Ch. 24 – Comparing Two Population Means

Def'n: Two samples drawn from two populations are independent if the selection of one sample from one population does not affect the selection of the second sample from the second population. Otherwise, the samples are dependent.

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

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 unequal to each other; that is, $\sigma_1 \neq \sigma_2$.

Checking the Assumptions:

The first assumption can be "checked" by analyzing the experimental design. The 2nd assumption can be "checked" just like in Ch. 23. The third should require a formal test that is highly sensitive, but for now, check if $\boxed{\frac{s_{\max}}{s_{\min}} \geq 2}$.

Hypotheses:

Although there are two population means (a.k.a. parameters) in our data structure and 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.
Thus, the standard error of $\bar{y}_1 - \bar{y}_2$ is

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

and the test statistic t_0 is

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

and t follows a *t*-distribution with a complicated *df* (see footnote on p. 657). Thus, we will instead use a conservative lower bound: $df \geq \min\{n_1 - 1, n_2 - 1\}$.

P-value: No different than how we calculated it in Ch. 23.

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.

The Pooled t-Test

Recall the three assumptions from the previous test. The 3rd assumption now changes to

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 assumption reverses as well.)

The consequence of this change is that the standard error of $\bar{y}_1 - \bar{y}_2$ now 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 now

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

Also, the *t*-distribution now has a much simpler parameter: $df = n_1 + n_2 - 2$.

No other changes occur for the testing process. The test statistic t_0 is still written as

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

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. 25 - 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: $\mu_d = 0$ 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

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

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.