研究設計與統計方法

林口長庚醫院 實證醫學中心

余光輝 醫師

學習目標 Objectives

- **Outlines**
  - 基本概念 General concept (slide 1~20)
  - 基礎統計 Basic statistics (slide 21~47)
  - SPSS 操作 Procedure (optional) (slide 47~77)
- **Understanding Statistical Term**
- **Selecting a Statistical Test**
- **SPSS 解讀**
- 進階學習 **(may skip)**

進行統計的五個步驟
Five Steps to Practice Statistics

1. **Formulate a question** **(Literature review, Experience, Hypothesis)**
   - **Important issue**, well design, well conducted (Data collection & record)
2. **Data 分類: 數值或類別型 (Numerical or Categorical Data)**
3. **Dependent variable (Outcome variable, Y)** $Y=a+bx_1+cx_2$
   - **Independent variable (explanatory, x)**
4. **Selecting a statistic test** (Type of data, Null hypothesis H_0)
   - **Parametric or Nonparametric test (有母數或無母數分析)**
     - **Nonparametric使用時機: 不是常態分布, 小樣本, outlier, 結果為類別型**
   - **比較組別: One, Two, ≧ 3 samples (比較 mean, %, difference)**
   - **樣本: Paired (dependent相關) or Unpaired (independent)**
   - **變異數: Variance equal or unequal (假設前提 assumption)**
   - **單尾或雙尾檢定: One tailed or Two tailed *p* value**
5. **Interpretation and Discussion (rationale)**

1. 數據的分類: 類別或數值型 (資料的類型)

- 變項(Variable)的種類: (可用統計方法不同: mean, variance, %, median, rank)
  - 名義、類別變項 (Nominal variable, or scale): 性別、人種
    - 無大小、順序等級之分
  - 次序、順序變項 (Ordinal variable): 教育程度、喜好等級
    - 能分出大小、順序等級，但差距不一定相同 (不知差距)
  - 等距、區間尺度 (Interval scale): 攝氏溫度 (差距相同)
    - 具有任意零點，不能算倍數，但兩數值間差距可以計算倍數
  - 等比、比率尺度 (Ratio scale): 絕對溫度、身高、體重
    - 具有絕對零點，可以計算倍數及比率
- 類別型 (Categorical data): Nominal, Ordinal variable
- 數值型 (Numerical data): 含資訊較多，統計方法多
  - 離散型 (Discrete variable)：(整數值)，家中小孩人數
  - 連續型 (Continuous variable)：(可插入小數)，身高、體重
  - 變項 (Variable) 數據間的轉換 (轉換成爲常態分佈、線性關係)
    - 數值型轉成類別型: BP 140/80 mmHg 分界爲: 正常 高血壓 (Binary categorical data)

2. Variable: 變項、變數

(先區分 y 與 x 的資料類型是屬於: 數值型或類別型)

Multiple Regression Analysis

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

y：SBP (數值) or 血壓高或正常 (類別) **x**：Age Sex Race BH BW CH TG FC(score)

Dependent variable 依變數
Response variable 應變數
Outcome variable 結果變數
Predicted

(Y: 因自變數改變而發生改變的 結果變數)

Independent variables 自變數
Explanatory variables 解釋變數
Covariates (in ANCOVA) 共變數
Predictor variables
Factor 因子 (in ANOVA)~ One way

生物統計之重點

- 資料呈現與解讀
  - 統計：資料收集、組織、量化、分析、說明
    - **Data collection, organization, interpretation, inference (to population)**
- 描述性統計及推論性統計
  - 描述性統計〔敘述性〕 (Descriptive statistics)：集中趨勢 離中趨勢 相關
  - 推論性統計〔做檢定〕 (Inferential statistics)
    - 由母群體抽樣的樣本(sample), 對母群體(population)做統計推論 (inference)
    - 推論是否有顯著的差異 (p value, 95% CI)
      - **Statistical significance or clinical significance?** ~ rationale
      - 是否具有因果關係 **(association or causality?)** ~ Hill's rule
- 分類數據
  - 類別型 (categorical data) 或 數值型 (numerical data)
    - Nominal 名義, ordinal 次序, interval 等距, ratio 等比 scale (尺度的轉換)
    - 變項 (Variable) 數據間的轉換 (轉換成常態分佈、線性關係)
      - 數值型可轉成類別型: BP140/80 mmHg 分界爲正常 or高血壓 (Binary categorical)

敘述性統計 Descriptive statistics

Mean ± SD (range), median (range, or interquartile range)

- 集中趨勢 (Central tendency)
  - **Mean** 平均數、受極端值影響 **x** (**x** bar) **=** Σx_i **/ n** ($\mu = \Sigma x_i/N$)
  - **Median** 中位數、不受極端值影響 (有極端值、非常態分布時用)
  - **Mode** 眾數、又稱流行值 (商業用途)
- 離散趨勢 (Dispersion 分散性, Variation 變異性)
  - **Range** (Min–Max) 全距, **Interquartile range** $Q_{1\text{-}3}$ (極端值: outlier)
  - Mean deviation 平均離差〔與平均值相差的絕對值之平均〕
  - Sum of square: $\Sigma (x_i\text{-}\mathbf{x})^2$ **x: x** bar〔樣本平均〕
  - **Variance** $\sigma^2 = \Sigma (x_i - \mu)^2/N = \Sigma x^2 - (\Sigma x)^2/N = \Sigma x^2 - N\mu^2$
    變異數 $s^2 = \Sigma (x_i\text{-}\mathbf{x})^2/n\text{-}1 = \Sigma x^2 - (\Sigma x)^2/n\text{-}1 = \Sigma x^2 - n\mathbf{x}^2$
  - **SD** 標準差 σ〔母群體〕or **s**〔樣本〕= $\sqrt{}$**Variance**
  - **CV** 變異係數 coefficient of variation (%) = **SD**/mean = **s**/**x**
    - **SPSS** 分析 描述性統計 預檢資料 (次數分配表 Q)
- 次數、比例、百分比 **%** ~ 用於名義、次序變項 **(**類別型資料**)**

3. 統計方法的選擇
Selecting a Statistical Test

	名義	數值
名義	大樣本-- 卡方檢定 X^2 test 小樣本-- Fisher's exact test 相關強度-- odds ratio	兩組平均值比較 - t test 三組平均值比較 - ANOVA
數值	x	相關分析 -- correlation coefficient (r) 線性迴歸 $y= a+b_1x_1+b_2x_2$

Z test 檢定, t test 檢定, 變異數分析 (ANOVA, F檢定), 相關分析 (Correlation analysis), 迴歸分析 (Simple or multiple regression analysis: number of x axis), 無母數分析 (Nonparametric analysis, 卡方X^2)

分類數據 (測量尺度, 資料種類)

- **類別型 (Categorical data)**
  - 名目變項 (Nominal variable): 性別、人種
  - 次序變項 (Ordinal variable): 教育、喜好程度
- **數值型 (Numerical data)**
  - 離散型 (Discrete)：(整數值)，家中小孩人數
  - 連續型 (Continuous)：(可插入小數)，身高

統計檢定：Z test 檢定, t test 檢定, 變異數分析 (ANOVA), 相關分析 (Correlation analysis), 回歸分析 (Regression analysis), 複回歸分析 (Multiple regression analysis), 無母數分析 (Nonparametric analysis, X^2)

Selecting a Statistical Test

Goal	Type of Data		
	Measurement (from Gaussian Population) 連續變項且爲常態分佈	Rank, Score, or Measurement (from Non-Gaussian Population)	Binomial (Two Possible Outcomes) 二項式變數
Describe one group	Mean, SD	Median (Q_2), interquartile range (Q_1-Q_3)	Proportion (%)
Compare one group to a hypothetical value	One-sample t test	Wilcoxon test	Chi-square or Binomial test
Compare two unpaired groups	Two-sample t test (unpaired t test)	Mann-Whitney test/ Wilcoxon rank-sum test	Fisher's test (chi-square for large samples)
Compare two paired groups	Paired t test	Wilcoxon signed-rank test	McNemar's test
Compare three or more unmatched groups (≥3)	One-way ANOVA	Kruskal-Wallis test	Chi-square test
Association between two variables	Pearson correlation	Spearman correlation	Contingency coefficients
Predict value from another measured variable (由x預測 y值)	Simple linear regression $y = a + bx$	Nonparametric regression	Simple logistic regression Cox regression (Survival analysis)
Predict value from several measured or binomial variables	Multiple linear regression		Multiple logistic regression

統計分析與檢定 (機率分布)

- 標準常態分佈 (Z 檢定)
- t 分佈
- F 分佈
- X^2卡方分佈
- 相關係數 r
- 迴歸分析　迴歸係數 b　$Y=a+b_1x_1+b_2x_2$
- 直線方程式 (transformation 轉換)
- 信賴區間 臨界值 信賴水準 ***p*** value (H_0)
- 符號等級排序中位數檢定 (無母數分析)
- **(比較 mean, variance, %, median, mean difference)**

4. 常見的統計檢定方法

Z 檢定, t 檢定, 變異數分析 (ANOVA, F檢定), 相關分析 (Correlation analysis), 迴歸分析 (simple or multiple regression analysis: x axis), 無母數分析 (Nonparametric analysis, X^2 檢定) (mean Var SD median % difference)

- **名義變項 vs. 名義變項**
  - 卡方檢定: X^2 (Pearson chi-square) test -- contingency table *(rxc)* 列聯表
    - 大樣本 (Pearson's chi-square p value, 2x2 table 用 Yate's continuity correction *p* value)
    - 小樣本 (有一格子期望值 E < 2 or 20%以上格子 < 5): 用 Fisher's exact test *p* 值
- **數值變項 vs. 數值變項**
  - 迴歸分析，相關分析〔Pearson **r**: 相關之強度與正負方向 -1~+1〕
- **名義變項 vs. 數值變項**
  - 兩組平均數比較: two sample **t** test~ paired or unpaired **t** test
    - **Paired** (dependent sample, 相關樣本): 前後測量、配對樣本 (平均數差*d*)
    - **Unpaired** (independent sample, 獨立樣本): **Variance equal or unequal**
      - 先看變異數相等或不相等的 **Levene / F檢定**的 p值大小, 再看用哪一列 **t test** 的 *p* 值
      - Levene的 **p>** 0.05= 看**Variance equal**的 (pool t test) or **p<** 0.05= **unequal** 的 *p* (separate t test)
  - ≥3組平均數比較: 變異數分析 **ANOVA (F)**: **post hoc** comparison if p<0.05
- **Non-parametric analysis 無母數分析:**
  - **When:** 小樣本 (small sample size), outlier, 不是常態分布, 結果爲類別變項
- 單一樣本 **Z** 檢定 (樣本是否具母群體代表性, i.e. 取樣是否偏頗, 很少使用到)

5. 有母數統計 對應的 無母數分析 (rank)
Nonparametric analysis

- One sample t test (樣本數小用無母數統計較有利, 用有母數分析通常不顯著)
  - Wilcoxon test or signed test (兩項式分布 檢定母體比率是否等於1/2)
- Paired t test (兩個相關樣本) (paired data)
  - **Wilcoxon signed-rank test** (Binary categorical paired data: McNemar's X^2)
- Two sample t test (兩個獨立樣本) (unpaired data)
  - **Wilcoxon rank-sum test** (排序等級總合雙樣本z檢定) or
  - Mann-Whitney U test (分爲小樣本U: all<10和大樣本Z: any one >10 兩種)
    - **SPSS** 分析 無母數分析 兩個獨立樣本檢定 檢定變數➢ 分組變數➢ 按定義組別
- One-way ANOVA (≧3組, K個獨立樣本檢定)
  - **Kruskal-Wallis test** (ps.≧**3 paired** data use Friedman ANOVA
- Pearson correlation coefficient (r)
  - **Spearman** correlation coefficient (r_s)

Selecting a Statistical Test

推論樣本　　假設檢定

- One sample mean
- Two sample means
- Three or more sample means
- Two continuous variables
- Two categorical variables
- **Multiple regression model**
  - $\mathbf{Y}=a+b_1x_1+b_2x_2+\ldots+b_nx_n$
  - Y = continuous or categorical
  - X = single or multiple x
    - ps. 在ANOVA, x稱爲因子, y稱爲依變數
    - b = regression coefficient = slop
    - a = intercept, y value when x=0

- Z test, t test (small sample or σ unknown)
- two sample t test (pair or unpaired)
- ANOVA (F test, Post hoc comparison)
- r (correlation), Regression (prediction)
- Chi-square test (X^2)
- **Y**= continuous: **M Linear** reg.
- **Y**= categorical: **M Logistic** reg.
- **Y**= categorical while consider time: ~ **Cox** Regression (Survival analysis, Cox Proportional Hazard Regression)
- Nonparametric analysis (rank median)
  - X^2, Wilcoxon, K-W test, Spearman r

6. 假設檢定(統計推論)

- **One** sample mean (單一樣本平均數檢定 用 **Z, t** 檢定)
  - Z = (x-μ)/ $\sigma \div \sqrt{n}$, μ之 95% CI = $\mathbf{x \pm 1.96}\,\sigma/\sqrt{n}$ (以樣本估計母體SE) **Z=1.96**, p<0.05
  - t = (x-μ)/ s$\div\sqrt{n}$ (SE= s/$\sqrt{n}$) , μ的 CI 區間估計= $x \pm t_{(\alpha/2,\,df)}\sigma_x = x \pm \mathbf{t}_{(\alpha/2,\,df)}\,s/\sqrt{n}$
- **Two** sample means (two samples **t test**)
  - **Paired** or **unpaired data (t)**
    - **Unpaired**: Variance equal or unequal (check Levene test p value: select pool or separate t test)
- **Three** or more sample means **(ANOVA, 變異數分析, F 檢定)**
  - F 檢定：SSB/SSW (組間變異/df 與 組內變異/df 的比值= MSB/MSW為 F分布)
- **Two continuous variables**: Pearson **r**, **regression analysis**
  - **t 檢定 = $(r-r_u)/s_c$** $s_c = \sqrt{(1-r^2)/n-2}$, df= n-2 Y=a+bx
- **Multiple regression model** (X軸: 多變項迴歸分析) (由x值預測y值)
  - Y= continuous or categorical variable or category with time to event
    - 1. Linear regression $Y=a+b_1x_1+b_2x_2+\ldots+b_nx_n$ (**b** = **slope=** 迴歸係數)
    - 2. Logistic regression (binary outcome): **OR** (adjusted Odds Ratio)
    - 3. Cox regression (binary & time, survival analysis): **HR** (Hazard ratio, RR)
- **Nonparametric analysis (無母數分析) (類別 %、次序資料 rank)**
  - 單一樣本二項式檢定、單一樣本分布次數之X^2卡方適合度檢定… (cont.)

7. 類別資料分析 (Categorical data)

- 單一比例 **(a single proportion) %**
  - Rationale: 遵守二項式分布，但若 np與 n(1-p)都大於5的話，幾乎等於常態分布, 估計平均值 **p**= x/n, 估計標準差爲 $\sigma_p = \sqrt{p(1-p)/n}$
  - π 之 **95% CI = p ± 1.96**$\sqrt{p(1-p)/n}$
    - 若要保證誤差範圍爲 ± d%， $1.96\sqrt{p(1-p)/n} \le 1.96\sqrt{1/4n} \le$ d%
      - **n:** sample size calculation, **n ≥ 4 (1.96/d%)²** vs. 連續型 **n**= $(Z_{\alpha/2}\ \sigma/d)^2$
- 兩個比例值 **(two proportions)**
  - 卡方檢定 (2x2 table): 視樣本相關或不相關及樣本大小而選用不同的p值
    - 不相關兩組 (independent groups data): Pearson X^2= [Σ ｜O-E｜ -1/2]² /E, df=1(½ :校正)
      - 95% CI = $(p_1-p_2) \pm 1.96\sqrt{p_1q_1/n_1 + p_2q_2/n_2}$ (比例數差的檢定) (2x2 table: **Yate's continuity correction**)
    - 相關組別 (related, paired): McNemar's test X^2 = (｜b-c｜ - 1)²/ b+c (只有 bc格改變)
    - 小樣本: 若有任一格期望值 < 5 (違背假設): Fisher's exact test (不須符合卡方分布)
  - 獨立: 用 X^2; 相關（兩次測量）用 McNemar's X^2; 小樣本 (一格期望值<5)用 Fisher's exact test
- 超過兩個類別 **(more than two categories)** (row x column)
  - $X^2 = \Sigma (O - E)^2 / E$, df= (r-1) (c-1) [r x c 列聯表, df >1 不必做連續性校正]
  - $X^2_{(r-1)(c-1)} = \Sigma_r \Sigma_c (O_{ij} - E_{ij})^2 / E_{ij}$ $E_{ij} = n_i \times n_j / N$ df= (r-1) (c-1)
    - 當觀察值(O)與期望值(E)的差異(X^2值)很大，則拒絕H_0 ($X^2 = 3.84 = 1.96^2 = Z^2$)
- 相關強度(風險估計): **Odds ratio** (勝算比), **Relative risk** (相對風險)

相關與迴歸 (目的:做預測)

線性相關: 測兩變項的相關強度; 線性迴歸: 以x來預測y值

- **相關係數 r** (Pearson's product moment) $= \Sigma(X-\mathbf{X})(Y-\mathbf{Y}) / \sqrt{\Sigma(X-\mathbf{X})^2 \Sigma(Y-\mathbf{Y})^2}$
  - 虛無假設 r (or 母體 rho) = 0, t 檢定= $\mathbf{r}/\mathbf{s_E}$ $\mathbf{s_E}=\sqrt{(1-r^2)/n-2}$, df= n-2 , r = 負 -1~+1 正相關
  - r 之計算不要求符合LINE前提 (但建議做殘差分析及影響力分析: 注意outlier)
  - r 沒有單位，不受X, Y的單位影響，可稱爲**標準化**迴歸係數
- **SLR迴歸係數(斜率b)與 相關係數 r 的關係**
  - **Y = a + bx + e (b 爲直線迴歸方程式的斜率, a 爲截距: intercept, e: error)**
    - **X**= x_o $\mathbf{Y_1} = a+b_1x_o$
    - **X**= x_o**+1** $\mathbf{Y_2} = a+b_1(x_o+1)$ (迴歸方程式：以 x來預測 y值)
    - 以上兩式相減: $\mathbf{Y_2-Y_1} = \mathbf{b_1}$
  - **slop b (當X每改變±1個單位時, Y的改變量爲 b)** $= \Sigma(x-\mathbf{x})(y-\mathbf{y}) / \sqrt{\Sigma(x-\mathbf{x})^2}$
  - Multivariates: 當$\mathbf{x_1}$增加1且其他**x** 維持不變時(控制control 或調整adjust), $\mathbf{b_1}$等於此時 **y** 變動的平均值。($b_{1,2...k}$: **partial regression coefficients**)
    - $\mathbf{r} = \Sigma(x-\mathbf{x})(Y-\mathbf{Y}) / \sqrt{\Sigma(X-\mathbf{X})^2 \Sigma(Y-\mathbf{Y})^2} = b\sqrt{\Sigma(X-\mathbf{X})^2} / \sqrt{\Sigma(Y-\mathbf{Y})^2} = \mathbf{b}\, \mathbf{S_X}/\mathbf{S_Y}$
    - i.e. $\mathbf{b} = \mathbf{r}\, S_Y/S_X$ or $\mathbf{r} = \mathbf{b}\, \mathbf{S_X}/\mathbf{S_Y}$ (標準化後之斜率)
  - 相關係數 r 之平方 = 簡單線性迴歸的 R^2
- **Pearson** correlation coefficient **r**: 線性相關強度, **矩陣散佈圖(r×c)**, 相關矩陣表
  - 論文報告可用: r (or r_s), matrix scatter plot, or correlation matrix 來表示

Interpretation of the regression coefficients (HR, same for OR, RR)

- An estimated hazard rate ratio greater than **1** indicates the covariate is associated with an increased hazard of having the event of interest
- An estimated hazard rate ratio less than **1** indicates the covariate is associated with an decreased hazard of having the event of interest
- Estimated hazard rate ratio of **1** indicates no association between covariate and hazard. *(Null hypo.)*

Selecting a Statistical Test I

Goal	Type of Data		
	Measurement (from Gaussian Population) 連續變項且爲常態分佈	Rank, Score, or Measurement (from Non-Gaussian Population)	Binomial 二項式變數 (Two Possible Outcomes)
Describe one group	Mean, SD	Median (Q_2), interquartile range (Q_1-Q_3)	Proportion (%)
Compare one group to a hypothetical value	One-sample t test	Wilcoxon test	Chi-square or Binomial test
Compare two unpaired groups	Two-sample t test (unpaired t test)	Mann-Whitney test/ Wilcoxon rank-sum test	Fisher's test (chi-square for large samples)
Compare two paired groups	Paired t test	Wilcoxon signed-rank test	McNemar's test
Compare three or more unmatched groups (≥3)	One-way ANOVA	Kruskal-Wallis test	Chi-square test
Association between two variables	Pearson correlation	Spearman correlation	Contingency coefficients
Predict value from another measured variable	Simple linear regression	Nonparametric regression	Simple logistic regression
Predict value from several measured or binomial variables	Multiple linear regression		Multiple logistic regression

Selecting a Statistical Test II

Scale of measurement	Type of Experiment (Treatment)				
	Two treatment groups consisting of different individuals	Three or more treatment groups consisting of different individuals	Before and after a single treatment in the same individuals	Multiple treatments in the same individuals	Association between two variables
Interval *	Unpaired t test	Analysis of variance (ANOVA)	Paired t test	Repeated measures ANOVA	Linear regression Pearson correlation coeff.
Nominal	Chi-square analysis of contingency table	Chi-square analysis of contingency table	McNemar's test	Cochrane Q	RR or OR
Ordinal	Mann-Whitney test/ Wilcoxon rank-sum test	Kruskal-Wallis test	Wilcoxon signed-rank test	Friedman ANOVA	Spearman rank correlation
Survival time	Log-rank test				

* If the assumption of normal distribution is not meet, rank the observations and use the method for ordinal scale.

Statistical Methods I

Data Type	Comparison	Association / Prediction	Data Reduction
Continuous (SBP mmHg)	**t-test** -- paired/ unpaired **(M) ANOVA** - Repeated measures Discriminant analysis	**Correlation** coefficient **Multiple linear regression** (slop b, regression coefficient)	Principal components analysis Factor analysis
Categorical (High or normal)	Chi-square	**Logistic regression** (Odds ratio) Log-linear model	Correspondence analysis
Censored (time to event)	Log rank test (K-M survival curve)	**Cox regression** (Hazard ratio, RR)	

Statistical Methods II

Dependent variable (Y)	Independent variable (X)		
	Binary (e.g.男生/女生)	**Continuous** (身高)	**Categorical** (人種)
Continuous (收縮壓 mmHg)	t-test -- paired / unpaired Anova -- repeated measures	相關 Correlation - Pearson (r) - Spearman (r_s) 迴歸分析:SLR, **MLR**	ANOVA ANCOVA
Categorical (including binary) (正常或高血壓)	Chi-sq Logistic/log-linear CART	**Logistic regression** (Odds ratio) Discriminant analysis	Chi-sq Log-linear CART
Censored (存活分析)	Log rank test	**Cox regression** (Hazard ratio, RR)	Log rank test
None		Factor analysis	Correspondence analysis

Scientific Paper

- **Identify problem: important issue (Title)**
  - Introduction (paper review, clinical experience)
- **Sound method (avoid bias)**
  - Material & Method (detail, note confounding factor)
- **Result**
  - Be brief and to the point
  - Non-biased information (include statistics)
  - Discussion: main findings & limitation
- **Logic, Clarity, and Precision**
  - By chance alone multiple associations (*p value*)
  - Clinical rationale, biological plausibility

統 計 (Statistics)

- 敘述性統計 (Descriptive statistics)
  - 集中趨勢 (Central tendency 中央趨勢): **mean**, median, mode
  - 離散趨勢 (Dispersion, Variability 變異): **SD**, **variance**, range, CV
  - 表格、圖形、關聯的測量 (association 變數之間是否相關)
- 推論性統計 (Inferential statistics)
  - 由樣本推論母群體的特徵 (機率分佈, 面積, 做檢定, p value)
    - Sample 樣本 vs. Population 母體
      - Statistics 樣本統計量 s, p vs. Parameter 母體參數 σ , π
      - **Data 數據資料:** 實驗數據 **experimental** data, 調查 **survey data**
      - **Type of data (資料的分類):** (~ 可使用的統計方法不同, Table)
        - 數值型 **(numerical data)** or 類別型 **(categorical data)**
  - **Variable: 變項、變數 Y= a+bx**
    - **X** 軸 **:** independent, predictor, explanatory, covariate (獨立, 自, 解釋, 共變數, 分組變數)
    - **Y** 軸 **: dependent**, **outcome**, or **response** variable (依變數, **結果**變數, **反應**, 檢定變數)

Variable 變數: 區分變項的種類

- 類別型或數值型資料 **(Data)** (統計方法: mean, variance, difference, %, median)
  - 類別型 **(Categorical data):** Nominal, Ordinal variable (%, median, rank)
  - 數值型 **(Numerical data):** 資訊較豐富, 統計方法多 (mean, variance)
    - 離散、間斷型 (Discrete variable)：點計不可分割的整數值，例如人數、床數
    - 連續型變數 (Continuous variable)：可插入小數，例如身高、體重

- 自變數與應變數 **(Independent or Dependent Variable)**
  - 自變或獨立變數 (Independent variable, x軸)
    - 由實驗者操弄、控制的變項
  - 應變或相依變數 **(Dependent variable, outcome variable)**
    - 因上述自變數改變，而發生改變的 結果變數 (Y axis)

- 變項 (Variable) 間的轉換 (包括轉換成線性關係)
  - 數值型可轉成類別型： 血壓 140/80 mmHg, 正常 不正常 (高)
    - **SPSS** 轉換 **重新編碼** 成不同變數 ➢ 舊值與新值

敘述性統計 Descriptive statistics

Mean ± SD (range), median (range, or interquartile range), %

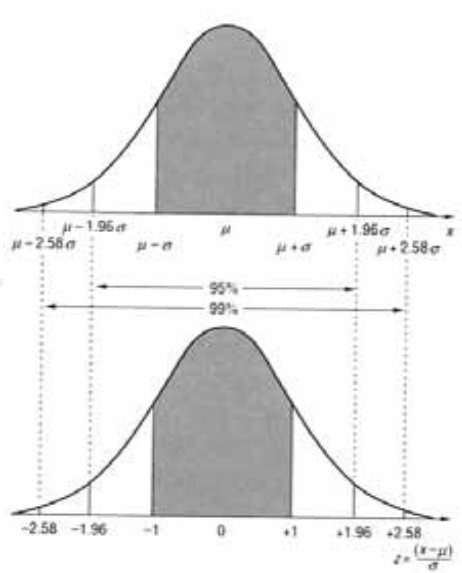
- 集中趨勢 (Central tendency)
  - **Mean** 平均數、受極端值影響 **x** (樣本平均 **x** bar) = $\Sigma x_i / n$ (母體 $\mu = \Sigma x_i/N$)
  - **Median** 中位數、不受極端值影響 (用於有極端值、非常態分布時)
  - Mode 眾數、又稱流行值 (商業用途) (極端值: outlier)
- 分散趨勢 (Dispersion 分散性, Variation 變異性)
  - **Range** (Min~Max or Max–Min) 全距, **Interquartile range $Q_{1\text{-}4}$**
  - Mean deviation 平均離差〔與平均值相差之絕對值的平均〕
  - Sum of square: $\Sigma (x_i - \mathbf{x})^2$
  - **Variance** 母體 $\sigma^2 = \Sigma (x_i - \mu)^2/N = \Sigma x^2 - (\Sigma x)^2/N = \Sigma x^2 - N\mu^2$
    變異數 樣本 $s^2 = \Sigma (x_i - \mathbf{x})^2/n\text{-}1 = \Sigma x^2 - (\Sigma x)^2/n\text{-}1 = \Sigma x^2 - n\mathbf{x}^2$
  - **SD 標準差** σ〔母群體〕or **s**〔樣本〕= $\sqrt{}$**Variance**
  - **CV** 變異係數 coefficient of variation (%) = **SD**/mean = **s/x**
  - **SPSS** 分析 描述性統計 預檢資料 (次數分配表) 〔連續資料分5-15組, 組距中點〕
- 次數 比例 百分比%: 用於名義及次序之類別型資料

離中趨勢 Spread, Dispersion, Variability

Mean ± SD (range), Median (range, or interquartile range)

- **Range** (min~max, or max-min 全距)
- **Quartiles,** Percentiles, Deciles (用於偏移**不對稱資料,** 有**outlier**)
  - **Interquartile range (Q2= median, Q_1-Q_3)** (可用盒形圖表示)
  - Interdecile range (D_{10}-D_{90})
  - **SPSS** 分析 描述性統計 次數分配表 變項➢ **按統計量** 勾**選項 (Q, $D_{10-90...}$)**
  - IQV (index of qualitative variation, 類別資料)
    - 觀察值的差異組合數/變異量最大可能數目，最大可能數目: n^2 (k-1)/2k
- **SD (s), Variance (s^2)** (用於**對稱資料,** 受極端值影響, 不適於偏移資料)
  - **Standard deviation 標準差, Variance 變異數**
- **CV** (coefficient of variation, 變異係數 CV = SD/mean)
  - **Coefficient of relative variation** (%) **= CRV = CV x 100%**
  - 用於生化檢驗數值，重複測量時的變異情形
  - 比較測量單位不相同之分布
  - 比較測量單位相同，但平均數差異很大之分布
    - SD 無法比較不同觀察值間的分布之分散程度 e.g. **5±1, 50 ±1** (SD皆是 ± 1)

Mean and SD 應用



- 常態分布　　面積
  - **Mean ± 1 SD**　　~ 68%　(68.3%)
  - **Mean ± 2 SD**　　~ 95%　(95.4%)
  - **Mean ± 3 SD**　　~ 99%　(99.7%)
  - **Q**: 由文章 mean ± SD看出 data分布範圍 (range), 及有無 outlier
    - **Mean ± 4 SD**　　~ 100% **= range (看有無 outlier)**
    - Whole range約有8個 SD, 所以知道範圍也可算出 **SD = range / 8**
  - **SPSS** 分析 描述統計 **預檢資料** 統計圖 機率圖附檢定 **(K-S, Shapiro-Wilk)**
  - **SPSS** 圖表 **直方圖** 選變數➢ 勾顯示常態曲線 (看資料是否**常態分布**)
  - 有極端值(outlier)或不對稱分布時, 改用 **Median** (interquartile range, Q_1~Q_3) 來表達資料的分布情形
    - 轉換: Normal distribution with skew to the right data 取 log 會接近常態分布
- **Normal distribution➢ Standardized normal (Z) distribution**
  - Z-score: 標準化常態分布下, 距離平均數幾個標準差. **z=** x_i**-x/ s** (樣本)
    - **Z=** $x_i - \mu / \sigma$ (母群體); Z~ N (μ =0, σ^2 =1)　分佈面積、百分比、機率 *p值*

敘述性統計: 兩變數之間的關係

- **A.** 表格、圖形
  - **Contingency table** 列聯表, cross table (用於**類別**資料)
    - SPSS 分析 描述統計 **交叉表** 選變項➢ 變項➢ 按**格** (儲存格顯示...勾%)
  - **Scatter plot** 散佈圖 (用於**連續**資料)
    - SPSS 圖形 **散佈圖** Y軸=依(結果)變數➢ X軸=自(獨立)變數➢ (or 盒形圖)
- **B.** 數字指標 **(關聯性指標) Measures of association**
  - **1.** 名義資料 (原理：proportional reduction in error, PRE 的測量)
    - **Lambda**, Goodman and Kruskal tau, **Cramer's V**=$\sqrt{x^2/N(K-1)}$ K:行列中較小者
    - SPSS 分析 描述統計 **交叉表** ➢ ➢ 按**統計量**再勾選**名義的**選項
  - **2.** 次序資料 (原理：排序上 concordant, discordant 之間的差異 D)
    - Somer's *d*, **Gamma**, Kendall's tau-b, Kendall's tau-c: SPSS...勾**次序**選項
    - **Spearman's correlation coefficient (r_s: -1~1, 0: 沒有相關)**
    - SPSS 分析 **相關 雙**變數 ➢➢勾選 Spearman's rank rho, $r_s=1-6\sum D^2/n(n^2-1)$
  - **3.** 等距, 等比資料: $\mathbf{b}=\sum(x-\mathbf{x})(y-\mathbf{y})/\sqrt{\sum(x-\mathbf{x})^2}$, $\mathbf{r}=\sum(x-\mathbf{x})(y-\mathbf{y})/\sqrt{\sum(x-\mathbf{x})^2\sum(y-\mathbf{y})^2}$
    - **Linear regression**: slop **b** = regression coefficient (正負相關, 做預測)
    - **Pearson's correlation coefficient :** r, r^2 (決定係數:選模式) (強度與方向)

兩變數之間的 關聯性指標

- 連續變數 $b=\sum(x-\mathbf{x})(y-\mathbf{y})/\sqrt{\sum(x-\mathbf{x})^2}$, $r=\sum(x-\mathbf{x})(y-\mathbf{y})/\sqrt{\sum(x-\mathbf{x})^2\ \sum(y-\mathbf{y})^2}$
  - **Scatter plot** 散佈圖 (連續性資料: 等距, 等比 data)
    - SPSS 圖形 **散佈圖 Y** 軸=依變數➢ **X** 軸=自(獨立)變數➢
      - Scatter plot按兩下..圖表..選項 勾全體式 (可**畫**出迴歸**直線**)
  - **Linear regression:** (slop **b** = regression coefficient, 方向)
    - SPSS 分析 迴歸 線性 依變數Y➢ 自變數X➢ (迴歸統計量)
      - least square method (觀察值和期望值之間差異的平方最小 的直線方程式)
      - 迴歸分析前提假設：linear relationship, independent, normal distribution, homoscedasticity (變異數同質性: 回歸線之誤差項的變異數-殘差-爲常數)
  - **Pearson's correlation coefficient** (**r**, 強度方向)
    - r (相關係數, 關係的強度及正負方向, -1~1) (0:沒有相關)
    - R^2 (決定係數, 預測的可信度, 可解釋變異的比例)
      - 迴歸線可解釋之Y_i到Y_{hat}預測值的變異量相對於Y_i到Y_{bar}平均值的變異
      - **Y**的總平方或變異量，有多少可被此直線迴歸模式所解釋 **(決定模型好壞的選擇)**
    - **SPSS** 分析 **相關** 雙變數 ➢➢ 點選 Pearson r (vs. 無母數 Spearman's rho)
- 二元變項 (Binary variables) 關係強度之測量: RR, OR

敘述性統計 Descriptive statistics
~ 表格與圖形 ~

- Frequency table 次數表
  - 一般分成 5-15 classes (class interval 組距)
- Graph 圖形
  - **Histogram** 直方圖, frequency polygon 次數多邊圖, cumulative frequency polygon 累積次數多邊圖, stem-and-leaf display 莖葉圖, bar chart 長條圖, pie chart 圓餅圖, **box** and whisker plot 盒鬚圖 (Quartiles) **(➢ 長條圖, 圓餅圖用於次序及名義資料, 直方圖用於連續資料)**
- Skewness 偏態, Kurtosis 峰度
  - **Skewed** to the **right** or **positively** skewed: Peak最高點不在中央, 尾巴在右邊稱. **(Mean> Median> Mode)**
    - Pearson偏態係數看對稱性, $SK_p = \mu - M_0 / \sigma$ (± 3之間, ± 0.5 if use s, 對稱爲 0)
  - Movement method 動差偏態係數 (判定對稱性)
    - 動差偏態係數 α_3 第三級動差 (立方)= 0-0.5 近似對稱, 0.5-1 略偏, >1 極偏 (絕對值)
    - 動差峰度係數 α_4 (四次方) = 3 常態 meso-, >3 高狹 lepto-, <3 低闊峰 platy-kurtosis

統計方法 統計檢定A

1. 名義變項 vs. 名義變項

- 卡方檢定〔列聯表, contingency table〕
  - 大樣本 (Pearson chi-square, p.s. 2x2 table 用 Yate's correction p value)
  - 小樣本 (一格期望值 E <2 or 20%以上格子 <5): 用 Fisher's exact test

2. 數值變項 vs. 數值變項

- 相關分析 (Pearson correlation coefficient, r or Speraman's rho)
- 迴歸分析 (regression analysis) ~ simple or multiple linear regression

3. 名義變項 vs. 數值變項

- 2組 平均數比較: t test~ paired (dependent) or unpaired (independent)
- 3組或以上 平均數比較: ANOVA (變異數分析, F檢定)

- Z 檢定, t 檢定, 變異數分析 (ANOVA, F test), 相關分析 (Correlation analysis), 迴歸分析 (Regression analysis), 無母數分析 (Nonparametric analysis, chi-square test X^2 卡方檢定)

統計方法 統計檢定B

- 名義變項 **vs.** 名義變項
  - 卡方檢定: X^2 (Chi-square) test- 列聯表 contingency table (r x c)
    - 大樣本 (Pearson's chi-square, but 2x2 table 用 Yate's continuity correction)
    - 小樣本 (有一格子期望值 E <2 or 20%以上格子 <5): 用 Fisher's exact test
- 數值變項 **vs.** 數值變項
  - 迴歸分析 (做預測)，相關分析 (Pearson **r**: 正或負相關, 相關強度)
- 名義變項 **vs.** 數值變項
  - **t test** (2組的 平均數做比較~ two sample **t**, paired or unpaired)
    - **Paired** (dependent, 相關樣本): 兩平均數差值(d)比較 (**d** 成爲單一樣本)
    - **Unpaired** (independent, 獨立樣本): **Variance equal** or unequal 公式不同
  - **ANOVA** (3組或3組以上 平均數比較: 變異數分析~ ANOVA, **F** test)
    - Analysis of variance (F test, if p<0.05 then do post hoc pairwise comparison analysis)
- 次序變項 **(類別變項)**
  - 無母數分析 **Non-**parametric analysis: 小樣本, outlier, 非常態分布, 類別%
- 單一樣本 Z 檢定 (樣本是否具母體代表性, 取樣是否偏頗)

有母數統計對應的無母數分析 A

- **t** test (two samples): **unpaired t** or **paired t test**
  - **Wilcoxon rank-sum test** (排序等級總合檢定) (**Unpaired** data)
    - $Z = (W_{小} - \mu_w) / \sigma_w$　$\mu_w = n_1(n_1+n_2+1)/2$,　$\sigma_w = \sqrt{n_1 n_2 (n_1+n_2+1)/12}$,　$n > 20$爲常態
      - n_1 爲兩樣本中個數較小的個數,　n_2 爲兩樣本中個數較大的個數, w = Wilcoxon rank sum
      - 描述統計值 Rank sum, mean rank (母群體中位數的區間估計)
    - **SPSS** 分析 無母數分析 兩個**獨立樣本**檢定 檢定(依)變數➢ 分組(自)變數➢ 按定義組別
    - **Mann-Whitney U test**　$U = n_1 n_2 + n_1(n_1+1)/2 - \sum R_1$　$\sum R_1$=人數較少一組的等級總合
      - $z = (U - \mu_u) / \sigma_u$　$\sigma_U = \sqrt{n_1 n_2 (n_1+n_2+1)/12}$　(與 Wilcoxon W test　z 檢定相同)
  - **Wilcoxon signed-rank test** (**paired** data) vs. **paired t test**
- **ANOVA** (3組或3組以上) (post hoc analysis if significant difference)
  - **Kruskal-Wallis test**
    - Kruskal-Wallis H (相當於多個 Wilcoxon test) 檢定是卡方檢定, 非z檢定
- Pearson correlation coefficient (**r**) 相關係數
  - Spearman correlation coefficient (r_s) 等級相關係數

無母數分析 B (正負符號、等級排序, 中位數)

- 無母數分析的使用時機
  - 1. 資料不是常態分布 (distribution free): 無母數分析 or 先轉換成常態
  - 2. 不足以形成常態分布之小樣本 或 有outlier值
  - 3. 結果爲類別變項: 卡方檢定 X^2 test
    - 單一樣本之二項式檢定 (binomial distribution, e.g. 有病或沒病, 男生或女生)
      - 樣本百分比的抽樣分布 (%抽樣分布之標準誤 $\sigma_p = \sqrt{pq/n}$)
        - Z檢定比較二元比例: $z_s = (P_s -/+0.5) - P_u / \sqrt{pq/n}$ (大樣本z: np ≧ and nq ≧ 5)
        - **SPSS** 分析 無母數分析 二項式檢定 ➢ 檢定比例%
        - 估計母體成功比例 CI = $P_s \pm z \sqrt{pq/n}$
        - 隨機性之連檢定 runs test: $z_s = [(R -/+0.5) - \mu_R] / \sigma_R$ (是否隨機抽取)(基因序列比對) $\mu_R = (2n_1n_2/n_1+n_2) + 1$, $\sigma_R = \sqrt{(n^2-2n)/4(n-1)}$
        - **SPSS** 分析 無母數分析 連檢定 ➢ 分割點勾選(中位數..) (中位數: 大於H, 小於L)
    - 單一樣本分布次數之卡方 **適合度檢定 (k組 次序資料: >二元時)**
      - $X^2 = \sum (f_{obs}-f_{exp})^2/f_{exp}$, df=k-1 (觀察值與期望值的差稱爲殘差 residual, 殘差越大越不合理: reject)
      - **SPSS** 分析 無母數分析 卡方檢定 項目 ➢
        - (也可測試等距/等比資料是否爲常態分布; ><2, 1-2, 1 SD之% Exp: 2,14,34,34,14,2 %)
    - 雙及多樣本卡方檢定稱爲 **獨立性檢定** (X^2 test of independence)

Selecting a Statistical Test

Goal	Type of Data		
	Measurement (from Gaussian Population) 連續變項且爲常態分佈	Rank, Score, or Measurement (from Non-Gaussian Population)	Binomial 二項式變數 (Two Possible Outcomes)
Describe one group	Mean, SD	Median, interquartile range (Q_1-Q_3)	Proportion (%)
Compare one group to a hypothetical value	One-sample t test	Wilcoxon test	Chi-square or Binomial test
Compare two unpaired groups	Two-sample t test (unpaired t test)	Mann-Whitney test/ Wilcoxon rank-sum test	Fisher's test (chi-square for large samples)
Compare two paired groups	Paired t test	Wilcoxon signed-rank test	McNemar's test
Compare three or more unmatched groups (≥3)	One-way ANOVA	Kruskal-Wallis test	Chi-square test
Quantify association between two variables	Pearson correlation	Spearman correlation	Contingency coefficients
Predict value from another measured variable	Simple linear regression	Nonparametric regression	Simple logistic regression
Predict value from several measured or binomial variables	Multiple linear regression		Multiple logistic regression

推論性統計 (Inferential statistics)

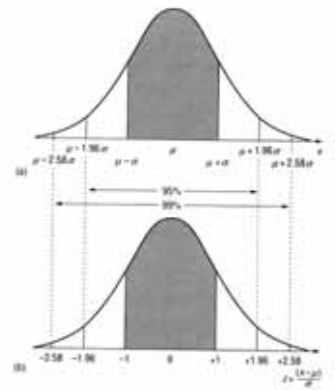
- 由樣本推論母群體的特徵
  - 樣本 sample vs. 母體 population
  - 樣本統計量 statistics, 母體參數 parameter
    - 前提爲需是隨機抽樣樣本，才不致造成推論的錯誤
    - **樣本能否代表母群體 (爲推論是否正確的前提) !?**
      - Selection bias ~ Uncertainty in medical decision making
- **數據 data, 變項 variable**
  - 調查數據 survey data, 實驗數據 experimental data
  - 決定: **Dependent variable, Independent variable**
    - **X軸: independent, predictor, explanatory, covariable, (factor)**
      (獨立變數, 自變數, 解釋變數, 共變數)
    - **Y軸: dependent, outcome, or response variable**
      (結果變數, 依變數, 應變數)
    - 先區分X, Y 變數分別爲數值型(連續、離散)或類別型(名義、次序)變項

統計推論 (Inferential statistics)

- 統計推論可分爲
  - 估計 **(Estimation)**
    - 點估計　20%
    - 區間估計 18-22%, 15-25%, 10-30%
    - 信賴區間 (confidence interval, 95% CI)
    - 機率分佈: Z, t, F, X^2 distribution…~ p value
  - 檢定 **(Test of Hypothesis)**

統計推論 (Inferential statistics)

- 統計推論可分為
  - 估計 **(Estimation)**
    - 點估計 20%
    - 區間估計 18-22%
    - 信賴區間 (confidence interval, 95% CI)
    - 機率分佈: Z, t, F, X^2 distribution… ~ p value
      - t 分配, df= n-1, 當 n≧30 時近似常態分配 (central limit theory)
        - N (μ, σ^2), 抽樣形成 t分配, t=x- μ /s_x, df= n-1 (Gosset)
  - 檢定 **(Test of Hypothesis)**
    - 列出虛無及對立假說 (H_o: Null hypothesis, H_1: Alternative hypothesis)
    - 選擇顯著性檢定方法 ($p< 0.05$, Reject H_o: statistical significance)
    - 計算樣本統計量 **z= x- μ / $\sigma \div \sqrt{}$ n,** t= **x- μ / s$\div \sqrt{}$ n** (n小 or σ unknown)
    - 建立臨界值及臨界域 significant level α, (1- α) x 100% CI, Z (t), *p* value 機率
    - 下決定 (比較樣本統計量與臨界值, 95%信賴區間, 樣本分布機率, p值)



檢定 (Test of Hypothesis)

Null hypothesis: H_0 檢定結果	真實情況	
	H_0 爲真 (沒有差異)	H_0 爲僞 (有差異)
接受 H_0 (Do not reject H_0)	正確判決	ß (第二型錯誤)
拒絕 H_0 (Reject H_0)	α (第一型錯誤)	1- ß (統計檢力)

1. α 機率 虛無假設爲真卻推翻它: 又稱 significant level, **type I error** (**0.05**, 0.01)
2. **Type II error (ß):** 虛無假設爲僞卻接受它, **ß** is relate to **power, sample size** calculation
3. **Statistical power 統計檢力** (H_1爲真時, 推翻 H_0, 即正確裁決): **1-ß** (**ß= 0.2**, 0.1)
4. 信賴區間: **(**1-α**)** x 100% C.I. (**95%** CI, 99% CI) (α = 1 - 信賴水準)

Confidence intervals (信賴區間, CI)

- 平均值的信賴區間 (CI for the mean)
  - 樣本數夠大，樣本平均值遵守 **常態分布** (常態且已知變異數)
    - **95% CI for the mean = sample mean (x) ± 1.96** x **SEM** (SEM= $\sigma \div \sqrt{n}$)
    - (2 SD includes 95% area under the normal distribution curve)
  - 非常態分布或變異數不知，則樣本平均值服從 **t 分布**
    - 平均值的 **95%信賴區間 = $x \pm t_{1-\alpha/2}\, s/\sqrt{n} = x \pm t_{0.975}$** x **SEM**, df= n-1
      - $t_{0.975}$ 爲 t 分布在 df = n-1 的 97.5 percentage point (percentile)
- 比率的信賴區間 (CI for the proportion)
  - 比例的取樣分布服從**二項式分布**
    - π 的 **95% CI = $p \pm 1.96 \sqrt{p(1-p)/n}$**
    - np or n (1-p) <5 時需用二項式分布計算, n 夠大則爲常態分布
- **P 值: 虛無假設對的機率 或 錯誤地推翻虛無假設的機率**
- 如何詮釋信賴區間~ **95%信心** (抽樣方法也會影響 p值)
  - **95% CI** 越窄表示估計值越精確，範圍較寬表示估計值較不精確
  - **p** 值爲犯 **type I error (α error)** 的機率 (α: significant level)
  - 標準誤**SEM** 影響 **C.I.** 的寬度，而 **SEM**又受 **sample size** 影響 (**n**大則 **CI**小)

統計推論 (Inferential statistics)

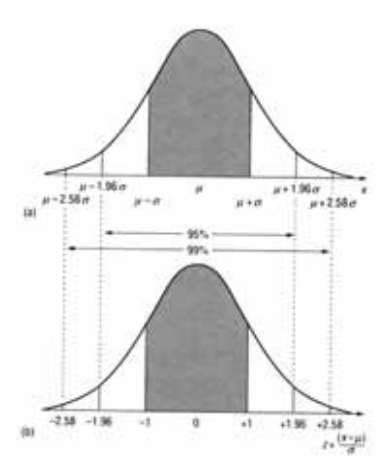
- 統計推論可分為
  - 估計 (Estimation)
    - 點估計 20%
    - 區間估計 18-22%
    - 信賴區間 (confidence interval, 95% CI) ~ α, n, σ, s 影響 C.I.
      - σ 已知, μ 之95% **CI** = **x** $\pm$ 1.96 $\sigma/\sqrt{n}$ **X** bar ~ N (μ, σ^2/**n**) Z~ N (0, 1)
        - μ 之 (1-α) x 100% **CI** = **x** $\pm$ $Z_{1-\alpha/2}\sigma_{\mathbf{x}}$ = **x** $\pm$ $\mathbf{Z}_{1-\alpha/2}\sigma/\sqrt{n}$
        - $Z_{1-\alpha/2}$ 是Z從負無限大到某數值下的面積, α =0.05, 95% CI, $\mathbf{Z}_{0.975}$ = 1.96
      - σ 未知, μ 之(1-α) x 100% CI = **x** $\pm$ $\mathbf{t}_{df,\ 1-\alpha/2}$ $s/\sqrt{n}$ (α =1-信賴水準)
        - 小樣本或 σ 未知, 抽樣形成 t 分配, t= (x-μ)/s_x, df= n-1 (Gosset)
        - 當 n≧30 時抽樣分佈近似常態分配
      - p_{hat} ~ N (π, π (1-π)/n), π 的 (1-α) x100% CI= p $\pm$ $\mathbf{Z}_{1-\alpha/2}\sqrt{p(1-p)/n}$
        - 應先檢查二項分佈是否近似常態: if np, nq, npq >5
  - 檢定 (Test of Hypothesis)
    - 信賴區間 臨界值 顯著水準 ~ ***p*** value ~ reject or accept null hypothesis

假設檢定(統計推論)

- **One** sample mean (單一樣本平均數 用 **z, t** 檢定)
  - $Z = (\mathbf{x}-\mu)/\ \sigma \div\sqrt{n}$ μ 之 95% CI = $\mathbf{x} \pm \mathbf{1.96}\,\sigma/\sqrt{n}$ (以樣本估計母體SE) **Z=1.96**, $p<0.05$
  - $t = (x-\mu)/\ s\div\sqrt{n}$ (SE= $s/\sqrt{n}$) μ 的區間估計= $x\pm t_{(\alpha/2,\ df)}\,\sigma_x = x\pm \mathbf{t}_{(\alpha/2,\ df)}\,s/\sqrt{n}$
- **Two** sample means (two sample **t test**)
  - **Paired** or **unpaired**, variance equal or unequal (Levene test *p*: pool t or separate t)
- **Three** or more sample means **(ANOVA, 變異數分析, F 檢定)**
  - F 檢定：SSB/SSW (組間變異/df 與 組內變異/df 的比值: 爲 F分布)
- **Two continuous variables**: Pearson **r**, **regression analysis**
  - **t** 檢定 $\mathbf{= (r-r_u)/s_c}$ $\mathbf{s_c}= \sqrt{(1-r^2)/n-2}$, df= n-2
- **Multiple regression model**
  - Y= 1. continuous or 2. categorical variable or 3. time to event
    - 1. Linear regression $Y=a+b_1x_1+b_2x_2+\ldots+b_nx_n$ (**b**= **slope=**迴歸係數)
    - 2. Logistic regression (binary outcome): **OR** (multivariate adjusted OR)
    - 3. Cox regression (consider time, survival analysis): **HR** (RR)
- **Nonparametric analysis** (無母數分析)
  - 單一樣本之二項式檢定、單一樣本分布次數之X^2卡方適合度檢定…

單一母群體平均數估計與檢定

- 單一母群體平均數 μ 的區間估計 (σ, n 決定 Z test, or t test)
  - 母群體 標準差 σ 已知 (很少見, z檢定)



    - μ 的區間估計= $x \pm Z_{\alpha/2}\sigma_{\bar{x}} = x \pm Z_{\alpha/2}\sigma/\sqrt{n}$ (α = 1-信賴水準)
      - Z= 1.65時 $x \pm 1.65\sigma_{\bar{x}}$ 稱爲 μ 的 90% CI (α = 0.1)
      - Z= 1.96時 $x \pm 1.96\sigma_{\bar{x}}$ 稱爲 μ 的 95% CI ($\mu \pm 1.96$ SE, $x = \mu$)
      - Z= 2.58時 $x \pm 2.58\sigma_{\bar{x}}$ 稱爲 μ 的 99% CI (α = 0.01)
        - standard error = $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ (樣本抽樣平均數的標準差稱標準誤, SE)
        - 標準常態分佈 standardized z score= $x_i - \mu / \sigma$ (母體); $z = x_i - \bar{x}/s$ (樣本)
        - 母體平均數之95% CI 估計 = $\bar{x} \pm Z\, s/\sqrt{n}$
        - μ 之 (1-α) x 100% CI = $\bar{x} \pm Z_{1-\alpha/2}\sigma_{\bar{x}} = \bar{x} \pm Z_{1-\alpha/2}\sigma/\sqrt{n}$
  - 母群體 標準差 σ 未知 (統計量 $t = (\bar{x} - \mu)/ s \div \sqrt{n}$, 隨**df**分布)
    - n<30時， μ 的區間估計= $x \pm t_{(\alpha/2,\, df)}\sigma_{\bar{x}} = x \pm t_{(\alpha/2,\, df)}\, s/\sqrt{n}$
      - 其中 $s = \sqrt{\Sigma(x-\bar{x})^2/n-1} = \sqrt{\Sigma x^2 - (\Sigma x)^2/n-1}$ (Student's t test, Gossett 1876)
    - **n≧30時，t分配近似常態分配** $\mu = \bar{x} \pm Z_{\alpha/2}\sigma_{\bar{x}} = \bar{x} \pm Z_{\alpha/2}\, s/\sqrt{n}$
      - μ 的區間估計，又稱 μ 的 100(1-α)% 信賴區間 **α : significant level**
      - **Central Limit Theory**, $\bar{x}$ bar= μ (不偏估計) 95% CI = $\mu \pm 1.96$ SE
      - 從母體無限次隨機抽樣 (n>30, 大樣本), 樣本平均數的抽樣分布會近於常態分佈

統計分析與檢定 (機率分布, 95% CI, p)

- 常態分佈、標準化常態分佈 (**Z** = x- μ / σ , Z = x_i -平均值 / 標準差) 母群體
  - 中央極限定理 Central limit theorem, n>30 時 樣本平均數之抽樣分佈接近常態分佈
  - 母體 X(μ , σ^2), 樣本**X**分布 **X** bar~ Normal (μ , σ^2 **/ n**), **Z** = **x**- μ **/** σ /$\sqrt{}$n~ N (0, 1), **t= x**- μ **/**s/$\sqrt{}$n (df=n-1)
  - μ 之 95% CI = **x** ± **1.96** σ /$\sqrt{}$n **(以樣本估計母體)** or (1- α)x100% **CI** = **x** ± $Z_{1-\alpha/2}\sigma_{\bar{x}}$ = **x** ± $\mathbf{Z}_{1-\alpha/2}\sigma$ /$\sqrt{}$n
- **t** 分佈
  - 若母體 σ 不知 (以s代替 σ), 或小樣本時: t 分佈 **t** = **x**- μ **/** s/$\sqrt{}$n (df 自由度 = n-1) (SE 標準誤 = s/$\sqrt{}$n)
  - 樣本平均數的標準差稱爲標準誤 SE: standard error is the standard deviation of the mean of the sample
  - μ 之 95% CI = **x** ± $\mathbf{t}_{n-1}$ s/$\sqrt{}$n (t值代替上式中常態分佈之1.96)
    - **SPSS** 分析 比較平均數法 **單一樣本**t檢定(μ) or **兩配對**d/**獨立**樣本t檢定 (**variance equal** or **unequal**)
- **F** 分佈 (變異數分析 ANOVA) (組間變異與組內變異比值的分布**= SSB**/k-1÷**SSW**/n-k =**MSB / MSW**)
- 卡方分佈 $\mathbf{X^2}$ (期望值與觀察值間差異的平方/期望值, $\mathbf{X^2 = (O-E)^2/E}$)
- 相關係數 (**r**) 迴歸係數 (**b**) 線性相關: 兩變項相關強度; 線性迴歸:用x來預測Y值
- 直線迴歸方程式(迴歸分析) **Y**: Linear, logistic, Cox regression
  - $Y=a+b_1x_1+b_2x_2+\ldots\ldots$ (迴歸的目的: 以 x來預測 y值)
- 符號等級中位數檢定 (無母數分析 Non-parametric analysis)
- 信賴區間, 臨界值, 信賴水準~ *p* value 下決策 (accept or reject H_0)

Selecting a Statistical Test

Goal	Type of Data		
	Measurement (from Gaussian Population) 連續變項且爲常態分佈	Rank, Score, or Measurement (from Non-Gaussian Population)	Binomial 二項式變數 (Two Possible Outcomes)
Describe one group	Mean, SD	Median, interquartile range (Q_1-Q_3)	Proportion (%)
Compare one group to a hypothetical value	One-sample t test	Wilcoxon test	Chi-square or Binomial test
Compare two unpaired groups	Two-sample t test (unpaired t test)	Mann-Whitney test/ Wilcoxon rank-sum test	Fisher's test (chi-square for large samples)
Compare two paired groups	Paired t test	Wilcoxon signed-rank test	McNemar's test
Compare three or more unmatched groups	One-way ANOVA	Kruskal-Wallis test	Chi-square test
Quantify association between two variables	Pearson correlation	Spearman correlation	Contingency coefficients
Predict value from another measured variable	Simple linear regression	Nonparametric regression	Simple logistic regression
Predict value from several measured or binomial variables	Multiple linear regression		Multiple logistic regression

SPSS 描述性統計

- **1.** 分析 描述性統計 次數分配表
  - 變項 ➢ 按統計量 勾選項 (Quartile, $D_{10\text{-}90...}$)
- **2.** 分析 描述性統計 **描述性統計量**
- **3.** 分析 描述性統計 預檢資料 統計圖 機率圖附檢定 (K-S test, Shapiro-Wilk test)
- **4.** 分析 描述性統計 交叉表 **(Chi-sq, Fisher exact, McNemar..)**

- **SPSS** 統計圖 **圓餅圖 or 長條圖** (類別變項資料使用)
- **SPSS** 統計圖 **盒形圖** (Median=Q_2, $Box_{lower\text{-}upper}$= Q_1~Q_3, $Outlier_{min\text{-}max}$)
- **SPSS** 統計圖 **散佈圖** Y=依(結果)變數➢ X=獨立變數➢ (連續變項)
  - Scatter plot按兩下..圖表..選項 勾全體式 (可畫出迴歸直線)
- **SPSS**是否常態分布**:** 分析 描述統計 **預檢資料** 統計圖 機率圖附檢定 (K-S test, Shapiro-Wilk test) and/or 圖表 **直方圖** 選變數➢ 勾顯示**常態曲線**

輸入資料、檢查錯誤、極端值

- **Type of data**
  - **Categorical data**
  - **Numerical data** (~ data transformation 轉換資料)
    - **SPSS** 轉換 **重新編碼** 成不同變數 ➢ 舊值與新值
- **Data entry**
  - Formats – ASCII, .txt file, or Microsoft Excel
  - Coding 編碼: handling (coding) missing values
  - Check errors and outliers
    - **Outlier – scatter plot** or **histogram**
      - 做兩次分析看踢除前後結果是否相同, if not, 改用無母數分析
  - Transformation 資料的轉換
    - 對數 (右偏)、平方根、倒數、平方 (左偏)、logistic (乙狀)

SPSS 比較平均數法

- SPSS 分析 比較平均數 單一樣本 t檢定 (student t test)
  - 分析 比較平均數 單一樣本t檢定 輸入檢定值 (μ)
- SPSS 分析 比較平均數 成對樣本 t檢定 (paired t test)
- SPSS 分析 比較平均數 獨立樣本 t檢定 (unpaired two sample t test)
  - 先看**變異數**相等或不相等(看變異數是否相等Levene檢定的p值),再看 **t檢定**的p值
    - 變異數未知 但 相等 用 pooled t test (pooled variance) (p>0.05)
    - 變異數未知 且 不相等 用 separate t test (p<0.05)
    - (變異數已知 用 Z test)
    - **(依 變異數σ已知或未知, σ相等或不相等, 樣本數大小而使用不同公式)**
- SPSS 分析 比較平均數 單因子變異數分析 (one-way ANOVA, ≧3組)
  - One-way analysis of variance (rather than use multi- two sample t tests)
  - F 檢定= 組間平均平方和/組內平均平方和，if p<0.05 再做事後(多重比較)檢定
  - 多重比較 (Post Hoc事後檢定:兩兩比較平均值差異) Tukey, LSD, Bonferroni correction...
- 非常態分佈或樣本數小於25(30): 用無母數統計 (排序法- ranking variable)
  - **SPSS** 看資料**是否常態分布**
    - SPSS 圖表 直方圖 選變數➢ 勾顯示**常態曲線**
    - SPSS分析 無母數檢定 單一樣本 K-S檢定 (Kolmogorov-Smirov) **(常態分布: p>0.05)**
    - 分析 描述性統計 預檢資料 圖形 常態機率圖附檢定 (K-S檢定, Sapiro-Wilk檢定)

A. 單一母群體平均數估計與檢定

- **Null hypothesis** 虛無假設, **Alternative hypothesis** 對立假設
- One-tailed 單尾 (已知優劣), Two-tailed test 雙尾檢定 (不知方向性)
- Type I (α) error, Type II (β) error
  - Type I error 無差異之虛無假說被推翻，但真正的情形是無差異 **(α = 0.05)**
  - Type II error 虛無假說是不正確的，但沒有推翻虛無假說 **(β = 0.2)**
- Significant level: 願意冒犯 type I error 的機率, α =0.05
  - Criteria value 判定值, or 用 observed significant level (O.S.L.) 機率
  - Acceptance region 接受區 or rejection region 棄卻區
- 單一母群體平均數 μ 的檢定 (單一樣本平均數 μ 檢定)
  - σ 已知 H_0: $\mu = \mu_0$, $Z_0 = (x-\mu_0) \div \sigma/\sqrt{n}$ (計算z值，查機率表) (z 檢定)
  - σ 未知，大樣本 ($n \geq 30$) 同上常態分配 $Z_0 = (x-\mu_0) \div \sigma/\sqrt{n}$ (z 檢定)
  - σ 未知，小樣本 ($n<30$) 根據 t分配 $t_0 = (x-\mu_0) \div s/\sqrt{n}$ (t test)
    - 並依單尾(>, <)或雙尾($\neq$)取 $Z_{\alpha\ (單尾)}$ or $Z_{\alpha/2\ (雙尾)}$ 或 $t_{\alpha,\ df\ (單尾)}$ or $t_{\alpha/2,\ df\ (雙尾)}$
  - **SPSS** 分析 比較平均數 單一樣本T檢定 選檢定變數 輸入檢定值 (μ)
- 樣本數 (n, sample size) 大小的決定 (估計樣本數的公式)
  - 誤差相當於信賴區間 $\pm d = \pm Z_{\alpha/2}\ \sigma/\sqrt{n}$ 故 $n = (Z_{\alpha/2}\ \sigma/d)^2$

兩個母群體平均數差的估計與檢定

I. 獨立樣本 (**unpaired** data) or 配對樣本 (**paired** data)

- **I. 獨立樣本 (Unpaired data): x_1-x_2**的抽樣分配，視變異數σ已知或未知, 變異數相等或不相等**(Levene test 檢定),** 及樣本數大小而使用不同公式
  - 1. $\sigma 1 = \sigma 2 = \sigma$ 且σ已知, x_1-x_2分配近似常態, $\mu_{(x1-x2)} = \mu_1 - \mu_2$
    - $\sigma_{(x1-x2)} = \sqrt{\sigma_{x1}^2 + \sigma_{x2}^2} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} = \sqrt{\sigma^2/n_1 + \sigma^2/n_2} = \sigma\sqrt{1/n_1 + 1/n_2}$
      - 1. $\mu 1 - \mu 2$之$(1-\alpha)$x100 % 信賴區間: $(x_1-x_2) \pm \mathbf{Z}_{\alpha/2}\,\sigma\sqrt{1/n_1+1/n_2}$ **(變異數已知用 Z test)**
      - 2. 檢定統計量 $Z_0 = (x_1-x_2) - (\mu_1-\mu_2) / \sigma\sqrt{1/n_1+1/n_2}$
  - **2. $\sigma 1 = \sigma 2 = \sigma$ 但σ未知**, (x_1-x_2) 爲 df= n_1+n_2-2 的 **t**分佈 $\mu_{(x1-x2)} = \mu_1 - \mu_2$
    - $\sigma_{(x1-x2)} = \mathbf{s_c}\,\sigma\sqrt{\mathbf{1/n_1 + 1/n_2}}$ $\quad$ $\mathbf{s_c = s_{pool}}$: common estimate of the SD
      $\mathbf{s_c} = \sqrt{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2/(n_1-1)+(n_2-1)}$ **(變異數未知 但相等用 pooled t test)**
      - 1. $\mu 1 - \mu 2$ 之$(1-\alpha)$x100 % 信賴區間: $(x_1-x_2) \pm \mathbf{t}_{(\alpha/2,\,df)}\,\mathbf{s_c}\sqrt{1/n_1+1/n_2}$
      - 2. 檢定統計量 $t_0 = (x_1-x_2) - (\mu_1-\mu_2) / s_c\sqrt{1/n_1+1/n_2}$ (= Z_0 表 if n_1, n_2 all≧30)
  - **3. σ未知$\sigma 1 \neq \sigma 2$,** x_1-x_2爲**df=** $[s_1^2/n_1 + s_2^2/n_2)]^2 / \{[s_1^2/n_1]^2/(n_1-1)\} + \{[s_2^2/n_2]^2/(n_2-1)\}$**t**分佈
    - $\mu_{(x1-x2)} = \mu_1 - \mu_2$, $\sigma_{(x1-x2)} = \sqrt{\mathbf{s_1^2/n_1 + s_2^2/n_2}}$
      - 1. $\mu 1 - \mu 2$ 之100$(1-\alpha)$% 信賴區間: $(x_1-x_2) \pm \mathbf{t}_{(\alpha/2,\,df)}\sqrt{s_1^2/n_1+s_2^2/n_2}$
      - 2. 檢定統計量 $t_0 = (x_1-x_2) - (\mu_1-\mu_2) / \sqrt{s_1^2/n_1+s_2^2/n_2}$ **(變異數未知且不相等 separate t test)**
  - **SPSS** 分析 比較平均數法 獨立樣本T檢定 檢定變數➢分組變數 ➢按定義組別
    - 註：檢定變數(test variable=依變數=數值型), 分組變數 (group variable=自變數, 要比較的組別)
    - ➢ **變異數相等的 Levene檢定**F檢定 p**>**0.05讀變異數相等的t值; **<0.05**讀假設變異數不相等的t值
    - **t分布 t=** $(\mathbf{x_1-x_2}) / \mathbf{s_{x_1-x_2}}$, 分散度 $\mathbf{s_{x_1-x_2}} = \sqrt{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2/(n_1+n_2-2)}\,\sqrt{1/n_1+1/n_2}$

兩個母群體平均數差的估計與檢定

II. 配對樣本 (paired data) (vs. 獨立樣本)

- **II. 配對樣本 (paired data): e.g. 前後測, 配對研究**
  - 成對樣本 t 檢定 (difference test): 看 x_1-x_2 差異(D)的平均X_D
  - **t = X_D/ S_D÷√n S_D= √[ΣD²-(ΣD)²/n]/ n-1** (n爲配對數 非總人數, df=n-1)
    - c.f. 獨立樣本 (unpaired data): 看 **x_1-x_2** 的差異
    - **SPSS** 分析 比較平均數法 **成對樣本t檢定** 配對變數➢➢

vs. Non-parametric test

- 成對樣本之無母數分析: **McNemar檢定(二項資料) / Wilcoxon符號等級t**
  - **McNemar 檢定(二項資料)** X_M^2 = $(n_1-n_2-1)^2/(n_1+n_2)$ (只看有改變的 b,c格卡方檢定)
    - **SPSS** 分析 無母數分析 二個相關樣本檢定 勾**McNemar** 配對變數➢➢
  - **Sign test 符號檢定** 利用二項分布之Z檢定來比較任一改變的比例與1/2(0.5)是否有差異
    - **SPSS** 分析 無母數分析 二個相關樣本檢定 勾**符號檢定** 配對變數➢➢
  - **Wilcoxon signed rank 符號等級檢定 (次序資料,** 排除前後分數沒改變的**)**
    - **SPSS** 分析 無母數分析 相關樣本檢定 勾**Wilcoxon** 配對變數➢➢
    - $z = (T-\mu_T)/\sigma_T$ $\mu_T = n(n+1)/4$ $\sigma_T = \sqrt{n(n+1)(2n+1)/24}$

SPSS 變異數分析 (ANOVA)
：比較三組或以上的平均數

F= SSB/k-1÷SSW/n-k(組間變異與組內變異比值)的分布= **MSB / MSE** 檢定 $S_b^2/S_w^2 = 1$

- **One-way** ANOVA 單因子變異數分析 (常用於實驗研究設計的統計)
  - 一個連續型依變數和一個類別的自變數 (= 因子 factor A_ level 層次 a_1 a_2 a_3)
    - **SPSS** 分析 **比較平均數法** 單因子變異數分析➢ 應變數(連續) 因子(類別自變數)
      - 選項: 統計量勾描述統計及**變異數同質性** (ANO前提:常態,獨立,變異數同質性. 若違反用K-W)
      - 勾 Post Hoc multiple pairwise comparison 事後多重比較
      - Tukey HSD人數相同, Scheffe人數不同保守, Bonferroni correction (non-pairwise, **planned**), LSD(寬鬆)
- **GLM**: **Two**-way ANOVA 雙因子變異數分析 (Design and Interaction)
  - 平行或非平行設計; 不同因子(factor,自變項)內的不同層次(level)是否有交互作用**(interaction)**
  - 一個連續依變數和兩個類別自變數 **(*看Main effect P>0.05 或 Interaction p<0.05)**
    - **SPSS** 分析 **一般線性模式** (general linear model, GLM) **單變量** (指應變數) ➢ 依變數 固定因子(=自變數) 亂數因子 共變量(共變數) 加權最小平方法之權數
      - 點選項: 勾敘述統計, 效果項大小估計effect size (eta 平方), 觀察的檢定能力 (power)
        - 看main effect (主因素AB), interaction (A*B交互作用剖面圖) **p**有沒有明顯差異
- **GLM**: **Three**-way ANOVA 三因子變異數分析 (..需要經驗天份)
  - 一個連續的依變數和三個類別的自變數 (main effect? **interactions**?)
    - **SPSS** 分析 **一般線性模式** (general linear model, GLM) **單變量** (univariate) …
      - **共變量**: 有一個與依變數有實質相關的數值型自變數,在實驗開始前,用來調整樣本間存在的差異(排除共變數的影響) (ANOVA變異數分析如果有共變數簡稱爲**ANCOVA**)

比較多個獨立母體平均數有沒有差異:

ANOVA (Analysis of Variance) 變異數分析(F)

TSS=SSB+SSW ~ 將總平方和 (似變異數) 分爲組間及組內平方和，做 F檢定

- TSS (total sum of square, 總平方和) **(先看變異數同質性檢定 Levene test)**
  - TSS $= \sum (x_i - \mathbf{x})^2 = \sum x_i^2 - n\mathbf{x}^2$ df= n-1 s^2 = TSS/n-1
- **SSB** (sum of square between, 組間平方和= explained variation)
  - SSB $= \sum n_s(\mathbf{x}_s - \mathbf{x})^2$ factor(自變項,組別) n_s 是每組的個數 df= k-1 (k=組數, level)
- **SSW** (sum of square within, 組內=殘差平方和 SSE= unexplained)
  - SSW $= \sum (x_i - \mathbf{x}_s)^2$ = TSS-SSB 隨機變異 $\mathbf{x}_s$ 是各組的平均值 df= n-k
- 用**F檢定**: $F_{樣本}$= **SSB**/k-1÷**SSW**/n-k **(組間變異與組內變異比值)**=**MSB/MSE**
  - **F** = 組間變異/自由度(k-1) ÷ 組內變異/自由度(n-k) = **MSB/MSE**
    (=mean square between迴歸平均平方和/mean square error殘差平均平方和)
    - 虛無假說爲母體平均數皆相等，若拒絕假說並不知哪些組別有差異，故需再進行事後 **(Post Hoc)**的多重比較，建議採用較保守的 Scheffe事後檢定
    - **SPSS** 比較平均數 單因子變異數分析 依變數 因子(自變數)➢ Post Hoc
      - **事後多重比較:** Tukey HSD, Scheffe, Bonferroni, SNK, Dunnett ... methods
      - 雙因子變異數分析 (**two** way ANOVA vs. one way ANOVA): 多了**交互作用** 平行設計 balanced design
      - **前提 Assumptions**: 常態 變異數相同 觀察值獨立 (若違反假設 改用無母數**K-W**分析)
      - **SPSS**: Source of variation (**B**etween=explained, **E**rror=Within=**un**explained, **T**otal) SS df MS **F**

變異數分析之事後檢定分析

3組或3組以上類別與數值相關的顯著性

Tukey HSD人數相同, Scheffe保守人數不同, Bonferroni (non-pairwise, **planned)**, LSD(較寬鬆)

- 事後多重比較: **Scheffe, Bonferroni, Tukey**, SNK, LSD.. (variance equal)
  - **Tukey**: 樣本n大小相等 (Tukey-Kramer procedure)
    - 兩組 Pairwise, **or balance design (**= equals ample size 各組的人數相同時**)**
  - **Scheffe**: 不論樣本大小是否相同 各組的人數不同 (paired or unpaired)
    - **the least powerful** 保守(不易找出有差異), **unplanned** **(p<0.05**才做多重比較 **F)**
  - **Bonferroni** correction: 類似 **LSD (correction:** α **/**檢定次數 **t)**
    - **non-pairwise and planned (**事先決定比較方式組合, 最有利**)** ➢ unplanned use Scheffe
  - LSD: least significant difference (寬鬆) (所有配對多重t檢定, multi-t tests increase type I error)
  - Student Newman-Keuls (SNK) ~ 類似Tukey的比較法
  - Dunnett (只與其中1組比較), Duncan's... (交互作用Interaction, 主要效果, fixed effect models)
  - **Factor**..**l**evel..**T**reatment..**D**esign: random, block (paired): complete, incomplete (missing data)
- **F = SSB**/k-1÷**SSW**/n-k = **MSB / MSE =** $S_1^2/S_2^2 = S_B^2 / S_W^2$ (離差平方和/df)
  - $S_t^2 = \sum (x_i - \mathbf{x})^2/n-1$ $S_B^2 = \sum n_i(\mathbf{x}_i - \mathbf{x}_t)^2/k-1$ $S_W^2 = \sum (x_{1i} - \mathbf{x}_1)^2 + (x_{2i} - \mathbf{x}_2)^2 + (x_{3i} - \mathbf{x}_3)^2 + ../n-k$
  - H_o: 所有組別平均值相等 $\mu_1 = \mu_2 = \mu_3 .. = \mu_n$ H_1: 至少兩組平均值不相等 (are not all equal)
  - Assumptions: random samples, independent, normally distributed, variance is the same
    - **SPSS 先看**檢定變異數同質性假設: Levene test 的 p 值 (>0.05 爲同質性)
    - Use nonparametric K-W test if against the above assumptions (相對的無母數分析法: K-W test)
    - X^2, block design or repeated measures ANOVA, **Kruskal-Wallis** (n<10 each), (weighted ANOVA)

SPSS ANOVA (常用於實驗研究設計的統計)

(Outcome依變數 爲連續數值變項) (自變數=因子爲類別型資料, 因子內有不同層次)

- **One way ANOVA** (比較平均數) **SPSS** 比較平均數 單因子變異數分析 依變數 因子(自變數)➢ Post Hoc
  - 先看檢定變異數同質性假設: Levene test, if p<0.05 alternative hypothesis➢ 改用相對的無母數分析法 K-W test
- **General Linear Model (GLM) 一般線性模式** (單變量多因子或多變量…分析)
  - Univariate **單變量** (≥**2 way ANOVA**雙自變項因子factors_levels: 先看**交互作用*p*** 平均數剖面圖 主效應)
    - 獨立樣本 (complete randomized design): 隨機分配到實驗及對照組
    - 相依樣本 (randomized block design受試者間設計: (Different study designs)
      - Repeated measures
      - Paired method **(balanced design vs. unbalanced design= missing data)**
      - Incomplete randomized block design (missing values: 細格人數不相等)
    - 混合以上兩種設計 (剖面圖線條平行爲no interaction, **p>0.05**➢直接看主效應main effect)
    - .. Nested design (巢狀設計=階層實驗=分隔設計), 拉丁方格實驗設計, 重複量數拉丁方格 (no interaction)
  - . Multivariate (依變數)多變量: MANOVA, MANCOVA (不能有線性相依的≥2個依變數的分析)
  - .. Repeated measures ANOVA 重複量數 (... 變異成分)
  - **共變量**: 有一個與依變數有實質相關的變數，在實驗開始前，用來調整樣本間存在的差異 (排除共變數的影響)。 (ps. 變異數分析如果有共變數通常簡稱爲ANCOVA)
  - **SPSS** 分析 **一般線性模式** (GLM) **單(依)變量** or多(依)變量➢ 依變數 固定因子(=自變數, one or two way= one or two factors) (亂數因子) **共變量**(共變數) (加權最小平方法之權數)
    - 圖形選項:敘述統計,效果項大小effect size(eta平方,可解釋多少變異),觀察的檢定能力 (power)
      - 看 main effect(主因素間), interaction(交互作用**A*B**交叉分析剖面圖)F有沒有明顯差異**p**

變異數分析之事後檢定分析

3組或3組以上類別與數值相關的顯著性

Tukey HSD人數相同, Scheffe保守人數不同, Bonferroni (non-pairwise, **planned)**, LSD(較寬鬆)

- 事後多重比較: **Scheffe, Bonferroni, Tukey**, SNK, LSD.. (var. equal)
  - **Tukey**: 樣本n大小相等 (Tukey-Kramer procedure)
    - 兩組 Pairwise, **or balance design (**= equals ample size 各組的人數相同時**)**
  - **Scheffe**: 不論樣本大小是否相同 各組的人數不同 (paired or unpaired)
    - **the least powerful** 保守(不易找出有差異), **unplanned (p<0.05**才做多重比較 **F)**
  - **Bonferroni** correction: 類似 **LSD (correction:** α **/**檢定次數 **t)**
    - **non-pairwise and planned** (事先決定比較方式組合, 最有利) ➢ unplanned use Scheffe
  - LSD: least significant difference (寬鬆) (所有配對多重t檢定, multi-t tests increase type I error)
  - Student Newman-Keuls (SNK) ~ 類似Tukey的比較法
  - Dunnett (只與其中1組比較), Duncan's... (交互作用Interaction 主要效果 fixed effect models)
  - **Factor**..**l**evel..**T**reatment..**D**esign: random, block (paired): complete, incomplete (missing data)
- **F = SSB**/k-1÷**SSW**/n-k = **MSB / MSE =** $S_1^2/S_2^2 = S_B^2 / S_W^2$ (離差平方和/df)
  - $S_t^2 = \sum (x_i - \mathbf{x})^2/n-1$ $S_B^2 = \sum n_i(\mathbf{x}_i - \mathbf{x}_t)^2/k-1$ $S_W^2 = \sum (x_{1i} - \mathbf{x}_1)^2 + (x_{2i} - \mathbf{x}_2)^2 + (x_{3i} - \mathbf{x}_3)^2 + ../n-k$
  - H_o: 所有組別平均值相等 $\mu_1 = \mu_2 = \mu_3 .. = \mu_n$ H_1: 至少兩組平均值不相等 (are not all equal)
  - Assumptions: random samples, independent, normally distributed, variance is the same
    - 先看檢定變異數同質性假設: Levene test 的 p 值 (>0.05 爲同質性)
    - Use nonparametric K-W test if against the above assumptions (相對的無母數分析法: K-W test)
    - X^2, block design or repeated measures ANOVA, **Kruskal-Wallis** (n<10 each), (weighted ANOVA)

有母數統計 對應的 無母數分析 (rank)
Nonparametric analysis

- One sample t test (樣本數小用無母數統計比較有利, 用有母數通常不顯著)
  - Wilcoxon test or signed test (兩項式分布檢定母體比率是否等於1/2)
- Paired t test (兩個相關樣本) (paired data)
  - Wilcoxon signed-rank test (Binary categorical paired data: McNemar's X^2)
- Two sample t test (兩個獨立樣本) (unpaired data)
  - Wilcoxon rank-sum test (排序等級總合雙樣本z檢定)
    - **SPSS** 分析 無母數分析 兩個獨立樣本檢定 檢定變數➢ 分組變數➢ 按定義組別
  - Mann-Whitney U test (分爲 小樣本**U**:all<10和大樣本z: any one >10 兩種)
- One-way ANOVA (≧3組, K個獨立樣本檢定)
  - Kruskal-Wallis test (是卡方,非z檢定) (c.f.≧3 paired data use Friedman ANOVA)
- Pearson correlation coefficient (r)
  - Spearman correlation coefficient (r_s)

SPSS 無母數統計

(When: 小樣本 n<30 (or 10 in each), Outlier, Skewness)

- Mann-Whitney U/ Wilcoxon rank-sum test 等級和檢定法, Signed test
  - 分析 無母數檢定 兩個獨立樣本檢定 ➢ 定義組別 (unpaired, **independent** data)
- Wilcoxon matched-pairs signed ranks test 符號排序檢定 (**paired** data)
  - 分析 無母數檢定 兩個相關樣本檢定 符號檢定 (or Wilcoxon signed rank test)
- Binomial test 兩項式檢定法 (判斷是否兩項分配 p=0.5)
  - 分析 無母數檢定 兩項式檢定➢ 檢定比例 0.5
- Runs test 連檢定法 (只有兩種結果的兩項分配是否隨機分配)
  - 分析 無母數檢定 連檢定➢ 自定 2 (不選中位數)
- Kolmogorov-Smirnov one-sample test (常態分配) (是適合度檢定)
  - 分析 無母數檢定 單一樣本 K-S檢定 ➢ 檢定分配 勾常態分配 **(Nor. D: p>0.05)**
- One-sample chi-square test (各組別人數是否有明顯差異 e=n/k)
  - 分析 無母數檢定 卡方分配
- Kruskal-Wallis test (K-sample median test) ~ k個獨立樣本中位數檢定
  - 分析 無母數檢定 k個獨立樣本檢定 定義組別 勾中位數 (or Kruskal-Wallis test)
- Friedman one-way ANOVA (用等級平均數做變異數分析) vs. repeated measure ANOVA
  - 分析 無母數檢定 多個相關樣本檢定 **Friedman** (multiple treatments in the same individuals)

類別資料分析 (Categorical data)

- **單一比例 (a single proportion)**
  - Rationale: 遵守二項式分布，但若 np與 n(1-p)都大於5的話，幾乎等於常態分布, 估計平均值 **p**= x/n, 估計標準差爲 $\sigma_p = \sqrt{p(1-p)/n}$
  - π 之 **95% CI = p ± 1.96$\sqrt{p(1-p)/n}$**
    - 若要保證誤差範圍爲 ± d%, $1.96\sqrt{p(1-p)/n} \le \mathbf{1.96\sqrt{1/4n}} \le$ d%
    - **n ≥ 4 (1.96/d%)²** (**n:** sample size calculation) vs. 數值型變數 $n = (Z_{\alpha/2}\ \sigma/d)^2$
- **兩個比例值 (two proportions)**
  - 卡方檢定 (2x2 table)
    - 不相關兩組 (independent groups data): $X^2 = [\Sigma\ |O-E| - 1/2]^2 / E$, df=1 (½ :校正)
      - 95% CI = $(p_1-p_2) \pm 1.96\sqrt{p_1q_1/n_1 + p_2q_2/n_2}$ (比例數差值的檢定) (2x2 table: use Yate's correction)
    - 相關組別 (related, paired): McNemar's test $X^2 = (|b-c| - 1)^2/(b+c)$ (只 bc格改變)
    - 若有任一格期望值<5(違背假設): Fisher's exact test (不須符合卡方分布)
  - 獨立: 用 X^2; **相關**(兩次測量)用McNemar's; 違背假設(**任一格期望值<5**)用Fisher's exact test
- 超過兩個類別 **(more than two categories)** (row x column)
  - $X^2 = \Sigma (O-E)^2 / E$, df= (r-1) (c-1)　　[r x c 列聯表, df >1 不必做連續性校正]
  - $X^2_{(r-1)(c-1)} = \Sigma_r \Sigma_c (O_{ij} - E_{ij})^2 / E_{ij}$　　$E_{ij} = n_i \times n_j / N$　df= (r-1) (c-1)
    - 當觀察值(O)與期望值(E)差異(X^2值)很大，則拒絕H_0
- 相關強度(風險估計值): Odds ratio = ad / bc (勝算比), Relative risk (相對風險)

SPSS 類別資料分析

- 卡方檢定
  - Pearson卡方
  - 連續性校正 **(Yate's** continuity correction**)**: 用於 **2x2 table** 需校正
  - 概似比: 風險
  - Fisher's exact test 費氏精確檢定: 用於 **small sample size,** 格子中有 **20% np<5**
  - McNemar's test (two paired data, off-diagonal cell b, c ≥ 20), McNemar exact test (small< 20)
  - 線性對線性的關連
  - **SPSS** 分析 描述性統計 預檢資料
  - **SPSS** 分析 描述性統計 交叉表 統計量(卡方, 風險, McNemar) 格(觀察值, 期望, 橫列%)
    - 轉換 **重新編碼** 成不同變數 變更 舊值與新值 新增
    - 資料 加權
    - McNemar's test: paired data (vs. unpaired: Chi-sq, or Fisher's exact~ for small sample)
    - $X^2 = \Sigma (f_{obs}-f_{exp})^2/f_{exp} = \Sigma (O-E)^2/E$, df= (r-1)(c-1) (觀察值與期望值的差稱爲殘差 residual, 殘差越大越不合理~ reject null hypothesis)
- **Logistic regression (Binary logistic regression)**
  - **SPSS** 分析 迴歸方法 二元Logistic 依變數共變量 類別 選項**Exp(B)**95%CI 方法

卡方檢定 (Chi-square test)

- **1. 兩類別變項的獨立性檢定 (test of independence)**
  - 檢定兩變項間有無相關 (獨立=不相關) (調查之後才知道 margin value: 不固定)
  - $X^2 = \Sigma (O-E)^2 / E$ df= (r-1) (c-1) = 1 , $\alpha = 0.05$時 $X^2 = 3.84$
  - $X^2 = n(ad-bc)^2/(a+c)(b+d)(a+b)(c+d)$, *if Yate's continuity correction: -0.5n*
- **2. 同質性檢定 (test of homogeneity)** Margin fixed: margin固定
  - 不同特質在不同組別之分佈是否相同 (分組變項之各組的百分比是否一樣)
  - H_0: $\pi_1 = \pi_2 = \pi_0$ H_A: $\pi_1 \neq \pi_2$ $p_0 = a+b/a+b+c+d$
  - $Z = (p_1-p_2) / \sqrt{p_0 (1-p_0) (1/n_1 + 1/n_2)}$ (雙樣本Z檢定: SPSS不提供) or $X^2 = \Sigma (O-E)^2 / E$
- **3. 兩個比例有沒有差異的顯著性檢定**
  - **A. Unpaired data**:
    - $X^2 = [\Sigma \mid O - E \mid - 1/2]^2 / E$, df=1 (½ :校正) (2x2 table use Yate's correction)
    - 95% CI = $(p_1-p_2) \pm 1.96\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$ (兩個比例數差別的檢定)
  - **B. Paired data (前後測): McNemar's test** $X^2 = (\mid b-c \mid - 1)^2 / b+c$ (僅bc格改變)
  - **C. 期望次數太少** (E<2 or 20% cell<5): 先合併格子或用 **Fisher's exact test**

卡方檢定 (Chi-square test)

- 單一樣本卡方檢定稱爲適合度檢定 **(Test of goodness-of-fit 分配)**
  - $X^2 = \Sigma (fobs-fexp)^2/fe$, df=k-1 (k組次序資料..>二元時) 各組**分配**情形(機率)是否一致
    - **SPSS** 分析 **無母數分析** 卡方檢定項目(觀察之分配跟一已知的機率分佈做比較 e.g. die)
- 雙及多樣本卡方檢定稱爲獨立性檢定 **(test of independence)**
  - 分析兩類別變數的獨立性 (不相關) (cross table, 卡方分佈)~ margin free
  - 比較觀察與期望個數(或比例)有沒有顯著差別 (無方向性)
    - $X^2 = \Sigma (\mathbf{f}obs-\mathbf{f}exp)^2/\mathbf{f}exp = \Sigma (O-E)^2/E$, df= (r-1)(c-1)(觀察值與期望值的差稱爲殘差 residual, 殘差越大越不合理) e.g. $X^2 = 3.841$ ($\alpha = 0.05$, $df = 1$)
    - **SPSS** 分析 描述性統計 **交叉表** 列➢ 行➢ 按**統計量** 勾卡方統計量 (在交叉表對話方塊 點**統計量**旁之儲存**格**顯示按鈕 勾**直行**百分比及**期望個數**有助於列聯表解讀)
      - 小樣本看註腳: 每一格子期望數<5時: 用 Fisher exact test or 先類別合併後再統計
        - *2x2 table use Yate's correction* $Xc^2 = \Sigma (| fo-fe | - 0.5)^2 / fe$
      - 大樣本的問題: 增加統計結果顯著的機會，但不見得具有實質意義。(用更細的類別, 看相對比例不看卡方值, 看相關指標gamma…)
- 兩個比例差異之假說檢定
  - 雙樣本z檢定是卡方檢定的特例，且其公式複雜，故仍建議用卡方檢定，又SPSS也未提供雙樣本z檢定計算。 $z = (P_1-P_2)/\sigma_{P1-P2}$

$$\sigma_{P1-P2} = \sqrt{P_u Q_u} \sqrt{(n_1+n_2)/n_1 n_2} \qquad P_u = n_1P_1+n_2P_2 / n_1+n_2 \text{ (加權)}$$

卡方檢定 (Chi-square test)

- **X^2 test of independence** df= (r-1)(c-1) (preliminary test) **(margin free)**
  - 兩個變項是否獨立: P (A and B) = P(A) x P(B) = Expected $X^2 = \sum (O-E)^2/E$
- **X^2 test of homogeneity** **(margin fixed, 與上述margin free 調查方法不同)**
  - 相同變項內不同組別間之同質性 (分組變項之**各組**的**百分比**是否一樣)
- **X^2 test for trend** (Mantel-Haenszel): 2 rows, many columns 對 columns做趨勢檢定
- 兩個比例有沒有差異之顯著性檢定
  - 治療組及對照組治療成功比例 (%)有沒有差異
- **McNemar's X^2 test** $X^2 = (b- c)^2 / b+c$ (repeated measures 只有算有改變 bc 格)
  - 比較相關(配對樣本)比例的麥氏檢定: 同一人前後測量血壓值
- **Fisher's exact test: small sample**
- **X^2 test of goodness-of-fit 適合度:** 樣本觀察次數與**理論分配**適合程度
- 相關強度的測量: **Relative Risk, Odds Ratio (X_{MH}^2)**
  - **Crammer's contingency coefficient …**
- ps.卡方分布是由常態分布轉換而來: df=1時 X_1^2=3.84=1.96^2= Z^2 =(x- μ / σ)^2 常態分布標準化, Z~ N (μ =0, σ^2 =1) 將各個Z值加以平方即爲卡方值(形成分布), Xn^2 分配的平均數爲n,眾數n-2,標準差$\sqrt{}$2n, n>30近似常態分布(漸進性顯著p)

研究分析重點

橫斷研究 (同一時間) →	世代研究	
	Exposed	Not exposed
Disease	a	b
No disease	c	d

病例對照

a/ a+b

c/ c+d

a/ a+c　　b/ b+d

橫斷研究: 暴露和疾病是否有關 (Margin free) ~ X^2 test

世代研究: 暴露組和非暴露組的發病率是否有差異

RR (Margin fixed) ~ two proportion t test

Risk ratio = Relative risk(intensity), Risk difference% = Attributable risk (intervention)

病例對照研究: 病例組和非病例組的暴露率是否有差異

OR (Margin fixed) ~ X_{MH}^2 test

類別資料分析 (Odds Ratio, Relative Risk)

- Breslow-Day Method **(Test for homogeneity)**
  - K個 2x2 tables, 其中每個之OR是否相同
    - H_0: $OR_1=OR_2=\ldots=OR_g$ $\quad X^2=\sum w_i(y_i-\mathbf{Y}) \sim X^2_{g-1}$
    - $OR_{hat} = a_id_i/b_ic_i$ $\quad y_i = \ln OR_{hat} = \ln a_id_i/b_ic_i$ (2x2 table)
    - $\mathbf{Y} = \sum w_iy_i/\sum w_i$ $\quad w_i = 1/(1/a_i+1/b_i+1/c_i+1/d_i)$
      - **If OR homogeneity, using common OR**
      - **If OR non-homogeneity, individual discussion**
- Mantel-Haenszel X^2 Method **(Test of association)**
  - H_0: test common **OR** equal to **1**
    - $X^2 = (\sum a_i - \sum m_i)^2/ \sum \sigma_i^2$
    - $m_i = M_{1i}N_{1i}/T_i$ $\quad \sigma_i^2 = M_{1i}M_{2i}N_{1i}N_{2i}/T_i^2(T_i-1)$ (2x2 table)
    - **If we reject null hypothesis, imply the exposure and disease have association and the common OR is not equal to 1.**
- 名義資料 的關連性指標 (association)
  - **e.g. Cramer's V contingency coefficient** $=\sqrt{x^2/n(K-1)}$ K:行列中較小者
  - SPSS 分析 描述統計 **交叉表** ➢ ➢ 按**統計量**再勾選**名義的**選項
- Berkson's fallacy, Simpson paradox ~ confounder (third factor)

Multiple linear regression

相關分析：決定兩個變項的**線性**關係之**強度及方向** (相關係數**r** = -1 ~ o ~ +1)

迴歸分析：求一方程式來描述兩個變項的關係及以x值來**預測y** **(斜率slop** β**)**

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

Variables: 變項、變數

y :

Dependent variable

Predicted

Response variable

Outcome variable

x_n :

Independent variables

Predictor variables

Explanatory variables

Covariables (in ANCOVA)

Factor (in ANOVA)

- **X**= x_o $\quad$ $Y_1 = a + b_1 x_o$
- **X**= $x_o + 1$ $\quad$ $Y_2 = a + b_1(x_o + 1)$
- 兩式相減 $\quad$ $Y_2 - Y_1 = b_1$

- **slop b (** β **): 當 X每改變1單位時，Y 的改變量爲 b**

簡單線性迴歸 Simple linear regression (相關與預測)

- $Y_{hat} = a+bx+e$ (simple linear regression, best fitted line)
  - **a** = intercept (y截距, **x**=0), b= slope (斜率), e= error (殘差) = $\mathbf{Y_i}$-$\mathbf{Y_{hat}}$
    - 用 least square method (SSE sum of square of error= $\Sigma(Y_i-\mathbf{Y}_{hat})^2$ 最小) 劃迴歸直線
    - **Scatter plot** 先看是否有爲線性相關, r, 迴歸直線 regression line (best fitted line)
  - **b** (slop): 代表 **x每增加1單位時, y會增加多少的平均值(b)** (做預測: prediction)
    - $\mathbf{X}=x_o$ $Y_1 = a+b_1x_o$
    - $\mathbf{X}=x_o+\mathbf{1}$ $Y_2 = a+b_1(x_o+\mathbf{1})$
    - 兩式相減 $Y_2 - Y_1 = \mathbf{b_1}$
  - $b = \Sigma(X-\mathbf{X})(Y-\mathbf{Y}) / \Sigma(X-\mathbf{X})^2$ (XY的交叉平方和 除以 X的總平方和) (least square method)
  - ANOVA 部分: **F檢定**看斜率 b是否爲零(水平, null, XY不相關), **F= MSR/MSE**
    - SST (總平方和) = $\Sigma(Y-\mathbf{Y})^2$, df= n-1 **Y=$\mathbf{Y_{bar}}$** SST = SSE + SSR
    - SSE (殘差平方和) = $\Sigma(Y-\mathbf{Y_{hat}})^2$, df= n-2 **MSE (殘差)** = $SSE/df_{SSE} = \Sigma(Y-\mathbf{Y_{hat}})^2/ n-2$
    - SSR (迴歸平方和) = $\Sigma(\mathbf{Y_{hat}}-\mathbf{Y})^2$, df= 1 **MSR** (**迴歸**平均平方和) = SSR/df_{SSR}
    - **R^2** (**判定**係數 0-1) = **SSR/SST** (Y的總平方和或變異量, 有多少可被直線迴歸模式所**解釋**的比例)
    - **Adjusted** R^2 (高: best fit model) = 1- $(1-R^2)$ n-1/n-p (p=迴歸係數的個數, SLR p=2)
    - 迴歸係數的**t**檢定(b=0): a, b ($F_{1, n-2} = t^2_{n-2}$), **t**= b_1-0/S_{b1}, df=n-2, $S_{b1} = \sqrt{MSE/\Sigma(X-\mathbf{X})^2}$
    - Y預測值的預測區間= $Y_{hat} \pm t_{n-2} S_{Y/X=x}$ $S_{Y/X=x} = \sqrt{MSE\ [(1+ 1/n+ (X-\mathbf{X})^2/\Sigma(X-\mathbf{X})^2]}$
- **X**與**Y**是否爲線性關係➢ 先做直線迴歸方程式, 由X預測Y➢ 再做假設前提的迴歸診斷

SPSS 相關係數與迴歸分析

數值變項 vs. 數值變項

- 相關分析 (Pearson correlation coefficient, r, product-moment r)
- 迴歸分析 (Regression analysis)

- 以**t**檢定來檢測**相關係數** H_o: rho值 = 0 (水平, 沒有線性關係), H_1: rho (ρ_{xy})≠0
  - **r** = $\Sigma(x-\bar{x})(y-\bar{y}) / \sqrt{\Sigma(x-\bar{x})^2 \Sigma(y-\bar{y})^2}$ = (負相關) **-1~+1** (正相關) 之間, **0**: 無相關
  - **t = (r-r_u) / s_c** **SE = s_c** = $\sqrt{(1-r^2)/ n-2}$, df= n-2
  - 檢查線性關係是否符合: 看(矩陣)散佈圖 or 執行線性迴歸
- **SPSS** 分析 相關 雙變數 ➢ ➢ 勾 Pearson or Spearman相關係數

- **線性迴歸**
  - **SPSS** 分析 迴歸 線性 依變數➢ 自變數(理論或經驗)➢ 方法 stepwise regression
    - 相關係數 (r, R) 決定係數 (r^2, **R^2**: amount of variance explained- %)(殘差: unexplained)
      - R^2 (改變): 所有自變數可以解釋依變數多少的變異量，用來決定一個模式 (Model) 的好壞
    - 迴歸係數 (SPSS係數中 B之估計值, slop b)
      - 每一個自變數的單位都不同，因此很難以迴歸係數的大小判定其重要性，用標準化係數 (Beta分配) 改善單位不同的影響。**(結果看Std coefficients Beta值大小，正負)**
    - 偏相關 (partial correlation)
    - 殘差 e更小 (複迴歸 multiple regression前提假設: 自變數間獨立, 沒有共線性 collinearity)
    - 檢定複迴歸模式的顯著性 可看標題**變異數分析**的 **F檢定**

線性迴歸的診斷 假設前提 (LINE) 的迴歸診斷

得到迴歸結果後，應做殘差及影響力分析，以對所得結果更具信心

- **Assumption**: 線性關係 觀察值獨立 常態分佈 變異數相等 **(LINE)**
  - **L**inearity, **I**ndependent, **N**ormal distribution, **E**qual variance **(LINE)**
    - 任一固定x, y均是常態分佈，y之變異數不會隨不同x值改變
- 殘差分析: 利用殘差做分析 (Residual analysis) ~ **L, E**
  - E (殘差) = $\mathbf{Y_i - Y_{hat}}$ $\mathbf{E_{bar} = \Sigma E/n = 0}$ $S_E^2 = \sigma^2_{y|x}$ (MSE爲$\sigma^2_{y|x}$的不偏估計, s=√MSE)
    - 將殘差標準化(standardized residuals)使得單位相同 $\mathbf{Z}= E/S_E \sim N(0, 1)$
    - **直方圖:**用殘差 **E**或標準化殘差 **Z±3**畫**直方圖**，若呈鐘型表示**常態**分佈 **~ N**
    - **殘差圖** (雙向度圖形): 縱軸放標準化殘差Z, 橫軸放X或Y_{hat}, no pattern圖形**亂**
    - **常態機率圖** (normal probability plot)
- 影響力分析 **(influence, outliers)**
  - Hat matrix (矩陣)
  - Centered leverage (離中心槓桿量數，SPSS稱影響量數) (>3/n 要注意)
  - Studentized residual, studentized deleted residual (> ±3 要注意)
  - Df適合度 (difference between the fitted value Y_{hat}), 標準化Df適合度 (>2)
  - Cook's distance (**>1**, index influence plot 影響力索引圖): outlier
  - 當少數觀察點 (outlier)的影響力非常大，先**偵錯**再決定是否要**刪除**此點
  - 分析 迴歸方法 線性➢統計圖$_{殘差分析}$ 儲存$_{影響力分析}$，統計圖 散佈圖➢簡單 定義$_{影響力索引圖}$= Cook's D

SPSS 相關與迴歸分析

線性相關: 測兩變項的相關強度; 線性迴歸: 以 x來預測 y值

- 分析 **相關** 雙變數 ➢ 勾 Pearson or Spearman 相關係數
  - Report: 用 scatter plot (散佈圖) or 文字敘述 or 相關矩陣 (correlation matrix)
- 分析 **迴歸**方法 **線性** 依變數 自變數➢方法 stepwise /forward /backward
  - (變異數分析 F檢定**slop b**, if p<0.05 then reject **b=0**, i.e. b≠0 有相關)
- **Linear regression: slop b (β)**
  - 線性迴歸之假設前提(LINE)的**迴歸診斷**
    - 得到迴歸結果後，應做殘差及影響力分析，以對所得結果更具信心。
    - Residual analysis 殘差分析: 直方圖及殘差圖
    - Influence 影響力分析 (check outlier):
      - **SPSS 分析** 迴歸方法 **線性** ➢ 統計圖殘差分析 儲存影響力分析
      - **SPSS** 統計圖 散佈圖 ➢ 簡單 定義影響力索引圖= Cook's Distance
- **Logistic regression (Binary二元類別資料): Odds Ratio= OR= Exp(B)**
  - **SPSS 分析** 迴歸方法 二元Logistic 依變量 共變量 類別 選項Exp(B)95%CI 方法
- **Cox regression**, Survival analysis (time to event) **: Hazard Ratio (RR)**
  - **分析** 存活分析 KM統計 時間 狀態(定義事件發生與否單一數值:1) 因子 比較因子方法 選項圖
  - **分析** 存活分析 Cox迴歸 共變量risk 方法 類別 參考類別 統計圖 選項95%C.I.

Binary Logistic Regression
羅吉斯回歸分析

- **Logistic regression** (依變數是二元變數時)
  - **Binary** dependent variable時，使用 logistic regression
    - 某人爲某種預後的機率 p, 將 p做 logistic transformation (轉換)
      - Logit (p)= ln p/1-p (轉換: 將患病機率的勝算比 p/1-p, 取自然對數)
      - Logit (p)= $a+b_1x_1+b_2x_2+\ldots+b_kx_k$ (bi=logistic regression coefficient)
        - X=0 (無致病因子), ln p_0/q_0= a
        - X=1 (有致病因子), ln p_1/q_1= a+b, 兩式相減得
        - ln p_1/q_1 - ln p_0/q_0 = b, so ln $p_1/q_1 \div p_0/q_0$ = b, 取指數成 $p_1/q_1 \div p_0/q_0$ = e^b
      - Odds ratio = e^{b1}, 這個**機率比比值**是**相對危險(RR)**的估計值
        - Wald test statistic: 檢定患病相對危險 (RR)爲1這個虛無假設
    - 電腦輸出 分析 迴歸方法 二元Logistic 依變量 共變量 類別 選項Exp(B)95%CI 方法
      - Wald test, $w=[\beta_i/SE_{\beta i}]^2$, H_o: β_i= 0 ($e^{\beta i}$ =1), H_1: $\beta_i \neq 0$ ($e^{\beta i} \neq 1$)
      - -2log likelihood爲卡方分布，表此模型與自變數不相和程度
      - Model Chi-square or Chi-square for covariates: 檢定模型中所有回歸係數爲零的虛無假設，另外也可用來比較擁有不同共變數的不同模型

Survival analysis 存活分析

(Binary dependent outcome while consider time to event)

- 存活資料分析有**兩個特點**:
  - 注重某事件發生所需時間 (**survival time**, time to event: disease, response, relapse..)
  - 資料常常是設限的 (**censored data +**, lost to follow-up or end of the study)
- **A. Kaplan-Meier survival curve (階梯函數)**
  - 某個體在**某時間點之累進存活機率** (survival probability%)~ median (50%) survival time
    - Ccensored data, 跑掉者不算死亡 ~ Product-limit (PL) method
  - **log-rank test** 是 univariate analysis, 比較兩條存活曲線 p 值有沒有<0.05
  - **log-rank test:** 比較兩組差異之X^2 檢定 (無法評估一個以上因素) $X^2 = [(O_1-E_1)/E_1]^2+[(O_2-E_2)/E_2]^2$
  - **SPSS** 分析 存活分析 KM統計 時間 狀態(定義事件發生與否單一數值:1) 因子 比較因子 選項
- **B. Regression model** (定量**存活與多個因子**關係, multi-variable analysis)
  - **Cox proportional hazards model** (I: incidence, PT: person time)
  - **Incidence density = I** (new case, incidence)**/PT = rate** (單位: 1/時間)
  - **IDR = ID_1/ID_0** (Incidence Density Ratio = 表達相對危險度 relative risk)
  - ln IDR = ln $ID_1/ID_0 = b_1x_1+b_2x_2+\ldots$, 取指數 **(cohort study e^{b1} = IDR = RR)**
  - Hazards 風險 = $\lambda_i(t) = \lambda_o(t)\ \exp^{(\beta 1x1+\beta 2x2+\ldots+\beta nxn)}$
  - 係數 $\beta_1\ldots\beta_n$的指數 **$e^{\beta 1}$... = Hazard Ratios (HR, RR)** or relative hazards
  - H_o: $\beta_1 = 0$, H_1: $\beta_1 \neq 0$, (檢定方法: Likelihood ratio, or Wald statistics)
  - **Nonparametric** method (vs. parametric: exponential, or Weibull model)

Functions of survival time

- **1. Survival function S**(t) ~ (t=∞) 0-1 (t=0)之間 (= exp [∫ -h(u)du])
  - Probability that one survives longer than t, P(T>t) (K-M survival curve ~ %)
- **2. Density function f**(t) ~ (= dF(t)/dt = - dS(t)/dt)
  - The probability that an individual fails in the short interval t to t+Δt per unit width Δt. (The probability of failure in a small interval per unit time)
  - **f**(t) = **P** (an individual dying in the interval t, t+Δ**t**) **/** Δ**t** (limit Δ**t** → 0)
  - **Incidence density = I** (new case, incidence) **/ PT**(人時) **= rate** (單位: 1/時間)
    - **IDR = ID_1/ID_0** (Incidence Density Ratio = 表達相對危險度 relative risk)
    - ln IDR = ln $ID_1/ID_0 = b_1x_1+b_2x_2+\ldots$, 再取指數 **($e^{b1}$ = IDR ≈ RR)**
    - Hazards 風險 = $\lambda_i(t) = \lambda_0(t)$ **exp**$^{(\beta 1x1+\beta 2x2+\ldots+\beta nxn)}$,
- **3. Hazard function h**(t)= **f**(t)**/S**(t)=$\mathbf{h}_0$(t) exp$^{(\beta_1x_1+\beta_2x_2+\ldots+\beta_nx_n)}$
  - The probability of failure during a very small time interval, assuming that the individual has survives to the beginning of the interval. (存活超過某時間的機率)(某人存活到某時間,但再下一刻瞬間死亡的機率)
  - **P**(an individual dying in the interval t, t+Δ**t**) / **P**(one has survived to time t)

Survival curve assumptions (K-M 存活曲線的前提假設)

- Random sample (事先檢查各 prognostic factors是否 homogeneity, p>0.05)
- **Independent observations** (被選入觀察機會均等)
- Consistent entry criteria
- **Consistent criteria** for defining "survival"
- Time of censoring is not unrelated to survival
  - only a small fraction of patients leave the study
  - check what fraction of the patients dropped out the study and why
  - assume that overall survival is not changing over time

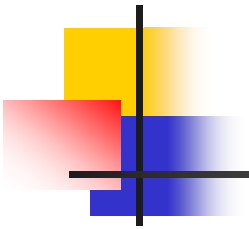
- Comparison of two survival distributions: **log-rank test**
- Compare of K (K>2) samples **(median survival time)**
  - Kruskal and Wallis H – test for uncensored data
  - Peto and Peto generalized H – test for censored data
- **Cox proportional hazard model:** regression approach (heterogeneity)
- Survival distribution: Exponential d. (λ) $f(t)=\lambda e^{-\lambda t}$, $S(t)= e^{-\lambda t}$, $h(t)=\lambda$, Weibull d.

What do the regression coefficients measure (β_j)?

- If we change the measurement of one variable by one and keep all the other variables unchanged, then the relative risk is

$$\frac{\lambda_0(t)\, e^{\beta xz + \ldots}}{\lambda_0(t)\, e^{\beta x(z-1) + \ldots}} = e^{\beta(z - z + 1)} = e^{\beta}$$

- **Thus, the coefficient β is the natural logarithm of the hazard rate ratio when z is increased by 1 unit**



The End