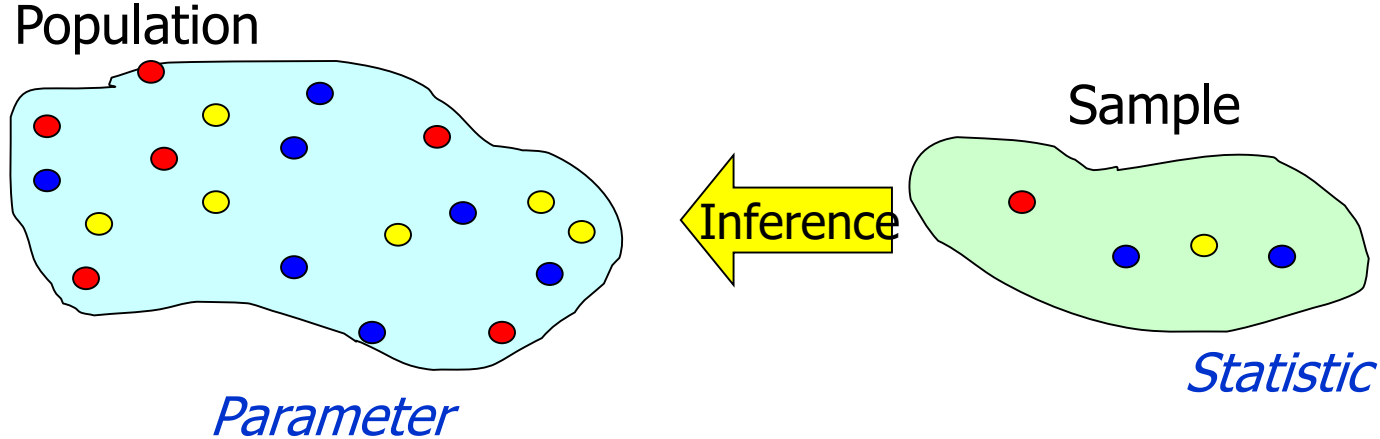


Foundation of Data Science and Analytics

Estimation using Confidence Intervals

Arun K. Timalsina

Estimation



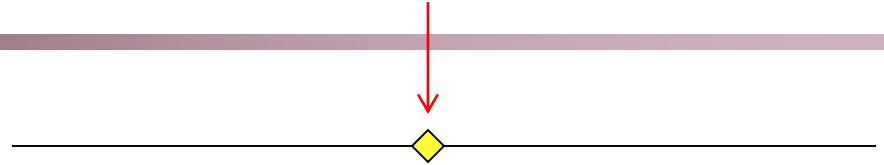
- There are two types of inference: estimation and hypothesis testing; **estimation** is introduced first.

E.g., the sample mean (\bar{x}) is used to **estimate** the population mean (μ).

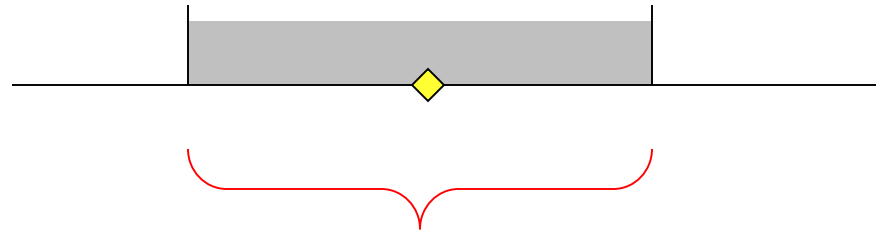
- The objective of estimation is to determine the **approximate value** of a population parameter on the basis of a sample statistic.

Point & Interval Estimation

1) Point Estimate



2) Interval Estimate



- **Point estimate:**

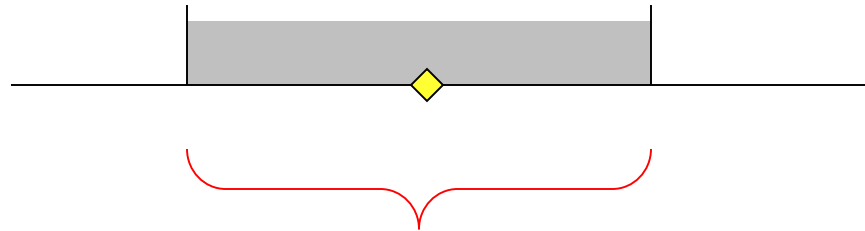
A point estimate is a single number,

- **Interval estimate**

Provides additional information about variability

Point & Interval Estimation

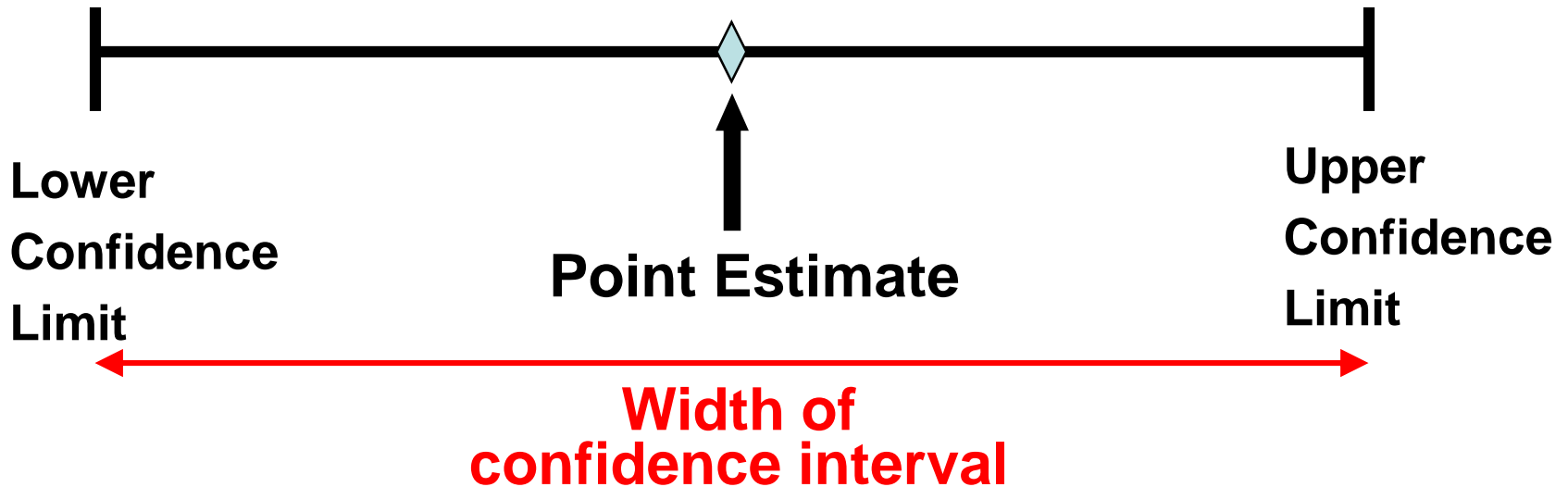
- An ***interval estimator*** draws inferences about a population by estimating the value of an unknown parameter using an interval.



- That is we say (with some ____% certainty) that the population parameter of interest is between some lower and upper bounds.

Point & Interval Estimation

- **A point estimate** is a rule or formula that tells us how to calculate a numerical estimate based on the measurements contained in the sample
 - A point estimate is a single number,
- **An interval estimator** is a formula that tell us how to use sample data to calculate an interval that estimates a population
 - A confidence interval provides additional information about variability



Point & Interval Estimation

For example, suppose we want to estimate the mean summer income of a class of business students. For $n=25$ students, \bar{x} is calculated to be 400 \$/week.

point estimate



interval estimate



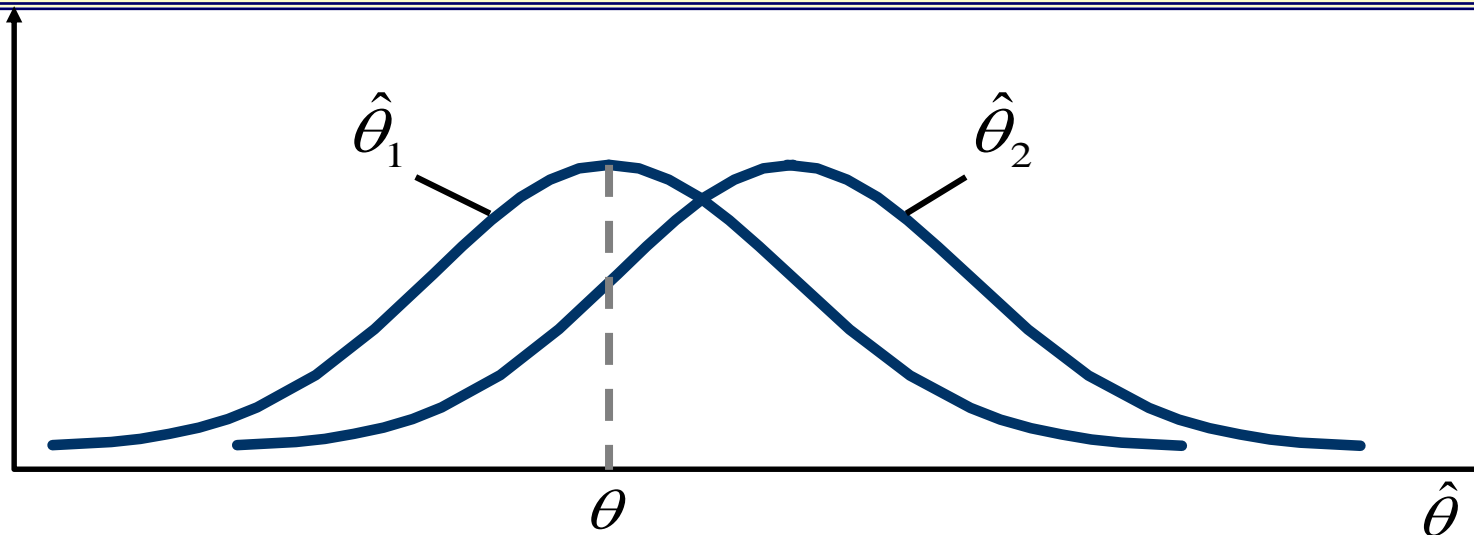
An alternative statement is:

The mean income is **between** 380 and 420 \$/week.

Unbiasedness

The point estimator $\hat{\theta}$ is said to be an unbiased estimator of the parameter θ if the expected value, or mean, of the sampling distribution of $\hat{\theta}$ is θ ; that is,

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



Minimum Variance Unbiased Estimator (MVUE)

Suppose there are several unbiased estimators of θ .

Then the unbiased estimator with the smallest variance is said to be **the most efficient estimator** or to be **the minimum variance unbiased estimator** of θ .

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two unbiased estimators of θ , based on the same number of sample observations. Then,

a) $\hat{\theta}_1$ is said to be more efficient than $\hat{\theta}_2$ if

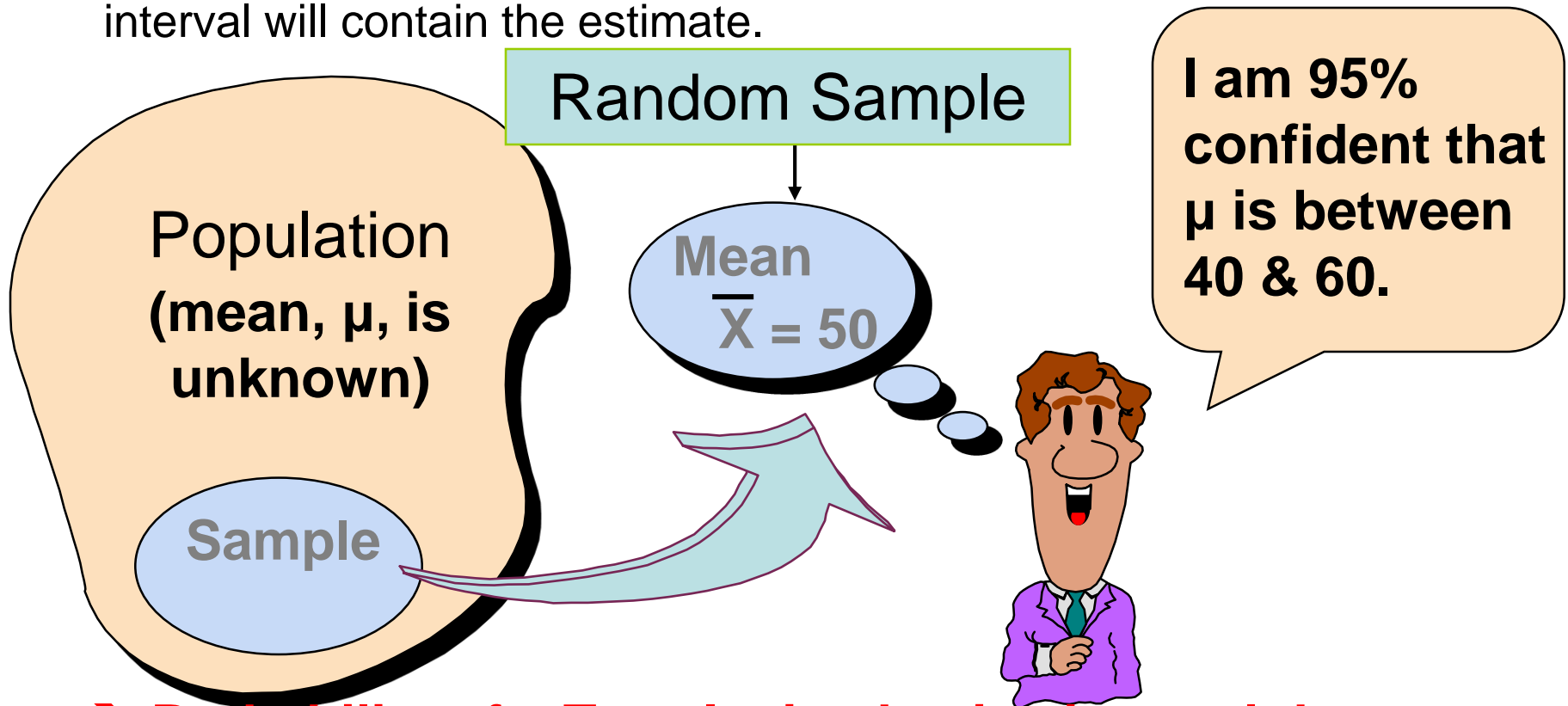
$$Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$$

b) The relative efficiency of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$ is the ratio of their variances; that is,

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

Confidence Coefficient

- Confidence coefficient: The probability that the random interval, prior to sampling, will contain the estimated parameter $(1 - \alpha)$.
- You specify α** \rightarrow $(1 - \alpha)$ 100% confidence that the constructed interval will contain the estimate.

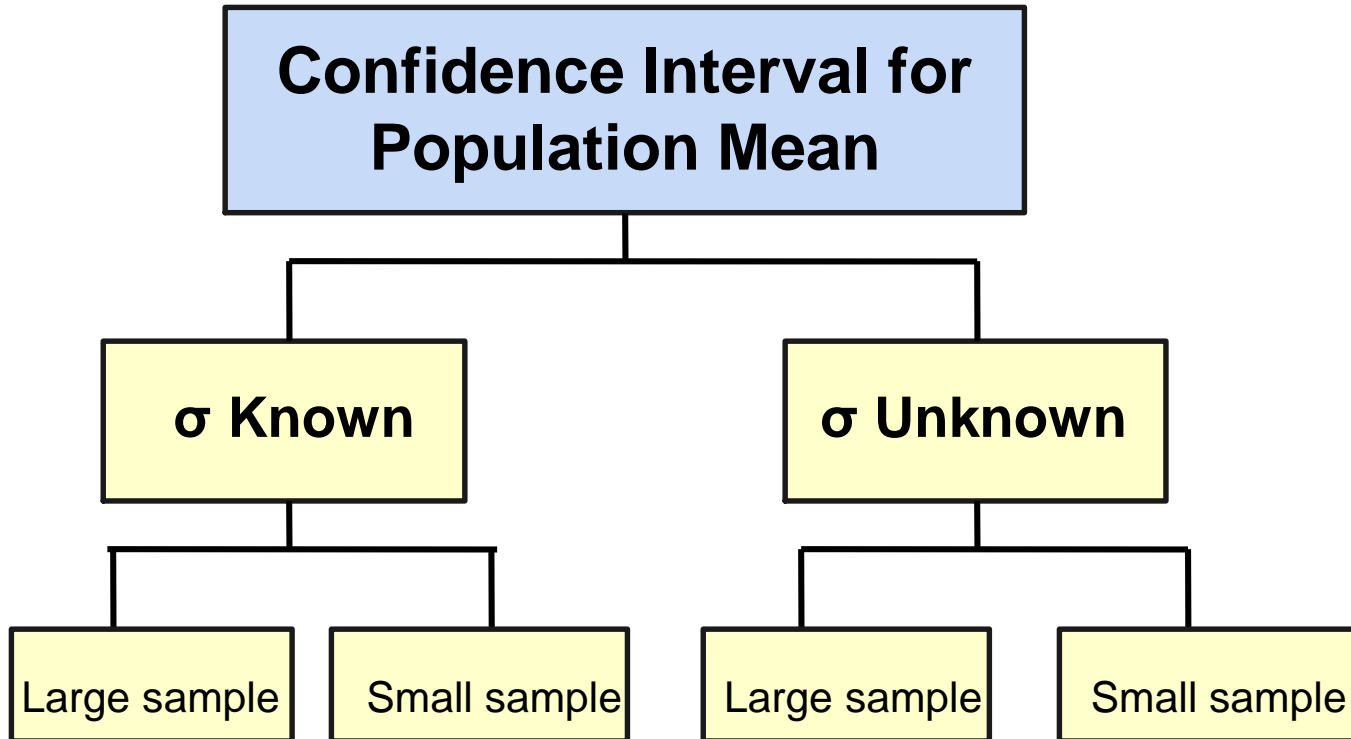


$\alpha \rightarrow$ Probability of a Type I mistake: i.e. interval does not actually contain parameter.

Confidence Coefficient

- Suppose **confidence level** = 95%
- Also written **$(1 - \alpha)$** = 0.95
- A relative frequency interpretation:
 - In the long run, 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

Confidence Intervals for Population Mean



Confidence Interval for Population Mean

- We can calculate an interval estimator from a sampling distribution, by:

Drawing a sample of size n from the population

Calculating its mean,

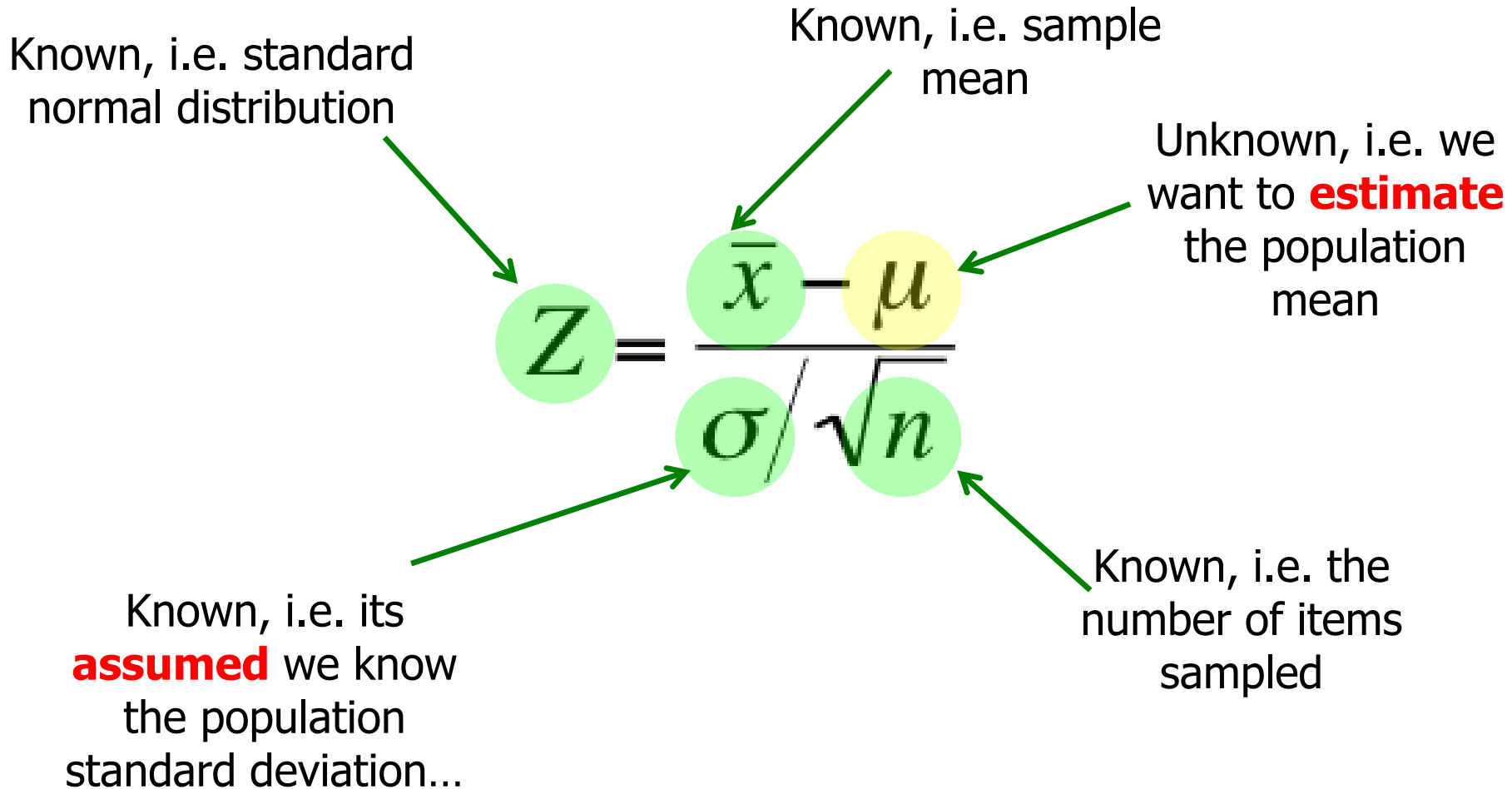
And, by the central limit theorem, we know that \bar{X} is normally (or approximately normally) distributed so...

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

- ...will have a standard normal (or approximately normal) distribution.

Confidence Interval for Population Mean

Looking at this in more detail...



Confidence Interval for Population Mean

$$P\left(\mu - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

- Thus, the **probability** that the interval:

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = \left\{ \bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\}$$

contains the population mean μ is $1 - \alpha$. This is a **confidence interval estimator for μ** .

Confidence Interval for μ (σ known)

For samples of size > 30 , the confidence interval is expressed as

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

$$\bar{x} \pm z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$$

Large Sample

For samples of size < 30 , the confidence interval is expressed as

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

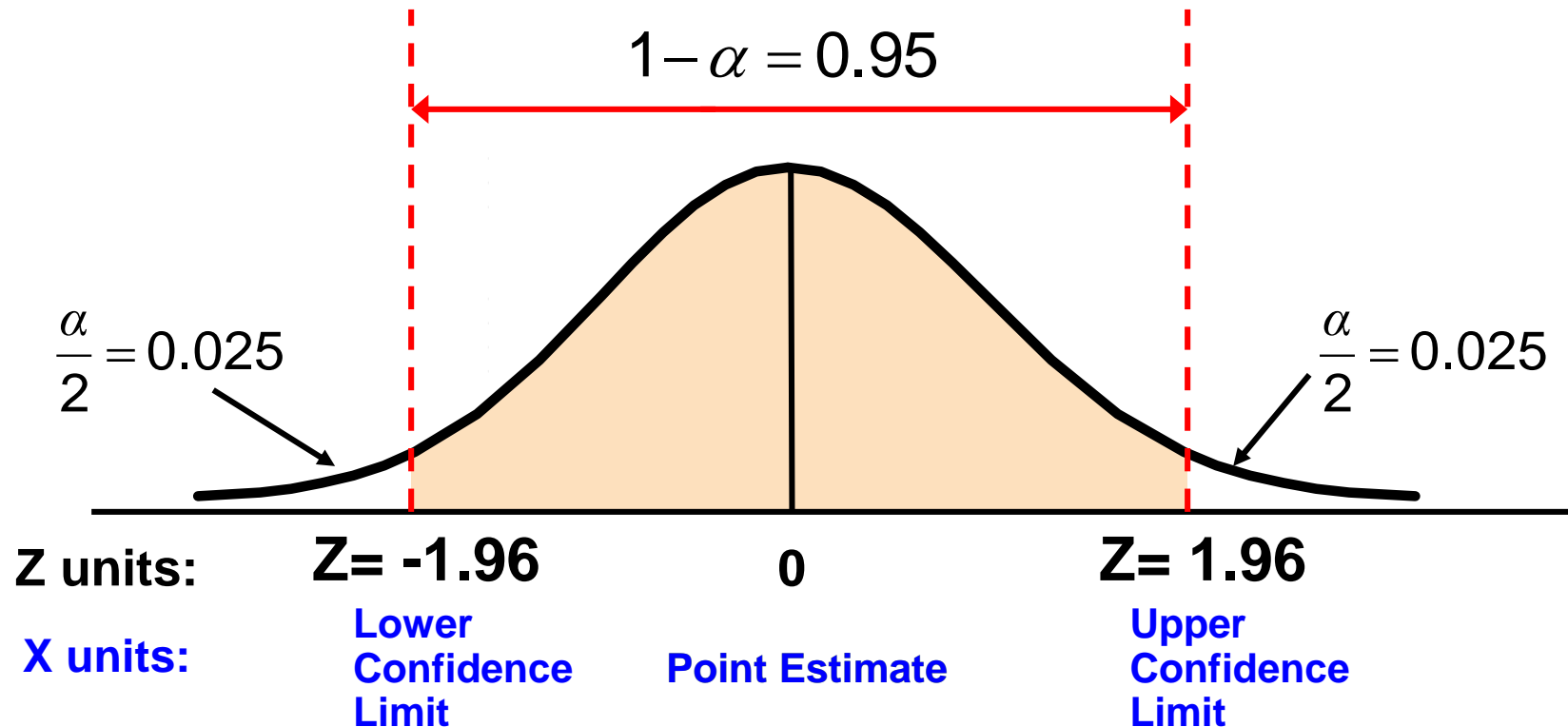
$$\bar{x} \pm t_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$$

Small Sample

Finding the Critical Value, Z (Example)

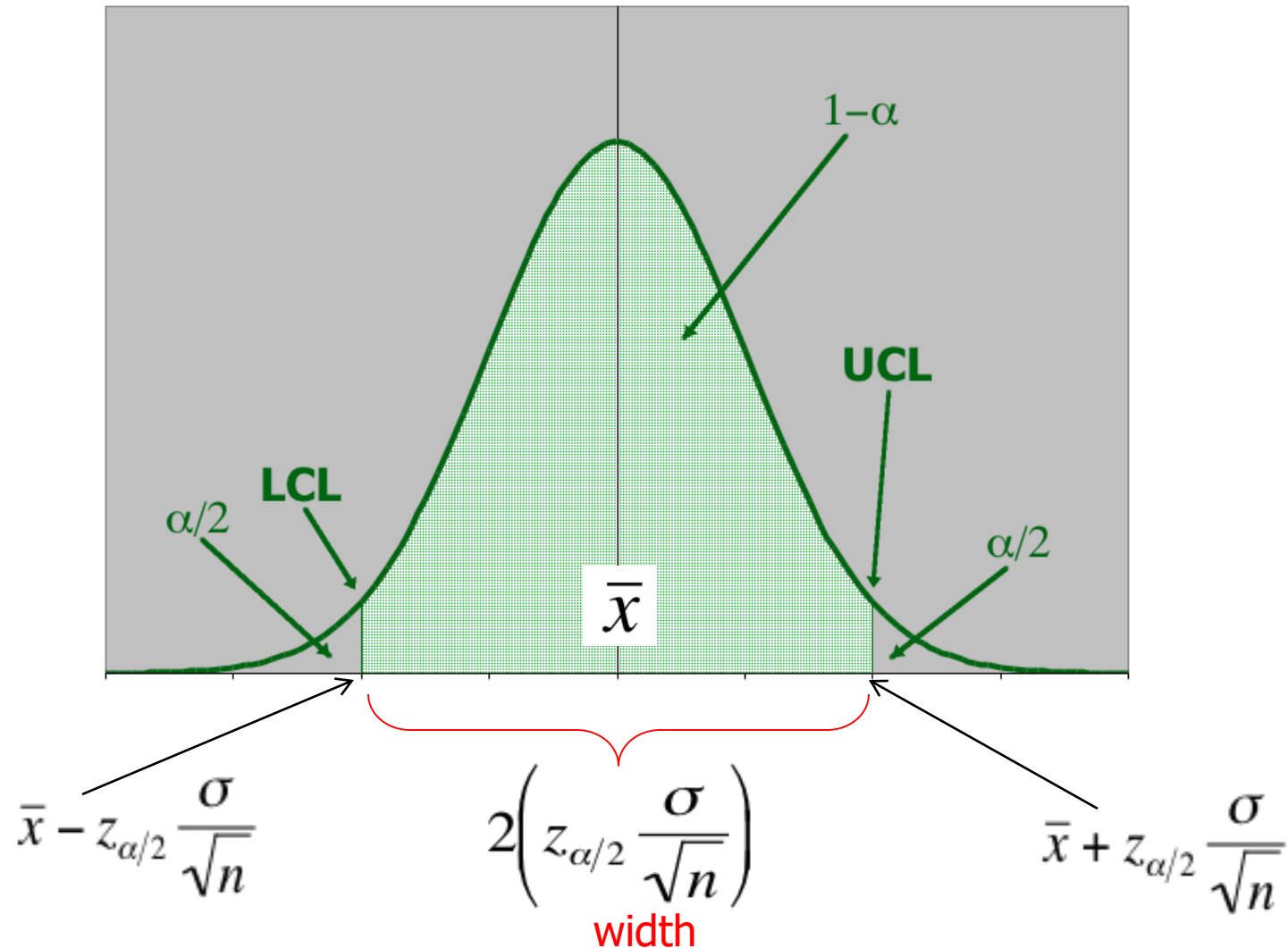
α	0.01	0.02	0.05	0.10
$Z_{\alpha/2}$	2.58	2.33	1.96	1.645
Confidence Level	99%	98%	95%	90%

Consider a 95% confidence interval: $Z = \pm 1.96$



Graphically...

...here is the confidence interval for μ :



Example

- A sample of 40 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 a 95% confidence interval for the true mean resistance of the population.

Example

$$\begin{aligned}\bar{X} \pm Z \frac{\sigma}{\sqrt{n}} \\ = 2.20 \pm 1.96 (0.35/\sqrt{40}) \\ = 2.20 \pm 0.108\end{aligned}$$

$$2.092 \leq \mu \leq 2.308$$

We are 95% confident that the true mean resistance is between 2.092 and 2.308 ohms

Although the true mean may or may not be in this interval, 95% of intervals formed in this manner will contain the true mean

Confidence Interval for μ (σ Unknown)

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

For samples of size > 30 , the confidence interval is expressed as

$$\bar{x} \pm z_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right)$$

Large Sample

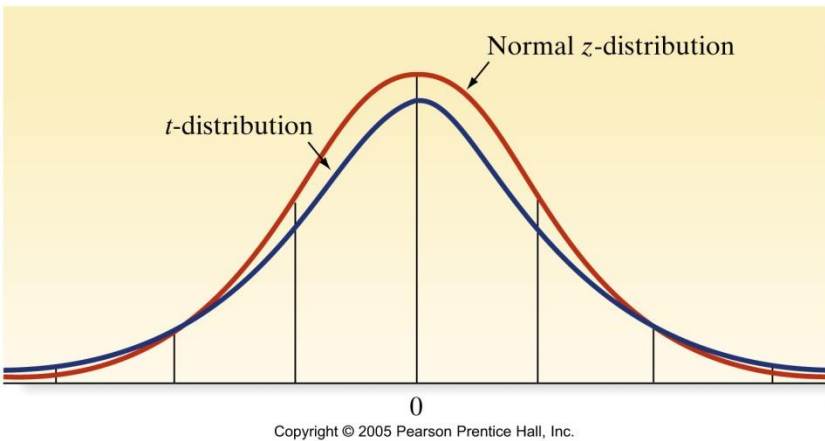
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

For samples of size < 30 , the confidence interval is expressed as

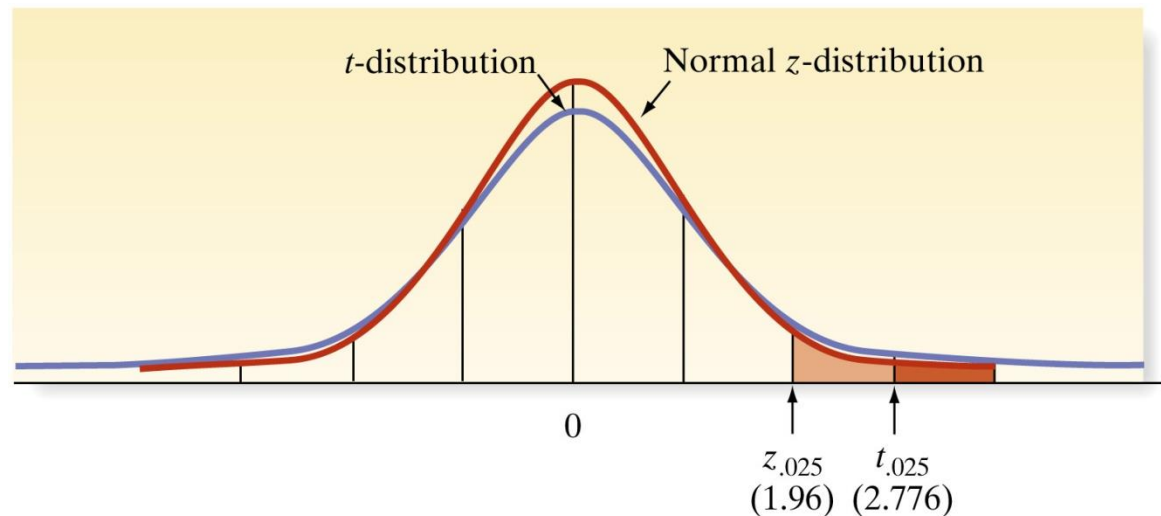
$$\bar{x} \pm t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right)$$

Small Sample

Small-Sample Confidence Interval for a Population Mean



- The t-statistic has: a sampling distribution very similar to z
- Variability dependent on n , or sample size.
- Variability is expressed as $(n-1)$ **degrees of freedom (df)**. As (df) gets smaller, variability increases

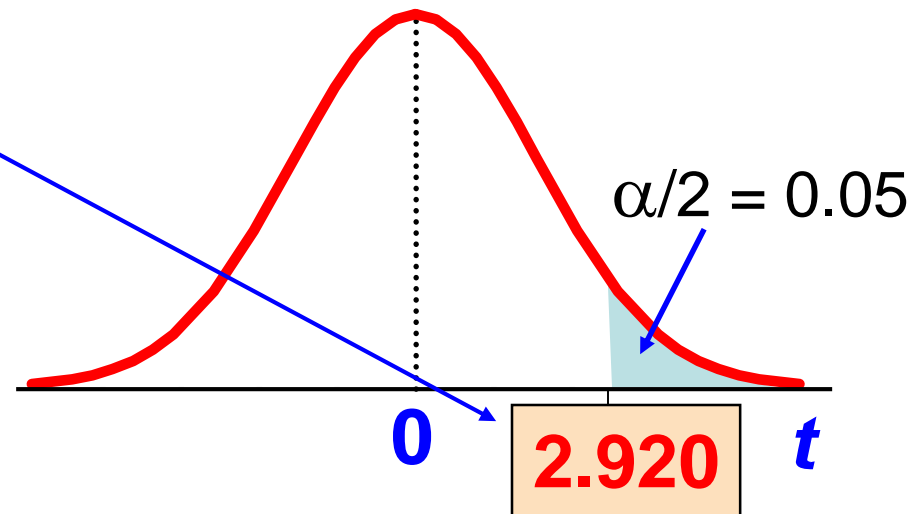


Student's t Table

Upper Tail Area			
df	.25	.10	.05
1	1.000	3.078	6.314
2	0.817	1.886	2.920
3	0.765	1.638	2.353

The body of the table contains t values, not probabilities

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



Example

A random sample of $n = 25$ has $\bar{X} = 50$ and $S = 8$. Form a 95% confidence interval for μ

– d.f. = $n - 1 = 24$, so $t_{\alpha/2, n-1} = t_{0.025, 24} = 2.0639$

The confidence interval is

$$\bar{X} \pm t_{\alpha/2, n-1} \frac{S}{\sqrt{n}} = 50 \pm (2.0639) \frac{8}{\sqrt{25}}$$

$$46.698 \leq \mu \leq 53.302$$

Confidence Interval for a Proportion

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

$$-z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \leq z_{\alpha/2}$$

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \sim N(0,1)$$

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

Example

In a random sample of 85 automobile engine crankshaft bearings, 10 have a surface finish that is rougher than the specifications allow. Therefore, a point estimate of the proportion of bearings in the population that exceeds the roughness specification is $\hat{p} = x/n = 10/85 = 0.12$. A 95% two-sided confidence interval for p is computed from

$$\hat{p} - z_{0.025} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \leq p \leq \hat{p} + z_{0.025} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

$$0.12 - 1.96 \sqrt{\frac{0.12(0.88)}{85}} \leq p \leq 0.12 + 1.96 \sqrt{\frac{0.12(0.88)}{85}}$$

$$0.05 \leq p \leq 0.19$$

Example

- In a random sample of 80 automotive crankshaft bearings, 15 of the bearings have a surface finish that is rougher than the specifications will allow.
 - The point estimate of the fraction nonconforming in the process: $15/80 = 0.1875$
 - Assuming that the normal approximation is appropriate
 - 95% two-sided confidence interval: $\alpha = 0.05$ $z_{\alpha/2} = 1.96$

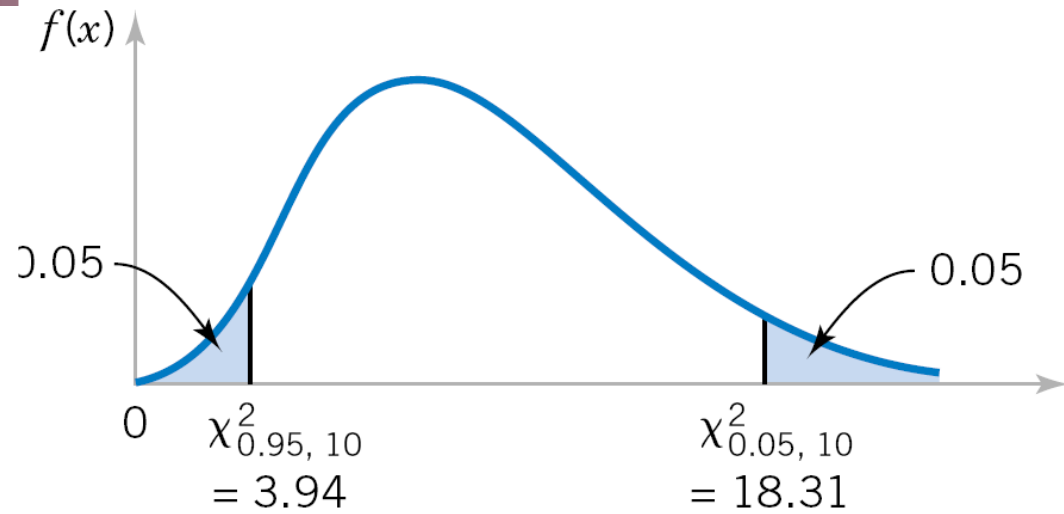
$$\hat{p} = 0.1875, z_{\alpha/2} = 1.96$$

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

$$0.1020 \leq p \leq 0.2730$$

Confidence Interval for the Variance

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$$



$$\chi_{1-\alpha/2, n-1}^2 \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi_{\alpha/2, n-1}^2$$

$$\frac{(n-1)s^2}{\chi_{\alpha/2, n-1}^2} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi_{1-\alpha/2, n-1}^2}$$

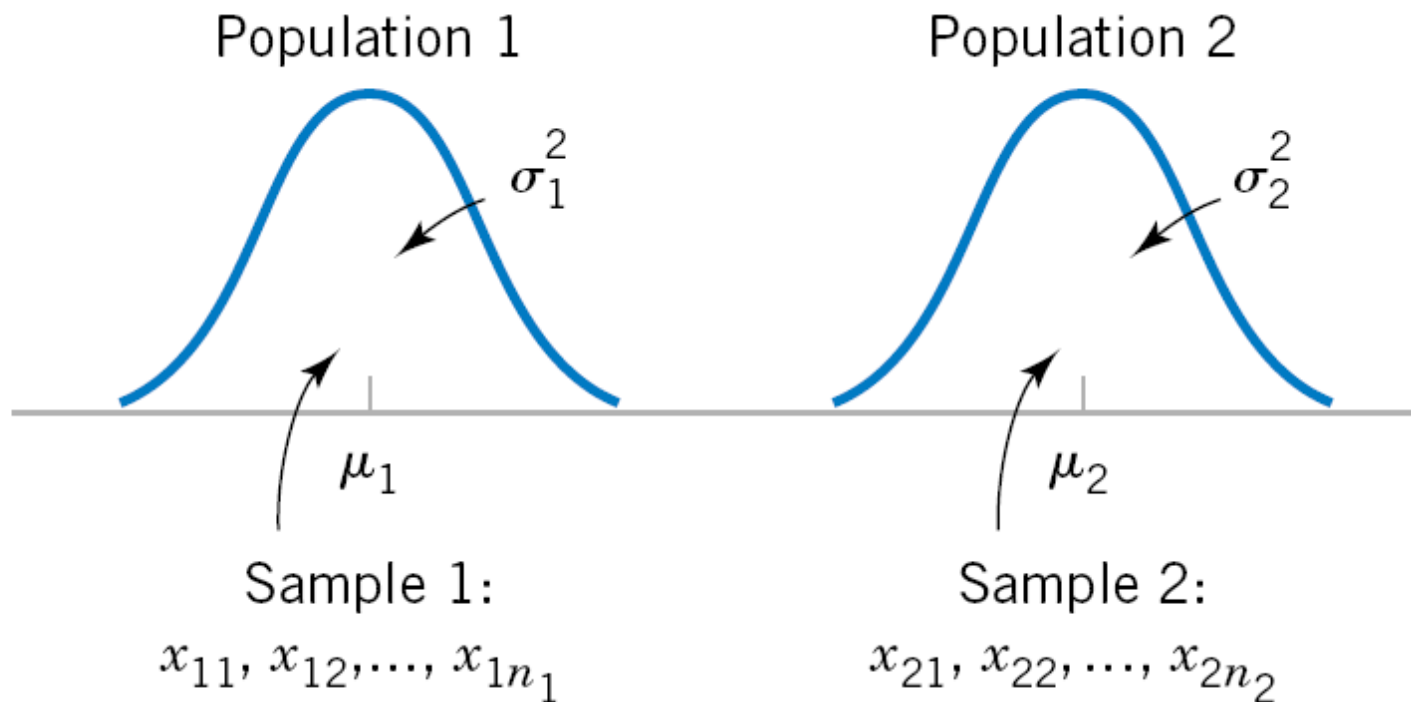
Example

An automatic filling machine is used to fill bottles with liquid detergent. A random sample of 20 bottles results in a sample variance of fill volume of $s^2 = 0.0153$ (fluid ounces)². If the variance of fill volume is too large, an unacceptable proportion of bottles will be under- or overfilled. We will assume that the fill volume is approximately normally distributed. A 95% upper-confidence interval is found from

$$\sigma^2 \leq \frac{(n - 1)s^2}{\chi_{0.95,19}^2}$$

$$\sigma^2 \leq \frac{(19)0.0153}{10.117} = 0.0287 \text{ (fluid ounce)}^2$$

Confidence Interval Between Two Populations



Confidence Interval for the Difference Between Two Means

Large Sample

σ Known

$$\bar{x}_1 - \bar{x}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Large Sample

σ Unknown

$$(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq (\mu_1 - \mu_2) \leq (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Confidence Interval for the Difference Between Two Means (σ unknown and equal)

$$\sigma_1^2 = \sigma_2^2$$

Small Sample

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$\begin{aligned} (\bar{X}_1 - \bar{X}_2) - t_{\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} &\leq (\mu_1 - \mu_2) \\ &\leq (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \end{aligned}$$

Example

An article in the journal *Hazardous Waste and Hazardous Materials* (Vol. 6, 1989) reported the results of an analysis of the weight of calcium in standard cement and cement doped with lead. Reduced levels of calcium would indicate that the hydration mechanism in the cement is blocked and would allow water to attack various locations in the cement structure. Ten samples of standard cement had an average weight percent calcium of $\bar{x}_1 = 90.0$, with a sample standard deviation of $s_1 = 5.0$, while 15 samples of the lead-doped cement had an average weight percent calcium of $\bar{x}_2 = 87.0$, with a sample standard deviation of $s_2 = 4.0$.

$$\begin{aligned}s_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\&= \frac{9(5.0)^2 + 14(4.0)^2}{10 + 15 - 2} \\&= 19.52\end{aligned}$$

Example (cont.)

$$\bar{x}_1 - \bar{x}_2 - t_{0.025,23} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t_{0.025,23} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$



$$90.0 - 87.0 - 2.069(4.4) \sqrt{\frac{1}{10} + \frac{1}{15}} \leq \mu_1 - \mu_2 \leq 90.0 - 87.0 + 2.069(4.4) \sqrt{\frac{1}{10} + \frac{1}{15}}$$



$$-0.72 \leq \mu_1 - \mu_2 \leq 6.72$$

Confidence Interval for the Difference Between Two Means (σ unknown and unequal)

Small Sample

$$\sigma_1^2 \neq \sigma_2^2$$

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2, v} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t_{\alpha/2, v} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$n_1 = n_2 = n \Rightarrow \quad v = n_1 + n_2 - 2$$

$$n_1 \neq n_2 \quad \Rightarrow \quad v = \frac{(s_1^2 / n_1 + s_2^2 / n_2)^2}{(s_1^2 / n_1)^2 / (n_1 - 1) + (s_2^2 / n_2)^2 / (n_2 - 1)}$$

Example

Tensile strength tests were performed on two different grades of aluminum spars used in manufacturing the wing of a commercial transport aircraft. From past experience with the spar manufacturing process and the testing procedure, the standard deviations of tensile strengths are assumed to be known. The data obtained are as follows: $n_1 = 10$, $\bar{x}_1 = 87.6$, $\sigma_1 = 1$, $n_2 = 12$, $\bar{x}_2 = 74.5$, and $\sigma_2 = 1.5$. If μ_1 and μ_2 denote the true mean tensile strengths for the two grades of spars, we may find a 90% confidence interval on the difference in mean strength $\mu_1 - \mu_2$ as follows:

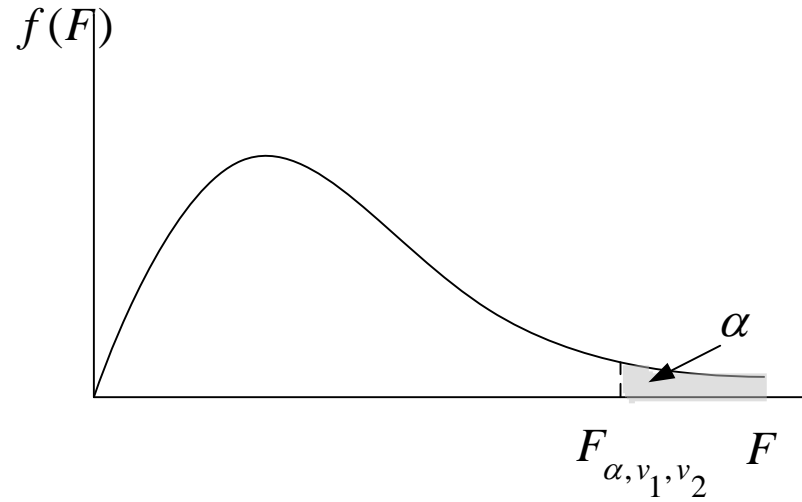
$$\bar{x}_1 - \bar{x}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$87.6 - 74.5 - 1.645 \sqrt{\frac{(1)^2}{10} + \frac{(1.5)^2}{12}} \leq \mu_1 - \mu_2 \leq 87.6 - 74.5 + 1.645 \sqrt{\frac{(1)^2}{10} + \frac{(1.5)^2}{12}}$$

$$12.22 \leq \mu_1 - \mu_2 \leq 13.98$$

Confidence Interval for the Ratio of Two Variance

$$\frac{s_1^2}{s_2^2} / \frac{\sigma_1^2}{\sigma_2^2} \sim F_{n_1-1, n_2-1}$$



$$F_{1-\alpha/2, n_1-1, n_2-1} \leq \frac{s_1^2}{s_2^2} / \frac{\sigma_1^2}{\sigma_2^2} \leq F_{\alpha/2, n_1-1, n_2-1}$$

$$\frac{s_1^2}{s_2^2} \left(\frac{1}{F_{\alpha/2, n_1-1, n_2-1}} \right) \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2}{s_2^2} \left(\frac{1}{F_{1-\alpha/2, n_1-1, n_2-1}} \right)$$

Confidence Interval for the Difference Between Two Binomial Proportions

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}_1(1 - \hat{p}_1)/n_1 + \hat{p}_2(1 - \hat{p}_2)/n_2}} \sim N(0,1)$$

$$-z_{\alpha/2} \leq \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}_1(1 - \hat{p}_1)/n_1 + \hat{p}_2(1 - \hat{p}_2)/n_2}} \leq z_{\alpha/2}$$

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \leq (p_1 - p_2)$$

$$\leq (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

MATCHED pairs

- In a matched pairs study, subjects are matched in pairs and the outcomes are compared within each matched pair.
 - Example: before and after studies

Modern Language Association listening scores for French teachers							
Teacher	Pretest	Posttest	Gain	Teacher	Pretest	Posttest	Gain
1	32	34	2	11	30	36	6
2	31	31	0	12	20	26	6
3	29	35	6	13	24	27	3
4	10	16	6	14	24	24	0
5	30	33	3	15	31	32	1
6	33	36	3	16	30	31	1
7	22	24	2	17	15	15	0
8	25	28	3	18	32	34	2
9	32	26	-6	19	23	26	3
10	20	26	6	20	23	26	3

***MATCHED* pairs**

Large Sample

$$\bar{d} \pm z_{\alpha/2} \left(\frac{\sigma_d}{\sqrt{n}} \right)$$

Small Sample

$$\bar{d} \pm t_{\alpha/2} \left(\frac{s_d}{\sqrt{n}} \right)$$

Example

Time in Seconds to Parallel Park Two
Automobiles

Subject	Automobile		Difference
	$1(x_{1j})$	$2(x_{2j})$	(d_j)
1	37.0	17.8	19.2
2	25.8	20.2	5.6
3	16.2	16.8	-0.6
4	24.2	41.4	-17.2
5	22.0	21.4	0.6
6	33.4	38.4	-5.0
7	23.8	16.8	7.0
8	58.2	32.2	26.0
9	33.6	27.8	5.8
10	24.4	23.2	1.2
11	23.4	29.6	-6.2
12	21.2	20.6	0.6
13	36.2	32.2	4.0
14	29.8	53.8	-24.0

Example Cont.

$$\bar{d} = 1.21 \qquad s_D = 12.68.$$

90% confidence interval

$$\bar{d} - t_{0.05,13}s_D/\sqrt{n} \leq \mu_D \leq \bar{d} + t_{0.05,13}s_D/\sqrt{n}$$

$$\begin{aligned} 1.21 - 1.771(12.68)/\sqrt{14} &\leq \mu_D \leq 1.21 + 1.771(12.68)/\sqrt{14} \\ -4.79 &\leq \mu_D \leq 7.21 \end{aligned}$$

Selecting the Sample Size...

- We can control the width of the interval by determining the sample size necessary to produce narrow intervals.
- Suppose we want to estimate the mean demand “to within 5 units”; i.e. we want the interval estimate to be: $\bar{x} \pm 5$

- Since:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- It follows that

$$z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 5$$

Solve for **n** to get requisite sample size!

Selecting the Sample Size...

- Solving the equation...

$$n = \left(\frac{z_{\alpha/2} \sigma}{5} \right)^2 = \left(\frac{(1.96)(75)}{5} \right)^2 = 865$$

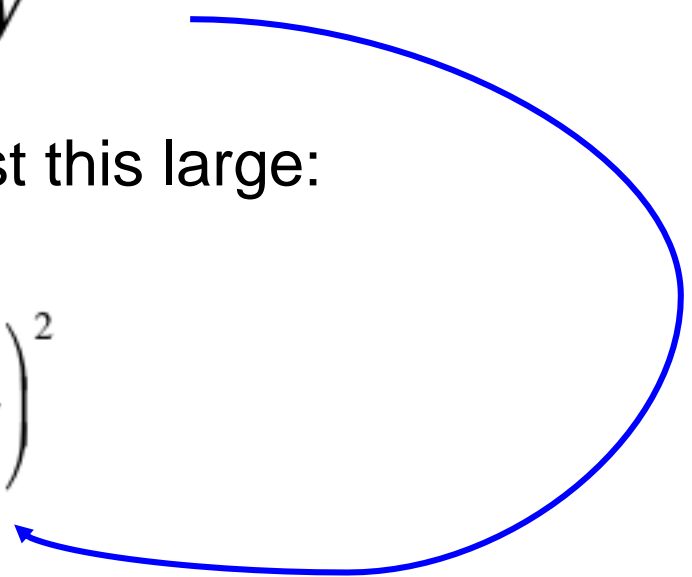
- that is, to produce a 95% confidence interval estimate of the mean (± 5 units), we need to sample 865 lead time periods (vs. the 25 data points we have currently).

Sample Size to Estimate a Mean...

- The general formula for the sample size needed to estimate a population mean with an interval estimate of:

$$\bar{x} \pm W$$

- Requires a sample size of at least this large:

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


Example 10.2...

- A lumber company must estimate the mean diameter of trees to determine whether or not there is sufficient lumber to harvest an area of forest. They need to estimate this to within 1 inch at a confidence level of 99%. The tree diameters are normally distributed with a standard deviation of 6 inches.
- How many trees need to be sampled?

- Things we know:
- Confidence level = 99%, therefore $\alpha = .01$

$1 - \alpha$	α	$\alpha / 2$	$z_{\alpha/2}$
.90	.10	.05	$z_{.05} = 1.645$
.95	.05	.025	$z_{.025} = 1.96$
.98	.02	.01	$z_{.01} = 2.33$
.99	.01	.005	$z_{.005} = 2.575$

- We want $\bar{x} \pm 1$ hence $W=1$. $z_{\alpha/2} = z_{.005} = 2.575$
- We are given that $\sigma = 6$.