Ch. 21 – Comparing Two Population Means

Notation: Two samples require appropriate subscripts → μ_1 and μ_2, n_1 and n_2

The Pooled t-Test

Assumptions:

1. The two samples are random and independent.
2. At least one of the following is also true:
   i. Both samples are large ($n_1 \geq 30$ and $n_2 \geq 30$)
   ii. If either one or both sample sizes are small, then both populations from which the samples are drawn are normally distributed.
3. The standard deviations σ_1 and σ_2 of the two populations are unknown and equal to each other; that is, $\sigma_1 = \sigma_2$.

Checking the Assumptions:

Assumption #1: Analyze the experimental design.

Assumption #2: Same method as Ch. 20.

Assumption #3: Officially requires a formal test that *can be* highly sensitive, however, side-by-side boxplots are helpful and an informal rule suggests that if the sample sizes are approximately equal and both sample sizes are large enough (≥15), check if

$$\frac{s_{\max}}{s_{\min}} < 2$$

Hypotheses:

Similar to Ch. 19, there are two population proportions (which are parameters) in our data structure, we consider them together as ONE parameter: $\mu_1 - \mu_2$. Thus, we have

H_0: $\mu_1 - \mu_2 = 0$ $\qquad$ H_A: $\mu_1 - \mu_2 \neq 0$

Note that we could use any value to compare to, but zero has a 'special' interpretation. Also, tests can be one-sided, too.

Test statistic:

If the assumptions hold, then we may use the *t*-distribution. Due to assuming equal variances, the standard error of $\bar{y}_1 - \bar{y}_2$ uses the *pooled sample standard deviation*, or s_p.

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

Thus, the standard error of $\bar{y}_1 - \bar{y}_2$ is

$$SE(\bar{y}_1 - \bar{y}_2) = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

The test statistic t_0 is

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)}{SE(\bar{y}_1 - \bar{y}_2)}$$

where t_0 follows a *t*-distribution with $df = n_1 + n_2 - 2$.

P-value: No different than how we calculated it in Ch. 20.

Conclusion: Reject/do not reject as in one-sample test; answer hypotheses/question posed.

Confidence Interval
The $(1 - \alpha)100\%$ CI for $\mu_1 - \mu_2$ is

$$\bar{y}_1 - \bar{y}_2 \pm t_{\alpha/2,\, df} \times SE(\bar{y}_1 - \bar{y}_2)$$

Assumptions: as per hypothesis test.

Notes: - CI tends to be more informative than a test.
- check if zero falls within the interval; check sign and magnitude.

<u>What if we can't assume the variances are equal?</u>
The pooled t-test works best when the sample sizes are small. Unfortunately, this is the hardest time to prove equal variances since s_1 and s_2 might under/overestimate σ_1 and σ_2 with such small sample sizes. The more general (and generally safer) method changes the 3rd assumption to the following.

3. The standard deviations σ_1 and σ_2 of the two populations are unknown.
Note: We don't assume $\sigma_1 \neq \sigma_2$; rather, we just don't assume $\sigma_1 = \sigma_2$.

If *any of the conditions* for the pooled variance procedure's 3rd assumption do NOT hold, then use this test where we do NOT assume equal variances.

Here, the standard error of $\bar{y}_1 - \bar{y}_2$ is

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The test statistic t_0 is still written as

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)}{SE(\bar{y}_1 - \bar{y}_2)}$$

where t_0 follows a t-distribution with a complicated df (see footnote on p. 657). **<u>For the exam</u>**, we will instead use a conservative lower bound: $df \geq \min\{n_1 - 1, n_2 - 1\}$.

P-value and conclusion processes do not change.

The $(1 - \alpha)100\%$ CI for $\mu_1 - \mu_2$ can still be written as

$$\bar{y}_1 - \bar{y}_2 \pm t_{\alpha/2,\, df} \times SE(\bar{y}_1 - \bar{y}_2)$$

where df is as above for the given confidence level.

Ch. 22 - Two Population Means for Paired Samples

Def'n: Two samples are said to be paired or matched samples when, for each value collected from one sample, there is a corresponding value collected from the second sample. In other words, these values are collected from the same source.

Notation: The value *d* denotes a paired difference.

The corresponding sample statistics are:

$$\bar{d} = \frac{\sum d_i}{n}$$

$$s_d^2 = \frac{1}{n-1}\left[\sum d_i^{\,2} - \frac{(\sum d_i)^2}{n}\right] \qquad \text{and} \qquad s_d = \sqrt{s_d^2}$$

Assumptions:

1. The samples are paired.
2. The n sample differences are viewed as a random sample from a pop'n of differences.
3. The sample size is large (generally ≥ 30) OR the population distribution is (approximately) normal.

Hypotheses:

Since we now have a "single sample" of differences, then we return to "ONE" parameter, but we need to define *d* first; it will be different for each situation.

H_0: $\mu_d = 0$ $\qquad$ H_A: $\mu_d \neq 0$

Again, we could use any value to compare to, but zero has a 'special' interpretation. Also, tests can still be one-sided.

Test statistic:

If the assumptions hold, then we may use the *t*-distribution. In fact, we return to one-sample inference, so $df = n - 1$ and our test statistic t is

$$t_0 = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$$

P-values and conclusions are found as before in this chapter.

Confidence Interval

The $(1 - \alpha)100\%$ CI for μ_d is

$$\bar{d} \pm t_{\alpha/2,\, n-1}\left(\frac{s_d}{\sqrt{n}}\right)$$

Assumptions: as per hypothesis test.

Ch. 21-22 Summary

- same pitfalls and subtleties that exist in Ch. 20 exist here, too.
- keep note of extensions to 2-sample data.
- not all assumptions can be checked graphically/statistically.