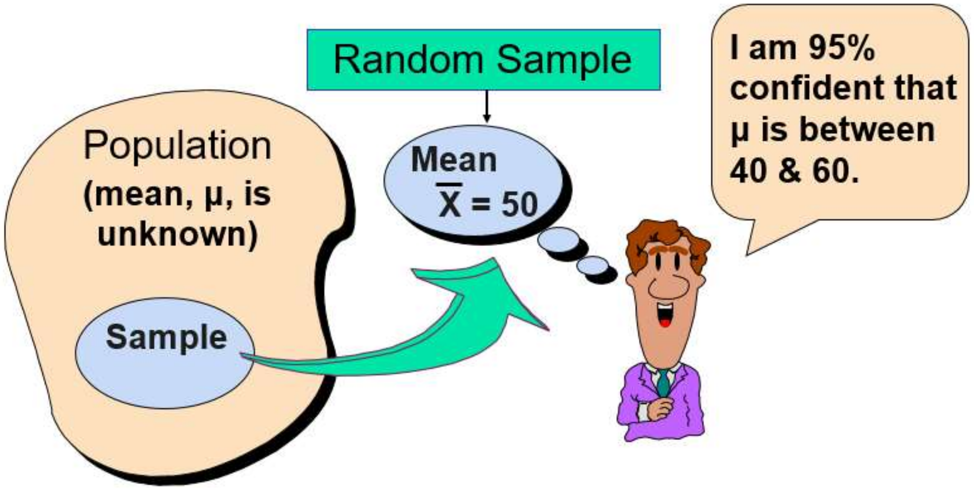
Confidence Interval Estimate

- How much uncertainty is associated with the point estimate of the population parameter? (temp Example)
  - An interval estimate (confidence interval) is a range of values within which the analyst can declare, with some confidence, the population parameter lies.
  - An interval estimate provides more information about a population characteristic than a point estimate does
  - Such interval estimates are called confidence intervals
  - It takes in to consideration variation in sample statistic from sample to sample
  - It gives the information about the closeness to unknown population parameters
  - Stated in terms of level of confidence
    - Can never be 100%confident

Confidence Interval and Confidence Level

- If $P(a < \theta < b) = 1 - \alpha$ then the interval from a to b is called a $100(1 — \alpha)\%$ confidence interval of θ.
- The quantity $(1- \alpha)$ is called the **confidence level** of the interval α between 0 and 1).
- In repeated samples of the population, the true value of the parameter θ would be contained in $100(1 — \alpha)\%$ of intervals calculated this way.
- The confidence interval calculated in this manner is written as $a < \theta < b$ with $100(1 — \alpha)\%$ confidence.

- **Estimation Process:**



Confidence Level, $(1-\alpha)$

- Suppose confidence level = 95%
- Also written $(1 - \alpha) = 0.95$
- A relative frequency interpretation:
  - From repeated samples, 95% of all the confidence intervals that can be constructed will contain the unknown true parameter
- A specific interval either will contain or will not contain the true parameter
  - No probability involved in a specific interval

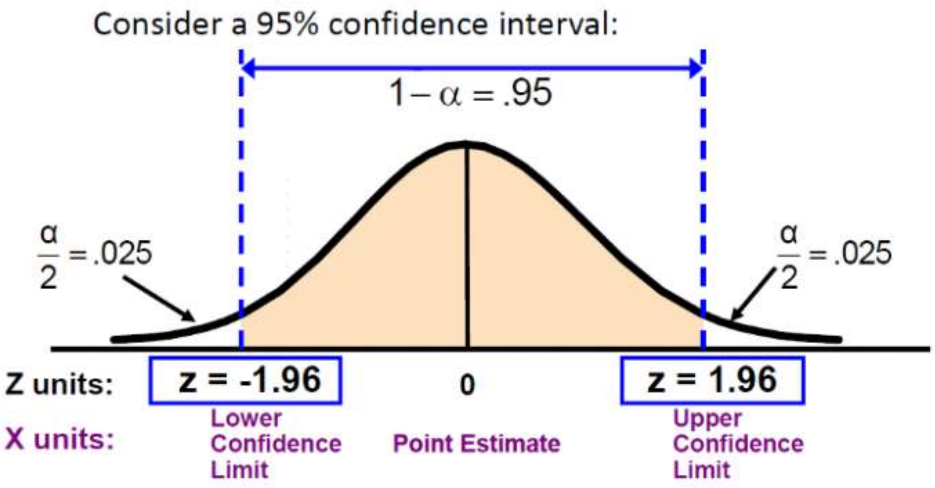
- The general formula for all confidence intervals is:

Point Estimate ±(Reliability Factor)(Standard Error)

- The value of the reliability factor depends on the desired level of confidence.

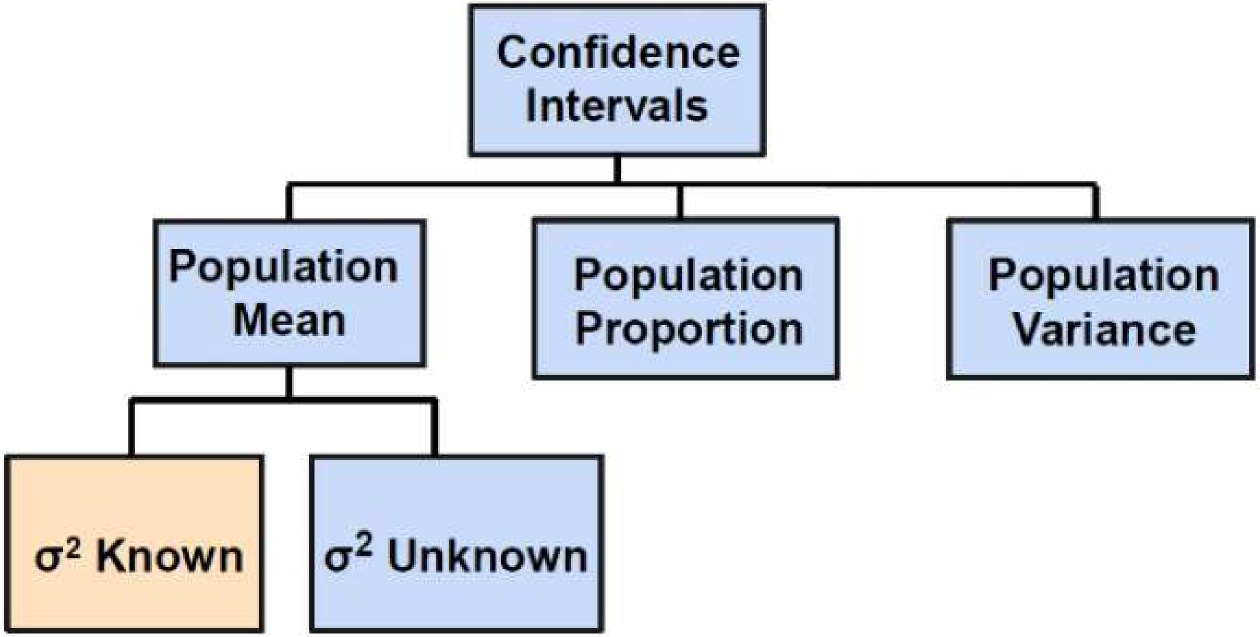
$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

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



Finding the Reliability Factor, z(α/2)

Confidence Intervals



Confidence Interval for μ (σ^2 Known)

- Assumptions
  - Population variance σ^2 is known
  - Population is normally distributed
  - If population is not normal, use large sample

- Confidence interval estimate:

$$\boxed{\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

$$\boxed{\bar{x} \pm ME}$$

(where $z_{\alpha/2}$ is the normal distribution value for a probability of $\alpha/2$ in each tail)

- Margin of Error

$$\boxed{ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

The **width**, *w*, is equal to twice the margin of error: $\boxed{w = 2(ME)}$

The **upper confidence limit (UCL)** is given by $\boxed{\text{UCL} = \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$

The **lower confidence limit (LCL**) is given by $\boxed{\text{LCL} = \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$

Reducing the Margin of Error

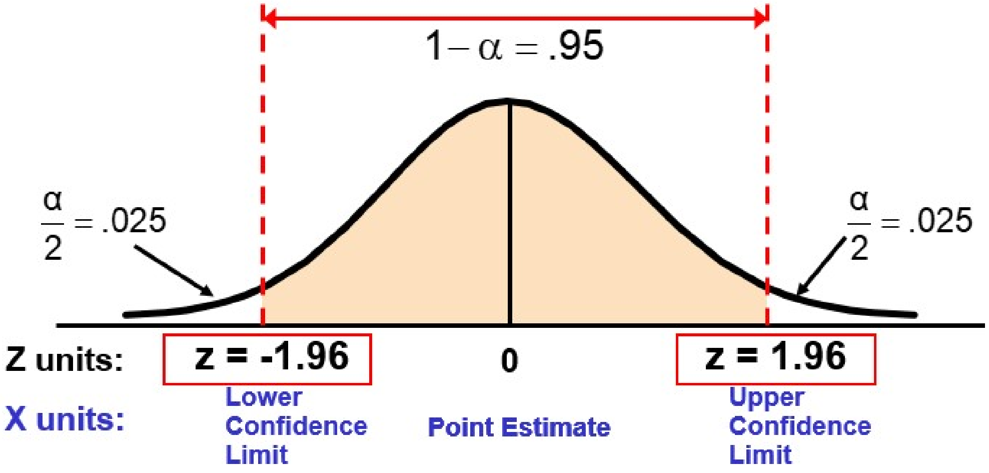
$$ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

The margin of error can be reduced if

- the population standard deviation can be reduced (σ↓)
- The sample size is increased (n↑)
- The confidence level is decreased, $(1 - \alpha)$ ↓

Finding the Reliability Factor, zα/2

Consider a 95% confidence interval:

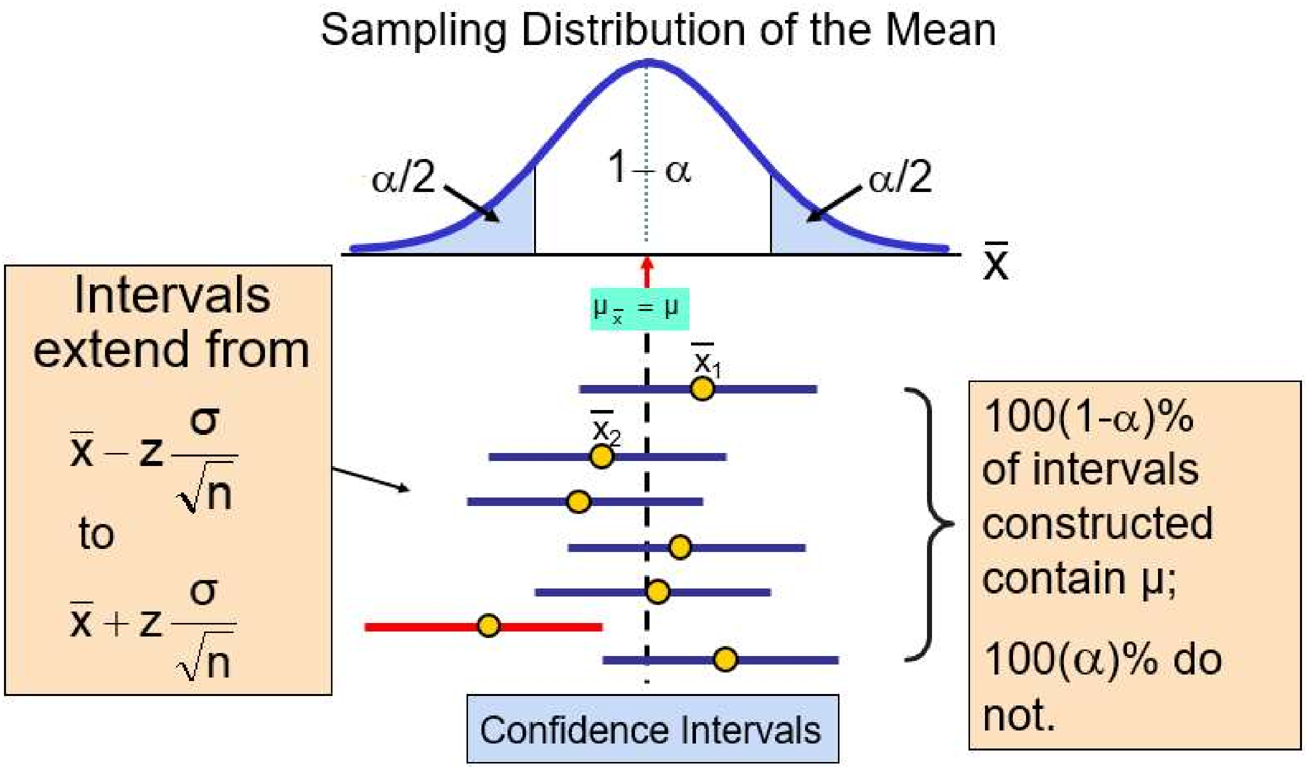


Find $z_{.025}=\pm1.96$ from the standard normal distribution table

Commonly used confidence levels are 90%, 95%, and 99%

Confidence Level	*Confidence Coefficient,* $1-\alpha$	$Z_{\alpha/2}$ *value*
80%	.80	1.28
90%	.90	1.645
95%	.95	1.96
98%	.98	2.33
99%	.99	2.58
99.8%	.998	3.08
99.9%	.999	3.27

Intervals and Level of Confidence



Sampling Distribution of the Mean

Intervals extend from

$$\bar{x} - z\frac{\sigma}{\sqrt{n}}$$

to

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

100(1-α)% of intervals constructed contain μ;

100(α)% do not.

Confidence Intervals

Example

A survey was taken of U.S. companies that do business with firms in India. One of the questions on the survey was: Approximately how many years has your company been trading with firms in India? A random sample of 44 responses to this question yielded a mean of 10.455 years. Suppose the population standard deviation for this question is 7.7 years. Using this information, construct a 90% confidence interval for the mean number of years that a company has been trading in India for the population of U.S. companies trading with firms in India

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

$$10.455 - 1.645\frac{7.7}{\sqrt{44}} \le \mu \le 10.455 + 1.645\frac{7.7}{\sqrt{44}}$$

$$10.455 - 1.910 \le \mu \le 10.455 + 1.910$$

$$8.545 \le \mu \le 12.365$$

The analyst is 90% confident that if a census of all U.S. companies trading with firms in India were taken at the time of this survey, the actual population mean number of years a company would have been trading with firms in India would be between 8.545 and 12.365. The point estimate is 10.455 years.

Confidence Intervals for the Population Proportion, p

➢An interval estimate for the population proportion (P) can be calculated by adding an allowance for uncertainty to the sample proportion ($\hat{p}$)

- Recall that the distribution of the sample proportion is approximately normal if the sample size is large, with standard deviation

$$\boxed{\sigma_P = \sqrt{\frac{P(1-P)}{n}}}$$

- Let $\hat{p}$ denote the observed proportion of "successes" in a random sample of *n* observations from a population with a proportion of successes *P*. Then, if *nP* (1 – *P*) >: 5, a 100(1 – a)% **confidence interval for the population proportion** is given by

$$\boxed{\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < P < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

- where
  - $z_{\alpha/2}$ is the standard normal value for the level of confidence desired
  - $\hat{p}$ is the sample proportion
  - n is the sample size

Example

- Management wants an estimate of the proportion of the corporation’s employees who favor a modified bonus plan. From a random sample of 344 employees, it was found that 261 were in favor of this particular plan. Find a 90% confidence interval estimate of the true population proportion that favors this modified bonus plan.

Solution:

- The sample proportion, $\widehat{p}$, and the reliability factor for a 90% confidence interval estimate (α = 0.102 of the true population proportion, *P*, are found to be

$$\hat{p} = 261/344 = 0.759$$
$$z_{\alpha/2} = z_{0.05} = 1.645$$

- 90% confidence interval for the population proportion is

$$0.759 \pm 1.645\sqrt{\frac{(0.759)(0.241)}{344}}$$
$$0.759 \pm 0.038$$

- The interval is 0.721, 0.797 (72.1% to 79.7%) with

Hypothesis Testing

- ➢Hypothesis testing can be used to determine whether a statement about the value of a population parameter should or should not be rejected
- A hypothesis is a claim (assumption) about a population parameter:
- ➢The null hypothesis, denoted by Ho, is a tentative assumption about a population parameter
- ➢The alternative hypothesis, denoted by Ha, is the opposite of what is stated in the null hypothesis
- ➢The hypothesis testing procedure uses data from a sample to test the two competing statements Indicated by Ho and Ha
- ➢**population mean**

Example: The mean monthly cell phone bill of this city is μ = $42

- ➢**population proportion**

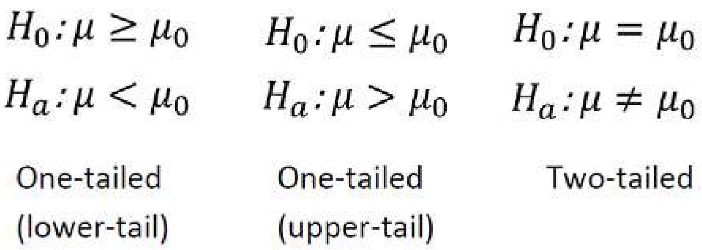
Example: The proportion of adults in this city with cell phones is p = .68

Developing Null and alternate hypothesis

- ➢It is not always obvious how the null and alternative hypothesis should be formulated
- ➢Care must be taken to structure the hypothesis appropriately so that the test conclusion provides the information the researches want
- ➢The context of the situation is very important in determining how the hypothesis should be started
- ➢In some cases it is easier to identify the alternate hypothesis first, in other cases null is easier
- ➢Correct hypothesis formulation will take practice

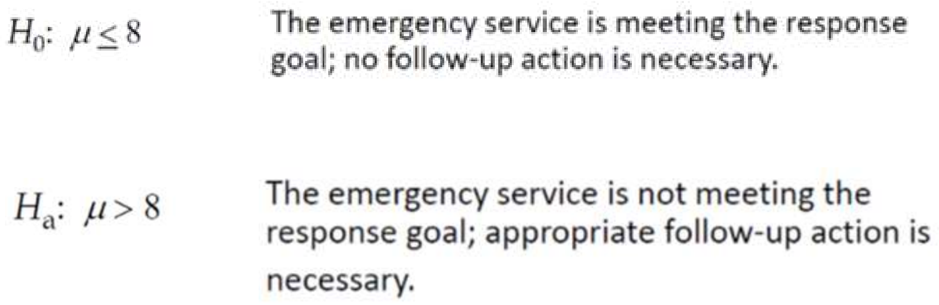
Null and Alternative Hypotheses about a Population Mean µ

- The equality part of the hypotheses always appears in the null hypothesis.
- In general, a hypothesis test about the value of a population mean µ must take one of the following three forms (where µo is the hypothesized value of the population mean):



$H_0: \mu \geq \mu_0$	$H_0: \mu \leq \mu_0$	$H_0: \mu = \mu_0$
$H_a: \mu < \mu_0$	$H_a: \mu > \mu_0$	$H_a: \mu \neq \mu_0$
One-tailed (lower-tail)	One-tailed (upper-tail)	Two-tailed

Null Hypothesis	Alternate Hypothesis
Refers to the status quo	Challenges the status quo
Always contains “=” , “≤” or “≥” sign	Never contains the “=” , “≤” or “≥” sign
May or may not be rejected	May or may not be supported

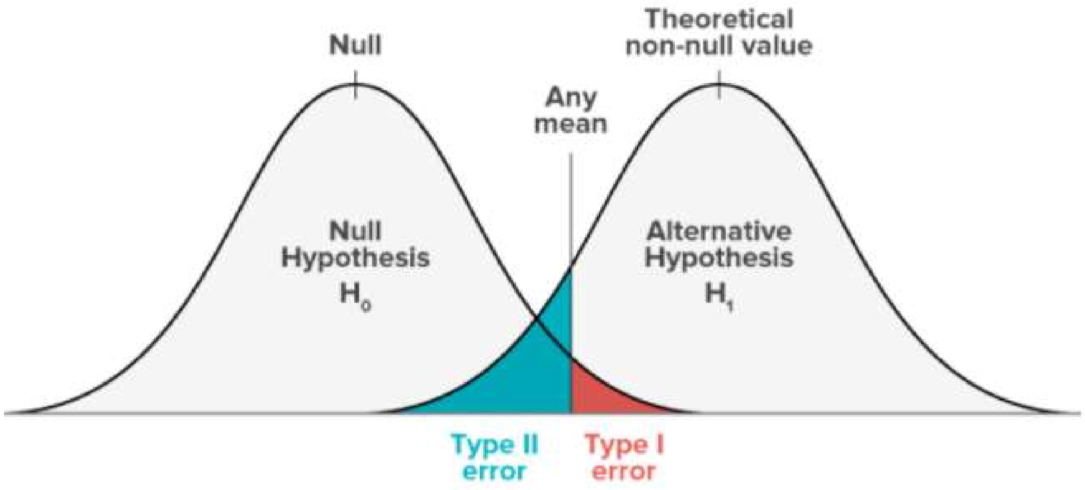


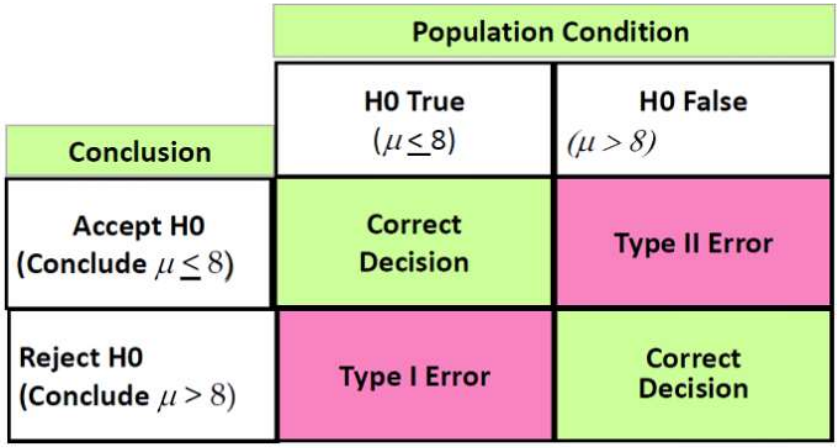
H_0: $\mu \leq 8$ The emergency service is meeting the response goal; no follow-up action is necessary.

H_a: $\mu > 8$ The emergency service is not meeting the response goal; appropriate follow-up action is necessary.

Errors in Hypothesis Testing

- Type I Error (α)
- hypothesis tests are based on sample data, we must allow for the possibility of errors.
- A Type I error is rejecting Ho when it is true.
- Type II Error (β)
- A Type II error is accepting Ho when it is false. It is difficult to control for the probability of making a Type II error.





	Population Condition	
Conclusion	H0 True ($\mu \leq 8$)	H0 False ($\mu > 8$)
Accept H0 (Conclude $\mu \leq 8$)	Correct Decision	Type II Error
Reject H0 (Conclude $\mu > 8$)	Type I Error	Correct Decision

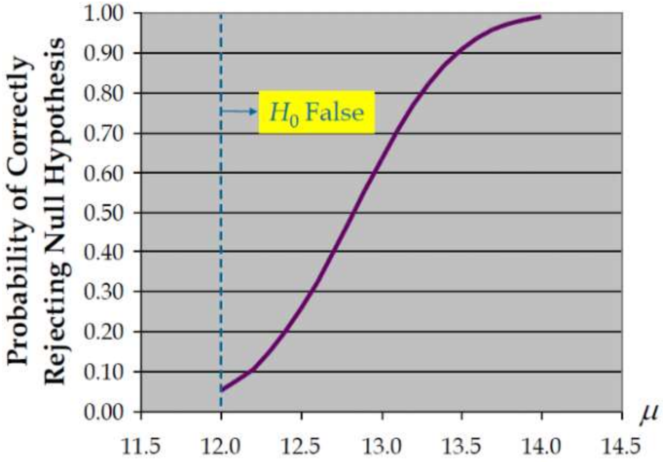
α is called the significance level

Then the probability of rejecting a false null hypothesis is (1 – β), which is called the power of the test.

Errors in Hypothesis Testing

- Factors Affecting Type II Error
- **True value of population parameter**: β-value increases when the difference between hypothesized parameter and its true value decrease.
- **Significance level α**: It increases when β decreases.
- **Population standard deviation σ**: It increases when β increases.
- **Sample size**: β-value increases when n decreases.

- **Power of the Test**
- The probability of correctly rejecting Ho when it is false is called the power of the test.
- For any particular value of μ, the power is 1 — β.
- We can show graphically the power associated with each value of μ.
- Such a graph is called a power curve.

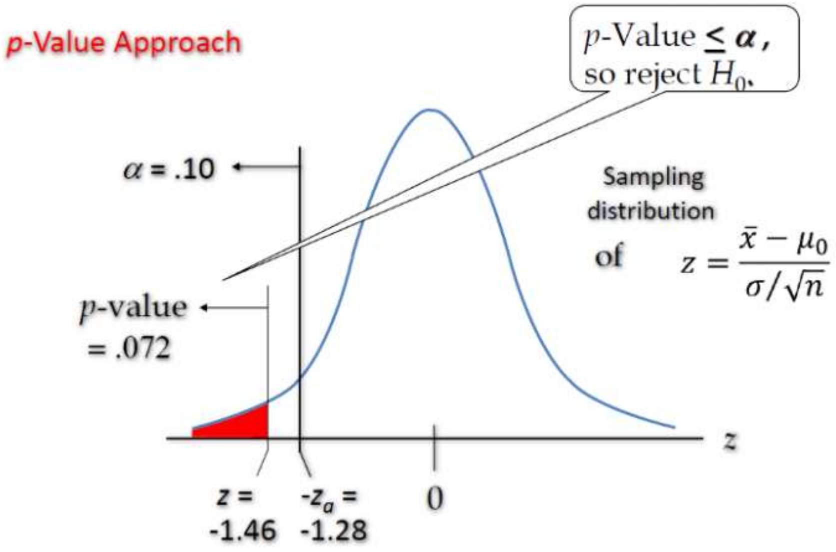


Level of Significance α

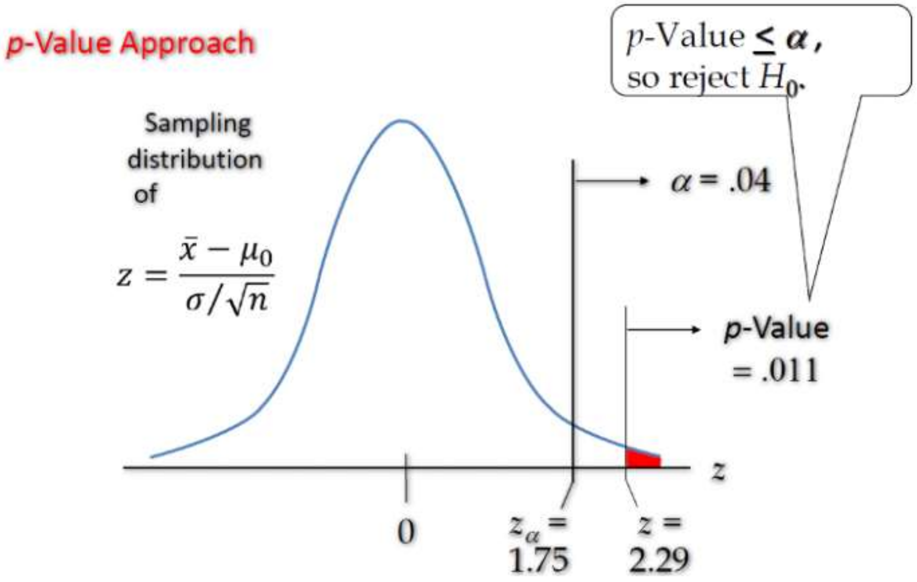
- **Defines the unlikely values of the sample statistic if the null hypothesis is true**
  - Defines rejection region of the sampling distribution
- Is designated by α , (level of significance)
  - Typical values are .01, .05, or .10
- Is selected by the researcher at the beginning
- Provides the critical value(s) of the test

Approaches for Hypothesis Testing: P Value

- ➢Step 1. Develop the null and alternative hypotheses.
- ➢Step 2. Specify the level of significance α.
- ➢Step 3. Collect the sample data and compute the test statistic.
- ➢Step 4. Use the value of the test statistic to compute the p-value.
- ➢Step 5. **Reject Ho if p-value < α.**



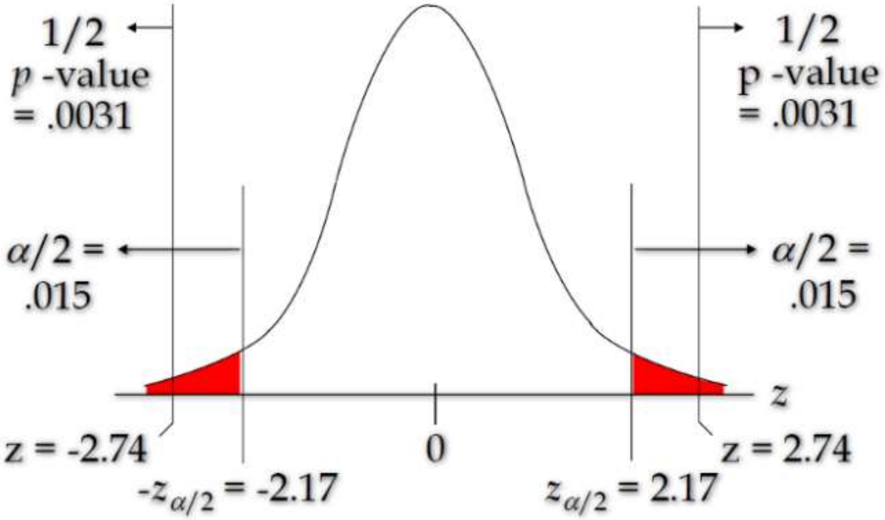
Lower-Tailed Test About a Population Mean: When σ is Known



Upper-Tailed Test About a Population Mean: When σ is Known

Approaches for Hypothesis Testing: P Value

- **p-Value Approach to Two-Tailed Hypothesis Testing**
- Compute the p-value using the following three steps:
- Compute the value of the test statistic z
- If z is in the upper tail (z > 0), find the area under the standard normal curve to the right of z
- If z is in the lower tail (z < 0), find the area under the standard normal curve to the left of z
- Reject Ho if the p-value <α



Two-Tailed Tests About a Population Mean: When σ is Known

Approaches for Hypothesis Testing: Critical Value

$$\text{reject } H_0 \text{ if } \bar{x} > \bar{x}_c = \mu_0 + z_\alpha \sigma/\sqrt{n}$$

- The value *xc* is often called the **critical value** for the decision. Note that for every value z_α obtained from the standard normal distribution, there is also a value x_c,

Reference

- Statistics for Business and Economics, Pearson edition (2013), William L Carlson, Paul Newbold, Betty M.Thorne
- statistics-for-business-and-economics-8th-edition, Paul newbold, William L.Carlson,Betty M.Thome
- Business statistics for contemporary Decision making , 6th edition by Ken Black
- https://medium.com/budding-data-scientist
- statistics-for-business-and-economics-11th edition, David R. Anderson,Dennis J. Sweeney,Thomas A. Williams

thank you
danke
謝謝
ngiyabonga
teşekkür ederim
спасибо
gracias
dank je
tapadh leat
bedankt
dziękuję
hvala
mauruuru
mochchakkeram
go raibh maith agat
obrigado
sukriya
kop khun krap
sagolun
terima kasih
ありがとう
arigato
takk
dakujem
grazie
мерси
merci
감사합니다
ευχαριστώ
dhanyavadagalu
nandri
நன்றி