ILRST/STSCI 2100 Discussion 10: Chapter 11 Comparing Two Populations

Comparing Two Sample Means:
Assumes that both samples are randomly sampled, independent, and approximately normally distributed. We will be using the t-test Welch-Satterthwaite correction for calculating the degrees of freedom.

The null hypothesis is: $H_0 : \mu_1 - \mu_2 = 0$
The alternative hypothesis is $H_a : \mu_1 - \mu_2 > 0$, could also be < 0 or $\neq 0$.

T-statistic: $\dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

This follows approximately the t-distribution with $df = \dfrac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$

Then you can find the p-value and make a decision to the hypothesis test

Confidence interval: $(\bar{X}_1 - \bar{X}_2) \pm t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Paired-sample t-test
We can compute differences for each pair of observations, and then compute the mean and standard deviation using the differences. Note that when looking at these samples, these are **not** independent.
Ex.

- Pre-test and post-test scores after online tutorial
- Comparing fertilizer vs no fertilizer on grass growth on the same patch of grass

Assuming that the population is approximately normal, or if $n > 30$, we can calculate a test statistic that follows the t-distribution with $df = n - 1$. Note that n is the number of pairs.

The null hypothesis would be: $H_0 : \mu_D = 0$

And the alternative hypothesis could be: $H_a : \mu_D > 0$, could also be < 0 or $\neq 0$

Use the t-test, and calculate the t statistic:

$$t_{obs} = \frac{\bar{D} - \mu_D}{SD_D/\sqrt{n}}, \text{ where } SD_D = \sqrt{\tfrac{1}{n-1} \sum (D_i - \bar{D})^2}$$

$\dfrac{SD_D}{\sqrt{n}}$ is also called the standard error.

Confidence interval $= \bar{D} \pm t_{\alpha/2} \dfrac{SD_D}{\sqrt{n}}$

Comparing Two Population Proportions:

You must first check that the sample sizes for each population is sufficiently large by making sure $n_1 * p_1$, $n_1 * (1 - p_1)$,$n_2 * p_2$, and $n_2 * (1 - p_2)$ are all greater than 10.

The null hypothesis is: $H_0 : p_1 - p_2 = 0$

The alternative would be: $H_a : p_1 - p_2 > 0$, can also be < 0 or $\neq 0$

The test statistic would be:

$Z = \dfrac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p_c}(1 - \hat{p_c})(\frac{1}{n_1} + \frac{1}{n_2})}}$, where $\hat{p}_c = \frac{X_1 + X_2}{n_1 + n_2}$, where X_1 and X_2 are number of success in population 1 and 2, respectively.

Find p-value and make a decision.

Examples:

1. A researcher is interested in comparing the average fish length (cm) in two lakes in a fishing community. People tend to fish in Lake A, whereas very few people fish in Lake B. Run a hypothesis test comparing two sample means (assume population variances are not equal).

Sample	n	$\bar{x}$	s^2
Lake A	32	15	2
Lake B	34	20	4

2. (modified from: http://stattrek.com/hypothesis-test/paired-means.aspx?tutorial=ap)
   44 students are randomly selected from a school and divided into 22 matched pairs, each pair having an equal GPA. One student in each pair was given a special online tutorial on a concept in statistics. Then, all students were tested. Conduct a hypothesis test to determine if the online tutorial helped students better understand the concept. Use $\alpha = 0.05$ and assume mean differences are normally distributed. Test results are summarized below.

Pair	Training	No training	Difference, d	$(d - \bar{d})^2$
1	95	90	5	16
2	89	85	4	9
3	76	73	3	4
4	92	90	2	1
5	91	90	1	0
6	53	53	0	1
7	67	68	-1	4
8	88	90	-2	9
9	75	78	-3	16
10	85	89	-4	25
11	90	95	-5	36

Pair	Training	No training	Difference, d	$(d - \bar{d})^2$
12	85	83	2	1
13	87	83	4	9
14	85	83	2	1
15	85	82	3	4
16	68	65	3	4
17	81	79	2	1
18	84	83	1	0
19	71	60	11	100
20	46	47	-1	4
21	75	77	-2	9
22	80	83	-3	16

$$\Sigma(d - \bar{d})^2 = 270$$
$$\bar{d} = 1$$

3. Calculate a 95% confidence interval for the data in problem 2. Is your answer consistent with the decision in problem 2?

Assume that both samples are randomly sampled, independent and check samples sizes $(n_1 = 32, n_2 = 34)$ are greater than 30.

Step 1: Define H_0
$H_0 : \mu_1 - \mu_2 = 0$
(Be sure to use the population parameter μ in hypotheses)

Step 2: Define H_1
$H_1 : \mu_1 - \mu_2 \neq 0$
(This is two-sided test)

Step 3: Calculate test statistic

$$t_{obs} = \frac{(15-20)-0}{\sqrt{\frac{2}{32}+\frac{4}{34}}} = -11.7$$

Calculate d.f $df = \dfrac{\left(\frac{2}{32}+\frac{4}{34}\right)^2}{\frac{2^2}{32^2(32-1)}+\frac{4^2}{34^2(34-1)}} = 59.5$

Step 4: Find p-value
Since this is a two-sided test:

$$\begin{aligned} P - \text{value} &= 2 * P(T > |t_{obs}|) \\ &= 2 * P(T > |-11.7|) \\ &= \approx 0 \end{aligned}$$

Step 5: Decision
The observed t-statistic is in the rejection region and p-value $< \alpha$ so we reject the null hypothesis. We conclude that there is a significant difference between fish lengths in Lake A and Lake B.

Step 1: Define H_0
$H_0 : \mu_D = 0$

Step 2: Define H_1
$H_0 : \mu_D \neq 0$
(Two-sided test)

Step 3: Calculate test statistic
Standard deviation of differences: $SD_D = \sqrt{\frac{\sum(d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{270}{(22-1)}} = 3.586$

$$t_{obs} = \frac{1-0}{\frac{3.586}{\sqrt{22}}} = 1.307$$

Step 4: Find p-value

$$\text{p-value} = 2 * P(T > |t_{obs}|) = 2 * P(T > |1.307|)$$

$df = n - 1 = 22 - 1$ (n is the number of pairs) If you look in Table 4, you'll see that the central area is <80%, meaning that the area in the two tails is greater than 20%. Area in the two tails is the p-value, so the p-value is greater than 20%.

Step 5: Decision
At $\alpha = 0.05$, p-value $> \alpha$ so we fail to reject the null hypothesis. There is no significant difference in performance on a test between those who have completed an online tutorial and those who have not.

Confidence interval $= \bar{D} \pm t_{\alpha/2} \frac{SD_D}{\sqrt{n}}$

95% confidence interval means $\alpha = 0.05$ so $t_{\alpha/2} = 2.08$

From problem 2, $SD_D = 3.586$

So the confidence interval is: $1 \pm 2.08 * \frac{3.586}{\sqrt{21}} = (-6.28, 2.63)$

Since 0 is included in the confidence interval, we conclude that the mean difference between the two groups is zero. So our results match the hypothesis test.