Confidence Intervals (Two Samples)

8.2.3-8.2.4 8.2.5-8.2.6, 8.4.2

Learning Objectives

At the end of this unit, students should be able to:

1. Describe situations in which one might want to compare two population means.
2. State the assumptions necessary for the comparison of two means confidence interval.
3. Compute the confidence interval for the difference between two means for different applications.
4. Adapt the confidence interval for the difference between two means when the population standard deviations are unknown.
5. Describe situations in which one might want to compare two population proportions.
6. Compute the confidence interval for the difference between two proportions for different applications.

Comparing Two Population Means

Often, we are interested in comparing the means of two populations. For example:

1.) One population receives a drug for a condition, and another pop. receives a sugar pill.

2.) One pop. receives 87 octane and another 85.

Comparing Two Population Means

How do two populations compare, in terms of their means?

Treatment Group

Pop #1

μ_1

σ_1^2

Pop #2

μ_2

σ_2^2

To try to answer this question, we collect samples from both populations and perform inference on both samples to draw conclusions about $\mu_1 - \mu_2$.

Comparing Two Population Means

Basic Assumptions:

1.) $X_1, \ldots, X_{n_1}$ is a random sample (iid) from pop. #1, w/ mean μ_1 and var. σ_1^2.

2.) $Y_1, \ldots, Y_{n_2}$ is a random sample (iid) from pop #2, w/ mean μ_2 and var. σ_2^2.

3.) X_i and Y_i are independent.

Note: We haven't made any distributional assumptions, for now.

Comparing Two Population Means

The natural estimator of $\mu_1 - \mu_2$ is $\underline{\quad \bar{X} - \bar{Y} \qquad\qquad}$.

Inferential procedures are based on standardizing estimators, so we'll need the mean and standard deviation of $\underline{\quad \bar{X} - \bar{Y} \qquad}$.

Comparing Two Population Means

Mean of $\boxed{\bar{X} - \bar{Y}}$:

- $E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_1 - \mu_2$

Standard deviation of $\bar{X} - \bar{Y}$:

$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) - \underbrace{2Cov(\bar{X}, \bar{Y})}_{=0}$$

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \Rightarrow sd(\bar{X} - \bar{Y}) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Comparing Two Population Means

<u>Normal Populations with known variances</u>:

If both populations are normal, both ___ $\bar{X}$ ___ and ___ $\bar{Y}$ ___ have normal distributions. Further if the samples are independent, then the sample means are independent of one another.

Thus, ___ $\bar{X} - \bar{Y}$ ___ is normally distributed with expected value ___ $\mu_1 - \mu_2$ ___ and standard deviation:

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$\bar{X} \pm Z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \Longleftrightarrow \boxed{\text{pt est} \pm (\text{z-score})(\text{st. error})}$

Comparing Two Population Means

Standardizing our estimator gives:

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

Therefore, the $(1 - \alpha) \times 100\%$ confidence interval is:

$$\boxed{(\bar{X} - \bar{Y}) \pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Comparing Two Population Means: Large Sample

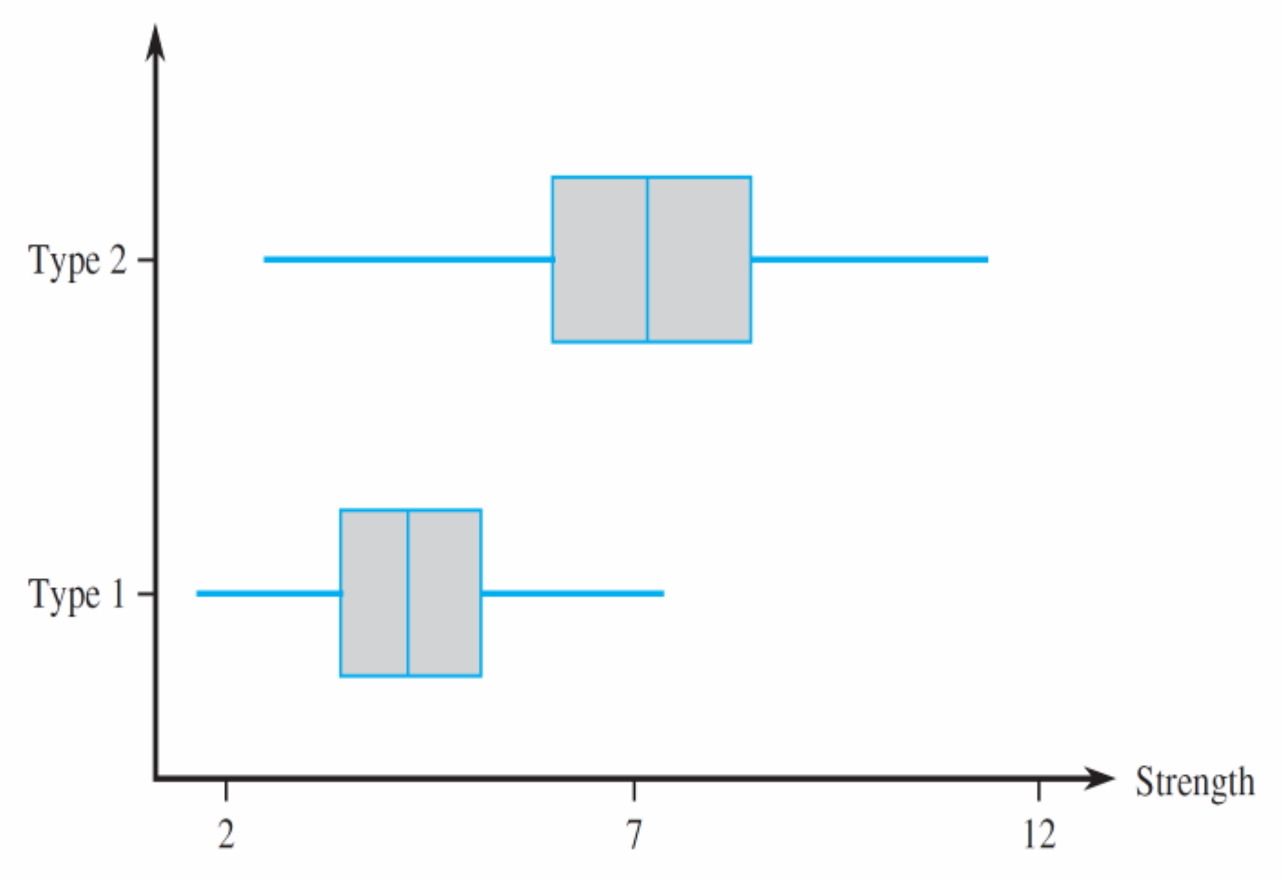
If both ~~m~~ n_1 and n_2 are large then the CLT implies that (≥ 30) our confidence interval is valid even without the assumption of normal populations. In this case, the confidence level is *approximately* $(1-\alpha)\times 100\%$.

Further, we can replace sample standard deviations for population standard deviations:

$$(\bar{X}-\bar{Y}) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

Comparing Two Population Means: Large Sample

An experiment carried out to study various characteristics of anchor bolts resulted in 78 observations on shear strength (kip) of 3/8-in. diameter bolts and 88 observations on the strength of 1/2-in. diameter bolts.



Comparing Two Population Means: Large Sample

The sample sizes, sample means, and sample standard deviations agree with values given in the article “Ultimate Load Capacities of Expansion Anchor Bolts” (*J. of Energy Engr.*, 1993: 139–158).

Variable	N $=n_1$	Mean $=\bar{x}$	Median	TrMean	StDev	SEMean
diam 3/8	78	4.250	4.230	4.238	1.300	0.147

Variable	Min	Max	Q1	Q3
diam 3/8	1.634	7.327	3.389	5.075

Variable	N $=n_2$	Mean $=\bar{y}$	Median	TrMean	StDev	SEMean
diam 1/2	88	7.140	7.113	7.150	1.680	0.179

Variable	Min	Max	Q1	Q3
diam 1/2	2.450	11.343	5.965	8.447

The summaries suggest that the main difference between the two samples is in where they are centered.

CI for $\mu_1 - \mu_2$

Comparing Two Population Means: Large Sample

Calculate a confidence interval for the difference between true average shear strength for 3/8-in. bolts (μ_1) and true average shear strength for 1/2-in. bolts (μ_2) using a confidence level of $\alpha = 0.05 \Leftarrow$ 95%.

Variable	N $= n_1$	Mean $= \bar{X}$	Median	TrMean	StDev	SEMean
diam 3/8	78	4.250	4.230	4.238	1.300	0.147

Variable	Min	Max	Q1	Q3
diam 3/8	1.634	7.327	3.389	5.075

Variable	N $= n_2$	Mean $= \bar{Y}$	Median	TrMean	StDev	SEMean
diam 1/2	88	7.140	7.113	7.150	1.680	0.179

Variable	Min	Max	Q1	Q3
diam 1/2	2.450	11.343	5.965	8.447

$$(\bar{X} - \bar{Y}) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \Leftrightarrow -2.89 \pm 1.96 \sqrt{\frac{(1.3)^2}{78} + \frac{(1.68)^2}{88}}$$

$$\Leftrightarrow (-3.3, -2.4)$$

Comparing Two Population Means: Small Sample

For large samples, the CLT allows us to use these methods we have discussed even when the two populations of interest are not normal. Skip

In practice, it can happen that ***at least one*** sample size is small and the population variances have unknown values.

Without the CLT at our disposal, we proceed by making specific assumptions about the underlying population distributions.

Comparing Two Population Means: Small Sample

When the population distribution are both normal, the standardized variable:

Skip

has approximately a *t* distribution with df *v* estimated from the data by:

Comparing Two Population Means: Small Sample

The **two-sample *t* confidence interval for $_1 - _2$** with confidence level 100(1 –) % is then

Skip

Comparing Two Population Means: Small Sample

Consider the following data on two different types of plainweave fabric:

skip

Fabric Type	Sample Size	Sample Mean	Sample Standard Deviation
Cotton	10	51.71	.79
Triacetate	10	136.14	3.59

Assuming that the porosity distributions for both types of fabric are normal, let's calculate a confidence interval for the difference between true average porosity for the cotton fabric and that for the acetate fabric, using a 95% confidence level

Comparing Two Population Proportions

Now consider the comparison of two population proportions. Just as before, an individual or object is a success if some characteristic of interest is present ("graduated from college", a refrigerator "with an ice-maker", etc.).

Let:
p_1 = the true proportion of successes in population 1
p_2 = the true proportion of successes in population 2

$\hat{p}_1 = \frac{X}{n_1}$, X = # of successes in sample 1; $\hat{p}_2 = \frac{Y}{n_2}$, Y = # of successes in sample 2

Comparing Two Population Proportions

(assume indep)

Mean of $\widehat{p}_1 - \widehat{p}_2$:

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

Standard deviation of $\widehat{p}_1 - \widehat{p}_2$:

$Var(\hat{p}_1 - \hat{p}_2) = Var(\hat{p}_1) + Var(\hat{p}_2) = \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}$

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

$\boxed{\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)}$ est $\pm\, z_{\alpha/2}$ s.e.

Comparing Two Population Proportions

So, $(1-\alpha) \times 100\%$ the confidence interval for $p_1 - p_2$ is:

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

Note: $\hat{q}_i = (1-\hat{p}_i)$

This interval can safely be used as long as

$$n_1\widehat{p}_1,\ n_1\widehat{q}_1,\ n_2\widehat{p}_2, \text{ and } n_2\widehat{q}_2$$

are all at least 10.

p-hacking

Comparing Two Population Proportions

The authors of the article "Adjuvant Radiotherapy and Chemotherapy in Node- Positive Premenopausal Women with Breast Cancer" (*New Engl. J. of Med.,* 1997: 956–962) reported on the results of an experiment designed to compare treating cancer patients with chemotherapy only to treatment with a combination of chemotherapy and radiation.

$\hat{p}_1 = \frac{76}{154}$, $\hat{p}_2 = \frac{98}{164}$

Of the 154 individuals who received the chemotherapy-only treatment, 76 survived at least 15 years, whereas 98 of the164 patients who received the hybrid treatment survived at least that long. What is the 99% confidence interval for this difference in proportions?

$(-0.247, 0.039)$

(See unit 6 R code)

Comparing Two Population Proportions

On occasion an inference concerning $p_1 - p_2$ may have to be based on samples for which at least one sample size is small.

Appropriate methods for such situations are not as straightforward as those for large samples, and there is more controversy among statisticians as to recommended procedures.

One frequently used test, called the Fisher–Irwin test, is based on the hypergeometric distribution.

Your friendly neighborhood statistician can be consulted for more information.