Section 11: Inferences Based on Two Samples

Comparison of Two Population Means: μ_1 and μ_2

The method of comparison depends on the design of the experiment

The samples will either be Independent or Dependent

Independent Samples – implies data values obtained from one sample are unrelated to the values from the other sample

Example

- Does a drug work? → Randomly assign patients in medical trials.
- Which teaching method is better? → Randomly assign students to teaching methods.

Dependent Samples (also called paired samples) – implies subjects are paired so that they are as much alike as possible within each pair

The purpose of pairing is to explain subject to subject variability. (In some studies we apply both treatments to the same subject)

Example

- Do Nike or Adidas shoes make you jump higher? → Each subject can jump in both shoes. Randomly assign which shoe they use first.
- Does this drug work in weight loss? → Pair subjects so they are as much alike as possible. Randomly assign one person from each pair to take the drug and the other to take placebo.

If subject to subject variability is large relative to the expected treatment differences then a dependent sample design should be considered. Remember that reducing the variability will reduce the error in the study.

Hypothesis Test and Confidence Interval for Dependent Samples

Do not focus too much on the notation, but it is important to have a basic understanding. Essentially, a subscript is added to the notation we are familiar with in order to identify the treatment group. Also, with dependent sample we will deal with the differences for each pair which is denoted by d.

$n = number\ of\ pairs\ (sample\ size)$

$x_{ij} = observation\ for\ treatment\ i\ in\ pair\ j$

$d_j = x_{1j} - x_{2j} = difference\ between\ treatment\ 1\ and\ treatment\ 2\ in\ pair\ j$

The table below identifies the proper notation and layout when dealing with dependent samples. At the bottom of the table, you will see how the descriptive statistics are denoted. You should pay special attention to the notation for the *d* values since this is the part of the data set we actually analyze.

Pair	Treatment 1	Treatment 2	Difference
1	x_{11}	x_{21}	$d_1 = x_{11} - x_{21}$
2	x_{12}	x_{22}	$d_2 = x_{12} - x_{22}$
3	x_{13}	x_{23}	$d_3 = x_{13} - x_{23}$
.	.	.	.
.	.	.	.
.	.	.	.
n	x_{1n}	x_{2n}	$d_n = x_{1n} - x_{2n}$
Sample Mean	$\bar{x}_1$	$\bar{x}_2$	$\bar{d} = \bar{x}_1 - \bar{x}_2$
Sample Variance	${s_1}^2$	${s_2}^2$	${s_d}^2$
Sample Standard Deviation	s_1	s_2	s_d

For a dependent sample design all analysis is based on the differences: $d_1, d_2, d_3, \cdots, d_n$. The differences are a sample from a distribution with mean $\mu_d = \mu_1 - \mu_2$ and unknown variance ${\sigma_d}^2$. The point Estimate for μ_d is $\bar{d}$ (just like $\bar{x}$ is the point estimate for μ). Because of this setup, we can conduct hypothesis tests and confidence intervals in the same way as was done with a single sample except now we are using *d* values instead of *x* values. Below are the needed formulas. If you compare these to what was done with a single sample, you should see that we are basically doing the same thing.

The test statistic for dependent samples is $t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$ with $n - 1$ degrees of freedom.

Therefore, to calculate a confidence interval we take $\bar{d} \pm E$ with $E = t\left(\frac{s_d}{\sqrt{n}}\right)$.

Another thing to keep in mind is the relationship between μ_1 and μ_2 when conducting inferential statistics on μ_d. Since $\mu_d = \mu_1 - \mu_2$ we know the following.

- If $\mu_d > 0$, then $\mu_1 > \mu_2$.
- If $\mu_d < 0$, then $\mu_1 < \mu_2$.
- If $\mu_d \neq 0$, then $\mu_1 \neq \mu_2$.

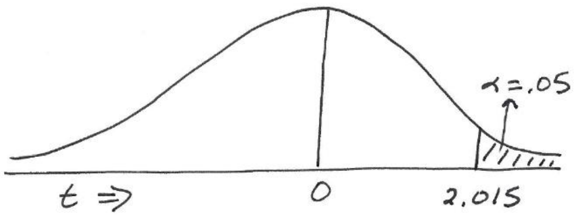
Example

A marketing expert wishes to prove that a new product display will increase sales over a traditional display. There are 12 stores available for the study. There is considerable variability from store to store so a dependent sample design will be used. The stores are divided into six pairs such that within each pair the stores are as alike as possible. The measurement will be the number of cases sold in a one month period. Using the data that follow, perform a hypothesis test using $\alpha = .05$.

	New Display	Traditional Display	
Pair	Treatment 1	Treatment 2	Difference
1	13	11	2
2	31	29	2
3	20	21	-1
4	19	17	2
5	42	39	3
6	26	22	4
Sample Mean			$\bar{d} = 2$
Sample Variance			${s_d}^2 = 2.8$
Sample Standard Deviation			$s_d = 1.6733$

1) $H_a\colon \mu_d > 0$ (Remember this is the same as testing $\mu_1 > \mu_2$)
   $H_0\colon \mu_d \le 0$

2) This is a *t* since we are doing a hypothesis test for μ_d with a dependent sample design. The critical region depends on H_a (since we are testing >, the critical region is to the right) and α (the area is α). Remember that degrees of freedom is $n - 1$.



Reject H_0 if $t \ge 2.015$

3) $t = \frac{\bar{d}-\mu_d}{{s_d}/{\sqrt{n}}} = \frac{2-0}{{1.6733}/{\sqrt{6}}} = 2.928$

4) Reject H_0 (since $2.928 \ge 2.015$, the test statistic is in the critical region)

5) Conclude with 95% confidence that the new method produces larger sales.

Example

Use the sales data presented in the previous example to construct and interpret a 95% confidence interval for μ_d.

Solution

$E = t\left(\frac{s_d}{\sqrt{n}}\right) = 2.571\left(\frac{1.6733}{\sqrt{6}}\right) = 1.756$
$\bar{d} \pm E$
2 ± 1.756
$(0.244, 3.756)$

We are 95% confident that the mean increase in sales is between 0.244 cases and 3.756 cases using the new product display.

Hypothesis Test and Confidence Interval for Independent Samples

With independent samples we cannot calculate *d* values since the samples are not paired. The process is still similar to the dependent case, but calculations are more cumbersome without the convenience of the *d* values. For independent samples we will estimate the parameter $\mu_1 - \mu_2$.

The table below identifies the proper notation for two independent samples.

	Treatment 1	Treatment 2
Sample Size	n_1	n_2
Data	$x_{11}, x_{12}, x_{13}, \cdots x_{1n_1}$	$x_{21}, x_{22}, x_{23}, \cdots x_{2n_2}$
Sample Mean	$\bar{x}_1$	$\bar{x}_2$
Sample Variance	${s_1}^2$	${s_2}^2$
Sample Standard Deviation	s_1	s_2

The quantity $\bar{x}_1 - \bar{x}_2$ is the point estimate for $\mu_1 - \mu_2$ with a standard error of $\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$.

Thus, if we sample from populations with means μ_1 and μ_2, and standard deviations of ${\sigma_1}^2$ and ${\sigma_2}^2$ respectively, then $Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$ is approximately standard normal for large n_1 and n_2.

As we should be accustomed to at this point, the issue with the above *Z* transformation is the fact we will not know ${\sigma_1}^2$ and ${\sigma_2}^2$ when estimating $\mu_1 - \mu_2$. Therefore, we must estimate ${\sigma_1}^2$ and ${\sigma_2}^2$ with the sample. The form of the test statistic depends on whether ${\sigma_1}^2 = {\sigma_2}^2$. For all of the problems we do in this class, we will assume ${\sigma_1}^2 = {\sigma_2}^2$.

If we sample from populations with means μ_1 and μ_2, standard deviations of ${\sigma_1}^2$ and ${\sigma_2}^2$ respectively, and ${\sigma_1}^2 = {\sigma_2}^2$, then t= $\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{{s_p}^2}{n_1}+\frac{{s_p}^2}{n_2}}}$ has a t-distribution with degrees of freedom $n_1 + n_2 - 2$.

The quantity ${s_p}^2 = \frac{(n_1-1){s_1}^2+(n_2-1){s_2}^2}{n_1+n_2-2}$ is the pooled variance which estimates ${\sigma_1}^2 = {\sigma_2}^2$.

Hence, when doing a hypothesis test for independent samples when ${\sigma_1}^2 = {\sigma_2}^2$ the test statistic is t= $\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{{s_p}^2}{n_1}+\frac{{s_p}^2}{n_2}}}$ with $n_1 + n_2 - 2$ degrees of freedom.

To calculate a confidence interval we take $(\bar{x}_1 - \bar{x}_2) \pm E$ with $E = t\sqrt{\frac{{s_p}^2}{n_1} + \frac{{s_p}^2}{n_2}}$.

Example

A study is designed to compare gas mileage with a fuel additive to gas mileage without the additive.

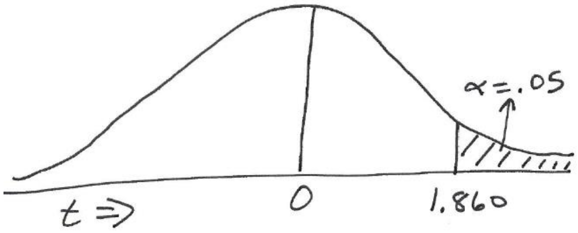
A group of 10 Ford Mustangs are randomly divided into two groups and the gas mileage is recorded for one tank of gas.

	With Additive Treatment 1	Without Additive Treatment 2
Sample Size	$n_1 = 5$	$n_2 = 5$
Data	26.3, 27.4, 25.1, 26.8, 27.1	24.5, 25.4, 23.7, 25.9, 25.7
Sample Mean	$\bar{x}_1 = 26.54$	$\bar{x}_2 = 25.04$
Sample Variance	${s_1}^2 = 0.813$	${s_2}^2 = 0.848$

Use the information provided to conduct a hypothesis to see if the additive increases gas mileage. Assume ${\sigma_1}^2 = {\sigma_2}^2$ and use a significance level of $\sigma = .05$.

1) $H_a: \mu_1 - \mu_1 > 0$ (Remember this is the same as testing $\mu_1 > \mu_2$)
$H_0: \mu_1 - \mu_2 \leq 0$

2) This is a *t* since we are doing a hypothesis test for $\mu_1 - \mu_2$ with an independent sample design. The critical region depends on H_a (since we are testing >, the critical region is to the right) and α (the area is α). Remember that *d.f.* is $n_1 + n_2 - 2$.



Reject H_0 if $t \geq 1.860$

3) $d.f. = n_1 + n_2 - 2 = 5 + 5 - 2 = 8$
$s_p{}^2 = \frac{(n_1-1)s_1{}^2+(n_2-1)s_2{}^2}{n_1+n_2-2} = \frac{(5-1)0.813+(5-1)0.848}{5+5-2} = 0.8305$
$\text{t}= \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_p{}^2}{n_1}+\frac{s_p{}^2}{n_2}}} = \frac{(26.54-25.04)-0}{\sqrt{\frac{.8305}{5}+\frac{.8305}{5}}} = 2.603$

4) Reject H_0 (since $2.603 \geq 1.860$, the test statistic is in the critical region)

5) Conclude with 95% confidence that the additive improves gas mileage.

<u>Example</u>
Use the gas mileage data presented in the previous example to construct and interpret a 95% confidence interval for $\mu_1 - \mu_2$.

<u>Solution</u>
$E = t\sqrt{\frac{s_p{}^2}{n_1} + \frac{s_p{}^2}{n_2}} = 2.306\sqrt{\frac{.8305}{5} + \frac{.8305}{5}} = 1.329$
$(\bar{x}_1 - \bar{x}_2) \pm E$
$(26.54 - 25.04) \pm 1.329$
1.50 ± 1.329
$(0.171, 2.829)$

We are 95% confident that the additive will increase gas mileage by an amount between 0.171 and 2.829 miles.

If ${\sigma_1}^2 \neq {\sigma_2}^2$ what should we do?

Without the assumption of equal variances, it makes no sense to compute ${s_p}^2$.

A reasonable transformation in this case is t$= \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}}}$.

Unfortunately, the distribution of this transformation is unknown.

For large n_1 and n_2, the distribution of *t* can be approximated by the standard normal distribution (just like we did with proportions).

For small n_1 and n_2, the distribution of *t* can be approximated by a t-distribution using a complicated formula for the degrees of freedom.

For this class, we will only do problems where we can assume ${\sigma_1}^2 = {\sigma_2}^2$. As mentioned above, when this assumption is not satisfied there are methods of conducting a hypothesis test and constructing a confidence interval. These methods are most commonly applied with the utilization of statistical software.