2.4 **Two Sample Estimation**

Many problems in statistics involve comparing two parameters.

Example

Suppose we want to study the effect of a course in reading comprehension. What do we do? Consider the following:

A. - Select n_1 students for the course, give them a test.
   - Select another group of n_2 students, give them a test.
   - Compare the scores for the two groups.

   OR

B. – select n students, give them a test before and after taking the course.
   – compare the scores
   Which approach is better? Why? The second approach (B) is better because it reduces variability among the subjects.
   In A we have 2 independent samples. In B we have 2 dependent samples.

Defn

Two samples are independent if the data values obtained from one are unrelated to the values from the other. The samples are dependent if each data value from one sample is paired in a natural way with a data value from the other sample.

2.4.1 Two Population Means (Dependent samples)

Suppose we want to estimate the difference between two population means $\mu_D = \mu_1 - \mu_2$ where μ_1 and μ_2 are population means of X_1 and X_2 respectively. We estimate μ_D by $\bar{d}$ where

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \quad , d_i = x_{1i} - x_{2i}$$

i.e. d_i are differences between paired observations.

Defn

If $\bar{d}$ and s_d are the mean and standard deviation of normally distributed differences of n random pairs of measurements, a $100(1-\alpha)\%$ confidence interval for $\mu_D = \mu_1 - \mu_2$ is

$$\bar{d} \pm t_{\frac{\alpha}{2},n-1}\frac{s_d}{\sqrt{n}}$$

where $t_{\frac{\alpha}{2},n-1}$ is t – value leaving an upper tail area of $\frac{\alpha}{2}$.

Example

Agricultural departments of 9 universities took part in a study to test the yield capabilities of two new varieties of wheat. Each variety was planted on plots of equal area at each university and the yields, in kilogrammes per plot, recorded as follows:

	University								
Variety	1	2	3	4	5	6	7	8	9
1	38	23	35	41	44	29	37	31	38
2	45	25	31	38	50	33	36	40	43

Assume that the differences in the yields are normally distributed.

(a) Find a 95% confidence interval for the mean difference between the yields of the varieties.
(b) Is there a difference in the mean yield of the two varieties?

Soln

(a)

$$d_i = x_1 - x_2 \ (i.e. variety\ 1 - variety\ 2\ at\ each\ university)$$

-7, -2, 4, 3, -6, -4, 1, -9, -5

$$\bar{d} = -2.78, \ \ s_d = 4.5765, n = 9, \alpha = 0.05$$

($\bar{d}$ and s_d are obtained directly from the calculator)

$$\bar{d} \pm t_{\frac{\alpha}{2},n-1}\frac{s_d}{\sqrt{n}}$$

$$-2.78 \pm t_{0.025,8}\frac{4.5765}{\sqrt{9}}$$

$$-2.78 \pm 2.306\left(\frac{4.5765}{3}\right)$$

$$(-6.3, 0.7)\ or\ -6.3 < \mu_1 - \mu_2 < 0.7$$

Note that if $d_i = x_2 - x_1$ then the confidence interval would be $(-0.7, 6.3)$

(b)

The confidence interval contains 0, therefore there's no difference in the mean yield of the two varieties.

Note that if there is no difference between the varieties then the mean difference would be 0, that is why we conclude that there is no difference if 0 is in the confidence interval.

2.4.2 Two population means (Independent samples)

If we have two independent samples of sizes n_1 and n_2 from populations X_1 and X_2, then an estimate for $\mu_1 - \mu_2$ is $\bar{x}_1 - \bar{x}_2$ ($\bar{X}_1 - \bar{X}_2$ is the estimator).
To construct a $100(1-\alpha)\%$ we consider three cases:

(I) $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ where σ_1^2 and σ_2^2 are known then

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

Note:
1. It doesn't matter whether the samples n_1 and n_2 are large or small the interval is based on the normal distribution.
2. This is an exact confidence interval.

(II) $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ but σ_1^2 and σ_2^2 are unknown

(a) Assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (i.e. the unkown variances are equal). Estimate σ^2 by the pooled variance

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

then

$$\bar{x}_1 - \bar{x}_2 \pm t_{\frac{\alpha}{2},\nu}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where $\nu = n_1 + n_2 - 2$ is the degrees of freedom for the t – distribution.

(b) If $\sigma_1^2 \neq \sigma_2^2$ then

$$\bar{x}_1 - \bar{x}_2 \pm t_{\frac{\alpha}{2},\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $t_{\frac{\alpha}{2},\nu}$ is a t – distribution with

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

degrees of freedom.

Note:

1. This case only applies to small samples.
2. Part(a) of this case gives an exact confidence interval while part(b) gives an approximate confidence interval.

(III) The distributions of X_1 and X_2 are unknown but the samples are large ($n_1 \geq 30,\ n_2 \geq 30$)

(a) If σ_1^2 and σ_2^2 are known then

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

(b) If σ_1^2 and σ_2^2 are unknown then

$$\bar{x}_1 - \bar{x}_2 \pm z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Note:

1. For this case the confidence intervals are approximate.
2. If the sample sizes are small the populations have to be normal and Cases (I) and (II) will apply. The normal assumption is not required if the samples are large and Case (III) will apply.