CHAPTER 9: CONFIDENCE INTERVAL ESTIMATION: FURTHER TOPICS

Confidence Interval Estimation of the Difference Between Two Normal Population Means: Dependent Samples

- To compare population means, random samples are drawn from the two populations and their means are compared.
  - Suppose a company receive shipments of a chemical from two suppliers and be concerned about the difference between the mean levels of impurity in the chemicals from the two sources

Ozan Eksi, TOBB-ETU

Dependent Samples

- ‘Dependent samples’ means that the values in one sample are influenced by the values in the other sample. They are either matched pairs (resemble one another as closely as possible) or the same individual or objects tested twice
  - Suppose that the effectiveness of a speed-reading course is to be measured. Record and compare the number of words per minute read by a sample of students before taking the course and and after taking the course

- Let's denote the means of two normally distributed populations by μ_X and μ_Y.
- Suppose the random sample of $x_1, x_2, ..., x_n$ are drawn from the first population, and the random sample of $y_1, y_2, ..., y_n$ are drawn from the second population
- The difference of observations is $d_i = x_i - y_i$, the mean and standart deviation of these differences is $\bar{d}$ and s_d

- If the population distribution of the differences is assumed to be normal, then a $100(1-\alpha)\%$ confidence interval for the differences between means is $(\mu_d = \mu_X - \mu_Y)$

$$\bar{d} - \frac{t_{n-1,\alpha/2} * s_d}{\sqrt{n}} < \mu_X - \mu_Y < \bar{d} + \frac{t_{n-1,\alpha/2} * s_d}{\sqrt{n}}$$

- Örnek (matched pairs): Farklı yıllarda üretilmiş aynı marka fakat farklı model 2 grup arabadan 8 tanesi seçilerek eşleştirilmiş, ve mile/gallon cinsinden aşağıdaki yakıt tüketimleri elde edilmiştir. Bu iki grup arabanın yakıt tüketimlerine ait populasyon ortalamalarının farkının %99 güven aralığını bulunuz

i	x_i	y_i		d_i	d_i^2
1	19.4	19.6		-0.2	.04
2	18.8	17.5		1.3	1.69
3	20.6	18.4		2.2	4.84
4	17.6	17.5		.1	.01
5	19.2	18.0		1.2	1.44
6	20.9	20.0		.9	.81
7	18.3	18.8		-.5	.25
8	20.4	19.2		1.2	1.44
			Sums	6.2	10.52

- Kullanacağımız formül

$$\bar{d} - t_{n-1,\alpha/2} * \frac{s_d}{\sqrt{n}} < \mu_X - \mu_Y < \bar{d} + t_{n-1,\alpha/2} * \frac{s_d}{\sqrt{n}}$$

- Örneklem ortalaması

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{6.2}{8} = .775$$

örneklem varyansı

$$s_d^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{10.52 - 8 * .775^2}{7}$$

ve standart sapması

$$s_d = \sqrt{.816} = .903$$

- Soruda verilen bilgiler: $\bar{d} = .775$ $\quad s_d = .903$ $\quad n = 8$ $\quad \alpha = .01$ $\quad t_{n-1,\alpha/2} = t_{7,.005} = 3.499$
- Formüle uyguladığımızda

$$.775 - \frac{3.499 * .903}{\sqrt{8}} < \mu_X - \mu_Y < .775 + \frac{3.499 * .903}{\sqrt{8}}$$

$$-.342 < \mu_X - \mu_Y < 1.892$$

- Yukarıdaki aralık 0'ı içerdiğinden populasyon ortalamaları arasında ciddi bir fark göremiyoruz.

Confidence Interval Estimation of the Difference Between Two Normal Population Means: Independent Samples

- In this case we develop confidence interval estimation when two samples are drawn independently from two normally distributed populations. We consider three situations

  1. Both population variances are known
  2. Both population variances are not known, but considered to be equal
  3. Both population variances are not known and not considered to be equal (will not cover)

1- Both population variances are known

- Suppose n_x observation are drawn from a population with mean μ_X and variance σ_X^2, and n_y observation are drawn from a population with mean μ_Y and variance σ_Y^2, let the respective sample means are $\bar{x}$ and $\bar{y}$

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_X - \mu_Y$$

since the samples are independent

$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) = \frac{\sigma_x^2}{n_x} + \frac{\sigma_Y^2}{n_y}$$

and it can be shown that in the case of indepence the differences are also normally distributed

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_Y^2}{n_y}}}$$

Then the $100(1-\alpha)\%$ confidence interval of the differences of population means can be formed as

$$(\bar{x}-\bar{y}) - z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_Y^2}{n_y}} < \mu_X - \mu_Y$$
$$< (\bar{x}-\bar{y}) + z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_Y^2}{n_y}}$$

- Örnek: Rassal bir şekilde seçilmiş 96 kişilik sigara içen bir grupta, işten kaytarma alışkanlığının 2.15 saat/ay olduğu, seçilen bağımsız 206 kişilik sigara içmeyen bir örneklemde ise 1.69 saat/ay olduğu ortaya çıkmıştır. Bu grupların populasyonlarındaki standart sapmanın, önceki yıllarda yapılan çalışmalardan sırasıyla 2.09 saat/ay ve 1.91 saat/ay olduğu bilinmektedir. Bu iki grubun populasyon ortalamarının arasındaki fark için %99 güven aralığını bulunuz

Soruda verilen bilgiler: $\bar{x} = 2.15$ $\sigma_x = 2.09$ $n_x = 96$ $\bar{y} = 1.69$ $\sigma_y = 1.69$ $n_y = 206$
Bu iki grup profesörün ortalama harcadıkları zaman aradındaki farkın %90 güven aralığını bulunuz

$*$ Kullanacımız formül

$$(\bar{x} - \bar{y}) - z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_Y^2}{n_y}} < \mu_X - \mu_Y < (\bar{x} - \bar{y}) + z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_Y^2}{n_y}}$$

$\alpha = 10$ için kulanacağımız z değeri

$$z_{\alpha/2} = z_{.005} = 2.575$$

$$(2.15 - 1.69) - 2.575\sqrt{\frac{2.09^2}{96} + \frac{1.91^2}{206}} < \mu_X - \mu_Y$$

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

$$-.19 < \mu_X - \mu_Y < 1.11$$

- Örnek: Bir grup işletme ve bir grup iktisat profesörü bir ders saati için gereken ders notlarının hazırlığı için kaç saat harcadıkları sorulmuştur. 321 tane iktisat hocasının ortalama 3.01 saat, 94 işletme hocasının da ortalama 2.88 saat harcadıkları görülmüştür. Daha önceki anket çalışmalarından iktisat hocaları populasyonunun bu soruya cevabının 1.09, işletme hocaları populasyonunun da 1.01 standart sapma gösterdikleri gözlenlenmiştir. Bu iki grup profesör populasyonunun ortalamala harcadıkları saatler arasındaki farkın %95 güven aralığını bulunuz

– İktisat profesörlerini X ile, işletme profesörlerini Y ile gösterelim.

– Soruda verilen bilgiler: $\bar{x} = 3.01$ $\sigma_x = 1.09$ $n_x = 321$ $\bar{y} = 2.88$ $\sigma_y = 1.01$ $n_y = 94$

$$(\bar{x} - \bar{y}) - z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_Y^2}{n_y}} < \mu_X - \mu_Y$$
$$< (\bar{x} - \bar{y}) + z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_Y^2}{n_y}}$$

$$z_{\alpha/2} = z_{.025} = 1.96$$

$$(3.01 - 2.88) - 1.96\sqrt{\frac{1.09^2}{321} + \frac{1.01^2}{94}} < \mu_X - \mu_Y$$

$$< (3.01 - 2.88) + 1.96\sqrt{\frac{1.09^2}{321} + \frac{1.01^2}{94}}$$

$$-.11 < \mu_X - \mu_Y < .37$$

2- Both population variances are not known, but considered to be equal

- Suppose we don't know the population variances of the independent populations

- If these unknown variances are assumed to be equal: $\sigma_x^2 = \sigma_y^2 = \sigma^2$, then we can use sample variances, $s_x^2 = s_y^2$, to estimate the common variance σ^2. In a way that

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

and as we did before when we use sample standart deviations we refer to Student t-distributions

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}}$$

As a result, the $100(1 - \alpha)\%$ confidence interval of the differences of population means can be formed as

$$(\bar{x}-\bar{y}) - t_{n_x+n_y-2,\alpha/2}\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}} < \mu_X - \mu_Y$$
$$< (\bar{x}-\bar{y}) + t_{n_x+n_y-2,\alpha/2}\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}$$

– Note that since we use two sample means, we lose two degrees of freedom, and the required degree of freedom you need to use is n_x+n_y-2

- Örnek: Planlamanın bankaların finansal performansı üzerindeki etkilerini araştıran bir çalışmada, rastgele bir şekilde 6 elemanlı bir 'profesyonel planlayıcılar' örnekleminde gelirlerin yıllık yüzde ortalama artışının 9.972, bunun standart sapmasınında 7.470 olduğu gözlemlenmiştir. 'Profesyonel olmayan planlayıcılar' olan 9 elemanlı bağımsız diğer bir rassal banka örnekleminde ise bu yüzde artışın ortalamasının 2.098, bunun standart sapmasınında 10.834 olduğu gözlemlenmiştir. Her iki grup populasyonunda normal dağıldığı ve aynı varyansa sahip olduğu düşünülürse, bu populasyonların ortalamaları arasındaki farkın %90 güven aralığını bulunuz

– Soruda verilen bilgiler: $\bar{x} = 9.972$ $s_x = 7.470$ $n_x = 6$ $\bar{y} = 2.098$ $s_x = 10.834$ $n_y = 9$

$$s_p^2 = \frac{(n_x - 1)s_X^2 + (n_y - 1)s_Y^2}{(n_x + n_y - 2)}$$

$$= \frac{5 * 7.47^2 + 8 * 10.834^2}{13} = 93.693$$

$$s_p = \sqrt{93.693} = 9.68$$

$$(\bar{x} - \bar{y}) - t_{n_x+n_y-2,\alpha/2} s_p \sqrt{\frac{n_x + n_y}{n_x n_y}} < \mu_X - \mu_Y$$

$$< (\bar{x} - \bar{y}) + t_{n_x+n_y-2,\alpha/2} s_p \sqrt{\frac{n_x + n_y}{n_x n_y}}$$

$$t_{n_x+n_y-2,\alpha/2} = t_{13,.05} = 1.771$$

$$
\begin{gathered}
(9.972 - 2.098) - 1.771 * 9.68\sqrt{\frac{6+9}{54}} \\
< \mu_X - \mu_Y \\
< (9.972 - 2.098) + 1.771 * 9.68\sqrt{\frac{6+9}{54}}
\end{gathered}
$$

$$1.161 < \mu_X - \mu_Y < 16.909$$

Confidence Interval Estimation of the Difference Between Two Population Proportion (Large Samples)

- Suppose we are are interested in the difference between the proportion of successes in two populations, and we use sample proportions two infer conclusions about that
- The distribution of the difference of the proportion of successes in two samples

$$Z = \frac{(\hat{p}_X - \hat{p}_Y) - E(\hat{p}_X - \hat{p}_Y)}{Var(\hat{p}_X - \hat{p}_Y)}$$

- The mean of the difference between sample proportions

$$E(\hat{p}_X - \hat{p}_Y) = E(\hat{p}_X) - E(\hat{p}_Y) = P_X - P_Y$$

the standart deviation of the difference between sample proportions

$$\begin{aligned} Var(\hat{p}_X - \hat{p}_Y) &= Var(\hat{p}_X) - Var(\hat{p}_Y) \\ &= \frac{P_X(1 - P_X)}{n_x} + \frac{P_Y(1 - P_Y)}{n_y} \end{aligned}$$

for the large sample size, the population proportions, P_X and P_Y, can be replaced by sample proportions, $\hat{p}_X$ and $\hat{p}_Y$, and the following random variable has a standard normal distribution

$$Z = \frac{(\hat{p}_X - \hat{p}_Y) - (P_X - P_Y)}{\sqrt{\dfrac{\hat{p}_X(1-\hat{p}_X)}{n_x} + \dfrac{\hat{p}_Y(1-\hat{p}_Y)}{n_y}}}$$

Fimally, the confidence interval can be formed as

$$(\hat{p}_X - \hat{p}_Y) - z_{\alpha/2}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_x} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_y}} < p_X$$

$$< (\hat{p}_X - \hat{p}_Y) + z_{\alpha/2}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_x} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_y}}$$

- Örnek: Tüketiciler arasında yapılan bir ankette A aletinin kullanımının zor olup olmadığı sorulmuş, ve 1. bölgeden seçilen 92 tüketiciden 49'u, 2. bölgeden seçilen 86 tüketicilenin 36'sı evet cevabını vermiştir. İki bölgedeki tüketici populasyonlarındaki evet cevabını verebilecek yüzde seçmen oranları arasındaki fark için %90 güven aralığını bulunuz

  – Soruda verilen bilgiler: $\hat{p}_X = 49/92 = .53$ $n_x = 92$ $\hat{p}_Y = 36/86 = .42$ $n_y = 86$

– Yukarıdaki formülde ihtiyacımız olan değer

$$z_{\alpha/2} = z_{.05} = 1.645$$

son olarak

$$(.53 - .42) - 1.645\sqrt{\frac{.53 * .47}{92} + \frac{.42 * .58}{86}} < p_X - p_Y < (.53 - .42) + 1.645\sqrt{\frac{.53 * .47}{92} + \frac{.42 * .58}{86}}$$

$$-.008 < p_X - p_Y < .236$$

- Örnek: Seçimlerdeki bir partinin seçmenler tarafından nasıl algılandığını belirlemek üzere pek çok tahmin yapılmıştır. A bölgesinden seçilmiş120 kişilik rassal bir örneklemde, 107 kişi bahsi geöen partiyi desteklediklerini belirtmiştir. B bölgesinden seçilmiş 141 kişilik rassal bir örneklemde, 73 kişi bahsi geçen partiyi desteklediklerini belirtmiştir. A ve B bölgelerindeki seçmen populasyonlarında partiyi destekleyen yüzde seçmen oranları arasındaki fark için %95 güven aralığını bulunuz

– Soruda verilen bilgiler: $\hat{p}_X = 107/120 = .89$
$n_x = 120$ $\hat{p}_Y = 73/141 = .52$ $n_y = 141$

– Kullacağımız formül

$$(\hat{p}_X - \hat{p}_Y) - z_{\alpha/2}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_x} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_y}}$$

$$< p_X - p_Y$$

$$< (\hat{p}_X - \hat{p}_Y) + z_{\alpha/2}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_x} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_y}}$$

ihtiyacımız olan değer

$$z_{\alpha/2} = z_{.025} = 1.96$$

dolayısıyla

$$(.89 - .52) - 1.96\sqrt{\frac{.89 * .11}{120} + \frac{.52 * .48}{141}} < p_X - p_Y$$

$$< (.89 - .52) + 1.96\sqrt{\frac{.89 * .11}{120} + \frac{.52 * .48}{141}}$$

$$.275 < p_X - p_Y < .473$$