Chapter 19, Fourth Edition : Two Sample Problems

Note : Here the 4-th and 5-th editions of the text have different chapters, but the material is the same.

This chapter extends the idea of comparing two populations by comparing their population means. Unlike the idea of matched pairs we cannot just modify the idea of inference for a population mean from a single simple random sample. It is much more common to have two independent random samples, say $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ from two populations. For example one may wish to compare some characteristic (i) for male students and for female students; (ii) IQ scores for two provinces (or two schooling or pedagogical methods); (iii) hardness or strength of two different steel manufacturing processes.

Conditions for Inference to Compare Two Population Means

- The two populations are normally distributed with with unknown means and variance. We also need the population variances to be close to each other in size.

  This later part indicates the two population distributions are of similar shape and spread. The similar spread will then allow us to interpret or simplify the comparison of the two populations in terms of the population means.

  In practice it is also not so important that the population distributions be normal, but that they are roughly symmetric and roughly bell shaped.

- The sample (or data) from the two populations are two independent simple random samples.

  This means that the data from the first sample cannot influence the data that is collected or observed for the second sample.

Next we need to names and terminology to keep the two samples and populations identifiable or distinct in our work.

Population parameters

Population	Random Variables	Population Mean	Population Standard Deviation
1	X_1	μ_1	σ_1
2	X_2	μ_2	σ_2

The sample or data will now be represented by

$$x_{1,1}, x_{1,2}, x_{1,3}, \ldots, x_{1,n_1}$$

and

$$x_{2,1}, x_{2,2}, x_{2,3}, \ldots, x_{2,n_2}$$

where the two random samples are of size n_1 for the sample from population 1 and of size n_2 for the sample from population 2. Notice that since there is no requirement of pairing the data in the two samples (recall they are independent samples) there simple random sample sizes n_1 and n_2 need not be the same. One may observe $n_1 = 10$ and $n_2 = 15$ from the two populations.

Sometimes the two samples are written as

$$x_{1,i}, i = 1, 2, \ldots, n_1$$

and

$$x_{2,j}, j = 1, 2, \ldots, n_2$$

Data summary form

Population	Sample size	Sample Mean	Sample Standard Deviation
1	n_1	$\bar{x}_1$	s_1
2	n_2	$\bar{x}_2$	s_2

Data for Example 19.2 Edition 4

This experiment measures strength of polyester after being buried for a certain number of weeks, in this case 2 or 16 weeks. This type of information is useful to determine how quickly this fabric decays in a dump or landfill.

The data is given below.

Sample	Data				
1	118	126	126	120	129
2	124	98	110	140	110

The data summary is

Population	Sample size	Sample Mean	Sample Standard Deviation
1	$n_1 = 5$	$\bar{x}_1 = 123.8$	$s_1 = 4.60$
2	$n_2 = 5$	$\bar{x}_2 = 116.4$	$s_2 = 16.09$

What is the problem that one wishes to study? What is the purpose of the data collection experiment?

The question of interest concerns whether or not polyester fabric decays in a dump or landfill. We need to translate or rewrite this in terms of parameters for a statistical model. This is because the decay rate is too complex to give a perfect prediction. Instead we will try formulate the question in terms of the average decay at various time periods. Thus we compare the population mean or average decay at 2 weeks versus the average decay at 16 weeks. The decay is measured by the amount of strength is takes to break (or pull apart the fabric. We use this to formulate the problem into terms of a statistical model. Specifically $X_{1,i}$ represents the random variables for population of the fabric strengths after 2 weeks buried, and $X_{2,j}$ represents the random variables for population of the fabric strengths after 16 weeks buried in the ground. We assume these population models are normal, mean μ_1, standard deviation σ_1 and normal, mean μ_2, standard deviation σ_2 respectively.

In terms of the two population means what is a natural null and alternative? We are really interested in determining if the breaking strength is equal at the two weeks versus is there evidence that the breaking strength at week 16 is less than at week 2. This means that the week 16 buried fabric is weaker than the week 2 buried fabric. We may continue to use parameters with subscripts 1 and 2, or we prefer to use notation with subscripts W2 and W16 for convenience.

Formulate the problem : We consider the null hypothesis $H_0 : \mu_1 = \mu_2$ (or $H_0 : \mu_{W2} = \mu_{W16}$) versus the alternative hypothesis $H_a : \mu_2 < \mu_1$ (or $H_a : \mu_{W16} < \mu_{W2}$).

Next we will need to consider what is an appropriate test statistic and what is an appropriate rejection region. Using our general ideas from Chapter 14 and the other one sample problems we will base the test statistic on the random variable $\bar{X}_1 - \bar{X}_2$. We will need to use its sampling distribution.

In Chapter 11, which deals with the sampling distribution of sample mean random variables, we had

$$\bar{X}_1 \sim \text{Normal, mean } = \mu_1, \text{ standard deviation } = \frac{\sigma_1}{\sqrt{n_1}}$$

and

$$\bar{X}_2 \sim \text{Normal, mean } = \mu_2, \text{ standard deviation } = \frac{\sigma_2}{\sqrt{n_2}}$$

It is also a fact that the population mean of $\bar{X}_1 - \bar{X}_2$ is given by $\mu_1 - \mu_2$. In particular this says that $\bar{X}_1 - \bar{X}_2$ is an unbiased estimator of $\mu_1 - \mu_2$.

An additional property for differences of independent random variables is that

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

and so

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

The standard error of $\bar{X}_1 - \bar{X}_2$, that is the estimator of the standard deviation of $\bar{X}_1 - \bar{X}_2$ is then given by

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The test statistic for testing H_0 versus H_a is then given by

$$t = \frac{\bar{X}_1 - \bar{X}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{1}$$

The centring in the numerator is 0 since IF the NULL hypothesis is true then the population mean of $\bar{X}_1 - \bar{X}_2$ is given by $\mu_1 - \mu_2 = 0$.

Next since there is a one sided alternative ($H_a : \mu_1 > \mu_2$) we reject $\bar{x}_1 - \bar{x}_2$ is much larger than the null hypothesized value of 0. This is equivalent to rejecting the null hypothesis H_0 in favour of the alternative H_a is

$$t > t^*$$

for an appropriate critical value t^*, and where the test statistic is t given by equation (1).

IF H_0 is true then the sampling distribution of t given by equation (1) is the Student's t distribution, with degrees of freedom given by a complicated formula. However we can use approximately the correct degrees of freedom (df) by using

$$\text{df} = \text{ smaller of } n_1 - 1 \text{ and } n_2 - 1$$

Aside : This procedure is *robust* in that the Student's t distribution with this degrees of freedom holds even if the population distributions are not quite normal.

Suppose we wish to perform our test of hypothesis at level $\alpha - 0.05$. The degrees of freedom are 5 - 1 = 4. Using Table C the critical value is 2.132.

This is because the upper tail area is 0.05. To use Table C we then need the upper and lower tail area each to be 0.05, thus the central area on this table will have to be 1 - 2*0.05 = 1 - 0.10 = 0.90.

The one sided rejection region is thus to reject if $t > t^* = 2.132$.

Observe $t_{\text{obs}} = 0.989$. Since this is less than $t^* = 2.132$ we do not reject H_0. Thus at significance level 0.05 there is no evidence against the decay at 2 weeks and 16 weeks are equal.

Confidence Interval

Recall that there is a correspondence between confidence intervals and two sided hypothesis tests.

The 100*(1-α)% confidence interval for $\mu_1 - \mu_2$ is given by

$$\bar{x}_1 - \bar{x}_2 \pm t^* SE \quad = \quad \bar{x}_1 - \bar{x}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where t^* is the critical value obtained from Table C where the central region is of area $C = 1 - \alpha$ and using the appropriate degrees of freedom.

In our example the 95% confidence interval, based on 5 - 1 = 4 degrees of freedom, is given by

$$\begin{aligned} \bar{x}_1 - \bar{x}_2 \pm t^* SE &= \bar{x}_1 - \bar{x}_2 \pm 2.776 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\ &= 123.8 - 116.4 \pm 2.776 * 7.483 \\ &= 7.4 \pm 20.773 \\ &= [-13.37, 28.17] \end{aligned}$$

Since the value of $\mu_1 - \mu_2 = 0$ is in this confidence interval there is no evidence at the 90% confidence level against μ_1 equal to μ_2.

In most statistical packages there are options in the two sample statistics called *equal* variance and *unequal* variance cases. The procedure discussed here is the one called the unequal variance procedure or case.

The equal variance case makes an additional assumption that $\sigma_1 = \sigma_2$. It then uses an appropriate estimator for the common value $\sigma = \sigma_1 = \sigma_2$. This is the so called pooled variance estimator of the common population variance. For this the corresponding Student's t statistic has degrees of freedom equal to $n_1 - 1 + n_2 - 1 = n_1 + n_2 - 2$.

This text does not consider this case. However in many applications in practice this is the case that is used. The student should be aware of which procedure is used.

There are also procedures for testing the assumption or hypothesis $H : \sigma_1 = \sigma_2$, but these are not discussed in the course. Note Edition 4 deals with the F distribution for this case and this is not discussed in Edition 5. As noted on the course outline this material is not included.