統計學（一）

第八章 信賴區間估計: 兩群體參數差異之估計
(Confidence Intervals Estimation: Further Topics)

授課教師：唐麗英教授

國立交通大學 工業工程與管理學系
聯絡電話：(03)5731896
e-mail：litong@cc.nctu.edu.tw

2013

本課程內容參考書目

- **教科書**
  - P. Newbold, W. L. Carlson and B. Thorne(2007). *Statistics for Business and the Economics*, 7^{th} Edition, Pearson.
- **參考書目**
  - Berenson, M. L., Levine, D. M., and Krehbiel, T. C. (2009). *Basic business statistics: Concepts and applications*, 11^{th} EditionPrentice Hall.
  - Larson, H. J. (1982). *Introduction to probability theory and statistical inference*, 3^{rd} Edition, New York: Wiley.
  - Miller, I., Freund, J. E., and Johnson, R. A. (2000). *Miller and Freund's Probability and statistics for engineers*, 6^{th} Edition, Prentice Hall.
  - Montgomery, D. C., and Runger, G. C. (2011). *Applied statistics and probability for engineers*, 5^{th} Edition, Wiley.
  - Watson, C. J. (1997). *Statistics for management and economics*, 5th Edition. Prentice Hall.
  - 唐麗英、王春和（2013），「從範例學MINITAB統計分析與應用」，博碩文化公司。
  - 唐麗英、王春和（2008），「SPSS 統計分析 」，儒林圖書公司。
  - 唐麗英、王春和（2007），「Excel 統計分析」，第二版，儒林圖書公司。
  - 唐麗英、王春和（2005），「STATISTICA與基礎統計分析」，儒林圖書公司。

兩相依(或配對)母體平均數差之估計 (Estimation of the Difference Between Two Population Means: Dependent or Paired Samples)

Interval Estimation of $\mu_d = (\mu_1 - \mu_2)$: Pair Samples

- **$(1-\alpha)100\%$ Confidence Interval for $\mu_d = (\mu_1 - \mu_2)$: Pair Samples**
  - Let $d_1, d_2, d_3, \ldots, d_n$ represent the differences between the pairwise observations in a random sample of n matched pairs, $\bar{d}$ = mean of the n sample differences, and s_d = standard deviation of the n sample differences.

σ is known (or large sample)	**σ is unknown**
$\bar{d} \pm z_{\alpha/2}\left(\frac{\sigma_d}{\sqrt{n}}\right)$	$\bar{d} \pm t_{\alpha/2}\left(\frac{s_d}{\sqrt{n}}\right)$
where σ_d is the population deviation of differences.	where $t_{\alpha/2}$ is based on (n-1) degrees of freedom.
Assumption: $n \geq 30$	Assumption: The population of paired differences is normally distributed.

Note: When σ_d is unknown (as is usually the case), use s_d to approximate σ_d.

Practical Interpretation of a Confidence Interval for $\mu_d = (\mu_1 - \mu_2)$

- **Let (LCL, UCL) represent a $(1-\alpha)100\%$ confidence interval for $(\mu_1 - \mu_2)$.**

1) If **LCL > 0** and **UCL > 0**, conclude $\mu_1 > \mu_2$

2) If **LCL < 0** and **UCL < 0**, conclude $\mu_1 < \mu_2$

3) If **LCL < 0** and **UCL > 0**, (i.e., the interval includes 0), conclude **no evidence of a difference between μ_1 and μ_2**

Interval Estimation of $\mu_d = (\mu_1 - \mu_2)$: Pair Samples

- **例 1: 請參考課本349頁 例8.1**

A medical study was conducted to compare the difference in effectiveness of two particular drugs in lowering cholesterol levels(膽固醇指數). The research team used a paired sample approach to control variation in reduction that might be due to factors other then the drug itself. Each member of a pair was matched by age, weight, lifestyle, and other pertinent factors. Drug X was given to one person randomly selected in each pair, and drug Y was given to the other individual in the pair. After a specified amount of time each person's cholesterol level was measured again. Suppose that a random sample of eight pairs of patients with known cholesterol problem is selected from the large populations of participants. Table gives the number of points by which each person's cholesterol level was reduced, as well as the differences, $d_i = x_i - y_i$, for each pair. Estimate with a 99% confidence level the mean difference in the effectiveness of the two drugs, X and Y, to lower cholesterol.

Interval Estimation of $\mu_d = (\mu_1 - \mu_2)$: Pair Samples

- **例 1:**

Pair	Drug X	Drug Y	Pair Difference (d_i)
1	29	26	3
2	32	27	5
3	31	28	3
4	32	27	5
5	32	30	2
6	29	26	3
7	31	33	-2
8	30	36	-6

Interval Estimation of $\mu_d = (\mu_1 - \mu_2)$: Pair Samples

- **[Ans]**

From the table, we obtain:

$n = 8$ $\quad \bar{d}=1.625 \quad$ and $\quad s_d = 3.777$

Thus the 99% confidence interval for $\mu_d = (\mu_x - \mu_y)$ is

$$\mu_x - \mu_y = \bar{d} \pm \frac{t_{n-1,\alpha/2}(s_d)}{\sqrt{n}}, \qquad \text{where } t_{n-1,\alpha/2} = t_{7,0.005} = 3.499.$$

$$\mu_x - \mu_y = 1.625 \pm \frac{(3.499)(3.777)}{\sqrt{8}}$$

$$= (-3.05, 6.30)$$

That is, $\quad -3.05 < \mu_x - \mu_y < 6.30$

Since the confidence interval contains the value of zero, it is not possible to determine if either drug is more effective in reducing one's cholesterol level.

Interval Estimation of $\mu_d = (\mu_1 - \mu_2)$: Pair Samples

- **例 2: 請參考課本350頁 例8.2**

Countless Web sites, study guides, software, on-line interactive courses, books, and classes promise to increase students' vocabulary, to refresh students' math skill, and to teach test-making strategies in order to improve SAT scores, which should help to enhance(改善) chances of college acceptance, or increase the possibilities of receiving certain scholarships. Similarly, the same types of offerings exist to improve GMAT scores, LSAT scores, MCAT scores, and other such standardized tests. One company randomly sampled 140 of its clients and collected data on each person's SAT score before taking the on-line course and each person's SAT score after taking the course. We obtain the following information:

$$\bar{d} = 77.7 \qquad and \qquad s_d = 43.68901$$

Estimate the difference in the same SAT scores before and after taking the on-line course. Using confidence interval estimation.

Interval Estimation of $\mu_d = (\mu_1 - \mu_2)$: Pair Samples

- **[Ans]**

We want to estimate $\mu_d = \mu_{after} - \mu_{before}$

$$n = 140, \quad \bar{d} = 77.7, \quad s_d = 43.68901$$

The 95% confidence interval for $\mu_d = \mu_{after} - \mu_{before}$ is

$$\mu_d = \mu_{after} - \mu_{before} = \bar{d} \pm \frac{t_{n-1,\alpha/2} S_d}{\sqrt{n}}$$

$$where\ t_{n-1,\alpha/2} = t_{139,0.025} \cong 1.96$$

$$\mu_d = 77.7 \pm \frac{(70.5)(43.68901)}{\sqrt{140}} = (70.5, 84.9)$$

i.e. $$70.5 < \mu_{after} - \mu_{before} < 84.9$$

Since the confidence interval does not contains zero, we can conclude that the on-line course can improve the SAT scores significantly.

兩獨立母體平均數差之估計
(Estimation of the Difference Between Two Population Means: Independent Samples)

Interval Estimation of $(\mu_1-\mu_2)$:Two Independent Samples

- **Case 1: σ_1^2 and σ_2^2 is known**
- **(1-α)100% Confidence Interval for $\mu_1-\mu_2$ using Independent Samples**

$$\left(\bar{x}_1-\bar{x}_2\right)\pm z_{\alpha/2}\sigma_{\bar{x}_1-\bar{x}_2}=\left(\bar{x}_1-\bar{x}_2\right)\pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\approx\left(\bar{x}_1-\bar{x}_2\right)\pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

Recall : **The Point Estimation of $(\mu_1-\mu_2)$is $\bar{x}_1-\bar{x}_2$ and $\sigma_{\bar{x}_1-\bar{x}_2}=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$**

- **Assumptions:**

1. The two random samples are selected in an **independent** manner from the target populations. That is, the choice of elements in one sample does not affect, and is not affected by, the choice of elements in the other sample.
2. The sample size n_1 and n_2 are sufficiently large for the central limit theorem to apply. (That is, $\mathbf{n_1}\geq 30$ and $\mathbf{n_2}\geq 30$.)
   (即兩樣本是獨立地抽自兩獨立母體，且$n_1\geq 30$ and $n_2\geq 30$)
3. 若有一 $n_i<30$, 則需假設兩個獨立群體服從**常態母體**

Interval Estimation of $(\mu_1-\mu_2)$:Two Independent Samples

- **例 1: 請參考課本354頁 Example8.3**

  Independent random samples of accounting professors and information systems (IS) professors were asked to provide the number of hours they spend in preparation for each class. The sample of **321 IS professors** had a mean time of 3.01 preparation hours, and the sample of **94 accounting professors** has a mean of 2.88 hours. From a similar past studies the population standard deviation for the IS professors is assumed to be 1.09, and, similarly, the population standard deviation for the accounting professors is 1.01. Denoting the population mean for IS professors by μ_x and the population mean for accounting professors by μ_y, find a 95% confidence interval for $(\mu_x - \mu_y)$.

Interval Estimation of $(\mu_1-\mu_2)$:Two Independent Samples

- **[Ans]**

$$\mu_x - \mu_y = (\bar{x} - \bar{y}) \pm Z\alpha_{/2}\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

with

$$n_x = 321 \quad \bar{x} = 3.01 \quad \sigma_x = 1.09$$
$$n_y = 94 \quad \bar{y} = 2.88 \quad \sigma_y = 1.01$$

and

$$Z\alpha_{/2} = Z_{0.025} = 1.96$$

$$\mu_x - \mu_y = (3.01 - 2.88) \pm 1.96\sqrt{\frac{(1.09)^2}{321} + \frac{(1.01)^2}{94}}$$

or

$$-0.11 < \mu_x - \mu_y < 0.37$$

This interval includes zero, indicating no evidence shows that the two population means are different.

Practical Interpretation of a Confidence Interval for $(\theta_1 - \theta_2)$

- **Let (LCL, UCL) represent a $(1-\alpha)100\%$ confidence interval for $(\theta_1 - \theta_2)$.**

1) If **LCL > 0** and **UCL > 0**, conclude $\boldsymbol{\theta_1 > \theta_2}$

2) If **LCL < 0** and **UCL < 0**, conclude $\boldsymbol{\theta_1 < \theta_2}$

3) If **LCL < 0** and **UCL > 0**, (i.e., the interval includes 0), conclude **no evidence of a difference between $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$**

Interval Estimation of $(\mu_1-\mu_2)$:Two Independent Samples

- **Case 2: When σ_1^2 and σ_2^2 are unknown, but assume $\sigma_1^2 = \sigma_2^2$.**
- **(1-α)100% Confidence Interval for $\mu_1 - \mu_2$ using Independent Samples**

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{s_p^2(\frac{1}{n_1}+\frac{1}{n_2})} \quad \text{where} \quad s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

($\mathbf{s_p^2}$ represents the **pooled variances**合併之變異數)

and the value of $t_{\alpha/2}$ is based on (n_1+n_2-2) degrees of freedom of a t-distribution.

- **Assumptions:**
  - Both of the populations from which the samples are selected have relative frequency distributions that are approximately normal.
  - The variance $\boldsymbol{\sigma_1^2}$ and $\boldsymbol{\sigma_2^2}$ of the two populations are equal.
  - The random samples are selected in an **independent** manners from the two populations.

  (即樣本是隨機抽自兩個**獨立**之**常態**群體，且 $\boldsymbol{\sigma_1^2 = \sigma_2^2}$)

Interval Estimation of $(\mu_1-\mu_2)$:Two Independent Samples

- **例 2: 請參考課本356頁 Example8.4**

The residents of City A complain that the local government charges for the removal of additional household rubbish in their city are higher than those charged by city B. City A council's refuse manager has agreed to study the problem and to indicate the complaints were reasonable. Independent random samples of the amounts paid by residents for household waste removal in each of two cities over the last three months were obtained. These amounts were as follows:

City A	100	125	135	128	140	142	128	137	156	142
City B	95	87	100	75	110	105	85	95		

Assuming an equal population variance, find a 95% confidence interval for the difference in the mean costs of speeding tickets in these two cities.

Interval Estimation of $(\mu_1-\mu_2)$:Two Independent Samples

- **[Ans] (1/2)**

Let the X population be City A and the Y population be city B. First, we get

$$n_x = 10 \qquad \bar{x} = 133.30 \qquad s_x^2 = 218.0111$$
$$n_y = 8 \qquad \bar{y} = 94.00 \qquad s_y^2 = 129.4286$$

The pooled sample variance is

$$s_p^2 = \frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x+n_y-2} = \frac{(10-1)(218.011)+(8-1)(129.4286)}{10+8-2}$$

$$= 179.2562$$

and

$$(\bar{x}-\bar{y}) = (133.30-94.00) = 39.30$$

The degrees of freedom result is $n_x+n_y-2 = 16$ and $t_{(16,0.0.25)} = 2.12$.

Interval Estimation of $(\mu_1-\mu_2)$:Two Independent Samples

- **[Ans] (2/2)**

$$(\bar{x}-\bar{y})-t_{16,0.025}\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}<\mu_x-\mu_y<(\bar{x}-\bar{y})+t_{16,0.025}\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}$$

$$39.3-(2.12)\sqrt{\frac{179.2562}{10}+\frac{179.2562}{8}}<\mu_x-\mu_y<39.3+(2.12)\sqrt{\frac{179.2562}{10}+\frac{179.2562}{8}}$$

$$25.84<\mu_x-\mu_y<52.76$$

The cost of refuse collection in city A is higher than that of city B, with a difference varying from as little as \$25.84 or as much as \$52.76.

Interval Estimation of $(\mu_1-\mu_2)$:Two Independent Samples

- **Case 3: When σ_1^2 and σ_2^2 are unknown, but assume $\sigma_1^2 \neq \sigma_2^2$**
- **(1-α)100% Confidence Interval for $\mu_1 - \mu_2$ using Independent Samples**

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

and the value of $t_{\alpha/2}$ is based on (ν) degrees of freedom.

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

- **Assumptions:**
  - Both of the populations from which the samples are selected have relative frequency distributions that are approximately **normal**.
  - The random samples are selected in **independent** manner from the two populations.

  (即兩樣本是獨立地抽自兩個常態母體)

Interval Estimation of $(\mu_1 - \mu_2)$:Two Independent Samples

- **例 3: 請參考課本358頁 Example8.5**

The Stryker accounting firm conducted a random sample of the accounts payable for the east and the west offices of Amalgamated Distributors. From these two independent samples the company wanted to estimate the difference between the population mean values of the payables. The sample statistics obtained were as follows:

	East Office Population X	West Office Population Y
Sample mean	$290	$250
Sample size	16	11
Sample standard deviation	15	50

We do not assume that the unknown population variances are equal. Estimate the difference between the mean values of the payables for the two offices. Use a 95% confidence level.

Interval Estimation of $(\mu_1-\mu_2)$:Two Independent Samples

- **[Ans] (1/2)**

$$(\bar{x}-\bar{y})-t_{11,0.025}\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}<\mu_x-\mu_y<(\bar{x}-\bar{y})+t_{11,0.025}\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}$$

then, we calculate the degrees of freedom.

$$\upsilon=\frac{\left[\left(\frac{s_x^2}{n_x}\right)^2+\left(\frac{s_y^2}{n_y}\right)^2\right]^2}{\frac{\left(\frac{s_x^2}{n_x}\right)^2}{n_x-1}+\frac{\left(\frac{s_y^2}{n_y}\right)^2}{n_y-1}}=\frac{\left[\frac{225}{16}+2500/11\right]^2}{\frac{\left(\frac{225}{16}\right)^2}{15}+\frac{\left(\frac{2500}{11}\right)^2}{10}}\approx 11$$

Interval Estimation of $(\mu_1-\mu_2)$:Two Independent Samples

- **[Ans] (2/2)**

$$(\bar{x}-\bar{y})-t_{11,0.025}\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}<\mu_x-\mu_y<(\bar{x}-\bar{y})+t_{11,0.025}\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}$$

$$(290-250)-34.19<\mu_x-\mu_y<(290-250)+34.19$$

$$5.81<\mu_x-\mu_y<74.19$$

In the long run, the mean accounts payable for the east office exceeds the mean accounts payable for the west office by as little as 5.81 or by as much as 74.19.

兩獨立母體比率差之估計
(Estimation of the Difference Between Two Population Proportions)

Interval Estimation of $(P_1 - P_2)$

- **$(1-\alpha)100\%$ Confidence Interval for $(P_1 - P_2)$**

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sigma_{(\hat{p}_1-\hat{p}_2)} \approx (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

  - where $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions of observations with the characteristic of interest.

- **Note:**
  - We have followed the usual procedure of substituting the sample value $\hat{\mathbf{p}}_1$, $\hat{\mathbf{q}}_1$, $\hat{\mathbf{p}}_2$, and $\hat{\mathbf{q}}_2$ for the corresponding population values required for $\boldsymbol{\sigma}_{(\hat{\mathbf{p}}_1-\hat{\mathbf{p}}_2)}$.

- **Assumptions:**
  - The samples are sufficiently large that the approximation is valid. As a general rule of thumb, we will require that $\mathbf{n_1\hat{p}_1} \geq 5$, $\mathbf{n_1\hat{q}_1} \geq 5$, $\mathbf{n_2\hat{p}_2} \geq 5$, and $\mathbf{n_2\hat{q}_2} \geq 5$.

Interval Estimation of $(P_1 - P_2)$

- **例 1: 請參考課本362頁 Example8.6**

  During a presidential election year many forecasts are made to determine how voters perceive a particular candidate. In a random sample of 120 registered voters in precinct(管理區) A ,107 indicated that they supported the candidate in question. In an independent random sample of 141 registered voters in precinct B, only 73 indicated support for the same candidate. If the respective population proportions are denoted P_A and P_B, find a 95% confidence interval for the population difference, $(P_A - P_B)$.

Interval Estimation of $(P_1 - P_2)$

- **[Ans]**

From the sample information it follow that

$$n_A = 120 \quad and \quad \hat{p}_A = \frac{107}{120} = 0.892; \quad n_B = 141 \quad and \quad \hat{p}_B = \frac{73}{141} = 0.518$$

For a 95% confidence interval is , therefore,

$$(0.892 - 0.518) - 1.96\sqrt{\frac{(0.892)(0.108)}{120} + \frac{(0.518)(0.482)}{141}} < P_A - P_B$$

$$< (0.892 - 0.518) + 1.96\sqrt{\frac{(0.892)(0.108)}{120} + \frac{(0.518)(0.482)}{141}}$$

or $$0.274 < P_A - P_B < 0.473$$

The fact that zero is well outside this interval suggests that there is a difference in the population proportions of registered voters in precinct A and precinct B who support this presidential candidate. In the ling run the difference is estimated to be as little as 27.4% or as high as 47.3%.

本單元結束