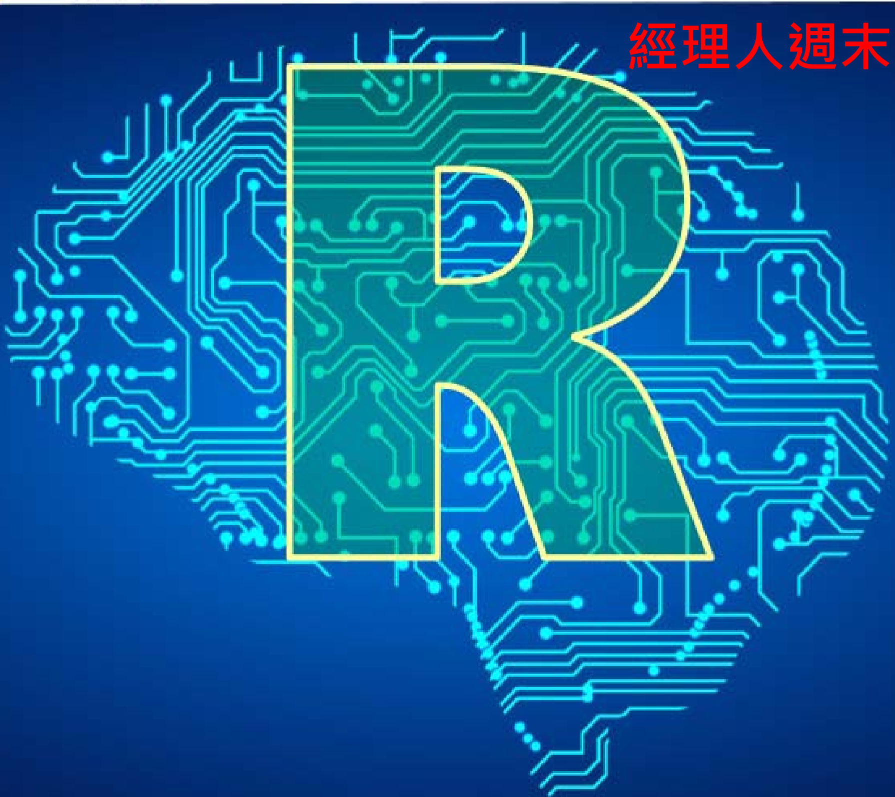
M2

台灣人工智慧學校

經理人週末研修班



參數估計
與假設檢定

吳漢銘

國立臺北大學 統計學系

http://www.hmwu.idv.tw

參數估計與假設檢定 - 大綱

- **主題1**
  - **點估計** (動差法、最大概似法、最小平方法)
    - 評斷準則: 不偏性、有效性、一致性、最小變異不偏性、充份性。
  - **區間估計**
- **主題2**
  - **貝式定理**
  - **貝式估計法**
- **主題3**
  - 統計假設檢定 (Hypothesis Testing)
  - 平均數檢定 (t檢定)
- **主題4**
  - 單因子變異數分析 (One-way Analysis of Variance, ANOVA)
  - R程式範例
- 主題5 [進階選讀]
  - Non-parametric Models
  - Non-parametric TestsL: Wilcoxon Signed-Rank Test (paired)
  - 事後比較檢定 (Post Hoc Tests): Tukey's HSD Test
- **主題6**
  - **常態分佈檢定 (Test for Normality)**
  - **卡方檢定 (Chi-Square Test)**

參數估計 (parameter estimation)
(利用樣本統計量及其抽樣分配來對母體參數進行推估, 以瞭解母體的特性)

概似函數 (The Likelihood Function)

1. Suppose the sample are iid from a distribution with density function $f(X|\theta)$, where θ is a parameter.

2. The **likelihood function** is the conditional probability of observing the sample , given θ

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta) .$$

   (a) The parameter could be a vector of parameters, $\theta = (\theta_1, \cdots, \theta_p)$.

   (b) The likelihood function regards the data as a function of the parameter θ .

   (c) The **log likelihood** function

$$l(\theta) = \log(L(\theta)) = \sum_{i=1}^{n} \log f(x_i|\theta) .$$

最大概似估計法
Maximum Likelihood Estimation (MLE)

The method of maximum likelihood was introduced by **R.A. Fisher** (1890-1962, English statistician).

(a) By maximizing the likelihood function $L(\theta)$ with respect to θ, we are looking for the most likely value of θ given the sample data.

(b) Θ: parameter space of possible values of θ.

(c) If the $\max L(\theta)$ exists and it occurs at a unique point $\hat{\theta} \in \Theta$, then $\hat{\theta}$ is called maximum likelihood estimator of θ.

$$\frac{\partial L(\theta)}{\partial \theta} = 0 \quad 且 \quad \frac{\partial^2 L(\theta)}{\partial \theta^2} < 0$$

點估計步驟：

1. 抽取代表性樣本
2. 選擇一個較佳的樣本統計量當估計式
3. 計算估計式的估計值
4. 以該估計值推論母體參數並作決策

MLE of (μ, σ^2) from a normal population

$$X_1, \ldots, X_n \sim \text{i.i.d. } N(\mu, \sigma^2). \qquad f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The probability density function for a sample of *n* independent identically distributed (iid) normal random variables (the likelihood) is

$$f(x_1, \ldots, x_n \mid \mu, \sigma^2) = \prod_{i=1}^{n} f(x_i \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right),$$

$$\mathcal{L}(\mu, \sigma) = f(x_1, \ldots, x_n \mid \mu, \sigma)$$

$$\log(\mathcal{L}(\mu, \sigma)) = (-n/2)\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$0 = \frac{\partial}{\partial\mu}\log(\mathcal{L}(\mu, \sigma)) = 0 - \frac{-2n(\bar{x}-\mu)}{2\sigma^2}. \quad \Rightarrow \quad \boxed{\hat{\mu} = \bar{x} = \sum_{i=1}^{n}\frac{x_i}{n}.} \qquad E\left[\widehat{\mu}\right] = \mu$$

https://en.wikipedia.org/wiki/Maximum_likelihood_estimation

MLE of (μ, σ^2) from a normal population

$$0 = \frac{\partial}{\partial\sigma} \log\left(\left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2}{2\sigma^2}\right)\right)$$

$$= \frac{\partial}{\partial\sigma}\left(\frac{n}{2}\log\left(\frac{1}{2\pi\sigma^2}\right) - \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2}{2\sigma^2}\right)$$

$$= -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2}{\sigma^3}$$

$$\boxed{E\left[\widehat{\sigma}^2\right] = \frac{n-1}{n}\sigma^2.}$$

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2. \qquad \mu = \widehat{\mu} \quad \Rightarrow \quad \boxed{\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

The maximum likelihood estimator (MLE) for $\theta = (\mu, \sigma^2)$ is

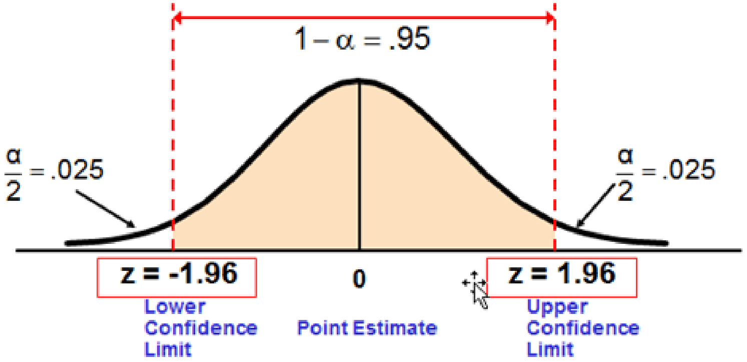
$$\boxed{\hat{\mu} = \bar{x} = \sum_{i=1}^{n}\frac{x_i}{n}.} \qquad \boxed{\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

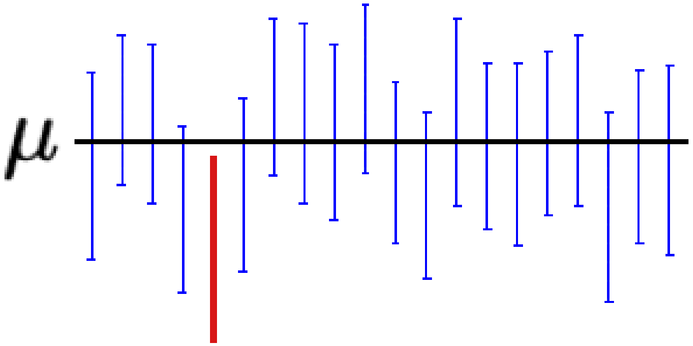
區間估計 (Interval Estimation)

- 區間估計是先對未知的母體參數求**點估計值**，然後在一信賴水準 (Confidence Level) 下，導出一個上下區間，此區間稱為信賴區間 (Confidence Interval)，信賴水準是指該區間**包含**母體參數的可靠度。
- 95% 信賴區間表示，做100 次信賴區間，區間約包含母體參數約95 次。

Interval Estimate of Population Mean

$$\bar{X} \sim N(\mu, \sigma^2/n) \Rightarrow Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \Rightarrow P(-z \le Z \le z) = 1 - \alpha = 0.95.$$





$$0.95 = P\left(-1.96 \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le 1.96\right)$$

$$= P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right).$$

A 95% confidence interval indicates that 19 out of 20 samples (95%) from the same population will produce confidence intervals that contain the population parameter.

範例: 老年人看電視的時間

根據行政院主計處調查，台灣地區15歲以上的人口中，以老年人(65歲以上)看電視的時間最長。現在新立傳播公司計畫推出老年人的電視節目，因此想要了解老年人看電視的時間，以決定電視節目的數量。新立公司於是採隨機抽樣法**抽取台北市100位老人**調查看電視的時數，結果得知，每星期看電視的**平均時間為 21.2小時**。假設根據過去數次調查的資料，已知每星期看電視時間的**標準差為8小時**，問在**95%信賴水準**下，每星期**看電視平均時間的信賴區間**為何？

信賴水準為95%，$\bar{X}$=21.2小時，σ =8小時， n =100

$\bar{X}$ 的抽樣分配為常態分配 $N(\mu,\sigma_{\bar{X}}^2)$ ⇒ $P(|\bar{X}-\mu|\le 1.96\sigma_{\bar{X}})=0.95$

$$\sigma_{\bar{X}}=\frac{\sigma}{\sqrt{n}}=\frac{8}{\sqrt{100}}=0.8$$

在 $1-\alpha$ 信賴水準下，母體平均數的信賴區間為

$$\bar{X}\pm Z_{\alpha/2}\sigma_{\bar{X}}$$

$$\bar{X}\pm Z_{\alpha/2}\sigma_{\bar{X}}=21.2\pm 1.96\times 0.8 \Rightarrow 19.632\le\mu\le 22.768$$

可推論：「老年人每星期平均看電視的時間在 19.632~22.768小時之間，而此一區間的可信度(信賴水準)為95%。」

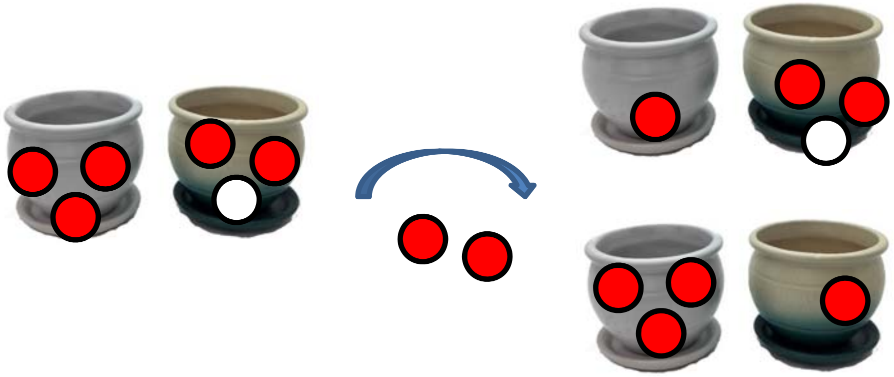
貝氏定理 (Bayes' Theorem)

$$P(A|B) = \frac{P(A\cap B)}{P(B)} = \frac{P(B|A)\times P(A)}{P(B)}$$

$$後驗機率 = \frac{可能性 \times 先驗機率}{標准化常量}$$

- $P(A|B)$: 已知在事件 B 發生的情況下事件 A 發生的機率。(稱作 A 的事後機率或後驗機率)(posterior probability)。
- $P(A)$, $P(B)$: A, B 的事前機率或先驗機率 (prior probability)。($P(A)\neq 0$, $P(B)\neq 0$)
- $P(B|A)$: 已知 A 發生後，B 的條件機率。(稱作概似函數 likelihood function)。

例子: 假設有兩個甕，第一個甕裡面有 3 顆紅球，第二個甕裡面有 2 顆紅球和 1 顆白球。我們隨機選擇一個甕，然後從中抽出 2 顆球。假設結果是 2 顆紅球，留在甕裡的那顆球是紅球的機率是多少？(https://ccjou.wordpress.com/)



樣本空間 $\Omega=\{r_1,r_2,r_3,r_4,r_5,w_1\}$。

令 $U_1=\{r_1,r_2,r_3\}$ 和 $U_2=\{r_4,r_5,w_1\}$。

A: 從一個甕中抽出 2 顆紅球之事件。

$$\begin{aligned} P(U_1|A) &= \frac{P(A|U_1)P(U_1)}{P(A|U_1)P(U_1)+P(A|U_2)P(U_2)} \\ &= \frac{1\cdot\frac{1}{2}}{1\cdot\frac{1}{2}+\frac{1}{3}\cdot\frac{1}{2}} = \frac{3}{4}. \end{aligned}$$

Bayesian Statistics
貝式統計

Bayes' Theorem

1. If A and B are events and $P(B) > 0$, then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

2. The distributional form of Bayes' Theorem for continuous random variables is

$$f_{X|Y=y}(x) = \frac{f_{Y|X=x}(y)f_X(x)}{f_Y(y)} = \frac{f_{Y|X=x}(y)f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X=x}(y)f_X(x)\ dx}$$

1. In the **frequentist approach** to statistics, the parameters of a distribution are considered to be fixed but unknown constants.

2. The **Bayesian approach** views the unknown parameters of a distribution as random variables.

   (a) In Bayesian analysis, probabilities can be computed for parameters as well as the sample statistics.

   (b) Bayes' Theorem allows one to revise the prior belief about an unknown parameter based on observed data.

Bayesian Statistics
貝式統計

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

3. Suppose that X has the density $f(x|\theta)$.

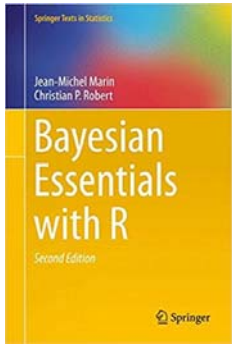
(a) $f_\theta(\theta)$: the pdf of the _prior distribution_ of θ.

(b) The conditional density of θ given the sample observations $x_1, \cdots, x_n$ is called the _posterior density_

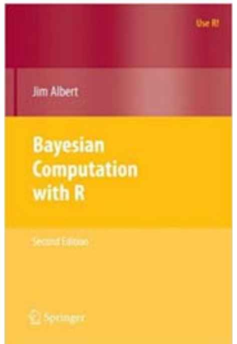
$$f_{\theta|x}(\theta) = \frac{f(x_1, \cdots, x_n|\theta) f_\theta(\theta)}{\int f(x_1, \cdots, x_n|\theta) f_\theta(\theta)\ d\theta}.$$

(c) The posterior distribution summarizes our modified belief about the unknown parameters, taking into account the observed data.

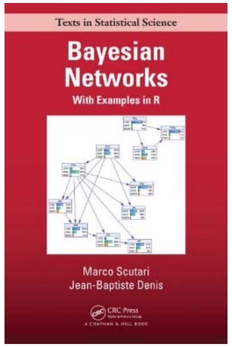
(d) One is interested in computing _posterior quantities_ such as posterior means, posterior modes, posterior standard deviations.



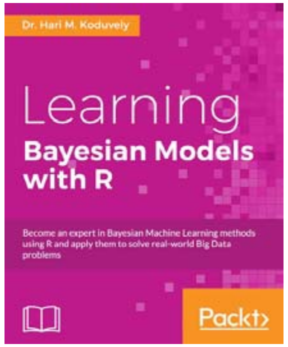
Springer Texts in Statistics
Jean-Michel Marin
Christian P. Robert
Bayesian Essentials with R
Second Edition
Springer



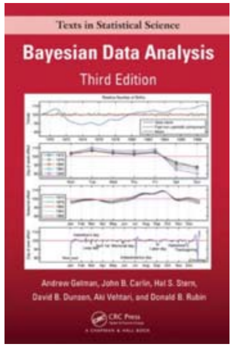
Use R!
Jim Albert
Bayesian Computation with R
Second Edition
Springer



Texts in Statistical Science
Bayesian Networks
With Examples in R
Marco Scutari
Jean-Baptiste Denis
CRC Press



Dr. Hari M. Koduvely
Learning Bayesian Models with R
Become an expert in Bayesian Machine Learning methods using R and apply them to solve real-world Big Data problems
Packt>



Texts in Statistical Science
Bayesian Data Analysis
Third Edition
Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin
CRC Press

Bayes Estimator for the Mean of a Normal Distribution

$X_1, X_2, \ldots, X_n$ be a random sample $\quad X_1, \ldots, X_n \sim$ i.i.d. $N(\mu, \sigma^2)$.

μ is unknown and σ^2 is known.

prior distribution for μ is normal with mean μ_0 and variance σ_0^2

$$f(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-(\mu-\mu_0)^2/(2\sigma_0^2)} = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-(\mu^2 - 2\mu_0 + \mu_0^2)/(2\sigma_0^2)}$$

The joint probability distribution of the sample

$$f(x_1, x_2, \ldots, x_n \mid \mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2\sigma^2)\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2\sigma^2)\left(\sum x_i^2 - 2\mu\sum x_i + n\mu^2\right)}$$

the joint probability distribution of the sample and μ is

$$f(x_1, x_2, \ldots, x_n, \mu) = \frac{1}{(2\pi\sigma^2)^{n/2}\sqrt{2\pi}\sigma_0} e^{-(1/2)\left[\left(1/\sigma_0^2 + n/\sigma^2\right)\mu^2 - \left(2\mu_0/\sigma_0^2 + 2\sum x_i/\sigma^2\right)\mu + \sum x_i^2\ /\sigma^2 + \mu_0^2/\sigma_0^2\ \right]}$$

$$= e^{-(1/2)\left[\left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n}\right)\mu^2 - 2\left(\frac{\mu_0}{\sigma_0^2} + \frac{\bar{x}}{\sigma^2/n}\right)\mu\right]} h_1(x_1, \ldots, x_n, \sigma^2, \mu_0, \sigma_0^2)$$

Bayes Estimator for the Mean of a Normal Distribution

$$f(x_1, x_2, \ldots, x_n, \mu) = e^{-(1/2)\left[\left(\frac{1}{\sigma_0^2}+\frac{1}{\sigma^2/n}\right)\mu^2 - 2\left(\frac{\mu_0}{\sigma_0^2}+\frac{\bar{x}}{\sigma^2/n}\right)\mu\right]} h_1(x_1, \ldots, x_n, \sigma^2, \mu_0, \sigma_0^2)$$

$$f\left(x_1, x_2, \ldots, x_n, \mu\right) = e^{-(1/2)\left(\frac{1}{\sigma_0^2}+\frac{1}{\sigma^2/n}\right)\left[\mu - \left(\frac{(\sigma^2/n)\,\mu_0}{\sigma_0^2+\sigma^2/n}+\frac{\bar{x}\sigma_0^2}{\sigma_0^2+\sigma^2/n}\right)\right]^2} h_2(x_1, \ldots, x_n, \sigma^2, \mu_0, \sigma_0^2)$$

$h_i(x_1, \ldots, x_n, \sigma^2, \mu_0, \sigma_0^2\ )$ is a function of the observed values and the parameters σ^2, μ_0, and σ_0^2.

because $f(x_1, \ldots, x_n)$ does not depend on μ,

$$f\left(\mu|x_1, \ldots, x_n\right) = e^{-(1/2)\left(\frac{1}{\sigma_0^2}+\frac{1}{\sigma^2/n}\right)\left[\mu - \left(\frac{\left(\sigma^2/n\right)\mu_0 + \sigma_0^2\bar{x}}{\sigma_0^2+\sigma^2/n}\right)\right]^2} h_3\left(x_1, \ldots, x_n, \sigma^2, \mu_0, \sigma_0^2\right)$$

a normal probability density function

posterior mean $\dfrac{\left(\sigma^2/n\right)\mu_0 + \sigma_0^2\,\bar{x}}{\sigma_0^2 + \sigma^2/n}$

posterior variance $\left(\dfrac{1}{\sigma_0^2} + \dfrac{1}{\sigma^2/n}\right)^{-1} = \dfrac{\sigma_0^2\left(\sigma^2/n\right)}{\sigma_0^2 + \sigma^2/n}$

假設檢定
Hypothesis Testing

Hypothesis Test

a procedure for determining if an assertion about a characteristic of a population is reasonable.

Example

"average price of a gallon of regular unleaded gas in Massachusetts is $2.5"

Is this statement true?

- find out every gas station.
- find out a small number of randomly chosen stations.

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

Sample average price was $2.2.

- Is this 30 cent difference a result of chance variability, or
- is the original assertion incorrect?

Hypothesis Testing

虛無假設 (Hull hypothesis):

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

擇一假設 (alternative hypothesis):

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

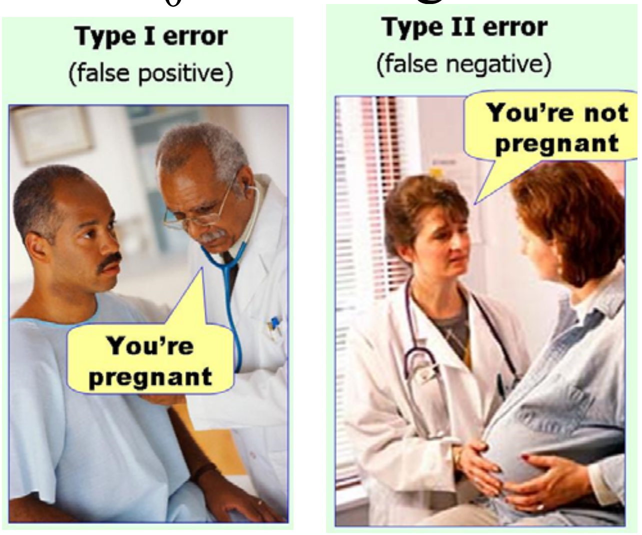
顯著顏準 (significance level)(alpha):

- Decide in advance.
- Alpha = 0.05: the probability of incorrectly rejecting the null hypothesis when it is actually true is 5%.

型一誤差、型二誤差

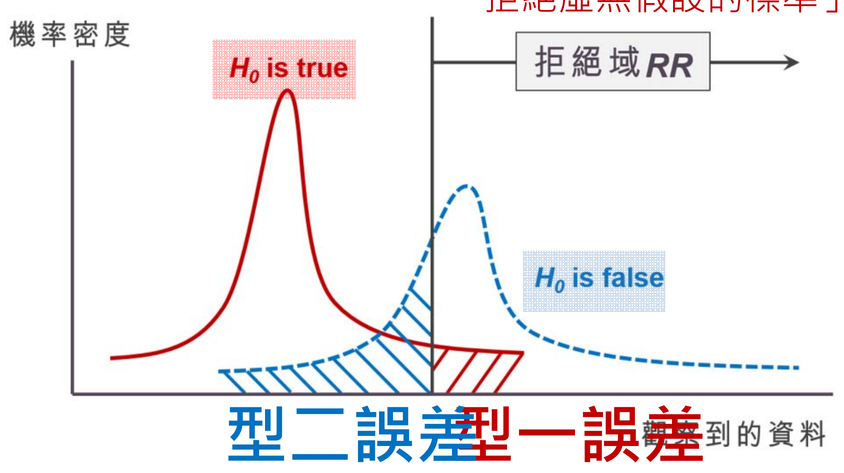
Hypothesis Testing		Truth	
		H_0	H_1
Decision	Reject H_0	Type I Error (α) (false positive)	Right Decision (true positive) Power = 1- β
	Fail to Reject H_0	Right Decision (true negative)	Type II Error (β) (false negative)

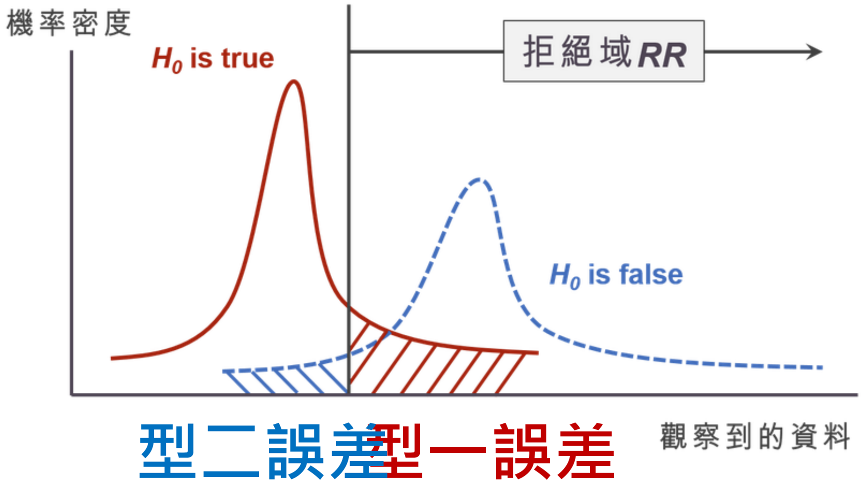
H_0: Not Pregnant



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

「拒絕虛無假設的標準」





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

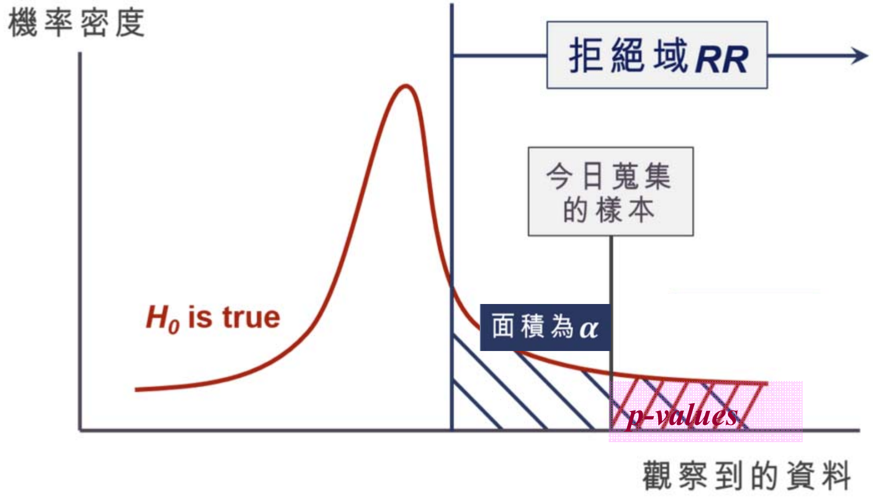
The *p*-values

p-values

- 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 $\mathbf{H_0}$ | $\mathbf{H_0}$ 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)。

Decision Rule

- Reject H_0 if *p-value* is less than alph
- $P < 0.05$ commonly used.
  (Reject $\mathbf{H_0}$, the test is significant)
- The lower the *p-value*, the more significant.



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

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

"**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。"

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**
	`t.test(x, mu = 0)`
		t-test
		`t.test(x-y, var.equal = TRUE)`
		`t.test(x, y, paired = TRUE, var.equal = TRUE)`
			t-test
			`t.test(x, y, var.equal = TRUE)`
				One-Way Analysis of Variance (ANOVA)
				`aov(x~g, data)`
				`oneway.test(x~g, data, var.equal = TRUE)`

Parametric (variance not equal)		**Welch t-test**
		`t.test(x-y)`
		`t.test(x, y, paired = TRUE)`
			Welch t-test
			`t.test(x, y)`
				Welch ANOVA
				`oneway.test(x~g, data)`

Non-Parametric (無母數檢定)	**Wilcoxon Signed-Rank Test**
	`wilcox.test(x, mu = 0)`
		Wilcoxon Signed-Rank Test
		`wilcox.test(x-y)`
		`wilcox.test(x, y, paired = TRUE)`
			Wilcoxon Rank-Sum Test (Mann-Whitney U Test)
			`wilcox.test(x, y)`
				Kruskal-Wallis Test
				`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

T檢定 (t-test)

One sample t-test

$H_0: \mu = \mu_0$
$H_1: \mu \neq \mu_0$ (two-tailed).
μ: population mean.
α: 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 μ: $\bar{X} - t_{\alpha/2} S/\sqrt{n} \leq \mu < \bar{X} + t_{\alpha/2} S/\sqrt{n}$
- p-value $= P_{H_0}(|\mathbf{T}| > t_0)$, $\mathbf{T} \sim t_{n-1}$.

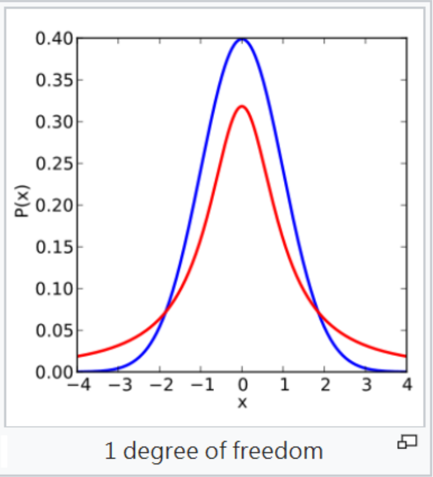
假設X是呈常態分布的獨立的隨機變量（隨機變量的期望值是μ，方差是σ^2但未知）。

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

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

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

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

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

Assumptions of t-test

Be Normal

- the distribution of the data must be **normal**.
- *How to Detect Normality*
  - Plots: Histogram, Density Plot, QQplot,…
  - Test for Normality: Jarque-Bera test, Lilliefors test, Kolmogorov-Smirnov test.

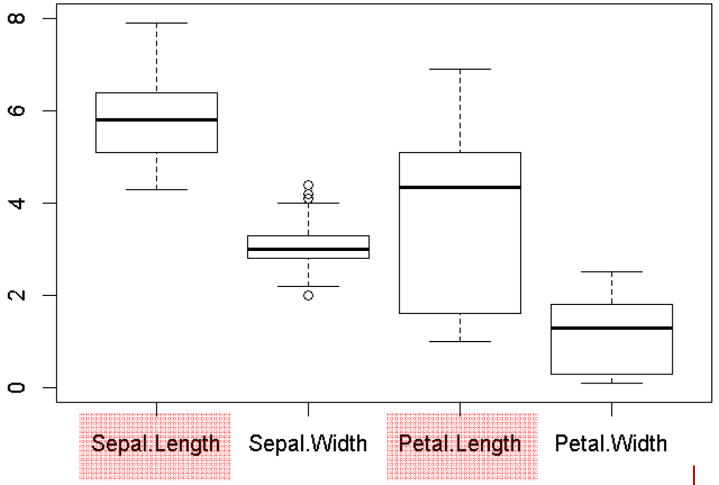
Homogeneous

- the variances of the two population are equal.
- Test for equality of the two variances: Variance ratio F-test.

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
```

Test Homogeneity of Variances

- `var.test {stats}`: an F test to compare the variances of two samples from normal populations.
- `bartlett.test {stats}`: a parametric test of the null that the variances in each of the groups (samples) are the same.
- `ansari.test {stats}`: Ansari-Bradley two-sample test for a difference in scale parameters. (testing for equal variance for non-normal samples)
- `mood.test {stats}`: another rank-based two-sample test for a difference in scale parameters.
- `fligner.test {stats}`: Fligner-Killeen (median) is a rank-based (nonparametric) k-sample test for homogeneity of variances.
- `leveneTest {car}`: Levene's test for homeogeneity of variance across groups.

- NOTE: Fligner-Killeen's and Levene's tests are two ways to test the ANOVA assumption of "equal variances in the population" before conducting the ANOVA test.
- Levene's is widely used and is typically the default in SPSS.

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

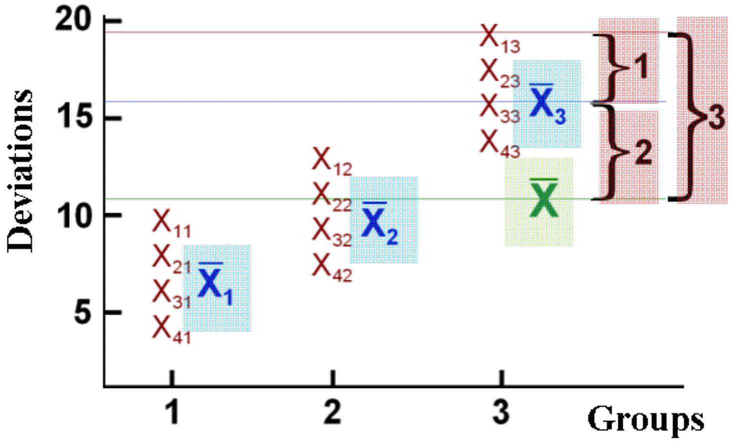
Groups

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	$N-1$			

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

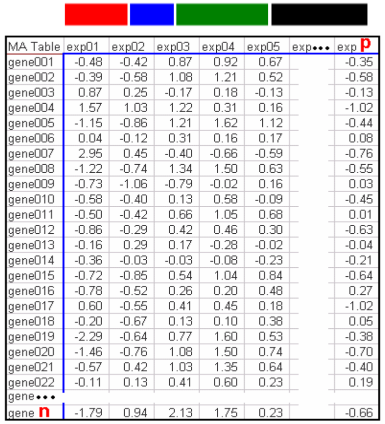
$$F = \frac{MS_{Between}}{MS_{Within}}$$

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

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/.

```
> source("https://bioconductor.org/biocLite.R")
> biocLite("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

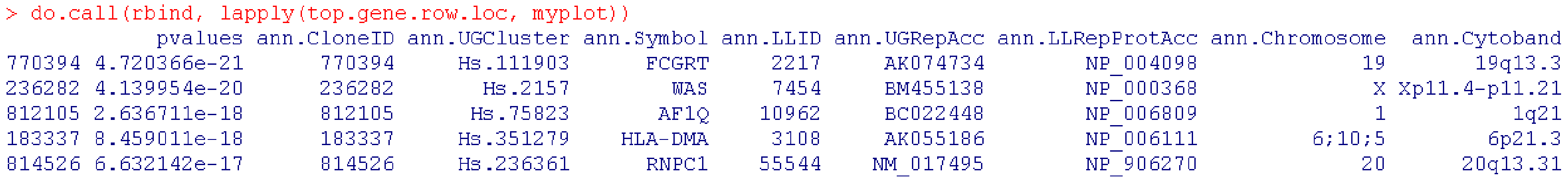
```
> # select the top 5 DE genes
> order.p <- order(SRBCT.aov.p)
> ranked.genes <- data.frame(pvalues=SRBCT.aov.p[order.p],
+                            ann=khan$annotation[order.p, ])
> top5.gene.row.loc <- rownames(ranked.genes[1:5,  ])
> # summarize the top5 genes
> summary(t(khan$train[top5.gene.row.loc, ]))
     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,]
+ }
```

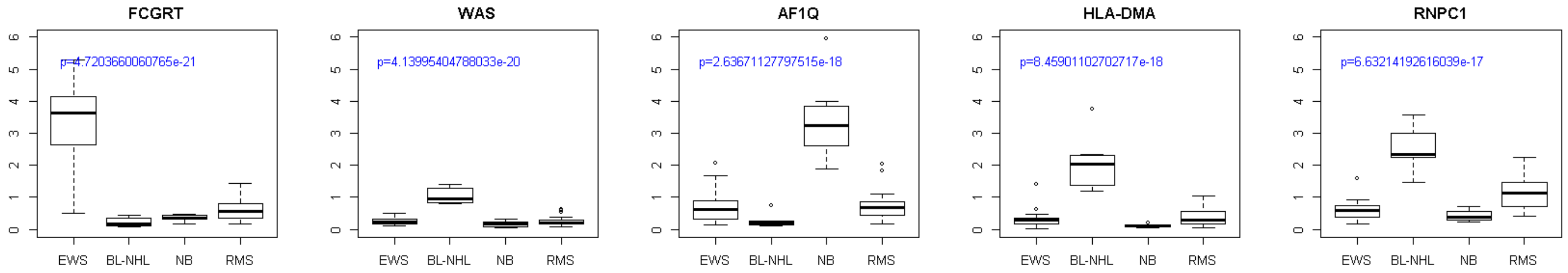
(重要技巧) 利用Key (gene.row.loc)
去連結多組資料(train, annotation)。

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
```

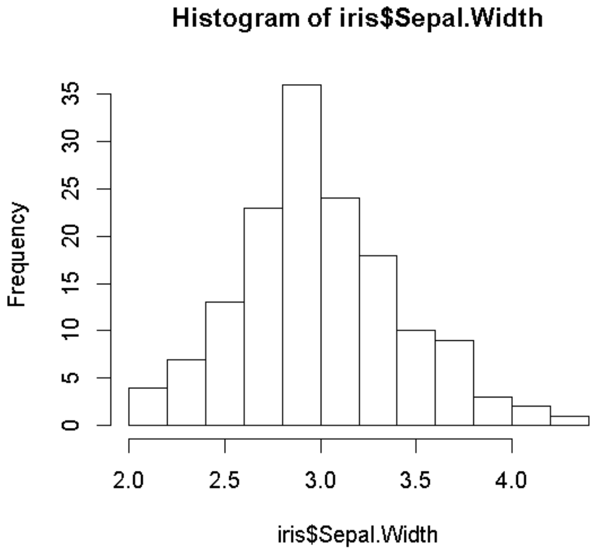


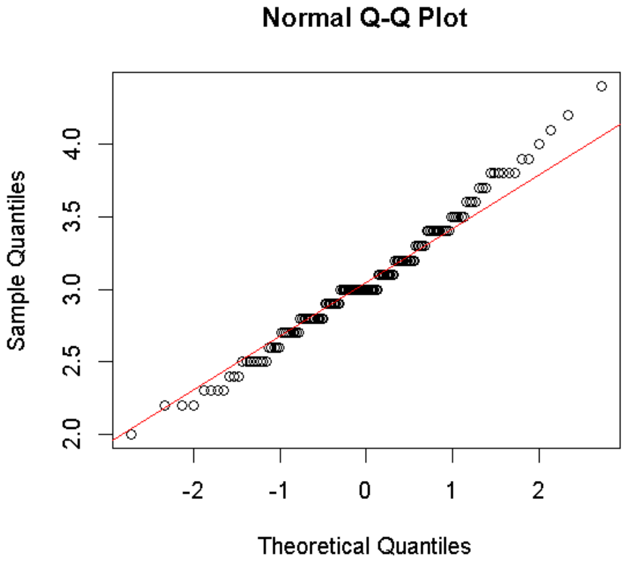
Formal Tests for Normality

H_0: The sample data are not significantly different than a normal population.
H_a: The sample data are significantly different than a normal population

```
hist(iris$Sepal.Width)
```

```
qqnorm(iris$Sepal.Width)
qqline(iris$Sepal.Width, col="red")
```





nortest Packages: Tests for Normality

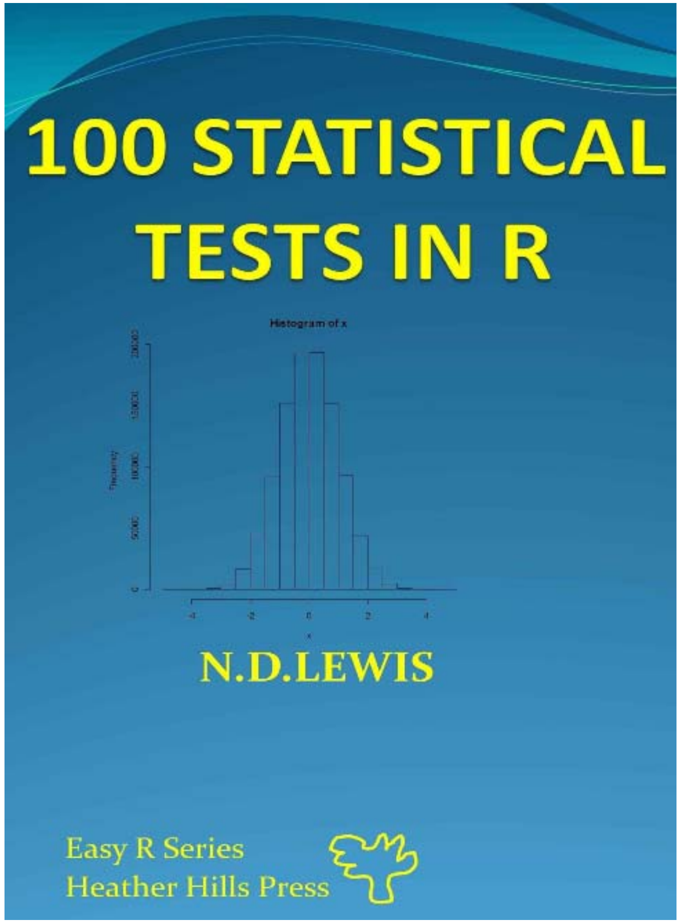
- nortest Packages: five omnibus tests for testing the composite hypothesis of normality: ad.test, cvm.test, lillie.test, pearson.test, sf.test
- Other tests:
  - Kolmogorov-Smirnov (K-S) test (Chakravarti et al., 1967).
  - The Shapiro-Wilk normality test (Shapiro and Wilk, 1965).

```
> library(nortest)
> ad.test(iris$Sepal.Width)

        Anderson-Darling normality test

data:  iris$Sepal.Width
A = 0.90796, p-value = 0.02023
```

```
> x <- iris$Sepal.Width
> ks.test(x, 'pnorm', mean(x), sd(x))

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.10566, p-value = 0.07023
alternative hypothesis: two-sided

Warning message:
In ks.test(x, "pnorm", mean(x), sd(x)) :
  ties should not be present for the Kolmogorov-Smirnov test
```

```
> shapiro.test(iris$Sepal.Width)

        Shapiro-Wilk normality test

data:  iris$Sepal.Width
W = 0.98492, p-value = 0.1012
```

卡方檢定: chisq.test



100 STATISTICAL TESTS IN R

N.D.LEWIS

Easy R Series
Heather Hills Press

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

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

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

$\mu_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$ 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?)**

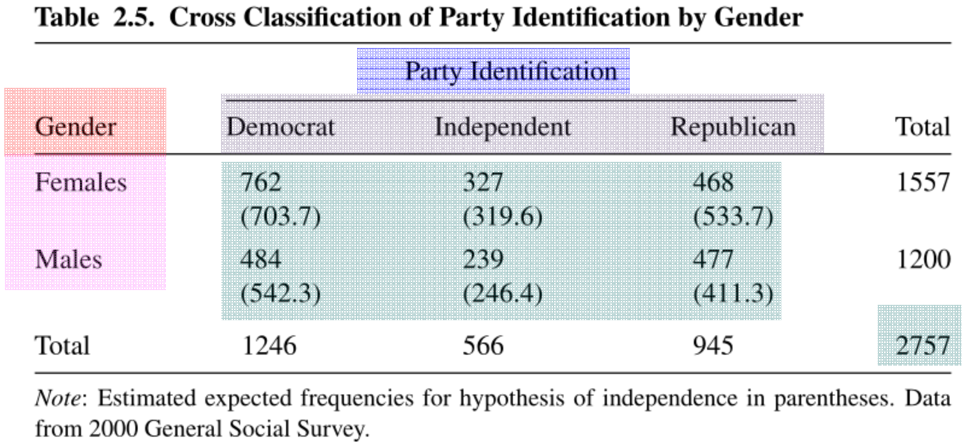


Table 2.5. Cross Classification of Party Identification by Gender

	Party Identification			
Gender	Democrat	Independent	Republican	Total
Females	762 (703.7)	327 (319.6)	468 (533.7)	1557
Males	484 (542.3)	239 (246.4)	477 (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
```

進階選讀

無母數統計
Non-parametric Statistics

- Nonparametric statistics is based on either
  - being distribution-free or having a specified distribution but with the distribution's parameters unspecified.
  - includes both descriptive statistics and statistical inference.
- **Non-parametric inferential statistical methods**: Sign test, Wilcoxon signed-rank test, Mann–Whitney U test, Kolmogorov–Smirnov test, Kruskal–Wallis one-way ANOVA,...
- **Non-parametric models**: kernel density estimation, non-parametric regression, ...

kernel regression

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)}$$

nonparametric regression

$$y_i = f_0(x_i) + \epsilon_i, \quad i = 1, \ldots n,$$

$\epsilon_1, \ldots \epsilon$ are still i.i.d. random errors with $\mathbb{E}(\epsilon_i) = 0$

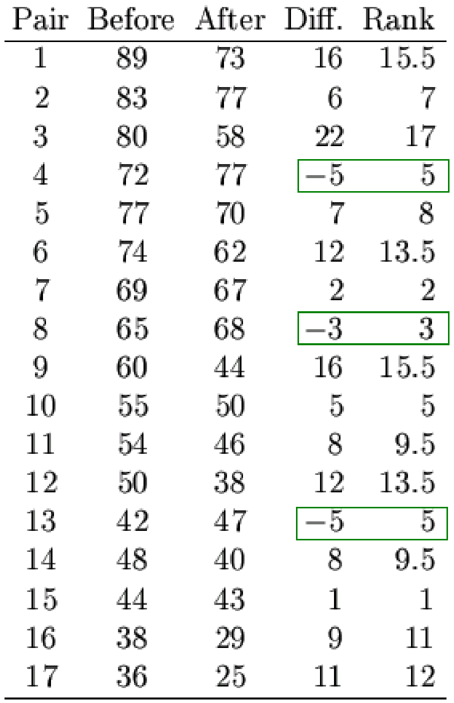
k-nearest-neighbors regression.

$$\hat{f}(x) = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x)} y_i$$

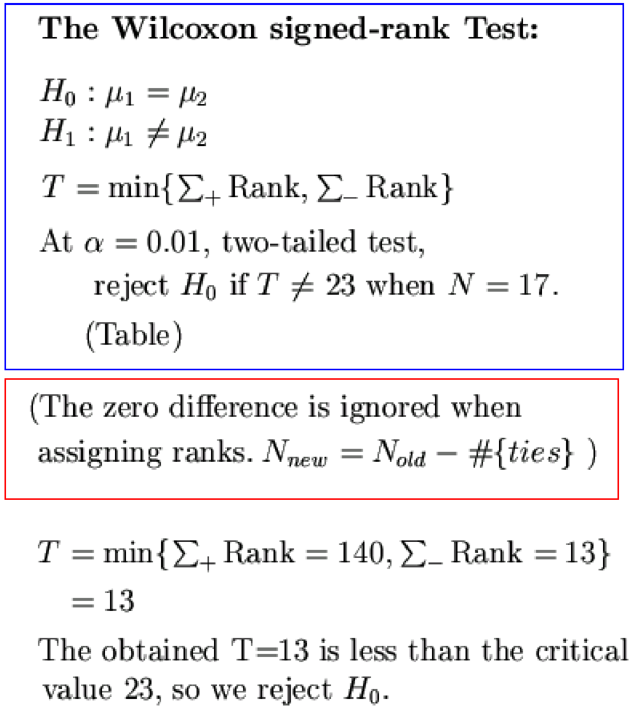
https://en.wikipedia.org/wiki/Nonparametric_statistics

Wilcoxon Signed-Rank Test (paired)

- Null hypothesis: the **population median** from which both samples were drawn is the same.
- The sum of the ranks for the "positive" values is calculated and compared against a precomputed table to a p-value.
- If the null hypothesis is true, the sum of the ranks of the positive differences should be about the same as the sum of the ranks of the negative differences.



Pair	Before	After	Diff.	Rank
1	89	73	16	15.5
2	83	77	6	7
3	80	58	22	17
4	72	77	−5	5
5	77	70	7	8
6	74	62	12	13.5
7	69	67	2	2
8	65	68	−3	3
9	60	44	16	15.5
10	55	50	5	5
11	54	46	8	9.5
12	50	38	12	13.5
13	42	47	−5	5
14	48	40	8	9.5
15	44	43	1	1
16	38	29	9	11
17	36	25	11	12



The Wilcoxon signed-rank Test:

$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$

$T = \min\{\sum_+ \text{Rank}, \sum_- \text{Rank}\}$

At $\alpha = 0.01$, two-tailed test,
reject H_0 if $T \neq 23$ when $N = 17$.
(Table)

(The zero difference is ignored when assigning ranks. $N_{new} = N_{old} - \#\{ties\}$)

$$T = \min\{\textstyle\sum_+ \text{Rank} = 140, \sum_- \text{Rank} = 13\}$$
$$= 13$$

The obtained T=13 is less than the critical value 23, so we reject H_0.

Parametric vs. Non-Parametric Test

Parametric Tests

- Assume that the data follows a certain distribution (normal distribution).
- Assuming equal variances and unequal variances.
- More powerful.
- Widely Implemented.
- Not appropriate for data with outliers.

Non-Parametric Tests

- When certain assumptions about the underlying population are questionable (e.g. normality).
- Does not assume normal distribution
- No variance assumption
- Less powerful.
- Widely Implemented.
- Decrease effects of outliers (Robust)
- Not recommended if there is less than 5 replicates per group.

事後檢定 (Post Hoc Tests)
Tukey′s Honestly Significant Difference (HSD) Test

- **Null hypothesis**: all means being compared are from the same population (i.e. $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$)

$$q_s = \frac{Y_A - Y_B}{SE},$$

Y_A is the larger of the two means being compared,
Y_B is the smaller of the two means being compared, and
SE is the standard error of the sum of the means.

- This q_s value can then be compared to a *q value* from the studentized range distribution.
- If the q_s value is larger than the critical value q_α obtained from the distribution, the two means are said to be significantly different at level α , 0 ≤ α ≤ 1.

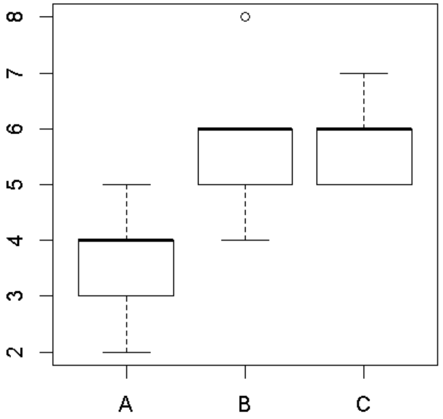
- **Assumptions for the test**
  - Observations are **independent** within and among groups.
  - The groups for each mean in the test are **normally distributed**.
  - **equal within-group variance** across the groups.
  - equal **sample sizes**.

範例: ANOVA + Post Hoc Test

- A drug company tested three formulations of a pain relief medicine for migraine headache sufferers. For the experiment 27 volunteers were selected and 9 were randomly assigned to one of three drug formulations. The subjects were instructed to take the drug during their next migraine headache episode and to report their pain on a scale of 1 to 10 (10 being most pain).

Drug A:	4	5	4	3	2	4	3	4	4
Drug B:	6	8	4	5	4	6	5	8	6
Drug C:	6	7	6	6	7	5	6	5	5

```
> pain <- c(4, 5, 4, 3, 2, 4, 3, 4, 4, 6, 8, 4, 5,
+  4, 6, 5, 8, 6, 6, 7, 6, 6, 7, 5, 6, 5, 5)
> drug <- c(rep("A", 9), rep("B", 9), rep("C", 9))
> migraine <- data.frame(pain, drug)
> plot(pain ~ drug, data=migraine)
> migraine.aov <- aov(pain ~ drug, data=migraine)
> summary(migraine.aov)
            Df Sum Sq Mean Sq F value   Pr(>F)
drug         2  28.22  14.111   11.91 0.000256 ***
Residuals   24  28.44   1.185
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # reject the null hypothesis of equal means for all three drug grou
```



```
> kruskal.test(pain ~ drug, data=migraine)
        Kruskal-Wallis rank sum test
data:  pain by drug
Kruskal-Wallis chi-squared = 14.395, df = 2, p-value = 0.0007483
```

Pairwise Comparisons

```
> pairwise.t.test(pain, drug, p.adjust="bonferroni")

        Pairwise comparisons using t tests with pooled SD

data:  pain and drug

  A       B
B 0.00119 -
C 0.00068 1.00000

P value adjustment method: bonferroni
>
> TukeyHSD(migraine.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = pain ~ drug, data = migraine)

$drug
         diff        lwr      upr     p adj
B-A 2.1111111  0.8295028 3.392719 0.0011107
C-A 2.2222222  0.9406139 3.503831 0.0006453
C-B 0.1111111 -1.1704972 1.392719 0.9745173
>
> # conclude that the mean pain is significantly different for drug A
```

Which Normality Test Should I Use?

- Kolmogorov-Smirnov test:
  - It is more sensitive near the center of the density than at the tails than other tests;
  - For data sets $n > 50$.
- The Anderson-Darling test:
  - A-D test is a modification of the K-S test and gives more weight to the tails of the density than does the K-S test.
  - It is generally preferable to the K-S test.
- Shapiro-Wilks test:
  - Doesn't work well if several values in the data set are the same.
  - Works best for data sets with $n < 50$, but can be used with larger data sets.
- W/S test (range(x)/sd(x)):
  - simple, but effective.
- Jarque-Bera test (`jarque.test {moments}`):
  - tests for skewness and kurtosis, very effective.
- D'Agostino test (`agostino.test{moments}`) :
  - powerful omnibus (skewness, kurtosis, centrality) test.

Which Normality Test Should I Use?

- Asghar Ghasemi and Saleh Zahediasl, Normality Tests for Statistical Analysis: **A Guide for Non-Statisticians**, *Int J Endocrinol Metab*. 2012 Spring; 10(2): 486–489.
    - assessing the normality assumption should be taken into account for using parametric statistical tests.
    - The KS test, should no longer be used owing to its low power.
    - It is preferable that normality be assessed both visually and through normality tests, of which the **Shapiro-Wilk test** is highly recommended.
- NOTE:
    - If the data are not normal, use non-parametric tests.
    - If the data are normal, use parametric tests.
    - If you have groups of data, you MUST test each group for normality.
    - It's common seen that a model is built from the training data and is then applied to the testing data. Did these two data sets follow the same distribution?