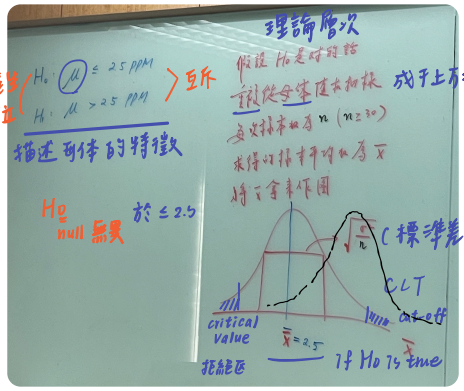
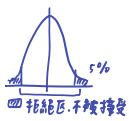
10.12
(四) W3

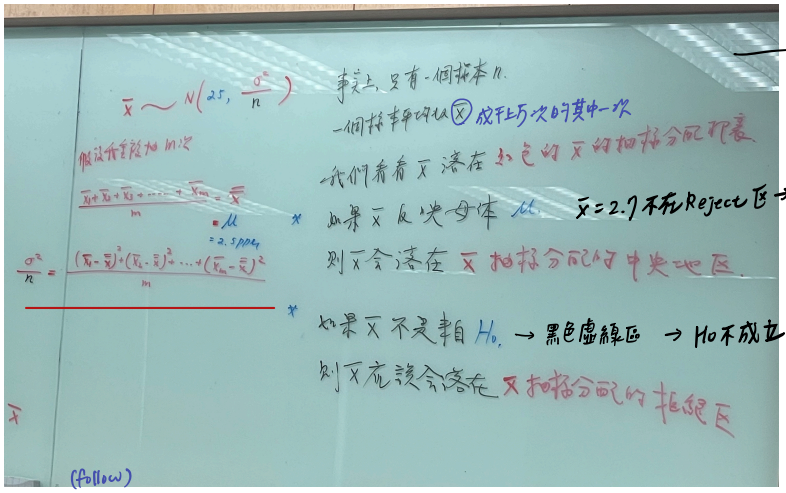
Review
如何用 CLT





紅字：理論層次：Reference
可以不停模擬 (Repeated)

一般而言，是要拒絕 H_0，要支持 $H_1: \mu > 2.5$ ppm → 收集資料去拒絕



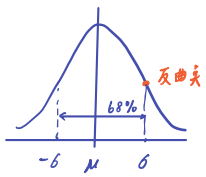
→ 代入 data. 看其落在何處
統計是反證法.回推

四個抽樣分配

$x \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$

i.i.d. 獨立同質分配

常態分配的2个參數



Z分配

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

常態 (高斯分配)

結束

$z = \frac{x-\mu}{\sigma}$ (標準化)

$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$

$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

(Repeated)

從 $x \overset{i.i.d.}{\sim} N(\mu,\sigma^2)$ 隨机抽样 n. 計算 $\bar{x}$

- 1st $x_1, x_2 \dots x_n \to \bar{x}_n$ $\frac{\bar{x}_n-\mu}{\frac{\sigma}{\sqrt{n}}}$
- 2nd $x_1, x_2 \dots x_n \to \bar{x}_n$ $\frac{\bar{x}_n-\mu}{\frac{\sigma}{\sqrt{n}}}$
- ⋮
- m th ⋮ $\bar{x}_m$ $\frac{\bar{x}_m-\mu}{\frac{\sigma}{\sqrt{n}}}$

卡方分配 χ^2
❗ 很重要
z^2 = 卡方

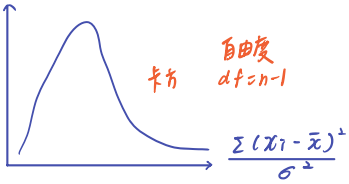
從 $x \overset{i.i.d.}{\sim} N(\mu,\sigma^2)$ 隨机抽样.

離均差平方和

1st $x_1, x_2 \cdots x_{1+n} \to \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 / \sigma^2$

2nd ⋮ ⋮

M th ⋮ $\to \sum_{i=1}^{n} (x_{mi} - \bar{x}_n)^2 / \sigma^2$



$S^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$
過去所学之變異数公式

$\frac{(n-1)S^2}{\sigma^2} = \frac{\Sigma(x_i-\bar{x})^2}{\sigma^2}$ → 卡方公式

Z分配平方是卡方

χ^2 <p.f>

$z_i = \frac{x_i - M}{\sigma}$ $z^2 = \left(\frac{x_i - M}{\sigma}\right)^2 = \chi^2_{(n)}$ $\sum_{i=1}^{n} z_i^2 = \sum_{i=1}^{n} \left(\frac{x_n - M}{\sigma}\right)^2 = \chi^2_{(n)}$

$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - M)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \boxed{\bar{x} + \bar{x}} - M)^2 = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - M)^2\right]$ (統計量)

$= \sum_{i=1}^{n}(x_i - \bar{x})^2 + \left(\frac{\bar{x} - M}{\frac{\sigma}{\sqrt{n}}}\right)^2$

$n-1$ $df=1$

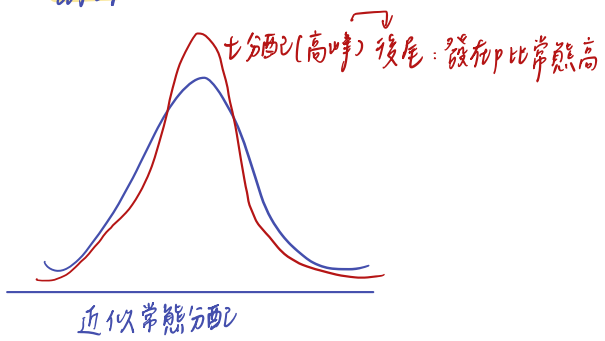
自由度 = E(X)

* 補充 回家查

布拉松分配 $z^2 = \frac{(O-E)^2}{?}$

t分配 從 $x \overset{iid}{\sim} N(M, \sigma^2)$ 隨機抽樣 $t = \frac{\bar{x} - M}{\frac{s}{\sqrt{n}}}$ (以前學的) 大

(TRUE) $t = \frac{z}{\sqrt{\frac{\chi^2}{df}}}$ $\left(\frac{\text{z分配}}{\sqrt{\frac{\text{卡方}}{\text{自由度}}}}\right) = \frac{\frac{\bar{x} - M}{\frac{\sigma}{\sqrt{n}}}\ (z)}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2}\ \text{卡方}}}$ $\frac{}{n-1}$



用於：迴歸分析

F分配 用於 ANOVA · 迴歸分析

來自2個常態分配

單因子變異數分析 (2個變異數相減)

SV	SS	MS	F
組間	SS_B	MS_B	$= s_B^2$
組內	SS_W	MS_W	$= s_W^2$
total			

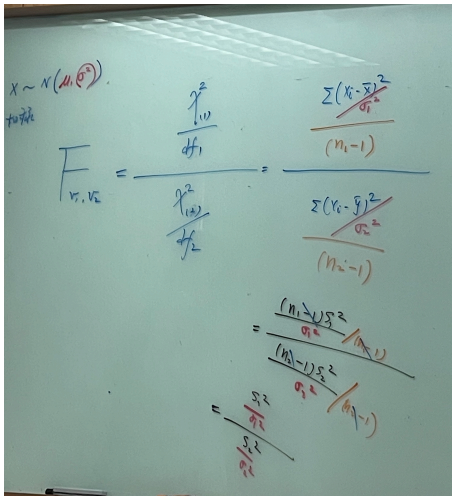
Mean Square = S^2(變異數)

$\sum (y_i - \bar{y})^2 = \underbrace{\sum (y_i - \bar{y})^2 + \sum\sum (\bar{y}_i - \bar{\bar{y}})^2}_{\text{互斥}}$

迴歸分析中的變異數分析

SV	df	MS	F
迴歸 $\sum(\hat{y}_i - \bar{y})^2 = SSR$	#	$\frac{SSR}{\#}$	$\div$
誤差 $\sum(y_i - \hat{y}_i)^2 = SSE$	$n-1-\#$	$\frac{SSR}{n-1-\#}$ 殘差	

$F_{v_1, v_2} = \frac{\frac{\chi^2_{(1)}}{df_1}}{\frac{\chi^2_{(2)}}{df_2}}$

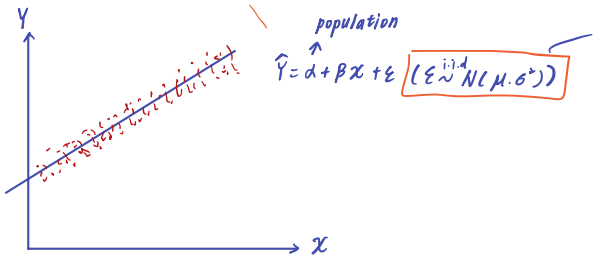


四個抽樣分配使用時機

z分配：$n \geq 30$，標準常態分配

t分配：$n \geq 30$，pic 隨 df 不同而改變

Sample



1st (my sample) $\hat{y}=a+bx$ $y=a+bx+e$

$b_1=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$

2nd

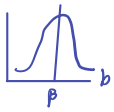
$x \to y$ 沒有效果

→ 身高男的線

3rd

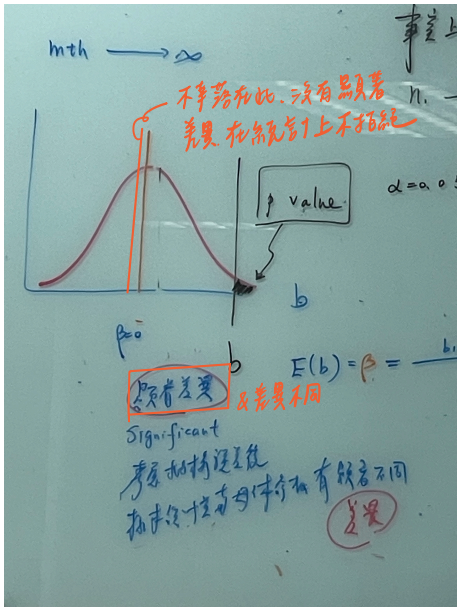
$E(b)=\beta$

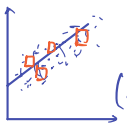
mth → ∞



$\frac{b_1+b_2+b_3+\cdots+b_m}{m}=\beta$

且為常態分配





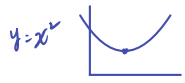
OLS 最小平方法

$\min \sum_{i=1}^{n}(\hat{y}_i-y_i)^2$

(觀測值 - 實際值)² 找最小

$\min \sum_{i=1}^{n}(\alpha+\beta x_i-y_i)^2$ 正方形面積

$=\min \sum_{i=1}^{n}\varepsilon_i^2$



$\hat{\beta}=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$ OLE / OLS Ordinary Least Square / Estimate

$=\frac{cov(x_i,y_i)}{Ver(x_i)}$

$=\frac{\text{長方形面積}(xy)}{\text{正方形面積}(x)}$

MLE = OLE

求解拙 α, β

一定有分配的假設

Today:
- 分配的求解
- 四種分配
- 抽樣分配

Q3.Q4 → 理想狀態

Q5 → TRUELY

Q5

$H_0: M \geq 100$ $n=16$ $\bar{x}=98.9375$

$H_1: M<100$ $s=\cancel{2.794}$ 2.886

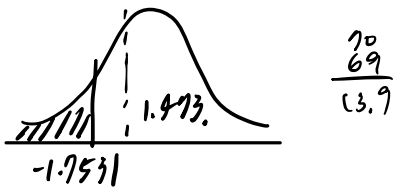
$t=\frac{\bar{x}-M}{s/\sqrt{n}}=\frac{98.9375-100}{2.886/4}=\frac{-1.0625}{0.7215}$

$=1.473$

查t表 ⇒

$t_{0.05,15}=1.7531$

$1.473<1.7531$ 不拒絕



70
69
13.9

10.26 統計一定會有「誤差項」

4.3 Measure of fit

使每個點相減平方和為 min

好的參數估計值包括 μ.

① 不偏性：$E(S^2)=\sigma^2$

$E\left(\frac{\sum(x_i-\bar{x})^2}{n-1}\right)=\frac{\sum(x_i-\mu)^2}{N}$

② 有效性

(有最小的變異誤 有min的標準誤)

③ 充份性：用到 all 樣本資訊

④ 一致性：$n\uparrow$, $B\to\beta$ 越接近母體 β

11.2

$y\text{-score}=\frac{\text{read}+\text{math}}{2}$

x = 生師比 STR

OR

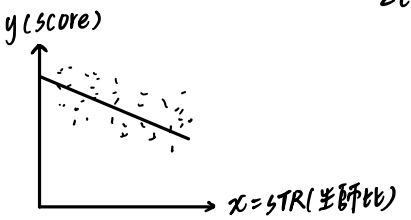
$x=\begin{cases}1 & \text{if } STR<20\\ 0 & \text{o.w.}\end{cases}$ 當參考組

連續

= dummy

= 兩個類別(0,1)

要去解釋單其變異



$y_i=a+bx_i+e_i=\hat{y}_i+e_i$

$\hat{y}_i=a+bx_i$ if $b=0$ $(=a=\bar{y})$

Sum of Square (y的變異)

$SST=SSY=SSR+SSE$

↳ $\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2$

$=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum(\hat{y}_i-\bar{y})^2$

df=2. 估a&b　df=1. from $\bar{x}$

$=SS_{ERROR}+SS_{Reg}$

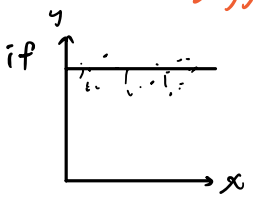
$s^2=\frac{SSy}{dfy}$ 自由度

$=\frac{\sum(y_i-\bar{y})^2}{n-1}$

= mean square (均方)

無論連續 or dummy

$b=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$ <不變>



if

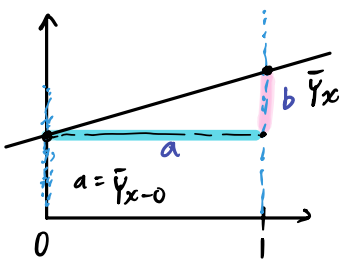
x對y沒有影響.

No patten

R^2 (可以解釋 y 的變異程度，決定係數)

$=\frac{SSR}{SST}$ or $1-\frac{SSE}{SST}$

Dummy



$=\bar{Y}_{x=1}-\bar{Y}_{x=0}$

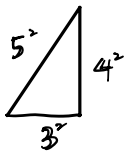
= 2個獨立樣本平均數的差

= t-test

適配度檢定
(F-Test)
1個卡方.獨立

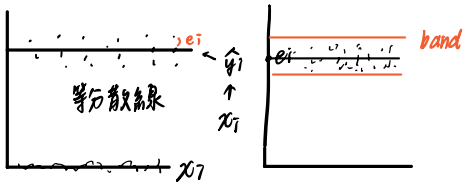
OVERALL TEST (卡方)

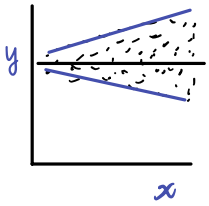
$$F_{df_R, df_E} = \frac{\frac{SSR}{df_R}}{\frac{SSE}{df_E}\ (S^2)} = \frac{\frac{(n-1)\cdot s_1^2}{df_1\ (p)}}{\frac{(n-1)\cdot s_2^2}{df_2\ (n-p-1)}}$$



變異數分析 → 2個S^2相除

sum of square





不符合等分散

lesson 2
信賴區間

If H_0是對的話.

至少有一Y變項.斜率≠0
但不知道是哪個.

(Overall)

F 用來檢定 H_0

$$SST = SSy = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$= \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

$$= SSE + SSR$$

Ch6 多元迴歸

x_1 = STR
x_2 = lunch
x_3 = English

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + e$$

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ H_1: 至少一個$\beta \neq 0$

null = 無異於0

迴歸係數 ⇒ partial regression coefficients
解釋 "Marginal"
設其他條件不變的時候.x增加一單位,y可以增加b1個單位

x_2, x_3, x_4 在model中.當2個人有相同

多元共線性：指的是x_i之間資訊影響的程度 or x_i之間彼此被Copy的程度

Another $\hat{x}_1 = \alpha_0 + \alpha_2x_2 + \alpha_3x_3 + \alpha_4x_4$

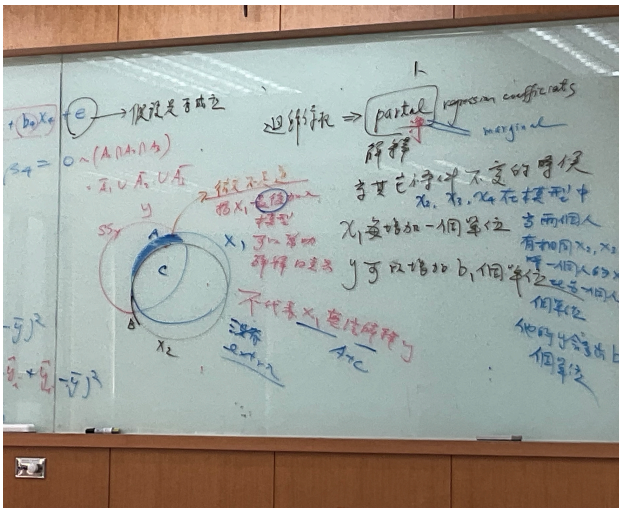
①.

$$VIF_{x1} = \frac{1}{1 - R^2_{x1|x2,x3,x4}}$$

有幾個解釋變數就有幾個 VIF.

(Variance Inflation factor)
變異數 膨脹

② CN C Condition Index) >30



$R^2_{adjusted} = R^{-2} =$ pseudo R^2

$$R^2 = 1 - \frac{SSE}{SST} \longrightarrow R_{adjust} = 1 - \frac{n-1}{n-p-1}(1-R^2) = \frac{s_y^2 - s_e^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

① 主效份迴歸
② 資迴歸
> 1.6 前 ,type "text" 才有報表.