六.樣品分佈與估計
(Sampling Distribution and Estimation)
(Chapter 6)

劉仁沛教授
國立台灣大學農藝學研究所生物統計組
國立台灣大學流行病學與預防醫學研究所
國家衛生研究院生物統計與生物資訊組
jpliu@ntu.edu.tw

綱要

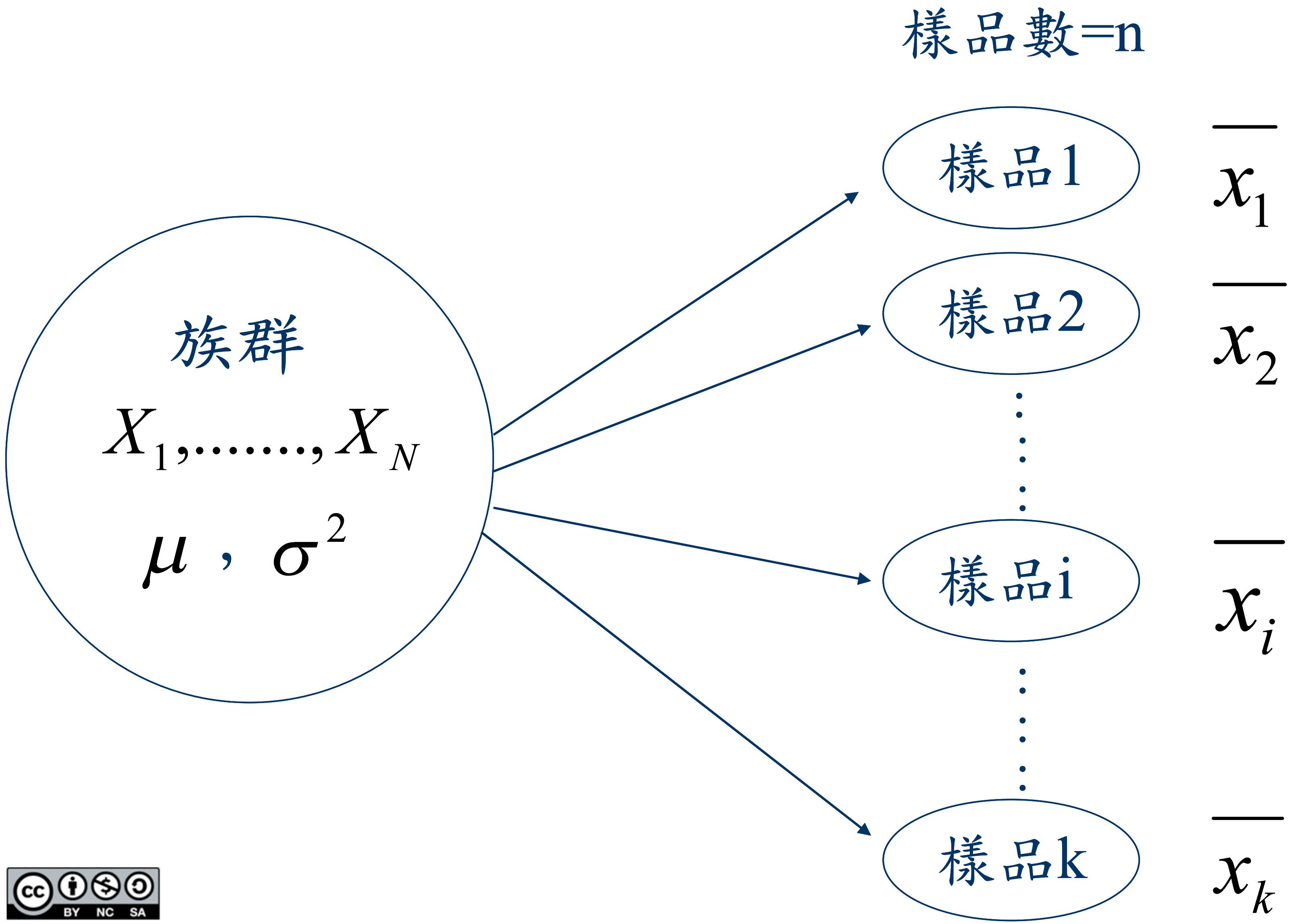
- 樣品統計量之分布(Sampling Distribution)
- 中央極限定理(Central Limit Theorem)
- 點估計(Point Estimation)
- 區間估計(Interval Estimation)

問題

1. 某品牌超大碗泡麵平均重量是否有足夠證據超過所宣稱500公克？
2. 新抗病害的品種其感染的比例低於5%
3. 治療憂鬱症新藥Hamilton憂鬱量表平均分數的下降
4. 某食品所含防腐劑平均含量低於政府所規定量
5. 瘦肉精零檢出是否保証無瘦肉精?

問題

- 自被射殺美國人權領袖金恩(Martin Luther King)屍體中所取出子彈碎片，是否發射自嫌犯James Earl Ray的步槍?
- 小孩DNA是否與某位宅男或某位敗犬女符合?
- 教授的論文是否抄襲自學生的報告?
- 某群基因的變異是否能預測某種疾病的發生或治療的效果(標靶治療)?


樣品數=n
族群
$X_1,\ldots\ldots,X_N$
μ，σ^2
樣品1
$\bar{x}_1$
樣品2
$\bar{x}_2$
樣品i
$\bar{x}_i$
樣品k
$\bar{x}_k$

歸還取樣n為2之樣品

族群

2，4，6

- 2，2　$\overline{x_1}=2$
- 2，4　$\overline{x_2}=3$
- 2，6　$\overline{x_3}=4$
- 4，2　$\overline{x_4}=3$
- 4，4　$\overline{x_5}=4$
- 4，6　$\overline{x_6}=5$
- 6，2　$\overline{x_7}=4$
- 6，4　$\overline{x_8}=5$
- 6，6　$\overline{x_9}=6$

樣品	1	2	3	4	5	6	7	8	9
	2	2	2	4	4	4	6	6	6
	2	4	6	2	4	6	2	4	6
平均	2	3	4	3	4	5	4	5	6

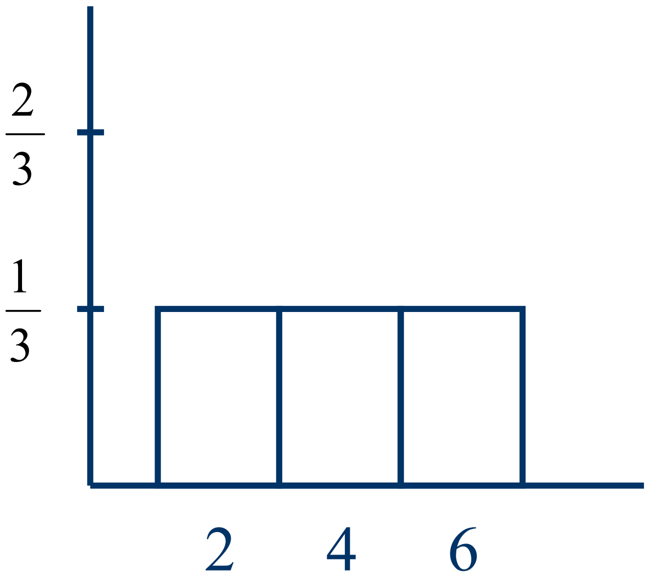
族群分佈

觀測值	機率
2	1/3
4	1/3
6	1/3

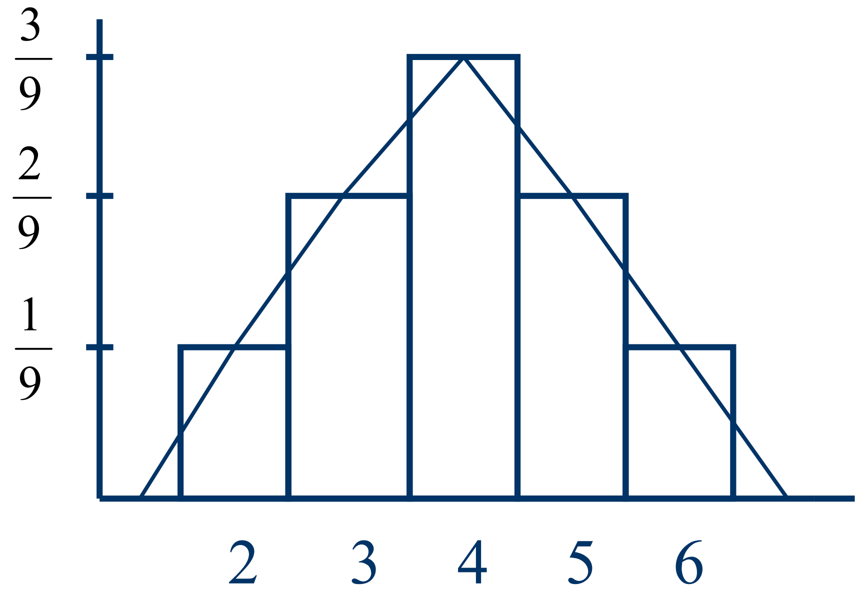
樣品平均值分佈

樣品平均	機率
2	1/9
3	2/9
4	3/9
5	2/9
6	1/9

族群分佈



樣品平均值分佈



族群平均值　　樣品平均值之平均值

觀測值 x	機率 $P(x)$	$x \cdot P(x)$
2	1/3	2/3
4	1/3	4/3
6	1/3	6/3
平均	$12/3=4=\mu_x$	

樣品平均值 $\bar{x}$	機率 $P(\bar{x})$	$\bar{x} \cdot P(\bar{x})$
2	1/9	2/9
3	2/9	6/9
4	3/9	12/9
5	2/9	10/9
6	1/9	6/9
平均		$36/9=4=\mu_{\bar{x}}$

族群平均值
=樣品平均值之平均值
此種性質稱為不偏性(Unbiasedness)

族群變方

觀測值 x	機率 $P(x)$	$(x-\mu_x)^2$	$(x-\mu_x)^2P(x)$
2	1/3	4	4/3
4	1/3	0	0/3
6	1/3	4	4/3

$8/3 = \sigma_x^2$

每次取樣所得樣品平均值均不同
必須估算平均值不確定性(Uncertainty)

樣品平均值之變方

樣品平均值 $(\bar{x})$	機率 $P(\bar{x})$	$(\bar{x}-\mu_{\bar{x}})^2$	$(\bar{x}-\mu_{\bar{x}})^2 \cdot P(\bar{x})$
2	1/9	4	4/9
3	2/9	1	2/9
4	3/9	0	0/9
5	2/9	1	2/9
6	1/9	4	4/9

$$\frac{12}{9}=\frac{4}{3}=\frac{8/3}{2}=\frac{\sigma_x^2}{n}$$

- 族群抽樣取得不同樣品其樣品平均值不同
- 樣品平均值之分佈稱為抽樣分佈(Sampling Distribution)
- 族群分佈=矩形分佈ÿ (2，4，6之機率均為1/3)
- 樣品平均值之分佈卻為對稱分佈，對稱於族群平均值
- 樣品平均值分佈之平均值=族群平均
- 樣品平均值分佈之變方為族群變方除以樣品觀測值數

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$\sigma_{\bar{x}}$ 即為樣品平均值的抽樣誤差

中央極限定理(Central Limit Theorem)

1. 當樣品觀測值數充分大時 (n≥30)
2. 樣品平均值之標準化值

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

的分佈會趨近於標準常態分佈

3. 不論母體觀測值分佈為何，上述結果均成立
4. 若ÿσ未知時，以樣品標準偏差S代之

$$Z' = \frac{\bar{x} - \mu}{S / \sqrt{n}}$$

之分佈亦趨近於標準常態分佈

中央極限定理之應用

問題：對某藥品有過敏的人約為10%，今隨機抽取650人，求至少有70人有過敏之機率？

二項分立隨機變數族群平均值為np變方為npq

其標準化值

$$Z = \frac{X - np}{\sqrt{npq}}$$

之分佈，當n大時，趨近於標準常態分佈

$X = 70 \quad n = 650 \quad P = 0.1$

$\mu = np = (650)(0.1) = 65$

$\sigma^2 = npq = (650)(0.1)(0.9) = 58.5$

標準化值

$$z = \frac{70-65}{\sqrt{58.5}} = 0.65$$

ÿ

$$\begin{aligned} P(X > 70) &= P(Z > 0.65) \\ &= 1 - P(Z < 0.65) \\ &= 1 - 0.7422 \\ &= 0.2578 \\ &= 25.78\% \end{aligned}$$

連續性矯正(Continuity Correction)

因二項分佈為分立隨機變數之分佈，而非連續隨機變數. 在n < 30，X < 5時以連續性的常態分佈. 求機率時會產生誤差，必須加以矯正

$$Z = \frac{(|X - np| - 0.5)}{\sqrt{npq}}$$

Jen-pei Liu, PhD

令X為二項分佈隨機變數 n=15 P=0.4 求 $P\left(2 \le Y \le 4\right)$

$$n \cdot p = (15)(0.4) = 6$$

$$npq = (15)(0.4)(0.6) = 3.6 \qquad \sqrt{npq} = \sqrt{3.6} = 1.9$$

$$P(a - 0.5 \le Y \le b + 0.5)$$

$$= P(2 - 0.5 \le Y \le 4 + 0.5) = P(1.5 \le Y \le 4.5)$$

$$= P\left(\frac{1.5 - 6}{1.9} \le \frac{Y - np}{\sqrt{npq}} \le \frac{4.5 - 6}{1.9}\right) = P(-2.37 \le Z \le -0.79)$$

$$= P(Z \le -0.79) - P(Z \le -2.37) = 0.2148 - 0.0089$$

$$= 0.2059$$

（表2值0.2121）

問題

題目：健康男性紅血球每立方公厘平均約為 5.4×10^{7} 個

今檢查一滴血均為 10^{-5} 公厘. 求紅血球在500至600間之機率？

中央極限定理在卜瓦松分佈之應用

卜瓦松分佈平均值

μ=λS

S=時間，空間，面積大小

λ在S中平均發生之個數

$$\sigma^2 = \lambda S$$

$$\mu = \lambda S = 5.4 \times 10^7 \cdot 10^{-5} = 5.4 \times 10^2 = 540$$

$$\sigma = \sqrt{540} = 23.4$$

$$P(500 \le Y \le 600)$$

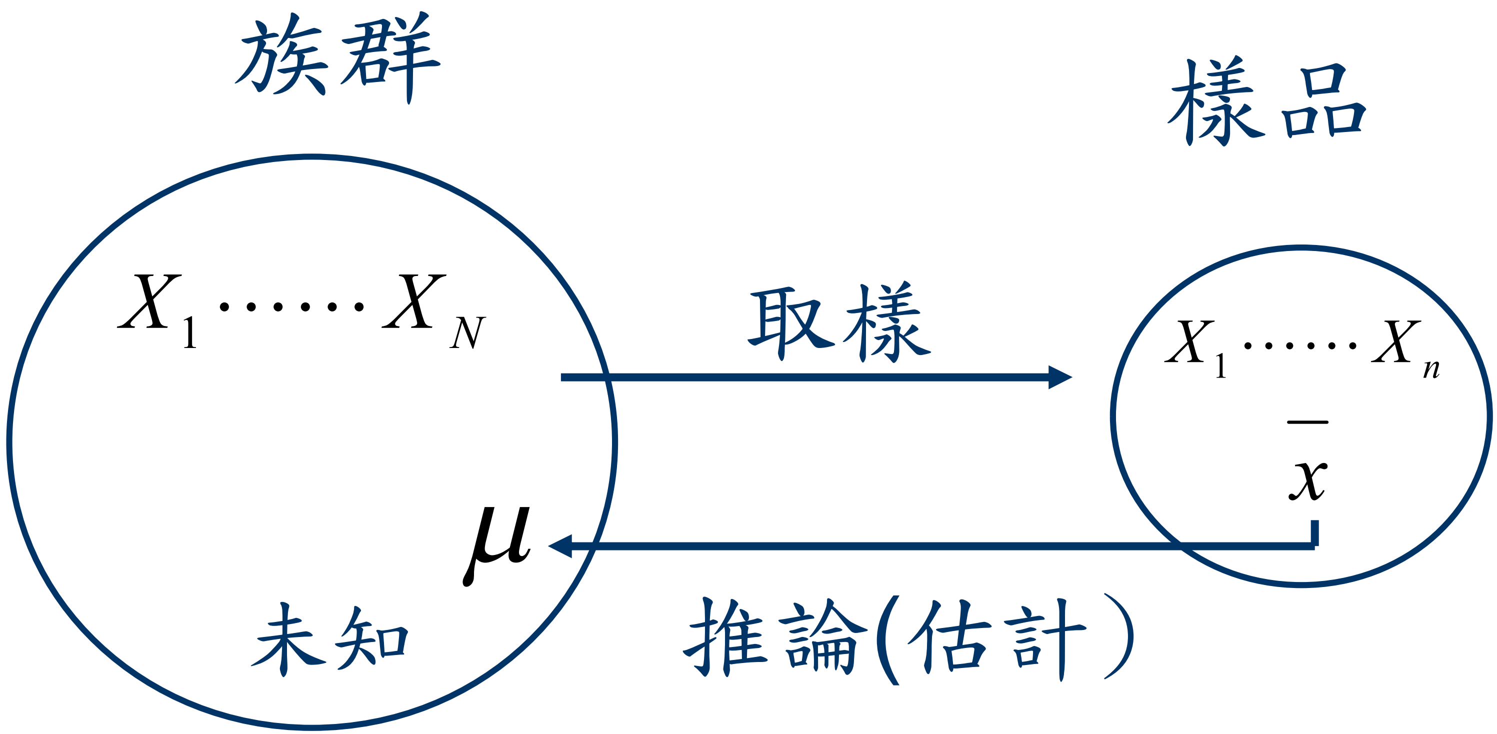
$$= P(499.5 \le Y \le 600.5)$$

$$= P(\frac{499.5 - 540}{23.4} \le Z \le \frac{600.5 - 540}{23.4}) = P(-1.74 \le Z \le 2.60)$$

$$= P(Z \le 2.60) - P\left(Z \le -1.74\right) = 0.9953 - 0.0409$$

$$= 0.9544$$

點估計值(Point Estimates)



- 族群母數不可獲，得以樣品n個觀測值所計算的統計量估算族群母數，此統計量稱為點估計值
- 隨機抽取台北市50位12歲學童之平均身高，估計台北市12歲學童之平均身高
- 通常將族群母數上加上帽子表示點估計值

$$\hat{\mu} = \overline{x}$$

服用某種藥物病人產生噁心的副作用比例

隨機抽取n位服用此藥物之病人X位發生噁心副作用

X為二項分佈~Bin(n，p)
發生副作用之比例點估計值： $\hat{p} = \dfrac{X}{n}$

如 $n = 50$ $\quad$ $\mathrm{X} = 10$ $\quad$ $\hat{p} = \dfrac{10}{50} = 0.2 = 20\%$

比例點估計值的抽樣分佈

- X為二項分佈~Bin(n, p)
- 比例點估計值: $Y = \hat{p} = X/n$
- 比例點估計值的期望值: $E(Y) = E(X/n) = E(np/n) = p.$
- 比例點估計值的變方:

$$\begin{aligned} Var(Y) &= Var(X/n) = (1/n^2)Var(X) \\ &= (1/n^2)(npq) = p(1\text{-}p)/n. \end{aligned}$$

- 比例點估計值的變方中p未知，可以比例點估計值$\hat{p}$估計
- 比例點估計值變方的估計值: $\hat{p}(1 - \hat{p})/n$
- 比例點估計值標準差的估計值:

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

點估計值應有的特性

1. 一致性(Consistency)

   當樣品數大時，點估計值應很接近族群真值。

2. 不偏性(Unbiased)

   所有可能樣品之點估計值之平均等於族群真值。

3. 充分性(Sufficiency)

   點估計值應包括樣品中所有觀測值之資訊(information)。

   30個觀測值的平均值較10個觀測值所含資訊為多。

   自常態分布$N(\mu, \sigma^2)$取得n個樣品，樣品平均數與樣品變方包含所有n個觀測值的資訊。

雖然所有可能樣品平均值＝族群平均值
但吾人只抽取一個樣品計算樣品平均值.
故所得樣品平均值可能與族群平均值差距很大.

族群｛2，4，6｝　　　　　$\mu = 4$
n＝2之可能樣品平均值：2，3，4，5，6

求得一個區間(Interval)而此區間包含族群母數機率達到吾人的最低要求

此一機率稱為信心水準(Confidence Level)，通常為90%，95%， 或99%

此一區間稱為(1－α)%信賴區間(Confidence Interval)

信賴區間是隨機，不同樣品計算的信賴區間不同

區間估計(Interval Estimation)

$$Z=\frac{\bar{x}-\mu}{\sigma_{\bar{x}}}=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$$

$$P[a\le Z\le b]=1-\alpha>0.5$$

因標準常態分佈對稱於0

$$-a=b=Z_{1-\alpha/2}$$

$Z_{1-\alpha/2}$ 為常態分佈的(1－ α /2)百分位

$$1-\alpha=0.90\Rightarrow Z_{0.95}=1.645$$

$$1-\alpha=0.95\Rightarrow Z_{0.975}=1.96$$

$$1-\alpha=0.99\Rightarrow Z_{0.995}=2.5758$$

$$P\left[-z_{1-\alpha/2} \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le z_{1-\alpha/2}\right] = 1-\alpha$$

$$\Rightarrow P\left[-z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \le \bar{x}-\mu \le z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

$$\Rightarrow P\left[\bar{x}-z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

下限： $\bar{x}-z_{1-\alpha/2} \cdot \dfrac{\sigma}{\sqrt{n}}$

上限： $\bar{x}+z_{1-\alpha/2} \cdot \dfrac{\sigma}{\sqrt{n}}$

設今以新飼料飼養25隻試驗兔子，兩星期後每隻平均增重600克，過去飼養同種類兔子之經驗，其標準偏差σ＝50克，試求族群均值μ之95%信賴區間？

由 $\bar{x}-1.96\dfrac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+1.96\dfrac{\sigma}{\sqrt{n}}$

下限 $\bar{x}-1.96\dfrac{\sigma}{\sqrt{n}} = 600-1.96\dfrac{50}{\sqrt{25}} = 580.4$

上限 $\bar{x}+1.96\dfrac{\sigma}{\sqrt{n}} = 600+1.96\dfrac{50}{\sqrt{25}} = 619.6$

故μ之95%之CI為 $580.4 \le \mu \le 619.6$

ÿ 20-30歲成人血液中尿酸濃度

3.8, 4.0, 4.2, 4.2, 4.5, 4.8, 5.4, 5.8, 6.2, 6.8, 7.2, 8.5

$\bar{x} = 5.45, \quad \sigma = 1.4 \qquad 95\%$ C.I.

$$L = \bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 5.45 - 1.96\frac{1.4}{\sqrt{12}}$$

$$= 5.45 - (1.96)(0.404)$$

$$= 4.66$$

$$\text{U} = \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 5.45 + 1.96\frac{1.4}{\sqrt{12}}$$

$$= 5.45 + (1.96)(0.404)$$

$$= 6.24$$

$95\%\text{C.I.} = (4.66,\ 6.24)$

95%信賴區間的意義

- 自族群取100個樣品
  計算100個95%信賴區間
  95個95%信賴區間會包含μ
  5個95%信賴區間不包含μ

⇒ 95%信心族群平均值會落在區間內

大樣本之信賴區間 $(n \geq 30)$

σ未知時可以S代之

$$L = \bar{x} - z_{1-\alpha/2} \frac{S}{\sqrt{n}}$$

$$U = \bar{x} + z_{1-\alpha/2} \frac{S}{\sqrt{n}}$$

比例之信賴區間 $(n \geq 30,\ \ n \cdot \hat{p} \geq 5)$

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

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

例：瞭解新殺蟲劑之效果.隨機抽取500隻某昆蟲結果375隻死亡.求新殺蟲劑之殺蟲率的95%信賴區間

$$n = 500 \qquad X = 375 \qquad 1-\alpha = 0.95 \qquad \alpha = 0.05$$

$$z_{1-\alpha/2} = z_{0.975} = 1.96$$

$$\hat{p} = 375/500 = 0.75$$

$$L = \hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.75 - 1.96\sqrt{\frac{(0.75)(0.25)}{500}} = 0.7121$$

$$U = \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.75 + 1.96\sqrt{\frac{(0.75)(0.25)}{500}} = 0.7879$$

殺蟲率的95%信賴區間=(0.7121, 0.7879)

總結(Summary)

抽樣分佈

中央極限定理

點估計 $x \to \mu$

$\hat{p} \to p$

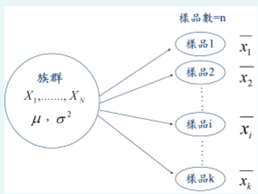
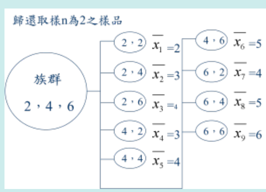
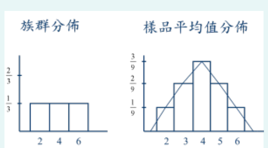
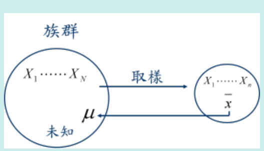
區間估計

$$\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad \bar{x} \pm z_{1-\alpha/2}\frac{S}{\sqrt{n}} \qquad p \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

習題：

- P. 130: 8
- P.142：2, 3, 5

版權聲明

頁碼	作品	授權條件	作者/來源
1-78			轉載自 Microsoft Office 2003 多媒體藝廊，依據 Microsoft 服務合約及著作權法第 46、52、65 條合理使用。
5		BY NC SA	國立臺灣大學 農藝系 劉仁沛 教授。
6		BY NC SA	國立臺灣大學 農藝系 劉仁沛 教授。
9		BY NC SA	國立臺灣大學 農藝系 劉仁沛 教授。
21		BY NC SA	國立臺灣大學 農藝系 劉仁沛 教授。