Comparing Two Means

Our goal is to compare the mean responses to two treatments, or to compare two population means (we have two separate samples).

We assume both populations are normally distributed (or "nearly" normal).

We're typically interested in the difference between the mean of population 1 (μ_1) and the mean of population 2 (μ_2).

We may construct a CI for $\mu_1 - \mu_2$ or perform one of three types of hypothesis test:

H_0: $\mu_1 = \mu_2$	H_0: $\mu_1 = \mu_2$	H_0: $\mu_1 = \mu_2$
H_a: $\mu_1 \neq \mu_2$	H_a: $\mu_1 < \mu_2$	H_a: $\mu_1 > \mu_2$

Note: H_0 could be written H_0: $\mu_1 - \mu_2 = 0$.

The parameter of interest is $\mu_1 - \mu_2$

Notation:

$\bar{X}_1$ = mean of Sample 1

$\bar{X}_2$ = mean of Sample 2
σ_1 = standard deviation of Population 1
σ_2 = standard deviation of Population 2
s_1 = standard deviation of Sample 1

s_2 = standard deviation of Sample 2
n_1 = size of Sample 1
n_2 = size of Sample 2

The point estimate of $\mu_1 - \mu_2$ is $\bar{X}_1 - \bar{X}_2$

This statistic has standard error $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

but we use $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ **since σ_1, σ_2 unknown.**

Since the data are normal, we can use the t-procedures for inference.

<u>Case I</u>: Unequal population variances ($\sigma_1^2 \neq \sigma_2^2$)

In the case where the two populations have <u>different variances</u>, the t-procedures are <u>only approximate</u>.

Formula for $(1 - \alpha)100\%$ CI for $\mu_1 - \mu_2$ is:

$$\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where the d.f. = the smaller of $n_1 - 1$ and $n_2 - 1$.

$\Leftrightarrow H_0: \mu_1 - \mu_2 = 0$

To test H_0: $\mu_1 = \mu_2$, the test statistic is:

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$|t| > t_{\alpha/2}$

H_a	Rejection region	P-value
$\mu_1 \neq \mu_2$	$t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$	2*(tail area)
$\mu_1 < \mu_2$	$t < -t_{\alpha}$	left tail area
$\mu_1 > \mu_2$	$t > t_{\alpha}$	right tail area

where the d.f. = the smaller of $n_1 - 1$ and $n_2 - 1$.

Case II: Equal population variances ($\sigma_1^2 = \sigma_2^2$)

In the case where the two populations have equal variances, we can better estimate this population variance with the pooled sample variance:

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

Our t-procedures in this case are exact, not approximate.

Formula for $(1 - \alpha)100\%$ CI for $\mu_1 - \mu_2$ is:

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2} \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}$$

where the d.f. = $n_1 + n_2 - 2$. ← different than in Case I

To test H_0: $\mu_1 = \mu_2$, the test statistic is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}}$$

H_a	Rejection region	P-value
$\mu_1 \neq \mu_2$	$t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$	2*(tail area)
$\mu_1 < \mu_2$	$t < -t_{\alpha}$	left tail area
$\mu_1 > \mu_2$	$t > t_{\alpha}$	right tail area

where the d.f. = $n_1 + n_2 - 2$.

Example: What is the difference in mean DVD prices at Best Buy and Walmart?

Let μ_1 = mean DVD price at Best Buy and let μ_2 = mean DVD price at Walmart.

Find 99% CI for $\mu_1 - \mu_2$.

Randomly sample 28 DVDs from Best Buy:

$\bar{X}_1 = 17.93$, $s_1 = 10.22$, $s_1^2 = 104.45$, $n_1 = 28$.

Randomly sample 20 DVDs from Walmart:

$\bar{X}_2 = 25.70$, $s_2 = 11.35$, $s_2^2 = 128.82$, $n_2 = 20$.

or CI for $\frac{\sigma_1^2}{\sigma_2^2}$

Does $\sigma_1^2 = \sigma_2^2$? Could test this formally using an F-test (Sec. 9.5) or could simply compare spreads of box plots for samples 1 and 2.

When in doubt, assume $\sigma_1^2 \neq \sigma_2^2$. Let's assume $\sigma_1^2 \neq \sigma_2^2$ here.

99% CI for $\mu_1 - \mu_2$:

$$\bar{X}_1 - \bar{X}_2 = 17.93 - 25.70 = -7.77$$

$$1 - \alpha = .99 \Rightarrow \alpha = .01 \Rightarrow \alpha/2 = .005$$

$s_1^2 = 104.45$

$s_2^2 = 128.82$

$t_{.005}$ (19 d.f.) = 2.861 (t-table)

$$-7.77 \pm (2.861)\sqrt{\frac{104.45}{28} + \frac{128.82}{20}}$$

$\Rightarrow -7.77 \pm 9.12 \Rightarrow$ 99% CI for $\mu_1 - \mu_2$: $(-16.89, 1.35)$

Interpretation: We are 99% confident that Best Buy's mean DVD price is between \$16.89 lower and \$1.35 higher than Walmart's mean DVD price.

Test: $H_0: \mu_1 = \mu_2$ vs. $H_a: \mu_1 < \mu_2$ (at $\alpha = .10$)

Assuming Case II: $\sigma_1^2 \neq \sigma_2^2$,

Test statistic:

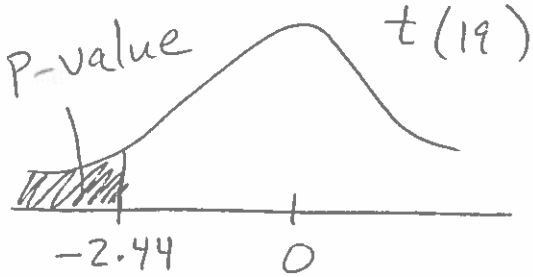
$$t = \frac{17.93 - 25.70}{\sqrt{\frac{104.45}{28} + \frac{128.82}{20}}} = \frac{-7.77}{3.19} = \boxed{-2.44}$$

Reject H_0 if $t < -t_{.10} = -1.328$ (t-table)
(19 d.f.)

Since $-2.44 < -1.328$, we reject H_0 (at $\alpha = .10$) and conclude Best Buy's population mean price is lower than Walmart's population mean price.

P-value: Area to left of -2.44 in t-distn with 19 d.f.

(P-value between .025 and .01) $\approx .013$



If we had assumed $\sigma_1^2 = \sigma_2^2$ here:

$$s_p^2 = \frac{(28-1)104.45 + (20-1)128.82}{28+20-2}$$

$$= \frac{5267.73}{46} = 114.516$$

99% CI for $\mu_1 - \mu_2$:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{.005}(46\ df)\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

$$-7.77 \pm 2.69\sqrt{\frac{114.516}{28} + \frac{114.516}{20}}$$

$$\Rightarrow -7.77 \pm 8.43 \Rightarrow (-16.2, 0.66)$$

Inference about Two Proportions (Sec. 9.4)

We now consider inference about $p_1 - p_2$, the difference between two population proportions.

Point estimate for $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2$

For large samples, this statistic has an approximately normal distribution with mean $p_1 - p_2$ and standard deviation $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$.

So a $(1-\alpha)100\%$ CI for $p_1 - p_2$ is

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

$\hat{p}_1$ = sample proportion for Sample 1

$\hat{p}_2$ = sample proportion for Sample 2

n_1 = sample size of Sample 1

n_2 = sample size of Sample 2

Requires large samples:

(1) Need $n_1 \geq 20$ and $n_2 \geq 20$.

(2) Need number of "successes" and number of "failures" to be 5 or more in both samples.

$\Leftrightarrow$ (2) Need: $n_1 \hat{p}_1 \geq 5$, $n_1(1-\hat{p}_1) \geq 5$,

$n_2 \hat{p}_2 \geq 5$, $n_2(1-\hat{p}_2) \geq 5$