台灣人工智慧學校

假設檢定 & 變異數分析

吳漢銘
國立臺北大學 統計學系

假設檢定 - 大綱

- **主題1**
  - 統計假設檢定 (Hypothesis Testing)
  - 平均數檢定 (t檢定)

- **主題2**
  - 單因子變異數分析 (One-way Analysis of Variance, ANOVA)
  - R程式範例

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.

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):

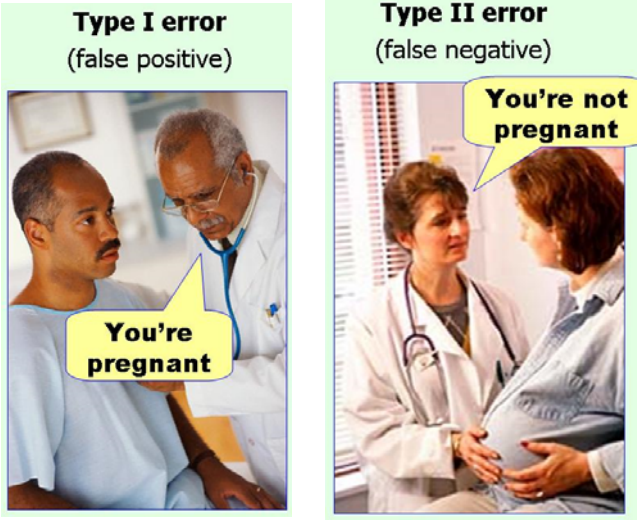
- Decide in advance.
- Alpha = 0.05: the probability of incorrectly rejecting the null hypothesis when it is actually true is 5%.

型一誤差、型二誤差

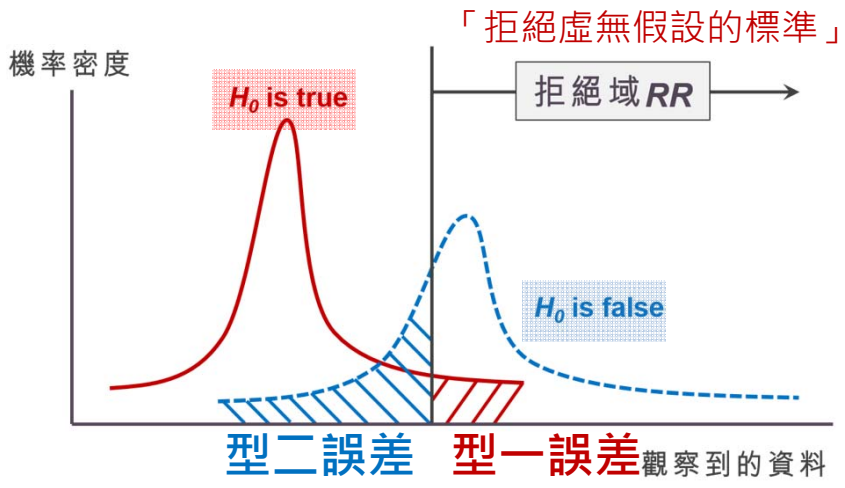
Hypothesis Testing		Truth	
		H_0	H_1
Decision	Reject H_0	**Type I Error** (α) (false positive)	Right Decision (true positive)
	Fail to Reject H_0	Right Decision (true negative)	**Type II Error** (β) (false negative)

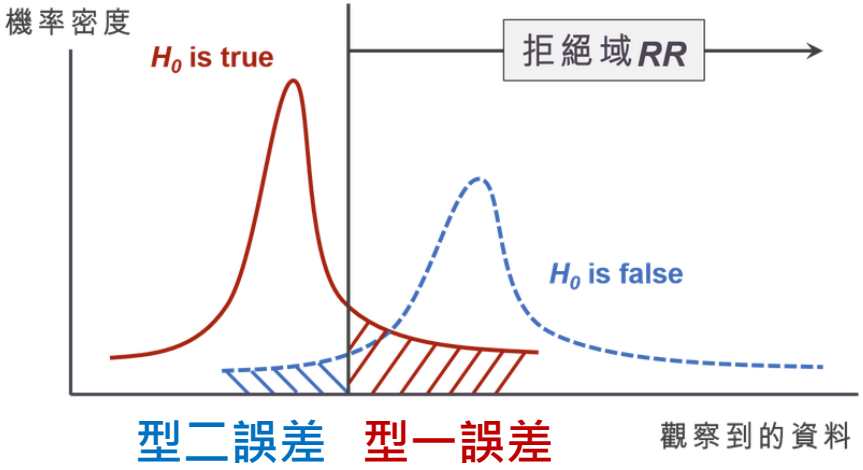
Power = 1- β

H_0: Not Pregnant



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





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

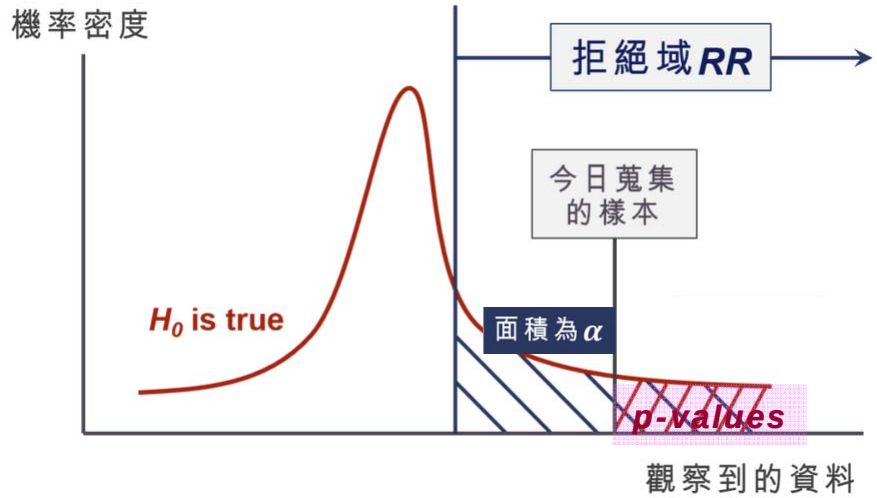
The *p*-values

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 alpha
- $P < 0.05$ 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/

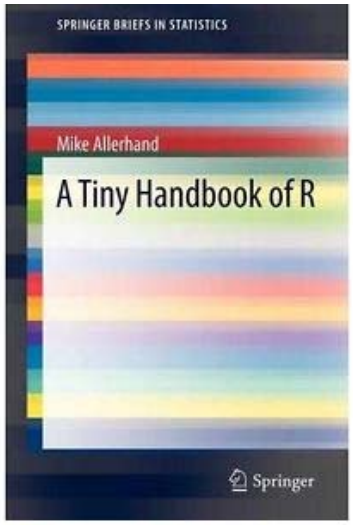
林澤民，看電影學統計: p值的陷阱
http://blog.udn.com/nilnimest/84404190
社會科學論叢2016年10月第十卷第二期

"**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。"

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, Spearman's, or Kendall's)
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



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

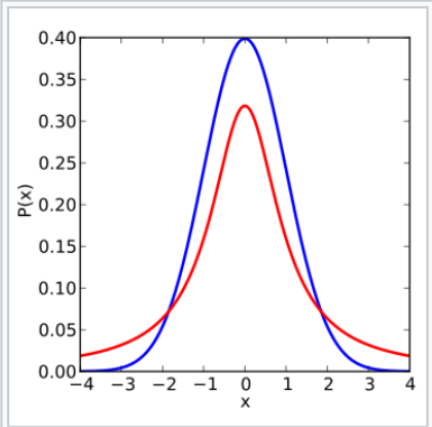
假設X是呈常態分布的獨立的隨機變量（隨機變量的期望值是μ，方差是σ^2但未知）。

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

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

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

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

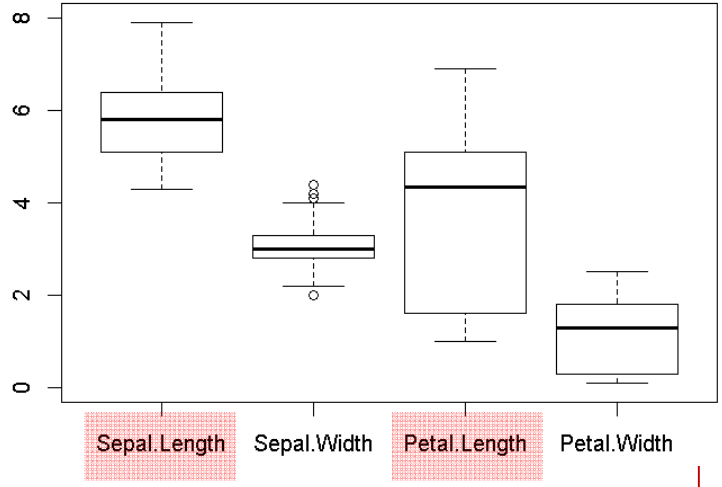
Description: Performs one and two sample t-tests on vectors of data.

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

假設檢定 - 大綱

- 主題1
  - 統計假設檢定 (Hypothesis Testing)
  - 平均數檢定 (t檢定)

- 主題2
  - 單因子變異數分析 (One-way Analysis of Variance, ANOVA)
  - R程式範例

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

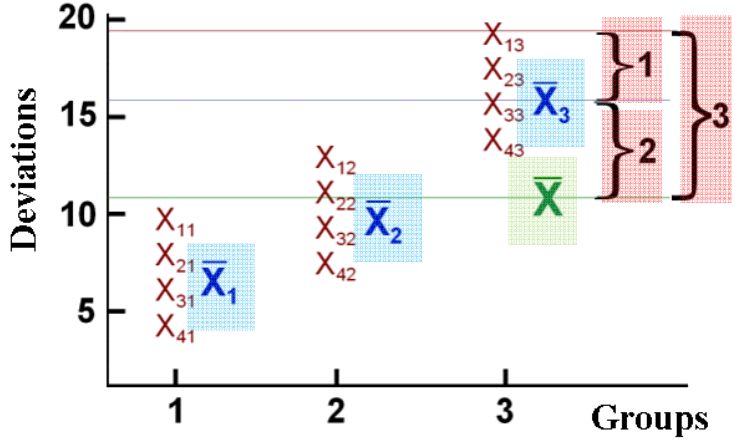
Groups

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}
$\vdots$	$X_{n_2 2}$	$\cdots$	$\vdots$	$\cdots$	$X_{n_k k}$
$X_{n_1 1}$			$X_{n_i j}$		

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

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

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



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

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

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

$$\epsilon_{ij} \sim N(0, \sigma^2) \qquad j = 1, \cdots, k$$

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

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

degree of freedom

SS/df

ANOVA Table

Source	SS	df	MS	F	p
Between	SS_B	$k-1$	MS_B	MS_B/MS_W	< 0.05
Within	SS_W	$N-k$	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/.

```
> 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
```