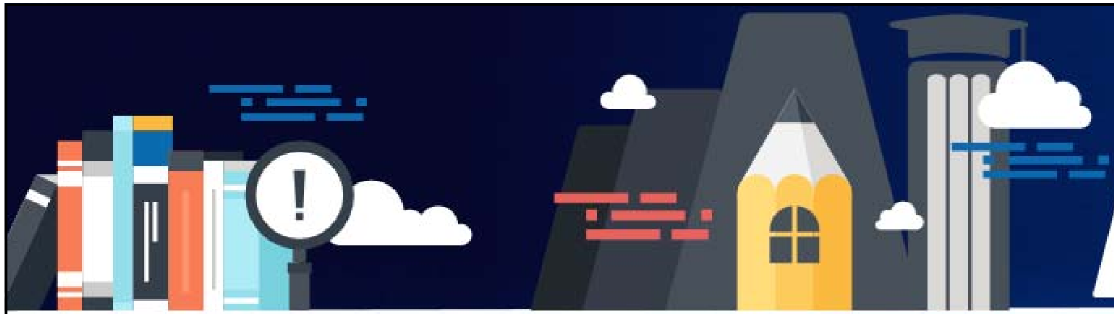
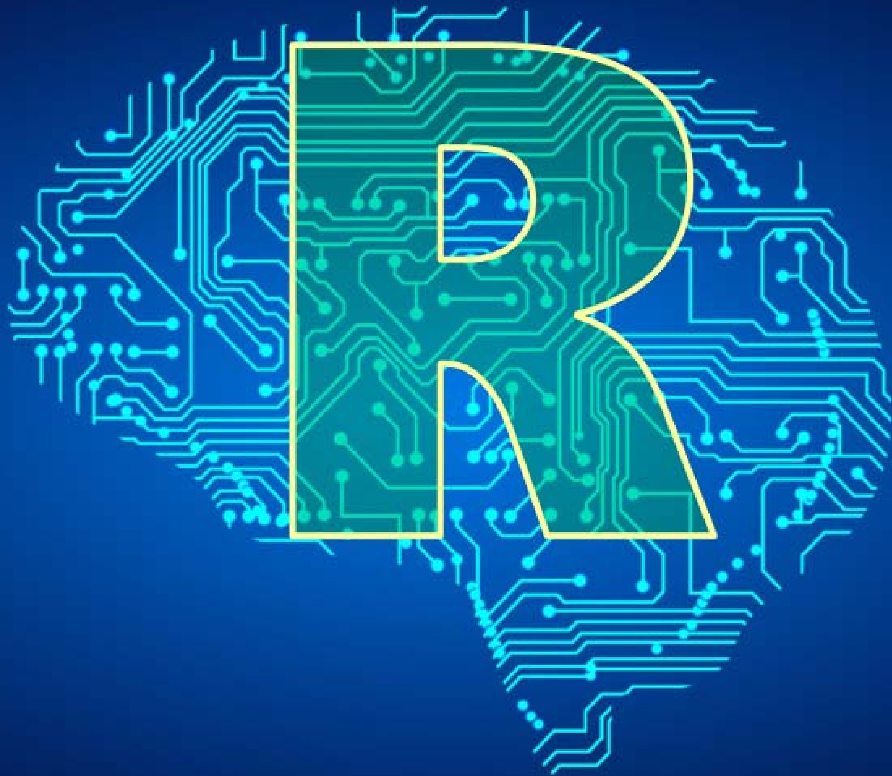


台灣人工智慧學校



無母數統計

吳漢銘
國立臺北大學 統計學系

本章大綱

- Non-parametric Models
- 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
- Test for Normality
- Permutation Tests
- Chi-Square Test

Non-parametric Statistics

- Do not assume that the data is normally distributed. 不需要常態分佈的假設
- Nonparametric statistics is based on either being distribution-free or having a specified distribution but with the distribution's parameters unspecified.
- Nonparametric statistics includes both descriptive statistics and statistical inference.

- Non-parametric models: kernel density estimation, non-parametric regression, ...
- Non-parametric inferential statistical methods: Kolmogorov–Smirnov test, Kruskal–Wallis one-way analysis of variance, Mann–Whitney U test, Sign test, Wilcoxon signed-rank test,...

Non-parametric Models

- Non-parametric models:
  - the model structure is not specified a priori or
  - the number and nature of the parameters are flexible and not fixed in advance.

- A **histogram** is a simple nonparametric estimate of **a probability distribution**.
- **Kernel density estimation** provides better estimates of the density than histograms.
- **Nonparametric regression** and **semiparametric regression** methods have been developed based on kernels, splines, and wavelets.

nonparametric regression

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

平均=0

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

k-nearest-neighbors regression.

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

kernel regression

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

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

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

Sign Test

- Given *n* pairs of data, the sign test tests the hypothesis that the median of the differences in the pairs is zero.

- The test statistic is the number of positive differences.

- If the null hypothesis is true, then the numbers of positive and negative differences should be approximately the same.

- In fact, the number of positive differences will have a Binomial distribution with parameters *n* and *p*.

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

The Sign Test:

when $n_1 = n_2 \leq 50$

$H_0 : P = Q = \frac{1}{2}$ 虛無假設

$H_1 : P \neq Q \neq \frac{1}{2}$ 擇一假設

$T = \#\text{"+"}$

At $\alpha = 0.01$, two-tailed test,
reject H_0 if $T \geq 14$ when $N = 17$.
(Binomial Probability)

$\#\text{"+"} = 14$
$\#\text{"−"} = 3$

The obtained T=14 is equal to the critical value, so we reject H_0.

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" (up-regulated) values is calculated and compared against a precomputed table to a p-value.
  - Sorting the absolute values of the differences from smallest to largest.
  - Assigning ranks to the absolute values.
  - Find the sum of the ranks of the positive differences.
- 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:

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

如果沒有差異，正的rank總和應該要跟負的rank 差不多

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

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.

Mann-Whitney Test

(Wilcoxon Rank-Sum Test, unpaired)

- The data from the two groups are combined and given ranks. (1 for the largest, 2 for the second largest,...)

- The ranks for the larger group are summed and that number is compared against a precomputed table to a p-value.

Group G_1	Group G_2	Rank G_1	Rank G_2
26	16	3	11
22	10	4	17
19	8	7.5	19
21	13	5.5	13.5
14	19	12	7.5
18	11	9	15.5
29	7	2	20
17	13	10	13.5
11	9	15.5	18
34	21	1	5.5
$n_1 = 10$	$n_2 = 10$	$R_1 = 69.5$	$R_2 = 104.5$

The Mann-Whitney U Test:

$H_0 : F_1 = F_2$

$H_1 : F_1 \neq F_2$

$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$

or

$U' = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$

$R_i = \sum_i \text{Rank}$

At $\alpha = 0.05$, two-tailed test for $n_1 = 10, n_2 = 10$, reject H_0 if $U \leq 23$ or $U' \geq 77$ (Table)

U: the number of times that a score from Group 1 is lower in rank than a score from Group 2.

$U = 85.5$, $U' = 14.5$

The obtained $U = 85.5$ is less than the critical value 77, so we reject H_0.

Kruskal-Wallis Test

- The Kruskal Wallis test can be applied in the one factor ANOVA case. It is a non-parametric test for the situation where the ANOVA normality assumptions may not apply.

- Each of the n_j should be at least 5 for the approximation to be valid.

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

$H_1 : \mu_i \neq \mu_j$ for at least one set of i and j

$$W = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1)$$

$W \sim \chi^2_{k-1}$ under H_0

Reject H_0 if $W > CHIPPF(\alpha, k-1)$, the chi-square percent point function

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

Rank Data

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

$$F(x) = P(X \leq x) = P(X \leq G(\alpha)) = \alpha$$

$$x = G(\alpha) = G(F(x))$$

The percent point function (ppf) is the inverse of the cumulative distribution function.

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.
- Not appropriate for data with outliers.

t-test	Non-parametric
Easy	Easy
Powerful	Robust
Widely Implemented	widely implemented
Not appropriate for data with outliers	Less powerful

Non-Parametric Tests

可能檢定不出來

- When certain assumptions about the underlying population are questionable (e.g. normality).
- Does not assume normal distribution
- No variance assumption
- Ranks the order of raw/normalized data across conditions for analyses
- Decrease effects of outliers (Robust)
- Not recommended if there is less than 5 replicates per group
- Needs a high number of replicates
- Less powerful

事後檢定 (Post Hoc Tests) Student-Newman-Keuls (SNK) Test

assuming
equal sample sizes and
homogeneity of variance

Group	A	B	C	D
Mean	2	3	7	8

alpha = 0.01
$n = 5$
$df = 16$

$$\sqrt{\frac{MSE}{n}} = \sqrt{\frac{.5}{5}} = 0.316$$

snk {mutoss}: {Unified Multiple Testing Procedures}
snk.test {GAD}: {Analysis of variance from general principles}
SNK.test {agricolae}: {Statistical Procedures for Agricultural Research}

"r" is the number of means spanned by a given comparison.

r, df, alpha → studentized range statistic q

1. $r = 4$, $q_{.01} = 5.19$

A vs D: $q = \frac{8-2}{0.316} = 18.99$, $p < 0.01$

2. $r = 3$, $q_{.01} = 4.79$

a. A vs C: $q = \frac{7-2}{0.316} = 15.82$, $p < 0.01$

b. B vs D: $q = \frac{8-3}{0.316} = 15.82$, $p < 0.01$

3. $r = 2$, $q_{.01} = 4.13$

a. A vs B: $q = \frac{3-2}{.316} = 3.16$, $p > 0.01$

b. B vs C: $q = \frac{7-3}{.316} = 12.66$, $p < 0.01$

c. C vs D: $q = \frac{8-7}{.316} = 3.16$, $p > 0.01$

Tukey's HSD Test

- To test all pairwise comparisons among means using the Tukey Honestly Significant Difference, calculate HSD for each pair of means using the following formula: $$\frac{M_1 - M_2}{\sqrt{MS_w\left(\frac{1}{n}\right)}}$$

(1) Mi – Mj is the difference between the pair of means.
(2) MS_w is the Mean Square Within, and n is the number in the group or treatment.

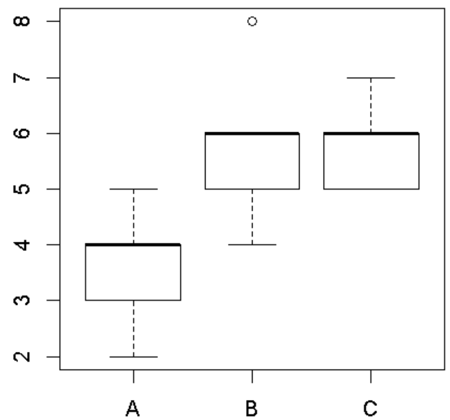
- **Steps:**
  - Step 1: Perform the ANOVA test. Assuming your F value is significant, you can run the post hoc test.
  - Step 2: Choose two means from the ANOVA output.
  - Step 3: Calculate the HSD statistic for the Tukey test using the formula.
  - Step 4: Find the score in Tukey's critical value table.
  - Step 5: Compare the score you calculated in Step 3 with the tabulated value you found in Step 4. If the calculated value from Step 3 is bigger than the critical value from the critical value table, the two means are significantly different.
- **Assumptions for the test**
  - Observations are independent within and among groups.
  - The groups for each mean in the test are normally distributed.
  - There is equal within-group variance across the groups associated with each mean in the test (homogeneity of variance).
- **Tukey's test and SNK test**
  - All alpha's in Tukey's test are compared to the same critical value.
  - All alpha's in SKN test are compared to a different critical value.
  - This test is more conservative (less powerful) than the SNK test.

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
```

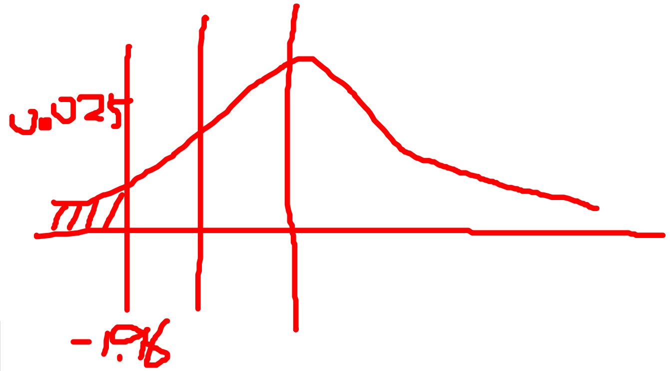
Formal Tests for Normality

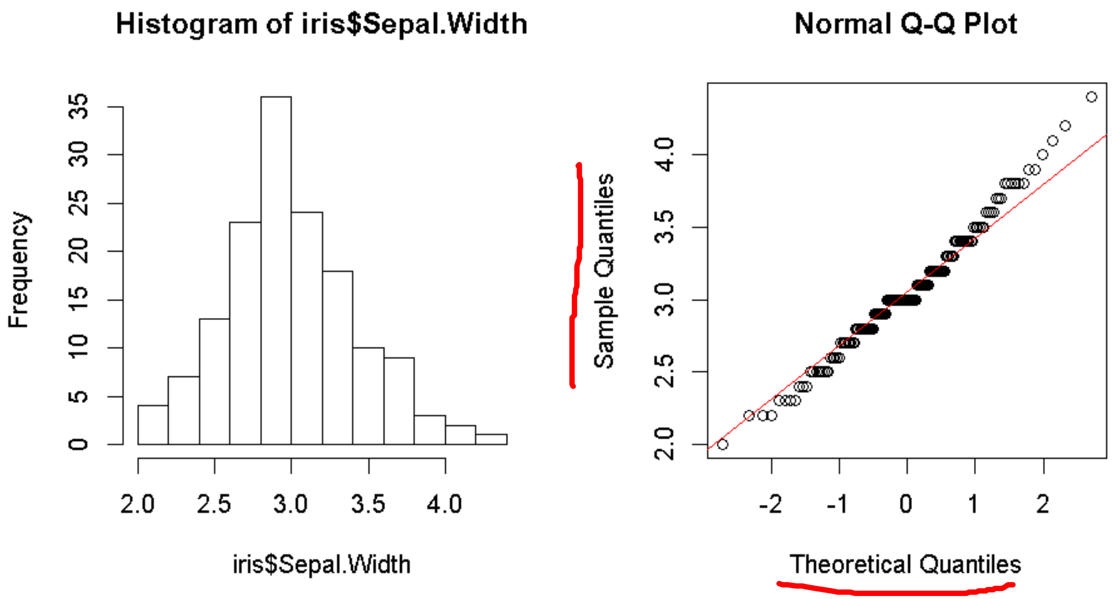
- The hypotheses used are:

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

```
> par(mfrow=c(1, 2))
> hist(iris$Sepal.Width)
> qqnorm(iris$Sepal.Width)
> qqline(iris$Sepal.Width, col="red")
```





Packages: **`nortest`**

Five omnibus tests for testing the composite hypothesis of normality:

`ad.test, cvm.test, lillie.test, pearson.test, sf.test`

ks.test, ad.test, shapiro.test

- Kolmogorov-Smirnov (K-S) test (Chakravarti et al., 1967).
- The Anderson-Darling test (Stephens, 1974).
- The Shapiro-Wilk normality test (Shapiro and Wilk, 1965).
- A large *p*-value (larger than, say, 0.05) indicates that the sample is not different from normal with the sample's mean and standard deviation.

```
> 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
```

Which Normality Test Should I Use?

- **Kolmogorov-Smirnov test**:
  - The test applies to continuous densities only.
  - 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 K-S 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?

Permutation Test
(randomization or re-randomization tests)

- The permutation test is a test where the null-hypothesis allows to reduce the inference to a randomization problem.
- The outcome data are analyzed many times (once for each acceptable assignment that could have been possible under H_0) and then compared with the observed result, without dependence on additional distributional or model-based assumptions.
- **Perform a permutation test (general):**
  1. Analyze the problem, choice of null-hypothesis
  2. Choice of test statistic **T**
  3. Calculate the value of the test statistic for the observed data: $\mathbf{t_{obs}}$
  4. Apply the randomization principle and look at all possible permutations, this gives the distribution of the test statistic **T** under H_0.
  5. Calculation of p-value:

$$p = P(T \geq t_{obs} \mid H_0) \approx \frac{\#\{t^* \geq t_{obs}\}}{\#\text{ permutations}}$$

Ref: Mansmann, U. (2002), Practical microarray analysis: resampling and the Bootstraap.Heidelberg.

Permutation Test

Coexpression of genes

H_0: Gene 1 and Gene 2 are not correlated.

Test statistic T:

Pearson (or Spearman) correlation coefficient, calculate t_{obs}

Randomization: Under H_0 it is possible to permute the values observed for Gene 2. There are n! possibilities.

p-value: $p = P(T \geq t_{obs} \mid H_0) \approx \dfrac{\#\{\, T^* \geq t_{obs} \,\}}{n!}$

Data

Gene1	Gene2
g_1^1	g_1^2
⋮	⋮
g_n^1	g_n^2

$g_{(1)}^1$	$g_{(1)}^2$
⋮	⋮
$g_{(n)}^1$	$g_{(n)}^2$

Random Permutation for group labels

Gene 1	Gene 2	Group	Group
1.4482	1.0709	1	2
0.4850	0.9324	1	1
1.1331	1.2379	1	4
		⋮	⋮
0.8015	0.6765	2	1
		⋮	⋮
1.3726	1.2373	3	4
		⋮	⋮
1.1030	1.735	4	2
0.5148	1.0015	4	3

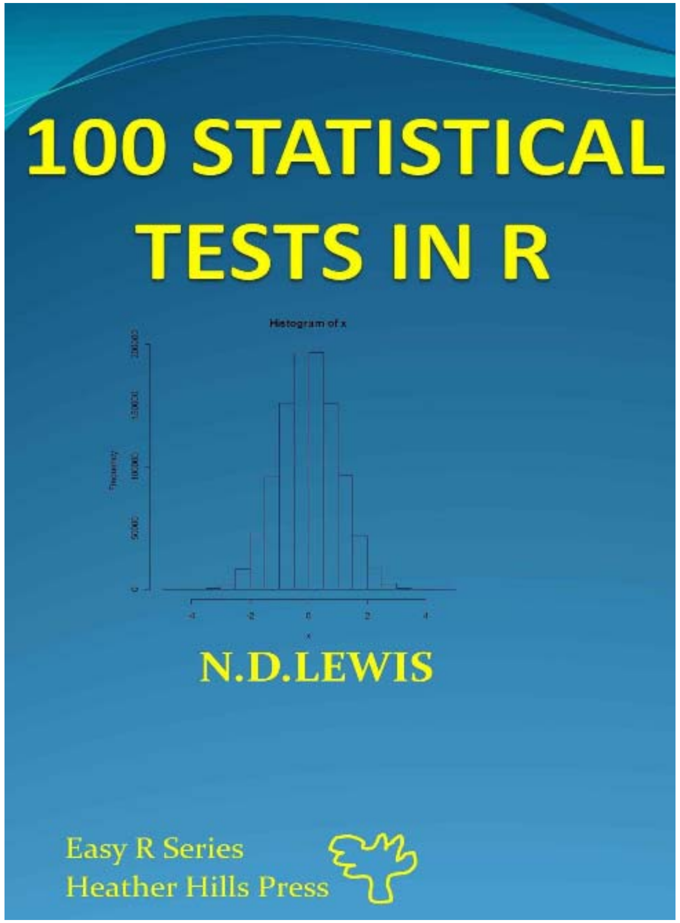
The permutation test allows determining the statistical significance of the score for every gene.

See also: the `coin` package and the `lmPerm` package:

`coin`: Conditional Inference Procedures in a Permutation Test Framework

`lmPerm`: Permutation Tests for Linear Models

卡方檢定: chisq.test



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.
- H_a: In the population, two categorical variables are dependent.

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

假設兩個變數是獨立的 P(I,J)=P(I)(J)

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

$\mu_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$ 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 (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
```