Chapter 8 : Inferences for Two Samples

In previous chapters, we had only one sample and we wanted to see whether its mean or variance might be above or below a certain value. In Chapter 8 we compare statistics from 2 populations, and we want to know whether one mean is larger than another, whether the means are different, etc. The techniques of this chapter are very useful for comparative studies.

Independent Samples Design:

There are a few different ways we can do an experiment. In an *independent samples design*, we have an independent sample from each population. The data from the two groups are independent.

$$\begin{aligned} \text{Sample 1:} \quad & x_1, \ldots x_{n_1} \\ \text{Sample 2:} \quad & y_1, \ldots y_{n_2} \end{aligned}$$

Here n_1 does not need to equal n_2, that is, the samples can be different sizes. The x_i's and y_i's are all statistically independent. The difference is that the y_i's receive the treatment and the x_i's do not. For example, the x_i's and y_i's are student grades. The first group is the control group, and the second group was taught by a different method.

Note: How might you compare 2 independent samples graphically? Let's say you wanted to find out if one sample had generally larger values than the other for instance?

Matched Pairs Design:

In the *matched pairs design*, the observations from each sample are paired. An example is that the x_i's are the student scores before a training program, and the y_i's are the scores of the same students after the training program.

$$\begin{aligned} \text{pair:} \quad & 1, \quad 2, \ldots, n \\ \text{Sample 1:} \quad & x_1, \quad x_2, \ldots, x_n \\ \text{Sample 2:} \quad & y_1, \quad y_2, \ldots, y_n \end{aligned}$$

Here the i^{th} observation in the first group is similar in some way to the i^{th} observation in the second group. The way in which they are similar is called

the *blocking factor*.

Note: How might you consider matched pairs graphically?

8.3 Comparing means of 2 populations

We will test whether the mean of one of the populations is different than the other by a difference of δ_0.

1) Independent Samples Design for Large Samples ($n_1, n_2 > 30$):

Sample 1: $x_1, \ldots, x_{n_1}$ from a population with unknown μ_1 and σ_1^2 with sample mean $\bar{x}$ and sample variance s_1^2.
Sample 2: $y_1, \ldots, y_{n_2}$ from a population with unknown μ_2 and σ_2^2 with sample mean $\bar{y}$ and sample variance s_2^2.

We are testing:

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = \delta_0 \\ H_1 &: \mu_1 - \mu_2 \neq \delta_0 \end{aligned}$$

So, if we want to test whether or not the means are different, we set $\delta_0 = 0$.

The main idea is that since n_1 and n_2 are both large, we are going to use the central limit theorem to say that the distribution of $\bar{X} - \bar{Y}$ is approximately normal. We can calculate:

$$\begin{aligned} \mathbf{E}(\bar{X} - \bar{Y}) &= \mu_1 - \mu_2 \\ \mathrm{Var}(\bar{X} - \bar{Y}) &= \mathrm{Var}(\bar{X}) + \mathrm{Var}(-\bar{Y}) + 2\mathrm{Cov}(\bar{X}, -\bar{Y}) \\ &= \mathrm{Var}(\bar{X}) + (-1)^2\mathrm{Var}(\bar{Y}) + 0 \\ &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \end{aligned}$$

From the CLT we know that Z is approximately $N(0,1)$ where:

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}.$$

This means we can do an α-level z-test for comparing the means using the test statistic:

$$z = \frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}},$$

where remember that for large samples $s_1 \approx \sigma_1$ and $s_2 \approx \sigma_2$.

To summarize, the statistic above is for conducting a 2-sample independent samples design z-test where both samples are large, and the goal is to compare the means. That's the basic idea, and the front page of your book has the summary written out for you to carry out the test. Basically if z is too big or too small, you'll reject H_0.

Example 1

2) Independent Samples Design for Small Samples ($n_1, n_2 \leq 30$):

We could create a z-test using rv Z if the populations are normal and if we know the population variances. But in most cases, we don't know this. In that case, we can't use the z-test since we have no variances and we also can't use the CLT to claim that Z defined above is approximately $N(0, 1)$. We'll have to assume the populations are normal and use the t-test.

Sample 1: $x_1, \ldots, x_{n_1} \sim N(\mu_1, \sigma_1^2)$ where μ_1 and σ_1^2 are unknown.
Sample 2: $y_1, \ldots, y_{n_2} \sim N(\mu_2, \sigma_2^2)$ where μ_2 and σ_2^2 are unknown.

Case 2a: $\sigma_1^2 = \sigma_2^2$. (You have to know this somehow ahead of time to use this test. Or you can check this assumption using an F-test that I'll show you in Chapter 8.4. It is also assumed that you don't necessarily know σ_1 or σ_2 in advance.)
Let $\sigma^2 := \sigma_1^2 = \sigma_2^2$. We are testing:

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = \delta_0 \\ H_1 &: \mu_1 - \mu_2 = \delta_0 \end{aligned}$$

I need to do some calculations to derive the test statistic. We need to know that:

$$\mathbf{E}(\bar{X} - \bar{Y}) = \mu_1 - \mu_2.$$

It's going to be a kind of t-test, so I'll need an estimator for the variance. To estimate the variance we could use either s_1^2 or s_2^2 but instead we use a combination so we get a better estimator:

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2 + \sum_i (y_i - \bar{y})^2}{(n_1 - 1) + (n_2 - 1)} = \text{“pooled variance.”}$$

It turns out (with some work required) that the following rv T has a t-distribution with d.f. $(n_1 - 1) + (n_2 - 1)$ which equals $n_1 + n_2 - 2$:

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_0)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

We can use the test statistic:

$$t = \frac{\bar{x} - \bar{y} - \delta_0}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

(where s is the square root of the pooled variance above) for a 2 sample t-test to compare the means for an independent samples design experiment where the samples have equal variance, d.f. $n_1 + n_2 - 2$.

Example 2

Case 2b: $\sigma_1^2 = \sigma_2^2$ and you don't know either of them. In this case, it is tempting to use the distribution of:

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

but T does not have a t-distribution. However, its distribution can be approximated by the t-distribution with d.f. ν where ν is computed according to the “Welch-Satterthwaite method:

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}},$$

where fractions are truncated to the nearest integer.
So to test:

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = \delta_0 \\ H_1 &: \mu_1 - \mu_2 = \delta_0, \end{aligned}$$

compute test statistic:

$$t = \frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

and comparing to $t_{\nu,\alpha/2}$ (2-sided) or $t_{\nu,\alpha}$ (1-sided) gives the approximate solution, using ν computed according to the Welch-S method. We can certainly also compute pvalues and confidence intervals, which are provided in the table in the front of the book.

The Welch-S method really makes a difference when:

1. s_1 and s_2 are very different
2. n_1 and n_2 are very different.

3) Matched Pairs Design:

Given n pairs:

$$\begin{aligned} \text{pair:} \quad & 1, \quad 2, \ldots, n \\ \text{Sample 1:} \quad & x_1, \quad x_2, \ldots, x_n \\ \text{Sample 2:} \quad & y_1, \quad y_2, \ldots, y_n \end{aligned}$$

Assume $X_i \sim N(\mu_1, \sigma_1^2)$ and $Y_i \sim N(\mu_2, \sigma_2^2)$ but X_i and Y_i are not independent, they are correlated. The pairs themselves are mutually independent (e.g. patients' temp before taking tylenol, patients' temp after tylenol). Let $\rho := corr(X_i, Y_i)$ (it's the same for all i).

Define $D_i = X_i - Y_i$. It turns out that the D_i's are independent normal rv's with:

$$\begin{aligned} \mu_D &= \mathbf{E}(D_i) = \mathbf{E}(X_i - Y_i) \\ {\sigma_D}^2 &= \mathrm{Var}(D_i) = \mathrm{Var}(X_i - Y_i) = \mathrm{Var}(X_i) + \mathrm{Var}(-Y_i) - 2\mathrm{Cov}(X_i, Y_i) \\ &= \sigma_1^2 + (-1)^2\sigma_2^2 - 2\rho\sigma_1\sigma_2. \end{aligned}$$

Since $\rho > 0$ when the pairs are matched, the variance we computed is smaller than that of the independent samples case.

Now we can actually reduce the whole thing to the single sample setting. Let $d_i = x_i - y_i$. To test:

$$\begin{aligned} H_0 &: \mu_D = \delta_0 \\ H_1 &: \mu_D = \delta_0, \end{aligned}$$

We assume $D_1, \ldots, D_n \sim N(\mu_D, \sigma_D^2)$. We calculated $\bar{d} = \frac{1}{n}\sum_i d_i$ and we also calculated $s_d = \sqrt{\frac{1}{n-1}\sum_i (d_i - \bar{d})^2}$. The test statistic is just:

$$t = \frac{\bar{x} - \bar{y} - \delta_0}{s_d/\sqrt{n}},$$

and we perform a t-test (this is called a "paired" t-test).

- Reject H_0 when $|t| > t_{n-1,\alpha_2}$
- Reject H_0 when pvalue$= 2P(T_{n-1} \geq |t|) < \alpha$
- Reject H_0 when $\delta \notin$ CI, where CI is $\delta \in \left[\bar{d} - t_{n-1,\alpha/2}\frac{s_d}{\sqrt{n}}, \bar{d} + t_{n-1,\alpha/2}\frac{s_d}{\sqrt{n}}\right]$.

We can adapt the power and sample size determinations from Chapter 7 even though the variables aren't normal to get approximate values. Pinning down $H_1 : \mu_D = \delta_1$,

$$\pi(\delta_1) \approx \Phi\left(-z_{\alpha/2} + \frac{\delta_1}{\sigma_D/\sqrt{n}}\right) + \Phi\left(-z_{\alpha/2} - \frac{\delta_1}{\sigma_D/\sqrt{n}}\right).$$

The following sample size calculation gives the sample size needed for a matched pairs test with α-risk α and power $1 - \beta$ to detect a difference in means of δ_1:

$$n = \left[\frac{(z_{\alpha/2} + z_\beta)\sigma_D}{\delta_1}\right]^2$$

(you can replace σ_D by s_d if the sample size is large enough in both of these formulas).

8.4 Comparing the Variances of 2 Populations:

The F-test for independent samples design (heavily requires normality) compares the variance of two populations where we have samples:

Sample 1: $x_1, \ldots, x_{n_1} \sim N(\mu_1, \sigma_1^2)$
Sample 2: $y_1, \ldots, y_{n_2} \sim N(\mu_2, \sigma_2^2)$

To compare the population variances, we consider σ_1^2/σ_2^2, estimated by s_1^2/s_2^2. We learned in Chapter 5 that the rv below has an F-distribution with d.f.'s $n_1 - 1$ and $n_2 - 1$:

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}.$$

So if we want to test:

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 \\ H_1 &: \sigma_1^2 = \sigma_2^2 \end{aligned}$$

we compute the test statistic

$$F = \frac{s_1^2}{s_2^2}$$

and since the upper and lower $\alpha/2$ critical points of the F-distribution are $f_{n_1-1,n_2-1,1-\alpha/2}$ and $f_{n_1-1,n_2-1,\alpha/2}$, then we:

- reject H_0 when $F < f_{n_1-1,n_2-1,1-\alpha/2}$ or $F > f_{n_1-1,n_2-1,\alpha/2}$.
- reject H_0 when $P(F < f_{n_1-1,n_2-1,1-\alpha/2}) < \alpha/2$ or $P(F > f_{n_1-1,n_2-1,\alpha/2}) > \alpha/2$.
- reject H_0 when $F \notin CI$.

  Let's derive the CI:

  $$f_{n_1-1,n_2-1,1-\alpha/2} \leq \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \leq f_{n_1-1,n_2-1,\alpha/2}.$$

  We need to solve for σ_1^2/σ_2^2. Just rewriting to make the notation simpler,

  $$f_- \leq \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \leq f_+.$$

Let's do the left equation first. Solving for the ratio $\frac{\sigma_1^2}{\sigma_2^2}$,

$$f_- \leq \frac{s_1^2}{s_2^2}\frac{\sigma_2^2}{\sigma_1^2}$$

$$\frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2}{s_2^2}\frac{1}{f_-}.$$

Then for the right equation, we'll have:

$$\frac{\sigma_1^2}{\sigma_2^2} \geq \frac{s_1^2}{s_2^2}\frac{1}{f_+}.$$

Putting it together:

$$\frac{1}{f_+}\frac{s_1^2}{s_2^2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2}{s_2^2}\frac{1}{f_-}.$$

So we'll reject H_0 when:

$$1 \notin \left[\frac{1}{f_{n_1-1,n_2-1,\alpha/2}}\frac{s_1^2}{s_2^2}, \frac{1}{f_{n_1-1,n_2-1,1-\alpha/2}}\frac{s_1^2}{s_2^2} \right].$$

(The 1-sided tests can be derived similarly).

15.075J / ESD.07J Statistical Thinking and Data Analysis

Fall 2011