巨量資料管理學院碩士在職專班

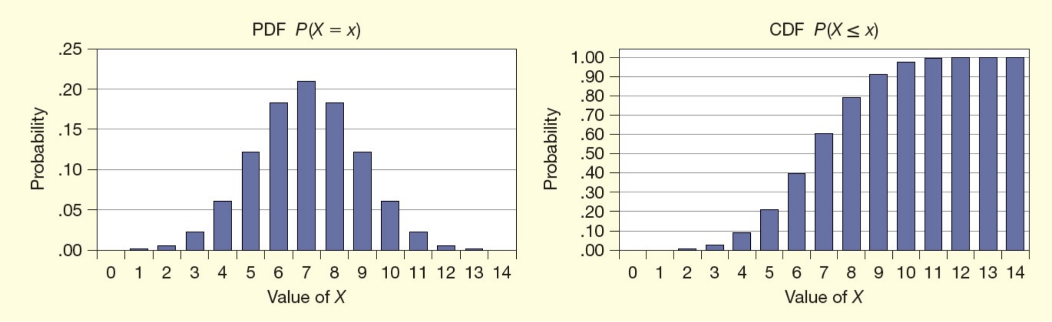
統計分析

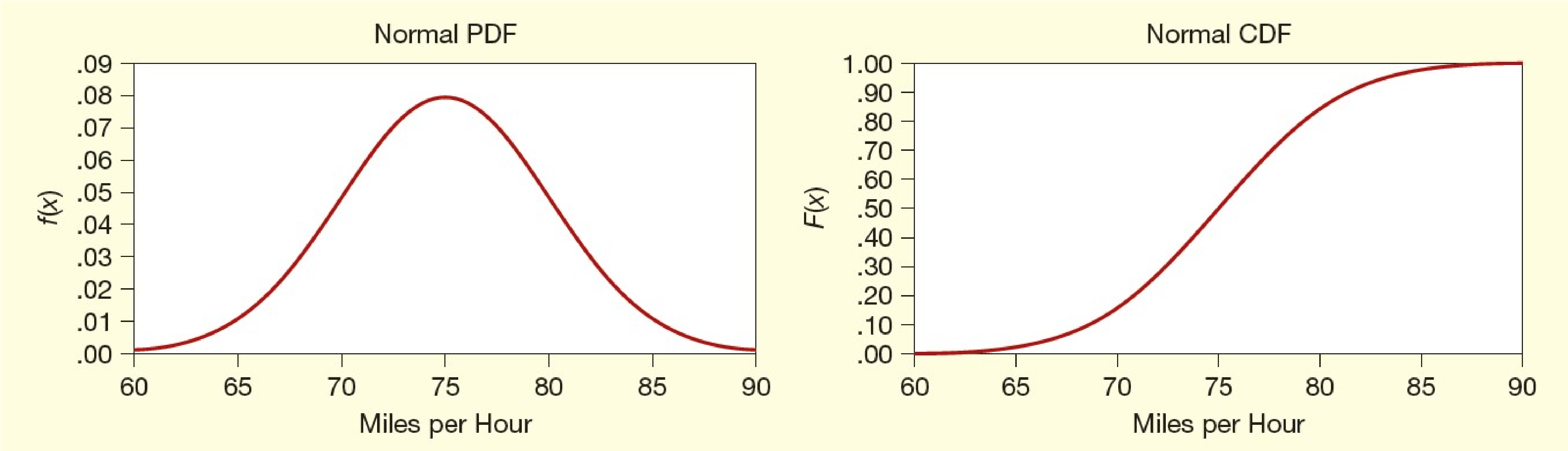
2022/10/28
陳光宏

機率分布 – 連續變項

連續變項的機率分布

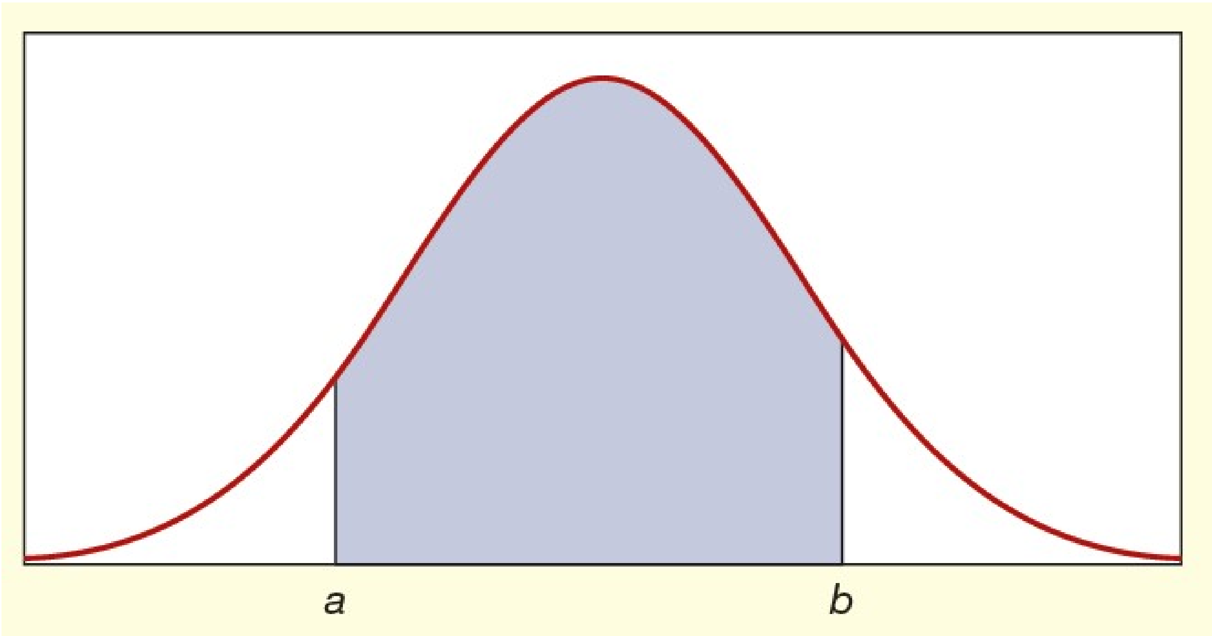
- PDF $f(x)$ 與 CDF $F(x)$





機率 = PDF的面積

- 連續型的隨機變數介於a和b之間的機率 $P(a < X < b)$



期望值與變異數

Mean $\quad E(X) = \mu = \int_{-\infty}^{+\infty} x f(x)\, dx$

Variance $\quad \mathrm{Var}(X) = \sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)\, dx$

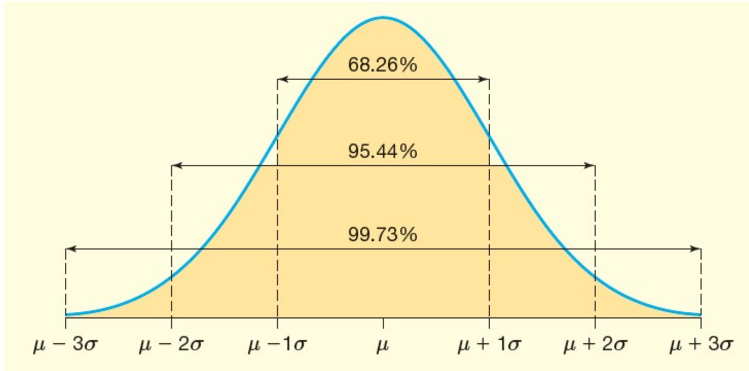
期望值與變異數

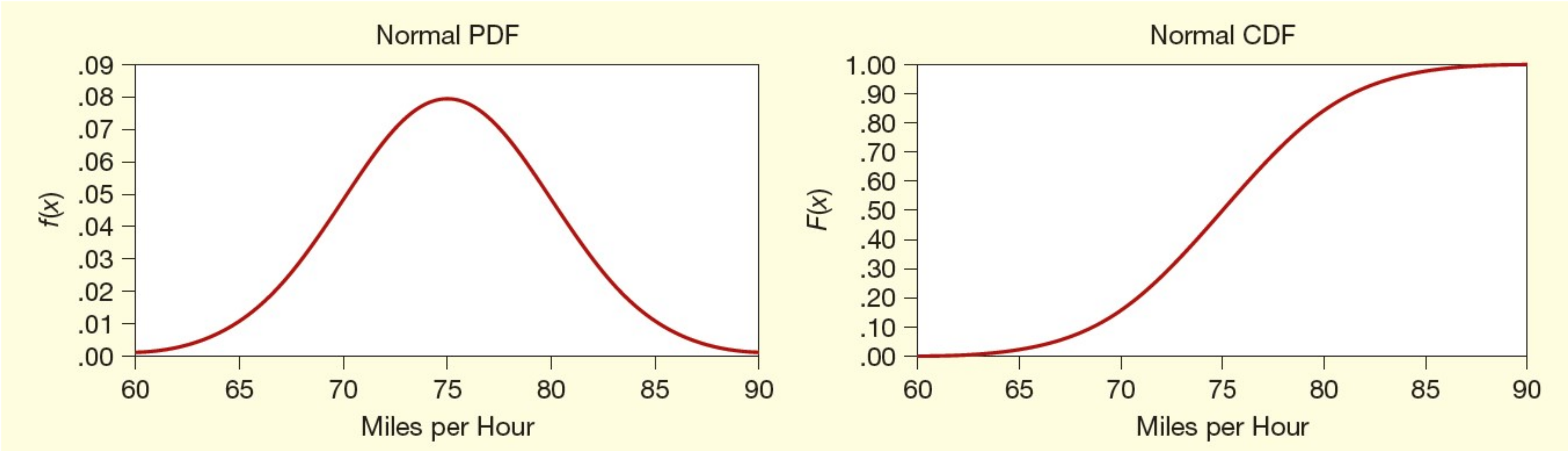
	Continuous Random Variable	*Discrete Random Variable*
Mean	$E(X) = \mu = \int_{-\infty}^{+\infty} x f(x)\, dx$	$E(X) = \mu = \sum_{\text{all } x} x\, P(x)$
Variance	$\text{Var}(X) = \sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)\, dx$	$\text{Var}(X) = \sigma^2 = \sum_{\text{all } x} [x - \mu]^2 P(x)$

常態分布 (Normal distribution)

- $N(\mu, \sigma^2)$
- $[\mu - 3\sigma, \mu + 3\sigma]$ 包含幾乎所有數值

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



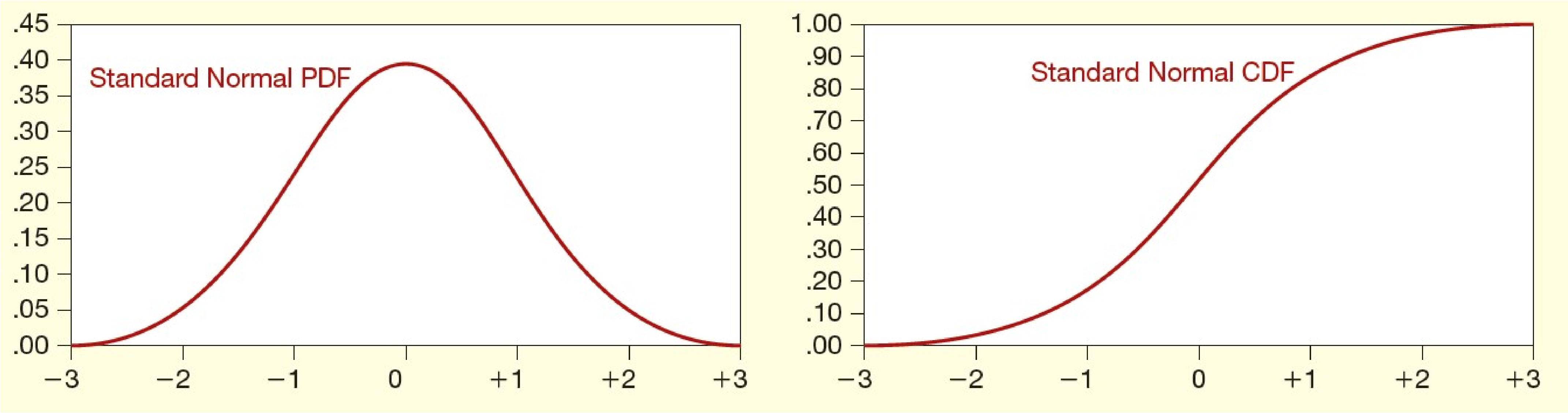


Parameters	μ = population mean σ = population standard deviation
PDF	$f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
Domain	$-\infty < x < +\infty$
Mean	μ
Std. Dev.	σ
Shape	Symmetric, mesokurtic, and bell-shaped.
PDF in Excel*	=NORM.DIST(x,μ,σ,0)
CDF in Excel*	=NORM.DIST(x,μ,σ,1)
Random data in Excel	=NORM.INV(RAND(),μ,σ)

標準常態分布

- 把隨機變數x標準化

$$z = \frac{x - \mu}{\sigma}$$

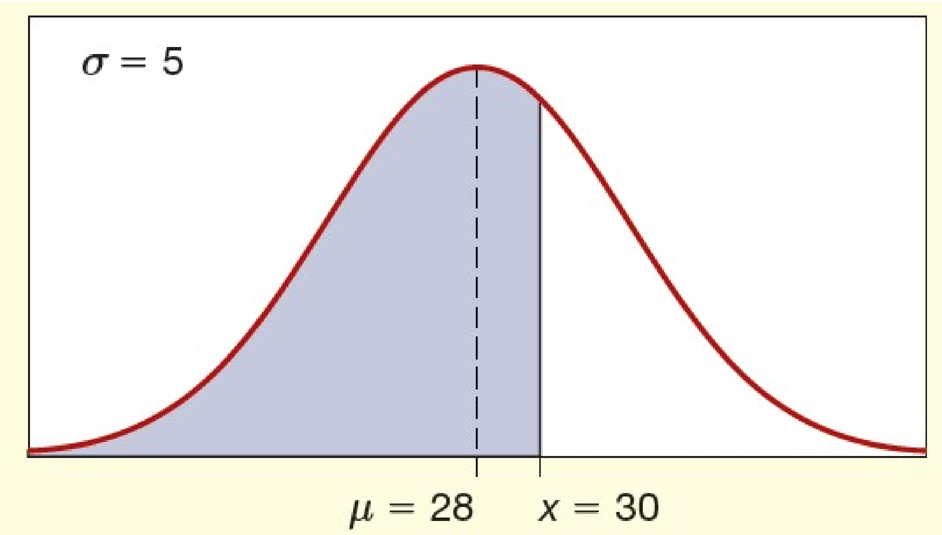


Parameters	μ = population mean σ = population standard deviation
PDF	$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ where $z = \frac{x - \mu}{\sigma}$
Domain	$-\infty < z < +\infty$
Mean	0
Standard deviation	1
Shape	Symmetric, mesokurtic, and bell-shaped.
CDF in Excel*	=NORM.S.DIST(z,1)
Random data in Excel	=NORM.S.INV(RAND())
Comment	There is no simple formula for a normal CDF, so we need normal tables or Excel to find areas.

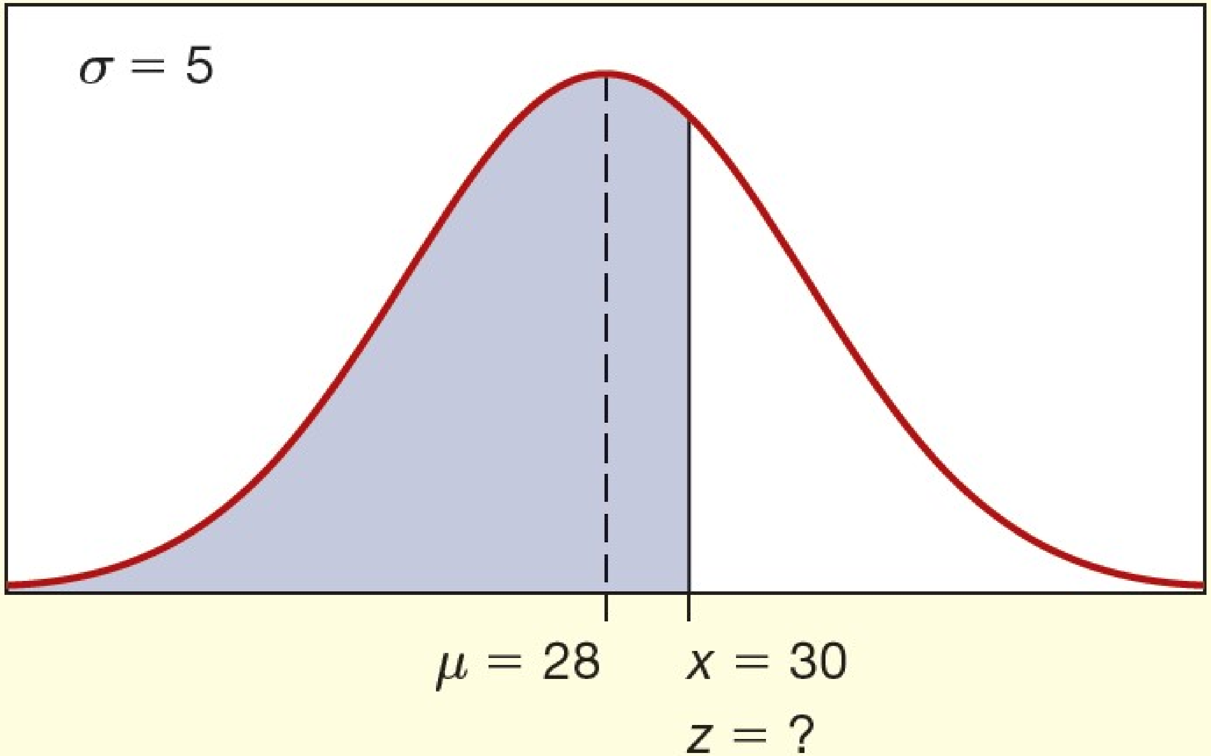
範例

某修車廠修車時間的PDF如右，平均時間是28分鐘，標準差5分鐘

1. 請問有多少比例的車，其修車時間會低於半小時？
2. 現在來了一台車，請問修這台車的時間超過40分鐘的機率是多少？
3. 老闆希望80%的車，修車時間不要超過半小時，請問平均修車時間必須是多少才能符合老闆的要求？



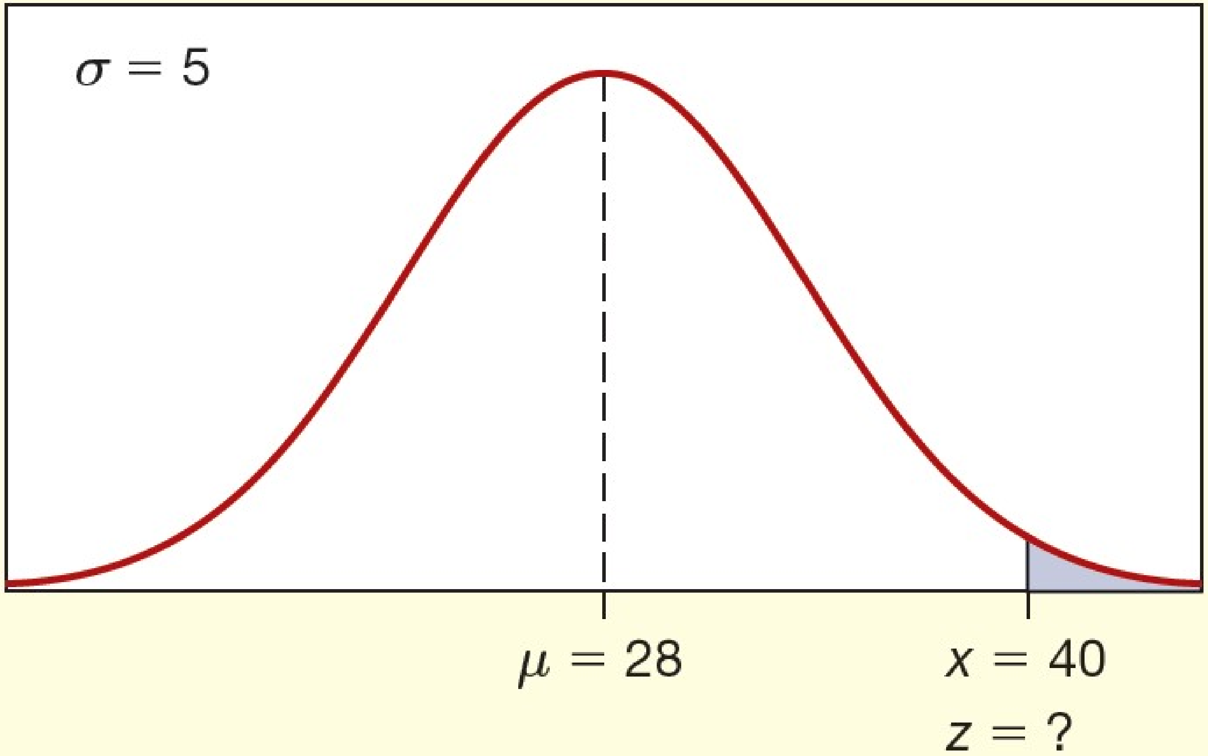
範例解答 (1)



Using Excel,
=NORM.DIST(30,28,5,1)
= .655422

$$z = \frac{30 - 28}{5} = 0.40$$

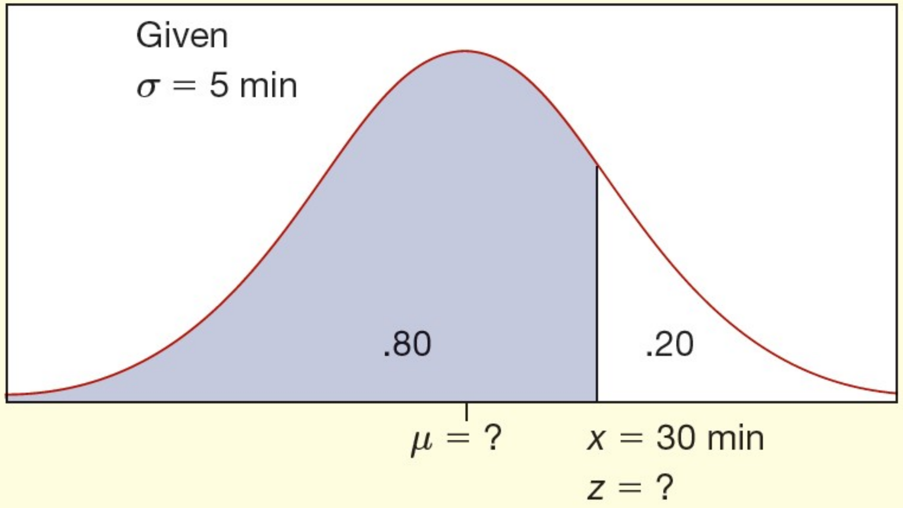
範例解答 (2)



Using Excel,
= 1 − NORM.DIST(40,28,5,1)
= .008198

$$z = \frac{40 - 28}{5} = 2.4$$

範例解答 (3)



Using Excel,
=NORM.S.INV(.80)
=.841621

$$z = \frac{x - \mu}{\sigma}$$

$$0.84 = \frac{30 - \mu}{5}$$

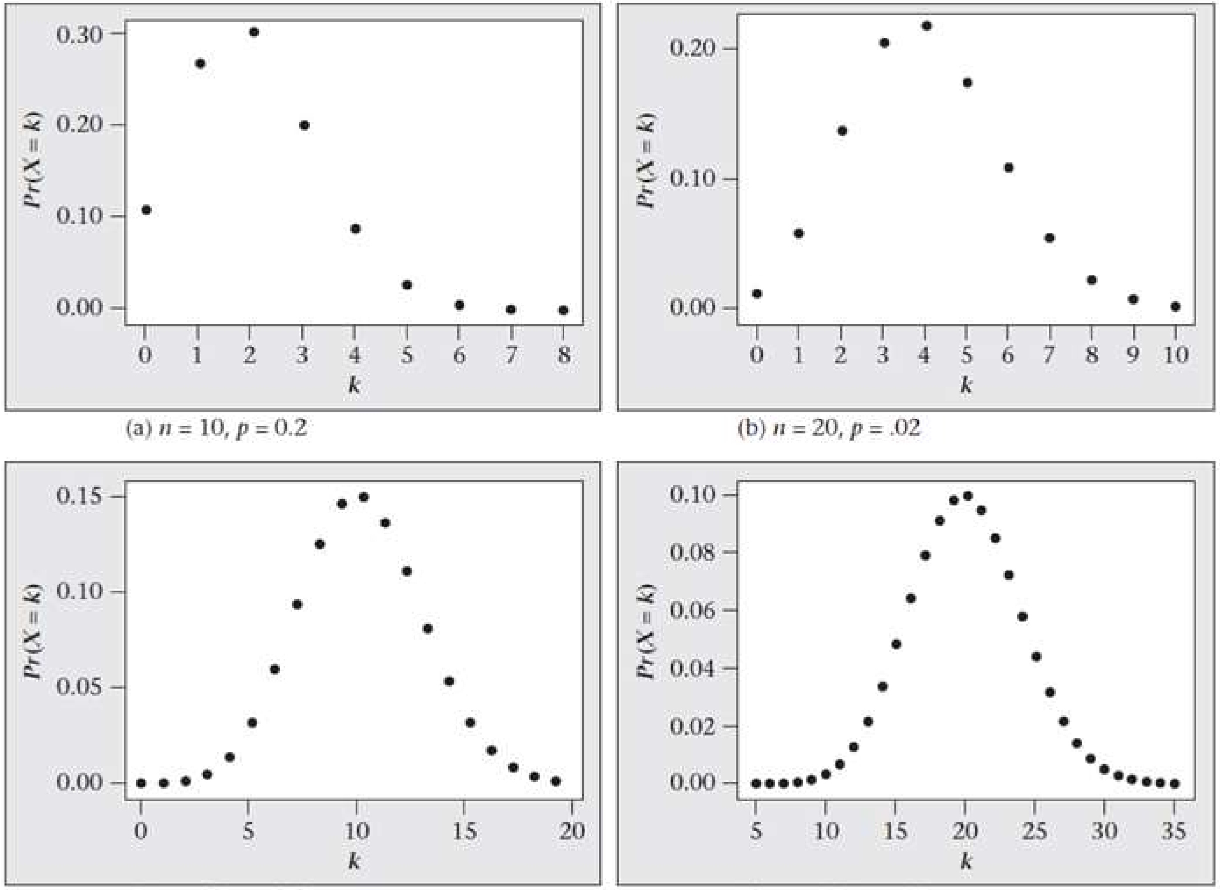
$$\mu = 30 - 0.84(5) = 25.8$$

Normal approximation to binomial

$$\mu = n\pi$$

$$\sigma = \sqrt{n\pi\,(1-\pi\,)}$$

when $n\pi \geq 10$ and $n(1-\pi) \geq 10$,



(a) $n = 10$, $p = 0.2$

(b) $n = 20$, $p = .02$

(c) $n = 50$, $p = 0.2$

(d) $n = 100$, $p = 0.2$

範例

丟一枚硬幣32次，請問出現超過17次正面的機率是多少？

1. 請問這個情況服從什麼機率分布？
2. 根據上述的機率分布，需要哪些參數？
3. 若基於常態分佈的假設，出現超過17次正面的機率是多少？

範例解答

1. 找出平均值 (期望值) 與標準差

$$\mu = n\pi = (32)(0.5) = 16$$

$$\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{(32)(0.5)(1-0.5)} = 2.82843$$

2. 計算z-score

$$z = \frac{x-\mu}{\sigma} = \frac{17.5-16}{2.82843} = .53$$

3. 找出 $P(X \geq 18)$的機率值

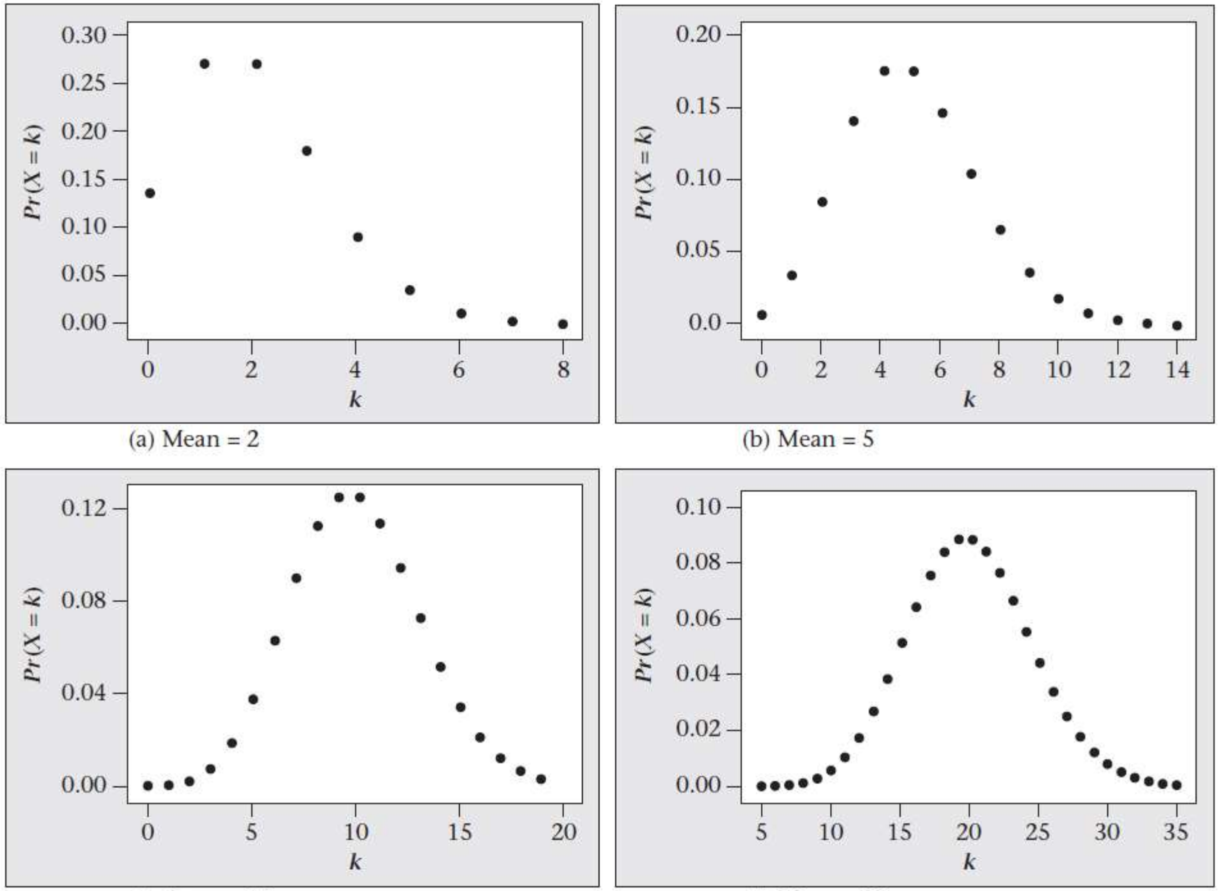
$$P(X \geq 18) = 1 - P(X \leq 17) = 1 - .7017 = .2983$$

Normal approximation to Poisson

$$\mu = \lambda$$

$$\sigma = \sqrt{\lambda}$$

when $\lambda \geq 10$



(a) Mean = 2

(b) Mean = 5

(c) Mean = 10

(d) Mean = 20

範例

週三上午10點到中午，平均每小時大概有42通客服電話詢問周年慶優惠活動，現在想評估一小時接到超過50通詢問周年慶活動的客服電話的可能性

1. 請問您認為這個情況是用什麼機率分布？參數是什麼？
2. 請問這個情況能否逼近常態分布？
3. 若基於常態分佈的假設，請問一小時接到超過50通詢問周年慶活動的客服電話，機率有多高？

範例解答

1. 找出平均值 (期望值) 與標準差

$$\mu = \lambda = 42$$
$$\sigma = \sqrt{\lambda} = \sqrt{42} = 6.48074$$

2. 計算z-score

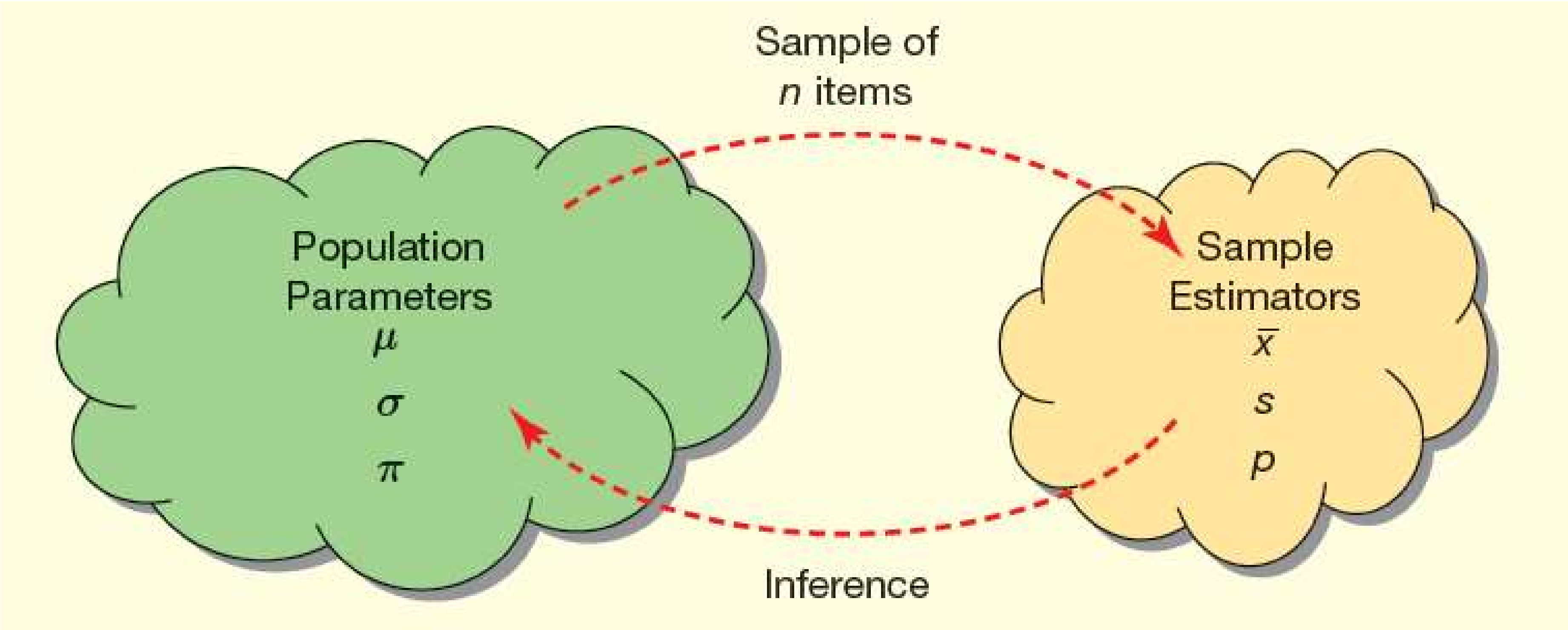
$$z = \frac{x - \mu}{\sigma} = \frac{50.5 - 42}{6.48074} \cong 1.31$$

3. 找出 $P(X \geq 51)$的機率值

$$P(X \geq 51) = 1 - P(X \leq 50) = .0951$$

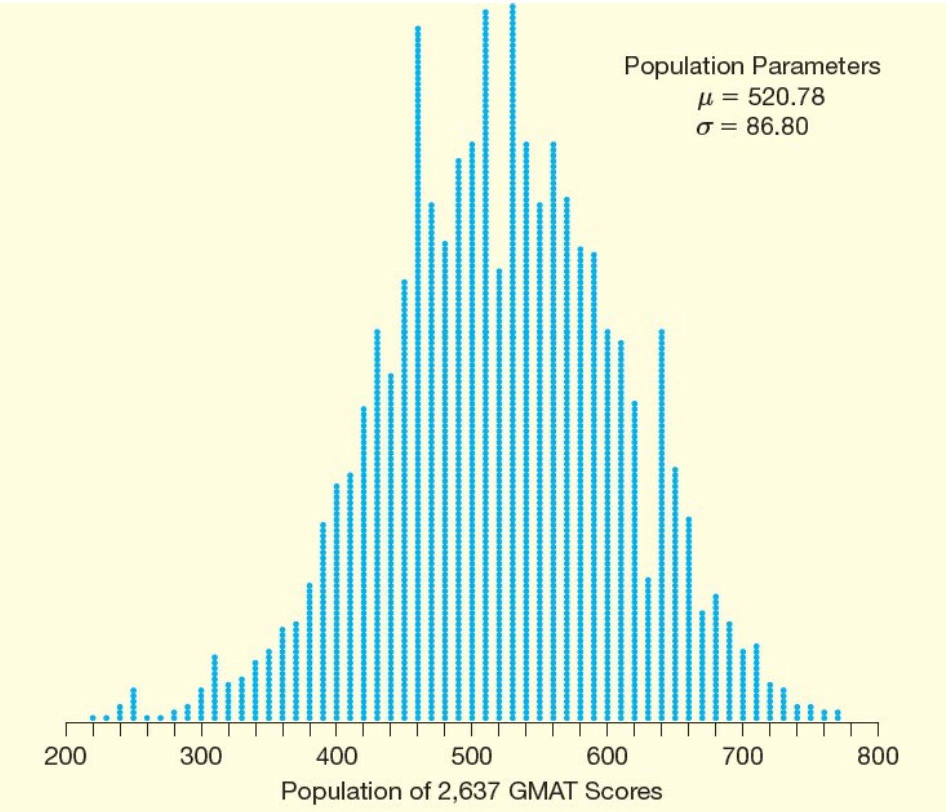
估計 (Estimation)

母體與樣本

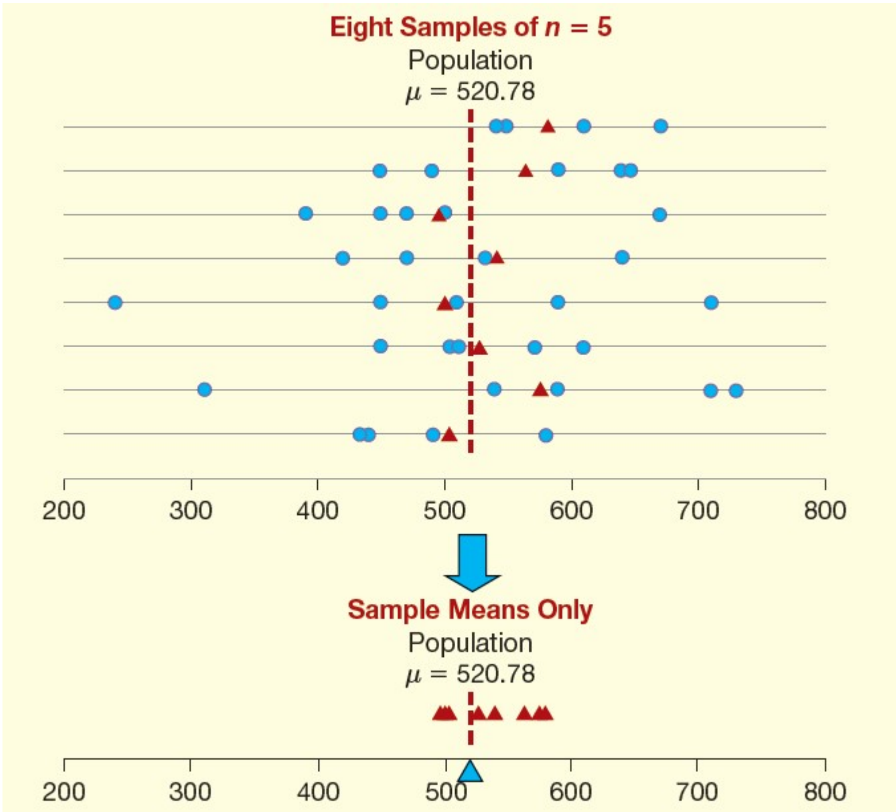

Sample of
n items
Population
Parameters
μ
σ
π
Sample
Estimators
x̄
s
p
Inference

我們只有一個樣本
這個樣本的估計值可以代表整個群體嗎？

抽樣 (Sampling)



Population Parameters
μ = 520.78
σ = 86.80
200
300
400
500
600
700
800
Population of 2,637 GMAT Scores



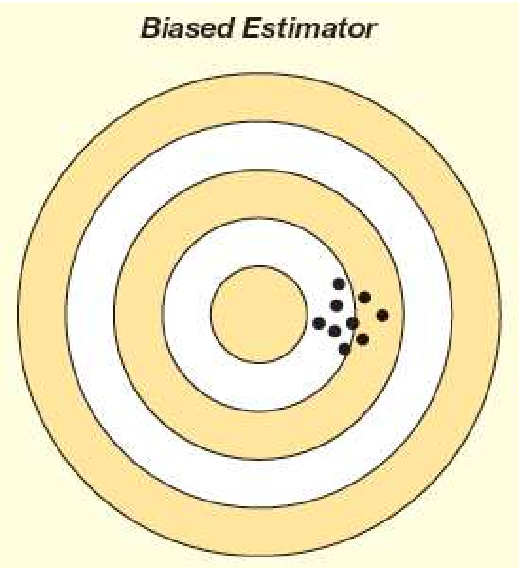
Eight Samples of n = 5
Population
μ = 520.78
200
300
400
500
600
700
800
Sample Means Only
Population
μ = 520.78
200
300
400
500
600
700
800

樣本估計值

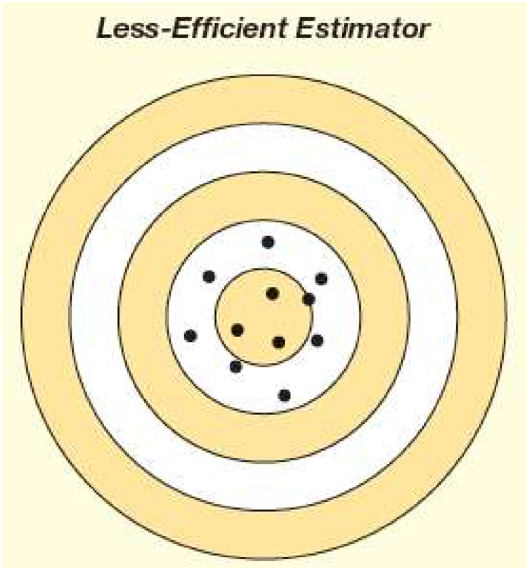
Estimator	***Formula***	***Parameter***
Sample mean	$\boxed{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i$ where x_i is the ith data value and n is the sample size	μ
Sample proportion	$\boxed{p} = x/n$ where x is the number of successes in the sample and n is the sample size	π
Sample standard deviation	$\boxed{s} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$ where x_i is the ith data value and n is the sample size	σ

抽樣誤差

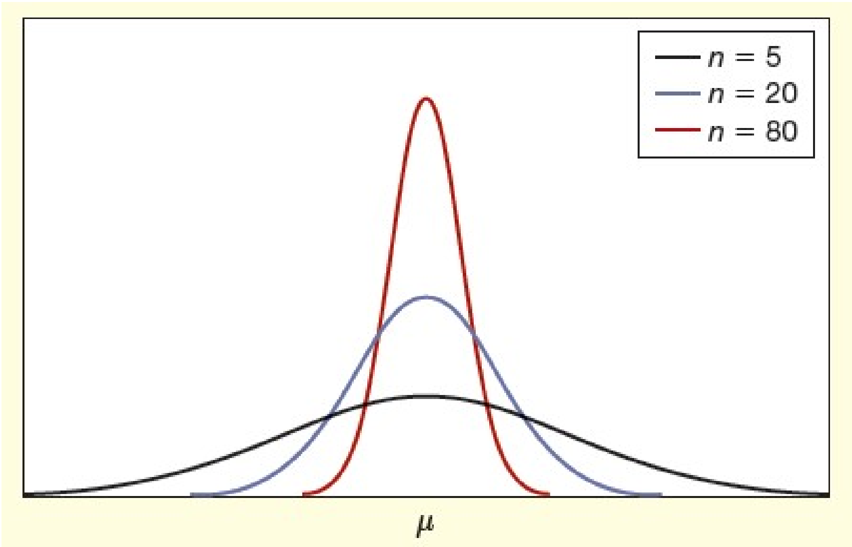
偏差 (bias)



效率 (efficiency)



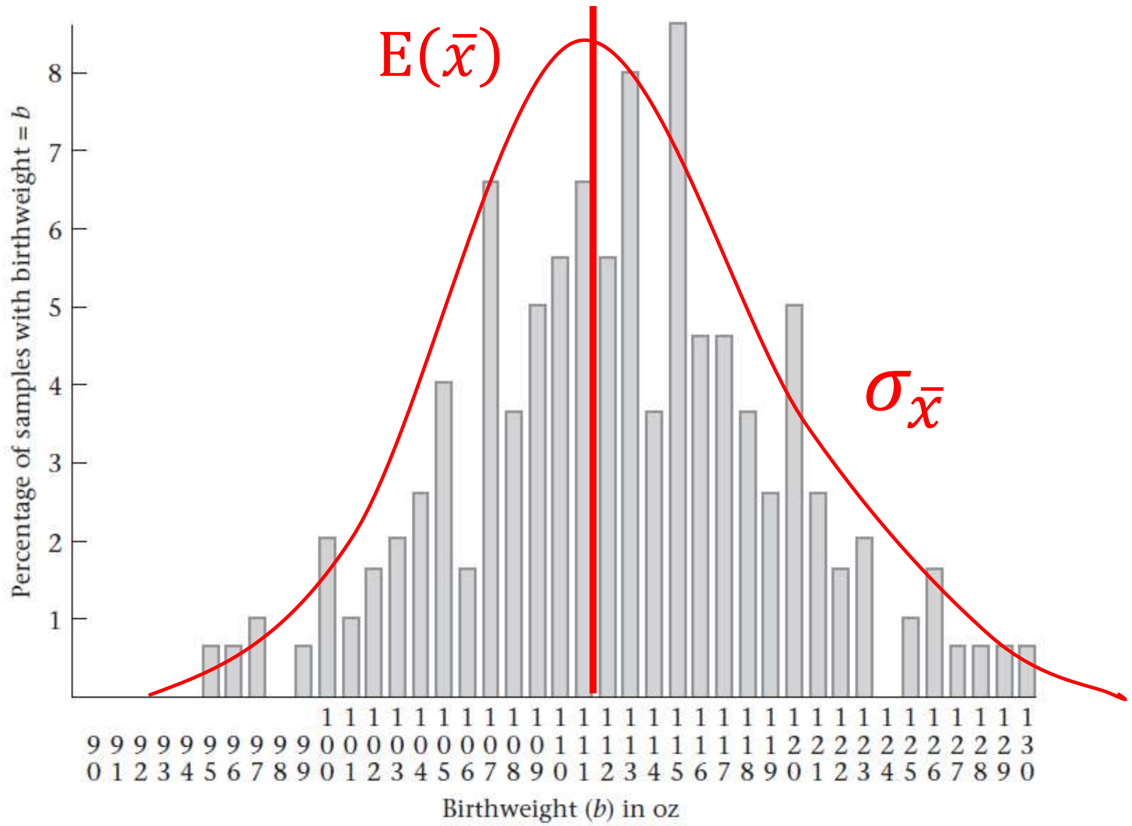
一致性 (consistency)



抽樣分布 (Sampling distribution)

- 假設可以重複抽樣，每次從母體抽 n 個樣本，平均值為 $\bar{x}$
- 將每次抽樣的 $\bar{x}$ 畫成右圖

Sampling distribution of $\overline{X}$ over 200 samples of size 10 selected from the population of 1000 birthweights given in Table 6.2 (100 = 100.0-100.9, etc.)



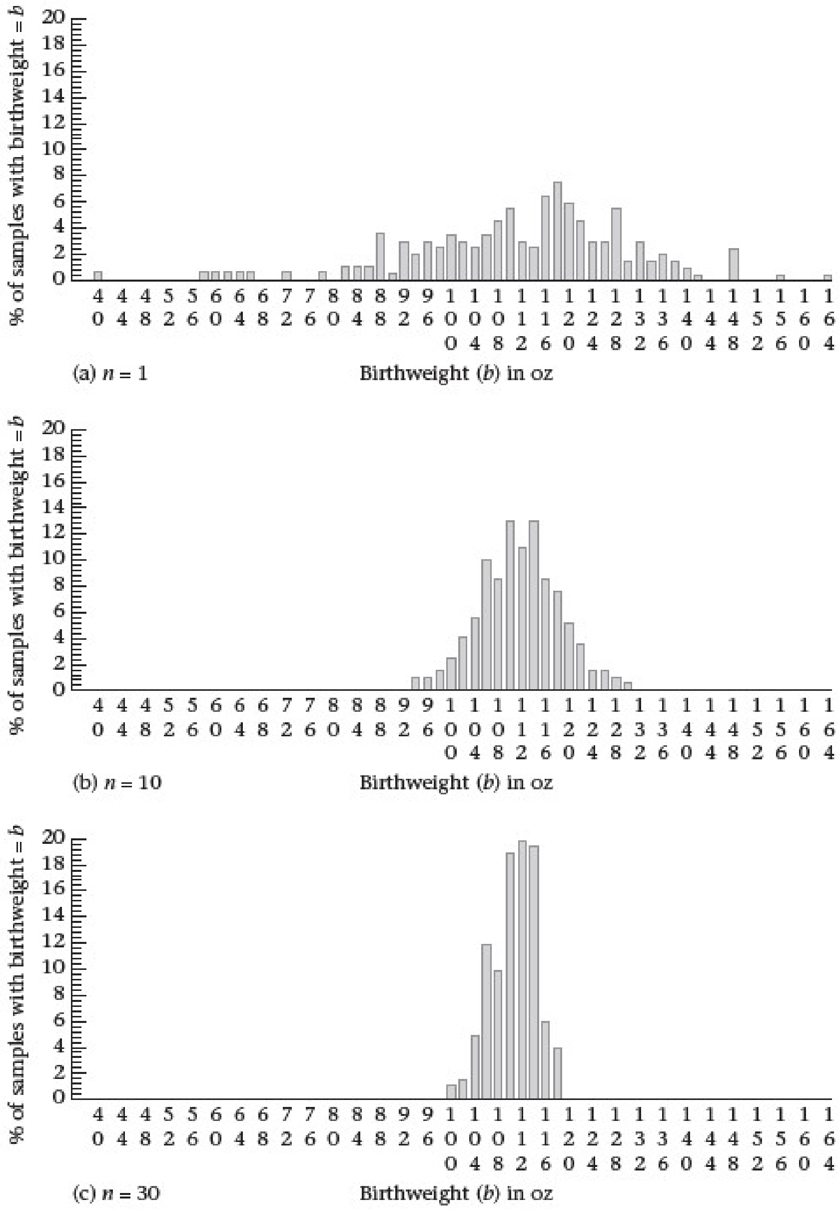
中央極限定理 (Central limit theorem)

- $\bar{x}$ 會服從常態分布　　$\bar{x} \sim N(\mu, \sigma/\sqrt{n})$

- 樣本平均 $\bar{x}$ 會是母體平均 μ 的估計值

- 樣本的標準誤差 (standard error) 為 $\sigma/\sqrt{n}$

- 當樣本數越來越大時，無論母體是什麼分布，最後都會趨近於常態分布

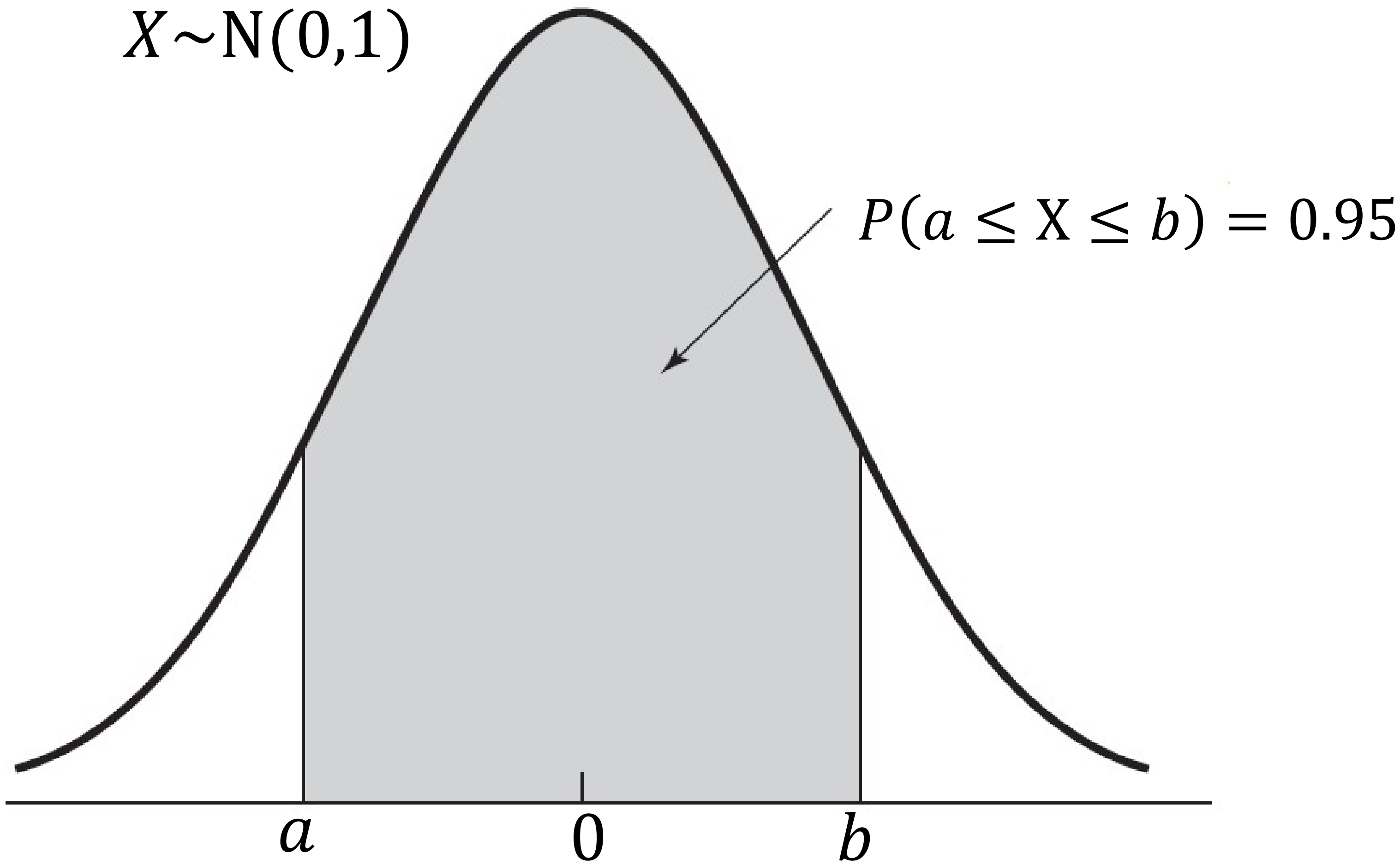


估計 (樣本推論母體)

- 利用中央極限定理，當樣本數夠大時，會趨近常態分布，不需考慮原本母體原本的分布 (normal approximation)

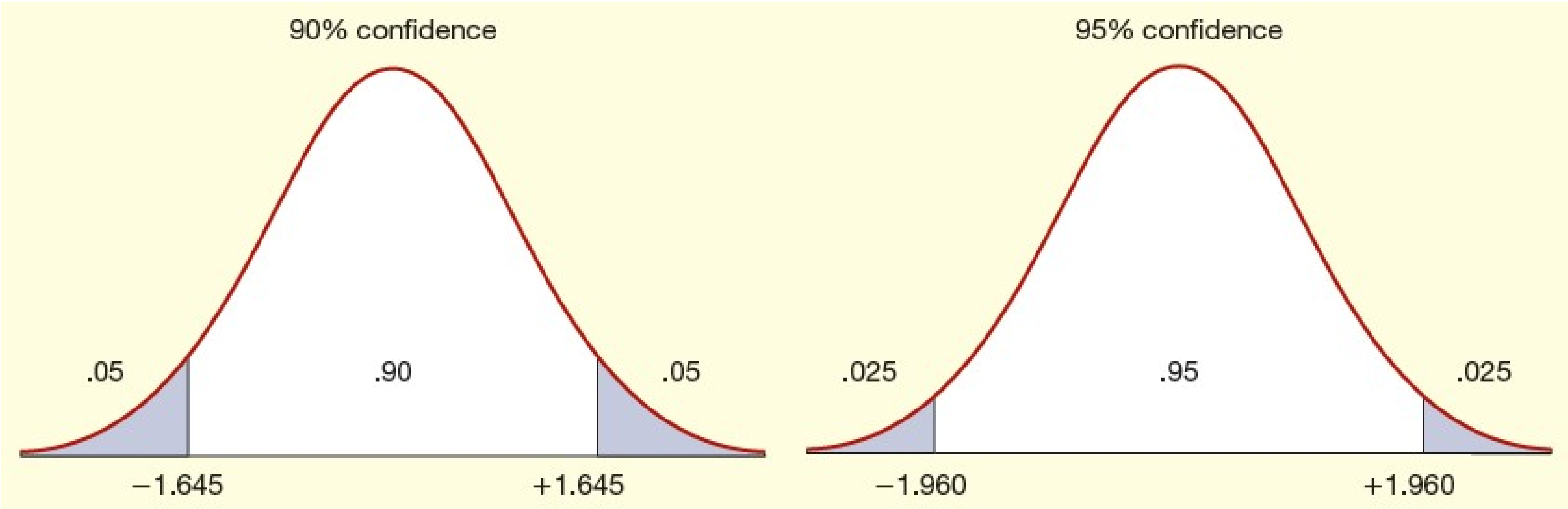
- 樣本平均 $\bar{x}$ 推論母體平均 μ
  - 點估計 (point estimate)

- 因為存在抽樣誤差，所以用標準誤差 (standard error, $\sigma/\sqrt{n}$) 來評估抽樣造成的不確定性
  - 區間估計 (interval estimate)
  - 建立一個區間，說明這個區間有多大的機率可以包含真正母體的平均值 μ



$X \sim \mathrm{N}(0,1)$
$P(a \leq \mathrm{X} \leq b) = 0.95$
a
0
b

信賴區間 (Confidence interval)

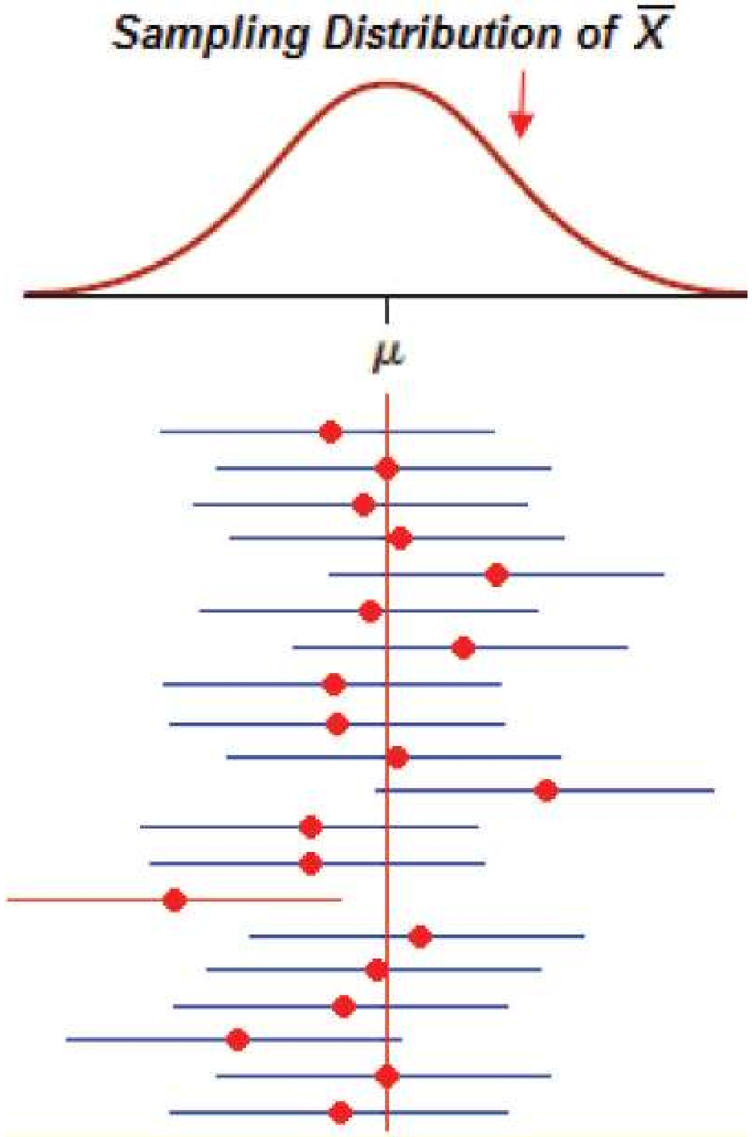


解讀「95%信賴區間」

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

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

- 假設已知母體變異數 σ，所以服從標準常態分布
- 假設重複從相同的母群體抽樣100次，每次抽樣都計算平均值與信賴區間，會有95次會包含真正的母體平均值



範例

在汽水工廠裡，取10瓶半升裝的瓶裝汽水，平均的容量是503.4mL。假設母群體中容量的標準差是1.2mL

1. 請問容量點估計值是多少？
2. 請問95%信賴區間為多少？

範例解答

1. 點估計值 = 503.4mL

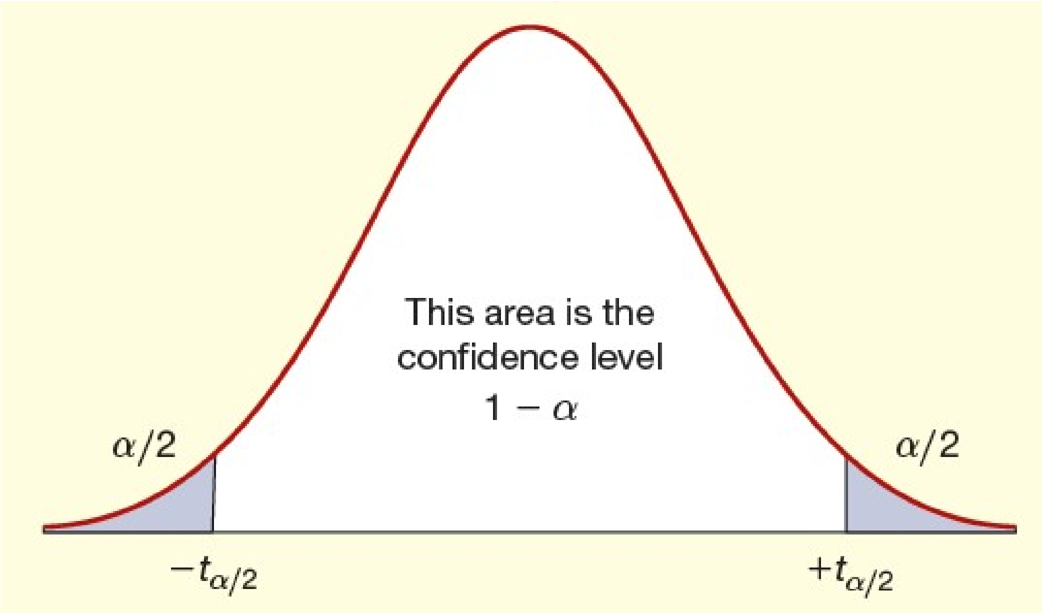
2. 95% 信賴區間 $\bar{x} \pm z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$

$$503.4 \pm 1.96\frac{1.20}{\sqrt{10}} \quad \text{or} \quad [502.66, 504.14]$$

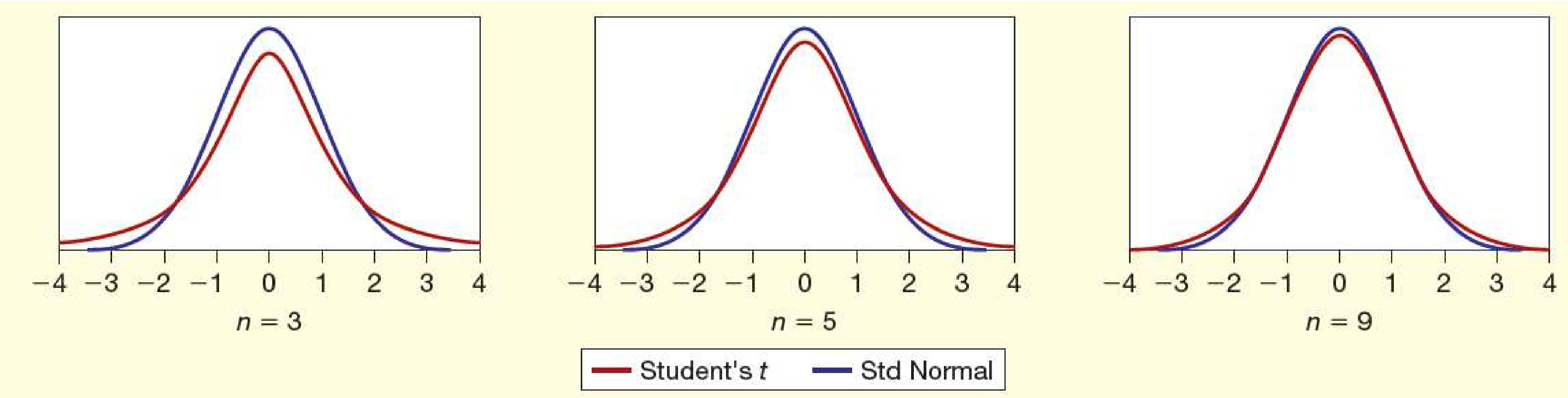
t分布

Confidence Interval for a Mean μ with Unknown σ

(8.8) $$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$ where $\frac{s}{\sqrt{n}}$ is the *estimated* standard error of the mean



標準常態分布與t分布



20	1.325	1.725	2.086	2.528	2.845
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
100	1.290	1.660	1.984	2.364	2.626
∞	1.282	1.645	1.960	2.326	2.576

標準常態分布與t分布

- T分布會受到自由度 (Degree of freedom)的影響，而改變分布的形狀
  - 自由度 = 樣本數 – 1 = n -1
- 當樣本數越大，越接近標準常態分布

範例

在汽水工廠裡，取11瓶半升裝的瓶裝汽水，平均的容量是503.4 mL，樣本標準差是1.2 mL

1. 請問容量點估計值是多少？
2. 請問95%信賴區間為多少？

範例解答

1. 點估計值 = 503.4 mL

2. 95% 信賴區間

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

$$503.4 \pm 2.228 \times \frac{1.2}{\sqrt{10}}$$

	Confidence Level				
d.f.	***80%***	***90%***	***95%***	***98%***	***99%***
1	3.078	6.314	12.706	31.821	63.656
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
10	1.372	1.812	2.228	2.764	3.169
20	1.325	1.725	2.086	2.528	2.845
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
100	1.290	1.660	1.984	2.364	2.626
∞	1.282	1.645	1.960	2.326	2.576