

Content

1-1 Descriptive and Inferential Statistics	3
1-2 Variables and Types of Data	3
1-3 Data Collection and Sampling Techniques	3
1-4 Experimental Design	4
2-1 Organizing Data	5
2-2 Histograms, Frequency Polygons, and Ogives	5
2-3 Other Types of Graphs	6
2-4 Paired Data and Scatter Plots	8
3-1 Measures of Central Tendency	9
3-2 Measures of Variation	10
3-3 Measures of Position	11
3-4 Exploratory Data Analysis	12
4-1 Sample Space and Probability	13
4-2 Addition Rules for Probability	14
4-3 Multiplication Rules and Conditional Probability	14
4-4 Counting Rules	15
5-1 Probability Distribution	16
5-2 Mean, Variance, Standard Deviation, Expectation	16
5-3 The Binomial Distribution	16
6-1 Normal Distribution	17
6-2 Applications of the Normal Distribution	17
6-3 Central Limit Theorem	18
7-1 Confidence Intervals	19
7-2 Confidence Intervals for Mean, σ is known	19
7-3 Confidence Intervals for Mean, σ is unknown	19
7-4 Confidence Intervals and Sample Size for proportion	20
7-5 Confidence Intervals for Variances and Standard Deviation	20
8 Hypothesis Testing	21
8-1 steps in hypothesis testing – traditional method	22
8-2 z test for a mean	22
8-3 t test for a mean	23
8-4 z test for a proportion	23
8-5 χ^2 test for a variance or standard deviation	23
8-6 additional topics	24
9-1 Testing the difference between two parameters	25
9-2 Testing the difference between two means: z test	25
9-3 Testing the difference between two means of independent sample: t test	25
9-4 Testing the difference between two means: dependent sample	26
9-5 Testing the difference between proportion	26
9-6 Testing the difference between two variance	27

10-1 Correlation	28
10-2 Regression.....	29
10-3 Coefficient of determination & standard error of the estate	29
11-1 Test for Goodness of Fit	30
11-2 Tests Using Contingency Table	31
11-3 ANOVA	31

1-1 Descriptive and Inferential Statistics

population (母體)	sample (樣本)
所有的研究對象	從母體中抽出的子集
欲研究全淡江學生的平均身高 母體：淡江所有的學生 樣本：某個班的全體學生、隨機訪問的淡江學生等	
descriptive statistics (敘述性統計)	inferential statistics (推論性統計)
描述資料的結果	利用樣本推論母體
收集全班的身高，經過計算得到平均身高是 160 cm	根據全班的平均身高，推論全校的平均身高是 160 cm

1-2 Variables and Types of Data

qualitative (屬值)	quantitative (屬量)	
可以根據特徵或屬性區分成不同類別	可以被計算或測量	
性別、地點	身高、體重、年齡	
	discrete (離散)	continuous (連續)
	被數出來的，不可再分割	被測量出來的，兩數值之間可以再細分
	◆ 人數(不會有 1.5 人這種) ◆ 美元(最小單位是分)	◆ 溫度 ◆ 體重

nominal (名目)	ordinal (順序)	interval (區間)	ratio (比率)
可以被分文互斥的類別，且類別之間不可排序	可以被分為可排序的類別，但類別之間的差距無意義	單位間的差距有意義，0 沒有意義 (可以加減、不能乘除)	基於已知的測量單位、0 有意義 (可以乘除)
◆ 顏色 ◆ 性別	◆ 成績(A, B, C, D, F) ◆ 評分(不好、普通、好)	◆ 溫度(0 只是測量尺度上的一個點，不代表沒有；不會說 40 度是 20 度的兩倍熱) ◆ IQ (IQ100 不代表比 IQ50 聰明兩倍) ◆ SAT 成績(對 3~5 題一個分數、6~8 題一個分數)	◆ 體重(0 公斤表示體重計上什麼都沒有) ◆ 薪水(時薪 50 美元是時薪 25 美元的兩倍)

1-3 Data Collection and Sampling Techniques

random (隨機)	systematic (系統)	stratified (分層)	cluster (集群)
母體中的所有成員被抽到的機率相同	每 k 個抽出一個	把母體根據特徵進行分割後，各組隨機抽出幾個 (組間變異大，組內變異小)	把母體分群後，隨機抽出一或多個群，使用該群所有的成員 (組間變異小，組內變異大)

1-4 Experimental Design

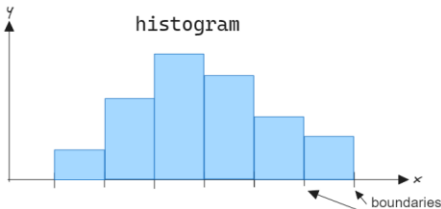
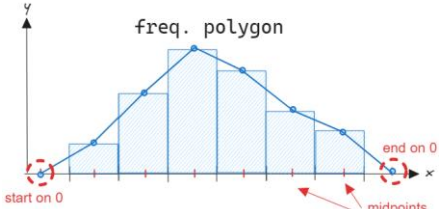
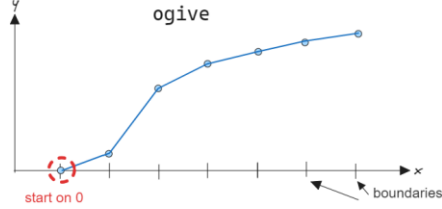
observational study	experimental study
實驗者僅觀察過去或正在進行的，基於這些觀察來得到結論	實驗者會控制某個變數來判斷它如何影響其他變數
independent / explanatory variable (獨立/解釋變數)	dependent / outcome variable (相依/結果變數)
研究者操縱的變數	研究者想比較的結果

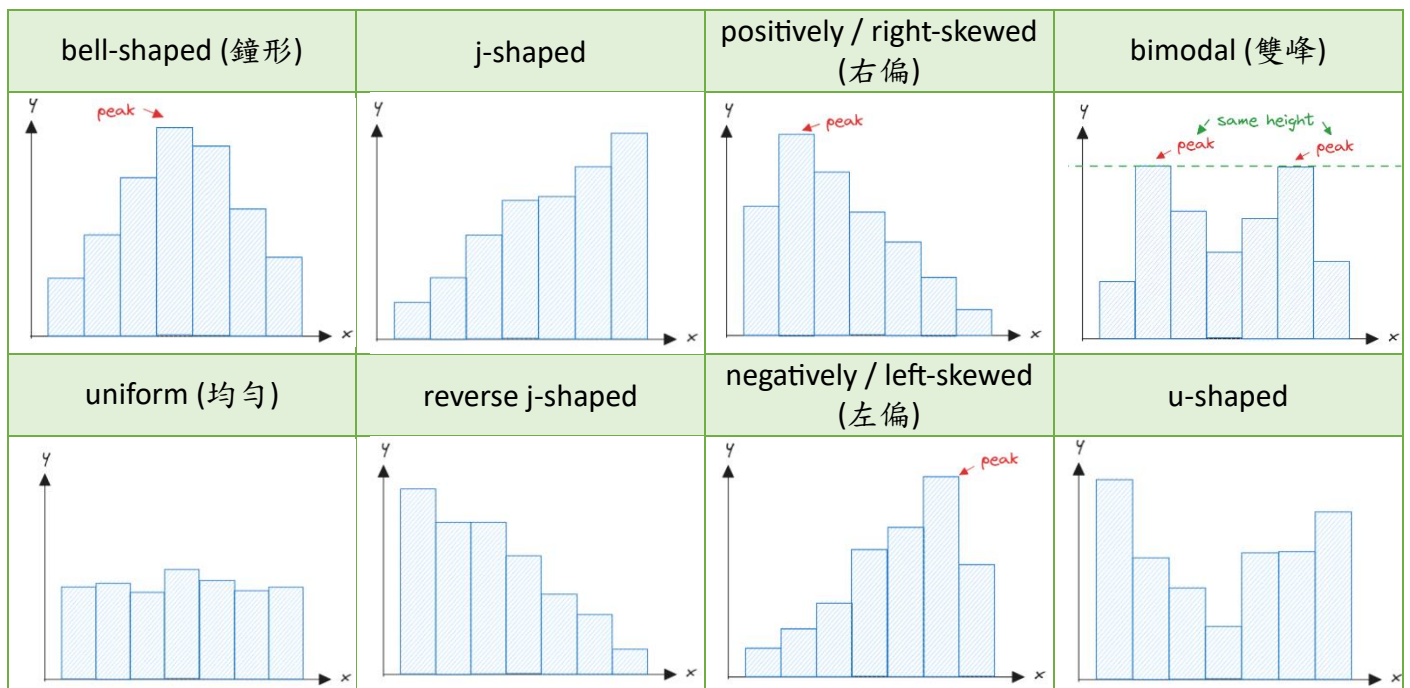
2-1 Organizing Data

frequency distribution (次數分配)											
透過分類和計算次數，將原始資料整理成表格資料											
categorical (類別)	grouped (分組)	open-ended (開放)	ungrouped								
當類別屬於名目 (nominal)、順序 (ordinal) 尺度時	1. highest (H), lowest (L) (最大最小值)	當第一類沒有下限或最後一類沒有上限時 eg. (p.46) <table><tr><th>Age</th><th>Freq.</th></tr><tr><td>10 – 20</td><td>3</td></tr><tr><td>21 – 31</td><td>6</td></tr><tr><td>32 up</td><td>10</td></tr></table>	Age	Freq.	10 – 20	3	21 – 31	6	32 up	10	當全距(range)範圍很小時 eg. (p.49) Example 2-3
	Age		Freq.								
	10 – 20		3								
	21 – 31		6								
	32 up		10								
2. range (R) = H – L (全距)											
3. number of classes											
4. width = R / number of classes round up (無條件進位)											
5. start point + width = lower limits start point: 通常是最小值或是方便的數字											
	6. upper limits										
	7. boundaries										
cumulative frequency distribution (累積次數分配)											
計算≤某個數值(通常為上界)的次數											

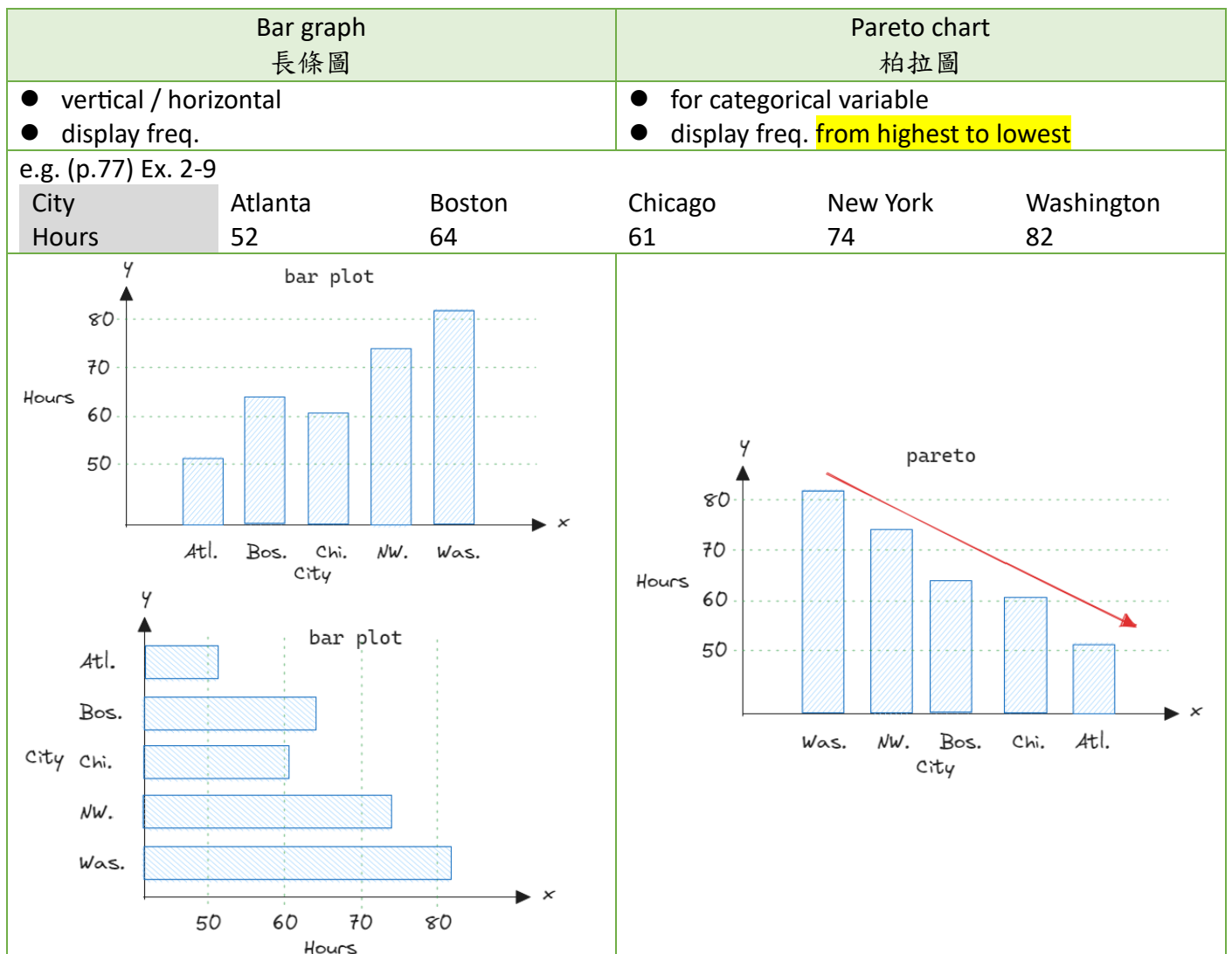
- lower / upper limits: 上下限
- lower / upper boundaries: 上下界
- freq. : frequency

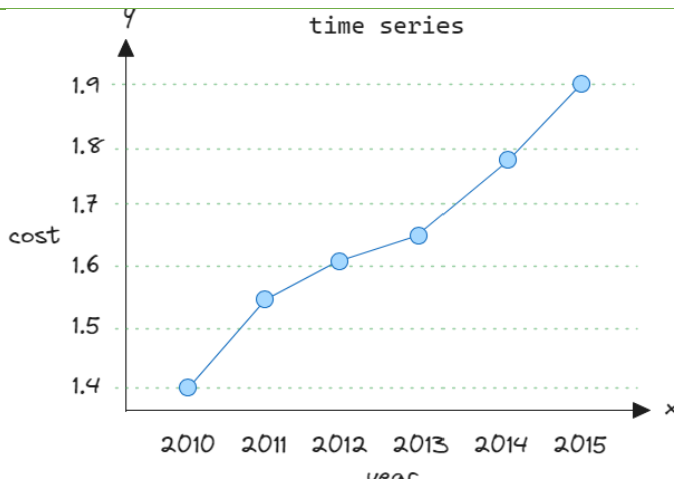
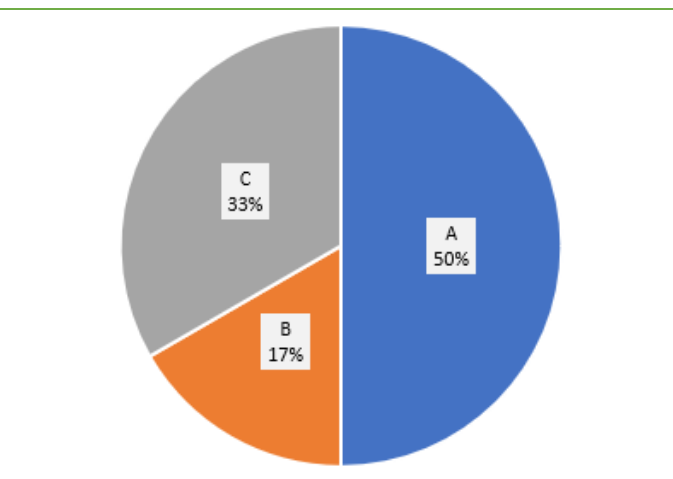
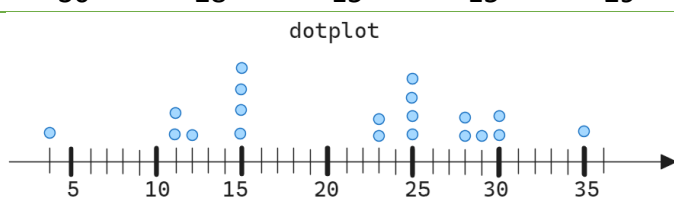
2-2 Histograms, Frequency Polygons, and Ogives

histogram 直方圖	frequency polygon 次數多邊圖	ogive (cumulative freq. graph) 肩形圖 (累積次數分配圖)
		
<ol style="list-style-type: none"> label x, y axis label class boundaries for histogram, ogive label midpoints for frequency polygon plot freq. for each class draw vertical bars for histogram draw lines for frequency polygon, ogive 		
frequency		relative frequency
計算每個類別的 次數		計算每個類別的 比例 當想比較的兩個類別的母體差很多時使用 eg. (p. 61)

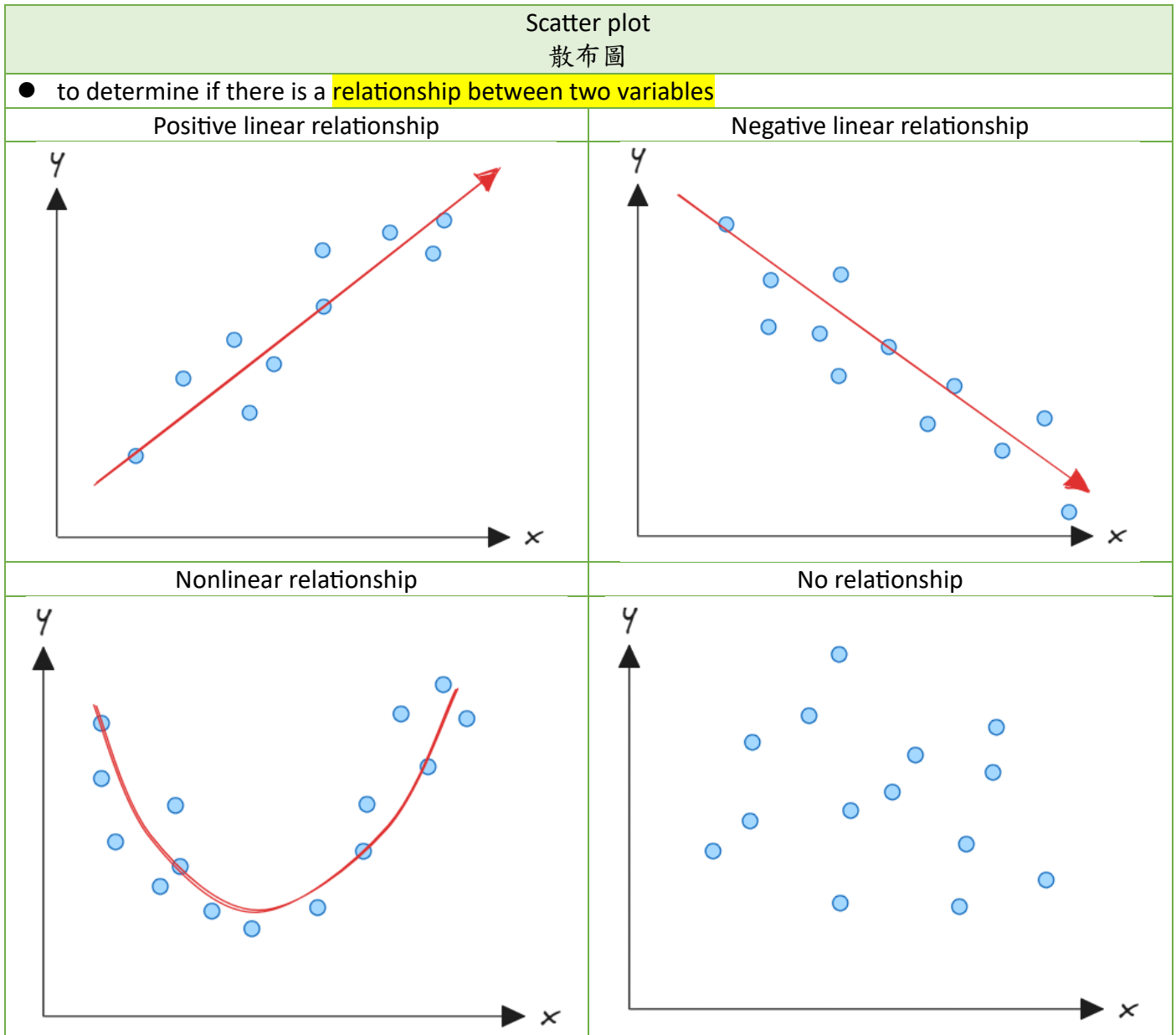


2-3 Other Types of Graphs



Time series graph 時間數列	Pie graph 圓餅圖																														
<ul style="list-style-type: none">display data over a specific period of timetrend or pattern: ascending / descendingslope or steepness (陡度)	<ul style="list-style-type: none">for categorical variabledisplay percentage of freq.																														
e.g. (p.78) Ex. 2-10	e.g.																														
<table><tr><th>Year</th><th>2010</th><th>2011</th><th>2012</th><th>2013</th><th>2014</th><th>2015</th></tr><tr><th>Cost</th><td>1.40</td><td>1.55</td><td>1.61</td><td>1.65</td><td>1.78</td><td>1.90</td></tr></table>	Year	2010	2011	2012	2013	2014	2015	Cost	1.40	1.55	1.61	1.65	1.78	1.90	<table><tr><th>Brand</th><th># of branches</th><th>Degrees =f/n*360°</th><th>% =f/n*100</th></tr><tr><td>A</td><td>15</td><td>180°</td><td>50%</td></tr><tr><td>B</td><td>5</td><td>60°</td><td>16.7%</td></tr><tr><td>C</td><td>10</td><td>120°</td><td>33.3%</td></tr></table>	Brand	# of branches	Degrees =f/n*360°	% =f/n*100	A	15	180°	50%	B	5	60°	16.7%	C	10	120°	33.3%
Year	2010	2011	2012	2013	2014	2015																									
Cost	1.40	1.55	1.61	1.65	1.78	1.90																									
Brand	# of branches	Degrees =f/n*360°	% =f/n*100																												
A	15	180°	50%																												
B	5	60°	16.7%																												
C	10	120°	33.3%																												
																															
Dotplot 點圖	Stem and leaf plot 莖葉圖																														
<ul style="list-style-type: none">plot as a point above horizontal axis	<ul style="list-style-type: none">combination of sorting and graphingretaining actual data																														
e.g.																															
<table><tr><td>4</td><td>11</td><td>12</td><td>23</td><td>25</td></tr><tr><td>30</td><td>28</td><td>15</td><td>15</td><td>29</td></tr></table>	4	11	12	23	25	30	28	15	15	29	<table><tr><td>11</td><td>15</td><td>25</td><td>30</td><td>35</td></tr><tr><td>28</td><td>23</td><td>25</td><td>15</td><td>25</td></tr></table>	11	15	25	30	35	28	23	25	15	25										
4	11	12	23	25																											
30	28	15	15	29																											
11	15	25	30	35																											
28	23	25	15	25																											
	<table><tr><td>0</td><td>4</td></tr><tr><td>1</td><td>1 1 2 5 5 5 5</td></tr><tr><td>2</td><td>3 3 5 5 5 5 8 8 9</td></tr><tr><td>3</td><td>0 0 5</td></tr></table>	0	4	1	1 1 2 5 5 5 5	2	3 3 5 5 5 5 8 8 9	3	0 0 5																						
0	4																														
1	1 1 2 5 5 5 5																														
2	3 3 5 5 5 5 8 8 9																														
3	0 0 5																														

2-4 Paired Data and Scatter Plots



Addition

Correlation vs. causation (相關性 vs. 因果關係)

- [Correlation Does Not Imply Causation: A One Minute Perspective on Correlation vs. Causation](#)
- [How Ice Cream Kills! Correlation vs. Causation](#)

相關性不能直接表示兩個變數之間有因果關係，例如冰淇淋的銷量和凶殺案的數量是正相關，不能表示冰淇淋會導致殺人；他們之間可能有隱藏的因素，例如天氣，會同時影響冰淇淋的銷量和凶殺案的數量。

- [Correlation CAN Imply Causation! | Statistics Misconceptions](#)

但相關性可以評估兩個變數之間的因果關係。

因果關係需要考慮事件發生的先後順序，以及可能會影響到他們的隱藏因素。

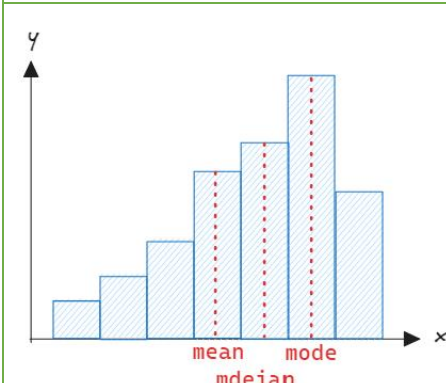
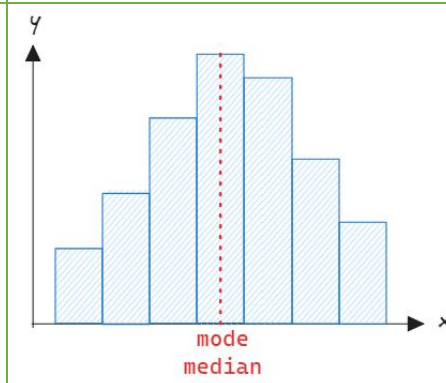
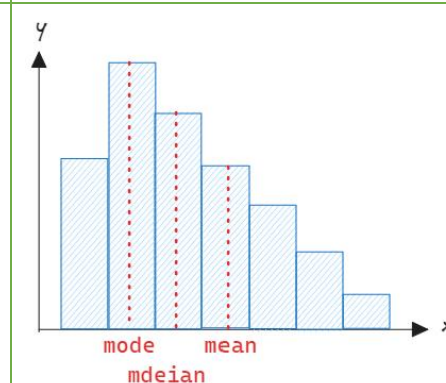
3-1 Measures of Central Tendency

Statistic 統計量	Parameter 參數
● 使用樣本資料計算得到	● 使用所有母體資料計算得到

Central Tendency		
Mean / arithmetic average 平均數 / 算術平均數	Weighted mean 加權平均數	Midrange 中程數
Sample mean: $\bar{X} = \frac{\sum X}{n}$ Population mean: $\mu = \frac{\sum X}{N}$	$\bar{X} = \frac{\sum wX}{\sum w}$	$MR = \frac{\max - \min}{2}$ ● Affected by extreme value
Median 中位數	Mode 眾數	
1. 對資料進行排序 2. 計算資料個數 n 3. n 是奇數，則中間值即為中位數 $MD = X_{\frac{n+1}{2}}$ n 是偶數，則中間的兩個數字取平均即為中位數 $MD = (X_{\frac{n}{2}} + X_{\frac{n}{2}+1}) / 2$	次數最多即為眾數 ● 一個眾數: unimodal (單峰) ● 兩個眾數: bimodal (雙峰) ● 多於兩個: multimodal 當沒有資料出現超過 1 次(重複出現)，即沒有眾數，而不是眾數為 0 ■ 以溫度為例，0 是一個實際值	

Rounding Rule

- 在得到最終答案時才進行四捨五入，如果在中間計算過程就簡化，答案和實際值的誤差會變大；但在實際計算過程中不方便保留所有的小數位數，因此會選擇簡化到足夠的位數(通常 3, 4 位)
 - 假設答案要求取到小數第 2 位，則中間計算過程至少需要取到小數第 3 位
- 平均數通常取到原始資料後一位
 - e.g. 原始資料 (21, 34, 54) → 平均 (36.3)
 - e.g. 原始資料 (20.3, 35.2, 51.7) → 平均 (35.73)

Distribution Shape		
Left-skewed Mode > median > mean	Symmetric Mode = median = mean	Right-skewed Mode < median < mean
		

3-2 Measures of Variation

Variation			
Range 全距	$R = \max - \min$		
	Population	Sample	Grouped data
Variance 變異數	$\sigma^2 = \frac{\sum(X - \mu)^2}{N}$	$s^2 = \frac{\sum(X - \bar{X})^2}{n-1}$ $= \frac{n(\sum X^2) - (\sum X)^2}{n(n-1)}$	$s^2 = \frac{n(\sum f \cdot X_m^2) - (\sum f \cdot X_m)^2}{n(n-1)}$
Standard deviation 標準差	$\sigma = \sqrt{\sigma^2}$ $= \sqrt{\frac{\sum(X - \mu)^2}{N}}$	$s = \sqrt{s^2}$ $= \sqrt{\frac{n(\sum X^2) - (\sum X)^2}{n(n-1)}}$	$s = \sqrt{\frac{n(\sum f \cdot X_m^2) - (\sum f \cdot X_m)^2}{n(n-1)}}$
Coefficient of variation 變異係數	$CVar = \frac{\sigma}{\mu} \cdot 100$	$CVar = \frac{s}{\bar{X}} \cdot 100$	<ul style="list-style-type: none"> ● 比較相同測量單位的 2 個樣本可以直接使用變異數和標準差 ● 比較不同單位的則要使用變異係數
Range rule of thumb 範圍經驗法則	$s \approx \frac{range}{4}$	$min = \bar{X} - 2s$	$max = \bar{X} + 2s$
● 只是粗略估計，當分配是單峰(僅一個眾數)且對稱時才能用			

Chebyshev's Theorem 柴比雪夫定理	落在 k 倍標準差內的比例至少有 $1 - \frac{1}{k^2}$ $k > 1$ (不一定要整數)	● 不限制分布形狀
Empirical rule 經驗法則	68% 的資料落在 1 倍標準差內 95% 的資料落在 2 倍標準差內 99.7% 的資料落在 3 倍標準差內	● 當分配對稱(或稱常態)時可用

- [推導] 變異數: $\frac{\sum(X - \bar{X})^2}{n-1} \rightarrow \frac{n(\sum X^2) - (\sum X)^2}{n(n-1)}$

$$\begin{aligned}
 & \frac{\sum(X - \bar{X})^2}{n-1} \\
 &= \frac{\sum(X^2 - 2X\bar{X} + \bar{X}^2)}{n-1} \quad (a-b)^2 = a^2 - 2ab + b^2 \\
 &= \frac{\sum X^2 - 2\bar{X}\sum X + n\bar{X}^2}{n-1} \quad \sum aX = a\sum X, a \text{ is constant} \\
 &= \frac{\sum X^2 - 2n\bar{X}^2 + n\bar{X}^2}{n-1} \quad \bar{X} = \frac{\sum X}{n} \rightarrow \sum X = n\bar{X} \\
 &= \frac{\sum X^2 - n\bar{X}^2}{n-1} \quad \bar{X}^2 = \frac{(\sum X)^2}{n^2} \rightarrow n\bar{X}^2 = \frac{(\sum X)^2}{n} \\
 &= \frac{n\sum X^2 - (\sum X)^2}{n(n-1)}
 \end{aligned}$$

● 線性轉換如何影響平均和標準差

$$Y = aX + b, a, b \text{ are constant}$$

$$\bar{X} = \frac{\sum X}{n}, s_X^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{\sum (aX + b)}{n} = \frac{a\sum X + nb}{n} = a \frac{\sum X}{n} + b = a\bar{X} + b$$

$$s_Y^2 = \frac{\sum (Y - \bar{Y})^2}{n - 1} = \frac{\sum [aX + b - a\bar{X} - b]^2}{n - 1} = \frac{\sum [a(X - \bar{X})]^2}{n - 1} = a^2 \frac{\sum (X - \bar{X})^2}{n - 1} = a^2 s_X^2$$

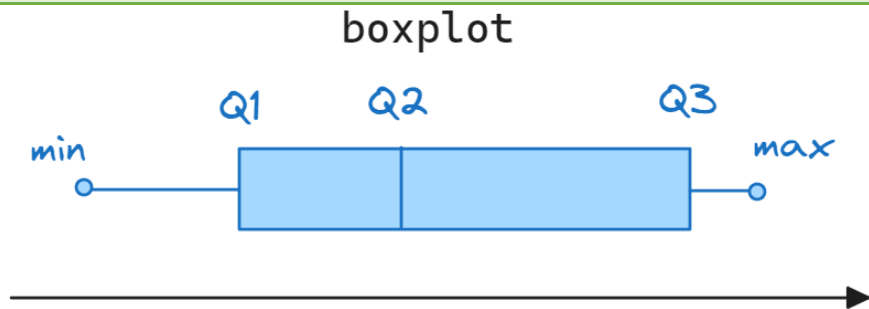
$$s_Y = \sqrt{s_Y^2} = \sqrt{a^2 s_X^2} = a s_X$$

3-3 Measures of Position

Measure of Position												
Standard score 標準分數	Percentile 百分位數		Quartile 四分位數	Decile 十分位數								
Sample: $z = \frac{X - \bar{X}}{s}$	Given data: $percentile = \frac{\text{比 } X \text{ 小的個數} + 0.5}{\text{總個數}}$		$\frac{Q_1 \quad Q_2 \quad Q_3}{p_{25} \quad p_{50} \quad p_{75}}$ $= median$ Note: 當 n 是奇數時， 使用百分位數算法， 小樣本誤差會變大 (ex. 28)	$D_1 = p_{10}$ $D_2 = p_{20}$ \vdots $D_9 = p_{90}$								
Population: $z = \frac{X - \mu}{\sigma}$	Given percentile: $c = \frac{n \cdot p}{100}, n = \text{總個數}, p = \text{百分位數}$ c 是整數，則數值為 c^{th} 和 $(c + 1)^{th}$ 的平均； c 不是整數，則無條件進位，數值為進位後的 c^{th}		Interquartile range: 四分位距 $IQR = Q_3 - Q_1$									
	Percentile graph: <table><tr><td>A</td><td>B</td><td>C</td><td>D</td></tr><tr><td>Class boundaries</td><td>Freq.</td><td>Cumulative Freq.</td><td>Cumulative percent</td></tr></table>		A	B	C	D	Class boundaries	Freq.	Cumulative Freq.	Cumulative percent	Outlier 離群值: $X < Q_1 - 1.5 \cdot IQR$ or $X > Q_3 + 1.5IQR$	
A	B	C	D									
Class boundaries	Freq.	Cumulative Freq.	Cumulative percent									

3-4 Exploratory Data Analysis

Boxplot 盒形圖 / 箱型圖



透過 boxplot 可以看資料分布的偏態，分為兩個部分

中位數接近箱子的中間，則資料接近對稱

中位數接近箱子的左邊，則資料右偏

中位數接近箱子的右邊，則資料左偏

兩端的線段長度大概相同，則資料接近對稱

左邊 < 右邊，則資料右偏

左邊 > 右邊，則資料左偏

如果兩部分分別得到不同的偏態，則以影響較大的為主

e.g. 上圖

中位數偏左 → 右偏，左邊線段較長 → 左偏

Min 到 Q1 的距離和 Q1 到 Q2 的距離相當；Q2 到 Q3 的距離大於 Q3 到 max 的距離

因此資料為右偏

4-1 Sample Space and Probability

Probability experiment 機率實驗	Outcome 結果	Sample space 樣本空間	Event 事件
一種隨機的過程，會導致明確的結果，稱為 outcome	單一機率實驗的結果	機率實驗中所有可能的結果的集合	包含結果的集合

e.g. (p.194)

Experiment	Sample space
擲一枚硬幣	頭, 尾
擲骰子	1, 2, 3, 4, 5, 6
擲兩枚硬幣	頭頭, 頭尾, 尾頭, 尾尾

Experiment	Event
投擲一骰子	拿到 6 點的事件稱 simple event 拿到奇數點(即 1,3,5 點)的事件稱 compound event

Tree diagram 樹狀圖	e.g. 擲兩枚硬幣，頭為 H、尾為 T
用來列出機率實驗中，所有可能的結果	<pre> graph LR A["first coin"] --> B["H"] A --> C["T"] B --> D["second coin"] B --> E["T"] C --> F["H"] C --> G["T"] D --> H["outcome"] E --> I["outcome"] F --> J["outcome"] G --> K["outcome"] H --- L["HH"] I --- M["HT"] J --- N["TH"] K --- O["TT"] </pre>

Classical probability 古典機率	Empirical probability 經驗機率	Subjective probability 主觀機率
所有在樣本空間的結果發生的機率相同，稱 equally likely events	經過重複多次相同的實驗後得到的結果來計算機率	根據個人經驗得到的機率
$P(E) = \frac{n(E)}{n(S)}$ $n(E)$ =E 事件中的結果個數 $n(S)$ =樣本空間的結果個數	$P(E) = \frac{f}{n}$ f =發生 E 事件的次數 n =總次數	
	Law of large numbers 當擲一枚硬幣一次時，拿頭的機率為 1/2，但擲一枚硬幣 50 次，很難剛好拿頭 25 次 當重複的次數增加時，經驗機率得到的機率會趨近於實際(即 $P(E) \rightarrow 1/2$)	
1. $0 \leq P(E) \leq 1$ 2. 所有樣本空間中的機率加總為 1 3. 事件 E 不會發生則機率為 0 4. 事件 E 一定會發生則機率為 1		

Complement 補集

事件 E 的補集記為 \bar{E} ，代表樣本空間中除了事件 E 以外的所有集合

$$P(\bar{E}) = 1 - P(E) \text{ or } P(E) = 1 - P(\bar{E}) \text{ or } P(E) + P(\bar{E}) = 1$$

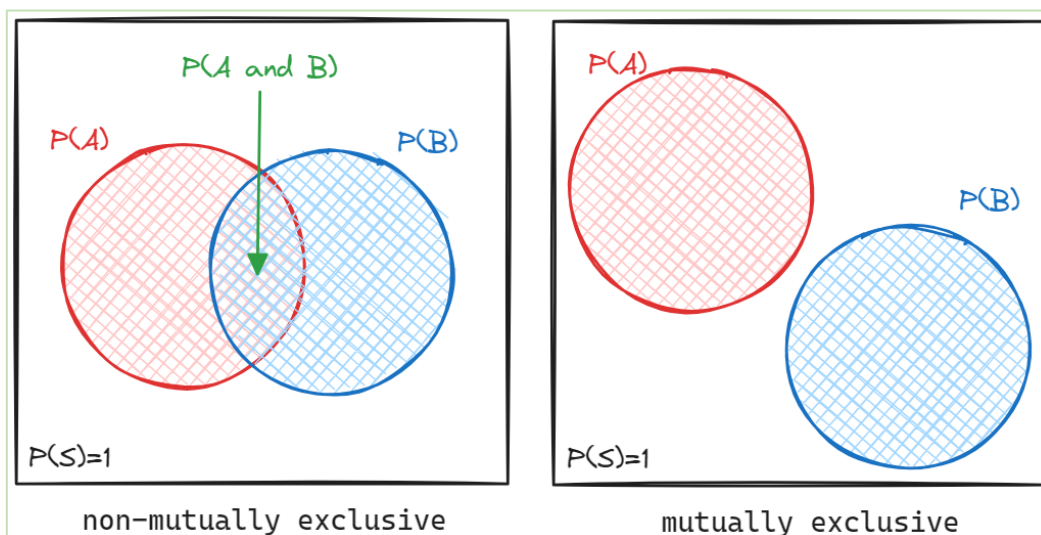
$$P(S)=1$$

$$P(\bar{E})=1-P(E)$$

$$P(E)$$

4-2 Addition Rules for Probability

Mutually exclusive events / Disjoint events 互斥事件	兩事件不會同時發生，即互斥
Addition rule 加法規則	$P(A \text{ or } B) = \begin{cases} P(A) + P(B) & , A, B \text{ are mutually exclusive} \\ P(A) + P(B) - P(A \text{ and } B) & , A, B \text{ are not mutually exclusive} \end{cases}$



4-3 Multiplication Rules and Conditional Probability

independent events 獨立事件	兩事件的發生不會互相影響，即 A 事件的發生不會影響 B 事件發生的機率 i.e. $P(A B) = P(A)$ (不管 B 事件有沒有發生，A 事件發生的機率都一樣)
Multiplication rule 乘法規則	$P(A \text{ and } B) = \begin{cases} P(A) \cdot P(B) & , A, B \text{ are independent} \\ P(A) \cdot P(B A) & , A, B \text{ are dependent} \end{cases}$
Conditional probability 條件機率	✦ 已知 A 事件發生的情況下，發生 B 事件的機率 $P(B A) = \frac{P(A \text{ and } B)}{P(A)}$

4-4 Counting Rules

Counting rule 計數法則	一個 n 個事件的序列，第一個事件有 k_1 個機率，第二個事件有 k_2 個機率.....，則該序列所有的機率個數為 $k_1 \cdot k_2 \cdots k_n$
Factorial 階乘	$n! = n(n-1)(n-2) \cdots 1$ $0! = 1$
Permutation 排列 (考慮排序的順序)	<p>★ n 個物件，一次抽 r 個，則排列的個數為</p> ${}_nP_r = \frac{n!}{(n-r)!}$ <p>★ n 個物件，r_1 個相同的物件，r_2 個相同的物件.....，r_p 個相同的物件，則排列的個數為</p> $\frac{n!}{r_1! r_2! \cdots r_p!}, r_1 + r_2 + \cdots r_p = n$
Combination 組合 (忽略排序的順序)	<p>★ n 個物件，一次抽 r 個，則組合的個數為</p> ${}_nC_r = \frac{n!}{(n-r)! r!} = {}_nC_{(n-r)}$

5-1 Probability Distribution

Random variable 隨機變數	一個變數，其值是隨機決定的
Discrete variable (Ch 5)	有限數量的可能值，或是無限數量的可數的值 (e.g. 1, 2, 3, etc.)
Continuous variable (Ch 6)	所有值落在任兩個給定的數值區間內，可以透過測量得到
Discrete probability distribution	<ol style="list-style-type: none"> 對所有結果(outcomes)做次數分配表 計算各自的機率 結果放 x 軸，機率放 y 軸 <p>★ 所有樣本空間內的事件，其機率加總為 1 (i.e. $\sum P(X) = 1$)</p> <p>★ 樣本空間內的事件，其機率介於 0 到 1 之間 (i.e. $0 \leq P(X) \leq 1$)</p>

5-2 Mean, Variance, Standard Deviation, Expectation

	Probability distribution	Population
Mean	$\mu = X_1 \cdot P(X_1) + X_2 \cdot P(X_2) + \dots + X_n \cdot P(X_n) = \sum X \cdot P(X)$ <p>X_1, X_2, \dots, X_n: outcomes $P(X_1), P(X_2), \dots, P(X_n)$: corresponding probabilities</p>	$\mu = \frac{\sum X}{N}$
Variance Standard deviation	$\sigma^2 = \sum [(X - \mu)^2 \cdot P(X)] = [\sum X^2 \cdot P(X)] - \mu^2$ $\sigma = \sqrt{\sigma^2}$	$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$
Expectation 期望值	$\mu = E(X) = \sum X \cdot P(X)$ <p>離散隨機變數的期望值為該變數的理論平均(theoretical average)</p>	

★ [推導]變異數

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N} = \frac{\sum X^2 - 2N\mu^2 + N\mu^2}{N} = \frac{\sum X^2 - N\mu^2}{N} = \frac{\sum X^2}{N} - \mu^2$$

$$\Rightarrow \sigma_{prob}^2 = [\sum X^2 \cdot P(X)] - \mu^2$$

5-3 The Binomial Distribution

Binomial experiment 二項實驗	<ol style="list-style-type: none"> 固定的試驗次數 只有兩個結果，成功或失敗 每次試驗都是獨立的 每次試驗中成功的機率是相同的
Binomial distribution 二項分配 $X \sim B(n, p)$	<p>$P(S) = p$: 成功的機率, $P(F) = q = 1 - p$: 失敗的機率 n: 總試驗次數 X: n次試驗中成功的次數</p> $P(X) = \frac{n!}{(n-X)!X!} \cdot p^X \cdot q^{n-X}$ <p>Mean: $\mu = n \cdot p$ Variance: $\sigma^2 = n \cdot p \cdot q$ Standard deviation: $\sigma = \sqrt{npq}$</p>

6-1 Normal Distribution

Normal distribution 常態分布 $X \sim N(\mu, \sigma)$ $y = \frac{e^{-(X-\mu)^2/(2\sigma)^2}}{\sigma\sqrt{2\pi}}$	5. 連續、鐘形、對稱 6. 曲線不會碰到 x 軸 (i.e. 無限接近 0，但不為 0) 7. 曲線下面積為 1 (or 100%) 8. 平均 $\pm K$ 倍標準差，曲線下面積： 甲、 $K=1$: 68% 乙、 $K=2$: 95% 丙、 $K=3$: 99.7%
Standard normal distribution 標準常態分布 $Z \sim N(0, 1)$ $z = \frac{X - \mu}{\sigma}$	平均為 0、標準差為 1 的常態分布
查表 Appendix A: TABLE E (p.650)	★ 注意表給你的是 z 值左邊的面積還是右邊的面積 給 z 值，查機率(曲線下面積) ▪ 左邊: 個位數和小數點第一位 ▪ 上面: 小數點第二位 ▪ 注意對準，不要看錯格 給機率，查 z 值 1. 如果值不在表裡，找最接近的 2. 如果老師要更精確的，用內插法

6-2 Applications of the Normal Distribution

$z = \frac{X - \mu}{\sigma}$	$X = z \cdot \sigma + \mu$
判斷常態	<ul style="list-style-type: none"> ▪ 畫直方圖 (histogram)，接近鐘形 \rightarrow 常態 ▪ Pearson Coefficient (PC) $PC = \frac{3(\bar{X} - median)}{s}$ $PC \in (-1,1) \rightarrow$ 常態 $\Leftrightarrow PC \geq 1$ or $PC \leq -1 \rightarrow$ 偏斜 (skewed) ▪ 找離群值，沒有 \rightarrow 常態

6-3 Central Limit Theorem

Distribution of Sample Means $X \sim N(\mu, \sigma)$ $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$	樣本平均的 1. 平均會和母體平均相同 2. 標準差會小於母體標準差，會等於母體標準差/ \sqrt{n}
Central limit theorem	當樣本數 n 越大，樣本平均的分布會接近常態分布
Finite population correction factor 有限母體修正因子	$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$ N 為母體大小， n 為樣本大小。 當 N 很大， n 很小則沒必要修正，因為 $\sqrt{\frac{N-n}{N-1}}$ 會趨近於 1

推導：

$$E(\bar{X}) = E\left(\frac{\sum X}{n}\right) = \frac{1}{n} \sum E(X) = \frac{1}{n} \cdot n \cdot \mu = \mu$$

$$Var(\bar{X}) = Var\left(\frac{\sum X}{n}\right) = \frac{1}{n^2} \sum Var(X) = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}$$

$$\Rightarrow SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

7-1 Confidence Intervals

Point estimate 點估計	一個特定的數值估計，用於對參數進行估算。 對母體平均 μ 的最佳點估計為樣本平均 \bar{X}
Properties of estimator	1. Unbiased (不偏) $E(\bar{X}) = \mu$ 2. Consistent (一致) $\bar{X} \rightarrow \mu, as n \rightarrow \infty$ 3. Relatively efficient (相對有效) $\min[Var(\bar{X})]$
Interval estimate 區間估計	一個區間或範圍的數值，這個區間可能會也可能不會包含要估計的參數
Confidence interval (CI) 信賴區間	區間估計會包含參數的機率 通常是 90%, 95%, 99%
Margin of error (E) / Maximum error of the estimate 邊際誤差	參數的估計值和實際值的最大可能差異 Eg. $21.9 < \mu < 22.7$ or 22.3 ± 0.4 ，則 0.4 為邊際誤差

7-2 Confidence Intervals for Mean, σ is known

Assumption	1. 樣本為隨機樣本 2. $n \geq 30$ 或 $n < 30$ 但母體服從常態
CI of μ with known σ	$\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ <div> 90% CI $z_{\frac{\alpha}{2}} = 1.65$ 95% CI $z_{\frac{\alpha}{2}} = 1.96$ 99% CI $z_{\frac{\alpha}{2}} = 2.58$ </div>
Margin of error	$E = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$
Minimum sample size	$n = \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{E} \right)^2$ <p>★ 無條件進位</p>

7-3 Confidence Intervals for Mean, σ is unknown

Student's t distribution (t distribution)	1. 鐘形 2. 對稱於平均 3. 平均數、中位數和眾數都等於 0，且位於分配的中間 4. 曲線會接近 x 軸但不會碰到 和標準常態分配的差異 1. 變異數 > 1 2. T 分配是一個基於自由度(degrees of freedom)的曲線家族，自由度和樣本數有關 3. 當樣本數增加，t 分配會接近標準常態分配
CI of μ with unknown σ	$\bar{X} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$

7-4 Confidence Intervals and Sample Size for proportion

Symbols used in proportion	$p = \text{population proportion}$ $\hat{p} = \text{sample proportion}$ $\hat{p} = \frac{X}{n} \text{ and } \hat{q} = 1 - \hat{p}$
CI of proportion	$\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$
Margin of error	$E = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$
Minimum sample size	$n = \hat{p}\hat{q} \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2$

★ 無條件進位

7-5 Confidence Intervals for Variances and Standard Deviation

Chi-square distribution 卡方分配	<ol style="list-style-type: none"> 卡方值 ≥ 0 基於自由度的曲線家族 曲線下面積為 1 右偏 (注意: 卡方不是對稱的)
CI of Variance	$\frac{(n-1)s^2}{\chi_{right}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{left}^2}$ <p style="text-align: center;">$d.f. = n - 1$</p>

8 Hypothesis Testing

Hypothesis testing 假設檢定	做決策的過程，用於評估有關母體的主張 e.g. 新藥物是否降低病人血壓、新的教學技巧是否提升學生成績
	步驟： 定義研究母體 → 陳述要研究的特定假設 → 給定顯著水準 → 從母體中挑選樣本 → 收集資料 → 計算檢定統計量 → 結論
	2 種統計檢定： z 檢定(σ 已知)、t 檢定(σ 未知)
	3 種方法： 傳統、p 值、信賴區間
Test value 檢定統計量	$\text{test value} = \frac{(\text{observed value}) - (\text{expected value})}{\text{standard error}}$
	<ul style="list-style-type: none"> Observed value: 統計量(例如樣本平均)，用樣本資料計算 Expected value: 參數(例如母體平均)，當虛無假設為真時預期得到的值(i.e. 假設值)

步驟

Step 1

假設: 列出 H_0, H_1

★ 注意包含等於的，要放在 H_0

臨界值方法

Step 2

找拒絕域: 題目會給你顯著水準 α ，利用查表找臨界值

Step 3

計算檢定統計量: 根據題目判斷要用 z 或 t 檢定，注意分子都是拿樣本資料計算的數值，減掉你要檢定的數值(寫 H_0 的時候放的數值)

Step 4

決策: 看你的檢定統計量有沒有落在拒絕域，決定是否拒絕 H_0

P 值方法

Step 2

計算檢定統計量

Step 3

找 p 值

★ 左尾: $p\text{-value}=P(Z<z)$

★ 右尾: $p\text{-value}=P(Z>z)$

★ 雙尾: $p\text{-value}=2*P(Z>|z|)$

Step 4

決策:

$p\text{-value} \leq \alpha \Rightarrow$ 拒絕 H_0

$p\text{-value} > \alpha \Rightarrow$ 不拒絕 H_0

Step 5

總結: 根據題目做出總結

宣稱在哪 (看包含 claim, conclude, test, hypothesis 或 report 的那串)

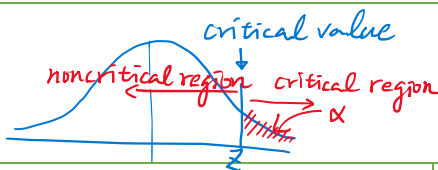
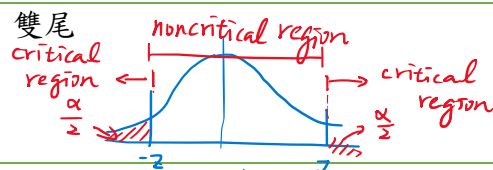
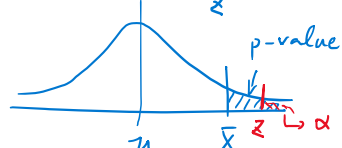
決策	H_0		H_1	
拒絕 H_0	is	reject	is	support
不拒絕 H_0	is not	reject	is not	support

There is
is not enough evidence to reject
support the claim.

8-1 steps in hypothesis testing - traditional method

Statistical hypothesis 統計假設	關於母體參數的猜想，這個猜想可能是真的或假的		
Null hypothesis 虛無假設 H_0	表明 參數和特定數值沒有差別 (e.g. $H_0: \mu = 1$)，或是 兩參數之間沒有差別 ($H_0: \mu_1 = \mu_2$)		
Alternative hypothesis 對立假設 H_1	表明 參數和特定數值有差別 ($H_1: \mu \neq 1$)，或是 兩參數之間有差別 ($H_1: \mu_1 \neq \mu_2$)		
$H_0: \mu = 1$ $H_1: \mu \neq 1$		$H_0: \mu \geq 1$ (or $\mu = 1$) $H_1: \mu < 1$	$H_0: \mu \leq 1$ (or $\mu = 1$) $H_1: \mu > 1$
★ 注意: 等號放在 H_0 H_0, H_1 後面是冒號，不要寫成等號 (e.g. $H_0 = \mu = 1 \rightarrow$ 是錯的) 結論只有“有足夠證據證明拒絕或不拒絕 H_0 ”，沒有接受 H_0 或接受 H_1			

我們用樣本資料來決定是否 拒絕虛無假設，因此有可能 做出錯誤的決策		$H_0: \text{true}$	$H_0: \text{false}$
	Reject H_0	Type I Error 型 I 誤差	Correct decision
	Do not reject H_0	Correct decision	Type II Error 型 II 誤差

Level of significance 顯著水準 (α) $P(\text{type I error}) = \alpha$	接受型 I 誤差的最大機率	
$P(\text{type II error}) = \beta$	β 不容易計算，但和 α 是一個增加，另一個減少的關係	
Critical/rejection region 拒絕域	單尾 	雙尾 
p-value p 值	當虛無假設為真時，所得到的樣本觀察結果或是更極端的结果出現的機率	

8-2 z test for a mean

z test for a mean when σ is known	
Assumption	1. 隨機樣本 2. $n \geq 30$ 或 $n < 30$ ，母體服從常態
檢定統計量	$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$

8-3 t test for a mean

t test for a mean when σ is unknown	
t 分配	和 z 分配很像 差異: 變異數大於 1、曲線根據自由度會有不同形狀 (簡單的互動式網頁)
Assumption	1. 隨機樣本 2. $n \geq 30$ 或 $n < 30$ ，母體服從常態
檢定統計量	$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$
Degrees of freedom 自由度	$d.f. = n - 1$
Mean 平均 Standard deviation 標準差	$\bar{X} = \frac{\sum X}{n}$ $s = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}} = \sqrt{\frac{n \sum X^2 - (\sum X)^2}{n(n - 1)}}$

8-4 z test for a proportion

z test for proportion	
Why z test	當二項分配的 $np \geq 5$ 且 $nq \geq 5$ 時會接近常態分配，因此使用 z 檢定
檢定統計量	$z = \frac{\hat{p} - p}{\sqrt{pq/n}}$

8-5 χ^2 test for a variance or standard deviation

Chi-square test for a single variance or standard deviation	
Assumption	1. 隨機樣本 2. 母體服從常態 3. 觀察值之間獨立
檢定統計量	$\chi^2 = \frac{(n - 1)s^2}{\sigma^2}$
卡方分配	1. 曲線根據自由度會有不同形狀 2. 右尾 (Note: 卡方分配不是對稱的)

8-6 additional topics

Confidence Interval (Sec 7)	
CI of μ with known σ	$\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$
CI of μ with unknown σ	$\bar{X} - t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}$
CI of Proportion	$\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$
CI of Variance	$\frac{(n-1)s^2}{\chi_{right}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{left}^2}$ d.f. = $n - 1$

9-1 Testing the difference between two parameters

	two-tailed	left-tailed	right-tailed
Hypothesis	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	$H_0: \mu_1 \geq \mu_2$ $H_1: \mu_1 < \mu_2$	$H_0: \mu_1 \leq \mu_2$ $H_1: \mu_1 > \mu_2$
	$H_0: \mu_1 - \mu_2 = 0$ $H_1: \mu_1 - \mu_2 \neq 0$	$H_0: \mu_1 - \mu_2 \geq 0$ $H_1: \mu_1 - \mu_2 < 0$	$H_0: \mu_1 - \mu_2 \leq 0$ $H_1: \mu_1 - \mu_2 > 0$

★ 注意: 要定義好 1 和 2 分別代表誰, 單尾檢定才知道假設怎麼寫, 檢定統計量要誰減誰

9-2 Testing the difference between two means: z test

z test for a two means	
Assumption	1. 兩組隨機樣本 2. 兩組樣本之間獨立 3. 母體標準差已知, 如果 $n < 30$ 則母體需要服從或接近常態分配
檢定統計量	$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
Confidence Interval	$(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

9-3 Testing the difference between two means of independent sample: t test

t test for a two means: independent	
Assumption	1. 兩組隨機樣本 2. 兩組樣本之間獨立 3. 母體標準差未知, 如果 $n < 30$ 則母體需要服從或接近常態分配
檢定統計量	<div> <div> ★ 假設變異數(標準差)不相等 $t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ $s = \frac{n\sum X^2 - (\sum X)^2}{n(n-1)}$ $d.f. = \min(n_1 - 1, n_2 - 1)$ </div> <div> ★ 假設變異數(標準差)相等 $t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p^2 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ $d.f. = n_1 + n_2 - 2$ or $d.f. = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$ </div> </div>
Confidence Interval	$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

9-4 Testing the difference between two means: dependent sample

t test for a two means: dependent	
Assumption	1. 隨機樣本 2. 樣本之間 不獨立 3. 如果 $n < 30$ 則母體需要服從或接近常態分配
檢定統計量	$t = \frac{\bar{D} - \mu_D}{s_D / \sqrt{n}}$ $\bar{D} = \frac{\sum D}{n}, s_D = \sqrt{\frac{n \sum D^2 - (\sum D)^2}{n(n-1)}}$ $d.f. = n - 1$
Confidence Interval	$\bar{D} - t_{\alpha/2} \cdot \frac{s_D}{\sqrt{n}} < \mu_D < \bar{D} + t_{\alpha/2} \cdot \frac{s_D}{\sqrt{n}}$

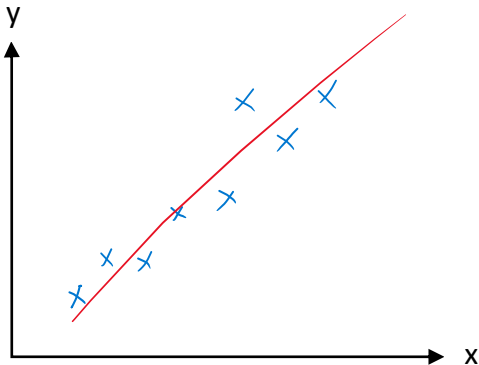
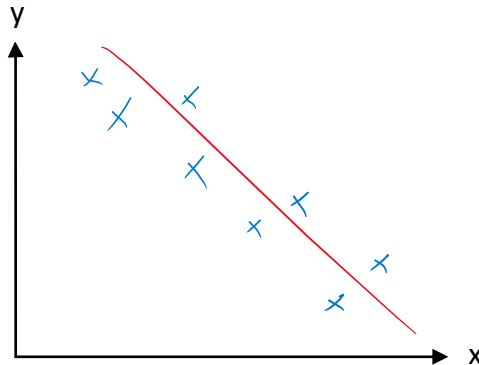
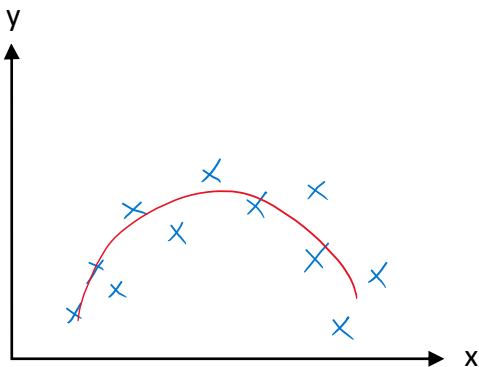
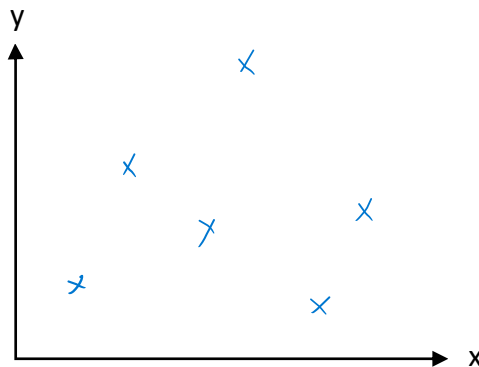
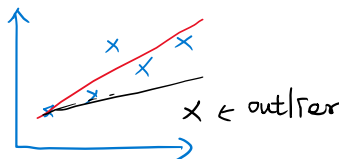
9-5 Testing the difference between proportion

z test for a two proportions	
Assumption	1. 隨機樣本 2. 樣本之間 獨立 3. $np \geq 5$ 且 $nq \geq 5$
檢定統計量	$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ $\hat{p}_1 = \frac{X_1}{n_1}, \hat{p}_2 = \frac{X_2}{n_2}$ <p>pooled sample proportion: $\bar{p} = \frac{X_1 + X_2}{n_1 + n_2}, \bar{q} = 1 - \bar{p}$ (under $H_0: \pi_1 = \pi_2$ is true)</p>
Confidence Interval	$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$

9-6 Testing the difference between two variance

F 分配	<ol style="list-style-type: none">1. 非負，因為變異數永遠大於或等於 02. 正偏分配3. 平均等於 14. 為一個家族，根據自由度有不同曲線
Assumption	<ol style="list-style-type: none">1. 隨機樣本2. 樣本之間獨立3. 樣本為常態分配
檢定統計量	$F = \frac{s_1^2}{s_2^2}$ <ul style="list-style-type: none">★ 變異數較大的放分子★ 雙尾檢定α要除 2，臨界值位於 F 曲線的右側

10-1 Correlation

scatter plot 散佈圖	正線性相關		負線性相關	
	曲線相關		無相關	
	correlation coefficients 相關係數	<div><ul style="list-style-type: none">• 衡量兩個屬量變數之間的線性關係的強度和方向• 母體: ρ ; 樣本: r• range: $[-1, 1]$, i.e. $-1 \leq r \leq 1$• 對離群值敏感<div>$x \leftarrow \text{outlier}$</div>$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$</div>		
	Assumption	<div><div>1. 隨機樣本且為屬量資料</div><div>2. 散佈圖看起來大約有線性相關</div><div>3. 沒有離群值</div><div>4. x, y 都來自常態分配的母體</div></div>		
Hypothesis	<div>$H_0: \rho = 0$$H_1: \rho \neq 0$</div>			
test value	<div>$t = \sqrt{\frac{n - 2}{1 - r^2}}$$d.f. = n - 2$<div>★ 損失 2 個自由度，因為計算 r 要用到 x 跟 y，各損失 1 個自由度</div></div>			

10-2 Regression

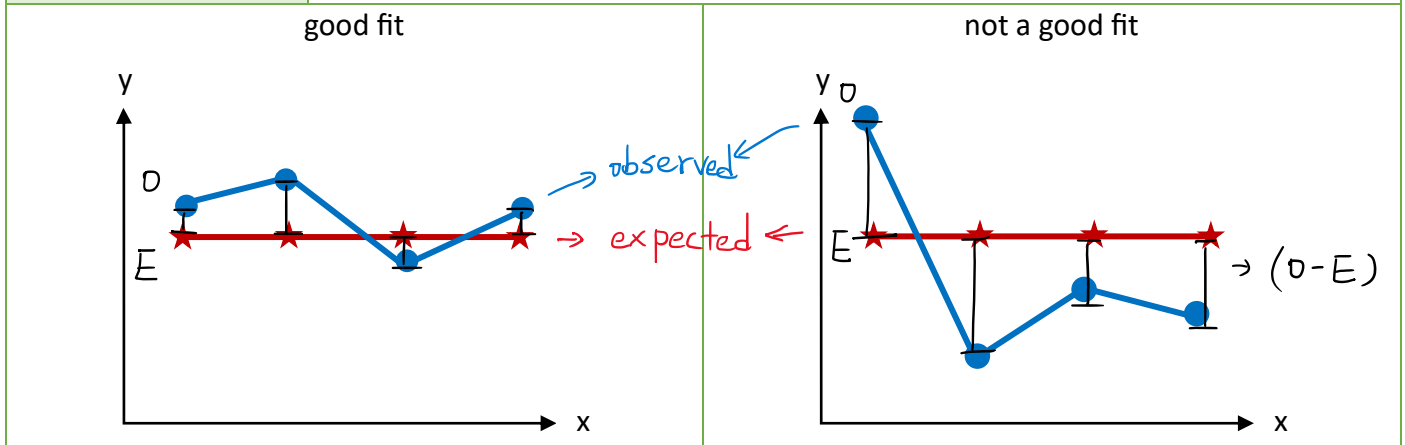
linear regression	$y' = a + b \cdot x$	
	截距 intercept	斜率 slope
	$a = \frac{\sum y \sum x^2 - \sum x \sum xy}{n \sum x^2 - (\sum x)^2}$	$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$

10-3 Coefficient of determination & standard error of the estate

coefficient of determination	衡量 Y 可以被迴歸線和 X 的變異 $r^2 = \frac{\text{explained variation}}{\text{total variation}}$
coefficient of nondetermination	$1 - r^2$
standard error of the estimate	$s_{est} = \sqrt{\frac{\sum (y - y')^2}{n - 2}}$ $= \sqrt{\frac{\sum y^2 - a \sum y - b \sum xy}{n - 2}}$
prediction interval	$y' \pm t_{\frac{\alpha}{2}} \cdot s_{est} \cdot \sqrt{1 + \frac{1}{n} + \frac{n(x - \bar{X})^2}{n \sum x^2 - (\sum x)^2}}$ $d.f. = n - 2$ <p>★ 損失 2 個自由度，因為截距和斜率各損失 1 個自由度</p>

11-1 Test for Goodness of Fit

goodness-of-fit test	檢定觀察到的次數分配是否和期望的次數分配一樣 (實際資料算出來的) (計算得到的)
hypothesis	H_0 : 資料服從特定的分配 H_1 : 資料不服從特定的分配 ★ 右尾檢定，因為 χ^2 小表示 H_0 : good fit； χ^2 大表示 H_1 : not a good fit
test statistic	$\chi^2 = \sum \frac{(O-E)^2}{E}$, O 表示觀察到的次數, E 表示期望的次數 $d.f. = k - 1$, k 表示類別的個數
assumption	<ul style="list-style-type: none"> 隨機樣本 每類的期望次數要≥ 5



test of normality	H_0 : 變數是常態分配 H_1 : 變數不是常態分配
test statistic	$z = \frac{X - \bar{X}}{s}$ $\bar{X} = \frac{\sum(f \cdot X_m)}{\sum f}, s = \sqrt{\frac{n \sum(f \cdot X_m^2) - (\sum f \cdot X_m)^2}{n(n-1)}}$

11-2 Tests Using Contingency Table

test for independence	檢定兩變數是否獨立
hypothesis	H_0 : <u>variable A</u> is independent to <u>variable B</u> H_1 : <u>variable A</u> is related to <u>variable B</u>
test for homogeneity of proportion	檢定不同母體中擁有特定特徵的比例是否相同
hypothesis	$H_0: p_1 = p_2 = p_3$ H_1 : at least one proportion is different (至少有一個不同) ★ H_1 不可以寫 $p_1 \neq p_2 \neq p_3$, 這表示 3 個都不同
assumption	<ul style="list-style-type: none"> 隨機樣本 每一格的期望值必須大於等於 5, 否則合併類別
test statistic	$\chi^2 = \sum \frac{(O-E)^2}{E}$, O 表示觀察到的次數, E 表示期望的次數 $E = \frac{(\text{row sum})(\text{column sum})}{\text{total}}$ $d.f. = (r-1)(c-1)$, r, c 分別表示列跟行的個數

	Column 1	Column 2	Column 3	total
Row 1	$C_{1,1}$	$C_{1,2}$	$C_{1,3}$	$n_{1,+}$
Row 2	$C_{2,1}$	$C_{2,2}$	$C_{2,3}$	$n_{2,+}$
total	$n_{+,1}$	$n_{+,2}$	$n_{+,3}$	n

$$E_{1,1} = \frac{n_{+,1} \cdot n_{1,+}}{n}$$

11-3 ANOVA

one-way ANOVA	利用樣本變異數檢定 3 組或以上的平均是否相同
assumption	<ul style="list-style-type: none"> 母體為常態或近似常態分配 獨立隨機樣本 母體變異數相同
hypothesis	$H_0: \mu_1 = \mu_2 = \mu_3$ H_1 : at least one mean is different (至少有一個不同)
test statistic	$F = \frac{\text{variance between groups}}{\text{variance within groups}} = \frac{s_B^2}{s_W^2}$
step	<ol style="list-style-type: none"> 計算每組的樣本平均(\bar{X}_i)和變異數(s_i^2) 計算總平均 $\bar{X}_{GM} = \frac{\sum X}{N}$ 計算組間變異 $s_B^2 = \frac{\sum n_i(\bar{X}_i - \bar{X}_{GM})^2}{k-1}$ 計算組內變異 $s_W^2 = \frac{\sum (n_i-1)s_i^2}{\sum (n_i-1)}$ 計算檢定統計量 $F = \frac{s_B^2}{s_W^2}$, $d.f. N = k-1$, $d.f. D = N-k$

Source	Sum of square (SS)	d.f.	Mean square (MS)	F
Between	$SS_B = \sum n_i(\bar{X}_i - \bar{X}_{GM})^2$	$k-1$	$MS_B = \frac{SS_B}{k-1} = s_B^2$	$F = \frac{MS_B}{MS_W}$
Within (error)	$SS_W = \sum (n_i-1)s_i^2$	$N-k$	$MS_W = \frac{SS_W}{N-k} = s_W^2$	
Total	$SST = SS_B + SS_W$	$N-1$		

k 為組別個數, N 為總樣本數

