9.3 Analysis of Independent Samples

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.

Ex9.1) μ_1 and μ_2, n_1 and n_2, $\bar{x}_1$ and $\bar{x}_2$

(Textbook uses μ_A and μ_B, n and m, $\bar{x}$ and $\bar{y}$.)

Assumptions:

1. The two samples are random and independent.
2. Both populations are normal.
3. Both variances are known.

Although there are two population means (a.k.a. parameters) in our data structure, we consider them together as ONE parameter: $\mu_1 - \mu_2$. The likely point estimator for this *single* parameter is $\bar{X}_1 - \bar{X}_2$. Subsequently,

$$E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

$$V(\bar{X}_1 - \bar{X}_2) = V(\bar{X}_1) + V(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

And $Z = \dfrac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0, 1)$

9.3.3 z-Procedure

Always make sure assumptions are holding before making any inferences.

Hypothesis Testing:

Assume $\mu_1 - \mu_2 = \delta$ (some value, but zero is unique). Then, H_0: $\mu_1 - \mu_2 = \delta$ and

$$z_0 = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

so that we find the p-value from $N(0, 1)$. The H_A can be one-sided or two-sided. Moreover, instead of μ_0 (as in Ch. 8), there is now δ; if $\delta > 0$, then $\mu_1 > \mu_2$.

Confidence Interval:

If $\bar{x}_1$ and $\bar{x}_2$ are the means of independent random samples of sizes n_1 and n_2 from two independent normal populations with known variances σ_1^2 and σ_2^2, respectively, then a $100(1 - \alpha)\%$ CI for $\mu_1 - \mu_2$

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Note: One-sided confidence "intervals" use a similar approach as above, but with z_α.

9.3.1 General Procedure

Hypotheses: No different than how we construct them in 9.3.3.

Assumptions:

1. The two samples are random and independent.
2. σ_1 and σ_2 of the two populations are unknown and unequal; that is, $\sigma_1 \neq \sigma_2$.
3. 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.

Checking the Assumptions:

The last assumption can be "checked" just like in Ch. 8. The first assumption can be "checked" by analyzing the experimental design. The second, however, can use "math".

→ "rule of thumb" about Assumption #2: "okay" if ratio of $s_{\max}/s_{\min} > 2$.

Test statistic:

Due to unknown population variances, the standard error of $\bar{x}_1 - \bar{x}_2$ is now

$$S.E.(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and the test statistic is

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{S.E.(\bar{x}_1 - \bar{x}_2)}$$

which has an approximate *t*-distribution with

$$v = \frac{(V_1 + V_2)^2}{\frac{V_1^2}{n_1 - 1} + \frac{V_2^2}{n_2 - 1}} \qquad \text{where } V_1 = \frac{s_1^2}{n_1} \text{ and } V_2 = \frac{s_2^2}{n_2}$$

Truncate the number of v (round down) to an integer value. In some cases, a conservative lower bound would suffice: $v \geq \min\{n_1 - 1, n_2 - 1\}$.

p-value: No different than how we calculated it in Ch. 8.

Conclusion: Reject/do not reject as per Ch. 8; answer hypotheses/question posed.

Confidence Interval

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

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2,v} \times S.E.(\bar{x}_1 - \bar{x}_2)$$

where the critical value uses v as above for the given confidence level.

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

9.3.2 Pooled Variance Procedure

Here, we assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Thus, $V(\overline{X}_1 - \overline{X}_2) = V(\overline{X}_1) + V(\overline{X}_2) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

Hypotheses: Again, still constructed the same as in 9.3.3.

From assumptions listed under 9.3.1, the 2nd assumption changes to

2. The standard deviations σ_1 and σ_2 of the two populations are unknown but assumed to be equal; that is, $\sigma_1 = \sigma_2$. (We now check the assumption for a ratio < 2.)

Test statistic:

The 2nd assumption's change allows the use of the *pooled variance estimate* of σ, 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 $\overline{x}_1 - \overline{x}_2$ is

$$S.E.(\overline{x}_1 - \overline{x}_2) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Tests can still be two- or one-sided and the test statistic "remains"

$$t_0 = \frac{\overline{x}_1 - \overline{x}_2 - \delta}{S.E.(\overline{x}_1 - \overline{x}_2)}$$

where the critical value uses $v = n_1 + n_2 - 2$.

p-value: No different than how we calculated it in Ch. 8.

Conclusion: Reject/do not reject as per Ch. 8; answer hypotheses/question posed.

Confidence Interval

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

$$\overline{x}_1 - \overline{x}_2 \pm t_{\alpha/2,v} \times S.E.(\overline{x}_1 - \overline{x}_2)$$

where the critical value uses v as above for the given confidence level.

9.2 Analysis for Paired Samples

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

Notation: A paired difference is $z_i = d_i = x_{1i} - x_{2i}$, $i = 1, 2, \ldots, n$.

The d_i's are assumed to be normally distributed with

$$\mu_d = E(d_i) = E(X_{1i} - X_{2i}) = E(X_{1i}) - E(X_{2i}) = \mu_1 - \mu_2$$

and variance σ_d^2, so any test is reduced to a one-sample t-test on μ_d.

The corresponding sample statistics are:

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

Hypotheses:
Since we now have a "single sample" of differences, then there's only ONE parameter, but we need to define d first since it will be different for each situation.

H_0: $\mu_d = \mu_0$ $\qquad$ H_A: $\mu_d \neq \mu_0$

Again, zero is unique and tests can still be one-sided.

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.

Test statistic:
If the assumptions hold, then we may use the t-distribution. In fact, we return to one-sample inference, so $v = n - 1$ and our test statistic t_0 is

$$t_0 = \frac{\bar{d} - \mu_0}{s_d / \sqrt{n}}$$

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

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

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

Ch. 9 Summary
- same pitfalls and subtleties that exist in Ch. 8 exist here, too.
- keep note of extensions to 2-sample data.
- not all assumptions can be checked graphically/statistically.