112-2 生物統計學一

# 考前複習

2024/05/21

助教: 廖振博

# R常見的物件/資料型態

1. 字串(character string)/數值(numeric):

```
> (aa <- "AA")
[1] "AA"
> (bb <- 100)
[1] 100
```

2. 向量(Vector): 用來儲存單一維度的資料，只能包含一種類型的資料(例如數值或字串)

```
> (cc <- c(1, 3, 4, 5, 6, 7))
[1] 1 3 4 5 6 7
```

3. 矩陣(Matrix): 用來儲存兩個維度的資料，只能包含一種類型的資料

```
> (dd <- matrix(cc, 2, 3, byrow = T))
     [,1] [,2] [,3]
[1,]    1    3    4
[2,]    5    6    7
```

4. 資料框(Data frame): 用來儲存兩個維度的資料，可以包含多種類型的資料

```
> (ee <- data.frame(Name = c("John", "Alice"), Age = c(25, 30)))
   Name Age
1  John  25
2 Alice  30
```

5. 列表(Data frame): 用來儲存很多個物件，可以包含向量、矩陣、資料框、甚至其他列表

```
> (ff <- list(c(1, 2, 3), c("A", "B")))
[[1]]
[1] 1 2 3

[[2]]
[1] "A" "B"
```

# 資料框(Data Frame)的CRUD

1. C (create): 建立物件

```
> id <- c(1:4)
> age <- c(25, 30, 35, 40)
> sex <- c("male", "male", "female", "female")
```

```
> (zz = data.frame(id, age, sex))
  id age    sex
1  1  25   male
2  2  30   male
3  3  35 female
4  4  40 female
```

2. R (read): 讀取物件裡的元素

```
> zz$sex
[1] "male"   "male"   "female" "female"
> zz[, 3]
[1] "male"   "male"   "female" "female"
```

3. U (update): 更新/修改物件裡的元素

```
> zz$age <- c(10, 10, 10, 10)
```

4. D (delete): 刪除物件或刪除物件裡的元素

```
> zz[, -1]
```

# 資料匯入與匯出(csv檔)

1. 匯入csv檔(逗點分隔)

```
> read.csv(file =  , header = TRUE, sep = ",")
```

- file(檔案路徑): 檔案路徑(注意斜線方向)，路徑需加上引號或使用file.choose()點選
- header(欄位標題): 第一列是否為標題
- sep(分隔符號): 檔案由什麼符號分隔

2. 匯出csv檔

```
> write.csv(x, file = , row.names = TRUE)
```

- x(資料表格): R內部資料的名稱
- file(檔案路徑): 匯出檔案路徑
- row.names(列標籤): 每列名稱是否要匯出
- sep(分隔符號): 檔案是由什麼符號分隔

# 查看資料內容

1. 查看整個資料

```
> View()
```

2. 顯示前/後幾筆資料，預設為6筆

```
> head() ; tail()
```

3. 查看資料欄列數

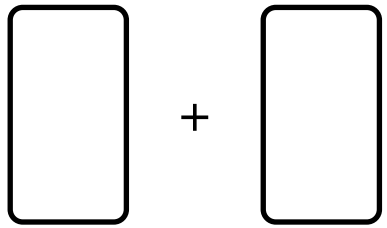
```
> dim()
```

4. 資料欄或列的名稱

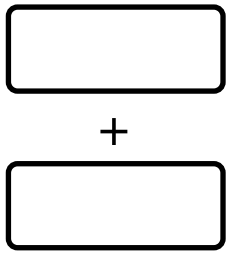
```
> colnames() ; rownames()
```

# 資料合併

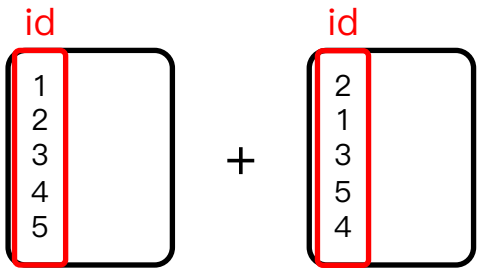
- cbind(資料1,資料2, …….): 行資料合併



- rbind(資料1,資料2, ……): 列資料合併



- merge(資料1, 資料2, by = “共同欄位”): 欄資料合併(根據共同欄位合併)



# 邏輯判斷

- 常見的邏輯判斷符號：
  - > 、 < : 大於、小於
  - >= 、 <= : 大於等於、小於等於
  - == 、 != : 等於、不等於
  - A %in% B : A 是否在 B 中
  - & : 交集(and)
  - | : 聯集(or)

- ifelse(條件, 滿足條件時的值, 不滿足條件時的值)

# 資料篩選

- which(邏輯判斷)

```
> which(data1$BMI > 40)
 [1]    6    8   11   35   42   63   84  103  124  134  144  187  193  231  232
[16]  251  257  292  296  312  321  324  334  343  346  376  384  389  402  420
[31]  480  508  517  519  548  550  601  629  640  644  652  695  742  752  767
[46]  771  781  826  836  837  838  863  868  884  907  949  987 1048 1054 1068
[61] 1082 1083 1096 1115 1162 1173 1230 1246 1256 1258 1282 1287 1382 1383 1387
[76] 1463 1475 1518 1529 1573 1582 1590 1594 1620 1642 1689 1721 1727 1914 2151
[91] 2253 2305 2330 2331 2338 2491 2493 2511 2646
```

- subset(資料, 邏輯判斷, select = 欲選取欄位)

```
> subset(data1, BMI > 40, select = c(age, diabetes, BMI))
     age diabetes   BMI
6     46      yes 46.26
8     46       no 40.62
11    47       no 44.75
35    51      yes 44.38
42    52       no 40.06
63    53       no 45.04
```

# 描述性統計

- 一般敘述統計

  ```
  > summary()
  ```

- 總合與個數

  ```
  > sum(); length(); dim()
  ```

- 平均數、標準差、變異數、中位數

  ```
  > mean(); sd(); var()
  > median()
  ```

- 百分位數(四分位數)及四分位差(IQR)

  ```
  > quantile(data, probs = c(0.25, 0.5, 0.75))
  > IQR()
  ```

- 單變數的次數分配表

  ```
  > table(x)
  ```

- 兩個變數的次數分配表

  ```
  > table(x1, x2)
  ```

- 比例分配表

  ```
  > prop.table(table(x))
  > prop.table(table(x1, x2), option)
  ```

  - option: 預設顯示每個表格佔所有表格總數的比例
  - option = 1 將顯示列比例
  - option = 2 將顯示行比例

# 常見統計圖形

- 描述連續型隨機變數
  - 莖葉圖(stem and leaf plot) `> stem(x)`
  - 直方圖(histogram) `> hist(x)`
- 描述類別型隨機變數
  - 圓餅圖(pie plot) `> pie(table(x))`
  - 長條圖(bar plot) `> barplot(table(x))`
- 描述兩個以上的隨機變數
  - 散佈圖(scatter plot) `> plot(x1, x2)`
  - 盒型圖(box plot) `> boxplot(x)`

# 單一樣本Z檢定(One-sample z-test)

- 檢定統計量： $Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$

```
> z_obs <- (mean(x) - mu0) / (sigma / sqrt(n))
```

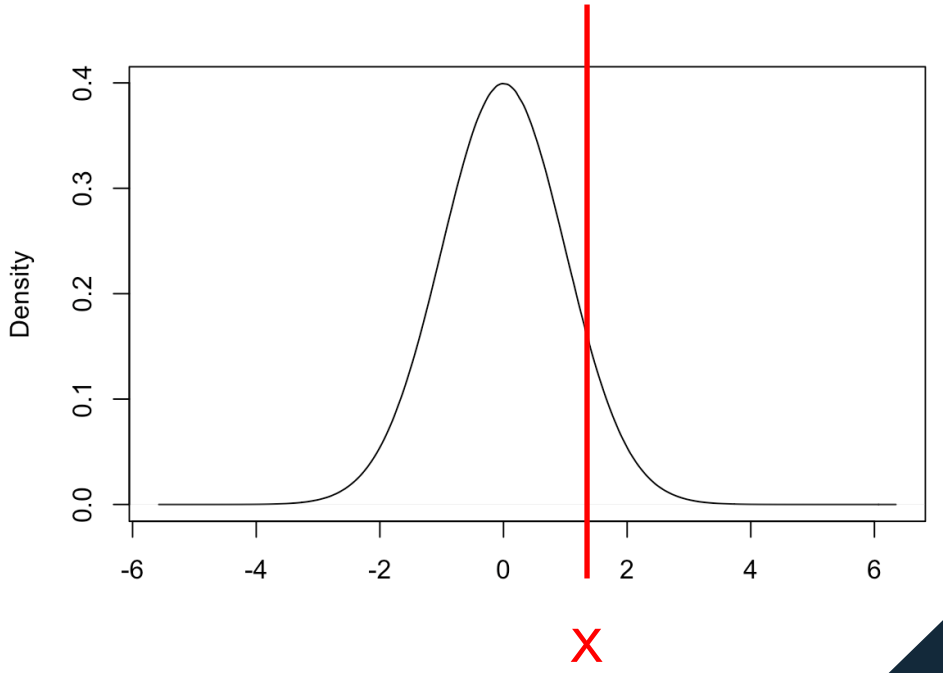
- 計算P-value(在$H_0$下，能得到比手上這筆樣本更極端的機率)：

```
> pnorm(x, 0, 1)
```

- 給定顯著水準$\alpha = 0.05$下：

1. 若p-value < 0.05，拒絕虛無假設，表示有足夠的證據可以推論$\mu \neq \mu_0$。($\mu$與$\mu_0$有達統計顯著上的差異)

2. 若p-value > 0.05，無法拒絕虛無假設，表示沒有足夠的證據可以推論$\mu \neq \mu_0$



# 單一樣本T檢定(One-sample t-test)

- 檢定統計量：$T = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}}$

```
> t_obs <- (mean(x) - mu) / (S / sqrt(n))
```

- 計算P-value：

```
> pt(x, df)
```

- 使用內建函數：

```
> t.test(studata1$Age, alternative = "two.sided", mu = 18)

        One Sample t-test

data:  studata1$Age
t = 2.2341, df = 43, p-value = 0.03073
alternative hypothesis: true mean is not equal to 18
95 percent confidence interval:
 18.13714 20.68105
sample estimates:
mean of x
 19.40909
```

# 兩獨立樣本T檢定(Two independent sample t-test)

- Step 1. 設立虛無與對立假說
- Step 2. 判斷變異數同質性

```
> var.test(y ~ x)
```

- Step 3. 依據變異數同質/異質，算出對應的檢定統計量 t、決定自由度
  - 若變異數同質 (相等) (equal variance)

    ```
    > t.test(y ~ x, var.equal = T)
    ```

  - 若變異數異質 (不相等) (unequal variance)

    ```
    > t.test(y ~ x, var.equal = F)
    ```

- Step 4. 判斷顯著性並下結論

# 配對樣本T檢定(Paired t-test)

```
> t.test(studata1$Credit_Last, studata1$Credit_Current, paired = T)

        Paired t-test

data:  studata1$Credit_Last and studata1$Credit_Current
t = 1.1757, df = 49, p-value = 0.2454
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.035608  3.955608
sample estimates:
mean of the differences
                   1.46
```

# 變異數分析(Analysis of variance, ANOVA)

- 適用情境：檢定三組或以上母體平均值是否相同

  (需檢查假設是否符合，包含變異數同質性假設)

```
> bartlett.test(y ~ x)
```

- 假說：

  $H_0$: 各組母體平均值均相等 ($\mu_1 = \mu_2 = \mu_3 = \ldots$)

  $H_1$: 各組母體平均值不完全相等 (At least two $\mu_i's$ are not equal)

$$F = \frac{\frac{SSB}{k-1}}{\frac{SSB}{n-k}} = \frac{MS_B}{MS_W}$$

```
> anova_test <- aov(y ~ as.factor(x), data)
> summary(anova_test)
```

若x為數字，需要將它轉換成因子(factor)，否則會被R當成數值，並執行迴歸分析

- 多重檢定(multiple testing): Bonferroni correction
  - 將每次個別檢定的顯數水準$\alpha$調整為較保守的$\alpha/n$
  - R的指令是將算出來的p-value乘上$n$後與$\alpha$比較，兩者等價

```
> pairwise.t.test(y, x, p.adjust.method = "bonf")
```

# 期末評鑑

- https://investea.aca.ntu.edu.tw/opinion/guide.asp