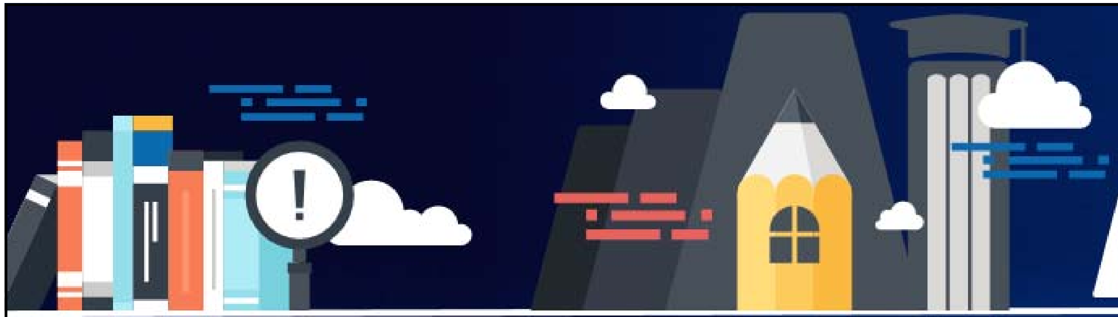
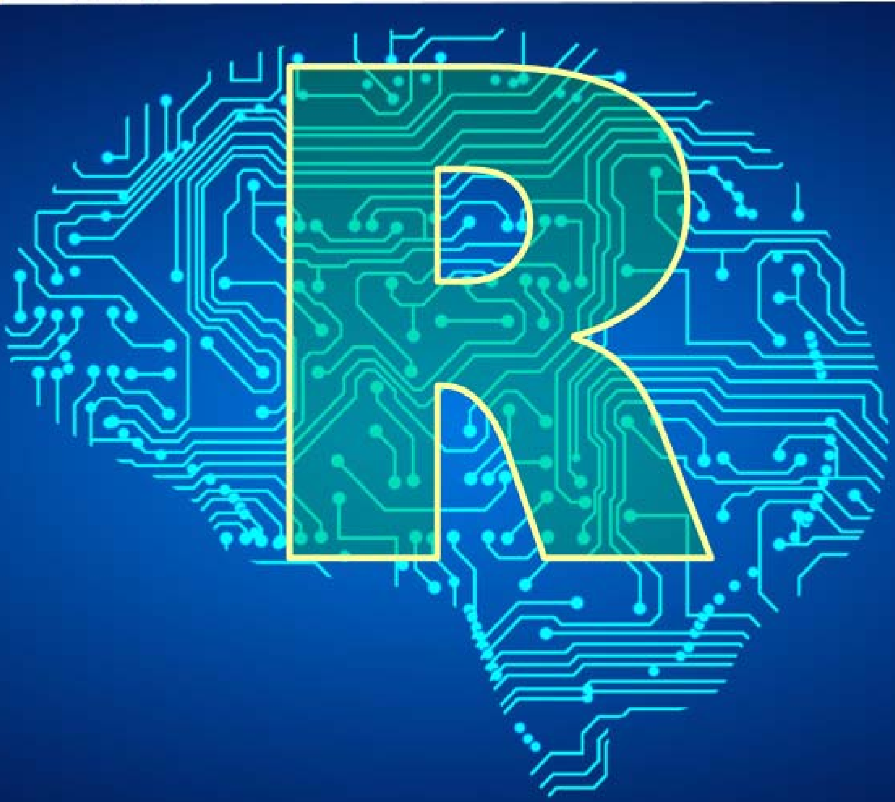


StatR-05

Basic

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

http://www.hmwu.idv.tw

無母數統計 - 大綱

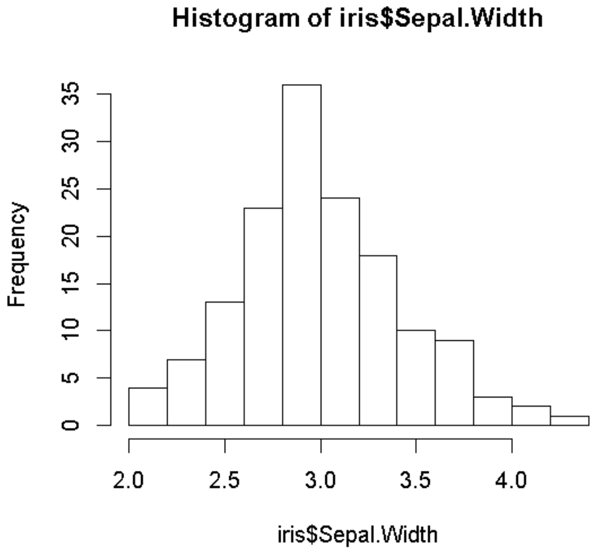
- 主題1
  - 常態分佈檢定 (Test for Normality)
  - 卡方檢定 (Chi-Square Test)

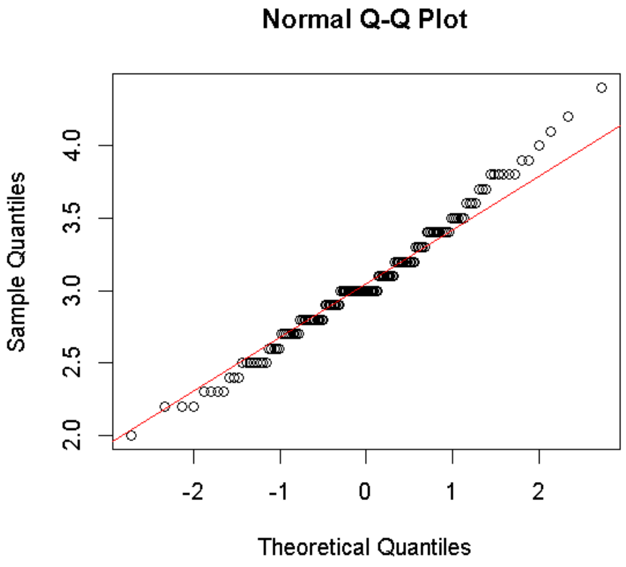
Formal Tests for Normality

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).

```
> 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
```

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

        Anderson-Darling normality test

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

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

        Shapiro-Wilk normality test

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

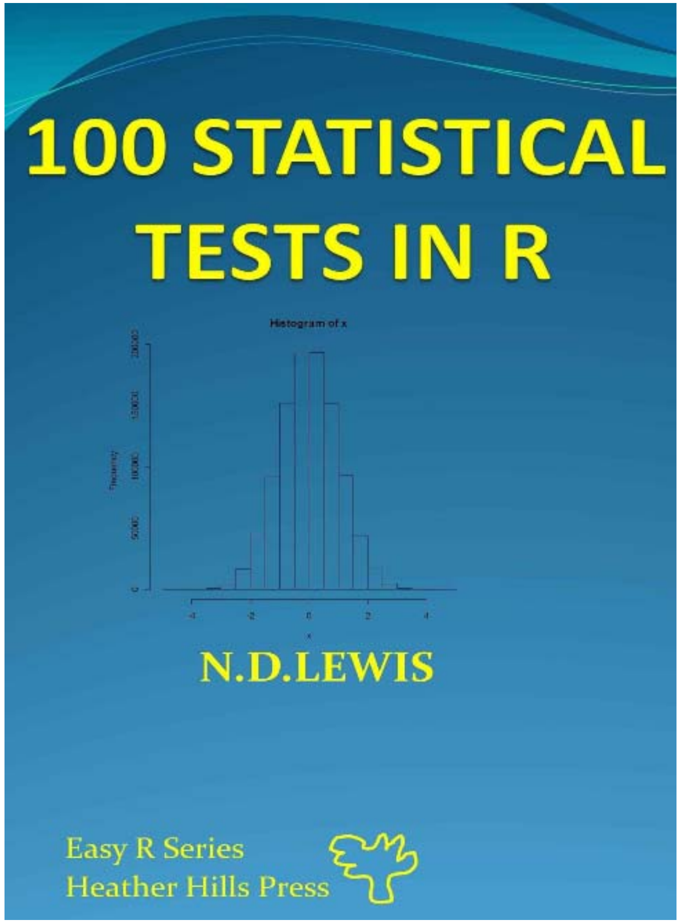
Which Normality Test Should I Use?

- Kolmogorov-Smirnov test:
  - 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?

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.**

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

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

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

estimated expected frequencies.

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

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

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

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

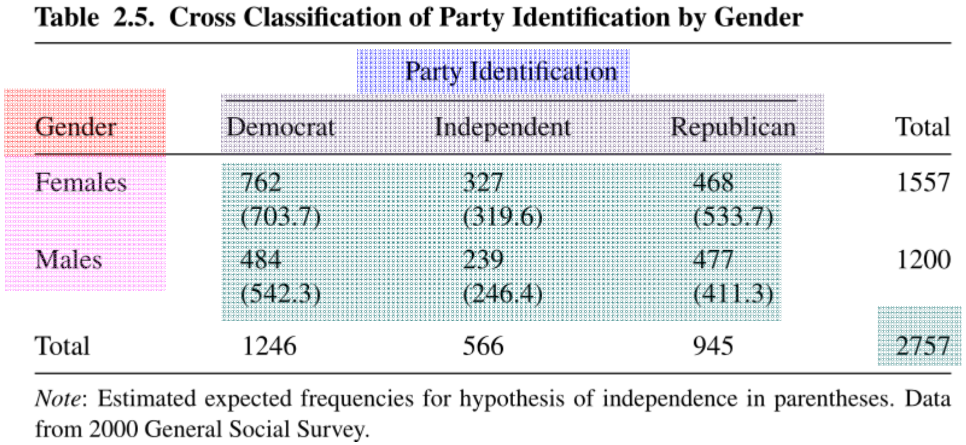


Table 2.5. Cross Classification of Party Identification by Gender

	Party Identification			
Gender	Democrat	Independent	Republican	Total
Females	762 (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
```