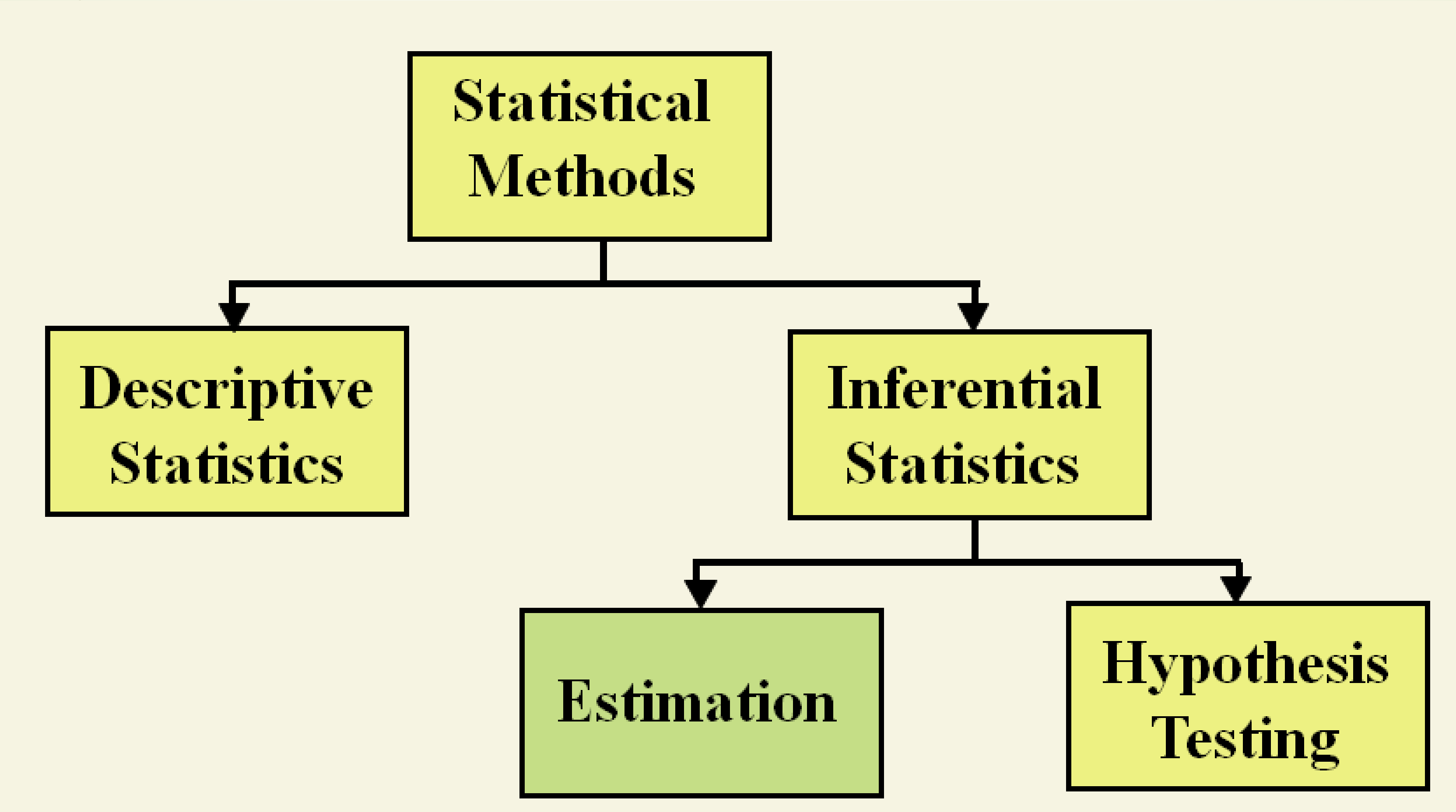
Week 7
Confidence Intervals

Statistical Methods



Statistical Inferences

- Estimation
  - Point Estimation
  - Interval Estimation
- Hypothesis Testing

Estimator versus Estimate

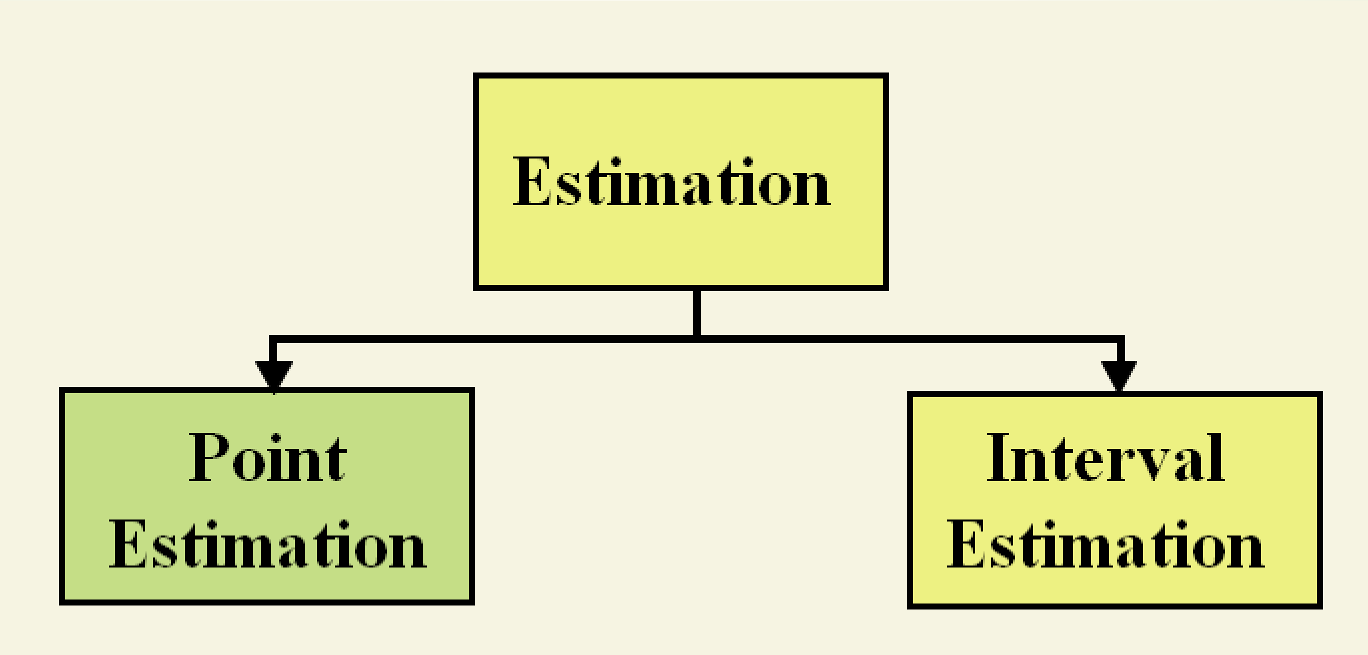
- **Estimator**
  - An estimator is any statistic used to estimate an unknown parameter value; it is a **random variable**.
- **Estimate**
  - An estimate is the numerical value of the estimator that results from a specific sample; it provides a best guess of an unknown parameter value; it is **fixed**, not random

3 Major Elements of the Estimator Required to Do Statistical Inferences

- Expected value of the estimator
- Standard error of the estimator
- Sampling distribution of the estimator

	Statistical Inferences about	
Target parameter	μ	p
Best point estimator	$\bar{X}$	$\hat{p}$
Expected value of the point estimator	$E(\bar{X}) = \mu$	$E(\hat{p}) = p$
Standard error of the estimator	$se(\bar{X}) = \frac{\sigma}{\sqrt{n}}$	$se(\hat{p}) = \sqrt{\frac{pq}{n}}$
CLT – sample size requirements	$n \geq 30$	$n > 25$, $np > 5$, and $nq > 5$
Sampling distribution	$\bar{X} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$	$\hat{p} \approx N\left(p, \sqrt{\frac{pq}{n}}\right)$

Estimation Methods



Point Estimation

- Provides a single value
  - Based on observations from one sample
- Gives no information about how close the value is to the unknown population parameter
- Example: Sample mean $\overline{X} = 3$ is **point estimate** of unknown population mean μ.

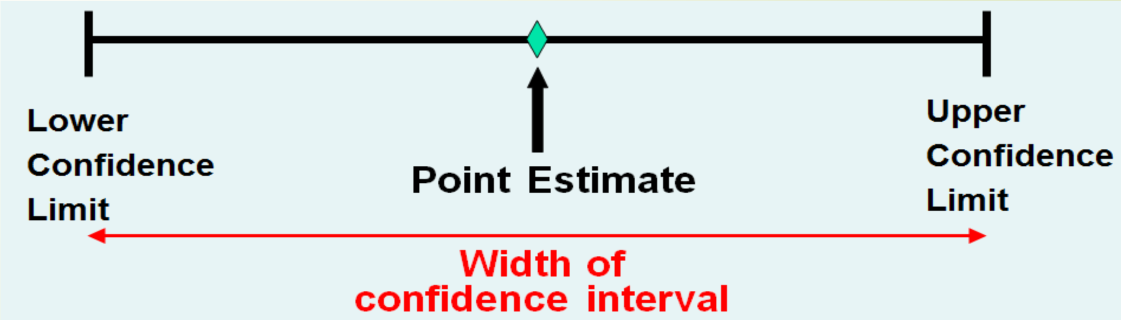
Interval Estimation

- Provides a **range of values**
  - Based on observations from one sample
- Gives information about closeness to unknown population parameter
  - Stated in terms of probability
  - Knowing exact closeness requires knowing unknown population parameter
- Example: unknown population mean lies between 50 and 70 with 95% confidence.
- The **general formula** for all interval estimator is

$$\boxed{\textbf{Point estimator} \pm \textbf{error bound}}$$

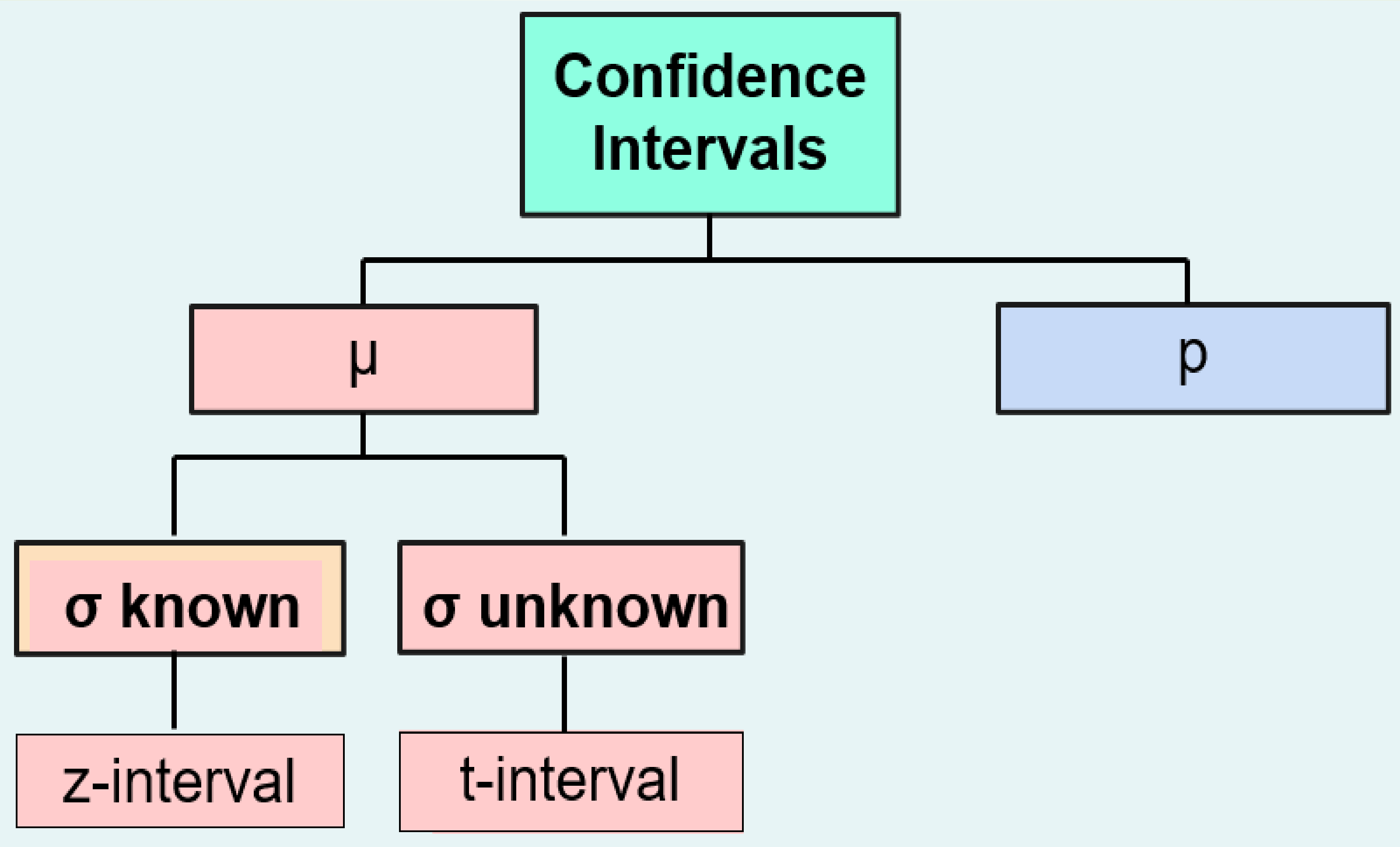
where

error bound = critical value * standard error



Terminology

- **Target parameter**
  - Is the unknown population parameter that we are interested in estimating
- **Confidence coefficient (1 – α)**
  - Is the probability that an interval estimator encloses the population parameter if the estimator is used repeatedly a very large number of times
- **Confidence level: 100(1 – α)%**
  - Is the confidence coefficient expressed as a percentage
  - Typical values are 90%, 95%, 99%
- α
  - Is the probability that target parameter is not within interval
- **Error bound / margin of error**
  - Is the sampling error that we are willing to tolerate


Confidence Intervals
μ
p
σ known
σ unknown
z-interval
t-interval

Derivation of the confidence estimator

- The sampling distribution of $\overline{X}$ for random samples of size n from a normal population with the mean μ and the variance σ^2 is a normal distribution with $\mu_{\overline{X}} = \mu$ and $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$.
  Thus, we can write
- $P(|Z| < z_{\alpha/2}) = 1 - \alpha$ where $Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ and $z_{\alpha/2}$ is such that the integral of the standard normal density from $z_{\alpha/2}$ to ∞ equals α/2. It follows that
- $P\left(\left|\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right| < z_{\alpha/2}\right) = 1 - \alpha$

$$P\left(|\overline{X} - \mu| < z_{\alpha/2}\ \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$P\left(-z_{\frac{\alpha}{2}}\ \frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < z_{\alpha/2}\ \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$P\left(\overline{X} - z_{\frac{\alpha}{2}}\ \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\frac{\alpha}{2}}\ \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

z-interval

Assumptions:

- **σ is known** and population is normal (or $n \geq 30$) and

- Interval estimator of μ:

$$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- Interval estimate of μ:
  - Numerical values of the interval estimator of μ

Factors affecting interval width - precision

- Define

  L = lower bound and U = upper bound of the confidence interval.

  E = error bound / margin of error

  W = width of confidence interval

  Then

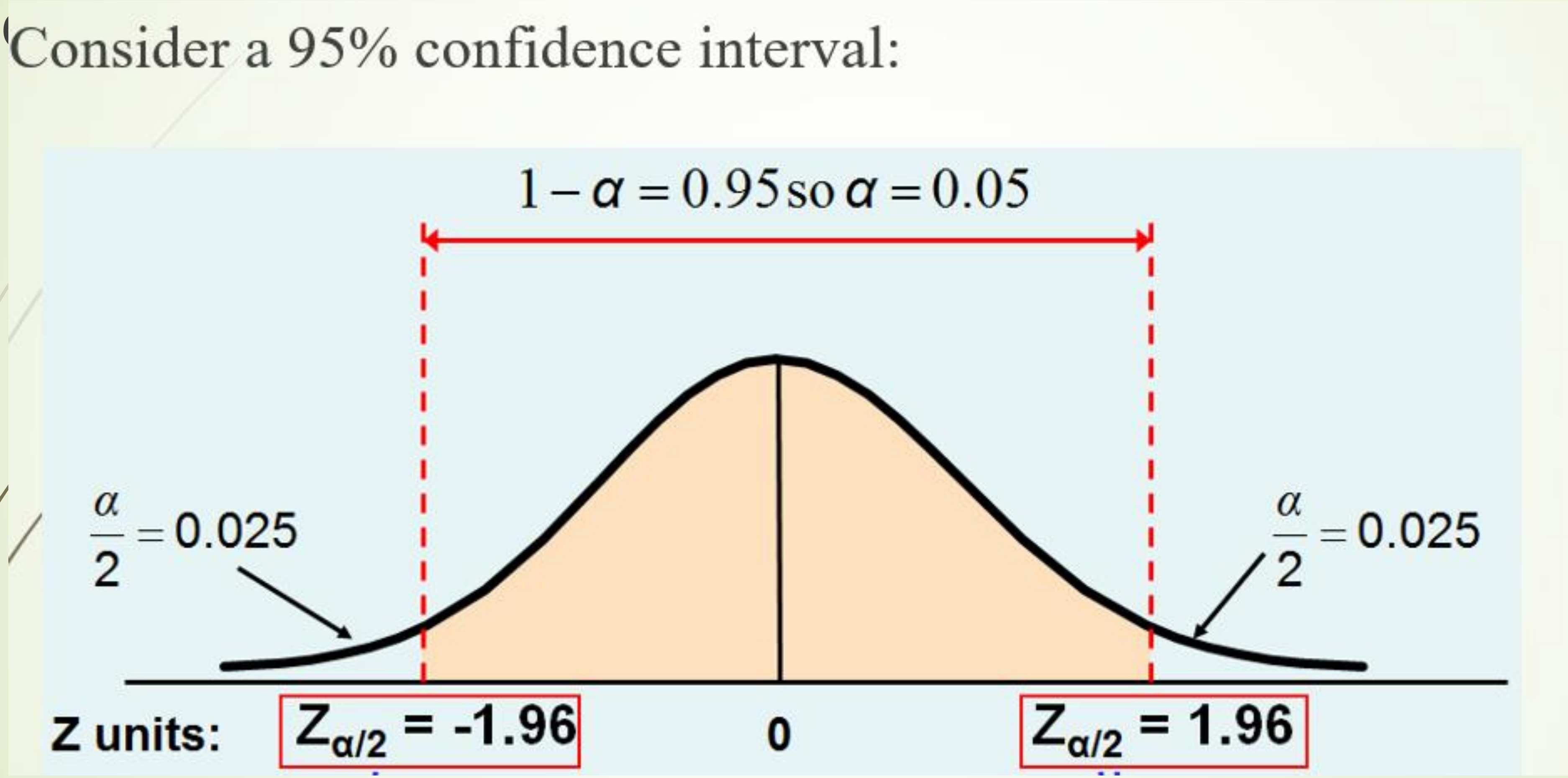
  $\frac{U+L}{2}$ = point estimator

  $\frac{U-L}{2}$ = E where E denotes the error bound

  W = 2E

- Data dispersion as measured by standard deviation
  - Standard deviation ↑ => W ↑
- Sample size
  - Sample size ↑ => W ↓
- Level of confidence
  - Confidence level ↑ => W ↑

Finding the critical value, $z_{\alpha/2}$



Example

A sample of 11 circuits from a large normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is 0.35 ohms. Determine and interpret a 90% confidence interval for the true mean resistance of the population.

$1 - \alpha = 0.9 \Rightarrow \alpha/2 = 0.05 \quad Z_{0.05} = 1.645$

	Area in Upper Tail					
df	0.2	0.1	0.05	0.025	0.01	0.005
$z = t_{\infty}$	0.842	1.282	1.645	1.96	2.326	2.576

$$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 2.2 \pm 1.645\frac{0.35}{\sqrt{11}} = 2.2 \pm 0.174 = [2.026, 2.374]$$

We are 90% confident that the true mean resistance is between 2.026 ohms and 2.374 ohms.

What if σ is unknown?

- If the population standard deviation σ is **unknown**, we can substitute the sample standard deviation, *s*
- This introduces extra uncertainty, since *s* is variable from sample to sample.
- If $X_i \sim N(\mu, \sigma)$, then

$$\mathbf{Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)} \text{ but}$$

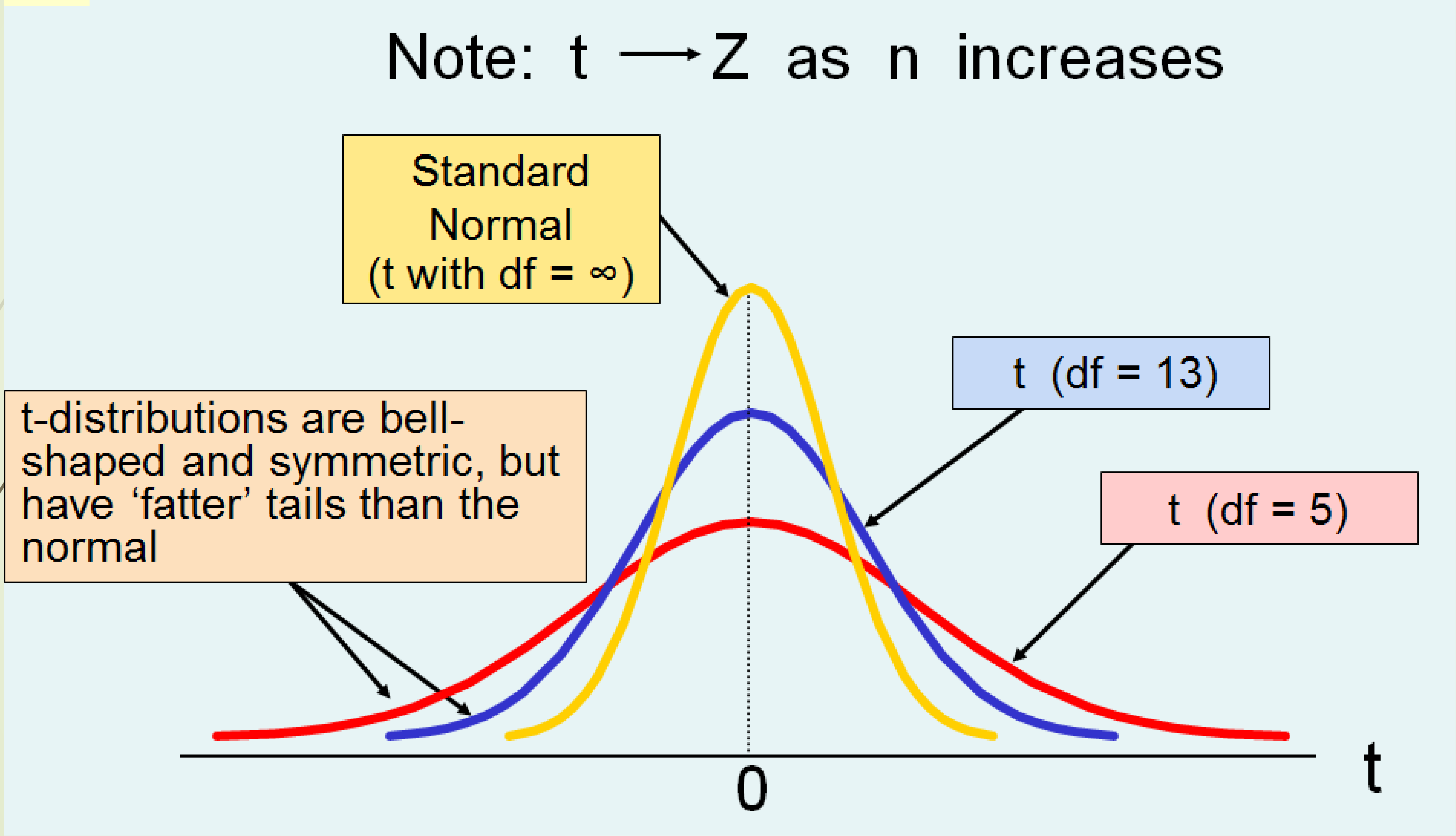
$$\mathbf{T = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}}$$

Characteristics of Student's t distribution

- It is based on the assumption that the population of interest is normal, or nearly normal
- It is a continuous distribution
- It is bell-shaped and symmetric
- There is not one t distribution, but rather a "family" of t distributions. All have the same mean of 0; i.e., $E(t) = 0$. However, their standard deviations differ according to the sample size n.
- The exact shape of the t distribution depends on a parameter called the degrees of freedom, ν
- $\mathrm{Var}(t) = \frac{\nu}{\nu-2} > 1$, so the t distribution is more spread out and flatter at the center than is the standard normal distribution. However, as n increases, the curve representing t distribution approaches the standard normal distribution; i.e.,

$$\mathbf{t_{\infty} = z}$$

z versus t



Note: t ⟶ Z as n increases
Standard Normal (t with df = ∞)
t (df = 13)
t (df = 5)
t-distributions are bell-shaped and symmetric, but have 'fatter' tails than the normal
0
t

t-interval

- Assumptions:
  - **σ is unknown** and population is normal (or $n \geq 30$)
- Interval estimator of μ:

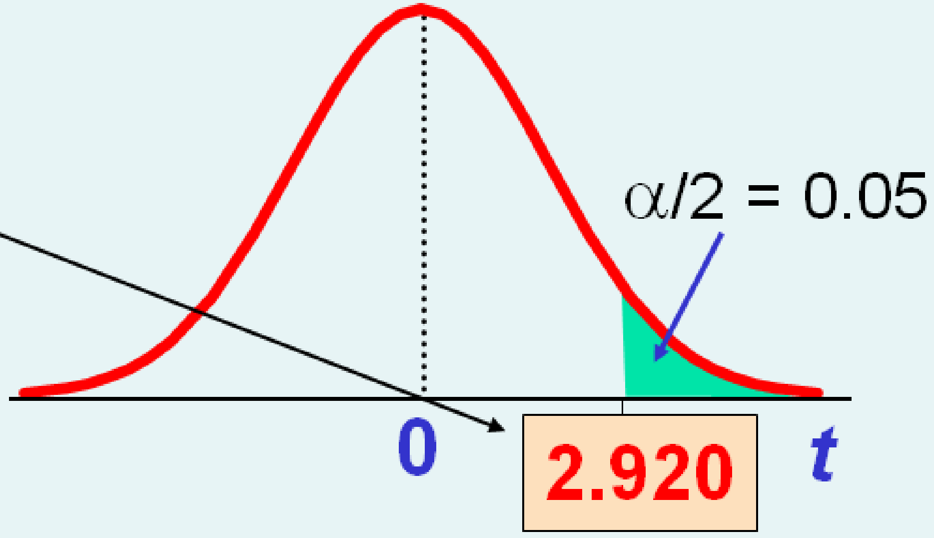
$\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$ where $t_{\alpha/2}$ is based on $df = n - 1$

Student's t Table

df	Upper Tail Area		
	.10	**.05**	.025
1	3.078	6.314	12.706
2	1.886	**2.920**	4.303
3	1.638	2.353	3.182

The body of the table contains t values, not probabilities

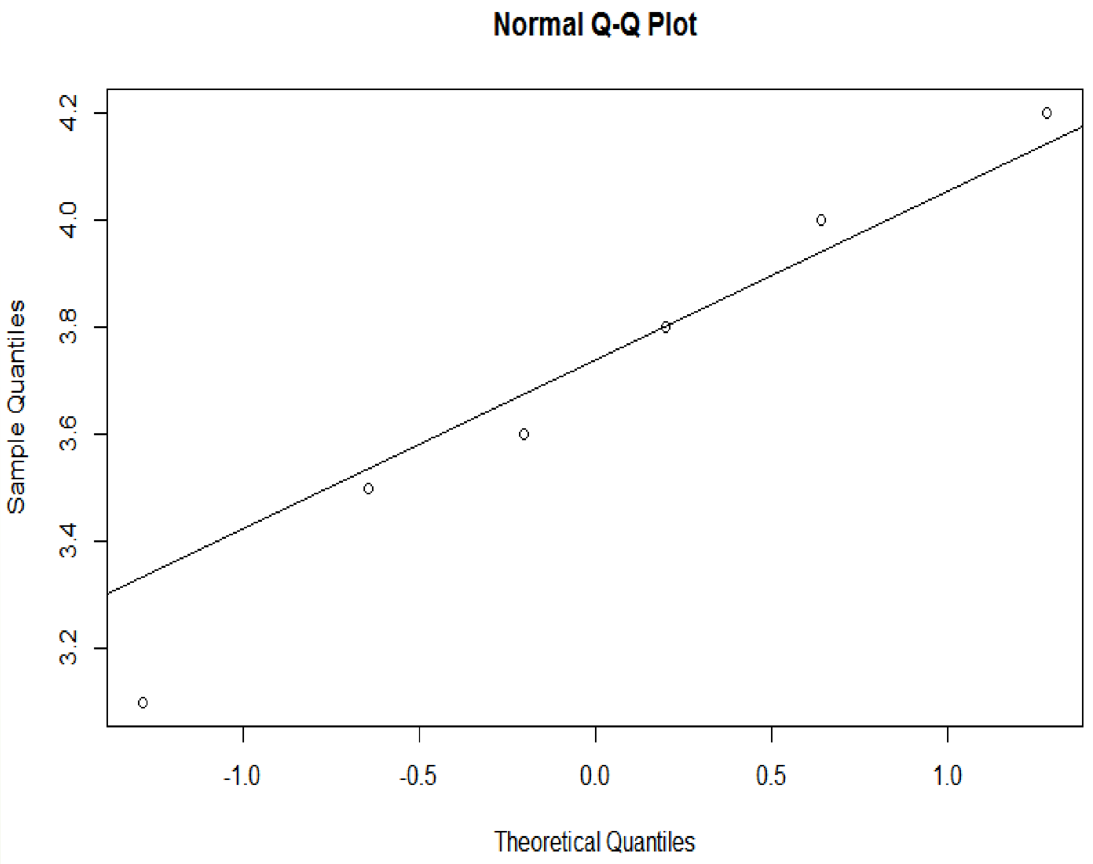
Let: n = 3
df = n - 1 = 2
α = 0.10
$\alpha/2$ = 0.05



Example

You are a time study analyst in manufacturing. You have recorded the following task times (min): 3.6, 4.2, 4.0, 3.5, 3.8, 3.1. Use the following normal Q-Q plot to check if we can assume that the task time is normally distributed. If so, construct a 99% confidence interval estimate of the population mean task time?

Since the dots fall very closely along the line, the task time can be assumed normal.



df = n – 1 = 6 – 1 = 5

1 – α = 0.99 => α/2 = 0.005 $t_{0.005;\,5} = 4.032$

$$\bar{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = 3.7 \pm 4.032\frac{0.3899}{\sqrt{6}}$$

$$= 3.7 \pm 0.642$$

$$= [3.058.\ 4.342]$$

Assessing normality

- Normal quantile plot / QQ plot
  - This is a plot of the data against its normal scores. If the plot is a straight line, then it suggests normality.
- Histogram or stem and leaf plot
  - Check if the histogram has a symmetric bell shape.
- The interquartile range should be close to 1.34898 times the standard deviation; i.e., $IQR \approx 1.34898s$

Z-Interval for *p*

- Assumptions:

  The binomial conditions have been met.

  - The sample data is the result of counts.
  - There are only 2 possible outcomes.
  - The probability of a success remains the same from one trial to the next
  - The trials are independent.

- The sample size is sufficiently large; i.e., **$n > 25$, $np > 5$, and $nq > 5$**. This condition allows us to invoke the central limit theorem and employ the standard normal distribution, that is, z, to complete a confidence interval.

- Interval estimator of p: $\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$

Example

You are a production manager for a newspaper. You want to find the % defective. Of 200 newspapers, 35 had defects. Let *p* be the true proportion defective.

a. Obtain the best point estimate of *p*. $\hat{p} = \frac{x}{n} = \frac{35}{200} = 0.175$

b. Is the sample size large enough to invoke the Central Limit Theorem?

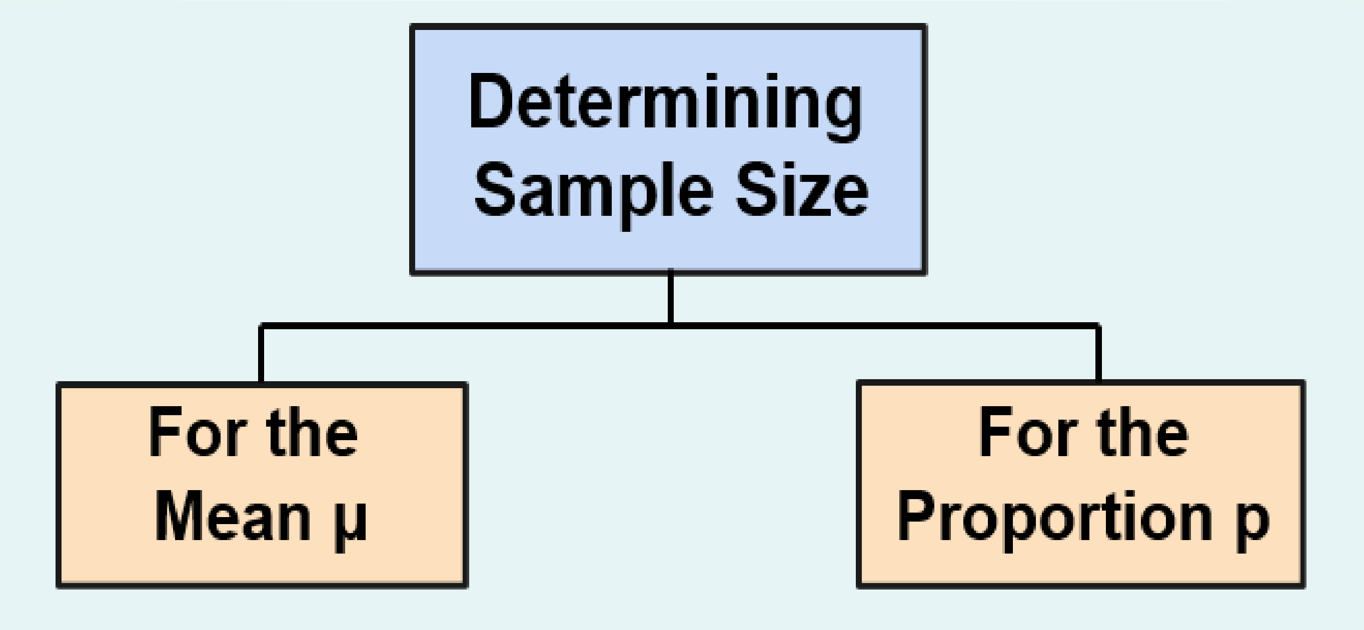
Yes, because $n = 200 > 25$, $n\hat{p} = 200(0.175) = 35$ and $n\hat{q} = 200(165)$

c. What is the 90% confidence interval estimate of the population proportion defective?

$1 - \alpha = 0.9 \Rightarrow \alpha/2 = 0.05 \quad Z_{0.05} = 1.645$

$$\hat{\mathbf{p}} \pm \mathbf{Z}_{\alpha/2}\sqrt{\frac{\hat{\mathbf{p}}\hat{\mathbf{q}}}{\mathbf{n}}} = 0.175 \pm 1.645\sqrt{\frac{0.175(0.825)}{200}} = 0.175 \pm 0.044 = [0.131, 0.219]$$

Determining Sample Size



Determining sample size - μ

Determining Sample Size

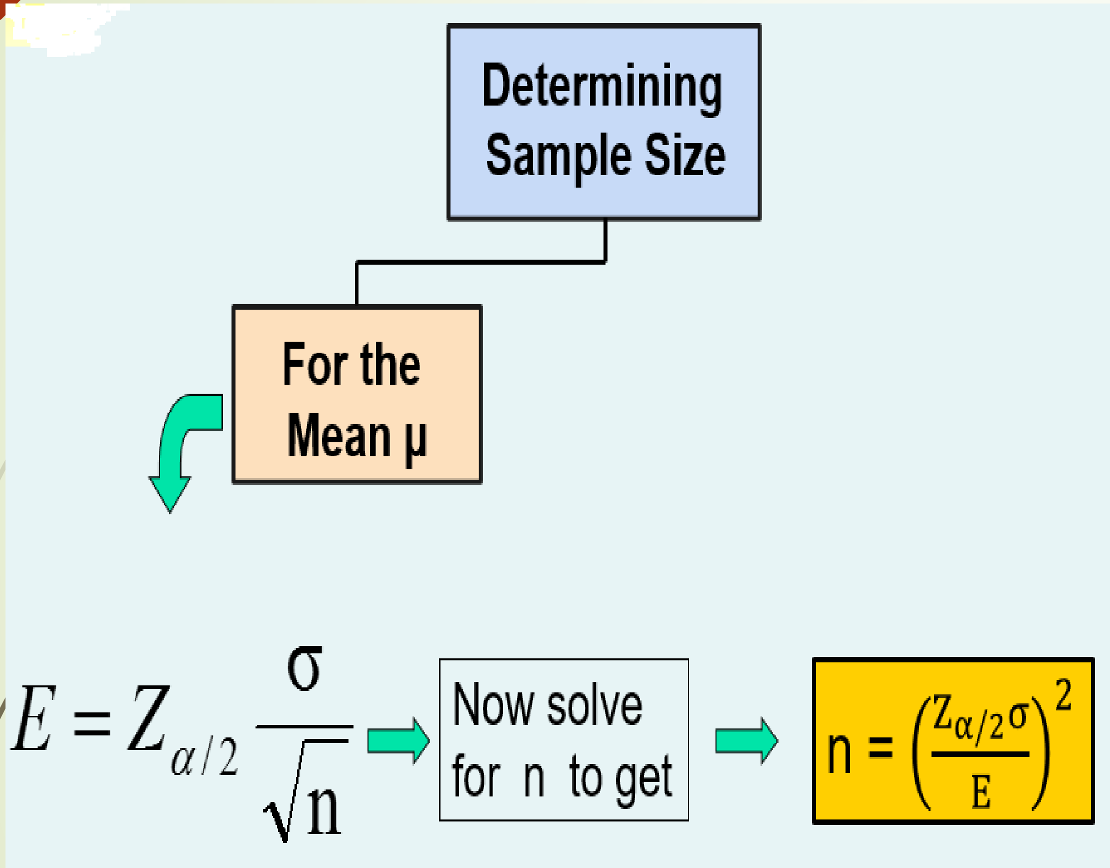
For the Mean

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Sampling error (margin of error)

$$E = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Determining sample size - μ



To determine the required sample size for μ, you must know:

- The desired level of confidence $(1 - \alpha)$, which determines the critical value, $Z_{\alpha/2}$
- The margin of error (or error bound), E
- The standard deviation, σ

If **σ is unknown**, it can be estimated by

- selecting a pilot sample and estimating σ with the sample standard deviation, *s*
- **$\sigma \approx \frac{\text{Range}}{4}$**

Always round up to the nearest integer.

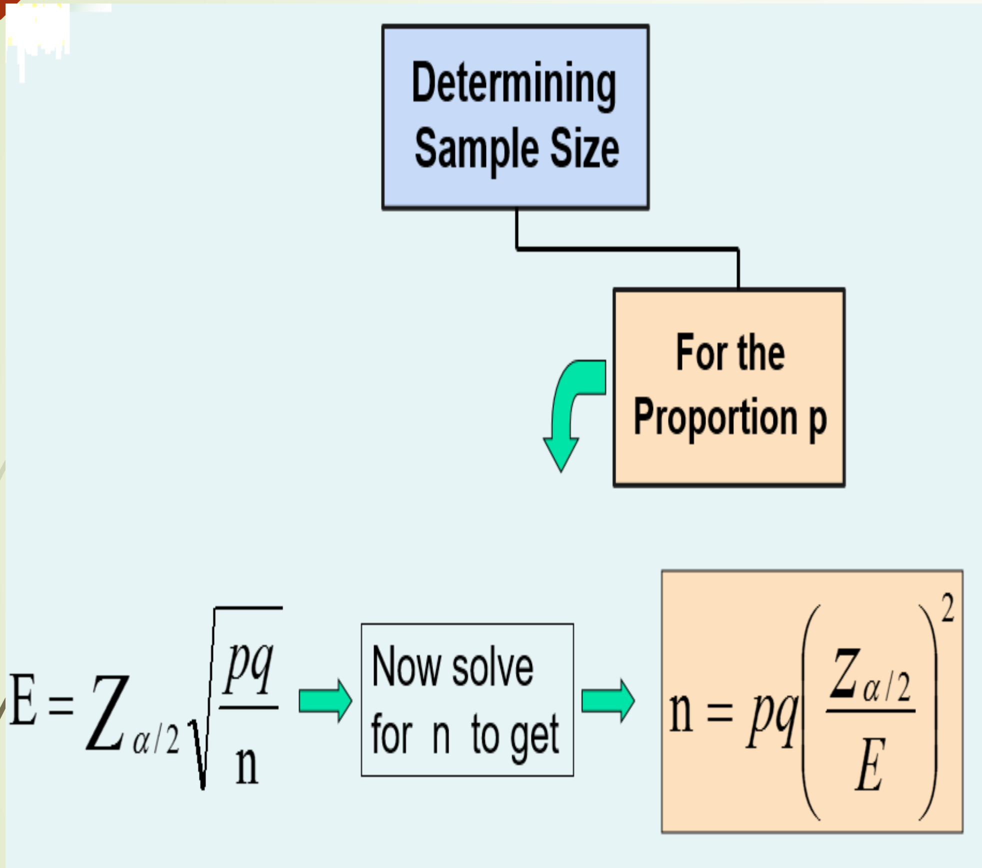
Example

If $\sigma = 45$, what sample size is needed to estimate the mean within ± 5 with 90% confidence?

$1 - \alpha = 0.9 \Rightarrow \alpha/2 = 0.05$ $\quad z_{0.05} = 1.645$

$$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{1.645 * 45}{5}\right)^2 = 219.188 \approx 220$$

Determining sample size - p



To determine the required sample size for *p*, you must know:

- The desired level of confidence (1 - α), which determines the critical value, $Z_{\alpha/2}$
- The margin of error (or error bound), E
- The true proportion of events of interest, *p*

If ***p* is unknown**, *p* can be estimated by

- selecting a pilot sample and estimating *p* with the sample proportion, $\hat{p}$
- conservatively using 0.5 as an estimate of *p*

Example

How large a sample would be necessary to estimate the true proportion defective in a large population within ±3%, with 95% confidence? (Assume a pilot sample yields $p = 0.12$)

$1 - \alpha = 0.95 \Rightarrow \alpha/2 = 0.025$ $z_{0.05} = 1.96$

$$n = pq\left(\frac{z_{\alpha/2}}{E}\right)^2 = 0.12(0.88)\left(\frac{1.96}{0.03}\right)^2 = 450.748 \approx 451$$

Determining Sample Size

For the Mean μ

$$n = \left(\frac{Z_{\alpha/2}\sigma}{E}\right)^2$$

Always round up to the next greater integer regardless of the decimal part

For the Proportion p

$$n = pq\left(\frac{Z_{\alpha/2}}{E}\right)^2$$

Always round up to the next greater integer regardless of the decimal part

Estimate μ	Estimate p
If the sample size is large enough; i.e., **$n \geq 30$**, then it is appropriate to use the normal distribution to approximate the sampling distribution of $\overline{X}$. By the CLT, $\overline{X} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ <u>Case 1</u>: σ is known and population is normal (or $n \geq 30$): z-confidence interval for μ: $\overline{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ <u>Case 2</u>: σ is unknown and population is normal (or $n \geq 30$): t-confidence interval for μ: $\overline{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$	If the sample size is large enough; i.e., **$n > 25$, $n\hat{p} > 5$ and $n\hat{q} > 5$**, then it is appropriate to use the normal distribution to approximate the sampling distribution of $\hat{p}$. By the CLT, $\hat{p} \approx N\left(p, \sqrt{\frac{pq}{n}}\right)$ z-confidence interval for p: : $\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$
Sample size required to estimate μ: $$n = \left(\frac{Z_{\alpha/2}\sigma}{E}\right)^2$$ If σ is unknown, use s. If s is unavailable, use $\sigma \approx \frac{\text{Range}}{4}$	Sample size required to estimate p: $$n = pq\left(\frac{Z_{\alpha/2}}{E}\right)^2$$ If p is unknown, use $\hat{p}$. If $\hat{p}$ is unavailable, take p = 0.5

Interval estimation: estimate $(\mu_1 - \mu_2)$

- Let $X_{11}, X_{12}, \ldots, X_{1n1}$ be a random sample from a distribution with μ_1, σ_1, and let $X_{21}, X_{22}, \ldots, X_{2n2}$ be another sample independent from the first one, from a distribution with μ_2, σ_2.
- $\widehat{\mu_1 - \mu_2} = \bar{X}_1 - \bar{X}_2$
- $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$
- $Var(\bar{X}_1 - \bar{X}_2) = Var(\bar{X}_1) + Var(\bar{X}_2) = \begin{cases} \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2} & \text{if } \sigma_1^2 \neq \sigma_2^2 \\ \sigma^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right) & \text{if } \sigma_1^2 = \sigma_2^2 = \sigma^2 \end{cases}$
- $(\bar{X}_1 - \bar{X}_2) \sim N(\mu_{\bar{X}_1 - \bar{X}_2}, \sigma_{\bar{X}_1 - \bar{X}_2})$

Interval estimation: estimate $(\mu_1 - \mu_2)$

- Two populations (independent) – σ_1^2 and σ_2^2 are known
- Assumptions:
  - The populations are normal or approximately normal (CLT for large n); i.e., $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$.
  - The samples are randomly and independently drawn from the respective populations
  - The population variances σ_1^2 and σ_2^2 are known
- z-confidence interval: $(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Interval estimation: estimate $(\mu_1 - \mu_2)$

- Two populations (independent) – σ_1^2 and σ_2^2 are known
- Assumptions:
  - The populations are normal or approximately normal (CLT for large n); i.e., $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$.
  - The samples are randomly and independently drawn from the respective populations
  - The population variances σ_1^2 and σ_2^2 are known
- z-confidence interval: $(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Interval estimation: estimate $(\mu_1 - \mu_2)$

- Two populations (independent) – σ_1^2 and σ_2^2 are unknown
- Case 1: $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (unknown)
- Pooled sample variance: $\hat{\sigma}^2 = S_p^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$
- t-confidence interval: $(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2}\sqrt{S_p^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}$ where df $= n_1 + n_2 - 2$
- Case 2: $\sigma_1^2 \neq \sigma_2^2$ (unknown)
- t-confidence interval: $(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2}\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$ where df $= \dfrac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_1^2}{n_1}\right)^2}{n_1-1} + \dfrac{\left(\dfrac{s_2^2}{n_2}\right)^2}{n_2-1}}$

Interval estimation: estimate $(\mu_1 - \mu_2)$

- Two populations (dependent) – σ_d unknown
- $\mu_d = \mu_1 - \mu_2$
- $d_i = X_{1i} - X_{2i}$ (eliminates variation among subjects)
- Assumptions:
  - Population of differences is normal or $n > 30$ with σ_d unknown
  - The differences are randomly selected from the population of difference.

- t-confidence interval for μ_d : $\bar{d} \pm t_{\alpha/2} \frac{s_d}{\sqrt{n}}$ where $df = n - 1$

- Note: t-confidence interval for μ: $\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$ where $df = n - 1$