Week 11

Confidence Intervals for the Difference between the means of Two Normal Populations

To compare two populations is encountered widely in statistics. For example, a company buys a chemical substance from two different suppliers and concerns about the difference between the mean levels of impurity present in the chemicals from the two sources of supply. Another example, a farmer may consider the use of two alternative fertilizers, his interest being the difference between the resulting mean crop yields per acre.

1.Matched Pairs

2.Independent Samples

Definition (Matched Pairs): In this design, the sample members are chosen in pairs, one from each population. The idea is that the members of these pairs should resemble one another as closely as possible so that the comparison of the interest can be made directly. For instance, we want to measure the effectiveness of a speed-reading course. One possible approach would be to record the number of words per minute read by a sample of students after completing the course. In this case, each pair of observations consists of before and after measurements on a single student.

Definition (Independent Samples): In this design, samples are drawn independently from two populations of interest so that the membership of one sample is not influenced by that of another. In the example of a company buying chemical from to suppliers, the independent sample are used since it is assumed that the first company providing chemicals are not influenced by the ones provided by the second company or visa versa.

1.Confidence Intervals Based on Matched Pairs

Suppose that the means of the two population are denoted by $\mu_X\ and\ \mu_{Y.}$ The sample sizes of n are drawn are denoted by $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$. These samples are called paired samples since they are either same observations or they are considered very close concerning interest of the research. To find the confidence interval between two populations, we need to compute two statistics that are defined by

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\sum_{i=1}^{n} x_i - y_i$$

$$s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$$

$$P\left(\bar{d} - \frac{s_d}{\sqrt{n}} t_{(n-1),\frac{\alpha}{2}} \leq \mu_x - \mu_y \leq \bar{d} - \frac{s_d}{\sqrt{n}} t_{(n-1),\frac{\alpha}{2}}\right) = 1 - \alpha$$

where they are the mean of difference and the variance of the difference.

Ex: Table 1 shows fuel consumption figures obtained from a random sample of eight cars from each of two different models. The sample cars are paired. Each member of a particular pair was driven over the same route by the same driver so that variability between drivers and routes could be laminated from the comparisons. Find the 99 percent confidence interval and interpret the result.

Table 1

İ	X Brand Automobiles	Y Brand Automobiles	d_i	$(d_i - \bar{d})^2$
1	19.4	19.6	-0.2	$(-0.2 - 1.03)^2$
2	18.8	17.5	1.3	$(1.3 - 1.03)^2$
3	20.6	18.4	2.2	$(2.2 - 1.03)^2$
4	17.6	17.5	0.1	$(0.1 - 1.03)^2$
5	1.2	18.0	1.2	$(1.2 - 1.03)^2$
6	20.9	20.0	0.9	$(0.9 - 1.03)^2$
7	18.3	18.8	-0.5	$(-0.5 - 1.03)^2$
8	20.4	19.2	1.2	$(1.2 - 1.03)^2$

$\bar{d} = \frac{1}{6}(-0.2 + 1.3 + 2.2 + 0.1 + 1.2 + 0.9 + (-0..5) + 1.2) = 1.03$

$s_d^2 = 0.82$ ve $s_d = 0.90$

$$P\left(\bar{d} - \frac{s_d}{\sqrt{n}} t_{(n-1),\frac{\alpha}{2}} \leq \mu_x - \mu_y \leq \bar{d} - \frac{s_d}{\sqrt{n}} t_{(n-1),\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$P\left(1.03 - \frac{0.90}{\sqrt{8}} 3.499 \leq \mu_x - \mu_y \leq 1.03 + \frac{0.90}{\sqrt{8}} 3.499\right) = 0.99$$

$$P(-0.34 \leq \mu_x - \mu_y \leq 1.89) = 0.99$$

Remark: The difference of the mean fuel consumption of two cars lies between -0.34 and 1.89, which implies that the population means are the same is not extremely strong since the interval includes 0.

2.Confidence Intervals Based on Independent Samples

Suppose that two random samples, n_x *and* n_y, not necessarily being equal are drawn from two different populations. Two populations have means and variances denoted by μ_X, μ_Y and σ_X^2, σ_Y^2. The confidence interval for the difference of two population means are defined by

$$1.P((\bar{x}_1 - \bar{x}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \leq \mu_x - \mu_y \leq ((\bar{x}_1 - \bar{x}_2) + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}) = 1 - \alpha$$

$$2. P((\bar{x}_1 - \bar{x}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} \leq \mu_x - \mu_y \leq ((\bar{x}_1 - \bar{x}_2) + z_{\frac{\alpha}{2}}\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}) = 1 - \alpha$$

While the first confidence interval is used when population variances, σ_X^2, σ_Y^2, are known, the second confidence interval is used when sample variances, $s_X^2,\ s_Y^2$, are known or computed.

Ex: A company wants to know whether there is a difference between employees who smoke cigarettes and who do not smoke cigarettes when concerning the absence from the work. Then, the company randomly selects 96 employees who smoke cigarettes and measures how many hours they are absent from the work. The mean and standard deviation are 2.15 and 2.09 hours, respectively. On the other hand, 206 employees are randomly chosen to determine how many hours they are absent from the work. The mean and standard deviation are 1.69 and 1.91 hours, respectively. Find the 99 percent confidence interval for the difference of means and interpret the result.

$$\bar{x} = 2.15, \bar{y} = 1.69$$

$$s_x = 2.09, s_y = 1.91$$

$$n_x = 96, n_y = 206$$

$$P((\bar{x}_1 - \bar{x}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} \le \mu_x - \mu_y \le ((\bar{x}_1 - \bar{x}_2) + z_{\frac{\alpha}{2}}\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}) = 1 - \alpha$$

$$P((2.15 - 1.69) - 2.575\sqrt{\frac{(2.09)^2}{96} + \frac{(1.91)^2}{206}} \le \mu_x - \mu_y$$

$$\le ((2.15 - 1.69) + 2.575\sqrt{\frac{(2.09)^2}{96} + \frac{(1.91)^2}{206}}) = 0.99$$

$$P(-0.19 \le \mu_x - \mu_y \le 1.11) = 0.99$$

Remark: 99 percent confidence interval for the mean difference of the populations between the employees smoking cigarettes and those not smoking cigarettes lies between -0.19 and 1.11. Since 0 is included in this interval, the evidence presented is not strong that the means of the two population is different.

Confidence Intervals for Difference Between the Means of Two Normal Populations: Independent Samples, Population Variances Equal

Suppose that we have independent random samples of $n_x\ and\ n_y$, not necessarily being equal that are drawn from two different populations. Two populations have means and variances denoted by $\mu_X,\ \mu_Y$ and σ_X^2, σ_Y^2. Besides, we assume that $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ olsun. The confidence interval for the difference of two population means are defined by

$1.P((\bar{x}_1 - \bar{x}_2) - t_{(n_x+n_y-2),\frac{\alpha}{2}} s_{ort}\sqrt{\frac{n_x+n_y}{n_x n_y}} \leq \mu_x - \mu_y \leq ((\bar{x}_1 - \bar{x}_2) - t_{(n_x+n_y-2),\frac{\alpha}{2}} s_{ort}\sqrt{\frac{n_x+n_y}{n_x n_y}}) = 1 - \alpha$

$$s_{Pool}^2 \quad = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x + n_y - 2)}$$

Ex: In the study of the effects of planning on the financial performance of banks, a random sample of six “partial formal planners” shows mean annual percentage increase in the net income of 9.972 and a standard deviation of 7.470. An independent random sample of nine banks with “no formal planning” system has a mean annual net income increase of 2.098 and standard deviation of 10.834. Assuming that the distribution is normal and the population variances are equal. Find 90 percent confidence interval and interpret the result.

$$\bar{x} = 9.97, \bar{y} = 2.098$$

$$s_x = 7.47, s_y = 10.83$$

$$n_x = 6, n_y = 9$$

$$s_{ort}^2 = \frac{(5)(7.47)^2 + (8)(10.83)^2}{(6 + 9 - 2)} = 93.69$$

$$s_{ort=9.68}$$

$$P((\bar{x}_1 - \bar{x}_2) - t_{(n_x+n_y-2),\alpha/2}\sqrt{\frac{n_x + n_y}{n_x n_y}} \leq \mu_x - \mu_y \leq ((\bar{x}_1 - \bar{x}_2) - t_{(n_x+n_y-2),\alpha/2}\sqrt{\frac{n_x + n_y}{n_x n_y}}) = 1 - \alpha$$

$$P\left((9.97 - 2.098) - (1.771)(9.68)\sqrt{\frac{6+9}{54}} \leq \mu_x - \mu_y \leq (9.97 - 2.098) + (1.771)(9.68)\sqrt{\frac{6+9}{54}}\right) = 0.90$$

$$P(-1.161 \leq \mu_x - \mu_y \leq 16.909) = 0.90$$

Remark: 90 percent confidence interval for the mean difference of the banks that have “partial formal planners” and no formal planning system lies between -1.161 and 16.909. Since 0 is included in this interval, the evidence presented is not strong that the means of the two population is different.

Intervals for The Difference Between Two Population Proportions (Large Samples)

Let $P_x \; and \; P_Y$ denote success rates in two populations concerning an interest. A random samples of $n_x \; and \; n_y$ are drawn form those populations and the sample success rates are denoted by $\hat{p}_x \; and \; \hat{p}_y$. The confidence interval for the difference of the proportions for those population are defined by

$$P\left((\hat{p}_x - \hat{p}_y) - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}} \le p_x - p_y \le (\hat{p}_x - \hat{p}_y) - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}} \right) = 1 - \alpha$$

Ex: The male and female sophomore students majoring in the field of accounting are asked to assess when they find a full time job after graduation. While 107 out of 120 males believes to find a full time job after 10 years, 73 out of 141 females expect the same to happen. Find the confidence interval of the difference of the two population portions with 95 percent and interpret the results.

$$n_x = 120 \; n_y = 141$$

$$\hat{p}_x = \frac{107}{120} = 0.892 \qquad \hat{p}_y = \frac{73}{141} = 0.518$$

$$P\left((\hat{p}_x - \hat{p}_y) - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}} \le p_x - p_y \le (\hat{p}_x - \hat{p}_y) - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}} \right) = 1 - \alpha$$

$$P\left((0892-0.518)-(1.96)\sqrt{\frac{(0.892)(0.108)}{120}+\frac{(0.518)(0.482)}{141}}\leq\left(\hat{p}_x-\hat{p}_y\right)\right.$$

$$\left.\leq(0892-0.518)+(1.96)\sqrt{\frac{(0.892)(0.108)}{120}+\frac{(0.518)(0.482)}{141}}\right)=0.95$$

$$P\left(0.275\leq\left(\hat{p}_x-\hat{p}_y\right)\leq 0.473\right)=095$$

Remark: 95 percent confidence interval of the difference of the two population portions lies between 0.275 and 0.473. The portion of the male students having higher expectancy than female students when having a full time job after 10 years of graduation could be expressed.