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

http://www.hmwu.idv.tw

本章大綱

- 簡介統計假設檢定 (Hypothesis Testing)
- 倍數變化 (Fold-Change)
- 平均數檢定 (t檢定)
  - 單樣本、成對樣本、雙樣本
- 單因子變異數分析 (One-way Analysis of Variance, ANOVA)
- 無母數檢定 (Non-parametric Tests)
  - Sign Test，Wilcoxon Signed-Rank Test (paired), Mann-Whitney Test, Kruskal-Wallis Test
- 事後比較檢定 (Post Hoc Tests)
  - Student-Newman-Keuls (SNK) Test, Tukey's HSD Test

假設檢定的應用:
Finding Differentially Expressed Genes

A B C
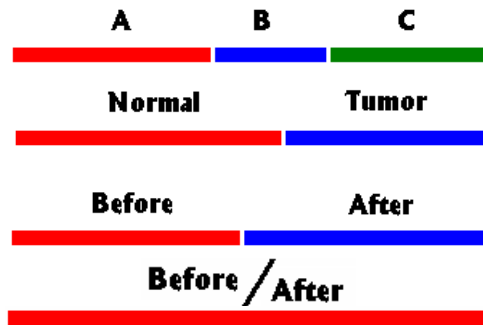
Normal Tumor → More than two samples

Before After → Two-sample (independent)

Before / After → Paired-sample (dependent)

Cy 5: treatment

Cy 3: control

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

p-values or Statistics

MA Table	exp01	exp02	exp03	exp04	exp05	exp•••	exp p		p-values	
gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35		0.067	
gene002	-0.39	-0.58	1.08	1.21	0.52		-0.58		0.052	
gene003	0.87	0.25	-0.17	0.18	-0.13		-0.13		0.013	*
gene004	1.57	1.03	1.22	0.31	0.16		-1.02		0.016	*
gene005	-1.15	-0.86	1.21	1.62	1.12		-0.44		0.112	
gene006	0.04	-0.12	0.31	0.16	0.17		0.08		0.017	*
gene007	2.95	0.45	-0.40	-0.66	-0.59		-0.76		0.059	
gene008	-1.22	-0.74	1.34	1.50	0.63		-0.55		0.063	
gene009	-0.73	-1.06	-0.79	-0.02	0.16		0.03		0.516	
gene010	-0.58	-0.40	0.13	0.58	-0.09		-0.45		0.009	*
gene011	-0.50	-0.42	0.66	1.05	0.68		0.01		0.068	
gene012	-0.86	-0.29	0.42	0.46	0.30		-0.63		0.030	*
gene013	-0.16	0.29	0.17	-0.28	-0.02		-0.04		0.002	*
gene014	-0.36	-0.03	-0.03	-0.08	-0.23		-0.21		0.423	
gene015	-0.72	-0.85	0.54	1.04	0.84		-0.64		0.084	
gene016	-0.78	-0.52	0.26	0.20	0.48		0.27		0.048	
gene017	0.60	-0.55	0.41	0.45	0.18		-1.02		0.018	*
gene018	-0.20	-0.67	0.13	0.10	0.38		0.05		0.538	
gene019	-2.29	-0.64	0.77	1.60	0.53		-0.38		0.053	
gene020	-1.46	-0.76	1.08	1.50	0.74		-0.70		0.074	
gene021	-0.57	0.42	1.03	1.35	0.64		-0.40		0.764	
gene022	-0.11	0.13	0.41	0.60	0.23		0.19		0.423	
gene•••										
gene n	-1.79	0.94	2.13	1.75	0.23		-0.66		0.723	

Microarray Data Matrix

gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35

ranking →

fix number

0.002	*
0.009	*
0.013	*
0.016	*
0.017	*
0.018	*
0.030	*
0.048	
0.052	
0.059	
0.063	
0.067	
0.068	
0.074	
0.084	
0.112	
0.423	
0.423	
0.516	
0.538	
0.723	
0.764	
..	

above some level

experimental   control

| For one gene | → | BG=100
S1=300 | BG=100
S2=200 | Fold-Change → |

$$\frac{cS1=200}{cS2=100} = {\color{red}2}$$
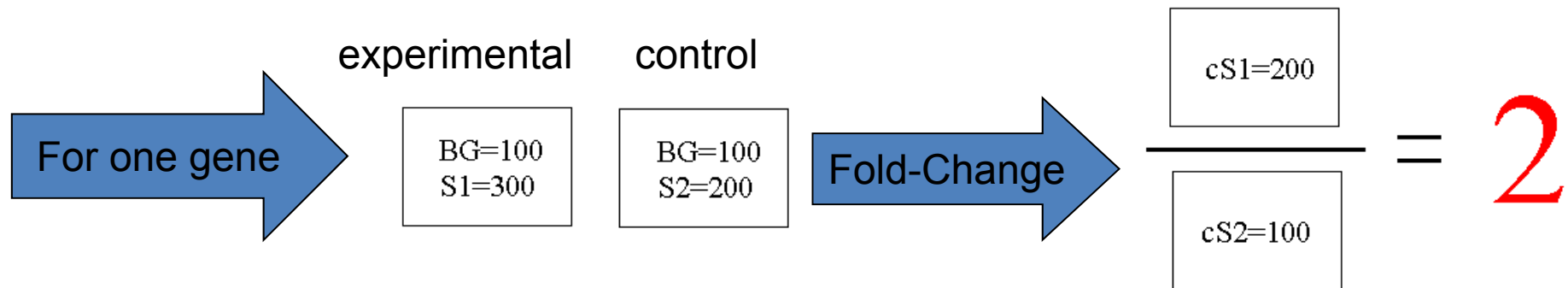
1) Calculate fold-change.

2) Rank the genes.

3) Select genes.

```
> exp.m <- apply(df[, index.exp], 1, mean)
> ctl.m <- apply(df[, index.ctl], 1, mean)
> plot(exp.m, ctl.m)
> abline(a=0, b=1)
> fc <- exp.m/crl.m
> no.genes <- 50
> sort(fc, decreasing = TRUE)[1:no.genes]
```

Method 1: Select genes based on Numbers

- average differential expression > FC.

Problems:

- FC is an arbitrary threshold.
- FC does not take into account individuals and sample size.

Example:

- s2 (200) close to BG (100), the difference could represent noise.
- credible: a gene is regulated 2-fold with 10000, 5000 units.

Method 2: Select genes based on %

- Choose 5% of genes that have the largest expression ratios.

Problems:

- Possible that no genes have statistically significantly different gene expression.

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 (2)

null hypothesis:

- H_0: $\mu = 2.5$. (the average price of a gallon of gas is \$2.5)
- H_0: $\mu_A - \mu_B = \mu_0$.

alternative hypothesis:

- H_a: $\mu > 2.5$. (gas prices were actually higher)
- H_a: $\mu < 2.5$.
- H_a: $\mu \mathrel{!=} 2.5$.

significance level (alpha):

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

Biological Question → *Statistical Formulation*

H_0: No differential expressed.

H_0: no difference in the mean gene expression in the group tested.

H_0: The gene will have equal means across every group.

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 (\ldots = \mu_n)$

H_0: no differential expressed.

- The test is significant

  = Reject H_0

- False Positive

  = (Reject H_0 | H_0 true)

  = concluding that a gene is differentially expressed when in fact it is not.

- A p-value=0.05 indicates that you would have only a 5% chance of drawing the sample being tested if the null hypothesis was actually true.

- The p-value is the smallest level of significance at which a null hypothesis may be rejected

The *p*-values

p-values

- probability of **false positives** (Reject H_0 | H_0 true).
- probability of observing your data under the assumption that the null hypothesis is true.
- *p-value* = 0.03: only a 3% chance of drawing the sample if the null hypothesis was true.

Decision Rule

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

p-value 的定義是：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。

p-value：若(前提) H_0 為真，則 test statistic 出現的可能性。(若p-value越小，表示抽樣樣本越(極端)不可能出現，因此推翻前提，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)。

林澤民，看電影學統計: p值的陷阱
(The Pitfalls of p-Values)
http://blog.udn.com/nilnimest/84404190
社會科學論叢2016年10月第十卷第二期
社會科學前沿課題論壇

"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。"

Type of Errors

Type I Error (alpha)
calling genes as differentially expressed when they are NOT
(when you see things that are not there.)

Type II Error
NOT calling genes as differentially expressed when they ARE
(when you dont see things that are there)

Hypothesis Testing		Truth	
		H0	H1
Decision	Reject H0	Type I Error (alpha) (false positive)	Right Decision (true positive)
	Don't Reject H0	Right Decision	Type II Error (beta)

Type I error (false positive)

You're pregnant

Type II error (false negative)

You're not pregnant

https://effectsizefaq.com/category/type-i-error/
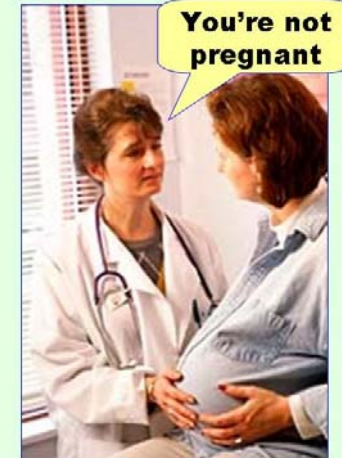
H_0: Not Pregnant

$$\text{Power} = 1 - \beta.$$

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

SPRINGER BRIEFS IN STATISTICS

Mike Allerhand

A Tiny Handbook of R

Springer
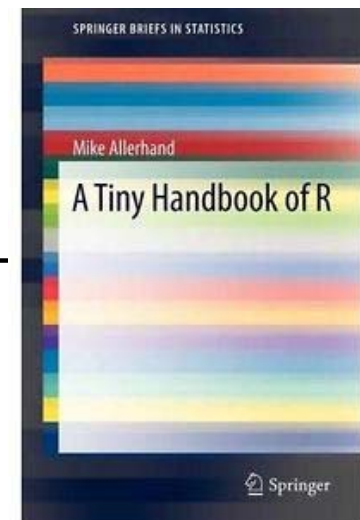
平均數檢定 in R

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

1. Determine the null and alternative hypothesis, using mathematical expressions if applicable.

2. Select a significance level (alpha).

3. Take a random sample from the population of interest.

4. Calculate a test statistic from the sample that provides information about the null hypothesis.

5. Decision

Hypothesis Testing:
 two-sided z-test & p-value

H_o: $\mu = 35$ null hypothesis
H_1: $\mu \neq 35$ alternative hypothesis ($\mu > 35$; $\mu < 35$)
  one-sided
α significant level: $= 0.05$

test statistic $z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

Reject H_0 if $|z| > z_{0.05}$

$$H_0 : \mu = m$$
$$H_1 : \mu \neq m$$
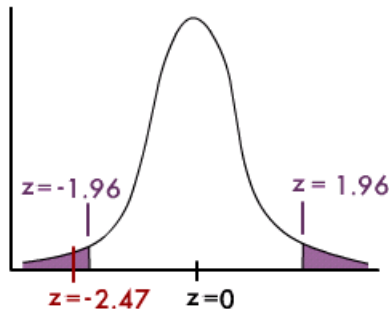$$\alpha = P_{H_0}(|Z| > z_{\alpha/2})$$

Sample Data: $= 33.6$
test statistic: $z = -2.47$

$(1-\alpha)100\%$ Confidence Interval:
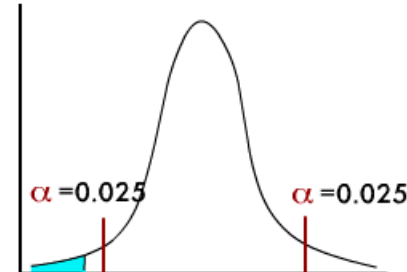$$P(z_{\alpha/2} < Z < z_{1-\alpha/2}) = 1 - \alpha$$

p-value $= P_{H_0}(|Z| > z_0)$, $z_0 = \frac{\bar{X} - m}{\sigma/\sqrt{n}}$

The Classical Approach

z = -1.96   z = 1.96

z = -2.47   z = 0

Conclusion: since the z value of the test statistic (-2.47) is less than the critical value of z= -1.96, we reject the null hypothesis.

The P-Value Approach

$\alpha = 0.025$ $\alpha = 0.025$

P-value = 0.0068 times 2 (for a 2-sided test) = 0.0136
Conclusion: since the P-value of 0.0136 is less than the significance level of $\alpha = 0.05$, we reject the null hypothesis.

檢定力 (Statistical Power)

- **Question**: What if I do a t-test on a pair of samples and **fail to reject the null hypothesis**--does this mean that there is no significant difference?

- **Answer**: Maybe yes, maybe no.

- For two-sample t-test, power is the probability of rejecting the hypothesis that the means are equal when they are in fact not equal. $P(RH_0 \mid \text{not } H_0) = \text{Power} = 1 - P(\text{Type-II error})$

- The power of the test depends upon the sample size, the magnitudes of the variances, the alpha level, and the actual difference between the two population means.

- Usually you would only consider the power of a test when you failed to reject the null hypothesis.

- High power is desirable (0.7 to 1.0): High power means that there is a high probability of rejecting the null hypothesis when the null hypothesis is false.

See also: `power.t.test {stats}`: Power calculations for one and two sample t tests.

Assumption: the variable is normally distributed.

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 $= 1 - \beta$.
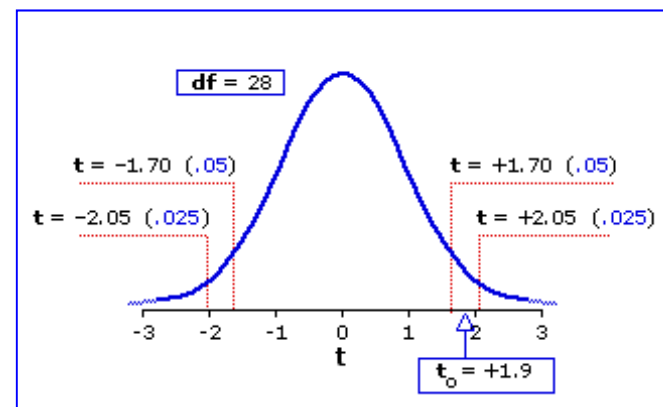- $(1 - \alpha)100\%$ Confidence Interval for μ:
  $\bar{X} - t_{\alpha/2}S/\sqrt{n} \leq \mu < \bar{X} + t_{\alpha/2}S/\sqrt{n}$
- $p\text{-value} = P_{H_0}(|\mathbf{T}| > t_0), \ \mathbf{T} \sim t_{n-1}$.

Question

■ whether a gene is differentially expressed for a condition with respect to baseline expression?

■ H_0: $\mu=0$ (log ratio)

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

df = 28

t = −1.70 (.05)   t = +1.70 (.05)

t = −2.05 (.025)   t = +2.05 (.025)

-3   -2   -1   0   1   2   3
t

$t_0 = +1.9$

Two Sample t-test

Paired Sample t-test

$H_0 : \mu_d = \mu_0$
$H_1 : \mu_d \neq \mu_0$ (two-tailed).
μ_d: mean of population differences.
α: significant level (e.g., 0.05).
Test Statistic:

$$T_d = \frac{\bar{d} - \mu_d}{S_d/\sqrt{n}}, \quad t_d = \frac{\bar{d} - \mu_0}{S_d/\sqrt{n}}$$

$\bar{d}$: average of sample differences.

S_d: standard deviation of sample difference

n: number of pairs.

- Reject H_0 if $|t_d| > t_{\alpha/2, n-1}$.

- Power $= 1 - \beta$.

- $(1 - \alpha)100\%$ Confidence Interval for μ_d:
  $\bar{d} - t_{\alpha/2} S/\sqrt{n} \leq \mu_d < \bar{d} + t_{\alpha/2} S/\sqrt{n}$

- $p\text{-}value = P_{H_0}(|\mathbf{T}| > t_d)$, $\mathbf{T} \sim t_{n-1}$.

Two Sample t-test (Unpaired)

$H_0 : \mu_x - \mu_y = \mu_0$
$H_0 : \mu_x - \mu_y \neq \mu_0$

α: significant level (e.g., 0.05).

Test Statistic:

$$t_0 = \frac{(\bar{X} - \bar{Y}) - \mu_0}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$

for homogeneous variances:
$df = n + m - 2$

for heterogeneous variances:
adjusted df

Reject H_0 if $|t_0| > t_{\alpha/2, df}$

Be Normal

- paired t-test,
  the distribution of the subtracted data that must be normal.

- unpaired t-test,
  the distribution of both data sets 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.

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
```
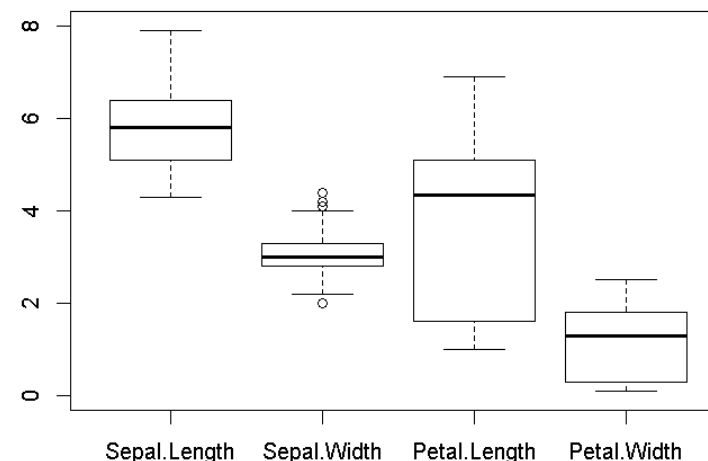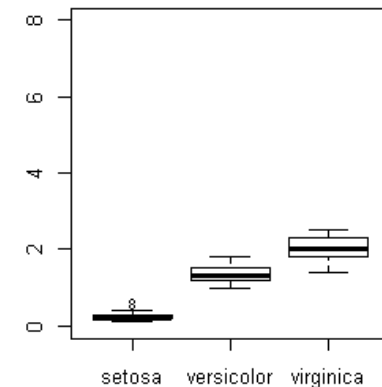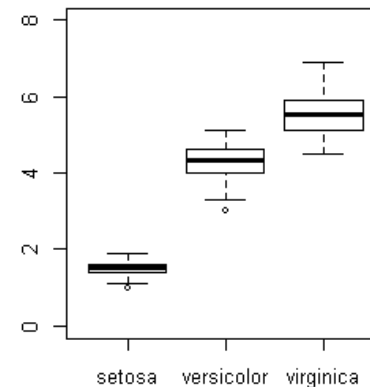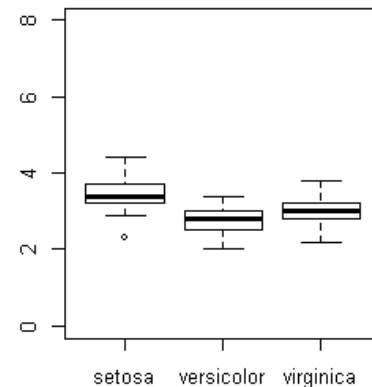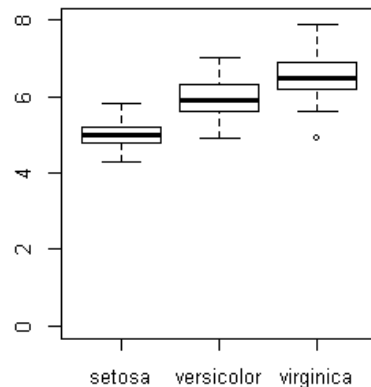

var.test {stats}: Performs an F test to compare the variances of two samples from normal populations.
H_0: the ratio of the variances of the populations from which x and y were drawn, or in the data to which the linear models x and y were fitted, is equal to ratio=1.

Using `t.test(x ~ g)`


```
> myData <- data.frame(value = iris$Sepal.Width[-(1:50)],
+   group <- iris[-(1:50), 5])
> alpha <- 0.05
> (bt <- bartlett.test(value ~ group, data=myData)$p.value <= alpha)
[1] FALSE
> t.test(value ~ group, data=myData, var.equal=!bt)
        Two Sample t-test

data:  value by group
t = -3.2058, df = 98, p-value = 0.001819
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.33028246 -0.07771754
sample estimates:
mean in group versicolor  mean in group virginica
                  2.770                    2.974
```

- `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 $t = \dfrac{\bar{M}}{\sqrt{(a+s^2)/n}}$, where a is estimated from the mean and standard deviation of the sample variances s^2.

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

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

Penalized t-statistic

Tusher et al (2001, PNAS, SAM)

Efron et al (2001, JASA)

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

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

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.
- ANOVA can test the following:
  - Are all the means from more than two populations equal?
  - Are all the means from more than two treatments on one population equal?
  - (This is equivalent to asking whether the treatments have any overall effect.)

One-Way ANOVA

- ### 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.

- ### Homogeneity of variance test

    - Bartlett's test (1937)

    - Levene's test (Levene 1960)
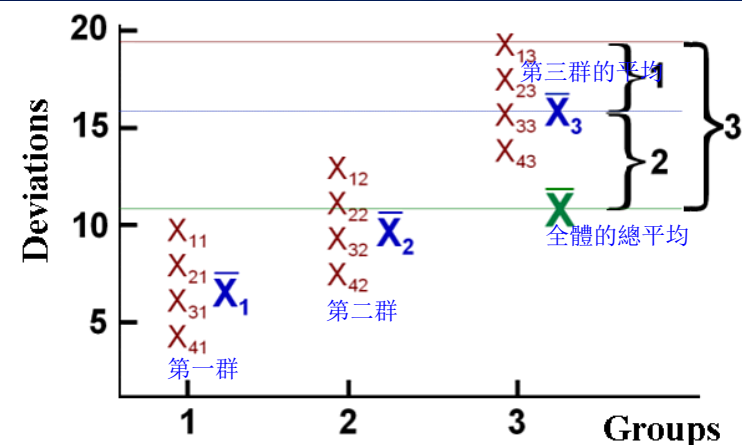
    - O'Brien (1979)

    - …

ANOVA Table

Groups

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

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

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

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

整體變異數

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

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$ 假設對之下，每一群平均要一樣

$$X_{ij} = \mu_j + \epsilon_{ij} \qquad \begin{array}{l} i = 1, \cdots, n_j \\ j = 1, \cdots, k \end{array}$$

觀察值　　　平均值

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

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

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

總變異是固定的　　　群中的平均　　　假設對之下，應該要是0

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	$N-1$			

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

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

F分布

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.

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

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

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

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

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

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

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

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

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

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

Small Round Blue Cell Tumors (SRBCT) Dataset

cDNA Microarrays

- *#Samples:* 63

  兒童小圓藍細胞腫瘤

  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	0.63		-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)                不要用for loop
+   SRBCT.aov.obj <- aov(x ~ khan$train.classes)
+   SRBCT.aov.info <- unlist(summary(SRBCT.aov.obj))
+   SRBCT.aov.info["Pr(>F)1"]
+ }                              回傳P-value
> # perform anova for each gene
> SRBCT.aov.p <- apply(khan$train, 1, Anova.pvalues)
                  一次傳出所有基因的P-value
```

```
> # 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))
> # lappay returns "list" and use rbind to convert it to "data.frame"
> # Try sapply?
```

```
> 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
```
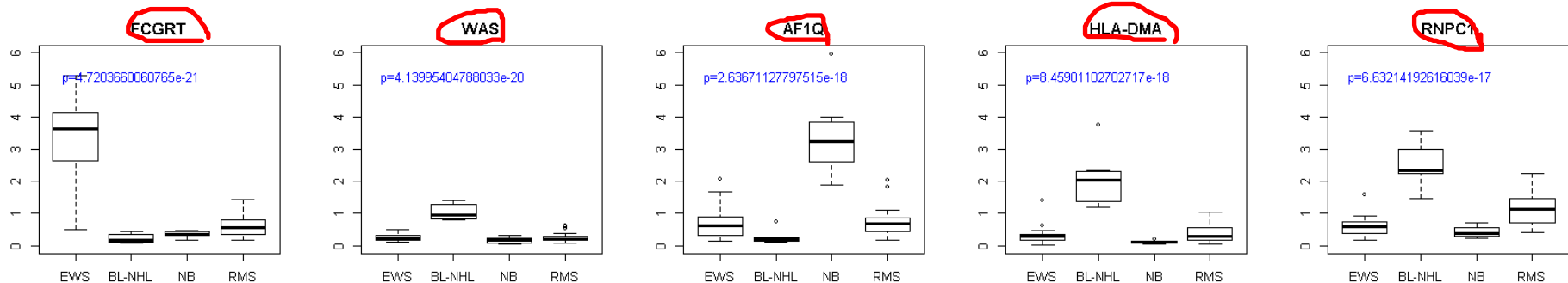

課堂練習: 試用Kruskal-Wallis Test重覆上述分析。