Ch 9 實習

本課程回顧:

統計是一門用科學化的方式，處理不確定性的問題

- **Stage 1: 確定假設 (C10, C11)**
  - Pockmon Go下載率會超過三個月 (C10)
  - Ho: u ≤3 month
  - H1:u >3 month
- **Stage 2: 確定母體/樣本的分配 （決定用 Z 、t、 F ）**
  - -sampling distribution (C9)
  - 母體未知，樣本已知(C10)
  - 不同題目，會有不同公式
- **Stage 3: 選擇critical point/p-value/CI**
  - ~(查表, 需要信心水準）
- **Stage 4: 拒絕 or 不拒絕 hypothesis**

學習目標

1. Sampling distribution & 中央極限定理(CLT)
2. Sampling求解三步驟

Example:

1. 一般狀況
2. Proposition
3. Difference (X1-X2)
4. 反推$\bar{X}$信賴區間

1. Sampling distribution and Central Limit Theorem…

Sampling distribution特色：

1. 母體為常態分配：

–If the population is normal, then $\bar{X}$ is normally distributed for all values of n. $\mu_{\bar{x}} = \mu \quad \sigma_{\bar{x}}^2 = \sigma^2 / n \quad and \quad \sigma_{\bar{x}} = \sigma / \sqrt{n}$

2. 母體不是常態分配（依據中央極限定理）：

–If the population is non-normal, then $\bar{X}$ is approximately normal only for larger values of n (n>30).

$$\mu_{\bar{x}} = \mu \quad \sigma_{\bar{x}}^2 = \sigma^2 / n \quad and \quad \sigma_{\bar{x}} = \sigma / \sqrt{n}$$

Example:

- 一般狀況
- Proposition (民意調查)
- Difference (玩candy crush人數>Pokémon人數)
- 反推求 $\overline{X}$的範圍

2. Sampling 求解三步驟

- Step 1: 將題目數量化/公式化
- Step 2:透過母體與sampling關係，求解Sampling的 $u_{\bar{x}}$ 與$\sigma_{\bar{X}}$

  $-\mu, \sigma \rightarrow u_{\bar{x}}, \sigma_{\bar{X}}$ (C9)

  $-u_{\bar{x}}$ 與$\sigma_{\bar{X}} \rightarrow \mu, \sigma$ (C10) 比較符合現實情況

- Step 3:計算題目所求 P($\overline{X_n} \geq 10$)?

  (變成常態了，就可以用標準常態來算)

- Step 4:反推u的範圍

反求X or u 的信賴區間

類型	母體 (已知)	樣本 (未知)	$Z=\frac{\bar{X}-u}{\sigma^{\square}}$
一般型	μ σ^2	$\mu_{\bar{x}}=\mu$ $\sigma^2_{\bar{x}}=\frac{\sigma^2}{n}$	$Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$
Proposition	μ=np $\sigma^2=np(1-p)$	$E(\hat{P})=p$ $V(\hat{P})=\sigma^2_{\hat{p}}=\frac{p(1-p)}{n}$	$Z=\frac{\hat{P}-p}{\sqrt{p(1-p)/n}}$
Difference (X1-X2)	u_1-u_2 $\sigma^2{}_{u_1-u_2}$	$\mu_{\bar{x}_1-\bar{x}_2}=\mu_1-\mu_2$ $\sigma_{\bar{x}_1-\bar{x}_2}=\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$	$Z=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$

類型一：一般型（類似1,2,3)

- The amount of time the university professors devote to their jobs per week is normally distributed with a mean of 52 hours and a standard deviation of 6 hours
  - –a. What is the probability that a professor works for more than 60 hours per week?

- Step1:把題目公式化
- P(X>60)
- Step2:求樣本的u,sd
- $u = 52,\ \sigma_u = 6$ (這裡是母體）

- Step3:正規化求解

$$P(X > 60) = P\left(\frac{X-\mu}{\sigma} > \frac{60-52}{6}\right) = P(Z > 1.33) = 1 - P(Z < 1.33) = 1 - .9082 = .0918$$

類型一：一般型

- The amount of time the university professors devote to their jobs per week is normally distributed with a mean of 52 hours and a standard deviation of 6 hours,.
  - –a. If they random chose four professors, what is the probability that a professor works for more than 60 hours per week?

- Step1:把題目公式化
- P($\bar{X}$>60)
- Step2:求樣本的u,sd
- $u_{\bar{x}} = 52,\ \sigma_{\bar{x}} = 6/\sqrt{4}$ (樣本)
- Step3:正規化求解
- $p(\bar{X} > 60) = p\left(\frac{\bar{X}-u}{\frac{\sigma}{\sqrt{n}}} > \frac{60-52}{\frac{6}{2}}\right) = P(Z > 2.66)$ =0.00383

類型二： Proportion(Proportion , Binomial)

- The parameter of interest for nominal data is the **proportion of times** a particular outcome (success) occurs.
- To estimate the population proportion 'p' we use the sample proportion.

母體參數 樣本統計量 The number of successes

$$\text{The estimate of } p = \hat{p} = \frac{x}{n}$$

類型二： Proportion

Using the laws of expected value and variance, we can determine the mean, variance, and standard deviation of $\hat{P}$.

(The standard deviation of $\hat{P}$ is called the ***standard error of the proportion***.)

$$E(\hat{P}) = p$$

$$V(\hat{P}) = \sigma^2_{\hat{p}} = \frac{p(1-p)}{n}$$

$$\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$$

為什麼是用P,不是$\widehat{p}$?

因為我們已知母體參數(parameter)，去推估樣本統計量 (Statistic)

*要搞清楚誰是p and $\widehat{p}$,才不會代錯

Sample proportions can be standardized to a standard normal distribution using this formulation:

$$Z = \frac{\hat{P} - p}{\sqrt{p(1-p)/n}}$$

類型二： Proportion (類似4,5,6)

- A psychologist believes that 80% of male drivers when lost continue to drive, hoping to find the location they seek rather than ask directions. To examine this belief, he took a random sample of 350 male drivers and asked each what they did when lost. If the belief is true, determine the probability that less than 75% said they continue driving.

Step1: 把題目公式化
$P(\hat{p} < 0.75)$

Step2:求樣本的u, sd

Step3:正規化求解

p=	p= 0.8
$\hat{p}$ =	$\hat{p}$ = 0.75
n=	n=350

類型二： Proportion

- Step1: 把題目公式化
- P($\hat{p}$ <0.75)
- Step2:求樣本的u, sd
- E($\hat{p}$)=p=0.8, $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.8(1-0.8)}{350}}$
- Step3:正規化求解

$$Z = \frac{\hat{P} - p}{\sqrt{p(1-p)/n}}$$

$$P(\hat{P} < .75) = P\left(\frac{\hat{P}-p}{\sqrt{p(1-p)/n}} < \frac{.75-.80}{\sqrt{(.80)(1-.80)/350}}\right) = P(Z < -2.34) = .0096$$

類型三： Difference of two means

The final sampling distribution introduced is that of the ***difference* between two sample means**. This requires:

➔ ***independent*** random samples be drawn from each of **two *normal*** populations

If this condition is met, then the sampling distribution of the ***difference*** between the two sample means, i.e. $\bar{X}_1 - \bar{X}_2$ will be normally distributed.

(note: if the two populations are **not** both normally distributed, but the sample sizes are "large" (>30), the distribution of $\bar{X}_1 - \bar{X}_2$ is *approximately* normal)

類型三：Difference of two means

Since the distribution of $\bar{X}_1 - \bar{X}_2$ is ***normal*** and has a

mean of $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$

and a standard deviation of $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

We can compute Z (standard normal random variable) in this way:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

類型三：Difference of two means (類似7,8)

- The manager of a restaurant believes that waiters and waitresses who introduce themselves by telling customers their names will get larger tips than those who don't. In fact, she claims that the average tip for the former group is 18% whereas that of the latter is only 15%.
- If tips are normally distributed with a standard deviation of 3 %, what is the probability that in a random sample of 10 tips recorded from waiters and waitresses who introduce themselves and 10 tips from waiters and waitresses who don't, the mean of the former will exceed that of the latter?

這題n<30, 能用常態分配嗎？Why?

可以，因為題目說兩個分配本身就是常態，所以兩樣本差異也會是常態

類型三：Difference of two means

- The manager of a restaurant believes that waiters and waitresses who introduce themselves by telling customers their names will get larger tips than those who don't. In fact, she claims that the average tip for the former group is 18% whereas that of the latter is only 15%.
- If tips are normally distributed with a standard deviation of 3 %, what is the probability that in a random sample of 10 tips recorded from waiters and waitresses who introduce themselves and 10 tips from waiters and waitresses who don't, the mean of the former will exceed that of the latter?

u1=　　n1=　　u1=0.18　　$\sigma_{u1} = \sigma_{u2} = 0.03$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

u2=　　n2=　　u2=0.15　　n1=10, n2=10

$\sigma_{u1} = \sigma_{u2}$

- Step1:把題目公式化
- P($\overline{X1} - \overline{X2}$>0)

- Step2:求樣本的u,sd

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- Step3:正規化求解

$$P(\bar{X}_1 - \bar{X}_2 > 0) = P\left(\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > \frac{0 - (18 - 15)}{\sqrt{\frac{3^2}{10} + \frac{3^2}{10}}}\right) = P(Z > -2.24) = 1 - P(Z < -2.24)$$

$$= 1 - .0125 = .9875$$

Example 4 (反推 sampling mean)

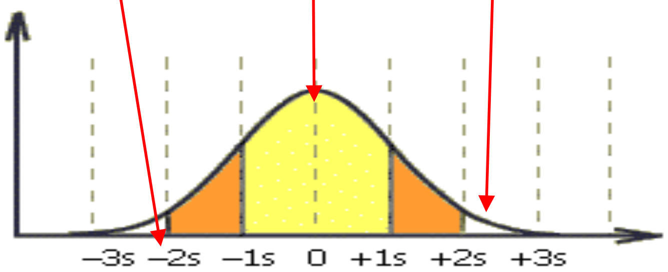
Here's another way of expressing the probability calculated from a sampling distribution.

$$P(-1.96 < Z < 1.96) = .95$$

Substituting the formula for the sampling distribution

$$P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} < 1.96\right) = .95$$

With a little algebra



$$P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = .95$$

$$P\left(u - Z_{\frac{a}{2}}\frac{\sigma}{\sqrt{n}} < \bar{X} < u + Z_{\frac{a}{2}}\frac{\sigma}{\sqrt{n}}\right)=1-a$$

a是什麼？
1-a是什麼？

a是type 1 error
1-a是信心水準

求 $\bar{x}$的信賴區間：u±幾倍標準差

Example 4 (反推 sampling mean)

Returning to the chapter-opening example where $\mu = 800, \sigma = 100$, and $n = 25$, we compute

$$P(800 - 1.96\frac{100}{\sqrt{25}} < \overline{X} < 800 + 1.96\frac{100}{\sqrt{25}}) = .95$$

or

$$P(760.8 < \overline{X} < 839.2) = .95$$

求 $\bar{x}$的信賴區間：u±幾倍標準差

This tells us that there is a 95% probability that a sample mean will fall between 760.8 and 839.2. Because the sample mean was computed to be \$750, we would have to conclude that the dean's claim is not supported by the statistic.

HW#6 (Proportion)

- 6. The Agency for Healthcare Research and Quality reports that medical errors are responsible for injury to 1 out of every 25 hospital patients in the United States. (Data extracted from M. Ozan-Rafferty, "Hospitals: Never Have a Never Event," The Gallup Management Journal, gmj. gallup.com, May 7, 2009.) These errors are tragic and expensive. Preventable health care–related errors cost an estimated $29 billion each year in the United States. Suppose that you select a sample of 100 U.S. hospital patients.
- a. What is the probability that the sample percentage reporting injury due to medical errors will be between 5% and 10%?
- b. The probability is 90% that the sample percentage will be within what symmetrical limits of the population percentage?
- c. The probability is 95% that the sample percentage will be within what symmetrical limits of the population percentage?

Step1: 把題目公式化
Step2:求樣本的u, sd
Step3:正規化求解

HW#6

- 6. The Agency for Healthcare Research and Quality reports that medical errors are responsible for injury to 1 out of every 25 hospital patients in the United States. (Data extracted from M. Ozan-Rafferty, "Hospitals: Never Have a Never Event," The Gallup Management Journal, gmj. gallup.com, May 7, 2009.) These errors are tragic and expensive. Preventable health care–related errors cost an estimated $29 billion each year in the United States. Suppose that you select a sample of 100 U.S. hospital patients.
- a. What is the probability that the sample percentage reporting injury due to medical errors will be between 5% and 10%?

Step1: 把題目公式化
P($0.05 < \hat{p} < 0.1$)

$$E(\hat{p}) = p = 0.04, \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.04(1-0.04)}{100}} = 0.0196$$

Step2:求樣本的u, sd
Step3:正規化求解

$$p\left(\frac{0.05-0.04}{0.0196}\right) < Z < p\left(\frac{0.1-0.04}{0.0196}\right) = p(0.5102 < Z < 3.0619)$$

HW#6

- Type equation here.6. The Agency for Healthcare Research and Quality reports that medical errors are responsible for injury to 1 out of every 25 hospital patients in the United States. (Data extracted from M. Ozan-Rafferty, "Hospitals: Never Have a Never Event," The Gallup Management Journal, gmj. gallup.com, May 7, 2009.) These errors are tragic and expensive. Preventable health care–related errors cost an estimated $29 billion each year in the United States. Suppose that you select a sample of 100 U.S. hospital patients.
- b. The probability is 90% that the sample percentage will be within what symmetrical limits of the population percentage?
- c. The probability is 95% that the sample percentage will be within what symmetrical limits of the population percentage?

$E(\hat{p})=p=0.04, \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.04(1-0.04)}{100}} = 0.0196$

求 $\bar{x}$的信賴區間：u±幾倍標準差

b. $P(A < \bar{X} < B) = P(0.0078 < \bar{X} < 0.0722)$
A: 0.04-1.645*0.0196=0.0078
B: 0.04+1.645*0.0196=0.0722

HW#7 (Difference of two means)

- 7. A factory worker (call him "Worker 1") has a productivity that is normally distributed, producing an average of 75 units per day, with a standard deviation of 20. Another worker (call him "Worker 2") also has a normally distributed productivity, with a mean of 65 units per day and a standard deviation of 21. Suppose both workers' productivities are independent of each other. What is the probability that during one week (5 working days) worker 1 will outproduce worker 2?
- Step1: 把題目公式化
- $P(\overline{X1} - \overline{X2} > 0)$
- Step2:求樣本的u, sd
- Step3:正規化求解

HW#7 (Difference of two means)

- 7. A factory worker (call him "Worker 1") has a productivity that is normally distributed, producing an average of 75 units per day, with a standard deviation of 20. Another worker (call him "Worker 2") also has a normally distributed productivity, with a mean of 65 units per day and a standard deviation of 21. Suppose both workers' productivities are independent of each other. What is the probability that during one week (5 working days) worker 1 will outproduce worker 2?
- Step1: 把題目公式化
- P($\overline{X1} - \overline{X2}$>0)
- Step2:求樣本的u, sd
- Step3:正規化求解

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

u1=75 $\sigma_{u1} = 20$

u2=65 $\sigma_{u1} = 21$

n1=n2=5

HW#7 (Difference of two means)

- 7. A factory worker (call him "Worker 1") has a productivity that is normally distributed, producing an average of 75 units per day, with a standard deviation of 20. Another worker (call him "Worker 2") also has a normally distributed productivity, with a mean of 65 units per day and a standard deviation of 21. Suppose both workers' productivities are independent of each other. What is the probability that during one week (5 working days) worker 1 will outproduce worker 2?
- Step3:正規化求解

$$u_1 - u_2 = 75 - 65$$

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\sigma_{\overline{X_1} - \overline{X2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{20^2}{5} + \frac{21^2}{5}} = 12.969$$

$$P(\overline{X_1} - \overline{X_2} > 0) = p(Z > \tfrac{0-10}{12.969}) = P(Z > -0.77106)$$

類型	母體 (已知)	樣本 (未知)	$Z=\frac{\bar{X}-u}{\sigma}$
一般型	μ σ^2	$\mu_{\bar{x}}=\mu$ $\sigma_{\bar{x}}^2=\frac{\sigma^2}{n}$	$Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$
Proposition	μ=np $\sigma^2=np(1-p)$	$E(\hat{P})=p$ $V(\hat{P})=\sigma_{\hat{p}}^2=\frac{p(1-p)}{n}$	$Z=\frac{\hat{P}-p}{\sqrt{p(1-p)/n}}$
Difference (X1-X2)	u_1-u_2 ${\sigma^2}_{u_1-u_2}$	$\mu_{\bar{x}_1-\bar{x}_2}=\mu_1-\mu_2$ $\sigma_{\bar{x}_1-\bar{x}_2}=\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$	$Z=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$