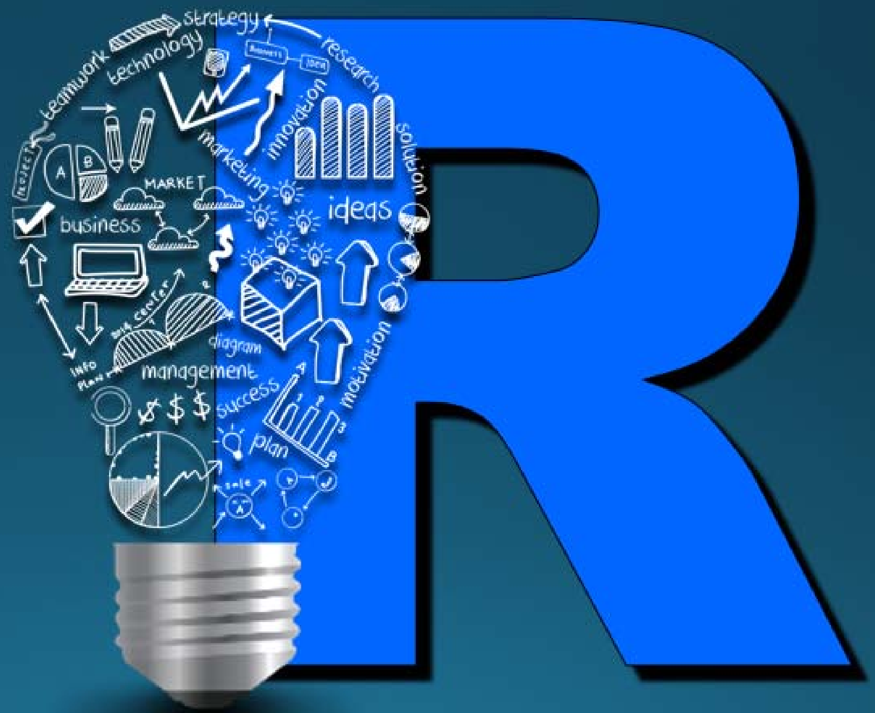
B04

# 假設檢定 &
# 變異數分析



**吳漢銘**

國立政治大學 統計學系

http://www.hmwu.idv.tw

# 本章大綱

- 統計假設檢定 (Hypothesis Testing)
- 型一誤差、型二誤差
- *p*-值
- 母體平均數檢定 (單一樣本t檢定)
- 單因子變異數分析 (One-way Analysis of Variance, ANOVA)
- 卡方檢定 (chi-square test)

# 假設檢定 (Hypothesis Testing)

***假設檢定***是一個用來決定母體特徵(參數)的命題是否為合理的程序。

***例子(1):***

"麻薩諸塞州(Massachusetts)的加油站平均一加崙的汽油(regular unleaded gas)價格是 $2.5 元"

***這個命題是對的嗎?***

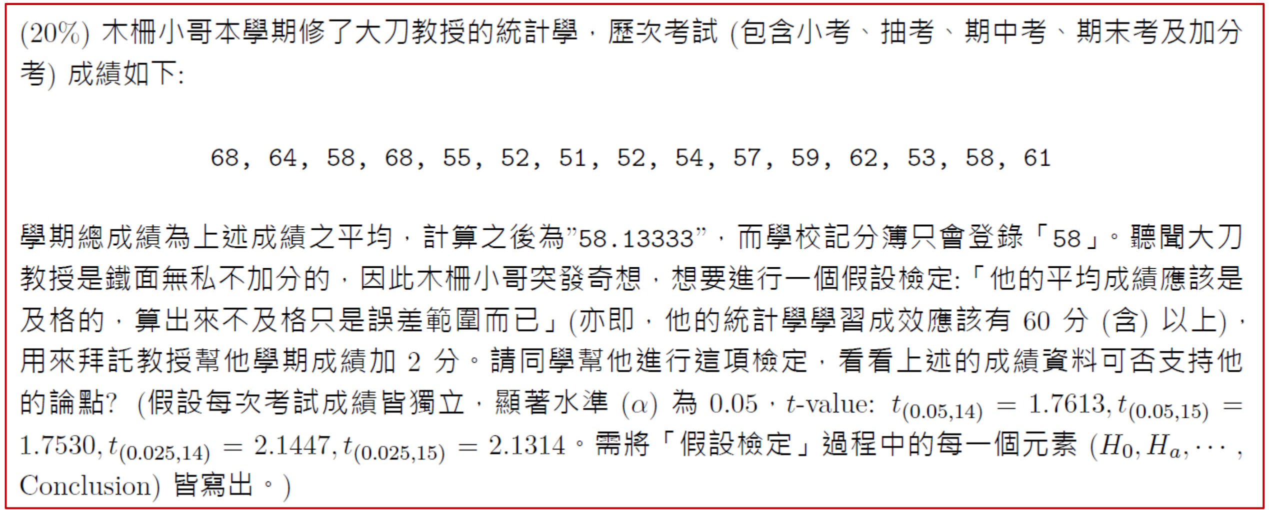
- 對所有加油站做調查。
- 隨機選一小部份加油站當樣本做調查。

CITGO
2.23 9/10 REGULAR UNLEADED
2.33 9/10 MID GRADE UNLEADED
2.43 9/10 PREMIUM UNLEADED

***若從樣本調查出的結果是平均價格為$2.2元.***

- 這30分的差異是隨機變異(chance variability)的結果，還是
- 原本的命題不對?

***例子(2):***



(20%) 木柵小哥本學期修了大刀教授的統計學，歷次考試 (包含小考、抽考、期中考、期末考及加分考) 成績如下:

68, 64, 58, 68, 55, 52, 51, 52, 54, 57, 59, 62, 53, 58, 61

學期總成績為上述成績之平均，計算之後為"58.13333"，而學校記分簿只會登錄「58」。聽聞大刀教授是鐵面無私不加分的，因此木柵小哥突發奇想，想要進行一個假設檢定:「他的平均成績應該是及格的，算出來不及格只是誤差範圍而已」(亦即，他的統計學學習成效應該有 60 分 (含) 以上)，用來拜託教授幫他學期成績加 2 分。請同學幫他進行這項檢定，看看上述的成績資料可否支持他的論點? (假設每次考試成績皆獨立，顯著水準 ($\alpha$) 為 0.05，$t$-value: $t_{(0.05,14)} = 1.7613, t_{(0.05,15)} = 1.7530, t_{(0.025,14)} = 2.1447, t_{(0.025,15)} = 2.1314$。需將「假設檢定」過程中的每一個元素 ($H_0, H_a, \cdots,$ Conclusion) 皆寫出。)

# 假設檢定

## *虛無假設 (Null hypothesis):*

- $H_0$: $\mu$ = 2.5. (the average price of a gallon of gas is $2.5)

## *擇一假設 ( alternative hypothesis):*

- $H_a$: $\mu$ > 2.5. (gas prices were actually higher)
- $H_a$: $\mu$ < 2.5.
- $H_a$: $\mu$ != 2.5. (雙尾檢定)

## *顯著水準 (significance level )(alpha):*

- 需事先決定。
- Alpha = 0.05: the probability of incorrectly rejecting the null hypothesis when it is actually true is 5%.

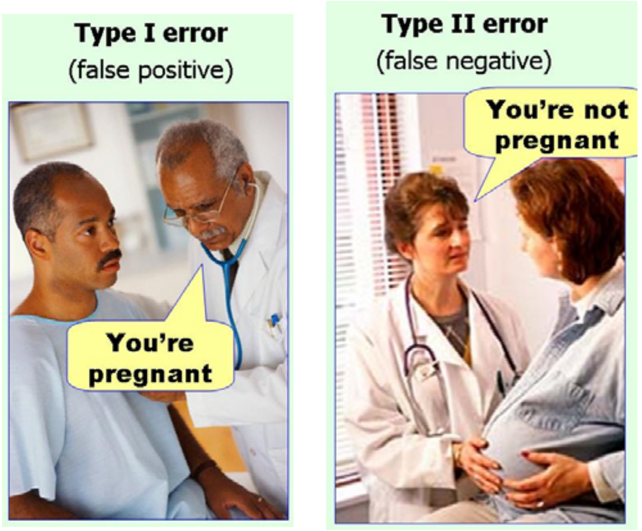
  (虛無假設對之下，拒絕虛無假設的機率)

  (錯誤地拒絕虛無假設的機率)

# 型一誤差、型二誤差

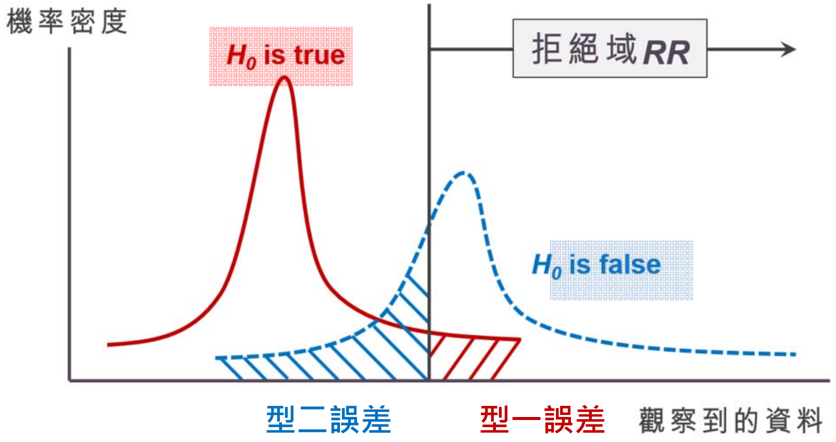
| 假設檢定 | | 真實 (Truth) | |
|---|---|---|---|
| | | $H_0$ | $H_1$ |
| 決策 (Decision) | Reject $H_0$ | Type I Error<br>(α)<br>(false positive) | Right Decision<br>(true positive)<br>Power = 1- β |
| | Fail to Reject $H_0$ | Right Decision<br>(true negative) | Type II Error<br>(β)<br>(false negative) |

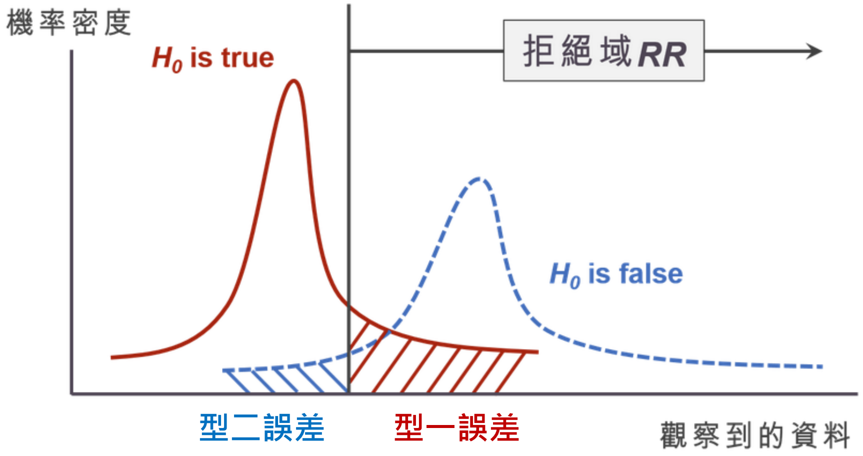
$H_0$: Not Pregnant



https://effectsizefaq.com/category/type-i-error/

「拒絕虛無假設的標準」





https://taweihuang.hpd.io/2017/01/11/poorpvalue/

# *p*-值 (The *p*-value)

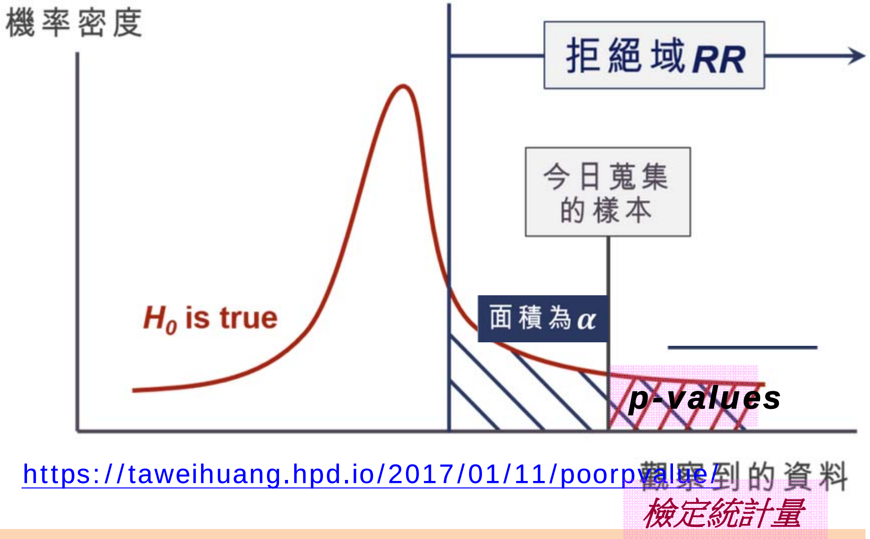
## *p-value:*

Harry Potter,
分類帽(Sorting Hat)

- 定義：在已知(現有)的抽樣樣本下，能棄卻 $H_0$(虛無假設)的最小顯著水準。(Reject $H_0$ | $H_0$ true)
- 若$H_0$ 為真，則檢定統計量出現(觀察到此樣本)的可能性。
  (若p-value越小，表示抽樣樣本越不可能出現，因此推翻假設，拒絕$H_0$)。
- p-value：以現有的抽樣所進行的推論，可能犯 type I error 的機率。
  (若p-value越小，表示拒絕$H_0$不太可能錯，因此拒絕$H_0$)。

## *決策法則:*

- 拒絕$H_0$ 若 *p-value* 比alpha小。
- *P* < 0.05 commonly used.
  (拒絕$H_0$，稱檢定是顯著的(significant)
- The lower the *p-value*, the more significant.

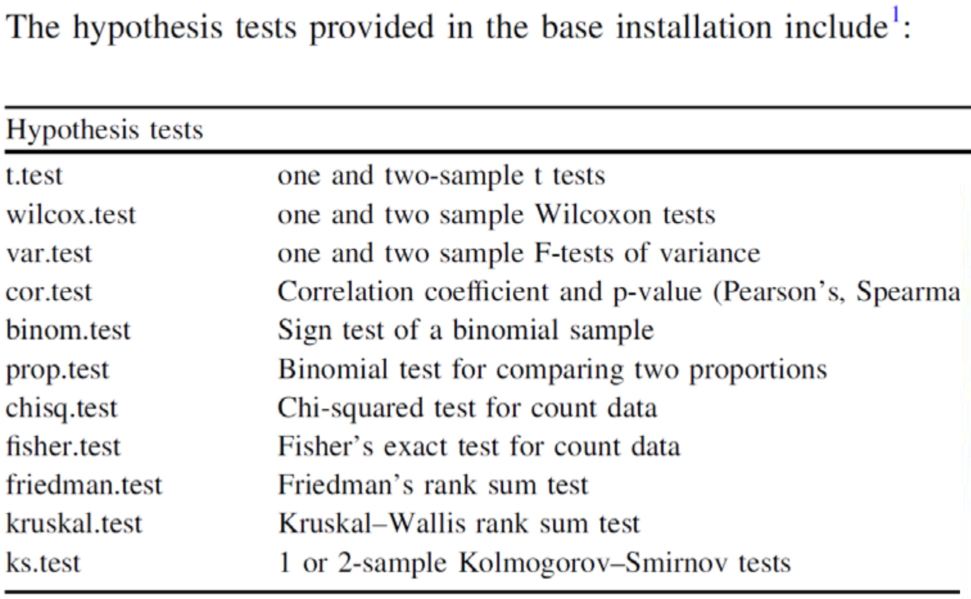


https://taweihuang.hpd.io/2017/01/11/poorpvalue/

"**只要是使用正確的意義，p-value並沒有問題，只是不要去誤用它**。不要只是著重在統計顯著性，因為model對錯的機率跟p-value不一樣。要使用p-value作檢定，要把它跟α來做比較，所以問題不只是p-value，而是α。界定了α之後，才知道結果是不是顯著。當得到一個顯著的結果以後，必須再來衡量偽陽性反機率的問題，也就是model後設機率的問題，這就不是p-value可以告訴你的。"

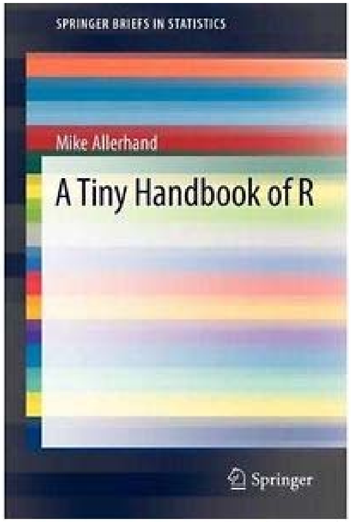
林澤民，看電影學統計: p值的陷阱
http://blog.udn.com/nilnimest/84404190
社會科學論叢2016年10月第十卷第二期

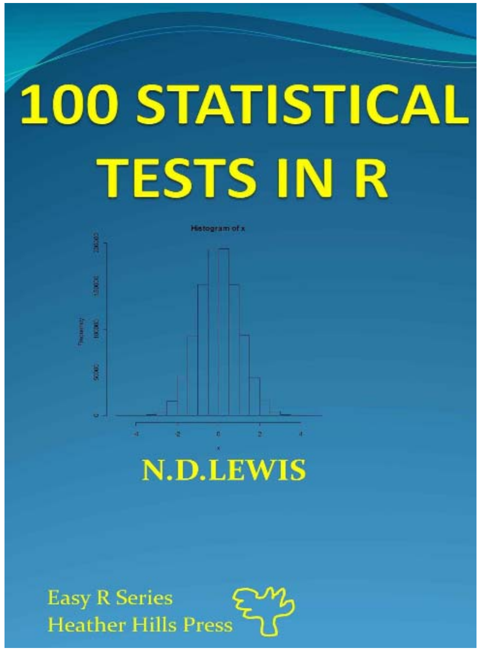
# The Hypothesis Tests in Base R



The hypothesis tests provided in the base installation include[1]:

| Hypothesis tests | |
|---|---|
| t.test | one and two-sample t tests |
| wilcox.test | one and two sample Wilcoxon tests |
| var.test | one and two sample F-tests of variance |
| cor.test | Correlation coefficient and p-value (Pearson's, Spearma |
| binom.test | Sign test of a binomial sample |
| prop.test | Binomial test for comparing two proportions |
| chisq.test | Chi-squared test for count data |
| fisher.test | Fisher's exact test for count data |
| friedman.test | Friedman's rank sum test |
| kruskal.test | Kruskal–Wallis rank sum test |
| ks.test | 1 or 2-sample Kolmogorov–Smirnov tests |





N.D Lewis, 100 Statistical Tests in R, Publisher: CreateSpace Independent Publishing Platform (April 15, 2013)

# 平均數檢定 in R

| Hypothesis Testing | One Sample | Two Samples | | > two Groups |
|---|---|---|---|---|
| | - | Paired data | Unpaired data | Complex data |
| **Parametric (variance equal)** | **t-test**<br>`t.test(x, mu = 0)` | **t-test**<br>`t.test(x-y, var.equal = TRUE)`<br>`t.test(x, y, paired = TRUE, var.equal = TRUE)` | **t-test**<br>`t.test(x, y, var.equal = TRUE)` | **One-Way Analysis of Variance (ANOVA)**<br>`aov(x~g, data)`<br>`oneway.test(x~g, data, var.equal = TRUE)` |
| **Parametric (variance not equal)** | | **Welch t-test**<br>`t.test(x-y)`<br>`t.test(x, y, paired = TRUE)` | **Welch t-test**<br>`t.test(x, y)` | **Welch ANOVA**<br>`oneway.test(x~g, data)` |
| **Non-Parametric** (無母數檢定) | **Wilcoxon Signed-Rank Test**<br>`wilcox.test(x, mu = 0)` | **Wilcoxon Signed-Rank Test**<br>`wilcox.test(x-y)`<br>`wilcox.test(x, y, paired = TRUE)` | **Wilcoxon Rank-Sum Test (Mann-Whitney U Test)**<br>`wilcox.test(x, y)` | **Kruskal-Wallis Test**<br>`kruskal.test(x, g)` |

`pairwise.t.test {stats}`: Calculate pairwise comparisons between group levels with corrections for multiple testing

`TukeyHSD {stats}`: Compute Tukey Honest Significant Differences

# 單一樣本t-檢定 (t-test)

## 可能的應用問題:

- 一家醫院想知道病患膽固醇值的平均數是否與目標值200mg不同?
- 消保官想了解能量棒上的標示「此能量棒含20公克的蛋白質」是否正確?

- 設定虛無假設及擇一假設。

  $H_0$: $\mu = \mu_0$
- 選定$\alpha$
- 收集資料: $x_1, x_2, \ldots, x_n$。
- 驗証假設。
- 計算平均數、變異數。
- 計算檢定統計量。
- 算*p*-值。
- 做決策。

*p*-value approach

Critical value approach

### One sample t-test

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$ (two-tailed).

$\mu$: population mean.

$\alpha$: significant level (e.g., 0.05).

Test Statistic:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}, \quad t_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$\bar{X}$: sample mean.

$S$: sample standard deviation.

$n$: number of observations in the sample.

- Reject $H_0$ if $|t_0| > t_{\alpha/2,n-1}$.
- Power $= 1 - \beta$.
- $(1-\alpha)100\%$ Confidence Interval for $\mu$:

  $\bar{X} - t_{\alpha/2}S/\sqrt{n} \leq \mu < \bar{X} + t_{\alpha/2}S/\sqrt{}$
- *p-value* $= P_{H_0}(|\mathbf{T}| > t_0)$, $\mathbf{T} \sim t_{n-1}$.

**雙尾檢定 (two-tailed test)**

**單尾檢定**

**左尾**
(Lower tail)

$H_0 : \mu \geq \mu_0$

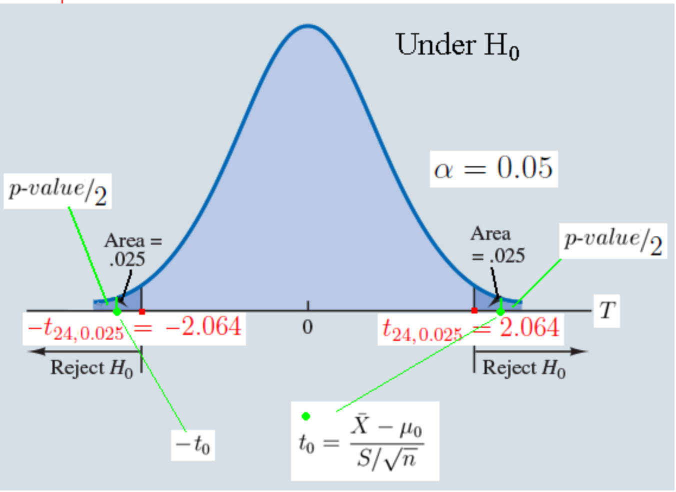
$H_a : \mu < \mu_0$

**右尾**
(Upper tail)

$H_0 : \mu \leq \mu_0$

$H_a : \mu > \mu_0$

***T*的抽樣分佈**
**(sampling distribution)**



# t檢定的假設 (Assumption)

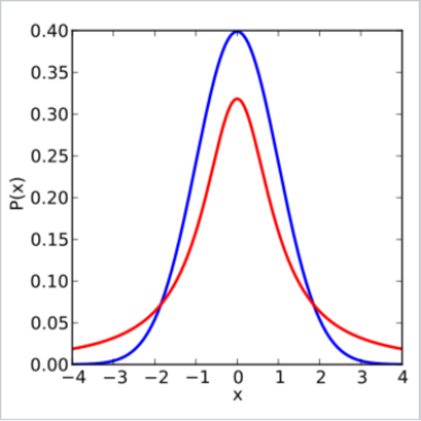
假設$X$是呈常態分布的獨立的隨機變量（隨機變量的期望值是$\mu$，方差是$\sigma^2$但未知）。

$$\overline{X}_n = (X_1 + \cdots + X_n)/n$$

$$S_n{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

$$T = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{(n-1)}$$

*t*-分布密度 (紅色曲線)
標準常態分布(藍色曲線).



1 degree of freedom

William Sealy Gosset, who developed the "*t*-statistic" and published it under the pseudonym of "Student".

- William Sealy Gosset, a chemist working for the Guinness brewery in Dublin, Ireland. "Student" was his pen name.
- 1908, Biometrika.

## 常態分佈 (Normal)

- 資料必需為常態分佈。
  (若不符合，有一些經驗法則(對稱分佈、樣本數很大、轉換)或
  改採用「無母數檢定」。)
- *如何檢測資料是否為常態?*
  - Plots: Histogram, Density Plot, QQplot,…
  - Test for Normality: Jarque-Bera test, Lilliefors test, Kolmogorov-Smirnov test, Shapiro-Wilk test.

## 同質性 (Homogeneous)

- (雙樣本t檢定) 兩母體的變異數要相同。
- Test for equality of the two variances: Variance ratio F-test.
- Tests in R: `var.test`, `bartlett.test`, `ansari.test`, `mood.test`, `fligner.test`, `leveneTest`.

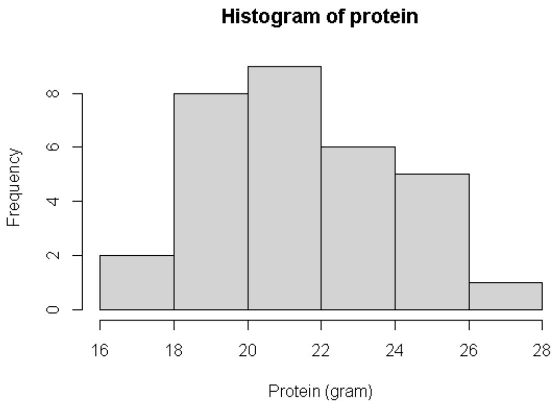
# 範例:消保官想了解能量棒上的標示「此能量棒含20公克的蛋白質」是否正確? (t檢定)

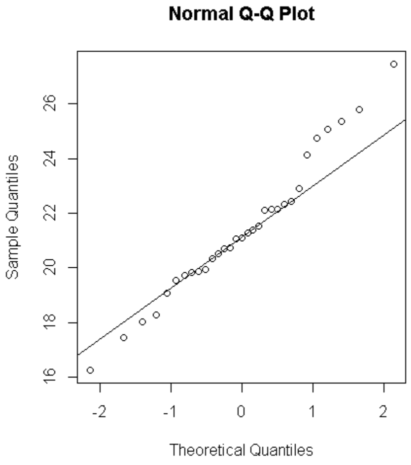
$$H_0\colon \mu = 20, \quad H_1\colon \mu \neq 20, \alpha = 0.05.$$

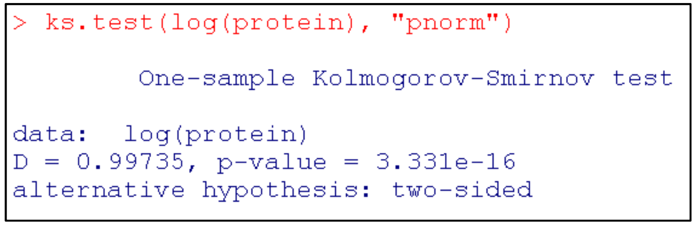
| 營養成分 每份(50克) | |
|---|---|
| 熱量 | 190大卡 |
| 蛋白質 | 20克 |
| 碳水化合物 | 17克 |
| 總脂肪 | 6克 |
| 飽和脂肪 | 3.5克 |
| 膽固醇 | 15毫克 |
| 鈉 | 180毫克 |
| 膳食纖維 | <1克 |
| 糖 | 2克 |
| 糖醇 | 8克 |

## 31根能量棒的蛋白質含量(克數):

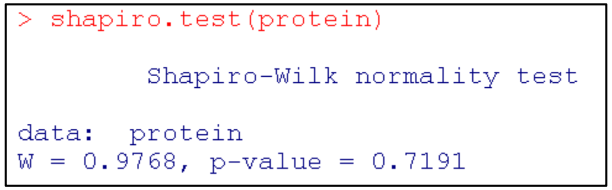
20.70, 27.46, 22.15, 19.85, 21.29, 24.75, 20.75, 22.91, 25.34, 20.33, 21.54, 21.08, 22.14, 19.56, 21.10, 18.04, 24.12, 19.95, 19.72, 18.28, 16.26, 17.46, 20.53, 22.12, 25.06, 22.44, 19.08, 19.88, 21.39, 22.33, 25.79



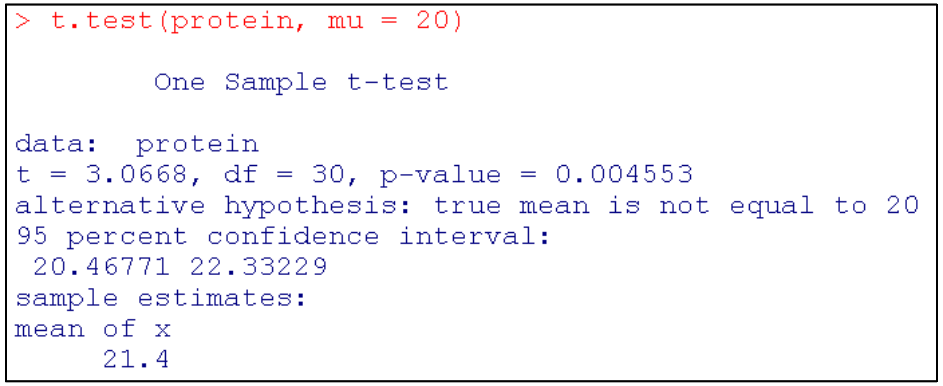




```
> ks.test(log(protein), "pnorm")

        One-sample Kolmogorov-Smirnov test

data:  log(protein)
D = 0.99735, p-value = 3.331e-16
alternative hypothesis: two-sided
```



```
> shapiro.test(protein)

        Shapiro-Wilk normality test

data:  protein
W = 0.9768, p-value = 0.7191
```



```
> t.test(protein, mu = 20)

        One Sample t-test

data:  protein
t = 3.0668, df = 30, p-value = 0.004553
alternative hypothesis: true mean is not equal to 20
95 percent confidence interval:
 20.46771 22.33229
sample estimates:
mean of x 
     21.4 
```

拒絕「平均蛋白質公克數等於 20」 的虛無假設。標示資訊不正確，且蛋白質公克數的母體實際上平均數大於 20 。

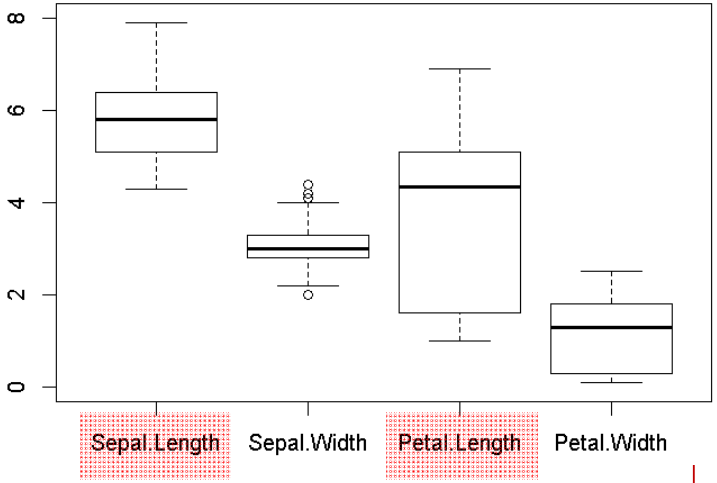
標籤資訊應該更新，或製造流程應該改善，以製造出平均含 20 公克蛋白質的能量棒。

# t.test {stats}: Student's t-Test

**Description**: Performs one and two sample t-tests on vectors of data.

**Usage**:
```
t.test(x, y = NULL,
        alternative = c("two.sided", "less", "greater"),
        mu = 0, paired = FALSE, var.equal = FALSE,
        conf.level = 0.95, ...)
```

```
> x <- iris$Sepal.Length
> y <- iris$Petal.Length
> alpha <- 0.05
> (vt <- (var.test(x, y)$p.value <= alpha))
[1] TRUE
> t.test(x, y, var.equal = !vt )

        Welch Two Sample t-test

data:  x and y
t = 13.098, df = 211.54, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.771500 2.399166
sample estimates:
mean of x mean of y
 5.843333  3.758000
```



# Other t-Statistics

## B-statistic

Lonnstedt and Speed, Statistica Sinica 2002: parametric empirical Bayes approach.

- B-statistic is an estimate of the posterior log-odds that each gene is DE.
- B-statistic is equivalent for the purpose of ranking genes to the penalized t-statistic $t = \frac{\bar{M}}{\sqrt{(a+s^2)/n}}$, where $a$ is estimated from the mean and standard deviation of the sample variances $s^2$.

$$M_{gj}|\mu_g, \sigma_g \sim N(\mu_g, \sigma_g^2)$$

$$B_g = \log \frac{P(\mu_g \neq 0|M_{gj})}{P(\mu_g = 0|M_{gj})}$$

Lonnstedt, I. and Speed, T.P. Replicated microarray data. *Statistica Sinica* , 12: 31-46, 2002

## Penalized t-statistic

Tusher et al (2001, PNAS, SAM)

Efron et al (2001, JASA)

$$t = \frac{\bar{M}}{(a+s)/\sqrt{n}}$$

## General Penalized t-statistic

(Lonnstedt et al 2001)

$$t = \frac{b}{s^* \times SE}$$

multiple regression model

## Penalized two-sample t-statistic

$$t = \frac{\bar{M}_A - \bar{M}_B}{s^* \times \sqrt{1/n_A + 1/n_B}}, \quad \text{where } s^* = \sqrt{a + s^2}$$

## Robust General Penalized t-statistic

# 單因子變異數分析 (One-Way ANOVA)

- ANOVA can be considered to be a generalization of the t-test, when
  - compare more than two groups (e.g., drug 1, drug 2, and placebo), or
  - compare groups created by more than one independent variable while controlling for the separate influence of each of them (e.g., Gender, type of Drug, and size of Dose).

- One-way ANOVA compares groups using one parameter.

- Assumptions
  - The subjects are sampled randomly.
  - The groups are independent.
  - The population variances are homogenous.
  - The population distribution is normal in shape.

- As with t-tests, violation of homogeneity is particularly a problem when we have quite different sample sizes.

# ANOVA Table

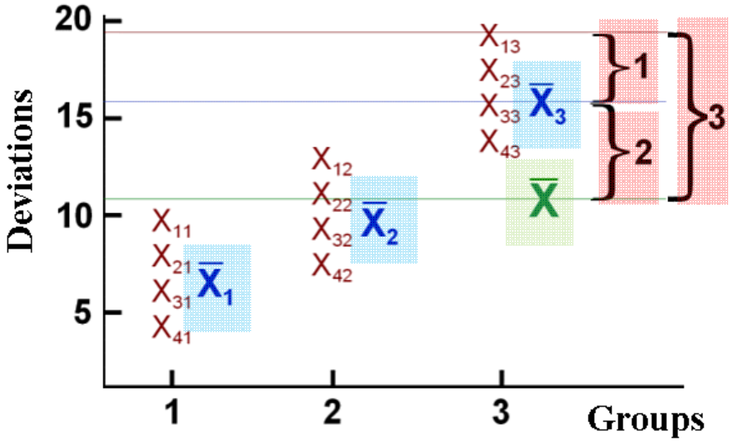
**Groups**

| 1 | 2 | … j … | k |
|---|---|---|---|
| $X_{11}$ | $X_{12}$ | $\cdots\ X_{1j}\ \cdots$ | $X_{1k}$ |
| $X_{21}$ | $X_{22}$ | $\cdots\ X_{2j}\ \cdots$ | $X_{2k}$ |
| | | $\cdots$ | |
| $X_{i1}$ | $X_{i2}$ | $\cdots\ X_{ij}\ \cdots$ | $X_{ik}$ |
| $\vdots$ | | | $X_{n_k k}$ |
| | $X_{n_2 2}$ | $\cdots\ \vdots\ \cdots$ | |
| $X_{n_1 1}$ | | $X_{n_j j}$ | |

$$T_j = \sum_{i=1}^{n_j} X_{ij} \qquad \bar{X}_j = \frac{T_j}{n_j}$$

$$T = \sum_{j=1}^{k} T_j \qquad \bar{X} = \frac{T}{N}$$

$$S^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j} \frac{(X_{ij} - \bar{X})^2}{N-1}$$



$$(X_{ij} - \bar{X}) = (X_{ij} - \bar{X}_j) + (\bar{X}_j - \bar{X})$$

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

$$X_{ij} = \mu_j + \epsilon_{ij} \qquad i = 1, \cdots, n_j$$

$$j = 1, \cdots, k$$

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}[(X_{ij} - \bar{X}_j) + (\bar{X}_j - \bar{X})]^2$$

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_j)^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{X}_j - \bar{X})^2$$

## ANOVA Table

| Source | SS | df | MS | F | p |
|---|---|---|---|---|---|
| Between | $SS_B$ | $p-1$ | $MS_B$ | $MS_B/MS_W$ | $< 0.05$ |
| Within | $SS_W$ | $N-p$ | $MS_W$ | | |
| Total | $SS_T$ | $N-1$ | | | |

$$SS_{Total} = SS_{Within} + SS_{Between}$$

$$F = \frac{MS_{Between}}{MS_{Within}}$$

Reject $H_0$, if $F_{obs} > F_{\{\alpha,k-1,N-k\}}$

# Welch ANOVA

## Welch's F Test

- Use when the sample sizes are unequal.
- Use when the sample sizes are equal but small.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

$$X_{ij} = \mu_j + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N(0, \sigma_j^2)$$

$$i = 1, \cdots, n_j$$

$$j = 1, \cdots, k$$

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2}{n_j - 1}$$

$$w_j = \frac{n_j}{s_j^2}$$

$$\bar{X}' = \frac{\sum_{j=1}^{k} w_j \bar{X}_j}{\sum_{j=1}^{k} w_j}$$

$$F' = \frac{\frac{\sum_{j=1}^{k} w_j (\bar{X}_j - \bar{X}')^2}{k-1}}{1 + \frac{2(k-2)}{k^2-1} \sum_{j=1}^{k} (\frac{1}{n_j - 1})(1 - \frac{w_j}{\sum_{j=1}^{k} w_j})^2}$$

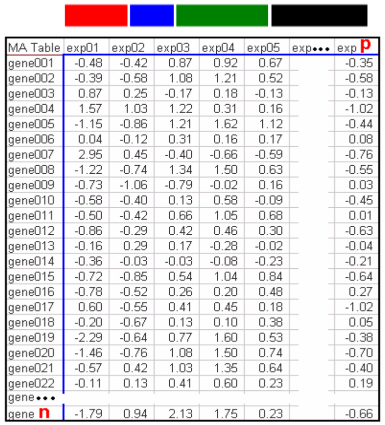
$$df' = \frac{k^2 - 1}{3 \sum_{j=1}^{k} (\frac{1}{n_j - 1})(1 - \frac{w_j}{\sum_{j=1}^{k} w_j})^2}$$

Reject $H_0$, if $F'_{obs} > F_{\{\alpha, k-1, df'\}}$

# 兒童小圓藍細胞腫瘤
# Small Round Blue Cell Tumors (SRBCT) Dataset

## *cDNA Microarrays*

- ***#Samples:*** 63
  four types of SRBCT of childhood:
  - Neuroblastoma (NB) (12),
  - Non-Hodgkin lymphoma (NHL) (8),
  - Rhabdomyosarcoma (RMS) (20)
  - Ewing tumours (EWS) (23).
- ***#Genes***: 6567 genes



| MA Table | exp01 | exp02 | exp03 | exp04 | exp05 | exp••• | exp p |
|---|---|---|---|---|---|---|---|
| gene001 | -0.48 | -0.42 | 0.87 | 0.92 | 0.67 | | -0.35 |
| gene002 | -0.39 | -0.58 | 1.08 | 1.21 | 0.52 | | -0.58 |
| gene003 | 0.87 | 0.25 | -0.17 | 0.18 | -0.13 | | -0.13 |
| gene004 | 1.57 | 1.03 | 1.22 | 0.31 | 0.16 | | -1.02 |
| gene005 | -1.15 | -0.86 | 1.21 | 1.62 | 1.12 | | -0.44 |
| gene006 | 0.04 | -0.12 | 0.31 | 0.16 | 0.17 | | 0.08 |
| gene007 | 2.95 | 0.45 | -0.40 | -0.66 | -0.59 | | -0.76 |
| gene008 | -1.22 | -0.74 | 1.34 | 1.50 | 0.63 | | -0.55 |
| gene009 | -0.73 | -1.06 | -0.79 | -0.02 | 0.16 | | 0.03 |
| gene010 | -0.58 | -0.40 | 0.13 | 0.58 | -0.09 | | -0.45 |
| gene011 | -0.50 | -0.42 | 0.66 | 1.05 | 0.68 | | 0.01 |
| gene012 | -0.86 | -0.29 | 0.42 | 0.46 | 0.30 | | -0.63 |
| gene013 | -0.16 | 0.29 | 0.17 | -0.28 | -0.02 | | -0.04 |
| gene014 | -0.36 | -0.03 | -0.03 | -0.08 | -0.23 | | -0.21 |
| gene015 | -0.72 | -0.85 | 0.54 | 1.04 | 0.84 | | -0.64 |
| gene016 | -0.78 | -0.52 | 0.26 | 0.20 | 0.48 | | 0.27 |
| gene017 | 0.60 | -0.55 | 0.41 | 0.45 | 0.18 | | -1.02 |
| gene018 | -0.20 | -0.67 | 0.13 | 0.10 | 0.38 | | 0.05 |
| gene019 | -2.29 | -0.64 | 0.77 | 1.60 | 0.53 | | -0.38 |
| gene020 | -1.46 | -0.76 | 1.08 | 1.50 | 0.74 | | -0.70 |
| gene021 | -0.57 | 0.42 | 1.03 | 1.35 | 0.64 | | -0.40 |
| gene022 | -0.11 | 0.13 | 0.41 | 0.60 | 0.23 | | 0.19 |
| gene••• | | | | | | | |
| gene n | -1.79 | 0.94 | 2.13 | 1.75 | 0.23 | | -0.66 |

6567 x 63

## *Interests:*

- To identify genes that are differentially expressed in one or more of these four groups.

*More on SRBCT:*
http://www.thedoctorsdoctor.com/diseases/small_round_blue_cell_tumor.htm

Khan J, Wei J, Ringner M, Saal L, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu C, Peterson C and Meltzer P. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine 2001, 7:673-679
Stanford Microarray Database

# Apply ANOVA to SRBCT data

- `khan {made4}`: Microarray gene expression dataset from Khan et al., 2001. Subset of 306 genes.
- http://svitsrv25.epfl.ch/R-doc/library/made4/html/khan.html
- Khan contains gene expression profiles of four types of small round blue cell tumours of childhood (SRBCT) published by Khan et al. (2001). It also contains further gene annotation retrieved from SOURCE at http://source.stanford.edu/.

```
if (!requireNamespace("BiocManager", quietly = TRUE))
install.packages("BiocManager")
BiocManager::install("made4")
```

```
> library(made4)
> data(khan)
> # some EDA works should be done before ANOVA
>
> # get the p-value from a anova table
> Anova.pvalues <- function(x){
+   x <- unlist(x)
+   SRBCT.aov.obj <- aov(x ~ khan$train.classes)
+   SRBCT.aov.info <- unlist(summary(SRBCT.aov.obj))
+   SRBCT.aov.info["Pr(>F)1"]
+ }
> # perform anova for each gene
> SRBCT.aov.p <- apply(khan$train, 1, Anova.pvalues)
```

# Apply ANOVA to SRBCT data

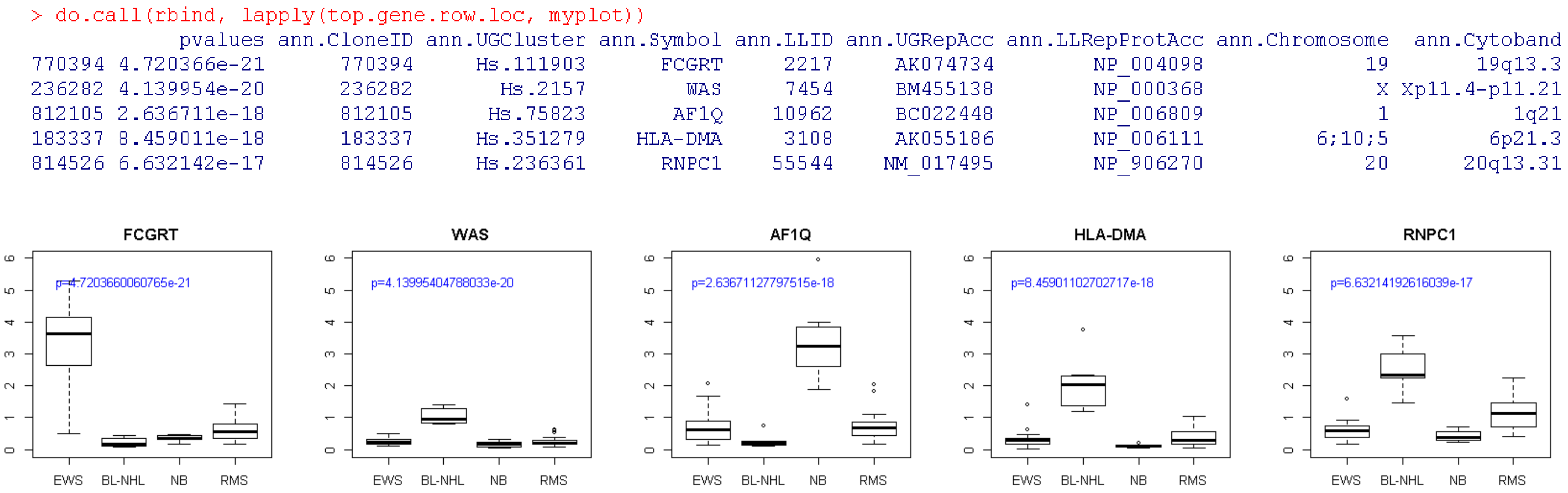
```
> # select the top 5 DE genes
> order.p <- order(SRBCT.aov.p)
> ranked.genes <- data.frame(pvalues=SRBCT.aov.p[order.p],
+                            ann=khan$annotation[order.p, ])
> top5.gene.row.loc <- rownames(ranked.genes[1:5,  ])
> # summarize the top5 genes
> summary(t(khan$train[top5.gene.row.loc, ]))
     770394           236282           812105           183337           814526      
 Min.   :0.0669   Min.   :0.0364   Min.   :0.1011   Min.   :0.0223   Min.   :0.1804  
 1st Qu.:0.3370   1st Qu.:0.1557   1st Qu.:0.3250   1st Qu.:0.1273   1st Qu.:0.4294  
 Median :0.6057   Median :0.2412   Median :0.7183   Median :0.2701   Median :0.6677  
 Mean   :1.5508   Mean   :0.3398   Mean   :1.1619   Mean   :0.5013   Mean   :0.9640  
 3rd Qu.:2.8176   3rd Qu.:0.3563   3rd Qu.:1.5543   3rd Qu.:0.5104   3rd Qu.:1.3620  
 Max.   :5.2958   Max.   :1.3896   Max.   :5.9451   Max.   :3.7478   Max.   :3.5809  
```

```
> # draw the side-by-side boxplot for top5 DE genes
> par(mfrow=c(1, 5), mai=c(0.3, 0.4, 0.3, 0.3))
> # get the location of xleft, xright, ybottom, ytop.
> usr <- par("usr")
> myplot <- function(gene){
+   # use unlist to convert "data.frame[1xp]" to "numeric"
+   boxplot(unlist(khan$train[gene, ]) ~ khan$train.classes,
+           ylim=c(0, 6), main=ranked.genes[gene, 4])
+   text(2, usr[4]-1, labels=paste("p=", ranked.genes[gene, 1],
+        sep=""), col="blue")
+   ranked.genes[gene,]
+ }
```

(重要技巧) 利用Key (gene.row.loc)
去連結多組資料(train, annotation)。

# Apply ANOVA to SRBCT data

```
> # print the top5 DE genes info
> do.call(rbind, lapply(top5.gene.row.loc, myplot))
```



```
> do.call(rbind, lapply(top.gene.row.loc, myplot))
             pvalues ann.CloneID ann.UGCluster ann.Symbol ann.LLID ann.UGRepAcc ann.LLRepProtAcc ann.Chromosome  ann.Cytoband
770394 4.720366e-21      770394     Hs.111903      FCGRT     2217     AK074734        NP_004098             19      19q13.3
236282 4.139954e-20      236282       Hs.2157        WAS     7454     BM455138        NP_000368              X Xp11.4-p11.21
812105 2.636711e-18      812105      Hs.75823       AF1Q    10962     BC022448        NP_006809              1          1q21
183337 8.459011e-18      183337     Hs.351279    HLA-DMA     3108     AK055186        NP_006111         6;10;5        6p21.3
814526 6.632142e-17      814526     Hs.236361      RNPC1    55544    NM_017495        NP_906270             20      20q13.31
```

# 卡方檢定: chisq.test

## 卡方檢定: chisq.test

- **適合度檢定**(test of goodness of fit): 檢定資料是否符合某個比例關係或某個機率分佈。
- **齊一性檢定**(test of homogeneity): 檢定幾個不同類別中的比例關係是否一致。
- **獨立性檢定**(test of independence): 檢定兩個分類變數之間是否互相獨立。

**chisq.test {stats}**: Pearson's Chi-squared Test for Count Data

**Description**:

chisq.test performs chi-squared contingency table tests and goodness-of-fit tests.

**Usage**:

```
chisq.test(x, y = NULL, correct = TRUE, p =
rep(1/length(x), length(x)), rescale.p = FALSE,
simulate.p.value = FALSE, B = 2000)
```

# Chi-Square Test for Independence

$H_0$: In the population, the two categorical variables are **independent.**

For testing independence in $I \times J$ contingency tables

$$H_0: \pi_{ij} = \pi_{i+}\pi_{+j} \quad \text{for all } i \text{ and } j$$

$\mu_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$ as the expected frequency.

*estimated expected frequencies.*

$$\hat{\mu}_{ij} = np_{i+}p_{+j} = n\left(\frac{n_{i+}}{n}\right)\left(\frac{n_{+j}}{n}\right) = \frac{n_{i+}n_{+j}}{n}$$

The *Pearson chi-squared statistic* for testing $H_0$ is

$$X^2 = \sum \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}$$

The $X^2$ statistic has approximately a chi-squared distribution, for large $n$. **(WHY?)**

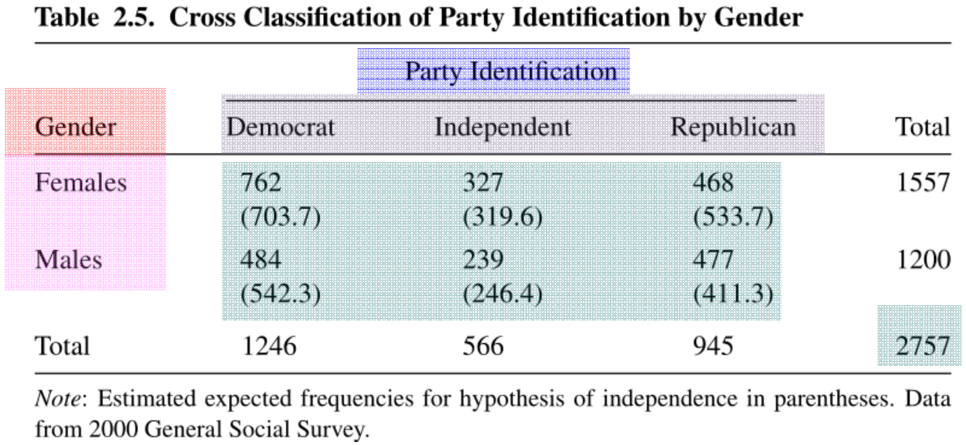


**Table 2.5. Cross Classification of Party Identification by Gender**

| | Party Identification | | | |
|---|---|---|---|---|
| Gender | Democrat | Independent | Republican | Total |
| Females | 762<br>(703.7) | 327<br>(319.6) | 468<br>(533.7) | 1557 |
| Males | 484<br>(542.3) | 239<br>(246.4) | 477<br>(411.3) | 1200 |
| Total | 1246 | 566 | 945 | 2757 |

*Note*: Estimated expected frequencies for hypothesis of independence in parentheses. Data from 2000 General Social Survey.

```
> M <- as.table(rbind(c(762, 327, 468),
                      c(484, 239, 477)))
> dimnames(M) <- list(gender = c("F", "M"),
+                     party = c("Democrat",
                                "Independent",
                                "Republican"))
> M
      party
gender Democrat Independent Republican
     F      762         327        468
     M      484         239        477
> (res <- chisq.test(M))
        Pearson's Chi-squared test

data:  M
X-squared = 30.07, df = 2, p-value = 2.954e-07
```