Chapter 9 One- and Two-Sample Estimation Problems

吳明龍 助理教授
成大資訊系/醫資所
minglong.wu@csie.ncku.edu.tw
辦公室:雲平大樓東棟 415 室

9.8 Two Samples: Estimating the Difference Between Two Means

- Two populations: $(\mu_1, \sigma_1), (\mu_2, \sigma_2)$
- The statictics $\overline{X_1} - \overline{X_2}$ is a point estimator for $\mu_1 - \mu_2$

The sampling distribution of $\overline{X_1} - \overline{X_2}$ will be approximated normally

with mean $\mu_{\overline{X_1}-\overline{X_2}} = \mu_1 - \mu_2$ and standard deviation $\sigma_{\overline{X_1}-\overline{X_2}} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$

$$\Rightarrow \text{standard normal variable } Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

$$P\left(-z_{\alpha/2} < \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < z_{\alpha/2}\right) = 1 - \alpha$$

$\mu_{\bar{x}_1} = \mu_1$; $\sigma_{\bar{x}_1}^2 = \frac{\sigma_1^2}{n_1}$

$\mu_{\bar{x}_2} = \mu_2$; $\sigma_{\bar{x}_2}^2 = \frac{\sigma_2^2}{n_2}$

$\mu_{a_1\bar{x}_1 + a_2\bar{x}_2} = a_1\mu_1 + a_2\mu_2$

$\sigma^2_{a_1\bar{x}_1 + a_2\bar{x}_2} = a_1^2\sigma_1^2 + a_2^2\sigma_2^2$

- Confidence interval

A $(1-\alpha)100\%$ confidence interval for mean $\mu_1 - \mu_2$ is given by

$$(\overline{x_1} - \overline{x_2}) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{x_1} - \overline{x_2}) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Two Samples: Estimating the Difference Between Two Means

- Example 9.10: An experiment was conducted in which two types of engines, *A* and *B*, were compared. Gas mileage in miles per gallon was measured.
  - Fifty experiments were conducted using engine type *A* and 75 experiments were done for engine type *B*.
  - The average gas mileage for engine *A* was 36 miles per gallon and the average for machine *B* was 42 miles per gallon.
  - Find a 96% confidence interval on $\mu_B - \mu_A$, where μ_B and μ_A are population mean gas mileage for machines *B* and *A*, respectively.
  - Assume that the population standard deviations are 6 and 8 for machines *A* and *B*, respectively.
  - **Solution**

Two Samples: Estimating the Difference Between Two Means

- Example 9.10: An experiment was conducted in which two types of engines, *A* and *B*, were compared. Gas mileage in miles per gallon was measured.
  - Fifty experiments were conducted using engine type *A* and 75 experiments were done for engine type *B*.
  - The average gas mileage for engine *A* was 36 miles per gallon and the average for machine *B* was 42 miles per gallon.
  - Find a 96% confidence interval on μ_B - μ_A , where μ_B and μ_A are population mean gas mileage for machines *B* and *A*, respectively.
  - Assume that the population standard deviations are 6 and 8 for machines *A* and *B*, respectively.
  - **Solution**

$$(\overline{x_1} - \overline{x_2}) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{x_1} - \overline{x_2}) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(42 - 36) - 2.05\sqrt{\frac{64}{75} + \frac{36}{50}} < \mu_1 - \mu_2 < (42 - 36) + 2.05\sqrt{\frac{64}{75} + \frac{36}{50}}$$

$$\therefore 3.43 < \mu_1 - \mu_2 < 8.57$$

Two Samples: Estimating the Difference Between Two Means with Unknown σ

- Variance unknown: If σ_1^2 and σ_2^2 are unknown, but $\sigma_1^2 = \sigma_2^2 = \sigma^2$, we obtain

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2[(1/n_1) + (1/n_2)]}}$$

$$\text{Two chi-squared random variables: } \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \text{ and } \frac{(n_2 - 1)S_2^2}{\sigma_2^2}$$

$$\text{Their sum: } V = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} + \frac{(n_2 - 1)S_2^2}{\sigma_2^2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2}$$

$$\text{has a chi-squared distribution with } v = n_1 + n_2 - 2 \text{ degrees of freedom}$$

Two Samples: Estimating the Difference Between Two Means with Unknown σ

- From theorem 8.5 : $T = \dfrac{Z}{\sqrt{V/v}}$

$$\therefore T = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2[(1/n_1) + (1/n_2)]}} \Bigg/ \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2(n_1 + n_2 - 2)}}$$

- Pooled estimator of the unknown common variance σ^2 :

$$S_P^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$$\therefore T = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{S_P\sqrt{(1/n_1) + (1/n_2)}}$$

Two Samples: Estimating the Difference Between Two Means with Unknown σ

- $P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha$

$$P\left[-t_{\alpha/2} < \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{S_p \sqrt{(1/n_1) + (1/n_2)}} < t_{\alpha/2} \right] = 1 - \alpha$$

- Confidence interval for $\mu_1 - \mu_2; \sigma_1^2 = \sigma_2^2$ but unknown :

$$(\overline{x_1} - \overline{x_2}) - t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\overline{x_1} - \overline{x_2}) + t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

s_p : the pooled estimate of the population standard deviation

$t_{\alpha/2}$: the t - value with $v = n_1 + n_2 - 2$ degrees of freedom

Two Samples: Estimating the Difference Between Two Means

- Example 9.11: Two independent sampling stations were chosen for the study of acid mine pollution.
  - For 12 monthly samples collected at the downstream station the species diversity index had a mean value $\bar{x}_1$ = 3.11 and a standard deviation s_1 = 0.771, while 10 monthly samples collected at the upstream station the species diversity index had a mean value $\bar{x}_2$ = 2.04 and a standard deviation s_2 = 0.448.
  - Find a 90% confidence interval for the difference between the population means for the two locations, assuming that the population are approximately normally distributed with equal variances.
  - **Solution**

$$s_P^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(12-1)0.771^2 + (10-1)0.448^2}{12+10-2} = 0.417 \Rightarrow S_P = 0.646$$

$$(3.11-2.04) - t_{0.05}(0.646)\sqrt{\tfrac{1}{12}+\tfrac{1}{10}} < \mu_1 - \mu_2 < (3.11-2.04) + t_{0.05}(0.646)\sqrt{\tfrac{1}{12}+\tfrac{1}{10}}$$

$$\therefore 0.593 < \mu_1 - \mu_2 < 1.547$$

Two Samples: Estimating the Difference Between Two Means with Unequal Variances

- Unequal Variances $\sigma 1 \neq \sigma 2$

$$T' = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{(S_1^2 / n_1) + (S_2^2 / n_2)}}$$

$$v = \frac{(s_1^2 / n_1 + s_2^2 / n_2)^2}{[(s_1^2 / n_1)^2 / (n_1 - 1)] + [(s_2^2 / n_2)^2 / (n_2 - 1)]} \text{ (degrees of freedom)}$$

四捨五入

$$P(-t_{\alpha/2} < T' < t_{\alpha/2}) \approx 1 - \alpha$$

- Confidence interval for $\mu_1 - \mu_2; \sigma_1^2 \neq \sigma_2^2$ and unknown :

$$(\overline{x_1} - \overline{x_2}) - t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{x_1} - \overline{x_2}) + t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Two Samples: Estimating the Difference Between Two Means with Unequal Variances

- Example 9.12: Orthophosphorus is measured in milligrams per liter. 15 samples were collected from station 1 had an average zinc content of 3.84 milligrams per liter and a standard deviation of 3.07 milligrams per liter, while the 12 samples from station 2 had an average zinc content of 1.49 milligrams per liter and a standard deviation of 0.80 milligrams per liter. Find a 95% confidence interval for the difference in the true average zinc contents at theses two stations, assuming that the observations came from normal population with different variance.
  - Solution

$$v = \frac{(3.07^2/15 + 0.80^2/12)^2}{[(3.07^2/15)^2/14] + [(0.80^2/12)^2/11]} = 16.3 \approx 16$$

For 95% confidence interval $\Rightarrow t_{0.025} = 2.120$ for $v = 16$

$$(3.84 - 1.49) - (2.120)\sqrt{\frac{3.07^2}{15} + \frac{0.80^2}{12}} < \mu_1 - \mu_2 < (3.84 - 1.49) + (2.120)\sqrt{\frac{3.07^2}{15} + \frac{0.80^2}{12}}$$

$$0.60 < \mu_1 - \mu_2 < 4.10$$

9.9 Paired Observations

- Consider that the samples are not independent and the variances of the two populations are not necessarily equal.
- If $\bar{d}$ and s_d are the mean and standard deviation of the normally distributed differences of n random pairs of measurements, a (1-a)100% confidence interval for $\mu_D = \mu_1 - \mu_2$ is

$$\bar{d} - t_{\alpha/2}\frac{s_d}{\sqrt{n}} < \mu_D < \bar{d} + t_{\alpha/2}\frac{s_d}{\sqrt{n}}$$

d bar：the mean of sample before and after difference

where $t_{\alpha/2}$ is the t-value with $v = n - 1$ degrees of freedom, leaving an area of $\alpha/2$ to the right.

- Example 9.13: For a study of dioxin, find a 95% confidence interval for $\mu_1 - \mu_2$, where μ_1 and μ_2 represent the true mean TCDD in plasma and in fat tissue, respectively. Assume the distribution of the differences to be approximately normal.

Paired Observations

Veteran	TCDD levels In plasma	TCDD levels In fat tissue	d_i	Veteran	TCDD levels In plasma	TCDD levels In fat tissue	d_i
1	2.5	4.9	−2.4	11	6.9	7.0	−0.1
2	3.1	5.9	−2.8	12	3.3	2.9	0.4
3	2.1	4.4	−2.3	13	4.6	4.6	0.0
4	3.5	6.9	−3.4	14	1.6	1.4	0.2
5	3.1	7.0	−3.9	15	7.2	7.7	−0.5
6	1.8	4.2	−2.4	16	1.8	1.1	0.7
7	6.0	10.0	−4.0	17	20.0	11.0	9.0
8	3.0	5.5	−2.5	18	2.0	2.5	−0.5
9	36.0	41.0	−5.0	19	2.5	2.3	0.2
10	4.7	4.4	0.3	20	4.1	2.5	1.6

Source: Schecter, A. et al. "Partitioning of 2, 3, 7, 8-chlorinated dibenzo-p-dioxins and dibenzofurans between adipose tissue and plasma lipid of 20 Massachusetts Vietnam veterans". *Chemosphere*, Vol. 20, Nos. 7–9, 1990, pp. 954–955 (Tables I and II).

– Solution

$$\bar{d} = -0.87, t_{0.025} = 2.093\,(v = 20-1 = 19), s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{168.4220}{19}} = 2.9773$$

$$-0.87 - (2.093)\frac{2.9773}{\sqrt{20}} < \mu_D < -0.87 + (2.093)\frac{2.9773}{\sqrt{20}}$$

$$\Rightarrow -2.2634 < \mu_D < 0.5234 \therefore \text{no sigificant difference}$$

9.10 Single Sample: Estimating a Proportion

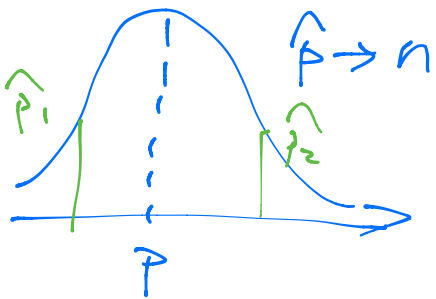
- A point estmator of the proportion p in a binomial experiment : $\hat{P} = X / n$

$$\mu_{\hat{p}} = E(\hat{P}) = E\left[\frac{X}{n}\right] = \frac{np}{n} = p$$

p head : unbiased estimation

$$\sigma^2_{\hat{p}} = \sigma^2_{X/n} = \frac{\sigma_x^2}{n^2} = \frac{npq}{n^2} = \frac{pq}{n}$$

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha, Z = \frac{\hat{P}-p}{\sqrt{pq/n}}$$



$$P(-z_{\alpha/2} < \frac{\hat{P}-p}{\sqrt{pq/n}} < z_{\alpha/2}) = 1-\alpha$$

$$P(\hat{P} - z_{\alpha/2}\sqrt{\frac{pq}{n}} < p < \hat{P} + z_{\alpha/2}\sqrt{\frac{pq}{n}}) = 1-\alpha$$

$$P(\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}) \approx 1-\alpha \text{ (when } n \text{ is large } \hat{p} = x/n \approx p)$$

error : p head - p

Single Sample: Estimating a Proportion

- If $\hat{p}$ is the proportion of successes in a random sample of size *n*, and $\hat{q} = 1 - \hat{p}$ an approximate (1-α)100% confidence interval for the binomial parameter *p* is given by is

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

where $z_{\alpha/2}$ is the *z*-value with leaving an area of $\alpha/2$ to the right.

Single Sample: Estimating a Proportion

- If $\hat{p}$ is the proportion of successes in a random sample of size *n*, and $\hat{q} = 1 - \hat{p}$ an approximate (1-α)100% confidence interval for the binomial parameter *p* is given by is

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

  where $z_{\alpha/2}$ is the *z*-value leaving an area of $\alpha/2$ to the right.

- $n\hat{p}, n\hat{q} \geq 5.$

Single Sample: Estimating a Proportion

- Ex 9.14: In a random of n = 500 families owning television sets in the city of Hamilton, Canada, it is found that x = 340 subscribed to HBO. Find a 95% confidence interval for the actual proportion of families in the city who subscribe to HBO.
  - **Solution**

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$\frac{340}{500} - 1.96\sqrt{\frac{0.68 \times 0.32}{500}} < p < \frac{340}{500} + 1.96\sqrt{\frac{0.68 \times 0.32}{500}}$$

$$0.64 < p < 0.72$$

Single Sample: Estimating a Proportion

- Theorem 9.3: If $\hat{p}$ is used as an estimate of *p*, we can be (1 - α)100% confident that the error will not exceed $z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$.

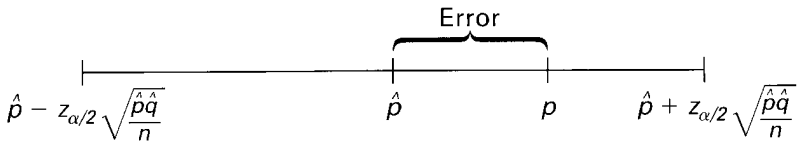


Figure 9.6 Error in estimating p by $\hat{p}$.

- Theorem 9.4: If $\hat{p}$ is used as an estimate of *p*, we can be (1 - α)100% confident that the error will be less than a specified amount *e* when the sample size is approximately $n = \frac{z_{\alpha/2}^2 \hat{p}\hat{q}}{e^2}$.

Single Sample: Estimating a Proportion

- Example 9.15: How large a sample is required in Example 9.14 if we want to be 95% confident that our estimate of *p* is within 0.02?
  - **Solution**

$$n = \frac{z_{\alpha/2}^2 \hat{p}\hat{q}}{e^2} = \frac{(1.96)^2(0.68)(0.32)}{(0.02)^2} = 2090.$$

Single Sample: Estimating a Proportion

- Upper bound of n is $\frac{z_{\alpha/2}^2}{4e^2}$, $\because \hat{p}\hat{q} = \hat{p}(1-\hat{p}) = \frac{1}{4} - (\hat{p} - \frac{1}{2})^2$
- Theorem 9.5: If $\hat{p}$ is used as an estimate of *p*, we can be at least (1 - α)100% confident that the error will not exceed a specified amount *e* when the sample size is approximately $n = \frac{z_{\alpha/2}^2}{4e^2}$.
- Example 9.16: How large a sample is required in Example 9.14 if we want to be at least 95% confident that our estimate of *p* is within 0.02?
  - **Solution**

$$n = \frac{z_{\alpha/2}^2}{4e^2} = \frac{(1.96)^2}{4 \cdot (0.02)^2} = 2401.$$

9.11 Two Samples: Estimating the Difference Between Two Proportions

- p_1 might be the proportion of smokers with lung cancer and p_2 the proportion of non-smokers with lung cancer.
- The sampling distribution of $\hat{P}_1 - \hat{P}_2$ will be approximated normally

with mean $\mu_{\hat{P}_1-\hat{P}_2} = p_1 - p_2$ and variance $\sigma^2_{\hat{P}_1-\hat{P}_2} = \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}$

$\Rightarrow$ standard normal variable $Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{p_1q_1/n_1 + p_2q_2/n_2}}$

$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$

$$P\left[-z_{\alpha/2} < \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{p_1q_1/n_1 + p_2q_2/n_2}} < z_{\alpha/2}\right] = 1 - \alpha$$

Two Samples: Estimating the Difference Between Two Proportions

- Large-sample confidence interval for p_1 - p_2:
  If $\hat{p}_1$ and $\hat{p}_2$ are the proportion of successes in a random sample of size n_1 and n_2, $\hat{q}_1 = 1 - \hat{p}_1$ and $\hat{q}_2 = 1 - \hat{p}_2$, an approximate (1-$\alpha$)100% confidence interval for the difference of two binomial parameters p_1 - p_2 is given by

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

  where $z_{\alpha/2}$ is the z-value leaving an area of $\alpha/2$ to the right.

- $n_1\hat{p}_1, n_1\hat{q}_1, n_2\hat{p}_2, n_2\hat{q}_2 \geq 5$

Two Samples: Estimating the Difference Between Two Proportions

- Example 9.17: A certain change in a process for manufacture of component parts is being considered.
  - Sample are taken using both the existing and new procedure so as to determine if the new process results in an improvement.
  - If 75 of 1500 items from the existing procedure were found to be defective and 80 of 2000 items from the new procedure were found to be defective.
  - Find a 90% confidence interval for the true difference in the fraction of defectives between the existing and the new process.
  - **Solution**

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$\hat{p}_1 = 75/1500 = 0.05,\ \hat{p}_2 = 80/2000 = 0.04,\ \hat{p}_1 - \hat{p}_2 = 0.05 - 0.04 = 0.01$$

$$0.01 - 1.645\sqrt{\frac{0.05 \cdot 0.95}{1500} + \frac{0.04 \cdot 0.96}{2000}} < p_1 - p_2 < 0.01 + 1.645\sqrt{\frac{0.05 \cdot 0.95}{1500} + \frac{0.04 \cdot 0.96}{2000}}$$

$$-0.0017 < p_1 - p_2 < 0.0217$$

無法判斷哪個製程技術較佳（因為p1-p2有一部分<0）

若p1 - p2 > 0，則可判斷第二個製程技術優於第一個製程技術

Exercise

- 9.35, 9.43, 9.53, 9.59