STATISTICS FOR BUSINESS AND ECONOMICS

CHAPTER 8 INTERVAL ESTIMATION (區間估計)-PART1

Outline

- 8.1. Interval estimation for population mean: σ known
  - Margin of error (邊際誤差) and the interval estimate
  - Confidence level (信賴水準)
  - Confidence coefficient (信賴區間)
- 8.2. Interval estimation for population mean: σ unknown
  - Margin of error and the interval estimate
  - Practical advice when σ is unknown
  - Summary of interval estimation procedures

Interval estimate (區間估計值)

- **Interval estimate (區間估計值)**
  - A point estimator (點估計量: $\bar{x}$, $\bar{p}$, s) is a sample statistic (樣本統計量) used to estimate a population parameter (母體參數: μ, p, σ).
  - <u>A point estimator cannot be expected to provide the exact value of the population parameter</u> (某一次樣本的點估計量的值不會恰好等於母體參數值)
    - An interval estimate is thus often computed by adding and subtracting a value, called **the margin of error**, to the point estimator.
  - The general form of an interval estimate is as follows:
    - **Point estimator $\pm$ Margin of error**.
    - The purpose of an interval estimate is to provide information about how close the point estimate, provided by the sample, is to the value of the population parameter.

Interval estimate (區間估計值)

- The general form of an interval estimate of a population mean (母體平均數μ的區間估計值) is :
  - $\bar{x} \pm$ Margin of error
- The general form of an interval estimate of a population proportion (母體比例p的區間估計值) is:
  - $\bar{p} \pm$ Margin of error

Interval Estimate of Population Mean: σ known

- **Ex.**, Each week Lloyd's Department Store selects a simple random of 100 customers in order to learn about the amount spent per shopping trip. With x representing the amount spent per shopping trip, the sample mean $\bar{x}$ provides a point estimate of me, the mean amount spent per shopping trip for the population of all Lloyd's customers. Lloyd's has been using the weekly survey for several years. Based on the historical data, Lloyd's now **assumes a known value of $\boldsymbol{\sigma = \$20}$** for the population standard deviation. The historical data also indicate that the population follows a normal distribution.
- During the most recent week, Lloyd's surveyed 100 customers **$(\boldsymbol{n = 100})$** and obtained **a sample mean of $\boldsymbol{\bar{x} = \$82}$**.
- How to compute the margin of error for this estimate and develop an interval estimate of the population mean μ

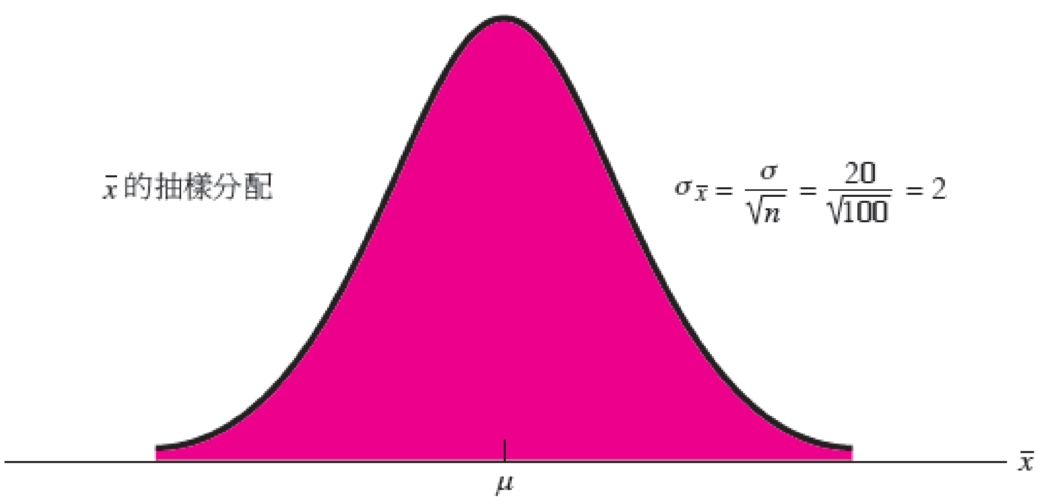
Interval Estimate of Population Mean: σ known

- **Note:** In order to develop an estimate of a population mean, either the population standard deviation σ or the sample standard deviation s must be used to compute the margin of error.
  - In most applications σ is not known, and s is used to compute the margin of error.
  - In some applications, large amounts of relevant historical data are available and can be used to estimate the population standard deviation **prior to sampling**.
    - In quality control applications where a process is assumed to be operating correctly, or "in control", it is appropriate to treat the population standard deviation as known (**如品質管制中, 如果程序進行已經很順利無錯誤, 將母體標準差視為已知是可以的**).

Interval Estimate of Population Mean: σ known

- Sampling distribution of sample mean amount spent from simple random samples of 100 customers.

圖 8.1 100 位顧客的簡單隨機樣本得到的購物花費樣本平均數的抽樣分配



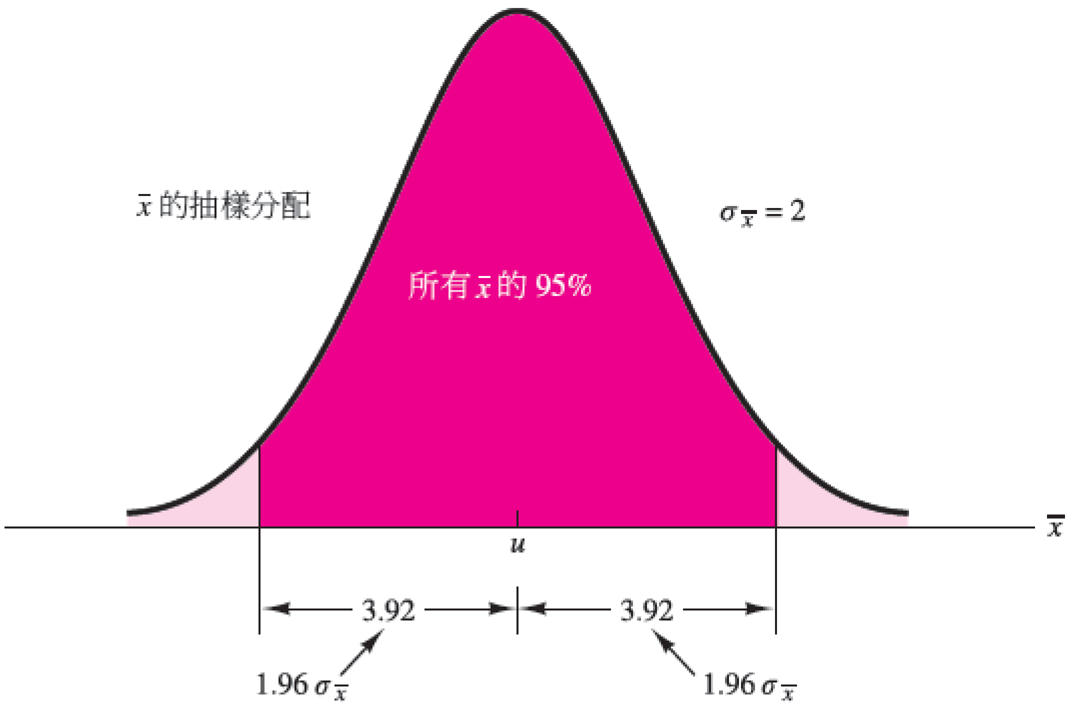
- **Note**: three important things about sampling distribution:
  - $E(\bar{x}) = \mu$
  - $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ (standard error, 抽樣分配的標準差, 又稱為標準誤)
  - Normal distributed, when n is large.

Interval Estimate of Population Mean: σ known

- We can conclude that the sampling distribution of $\bar{x}$ follows a normal distribution with a standard error of $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = 2$, and a mean of μ.
  - **Note:** according to the **standard normal probability table**, we find that 95% of the values of any normally distributed random variable are within ± 1.96 standard deviations of the mean.
- Thus, 95% of the $\bar{x}$ values must be within $\mu \pm 1.96\sigma_{\bar{x}}$.
- In the Lloyd's example, the sampling distribution of $\bar{x}$ is normally distributed with a standard error of $\sigma_{\bar{x}} = 2$.
- Because $\pm 1.96 \times 2 = \pm 3.92$, we can conclude that 95% of all $\bar{x}$ values obtained using a sample size of $n = 100$ will be within ± 3.92 of the population mean μ.
  - 洛伊德顧客平均購物的金額，在樣本大小為$n = 100$所得到的樣本，其樣本平均數有95%落在母體平均數±1.96*2=3.92的範圍內.

Interval Estimate of Population Mean: σ known

圖 8.2 $\bar{x}$ 的抽樣分配，與母體平均數 μ 的距離在 3.92 之內的樣本平均數所在的區域

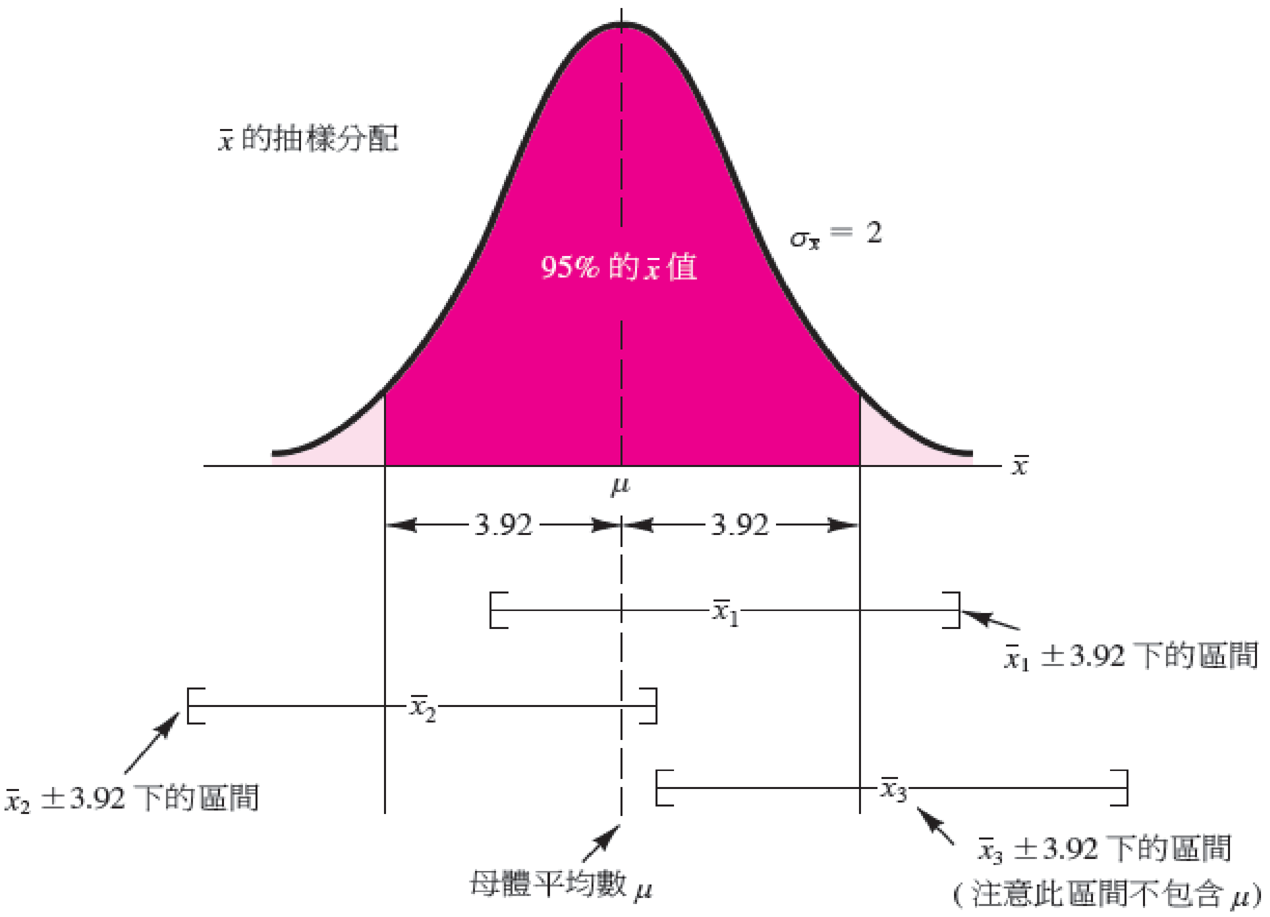


Interval Estimate of Population Mean: σ known

- The general form of an interval estimate of the population mean μ is:
  - $\bar{x} \pm$ Margin of error
  - Assume that we obtained a sample with a sample mean $\bar{x} = \$82$, then the interval estimate of the population mean is:
    - $\$82 \pm 3.92$
- The meaning of the interval estimate of the population mean μ:
  - For any sample mean $\bar{x}$, 95% of all possible intervals formed by subtracting 3.92 from $\bar{x}$ and adding 3.92 to $\bar{x}$ will include the population mean μ.

Interval Estimate of Population Mean: σ known

圖 8.3 樣本平均數 $\bar{x}_1$、$\bar{x}_2$ 和 $\bar{x}_3$ 形成的區間



Interval Estimate of Population Mean: σ known

- Assume that we obtained a sample with a sample mean $\bar{x} = \$82$, the interval estimate is:
  - $82 \pm 3.92 = (78.08, 85.92)$
- We are 95% confident that the interval 78.08 to 85.92 includes the population mean μ.
  - The interval has been established at the **95% confidence level (95%信賴水準)**.
  - The value 0.95 is referred to as the **confidence coefficient (信賴係數).**
  - The interval (78.08, 85.92) is called the **95% confidence interval (95%信賴區間)**.

Interval Estimate of Population Mean: σ known

- Interval estimate of a population μ when σ is known:
  - $\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
  - where $1-\alpha$ is the confidence coefficient.
  - and $z_{\alpha/2}$ is the z value providing an area of $\alpha/2$ in the upper tail of the standard normal probability distribution.
- **Ex.**, In the Lloyd's example, if we want to derive 95% confidence interval, and we know that $\bar{x} = 82$, $\sigma = 20$, and $n = 100$, then:
  - Step1: For a 95% confidence level, the confidence coeeficient is $1-\alpha = 0.95$, and thus, $\alpha = 0.05$.
  - Step2: Using the standard normal probability table (**標準常態機率表**), an area of $\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$ in the upper tail provides $z_{0.025} = 1.96$.
  - Step3: Compute the standard error of the sampling distribution: $\frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{100}} = 2$
  - Thus, the margin of error is $z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 3.92$.
  - The 95% confidence interval is $\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 82 \pm 1.96 * 2 = 82 \pm 3.92$, which is (78.08, 85.92)

Interval Estimate of Population Mean: σ known

- The **90%** confidence interval for the Lloyd's example is:
  - $\bar{x} \pm z_{\frac{0.1}{2}} \frac{\sigma}{\sqrt{n}} = 82 \pm 1.645 * 2 = 82 \pm 3.29$
  - The margin of error is 3.29, the 90% confidence interval is (78.71, 85.29)
- The **99%** confidence interval for the Lloyd's example is:
  - $\bar{x} \pm z_{\frac{0.01}{2}} \frac{\sigma}{\sqrt{n}} = 82 \pm 2.576 * 2 = 82 \pm 5.15$
  - The margin of error is 5.15, the 99% confidence interval is (76.85, 87.15)

表 8.1 常用的信賴水準及所對應的 $z_{\alpha/2}$ 值

信賴水準	α	$\alpha/2$	$z_{\alpha/2}$
90%	0.10	0.05	1.645
95%	0.05	0.025	1.960
99%	0.01	0.005	2.576

Practical Advice

- If the population follows a normal distribution, the confidence interval provided by $\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ is exact.
  - If we repeatedly generate 95% confidence intervals (一直重複以$n = 100$抽出樣本並計算信賴區間), exactly 95% of the intervals generated would contain the population mean.
- If the population does not follow a normal distribution, the confidence interval will be approximate.
  - In this case, the quality of the approximation depends on **both the distribution of the population and the sample size**.
- In most applications, a sample size of $n \geq 30$ is adequate when using to develop an interval estimate of a population mean.
  - If the population is not normally distributed but is **roughly symmetric**, sample sizes as small as 15 can be expected to provide good approximation.

Interval Estimate of Population Mean: σ unknown

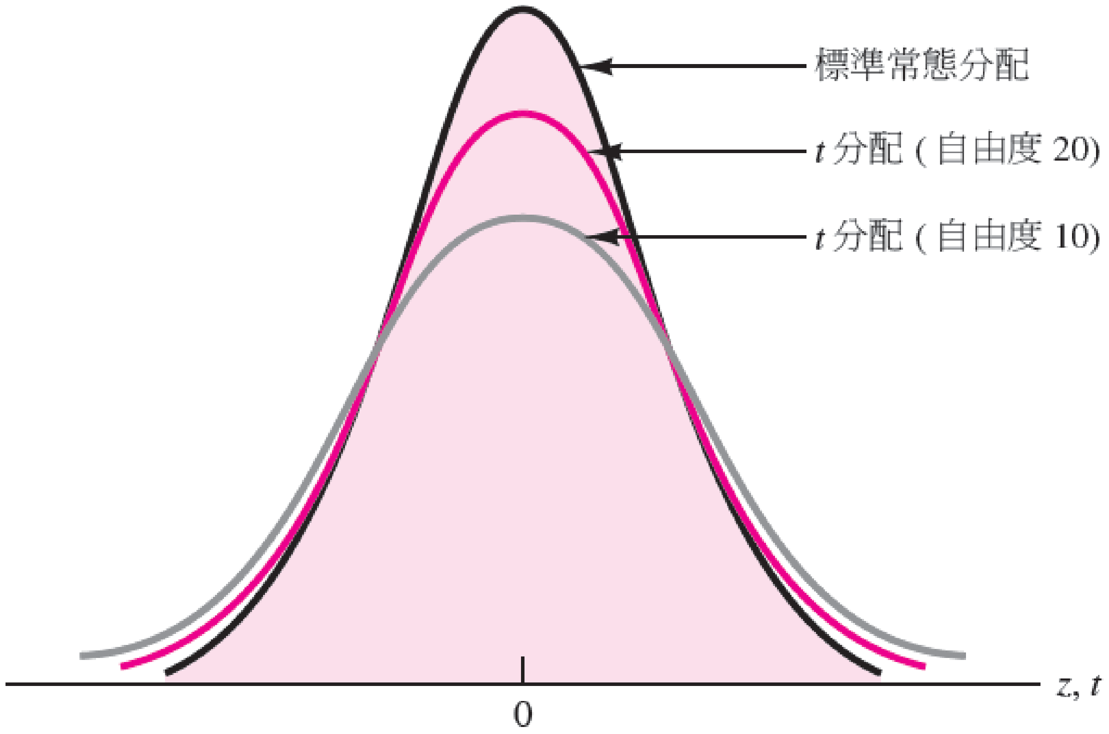
- When developing an interval estimate of a population mean, we usually do not have a good estimate of the population standard deviation.
  - In this case, we must use **the same sample** to estimate both μ and σ.
  - This situation represents the ***σ* unknown case**.
  - When s is used to estimate σ, the margin of error and the interval estimate for the population mean are based on a probability distribution known as the ***t* distribution**.
    - **Recall that** when σ is known, the interval estimate for the population mean is:
      - $\bar{x} \pm$ Margin of error
      - i.e., $\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

t distribution

- The mathematical development of the t distribution is based on the assumption of a normal distribution for the population we are sampling from.
- However, research shows that the t distribution can be applied in many situation where the population deviates significantly from normal.
- The t distribution is a family of similar probability distributions, with a specific t distribution depending on a parameter known as the degrees of freedom.
  - As the number of degrees of freedom increases, the difference between the t distribution and the standard normal distribution becomes smaller and smaller.
  - Note that a t distribution with **more degrees of freedom exhibits less variability** and more closely resembles the standard normal distribution.
  - The mean of any t distribution is **zero**.
  - We place a subscript on t to indicate the area in the upper tail of the t distribution.
    - $t_{0.025}$ indicates a 0.025 area in the upper tail of a t distribution.
    - $z_{0.025}$ indicates a 0.025 area in the upper tail of a standard normal distribution.
    - $t_{\alpha/2}$ represents a t value with an area of $\alpha/2$ in the upper tail of the t distribution.

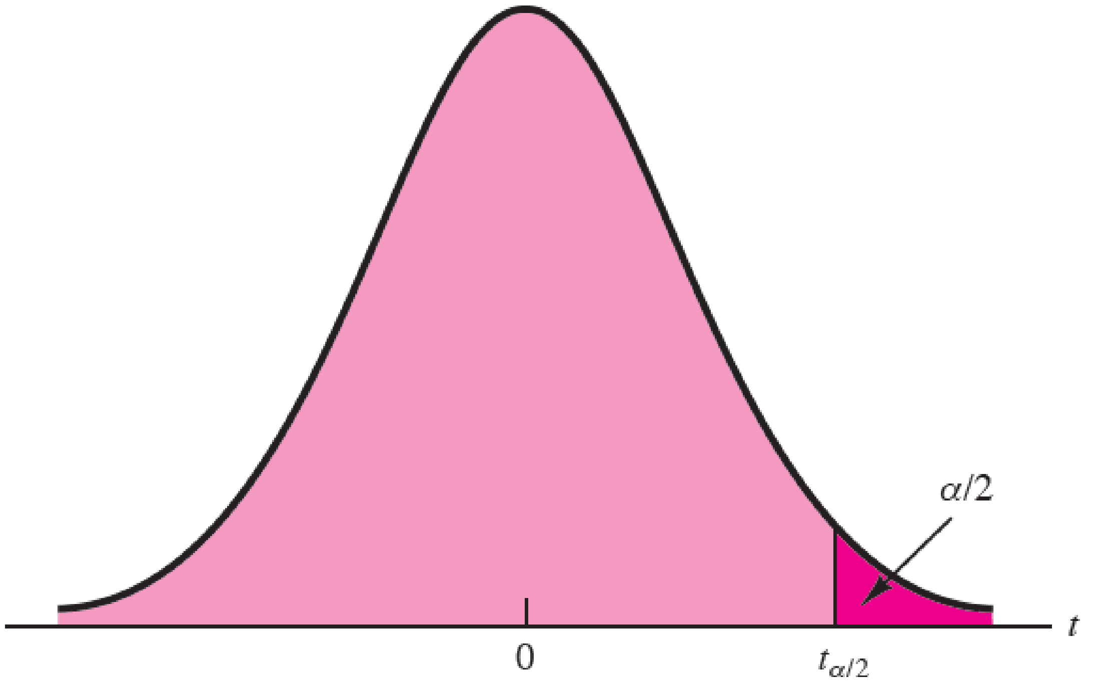
t distribution

圖 8.4 標準常態分配和 t 分配在自由度 10 與 20 時的比較圖



t distribution

圖 8.5 右尾面積或機率為 $\alpha/2$ 的 t 分配



t distribution

- We can look up t value using the t distribution table just as the standard normal distribution table.
  - For a t distribution with 9 degrees of freedom, $t_{0.025} = 2.262$
  - For a t distribution with 60 degrees of freedom, $t_{0.025} = 2.000$
  - As the degrees of freedom continue to increase, $t_{0.025} \to z_{0.025} = 1.96$
  - If the degree of freedom exceed 100, the standard normal z value provides a good approximation to the t value.

表 8.2 部分 t 分配表 *

	面積或機率（陰影部分）						
自由度	0.20	0.10	0.05	0.025	0.01	0.05	0.005
1	1.376	3.078	6.314	12.706	31.821	31.821	63.656
2	1.061	1.886	2.920	4.303	6.965	6.965	9.925
3	0.9780	1.638	2.353	3.182	4.541	4.541	5.841
4	0.9410	1.533	2.132	2.776	3.747	3.747	4.604
5	0.9200	1.476	2.015	2.571	3.365	3.365	4.032
6	0.9060	1.440	1.943	2.447	3.143	3.143	3.707
7	0.8960	1.415	1.895	2.365	2.998	2.998	3.499
8	0.8890	1.397	1.860	2.306	2.896	2.896	3.355
9	0.8830	1.383	1.833	2.262	2.821	2.821	3.250

t distribution

表 8.2 部分 t 分配表 *

60	0.8480	1.296	1.671	2.000	2.390	2.390	2.660
61	0.8480	1.296	1.670	2.000	2.389	2.389	2.659
62	0.8470	1.295	1.670	1.999	2.388	2.388	2.657
63	0.8470	1.295	1.669	1.998	2.387	2.387	2.656
64	0.8470	1.295	1.669	1.998	2.386	2.386	2.655
65	0.8470	1.295	1.669	1.997	2.385	2.385	2.654
66	0.8470	1.295	1.668	1.997	2.384	2.384	2.652
67	0.8470	1.294	1.668	1.996	2.383	2.383	2.651
68	0.8470	1.294	1.668	1.995	2.382	2.382	2.650
69	0.8470	1.294	1.667	1.995	2.382	2.382	2.649
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

表 8.2 部分 t 分配表 *

90	0.8460	1.291	1.662	1.987	2.368	2.368	2.632
91	0.8460	1.291	1.662	1.986	2.368	2.368	2.631
92	0.8460	1.291	1.662	1.986	2.368	2.368	2.630
93	0.8460	1.291	1.661	1.986	2.367	2.367	2.630
94	0.8450	1.291	1.661	1.986	2.367	2.367	2.629
95	0.8450	1.291	1.661	1.985	2.366	2.366	2.629
96	0.8450	1.290	1.661	1.985	2.366	2.366	2.628
97	0.8450	1.290	1.661	1.985	2.365	2.365	2.627
98	0.8450	1.290	1.661	1.984	2.365	2.365	2.627
99	0.8450	1.290	1.660	1.984	2.364	2.364	2.626
100	0.8450	1.290	1.660	1.984	2.364	2.364	2.626
∞	0.8420	1.282	1.645	1.960	2.326	2.326	2.576

* 註：完整的 t 分配表請參考附錄 B 的表 2。

Interval Estimate of Population Mean: σ unknown

- Interval estimate of population mean for the **σ known case** is:
  - $\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
    - where σ is the population standard deviation, $(1-\alpha)$ is the confidence coefficient, $z_{\alpha/2}$ is the z value providing an area of $\alpha/2$ in the upper tail of the standard normal probability distribution.
- Interval estimate of population mean for the **σ unknown case** is:
  - $\bar{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$
    - Where s is the sample standard deviation, $(1-\alpha)$ is the confidence coefficient, $t_{\alpha/2}$is the t value providing an area of $\alpha/2$ in the upper tail of the t distribution with $n-1$ degrees of freedom.
    - The expression for the sample standard deviation is: $s = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$
      - Degrees of freedom refer to the number of independent pieces of information that go into the computation of $\sum(x_i-\bar{x})^2$.
      - 知道$n-1$個$x_i-\bar{x}$的值, 最後一個值可由$\sum(x_i-\bar{x})=0$而得, 只有$n-1$個$x_i-\bar{x}$值是獨立的, 因此自由度只有$n-1$.

Interval Estimate of Population Mean: σ unknown

- **Ex.**, Consider a study designed to estimate the mean credit card debt for the population of U.S. households. A sample of $n = 70$ households provided the credit card balances shown in Table 8.3.
  - For this situation, **no previous estimate** of the population standard deviation σ is available. Thus, the sample data must be used to estimate both the population mean and the population standard deviation.

表 8.3 70 個家庭的信用卡帳戶餘額

9430	14661	7159	9071	9691	11032
7535	12195	8137	3603	11448	6525
4078	10544	9467	16804	8279	5239
5604	13659	12595	13479	5649	6195
5179	7061	7917	14044	11298	12584
4416	6245	11346	6817	4353	15415
10676	13021	12806	6845	3467	15917
1627	9719	4972	10493	6191	12591
10112	2200	11356	615	12851	9743
6567	10746	7117	13627	5337	10324
13627	12744	9465	12557	8372	
18719	5742	19263	6232	7445	

Interval Estimate of Population Mean: σ unknown

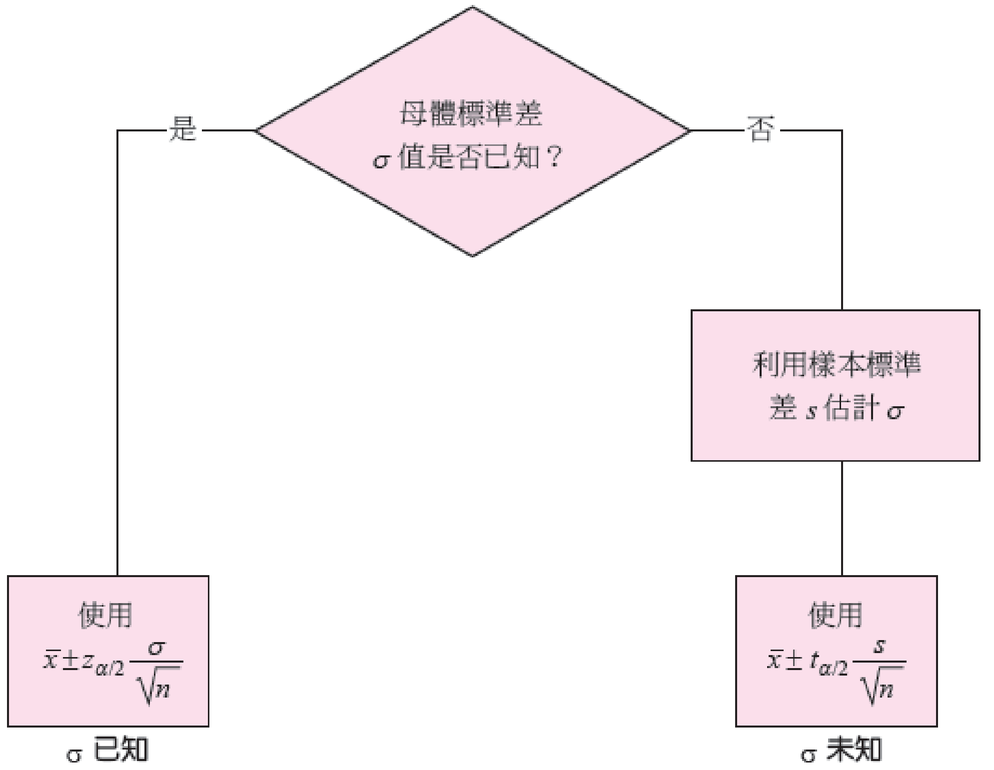
- The confidence level (信賴水準) is 95%, the degree of freedom is $n-1=69$:
  - $(1-\alpha)=95\%$, $\alpha=5\%$, $\frac{\alpha}{2}=2.5\%$.
  - Using the t distribution table: $t_{0.025}=1.995$.
- The interval estimate of population mean is:
  - $\bar{x} \pm t_{0.025}\frac{s}{\sqrt{n}} = 9312 \pm 1.995 * \frac{4007}{\sqrt{70}} = 9312 \pm 955$
  - Using table 8.3, we compute the sample mean is $\bar{x}=\$9312$, and the sample standard deviation $s=\$4007$.
  - Thus, the margin of error is 955, the 95% confidence interval (信賴區間) is (8357, 10267).
  - That is, we are 95% confident that the mean credit card balance for the population of **all households** is between $8357 and $10267.

Practical Advice

- If the population follows a normal distribution, the confidence interval provided by $\bar{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$ is exact and can be used **for any sample size**.
- If the population does not follow a normal distribution, the confidence interval provided by $\bar{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$ will be approximate.
  - In this case, the quality of the approximation depends on **both the distribution of the population and the sample size**.
- In most applications, a sample size of $n \geq 30$ is adequate when using to develop an interval estimate of a population mean.
  - However, if the population is highly skewed or contain outliers, most statisticians would recommend increasing the sample size to 50 or more.
  - If the population is not normally distributed but is roughly symmetric, sample sizes as small as 15 can be expected to provide good approximate confidence intervals.

區間估計程序

圖 8.7 母體平均數區間估計程序摘要



Exercise (Using a Small Sample)

- **Ex.**, Scheer Industries is considering a new computer-assisted program to train maintenance employees to do machine repairs. In order to fully evaluate the program, the director of manufacturing requested an estimate of the population mean time required for maintenance employees to complete the computer-assisted training.
- A sample of 20 employees is selected, with each employee in the sample completing the training program. Data on the training time in days for the 20 employees are shown in Table 8.4. A histogram of the sample data appears in Figure 8.6.

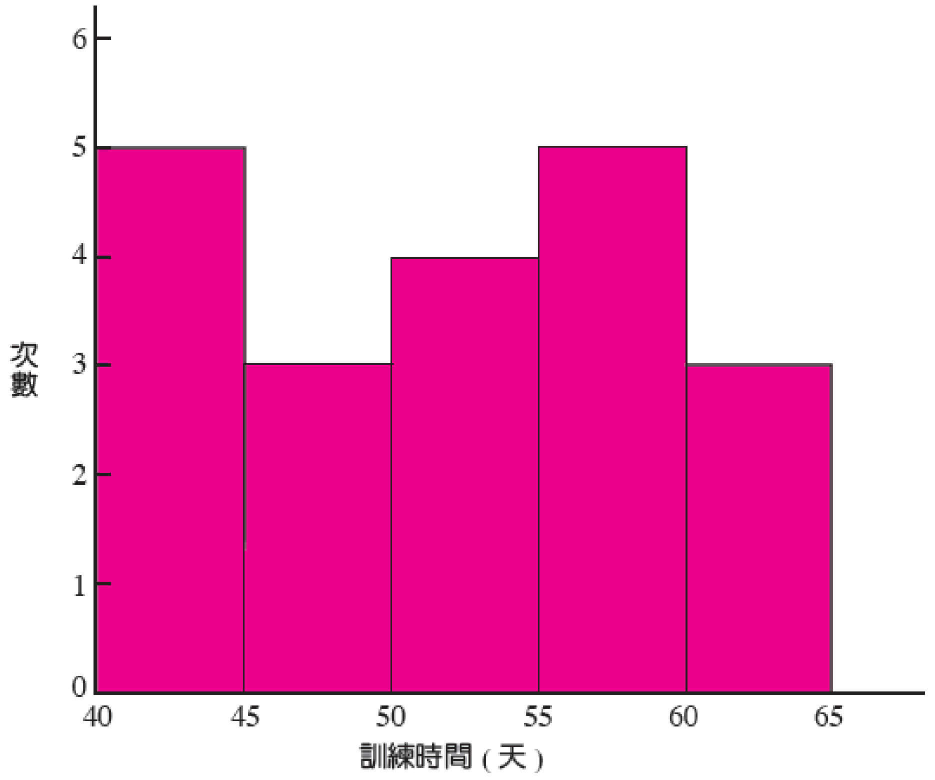
表 8.4 Scheer 工業 20 個樣本的電腦輔助訓練計畫的天數

52	59	54	42
44	50	42	48
55	54	60	55
44	62	62	57
45	46	43	56

Exercise (Using a Small Sample)

- Histogram of Training Times for the Scheer Industries Sample

圖 8.6 Scheer 工業電腦輔助訓練計畫天數的直方圖



Exercise (Using a Small Sample)

- The population distribution is **not normal**, yet we do not see any evidence of skewness or outliers (**roughly symmetric**). Therefore, an interval estimate based on the t distribution appears acceptable for the sample of 20 employees.
- The sample mean: $\bar{x} = \frac{\sum x_i}{n} = \frac{1030}{20} = 51.5$ days
- The sample standard deviation: $s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{\boldsymbol{n-1}}} = \sqrt{\frac{889}{20-1}} = 6.84$ days
- For a 95% confidence interval: $\boldsymbol{(1-\alpha) = 0.95, \alpha = 0.05}$
- The degree of freedom is 19, using t distribution table: $t_{\frac{\alpha}{2}} = t_{0.025} = 2.093$
- The 95% confidence interval is $\bar{x} \pm t_{0.025} \frac{s}{\sqrt{\boldsymbol{n}}}$
  - $51.5 \pm 2.093 \left(\frac{6.84}{\sqrt{20}}\right) = 51.5 \pm 3.2$
  - The 95% confidence interval is (48.3, 54.7)