Content

[1-1 Descriptive and Inferential Statistics 3](#_Toc171274101)

[1-2 Variables and Types of Data 3](#_Toc171274102)

[1-3 Data Collection and Sampling Techniques 3](#_Toc171274103)

[1-4 Experimental Design 4](#_Toc171274104)

[2-1 Organizing Data 5](#_Toc171274105)

[2-2 Histograms, Frequency Polygons, and Ogives 5](#_Toc171274106)

[2-3 Other Types of Graphs 6](#_Toc171274107)

[2-4 Paired Data and Scatter Plots 8](#_Toc171274108)

[3-1 Measures of Central Tendancy 9](#_Toc171274109)

[3-2 Measures of Variation 10](#_Toc171274110)

[3-3 Measures of Position 11](#_Toc171274111)

[3-4 Exploratory Data Analysis 12](#_Toc171274112)

[4-1 Sample Space and Probability 13](#_Toc171274113)

[4-2 Addition Rules for Probability 14](#_Toc171274114)

[4-3 Multiplication Rules and Conditional Probability 14](#_Toc171274115)

[4-4 Counting Rules 15](#_Toc171274116)

[5-1 Probability Distribution 16](#_Toc171274117)

[5-2 Mean, Variance, Standard Deviation, Expectation 16](#_Toc171274118)

[5-3 The Binomial Distribution 16](#_Toc171274119)

[6-1 Normal Distribution 17](#_Toc171274120)

[6-2 Applications of the Normal Distribution 17](#_Toc171274121)

[6-3 Central Limit Theorem 18](#_Toc171274122)

[7-1 Confidence Intervals 19](#_Toc171274123)

[7-2 Confidence Intervals for Mean, is known 19](#_Toc171274124)

[7-3 Confidence Intervals for Mean, is unknown 19](#_Toc171274125)

[7-4 Confidence Intervals and Sample Size for proportion 20](#_Toc171274126)

[7-5 Confidence Intervals for Variances and Standard Deviation 20](#_Toc171274127)

[8 Hypothesis Testing 21](#_Toc171274128)

[8-1 steps in hypothesis testing – traditional method 22](#_Toc171274129)

[8-2 z test for a mean 22](#_Toc171274130)

[8-3 t test for a mean 23](#_Toc171274131)

[8-4 z test for a proportion 23](#_Toc171274132)

[8-5 test for a variance or standard deviation 23](#_Toc171274133)

[8-6 additional topics 24](#_Toc171274134)

[9-1 Testing the difference between two parameters 25](#_Toc171274135)

[9-2 Testing the difference between two means: z test 25](#_Toc171274136)

[9-3 Testing the difference between two means of independt sample: t test 25](#_Toc171274137)

[9-4 Testing the difference between two means: dependt sample 26](#_Toc171274138)

[9-5 Testing the difference between proportion 26](#_Toc171274139)

[9-6 Testing the difference between two variance 27](#_Toc171274140)

[10-1 Correlation 28](#_Toc171274141)

[10-2 Regression 29](#_Toc171274142)

[10-3 Coefficient of determination & standard error of the estate 29](#_Toc171274143)

[11-1 Test for Goodness of Fit 30](#_Toc171274144)

[11-2 Tests Using Contingency Table 31](#_Toc171274145)

[11-3 ANOVA 31](#_Toc171274146)

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美元是時新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) 2. range (R) = H – L (全距) 3. number of classes 4. width = R / number of classes   round up (無條件進位)   1. start point + width = lower limits   start point: 通常是最小值或是方便的數字   1. upper limits 2. boundaries | 當第一類沒有下限或最後一類沒有上限時  eg. (p.46)   |  |  | | --- | --- | | 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 |
|  | 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) |
|  |  | |  |
| 1. label x, y axis 2. label class boundaries for histogram, ogive   label midpoints for frequency polygon   1. 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) | |

|  |  |  |  |
| --- | --- | --- | --- |
| bell-shaped (鐘形) | j-shaped | positively / right-skewed (右偏) | bimodal (雙峰) |
|  |  |  |  |
| uniform (均勻) | reverse j-shaped | negatively / left-skewed (左偏) | u-shaped |
|  |  |  |  |

2-3 Other Types of Graphs

|  |  |
| --- | --- |
| Bar graph  長條圖 | Pareto chart  柏拉圖 |
| * vertical / horizontal * display freq. | * for categorical variable * display freq. from highest to lowest |
| e.g. (p.77) Ex. 2-9   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | City | Atlanta | Boston | Chicago | New York | Washington | | Hours | 52 | 64 | 61 | 74 | 82 | | |
|  |  |

|  |  |
| --- | --- |
| Time series graph  時間數列 | Pie graph  圓餅圖 |
| * display data over a specific period of time * trend or pattern: ascending / descending * slope or steepness (陡度) | * for categorical variable * display percentage of freq. |
| e.g. (p.78) Ex. 2-10   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | Year | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | | Cost | 1.40 | 1.55 | 1.61 | 1.65 | 1.78 | 1.90 | | e.g.   |  |  |  |  |  | | --- | --- | --- | --- | --- | | 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 |
| * plot as a point above horizontal axis | * combination of sorting and graphing * retaining actual data |
| e.g.   |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | 4 | 11 | 12 | 23 | 25 | 11 | 15 | 25 | 30 | 35 | | 30 | 28 | 15 | 15 | 29 | 28 | 23 | 25 | 15 | 25 | | |
|  | |  |  | | --- | --- | | 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

|  |  |
| --- | --- |
| Scatter plot  散布圖 | |
| * to determine if there is a relationship between two variables | |
| Positive linear relationship | Negative linear relationship |
|  |  |
| Nonlinear relationship | No relationship |
|  |  |

**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*](https://www.youtube.com/watch?v=mQfacqVvOEM)
* [*How Ice Cream Kills! Correlation vs. Causation*](https://www.youtube.com/watch?v=VMUQSMFGBDo)

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*](https://www.youtube.com/watch?v=HUti6vGctQM)

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 Tendancy

|  |  |
| --- | --- |
| 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 / arithemetic average  平均數 / 算術平均數 | Weighted mean  加權平均數 | | Midrange  中程數 |
| Sample mean:  Population mean: |  | | * 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 3. 是奇數，則中間值即為中位數   是偶數，則中間的兩個數字取平均即為中位數 | | 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 | Symmetric | Right-skewed |
| Mode > median > mean | Mode = median = mean | Mode < median < mean |
|  |  |  |

3-2 Measures of Variation

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variation** | | | | | |
| Range全距 |  | | | | |
|  | Population | Sample | | Grouped data | |
| Variance  變異數 |  |  | |  | |
| Standard deviation  標準差 |  |  | |  | |
| Coefficient of variation  變異係數 |  |  | | * 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 |  | |  | |  |
| * 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  (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:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

* 線性轉換如何影響平均和標準差

|  |  |
| --- | --- |
|  |  |
|  | |
|  | |

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**: | **Given data**: | |  |  |  | | --- | --- | --- | |  |  |  | |  |  |  |   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)](#Q28) |  |
| **Population**: | **Given percentile**:  c是整數，則數值為和的平均；  c不是整數，則無條件進位，數值為進位後的 | **Interquartile range**:  四分位距 |  |
|  | **Percentile graph:**   |  |  |  |  | | --- | --- | --- | --- | | A | B | C | D | | Class  boundaries | Freq. | Cumulative  Freq. | Cumulative  percent | | **Outlier**離群值: |  |

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)

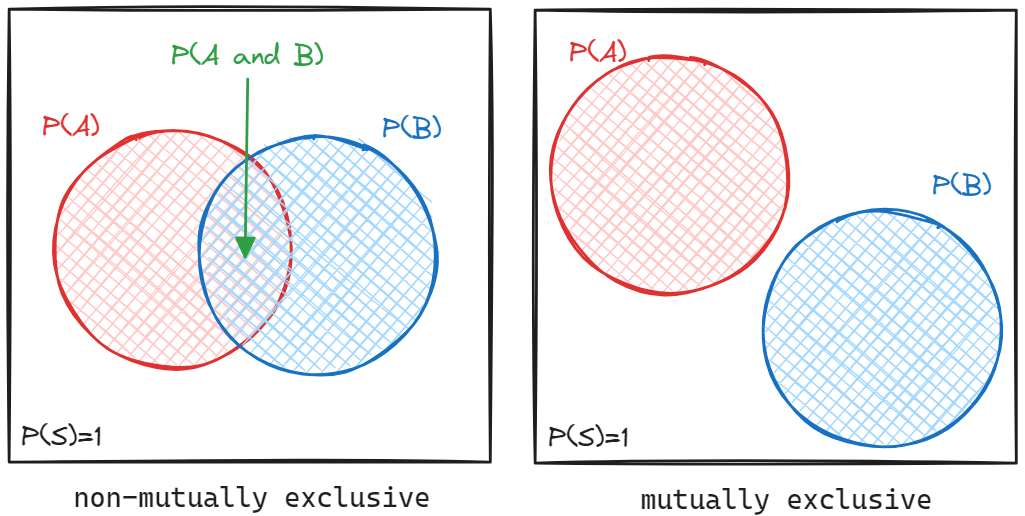
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Experiment | Sample space |  | Experiment | Event |
| 擲一枚硬幣  擲骰子  擲兩枚硬幣 | 頭, 尾  1, 2, 3, 4, 5, 6  頭頭, 頭尾, 尾頭, 尾尾 |  | 投擲一骰子 | 拿到6點的事件稱simple event  拿到奇數點(即1,3,5點)的事件稱compound event |

|  |  |
| --- | --- |
| 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 |

|  |  |  |  |
| --- | --- | --- | --- |
| 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 |
| =E事件中的結果個數  =樣本空間的結果個數 | =發生E事件的次數  =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) | |  |
| 1. 所有樣本空間中的機率加總為1 2. 事件E不會發生則機率為0 3. 事件E一定會發生則機率為1 | | | |
| **Complement 補集**  事件的補集記為，代表樣本空間中除了事件E以外的所有集合 | |  | |

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** |  |



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.  (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** |  |
| **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 |

4-4 Counting Rules

|  |  |
| --- | --- |
| **Counting rule**  **計數法則** | 一個n個事件的序列，第一個事件有個機率，第二個事件有個機率……，則該序列所有的機率個數為 |
| **Factorial**  **階乘** |  |
| **Permutation**  **排列**  **(考慮排序的順序)** | * n個物件，一次抽r個，則排列的個數為 * n個物件，個相同的物件，個相同的物件……，個相同的物件，則排列的個數為 |
| **Combination**  **組合**  **(忽略排序的順序)** | * n個物件，一次抽r個，則組合的個數為 |

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** | 1. 對所有結果(outcomes)做次數分配表 2. 計算各自的機率 3. 結果放x軸，機率放y軸  * 所有樣本空間內的事件，其機率加總為1 (i.e. ) * 樣本空間內的事件，其機率介於0到1之間 (i.e. ) |

5-2 Mean, Variance, Standard Deviation, Expectation

|  |  |  |
| --- | --- | --- |
|  | **Probability distribution** | **Population** |
| **Mean** |  |  |
| **Variance**  **Standard deviation** |  |  |
| **Expectation**  **期望值** | 離散隨機變數的期望值為該變數的理論平均(theoretical average) | |

* [推導]變異數

5-3 The Binomial Distribution

|  |  |
| --- | --- |
| **Binomial experiment**  **二項實驗** | 1. 固定的試驗次數 2. 只有兩個結果，成功或失敗 3. 每次試驗都是獨立的 4. 每次試驗中成功的機率是相同的 |
| **Binomial distribution**  **二項分配** | : 成功的機率, : 失敗的機率  : 總試驗次數  : 次試驗中成功的次數   |  |  |  | | --- | --- | --- | | Mean: | Variance: | Standard deviation: | |

6-1 Normal Distribution

|  |  |
| --- | --- |
| **Normal distribution**  **常態分布** | 1. 連續、鐘形、對稱 2. 曲線不會碰到x軸 (i.e. 無限接近0，但不為0 3. 曲線下面積為1 (or 100%) 4. 平均倍標準差，曲線下面積:    1. K=1: 68%    2. K=2: 95%    3. K=3: 99.7% |
| **Standard normal distribution**  **標準常態分布** | 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

|  |  |  |
| --- | --- | --- |
|  | |  |
| **判斷常態** | * 畫**直方圖** (histogram)，接近鐘形常態 * **Pearson Coefficient (PC)**   常態 偏斜 (skewed)   * 找**離群值**，沒有常態 | |

6-3 Central Limit Theorem

|  |  |
| --- | --- |
| **Distribution of Sample Means** | 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/ |
| **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** | 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 會趨近於1 |

推導:

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。  對母體平均的最佳點估計為樣本平均 |
| **Properties of estimator** | 1. Unbiased (不偏) 2. Consistent (一致) 3. Relatively efficient (相對有效) |
| **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 ot 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 < < 22.7 or 22.3 0.4，則0.4為邊際誤差 |

7-2 Confidence Intervals for Mean, is known

|  |  |
| --- | --- |
| **Assumption** | 1. 樣本為隨機樣本 2. 或 但母體服從常態 |
| **CI of with known** | |  |  |  | | --- | --- | --- | |  | 90% CI |  | | 95% CI |  | | 99% CI |  | |
| **Margin of error** |  |
| **Minimum sample size** | * 無條件進位 |

7-3 Confidence Intervals for Mean, 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 with unknown** |  |

7-4 Confidence Intervals and Sample Size for proportion

|  |  |
| --- | --- |
| **Symbols used in proportion** |  |
| **CI of proportion** |  |
| **Margin of error** |  |
| **Minimum sample size** | * 無條件進位 |

7-5 Confidence Intervals for Variances and Standard Deviation

|  |  |
| --- | --- |
| **Chi-square distribution**  **卡方分配** | 1. 卡方值 2. 基於自由度的曲線家族 3. 曲線下面積為1 4. 右偏 (注意: 卡方不是對稱的) |
| **CI of Variance** |  |

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:  檢定(已知)、 檢定(未知) |
| 3種方法:  傳統、值、信賴區間 |
| **Test value**  **檢定統計量** | * 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   * 注意包含等於的，要放在 | |
| * 臨界值方法   Step 2  找拒絕域: 題目會給你顯著水準，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 | * P值方法   Step 2  計算檢定統計量  Step 3  找p值   * 左尾: p-value=P(Z<z) * 右尾: p-value=P(Z>z) * 雙尾: p-value=2\*P(Z>|z|)   Step 4  決策:  p-value 拒絕  p-value 不拒絕 |
| 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 |  | |  | | | 拒絕 | is | reject | is | support | | 不拒絕 | 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** | 表明  參數和特定數值沒有差別 (e.g. )，或是  兩參數之間沒有差別 () | | |
| **Alternative hypothesis**  **對立假設** | 表明  參數和特定數值有差別 ()，或是  兩參數之間有差別 () | | |
|  | |  |  |
| * 注意:   等號放在  後面是冒號，不要寫成等號 (e.g. 是錯的)  結論只有”有足夠證據證明拒絕或不拒絕”，沒有接受或接受 | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| 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 |  |  |  |
| Reject | Type I Error  型I誤差 | Correct decision |
| Do not reject | Correct decision | Type II Error  型II誤差 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Level of significance**  **顯著水準 ()** | 接受型I誤差的最大機率 | | |
|  | 不容易計算，但和是一個增加，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結果出現的機率 |  | |

8-2 z test for a mean

|  |  |
| --- | --- |
| **z test for a mean when is known** | |
| **Assumption** | 1. 隨機樣本 2. 或 ，母體服從常態 |
| **檢定統計量** |  |

8-3 t test for a mean

|  |  |
| --- | --- |
| **t test for a mean when 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](https://miyucat.shinyapps.io/distribution/)) |
| **Assumption** | 1. 隨機樣本 2. 或 ，母體服從常態 |
| **檢定統計量** |  |
| **Degrees of freedom**  **自由度** |  |
| **Mean**  **平均**  **Standard deviation**  **標準差** |  |

8-4 z test for a proportion

|  |  |
| --- | --- |
| **z test for proportion** | |
| **Why z test** | 當二項分配的且時會接近常態分配，因此使用z檢定 |
| **檢定統計量** |  |

8-5 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** |  |
| **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 with known** |  |
| **CI of with unknown** |  |
| **CI of Proportion** |  |
| **CI of Variance** |  |

9-1 Testing the difference between two parameters

|  |  |  |  |
| --- | --- | --- | --- |
|  | two-tailed | left-tailed | right-tailed |
| **Hypothesis** |  |  |  |
|  |  |  |

* 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** |  |
| **Confidence Interval** |  |

9-3 Testing the difference between two means of independt 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** | * 假設變異數(標準差)不相等 | * 假設變異數(標準差)相等 |
| **Confidence**  **Interval** |  |  |

9-4 Testing the difference between two means: dependt 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** |  |
| **Confidence Interval** |  |

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. 且 |
| **檢定統計量** | pooled sample proportion:  (under is true) |
| **Confidence Interval** |  |

9-6 Testing the difference between two variance

|  |  |
| --- | --- |
| **F 分配** | 1. 非負，因為變異數永遠大於或等於0 2. 正偏分配 3. 平均等於1 4. 為一個家族，根據自由度有不同曲線 |
| **Assumption** | 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** | * 變異數較大的放分子 * 雙尾檢定要除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  x  y | 負線性相關  x  y | | 曲線相關  x  y | 無相關  x  y | |
| **correlation coefficients**  **相關係數** | * 衡量兩個**屬量**變數之間的**線性關係**的**強度**和**方向** * 母體: ; 樣本: * range: , i.e.  * 對離群值敏感 |
| **Assumption** | 1. 隨機樣本且為屬量資料 2. 散佈圖看起來大約有線性相關 3. 沒有離群值 4. x, y都來自常態分配的母體 |
| **Hypothesis** |  |
| **test value** | * 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 |

10-2 Regression

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **linear regression** | |  |  | | --- | --- | | 截距intercept | 斜率slope | |  |  | |

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 |
| **coefficient of**  **nondetermination** |  |
| **standard error of the estimate** |  |
| **prediction interval** | * 損失2個自由度，因為截距和斜率各損失1個自由度 |

11-1 Test for Goodness of Fit

|  |  |  |
| --- | --- | --- |
| **goodness-of-fit test** | 檢定觀察到的次數分配使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** | 資料服從特定的分配  資料不服從特定的分配   * 右尾檢定，因為小表示good fit；大表示not a good fit | |
| **test statistic** | , O表示觀察到的次數, E表示期望的次數  , k表示類別的個數 | |
| **assumption** | * 隨機樣本 * 每類的期望次數要 | |
| good fit  x  y | | not a good fit  x  y |

|  |  |
| --- | --- |
| **test of normality** | 變數是常態分配  變數不是常態分配 |
| **test statistic** | , |

11-2 Tests Using Contingency Table

|  |  |
| --- | --- |
| **test for independence** | 檢定兩變數是否獨立 |
| **hypothesis** | variable A is independent to variable B  variable A is related to variable B |
| **test for homogenity**  **of proportion** | 檢定不同母體中擁有特定特徵的比例是否相同 |
| **hypothesis** | at least one proportion is different (至少有一個不同)   * 不可以寫，這表示3個都不同 |
| **assumption** | * 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** | , O表示觀察到的次數, E表示期望的次數  , r,c分別表示列跟行的個數 |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **Column 1** | **Column 2** | **Column 3** | **total** | | **Row 1** |  |  |  |  | | **Row 2** |  |  |  |  | | **total** |  |  |  |  | | |

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** | * 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** | at least one mean is different (至少有一個不同) |
| **test statistic** |  |
| **step** | 1. 計算每組的樣本平均()和變異數() 2. 計算總平均 3. 計算組間變異 4. 計算組內變異 5. 計算檢定統計量 , |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Source** | **Sum of square (SS)** | **d.f.** | **Mean square (MS)** | **F** | | **Between** |  |  |  |  | | **Within (error)** |  |  |  | | **Total** |  |  |  |  |   為組別個數, 為總樣本數 | |