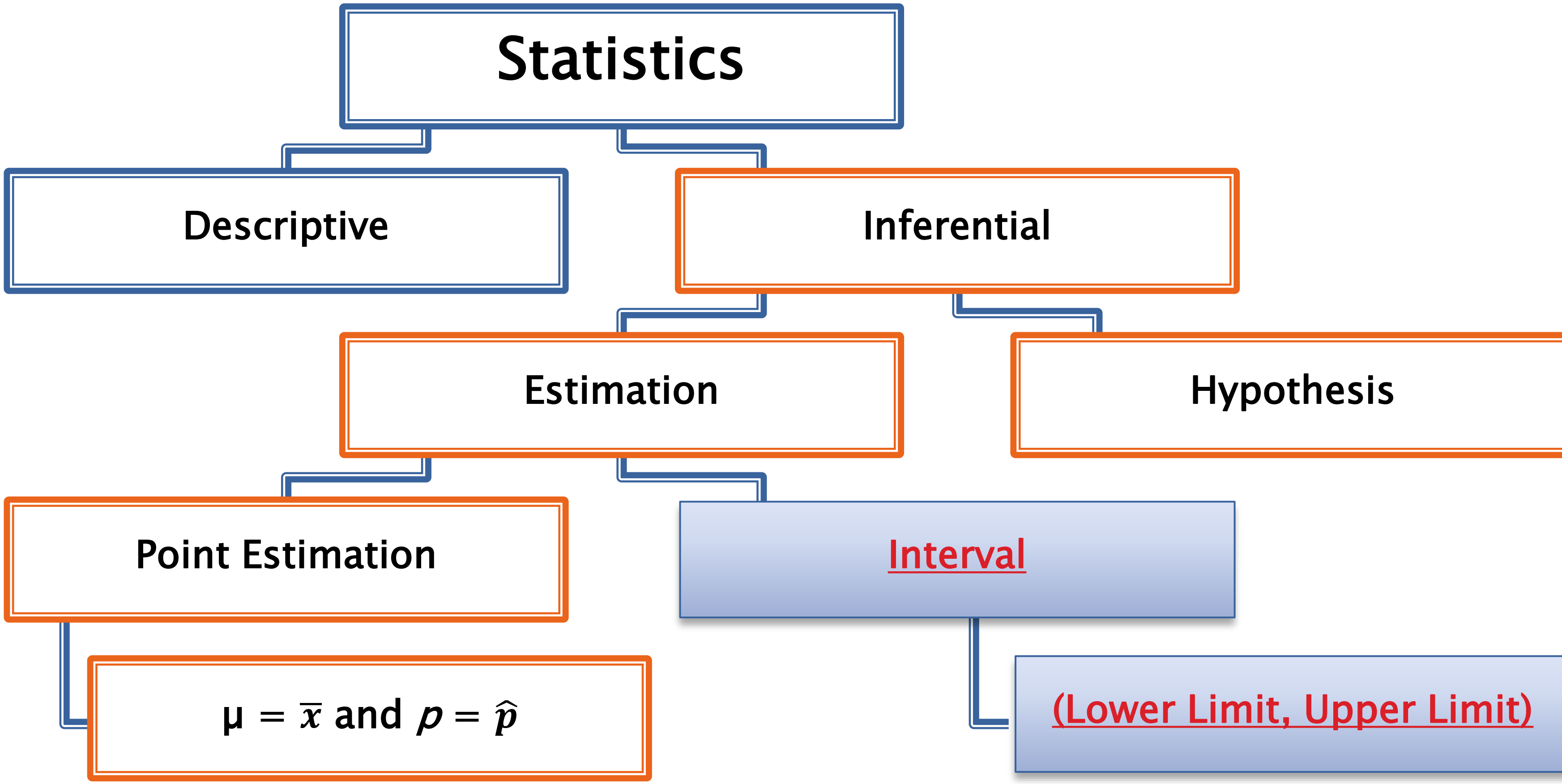
STATISTICAL ANALYSIS - LECTURE 09

Dr. Mahmoud Mounir

mahmoud.mounir@cis.asu.edu.eg

STATISTICAL INFERENCE

Recap



Statistics
Descriptive
Inferential
Estimation
Hypothesis
Point Estimation
Interval
$\mu = \bar{x}$ and $p = \hat{p}$
(Lower Limit, Upper Limit)

Statistical Inference

- The statistical inference is the process of making judgment about a population based on the properties of a random sample from the population.

Estimators

Estimator (Sample Statistic)		Population Parameter
$\bar{x}$	ESTIMATES	μ
S^2	ESTIMATES	σ^2
s	ESTIMATES	σ
$\hat{p}$	ESTIMATES	p
$\bar{x}_1 - \bar{x}_2$	ESTIMATES	$\mu_1 - \mu_2$
$\hat{p}_1 - \hat{p}_2$	ESTIMATES	$p_1 - p_2$

Central Limit Theorem

Sampling Distribution of Sample Mean = Distribution of $\bar{x}$

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

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

$$\bar{X} \text{ follows Approximately } N(\mu, \frac{\sigma}{\sqrt{n}})$$

$$\text{Standardize: Z-score of } \bar{X} = \frac{\bar{X} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

Sampling Distribution of sample Proprtion = Distribution of $\hat{p}$

$$\mu_{\hat{p}} = p$$

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

$$\hat{p} \text{ follows Approximately } N(p, \sqrt{\frac{p(1-p)}{n}})$$

$$\text{Standardize: Z-score of } \hat{p} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}};$$

Central Limit Theorem

Difference Between Two Sample Means

$$\mu_{\overline{X}_1-\overline{X}_2} = \mu_1 - \mu_2$$

$$\sigma_{\overline{X}_1-\overline{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\overline{X}_1 - \overline{X}_2 \sim \mathrm{N}(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

$$\Leftrightarrow$$

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathrm{N}(0,1)$$

Difference Between Two Sample Proportions

$$\mu_{(\hat{p}_1-\hat{p}_2)} = p_1 - p_2$$

$$\sigma_{(\hat{p}_1-\hat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$z_{\hat{p}_1-\hat{p}_2} = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

Estimation

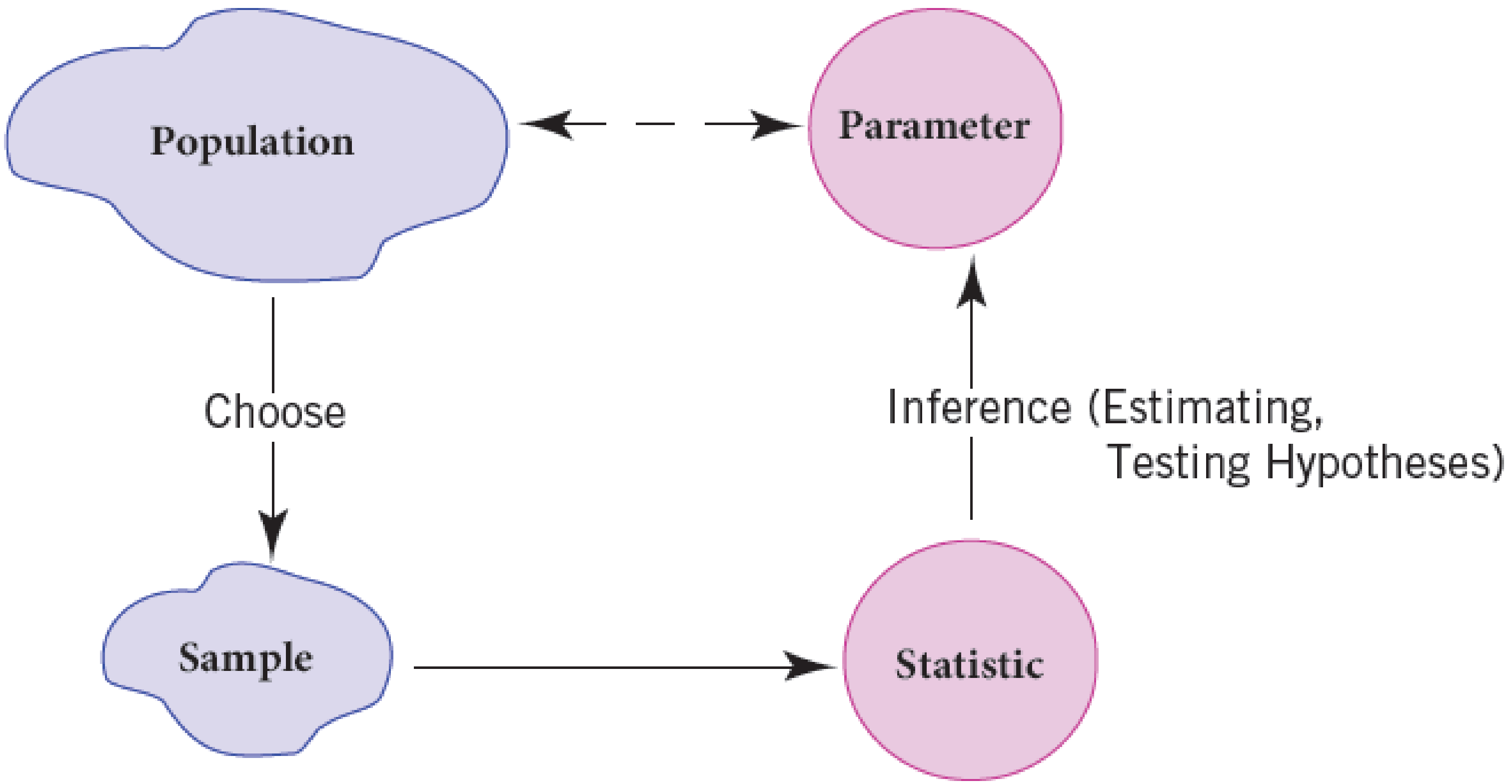
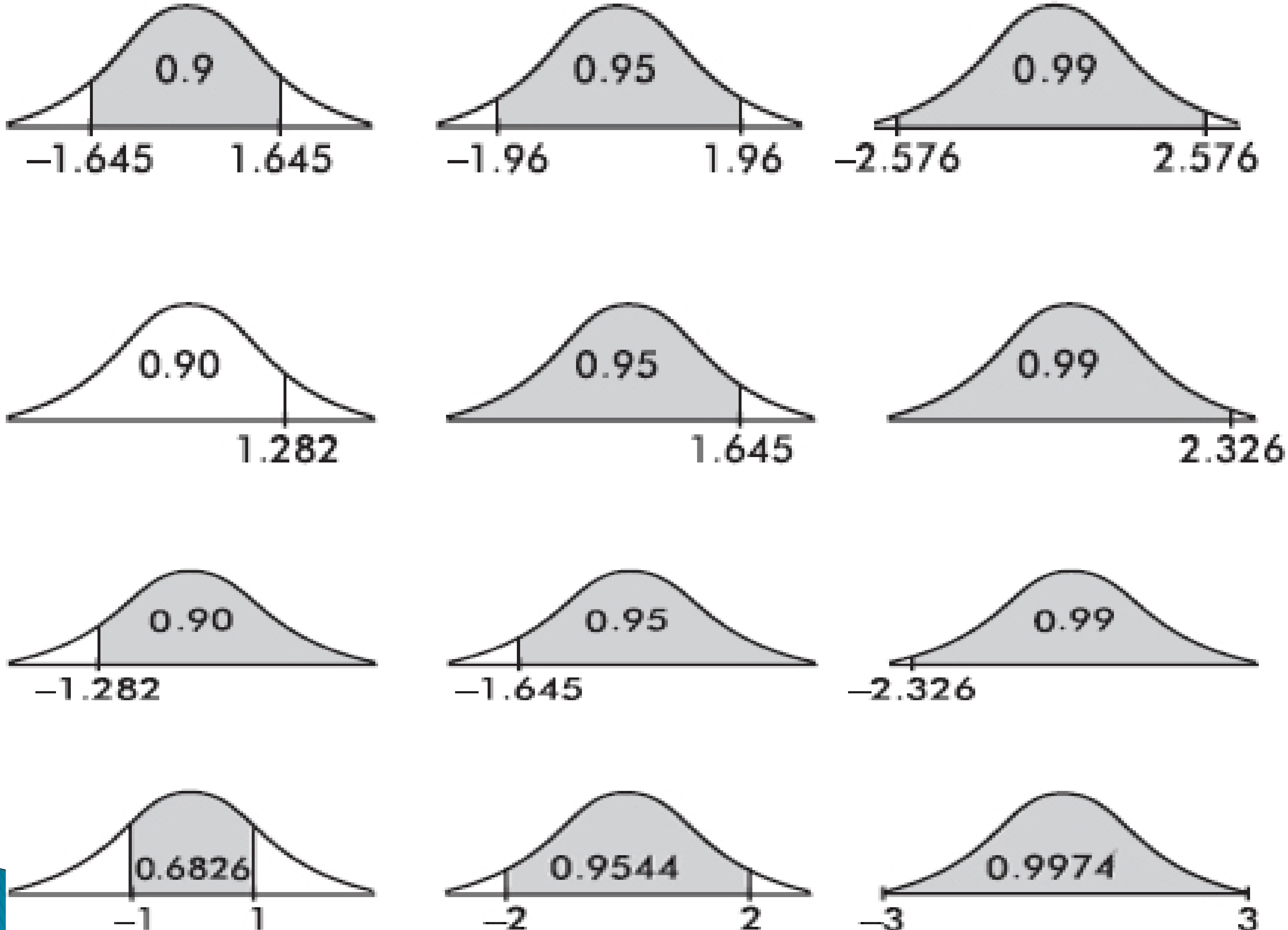


Figure: The steps required to draw the statistical inference (estimation) of an unknown parameter.

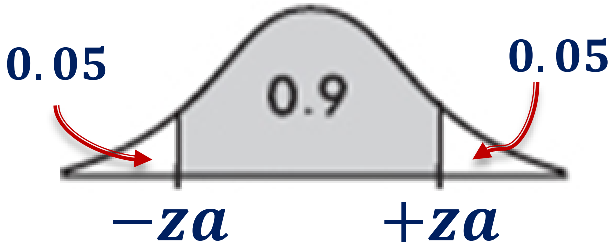

0.9
−1.645
1.645
0.95
−1.96
1.96
0.99
−2.576
2.576
0.90
1.282
0.95
1.645
0.99
2.326
0.90
−1.282
0.95
−1.645
0.99
−2.326
0.6826
−1
1
0.9544
−2
2
0.9974
−3
3

Recap

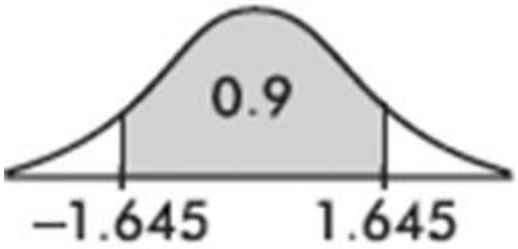
$P\ (Z < -Z_a) = P\ (Z > +Z_a)$

$P\ (Z < -Z_a) = 0.05$

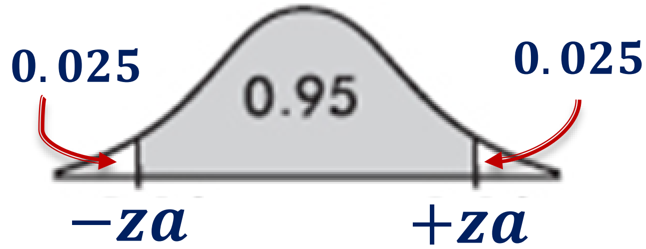
From the Z- Table, we can find the value of $-Z_a$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455



Recap



$P(Z < -Z_a) = P(Z > +Z_a)$

$P(Z < -Z_a) = 0.025$

From the Z- Table, we can find the value of $–Z_a$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233

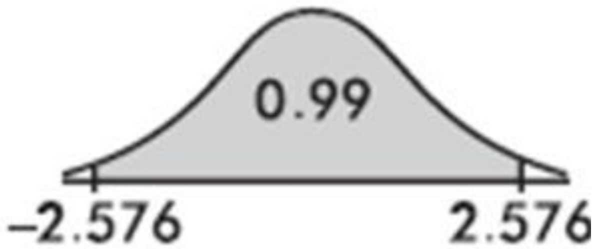
Recap

$P\,(Z < -Z_a) = P\,(Z > +Z_a)$

$P\,(Z < -Z_a) = 0.005$

From the Z- Table, we can find the value of $-Z_a$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048



Standard Z-Score (Z_α)

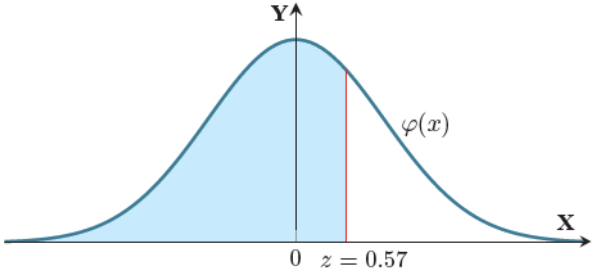
$$Z = \frac{X - \mu}{\sigma}$$

For $z_\alpha = 0.57$ we find from the table of standard normal distribution

$$P(Z < z_\alpha) = \Phi(z_\alpha) = \Phi(0.57)$$

z	0.00	0.01	...	0.07	0.08	0.09
0.0	0.5000	0.5040		0.5279	0.5391	0.5359
⋮						
0.5	0.6915	0.6950		0.7157	0.7190	0.7224
0.6	0.7257	0.7291		0.7486	0.7517	0.7549
0.7	0.7580	0.7611		0.7794	0.7823	0.7852

So we find $P(Z < z_\alpha) = \Phi(0.57) = 0.7157$



Standard Z-Score (Z_α)

$$Z = \frac{X - \mu}{\sigma}$$

$$Z\,\sigma = X - \mu$$

$$\mu = X - Z\,\sigma$$

Standard Z-Score (Z_α)

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

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

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

Confidence Interval (CI)

- Is a range of values that likely would contain an unknown population parameter.

- This means that if we used the same sampling method to select different samples and computed an interval estimate for each sample, we would expect the true population parameter to fall within the interval estimates 95% of the time.

Confidence Interval (CI)

$$\mu = \bar{x} \pm Z\,\sigma/\sqrt{n}$$

Population Parameter = Estimator ± Margin of Error (d)

Where:

Margin of Error (d) = Critical Value of Z * Standard Error

$$\mu \in \left[\bar{x} - Z\,\sigma/\sqrt{n}\ ,\ \bar{x} + Z\,\sigma/\sqrt{n}\right]$$

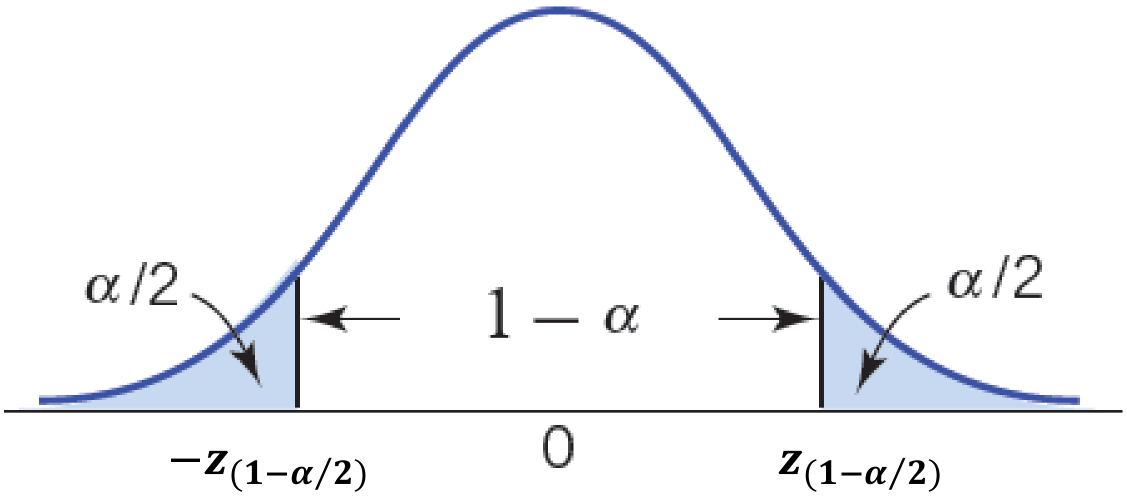
Population Parameter ∈

[Estimator − Margin of Error (d) , Estimator + Margin of Error (d)]

Standard Error for Different Sample Statistics

Estimator (Sample Statistic)		Standard Error
$\bar{x}$		$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
$\hat{p}$		$\sigma_{\hat{p}} = \sqrt{\frac{p*(1-p)}{n}}$
$\bar{x}_1 - \bar{x}_2$		$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
$\hat{p}_1 - \hat{p}_2$		$\sigma_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Confidence Interval (CI)

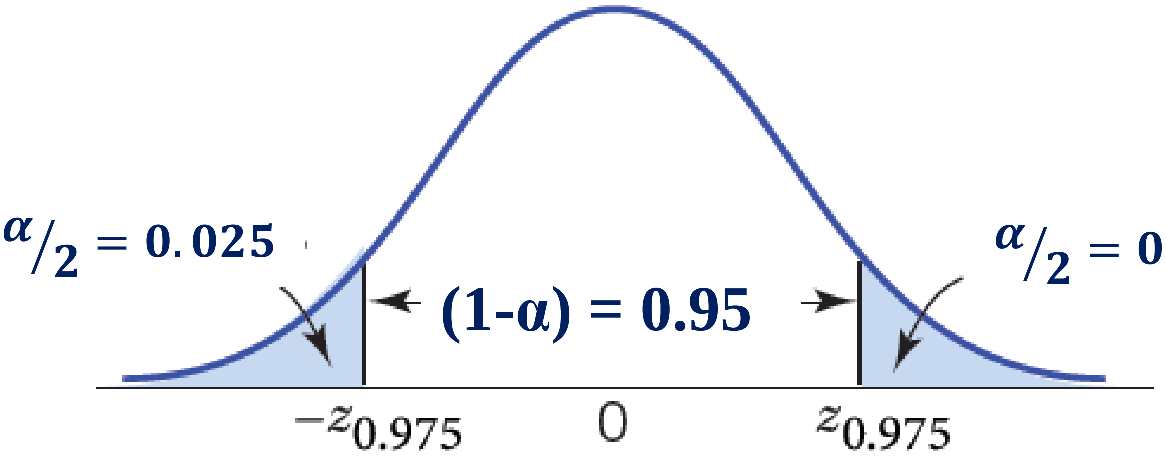


(1-α) **Confidence Level**

α **Confidence Coefficient**

Confidence Interval (CI)

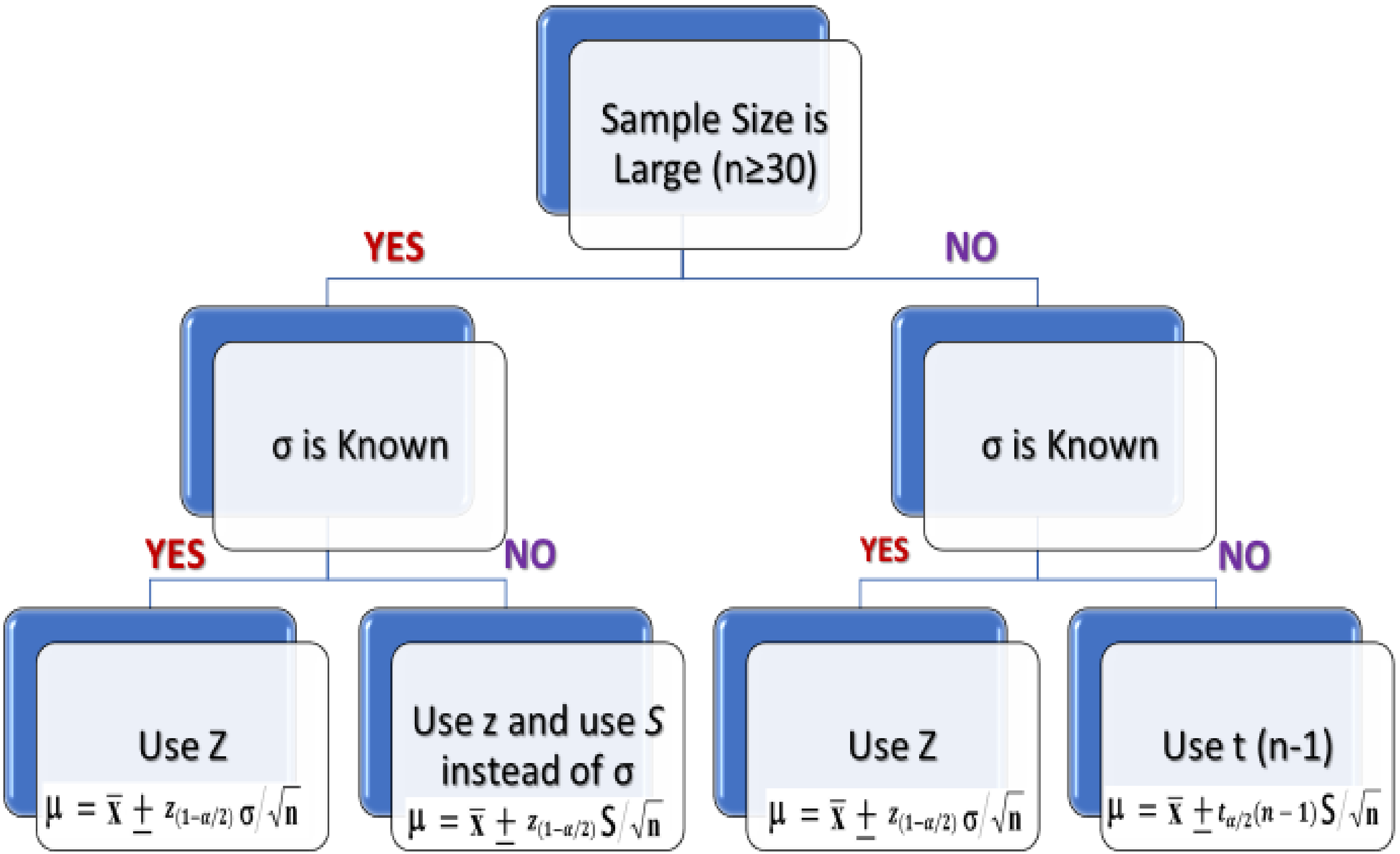
If (1-α) = 95% or 0.95



Critical Value for Z:

1- α	$100(1-\alpha)\%$	α	1- $\alpha/2$	$z_{1-(\alpha/2)}$
0.90	90%	0.10	0.95	1.65
0.95	95%	0.05	0.975	1.96
0.98	98%	0.02	0.990	2.33
0.99	99%	0.01	0.995	2.58

Confidence Interval for Population Mean



Example (1)

We wish to estimate the average number of heart beats per minute for a certain population. The average number of heart beats per minute for a sample of 49 subjects was found to be 90. Assume that these 49 patients constitute a random sample, and that the population is normally distributed with a standard deviation of 10. Construct 90, 95, 99 percent confidence intervals for the population mean.

Solution

$n = 49$	$\bar{x} = 90$	$\sigma = 10$

Confidence level	$(1-\alpha)$
Confidence coefficient	α
$(1-\alpha/2) \rightarrow \phi(z_{1-\alpha/2}) = p(z < \alpha/2)$	

90% confidence interval		
$(1-\alpha) = 0.9$	$\alpha = 0.1$	$\frac{\alpha}{2} = .05$
$1 - \frac{\alpha}{2} = .95$	$z_{1-\alpha/2} = p(z < \alpha/2)$ $= p(z < 0.05) = 1.645$	
	$\mu = \bar{x} \pm z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$	
	$\mu = 90 \pm 1.645\left(\frac{10}{\sqrt{49}}\right)$ $= 90 \pm 2.35$	
	$\mu \in [87.65, 92.35]$	

Solution

$n = 49$	$\bar{x} = 90$	$\sigma = 10$

Confidence level	$(1-\alpha)$
Confidence coefficient	α
$(1-\alpha/2) \rightarrow \phi\left(z_{1-\alpha/2}\right) = p(z < \alpha/2)$	

95% confidence interval		
$(1-\alpha) = 0.95$	$\alpha = 0.05$	$\frac{\alpha}{2} = .025$
$1 - \frac{\alpha}{2} = .975$	$z_{1-\alpha/2} = p(z < \alpha/2)$ $= p(z < 0.025) = 1.960$	
	$\mu = \bar{x} \pm z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$	
	$\mu = 90 \pm 1.960\left(\frac{10}{\sqrt{49}}\right)$ $= 90 \pm 2.8$	
	$\mu \in [87.2, 92.8]$	

Solution

$n = 49$	$\bar{x} = 90$	$\sigma = 10$

Confidence level	$(1-\alpha)$
Confidence coefficient	α
$(1-\alpha/2) \rightarrow \phi\left(z_{1-\alpha/2}\right) = p(z < \alpha/2)$	

99% confidence interval		
$(1-\alpha) = 0.99$	$\alpha = 0.01$	$\frac{\alpha}{2} = .005$
$1 - \frac{\alpha}{2} = .995$	$z_{1-\alpha/2} = p(z < \alpha/2)$ $= p(z < 0.005) = 2.576$	
	$\mu = \bar{x} \pm z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$	
	$\mu = 90 \pm 2.576\left(\frac{10}{\sqrt{49}}\right)$ $= 90 \pm 3.68$	
	$\mu \in [86.3, 93.7]$	

Example (2)

In a simple random sample of machine parts, 18 out of 225 were found to have been damaged in shipment. Establish a 95% confidence interval estimate for the proportion of machine parts that are damaged in shipment.

Suppose there are 50,000 parts in the entire shipment. How can we translate from the proportions to actual numbers?

Solution

$n = 255$	$x = 18$	$\hat{p} = \frac{18}{255} = 0.08$

$(1-\alpha) = 0.95$	$\alpha = 0.05$	$\frac{\alpha}{2} = .025$
$1 - \frac{\alpha}{2} = .975$	$z_{1-\alpha/2} = p(z < \alpha/2) = p(z < 0.025) = 1.960$	
	$p = \hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	
	$p = 0.08 \pm 1.960\left(\sqrt{\frac{0.08(0.92)}{225}}\right)$ $= 0.08 \pm 1.960(0.01699)$	
	$= 0.08 \pm 0.035$	
	$P \in [0.045, 0.115]$	

Thus, we are 95% certain that the population of machine parts damaged in shipment is between 0.045 *and* 0.115

Solution

If there are 50,000 parts in the entire shipment N=50,000 then

$$(0.045)(50000) = 2250$$

$$(0.115)(50000) = 5750$$

So, we can be 95% confident that there are between 2250 and 5750 defective parts in the whole shipment