

■ 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)

- 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
- 貝式估計法 (X)

■ 統計假設檢定 (Hypothesis Testing)

- 型一誤差、型二誤差、 p -值
- 母體平均數檢定 (**單一樣本t檢定**)
- 單因子變異數分析 (One-way Analysis of Variance, ANOVA)
- **卡方檢定** (chi-square test): 齊一性檢定、獨立性檢定、適合度檢定



可能性函數、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, $\theta = \underline{(\theta_1, \dots, \theta_p)}$.
- (b) The likelihood function regards the data as a function of the parameter θ .
- (c) The **log likelihood** function

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



Maximum Likelihood Estimation

1. 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 .

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

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



範例: 估計最有可能中獎的機率

假設有一台抽獎機，每次抽的中獎機率都不會改變，也就是說每次抽中與否，都與前一次是否抽中無關，表示每次抽都是獨立事件。

假設此抽獎機連抽 5 次，只有第 1 次和第 4 次中獎，其他 3 次沒有中獎。若每次中獎機率為 p ，請推測最有可能的 p 值為多少？

抽獎機的機率模型

將隨機變數 X_i 定義為：
$$X_i = \begin{cases} 1 & (\text{中獎}) \\ 0 & (\text{沒中獎}) \end{cases}$$

每次中獎的機率為 p

沒中獎的機率為 $(1 - p)$

想要推估的參數就是 p 的值

則抽 5 次的中獎機率可分別寫為：

$$\begin{aligned} &P(X = X_1) \cdot P(X = X_2) \cdot P(X = X_3) \cdot P(X = X_4) \cdot P(X = X_5) \\ &= p \cdot (1 - p) \cdot (1 - p) \cdot p \cdot (1 - p) \\ &= p^2 \cdot (1 - p)^3 \end{aligned} \tag{6.3.2}$$

- 式子(6.3.2)稱為**概似函數 (Likelihood function)**。
- 只要找出能讓概似函數出現極大值的 p 就是最能符合此抽獎機率模型的答案。
- 要找出極大值，就是找出概似函數微分後等於 0 的 p ，且此 p 可以讓概似函數出現極大值。
- 概似函數習慣上會用 L (Likelihood) 做為函數名稱，但許多機器學習的書中習慣用 L 表示損失函數 (Loss function)，應避免混淆。

$$\begin{aligned}
 &P(X = X_1) \cdot P(X = X_2) \cdot P(X = X_3) \cdot P(X = X_4) \cdot P(X = X_5) \\
 &= p \cdot (1 - p) \cdot (1 - p) \cdot p \cdot (1 - p) \\
 &= p^2 \cdot (1 - p)^3
 \end{aligned}$$

$$\log(p^2(1 - p)^3) = 2 \log p + 3 \log(1 - p)$$

$$\frac{2}{p} + \frac{3 \cdot (-1)}{1 - p} = 0$$

$$\Leftrightarrow 2(1 - p) - 3p = 0$$

$$\Leftrightarrow 5p = 2$$

$$\Leftrightarrow p = \frac{2}{5} = 0.4$$

最大概似估計量
(maximum likelihood estimator, MLE)

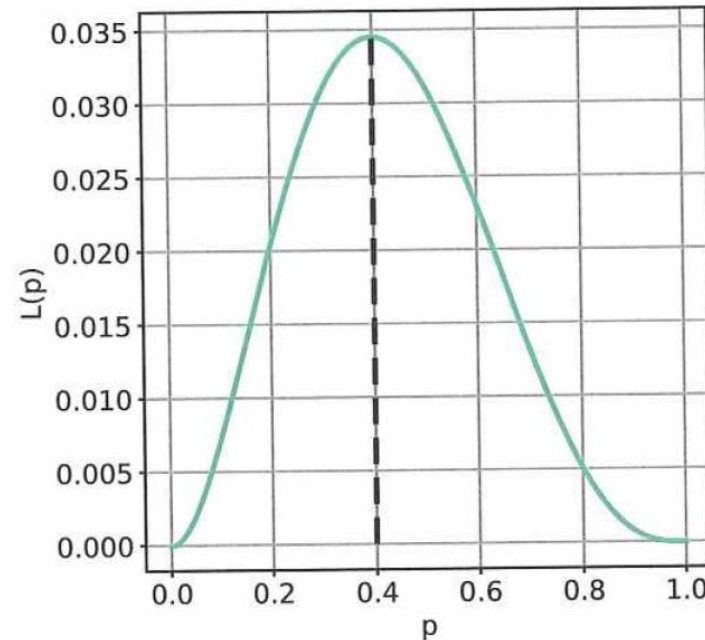


圖 6-13 橫軸為 p ，縱軸為概似函數的值

為何概似函數的極值是求最大值，而不是最小值？

- 最大概似估計法是找出「概似函數微分等於 0」的參數值。照理講，找出的參數也有可能讓概似函數出現極小值或無極值。
- 概似函數是由各已知事件的機率(介於0~1)相乘而來，數值只會大於等於0，而等於 0 就是極小值，也就是此機率模型最不可能發生的情況。
- 我們希望的是此機率模型最可能發生的情況，因此能產生極大值的參數才是我們要的。

Suppose Y_1, Y_2 are iid with density $f(y) = \theta e^{-\theta y}$, $y > 0$. Find the MLE of θ .

By independence, $L(\theta) = \frac{(\theta e^{-\theta y_1})(\theta e^{-\theta y_2})}{\theta^2 e^{-\theta(y_1+y_2)}}$.

(a) Thus $\ell(\theta) = \frac{2 \log \theta - \theta(y_1 + y_2)}{\theta}$ and the log-likelihood equation to be solved is

$$\frac{d}{d\theta} \ell(\theta) = \frac{2}{\theta} - (y_1 + y_2) = 0, \quad \theta > 0$$

(b) The unique solution is $\hat{\theta} = \frac{2}{(y_1 + y_2)}$, which maximizes $L(\theta)$.

(c) Therefore the MLE is the reciprocal of the sample mean in this example.

(a) The `mle` function takes as its first argument the function that evaluates $-\ell(\theta) = -\log(L(\theta))$.

(b) The negative log-likelihood is minimized by a call to `optim`, an optimization routine.

```
> y <- c(0.04304550, 0.50263474)
> theta_hat <- length(y) / sum(y)
> theta_hat
[1] 3.66515
>
> mlogL <- function(theta = 1) {
+   n <- length(y)
+   f <- -(n * log(theta) - theta * sum(y))
+   f
+ }
>
> library(stats4)
> fit <- mle(mlogL)
```

```
> summary(fit)
Maximum likelihood estimation

Call:
mle(minuslogl = mlogL)

Coefficients:
      Estimate Std. Error
theta  3.66515   2.591652

-2 log L: -1.195477
```

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

題目: 若 $X_1, \dots, X_n \sim \text{i.i.d. } N(\mu, \sigma^2)$. 求 (μ, σ^2) 之 MLE。

$$f(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

解:

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

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

$$\mathcal{L}(\mu, \sigma) = f(x_1, \dots, x_n | \mu, \sigma)$$

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

$$0 = \frac{\partial}{\partial \mu} \log(\mathcal{L}(\mu, \sigma)) = 0 - \frac{-2n(\bar{x} - \mu)}{2\sigma^2}.$$



$$\hat{\mu} = \bar{x} = \sum_{i=1}^n \frac{x_i}{n}.$$

$$E[\hat{\mu}] = \mu$$

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

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

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

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

$$E[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2.$$

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

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

$$\hat{\mu} = \bar{x} = \sum_{i=1}^n \frac{x_i}{n}.$$

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



MLE using `optim {stats}`

```
> loglikefun <- function(x, par){
+   mu <- par[1]
+   sigma <- par[2]
+   n <- length(x)
+   loglikelihood <- - (n / 2)*(log(2 * pi * sigma^2)) +
+     (-1/(2 * sigma^2)) * sum((x - mu)^2)
+   # return the negative to maximize rather than minimize
+   - loglikelihood
+ }
>
> set.seed(1123)
> x <- rnorm(100)
> x <- x/sd(x) * 8 # sd of 8
> x <- x - mean(x) + 10 # mean of 10
> cat("mean(x) =", mean(x), ", sd(x) =", sd(x))
mean(x) = 10 , sd(x) = 8
> optim(par = c(0.5, 0.5), fn = loglikefun, x = x)
$par
[1] 10.001693  7.975965

$value
[1] 349.3359

$counts
function gradient
          95      NA

$convergence
[1] 0

$message
NULL
```

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

(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95 次

Interval Estimate of Population Mean

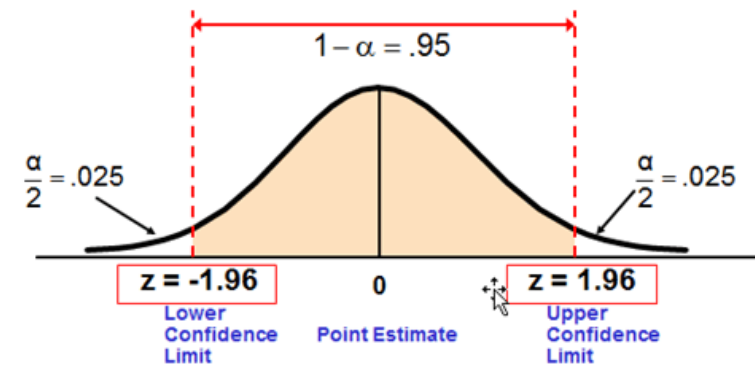
若大樣本($n > 30$)、母體 σ 已知，
由中央極限定理知 $\bar{X} \sim N(\mu, \sigma^2/n)$

$$\Rightarrow Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

$$P(-z \leq Z \leq z) = 1 - \alpha = 0.95.$$

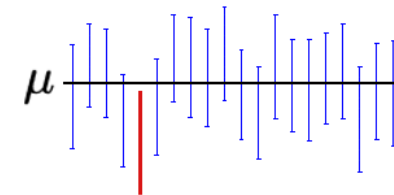
$$\Phi(z) = P(Z \leq z) = 1 - \frac{\alpha}{2} = 0.975,$$

$$z = \Phi^{-1}(\Phi(z)) = \Phi^{-1}(0.975) = 1.96,$$



$$0.95 = 1 - \alpha = P(-z \leq Z \leq z) = P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right)$$

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



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， $\sigma=8$ 小時， $n=100$

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

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

在 $1-\alpha$ 信賴水準下，母體平均數的信賴區間為

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

$$19.632 \leq \mu \leq 22.768$$

$$\bar{X} \pm Z_{\alpha/2} \sigma_{\bar{X}} = 21.2 \pm 1.96 \times 0.8$$

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%。」

```
> alpha <- 0.05
> xbar <- 21.2
> sigma <- 8
> n <- 100
> v <- qnorm(1-alpha/2)*(sigma/sqrt(n))
> c(xbar - v, xbar + v)
[1] 19.63203 22.76797
```

假設檢定 (Hypothesis Testing)

假設檢定 是一個用來決定母體特徵(參數)的命題是否為合理的程序。

例子(1):

“麻薩諸塞州(Massachusetts)的加油站平均一加崙的汽油(regular unleaded gas)價格是 \$2.5 元”

這個命題是對的嗎?

- 對所有加油站做調查。
- 隨機選一小部份加油站當樣本做調查。

若從樣本調查出的結果是平均價格為\$2.2元。

- 這30分的差異是隨機變異(chance variability)的結果，還是
- 原本的命題不對?



例子(2):

(20%) 木柵小哥本學期修了大刀教授的統計學，歷次考試 (包含小考、抽考、期中考、期末考及加分考) 成績如下:

68, 64, 58, 68, 55, 52, 51, 52, 54, 57, 59, 62, 53, 58, 61

學期總成績為上述成績之平均，計算之後為“58.13333”，而學校記分簿只會登錄「58」。聽聞大刀教授是鐵面無私不加分的，因此木柵小哥突發奇想，想要進行一個假設檢定:「他的平均成績應該是及格的，算出來不及格只是誤差範圍而已」(亦即，他的統計學學習成效應該有 60 分 (含) 以上，用來拜託教授幫他學期成績加 2 分。請同學幫他進行這項檢定，看看上述的成績資料可否支持他的論點? (假設每次考試成績皆獨立，顯著水準 (α) 為 0.05， t -value: $t_{(0.05,14)} = 1.7613$, $t_{(0.05,15)} = 1.7530$, $t_{(0.025,14)} = 2.1447$, $t_{(0.025,15)} = 2.1314$ 。需將「假設檢定」過程中的每一個元素 (H_0, H_a, \dots , Conclusion) 皆寫出。)

虛無假設 (*Null hypothesis*):

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

擇一假設 (*alternative hypothesis*):

- $H_a: \mu > 2.5$. (gas prices were actually higher)
- $H_a: \mu < 2.5$.
- $H_a: \mu \neq 2.5$. (雙尾檢定)

顯著水準 (*significance level*)(*alpha*):

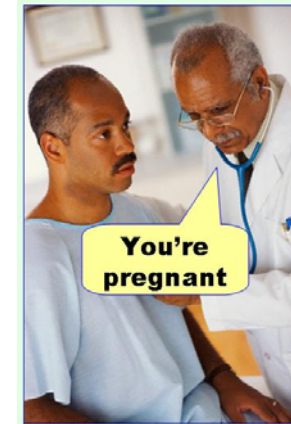
- 需事先決定。
- Alpha = 0.05: the probability of **incorrectly rejecting the null hypothesis** when it is actually true is 5%.
(虛無假設對之下，拒絕虛無假設的機率)
(錯誤地拒絕虛無假設的機率)

假設檢定		真實 (Truth)	
		H_0	H_1
決策 (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)

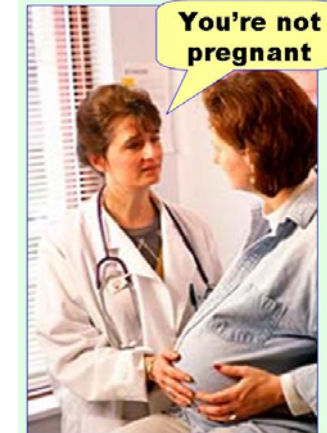
Power = $1 - \beta$

H_0 : Not Pregnant

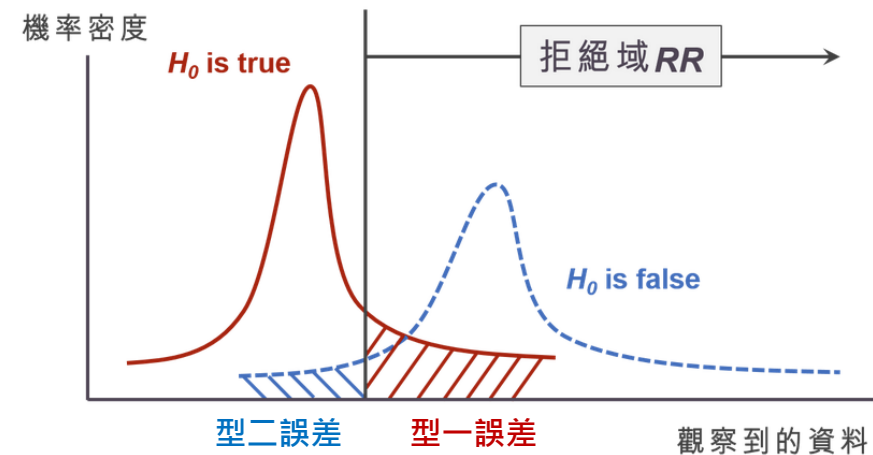
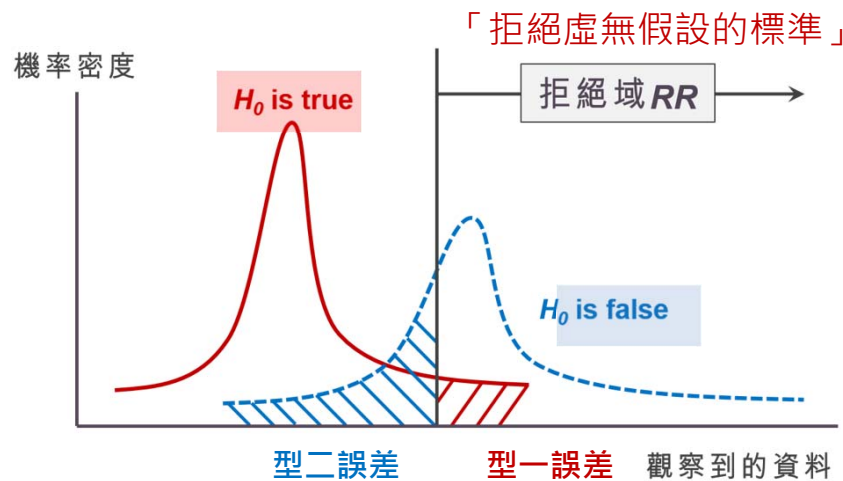
Type I error
(false positive)



Type II error
(false negative)



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



p-值 (The p-value)

16/25



p-value:

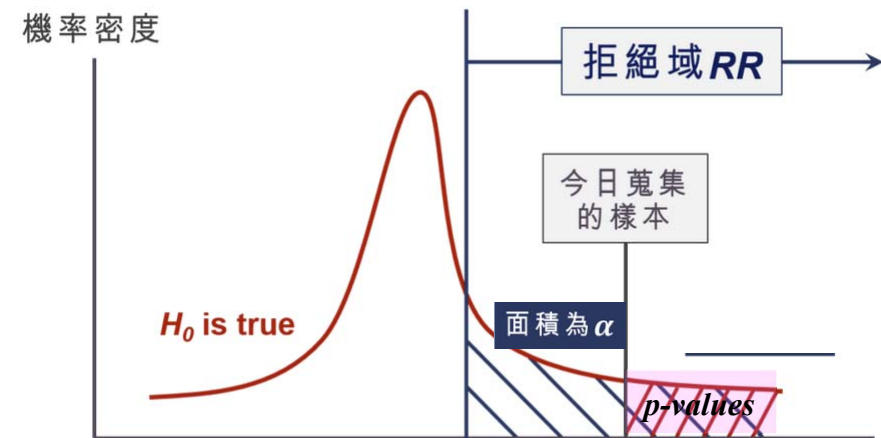
- 定義：在已知(現有)的抽樣樣本下，能棄卻 H_0 (虛無假設) 的最小顯著水準。(Reject H_0 | H_0 true)
- 若 H_0 為真，則檢定統計量出現(觀察到此樣本)的可能性。
(若 p-value 越小，表示抽樣樣本越不可能出現，因此推翻假設，拒絕 H_0)。
- p-value：以現有的抽樣所進行的推論，可能犯 type I error 的機率。
(若 p-value 越小，表示拒絕 H_0 不太可能錯，因此拒絕 H_0)。

Harry Potter,
分類帽(Sorting Hat)



決策法則:

- 拒絕 H_0 若 *p-value* 比 alpha 小。
- $P < 0.05$ commonly used.
(拒絕 H_0 ，稱檢定是顯著的(significant))
- The lower the *p-value*, the more significant.



<https://tawei Huang.hpdl.io/2017/01/11/poorpvalue/>

觀察到的資料
檢定統計量

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。"

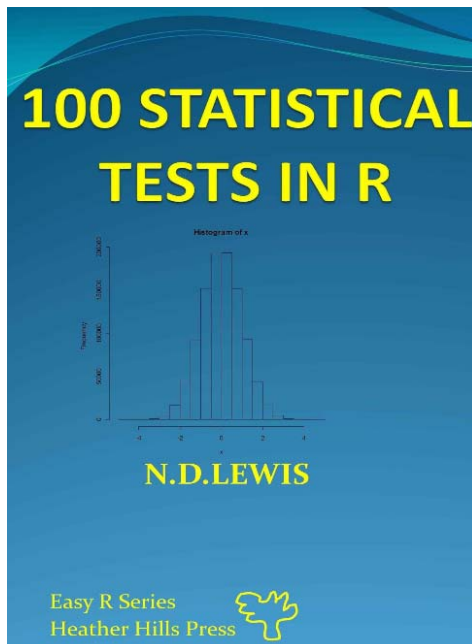
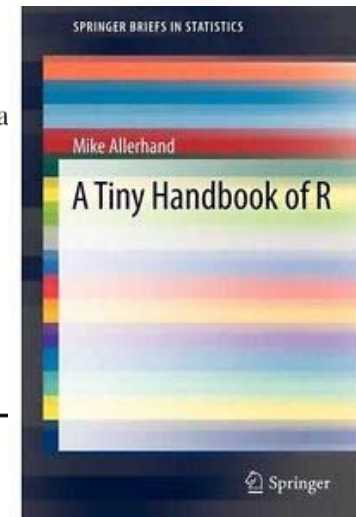


The Hypothesis Tests in Base R ^{17/25}

The hypothesis tests provided in the base installation include¹:

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)
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



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



Hypothesis Testing	One Sample	Two Samples		> two Groups
	-	Paired data	Unpaired data	Complex data
Parametric (variance equal)	t-test <code>t.test(x, mu = 0)</code>	t-test <code>t.test(x-y, var.equal = TRUE)</code> <code>t.test(x, y, paired = TRUE, var.equal = TRUE)</code>	t-test <code>t.test(x, y, var.equal = TRUE)</code>	One-Way Analysis of Variance (ANOVA) <code>aov(x~g, data)</code> <code>oneway.test(x~g, data, var.equal = TRUE)</code>
Parametric (variance not equal)		Welch t-test <code>t.test(x-y)</code> <code>t.test(x, y, paired = TRUE)</code>	Welch t-test <code>t.test(x, y)</code>	Welch ANOVA <code>oneway.test(x~g, data)</code>
Non-Parametric (無母數檢定)	Wilcoxon Signed-Rank Test <code>wilcox.test(x, mu = 0)</code>	Wilcoxon Signed-Rank Test <code>wilcox.test(x-y)</code> <code>wilcox.test(x, y, paired = TRUE)</code>	Wilcoxon Rank-Sum Test (Mann-Whitney U Test) <code>wilcox.test(x, y)</code>	Kruskal-Wallis Test <code>kruskal.test(x, g)</code>

`pairwise.t.test {stats}`: Calculate pairwise comparisons between group levels with corrections for multiple testing

`TukeyHSD {stats}`: Compute Tukey Honest Significant Differences



單一樣本t-檢定 (t-test)

可能的應用問題:

- 一家醫院想知道病患膽固醇值的平均數是否與目標值200mg不同?
- 消保官想了解能量棒上的標示「此能量棒含20公克的蛋白質」是否正確?

- 設定虛無假設及擇一假設。

$$H_0: \mu = \mu_0$$

- 選定 α
- 收集資料: x_1, x_2, \dots, x_n 。
- 驗證假設。
- 計算平均數、變異數。
- 計算檢定統計量。
- 算p-值。
- 做決策。

p-value approach

Critical value approach

One sample t-test

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0 \text{ (two-tailed).}$$

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 :

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

雙尾檢定 (two-tailed test)

單尾檢定

左尾
(Lower tail)

$$H_0: \mu \geq \mu_0$$

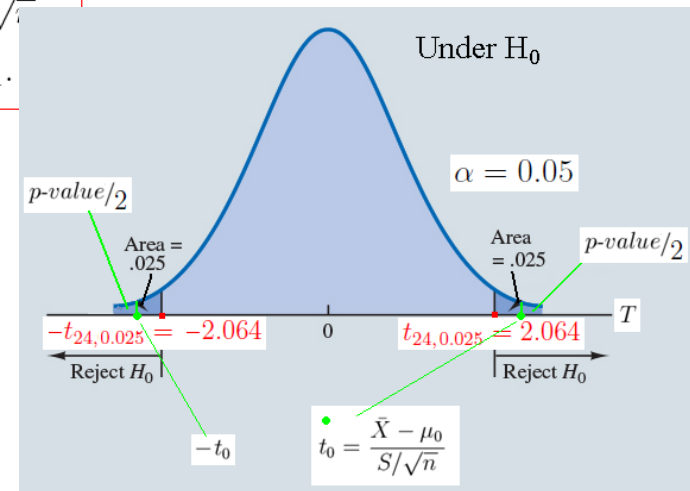
$$H_a: \mu < \mu_0$$

右尾
(Upper tail)

$$H_0: \mu \leq \mu_0$$

$$H_a: \mu > \mu_0$$

T的抽樣分佈 (sampling distribution)



t檢定的假設 (Assumption)

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 ,

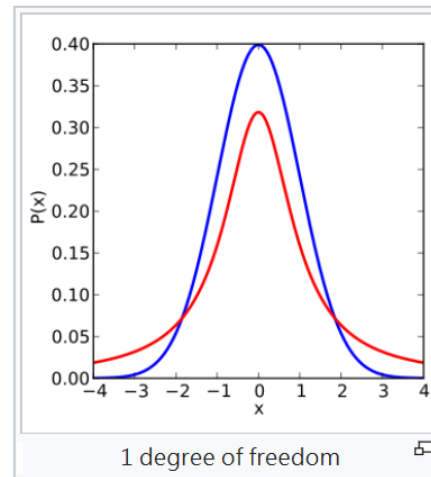
方差是 σ^2 但未知) 。

$$\bar{X}_n = (X_1 + \dots + X_n)/n$$

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

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

t-分布密度 (紅色曲線)
標準常態分布(藍色曲線)。



常態分佈 (Normal)

- 資料必需為常態分佈。
(若不符合，有一些經驗法則(對稱分佈、樣本數很大、轉換)或改採用「無母數檢定」。)
- **如何檢測資料是否為常態?**
 - **Plots:** Histogram, Density Plot, QQplot,...
 - **Test for Normality:** Jarque-Bera test, Lilliefors test, Kolmogorov-Smirnov test, Shapiro-Wilk test.

同質性 (Homogeneous)

- (雙樣本t檢定) 兩母體的變異數要相同。
- Test for equality of the two variances: Variance ratio F-test.
- Tests in R: `var.test`, `bartlett.test`, `ansari.test`, `mood.test`, `fligner.test`, `leveneTest`.

- 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".

範例:消保官想了解能量棒上的標示 「此能量棒含20公克的蛋白質」是否正確? (t檢定)

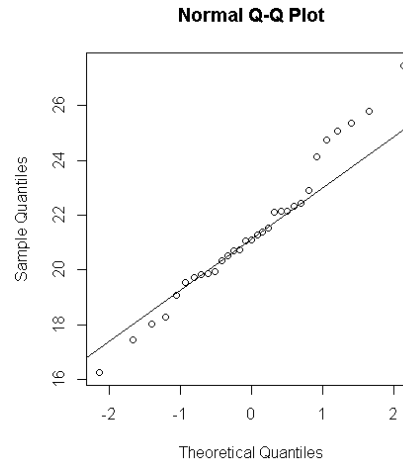
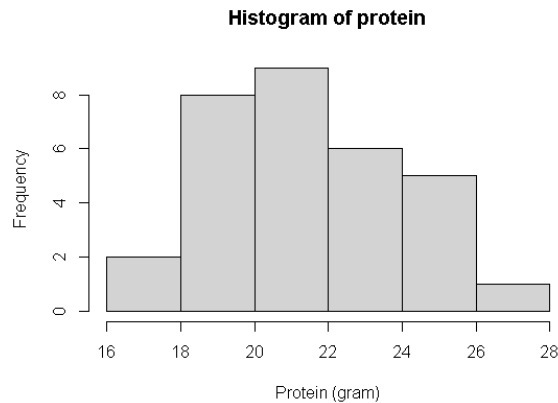
$$H_0: \mu = 20, \quad H_1: \mu \neq 20, \quad \alpha = 0.05.$$

31根能量棒的蛋白質含量(克數):

20.70, 27.46, 22.15, 19.85, 21.29, 24.75, 20.75, 22.91, 25.34, 20.33, 21.54, 21.08, 22.14, 19.56, 21.10, 18.04, 24.12, 19.95, 19.72, 18.28, 16.26, 17.46, 20.53, 22.12, 25.06, 22.44, 19.08, 19.88, 21.39, 22.33, 25.79



營養成分 每份(50克)	
熱量	190大卡
蛋白質	20克
碳水化合物	17克
總脂肪	6克
飽和脂肪	3.5克
膽固醇	15毫克
鈉	180毫克
膳食纖維	<1克
糖	2克
糖醇	8克



```
> ks.test(log(protein), "pnorm")

One-sample Kolmogorov-Smirnov test

data:  log(protein)
D = 0.99735, p-value = 3.331e-16
alternative hypothesis: two-sided
```

```
> shapiro.test(protein)

Shapiro-Wilk normality test

data:  protein
W = 0.9768, p-value = 0.7191
```

```
> t.test(protein, mu = 20)

One Sample t-test

data:  protein
t = 3.0668, df = 30, p-value = 0.004553
alternative hypothesis: true mean is not equal to 20
95 percent confidence interval:
 20.46771 22.33229
sample estimates:
mean of x
 21.4
```

拒絕「平均蛋白質公克數等於 20」的虛無假設。標示資訊不正確，且蛋白質公克數的母體實際上平均數大於 20。

標籤資訊應該更新，或製造流程應該改善，以製造出平均含 20 公克蛋白質的能量棒。



t.test {stats}:

Student's t-Test

22/25

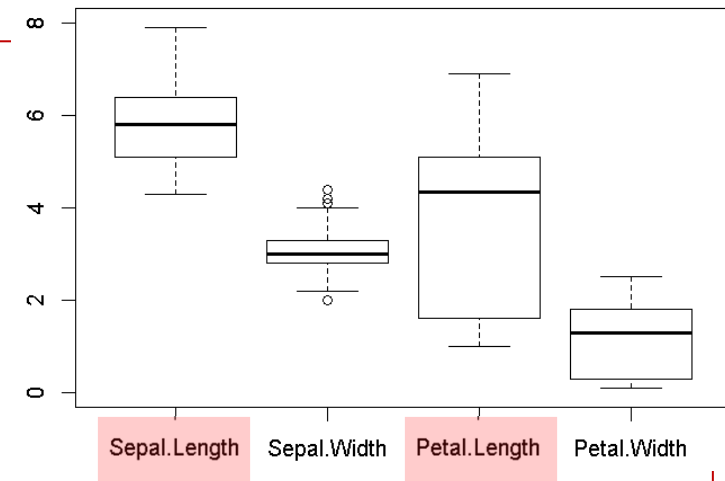
Description: Performs one and two sample t-tests on vectors of data.

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
```



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 .

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

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

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

$$t = \frac{\bar{M}_A - \bar{M}_B}{s^* \times \sqrt{1/n_A + 1/n_B}}, \text{ where } s^* = \sqrt{a + s^2}$$

Robust General Penalized t-statistic

卡方檢定: `chisq.test`

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

Gender	Party Identification			Total
	Democrat	Independent	Republican	
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
```