Engineering Statistics

Testing Statistical Hypotheses

Dr. Vvn Weian Chao (趙韋安)
https://ce.nctu.edu.tw/member/teachers/23
Department of Civil Engineering, National Yang Ming Chiao Tung University, Taiwan

Purpose

假設檢定

- 點估計值與估計信賴區間並未直接作出決策，然而假設檢定即可以做到決策

對母體參數做出一個適當地假設，然後根據隨機抽樣之樣本，因應樣本統計量的抽樣分佈來決定接受或拒絕假設的過程

R: Useful function

“BSDA” package

z.test().-z檢定

t.test().-t檢定

var.test().-F檢定

shapiro.test()-常態分佈檢定

Outline

-Errors in Hypothesis Testing
-Test Statistics & *p*-values
-Tests Concerning a Difference between two Means
-Test to Compare two Variances, F-test

Errors in Hypothesis Testing

- 虛無假設(H_0)及對立假設(H_a)
- 假設檢定有兩種可能的結論，為拒絕H_0或無法拒絕H_0

- **Null hypothesis** (H_0) – the assertion that is initially assumed to be true.
- **Alternative hypothesis** (H_a) – the claim that is contradictory to H_0.
- The null hypothesis will be rejected in favor of the alternative hypothesis only if sample evidence suggests that H_0 is false.
- The two possible conclusions from a hypothesis-testing analysis are *reject H_0* or *fail to reject H_0*.

Errors in Hypothesis Testing

- 利用隨機抽樣樣本來拒絕H_0
- 若無任何證據拒絕H_0，並非指H_0成立，只是說沒有足夠的證據去拒絕H_0
- 通常H_0的假設習慣用 "=" ">=" "<="
- H_a的假設習慣用 "/=" "<" ">"

- A random sample is used to "*reject* H_0"
- If we conclude 'do not reject H_0', this does not necessarily mean that the null hypothesis is true, it only suggests that there is not sufficient evidence to reject H_0. If we reject the null hypothesis, then it suggests that the alternative hypothesis may be true.
- Equality is always part of H_0 (e.g. "$=$" , "$\geq$" , "$\leq$"). "$\neq$" "$<$" and "$>$" is always part of H_1

Errors in Hypothesis Testing

- A decision rule used to determine whether H_0 should be rejected is called a **test procedure**.
- **Type I error** – is the error of rejecting H_0 when H_0 is actually true.
- **Type II error** – consists of *not* rejecting H_0 when H_0 is false.
- The probability of making a type I error is denoted by α and is called the **significance level** of the test.
- A test with α = .01 is said to have a significance level of .01.
- This is, if H_0 is actually true and the test procedure is used repeatedly on different samples selected from the population or process, in the long run H_0 would be incorrectly rejected only 1% of the time.
- The probability of a type II error is denoted by β.

Errors in Hypothesis Testing

- **Type I error**: 當H_0為真，但是卻拒絕H_0。發生此現象的機率為**α**，稱為**顯著水準**，1- α為**信心水準**
- **Type II error**: 當H_0為假，但是卻不拒絕H_0。發生此現象的機率為**β**，1-β定義為**檢定力**。

Type I error: (**偽陽**)

H_0~我不是孕婦 醫生卻說我是孕婦

Type II error: (**偽陰**)

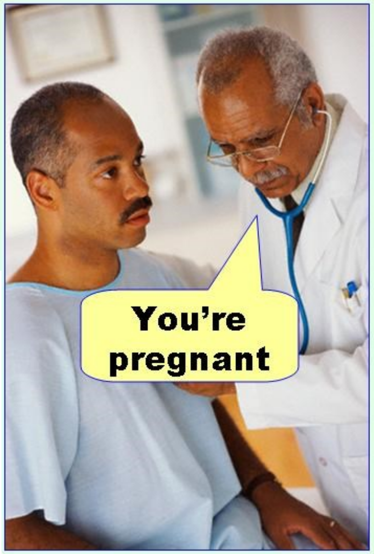
H_0~我不是孕婦 醫生也說我不是孕婦

但是我是孕婦

Type I error | Effect Size FAQs

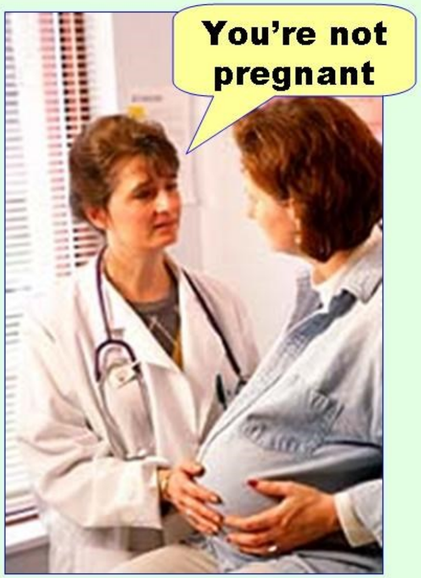
Type I error

(false positive)



Type II error

(false negative)



Scientist's Decision

	Reject null hypothesis	Fail to reject null hypothesis
Null hypothesis is true	Type 1 Error probability = α	Correct Decision Probability = 1- α
Null hypothesis is false	Correct decision probability = 1 - β	Type 2 Error probability = β

觀測顯著水準*p*-value

若***p*值足夠小**於設定之顯著水準(α)，則可以拒絕H_0

($1-\alpha$)為信心水準

Test Statistics & p-values

- A test of hypotheses is carried out by employing a **test statistic.**
- The ***p*-value,** or **observed significance level** (OSL), is the probability of obtaining a test statistic value at least as contradictory to H_0 as the value that actually resulted.
- The smaller the ***p*-value**, the more contradictory is the data to H_0.
- The null hypothesis should then be rejected if the ***p*-value** is sufficiently small.
- The following decision rule specifies a test with the desired significance level (type I error probability):
  - **Reject H_0 if p-value $\le \alpha$.**
  - **Do not reject H_0 if p-value $> \alpha$.**

觀測顯著水準*p*-value

以現有的抽樣所進行的推論，可能犯type I error的可能性 (若*p*-value越小，表示拒絕H_0不太可能錯，因此拒絕H_0)

Test Statistics & *p*-values

檢定方式(計算可能性)與對立假設條件H_a有關

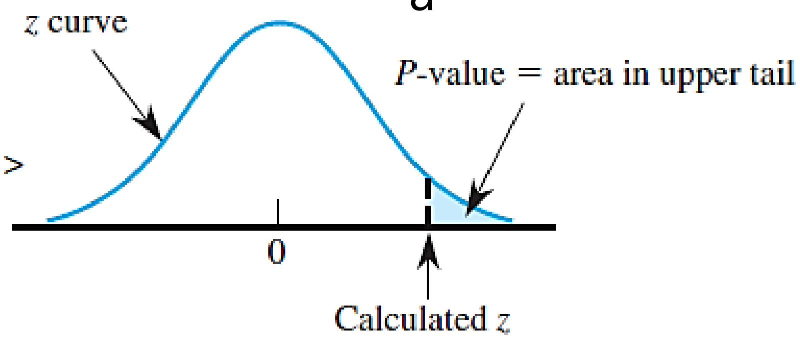
H_0為真時而拒絕H_0的可能性，其實就是對立假設H_a出現的可能性有關

Test Statistics & *p*-values

檢定方式(計算可能性)與對立假設條件H_a有關

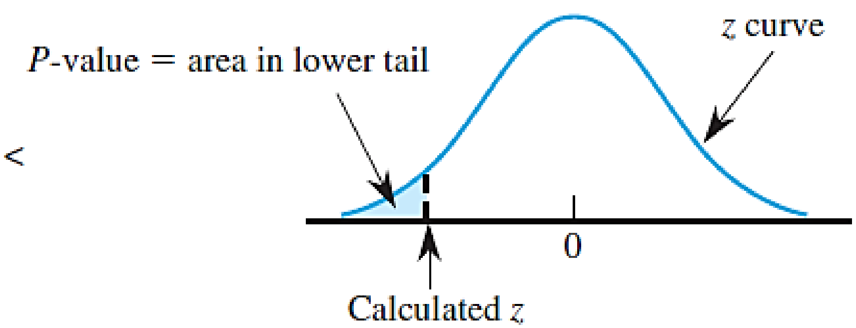
上（右）尾檢定

1. **Upper-tailed test**
   H_a contains the inequality $>$



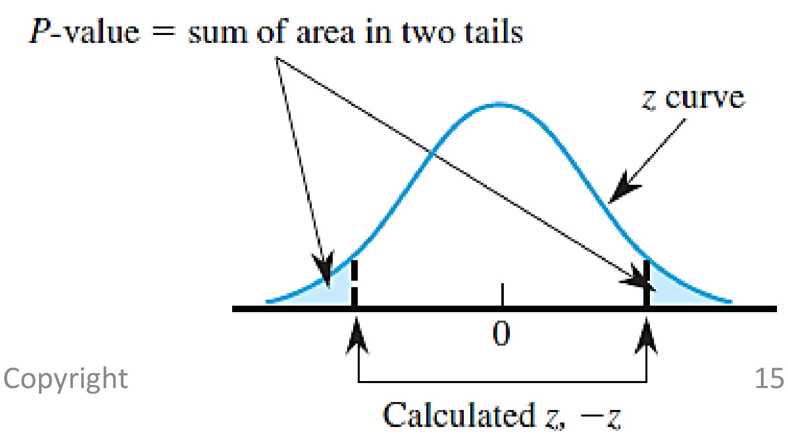
下（左）尾檢定

2. **Lower-tailed test**
   H_a contains the inequality $<$



雙尾檢定

3. **Two-tailed test**
   H_a contains the inequality $\neq$



Test Statistics & *p*-values

Lower-tailed test (左尾檢定)

The recommended daily dietary allowance (RDA) for zinc among males older than 50 years is 15 mg/day (*World Almanac*, 1992). The article "Nutrient Intakes and Dietary Patterns of Older Americans: A National Study" (*J. of Gerontology*, 1992: M145–M150) reported the following data on zinc intake for a sample of males age 65–74 years:

$$n = 115 \qquad \bar{x} = 11.3 \qquad s = 6.43$$

Does this data suggest that μ, the average daily zinc intake for the entire population of males age 65–74, is less than the RDA? The relevant hypotheses are

Figure 8.1 shows a boxplot of data consistent with the given summary quantities. Roughly 75% of the sample observations are smaller than 15 (the top edge of the box is at the upper quartile). Furthermore, the observed $\bar{x}$ value, 11.3, is certainly smaller than 15, but this could be just the result of sampling variability when H_0 is true. Is it plausible that a sample mean this much smaller than what was expected if H_0 were true occurred as a result of chance variation, or is $\mu < 15$ a better explanation for what was observed?

Test Statistics & *p*-values

樣本數n=115，可以使用常態分佈利用**z-curve**計算*p*值。

假設顯著水平α=0.01

- The appropriate test statistic for testing the stated hypotheses is

$$z = \frac{\bar{x} - 15}{s/\sqrt{n}} = \frac{11.3 - 15}{6.43/\sqrt{115}} = -6.17$$

- Values of z at least as contradictory to H_0 as this are those even smaller than –6.17

  p-value = $P(z < -6.17$ **when H_0 is true**)
  = area under the standard normal (z) curve to the left of –6.17
  ≈ 0

- If a significance level of .01 is used, then

$$P\text{-value} \approx 0 \leq .01 = \alpha$$

- So the null hypothesis should be **rejected**
- The data is much more consistent with the **conclusion that true average intake is in fact smaller than the RDA**

R: Hypothesis Testing

```
t.test(x,
mu=,
alternative=,
conf.level=).
```

R: Hypothesis Testing

alternative= c(‘two.sided’ , ‘less’ , ‘greater’)

R: Hypothesis Testing

conf.level= 0.95...)

TRY
it
in
R

R: Hypothesis Testing

R_testing_a.R

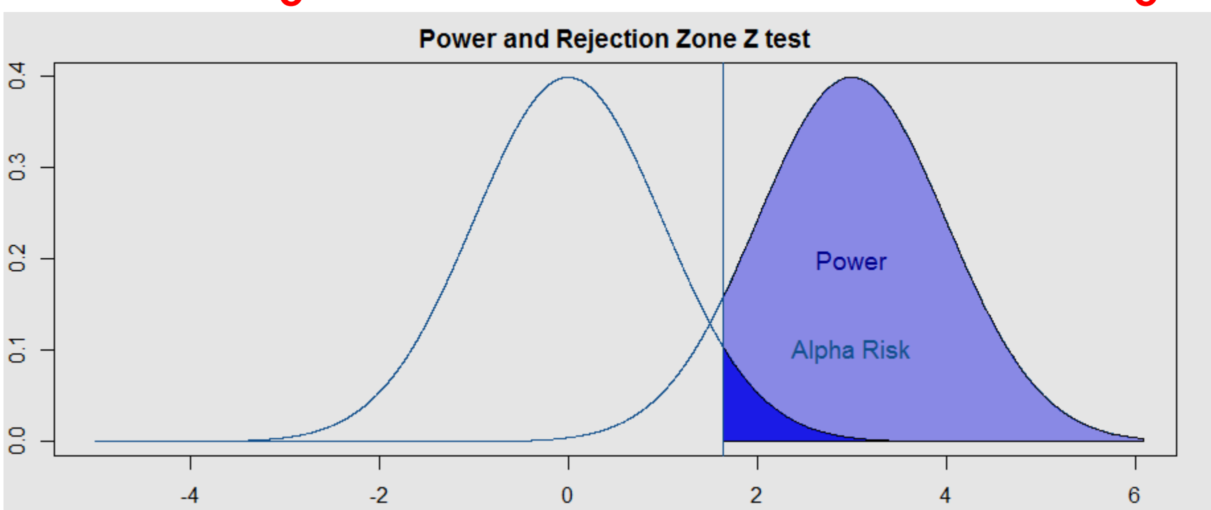
Test Statistics & p-values: power

Statistical Power (檢定力)=**(1-β)**

檢定力受到樣本數量(sample size)、變異大小(magnitudes of the variances)、顯著水平、兩個母體間的平均值差異

高檢定力(power 0.7~1.0):

代表若虛無假設H_0是false時檢定可以拒絕H_0的可能性



R: Hypothesis Testing

power.t.test(n,
delta=,
sd=,
sig.level=).

R: Hypothesis Testing

power.t.test(
power=NULL,
type=,
alternative=).

R: Hypothesis Testing

type=

c(‘two.sample’ , ‘one.sample’ , ‘paired’)

R: Hypothesis Testing

alternative=
c(‘two.sided’ ,
‘one.sided’)

TRY it in R

R: Hypothesis Testing

R_testing_a.R

Quick Summary

一般從批次樣本隨機抽樣的結果來做檢定。此時擁有幾個問題需要回答：

Q1. 如何提出檢定的假設 (H_0, H_a)?

Q2. 如何根據樣本資訊做出決策 (拒絕/無法拒絕)?

Q3. 做出這一決策有可能犯什麼錯誤 (α,β)?

Q4. 決策的結論該如何解釋?

Tests Concerning a Difference Between two Means: Independent Samples

一般用來檢驗兩個母體、程序的差異

- Hypothesis testing is often used as a basis for comparing two populations, processes, or treatments.

The Two-Sample *t* Test

- Null hypothesis: H_0: $\mu_1 - \mu_2 = \Delta$ (Δ denotes the null value)
- Test statistic: $t = \dfrac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$
- *p*-value: When H_0 is true, the test statistic has approximately a *t* distribution with

$$df = \frac{[(se_1)^2 + (se_2)^2]^2}{\dfrac{(se_1)^4}{n_1 - 1} + \dfrac{(se_2)^4}{n_2 - 1}}$$

- Where $se = s / \sqrt{n}$

(Note: df rounded down to nearest integer.)

Tests Concerning a Difference Between two Means: Independent Samples

兩組隨機樣本互相獨立，且皆來自常態分佈的母體，如果樣本個數n皆大於30，則無需常態分佈的假設

可使用z-test

Assumptions:

- The two random samples are selected independently, both from underlying normal population, process, or treatment response distributions.
- If the sample sizes are large (usually both $n_1 > 30$ and $n_2 > 30$ will suffice), the Central Limit Theorem implies that the normality assumption is no longer necessary. In this case, the test statistic can be denoted by *z*, and the *p*-value calculated by reference to the *z* curve.

Tests Concerning a Difference Between two Means: Independent Samples: An Example

市區管線惡化情況: 案例探討Fusion Process對於管線張力強度影響(tensile strength)

Deterioration of municipal pipeline networks

- The data on tensile strength (psi) of liner specimens both when a certain fusion process was used and when this process was not used:

1. No fusion:	2748	2700	2655	2822	2511			
	3149	3257	3213	3220	2753			
	$n_1 = 10$	$\bar{x}_1 = 2902.8$	$s_1 = 277.3$	$se_1 = 87.69$				
2. Fused:	3027	3356	3359	3297	3125	2910	2889	2902
	$n_2 = 8$	$\bar{x}_2 = 3108.1$	$s_2 = 205.9$	$se_2 = 72.80$				

Tests Concerning a Difference Between two Means: Independent Samples: An Example

The deterioration of many municipal pipeline networks across the country is a growing concern. One technology proposed for pipeline rehabilitation uses a flexible liner threaded through existing pipe. The article "Effect of Welding on a High-Density Polyethylene Liner" (*J. of Materials in Civil Engr.*, 1996: 94–100) reported the following data on tensile strength (psi) of liner specimens both when a certain fusion process was used and when this process was not used:

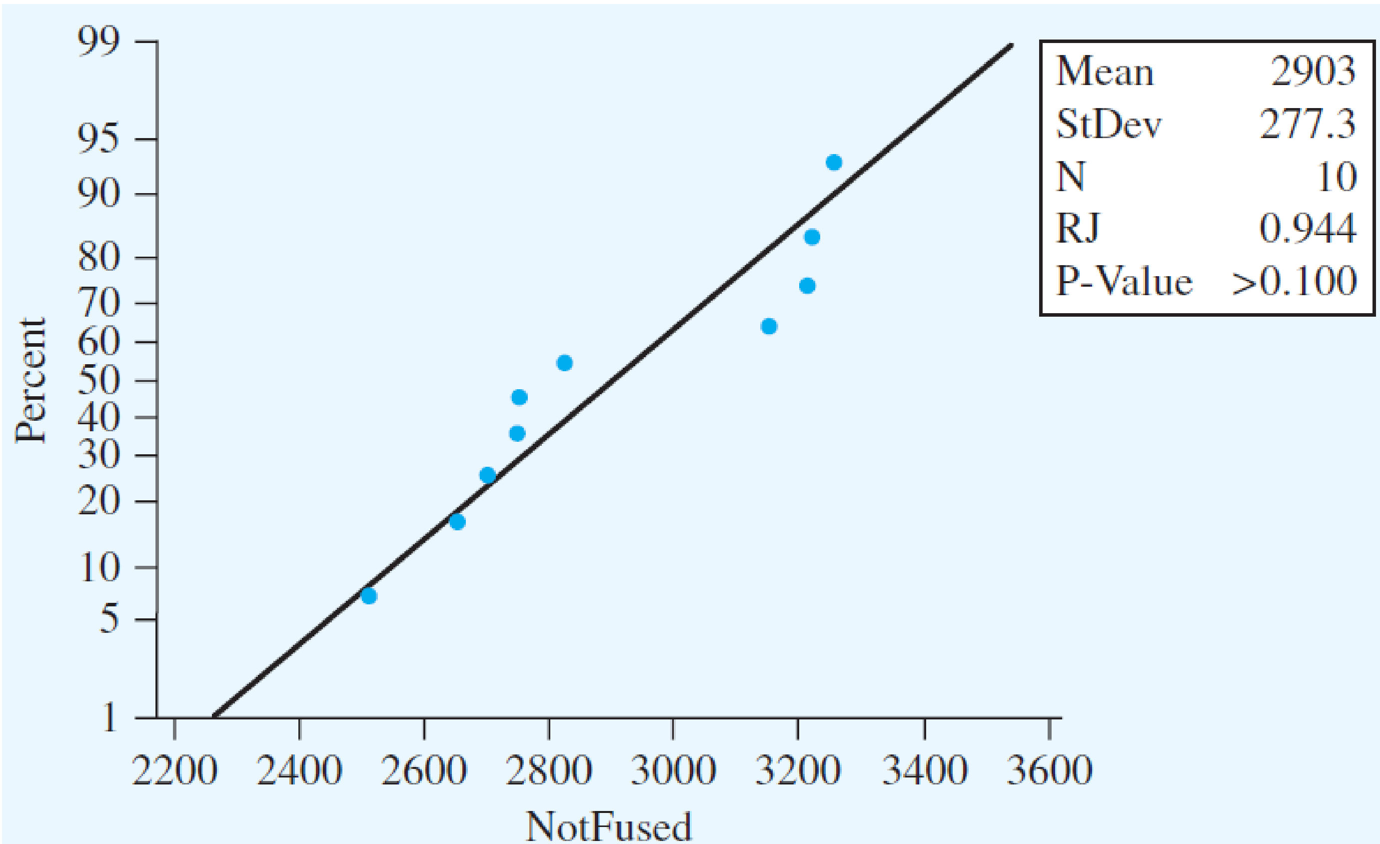
1. No fusion: 2748 2700 2655 2822 2511
   3149 3257 3213 3220 2753
   $n_1 = 10 \quad \bar{x}_1 = 2902.8 \quad s_1 = 277.3 \quad se_1 = 87.69$
2. Fused: 3027 3356 3359 3297 3125 2910 2889 2902
   $n_2 = 8 \quad \bar{x}_2 = 3108.1 \quad s_2 = 205.9 \quad se_2 = 72.80$

Figure 8.5 shows *normal probability plots* from Minitab. These plots employ a probability scale rather than the normal quantiles discussed previously, but the critical issue is the same: Is the pattern of plotted points reasonably close to linear? There certainly is some wiggling in these plots, but not enough to suggest that the normality assumption is implausible. Furthermore, the *P*-values that appear along with the plots are for formal tests of the assertion that the underlying distributions are normal (we discuss this test in Section 8.4). Because each *P*-value exceeds .1, the hypothesis of normality cannot be rejected.

Tests Concerning a Difference Between two Means: Independent Samples: An Example

Is the distribution Normal? 常態分佈檢定

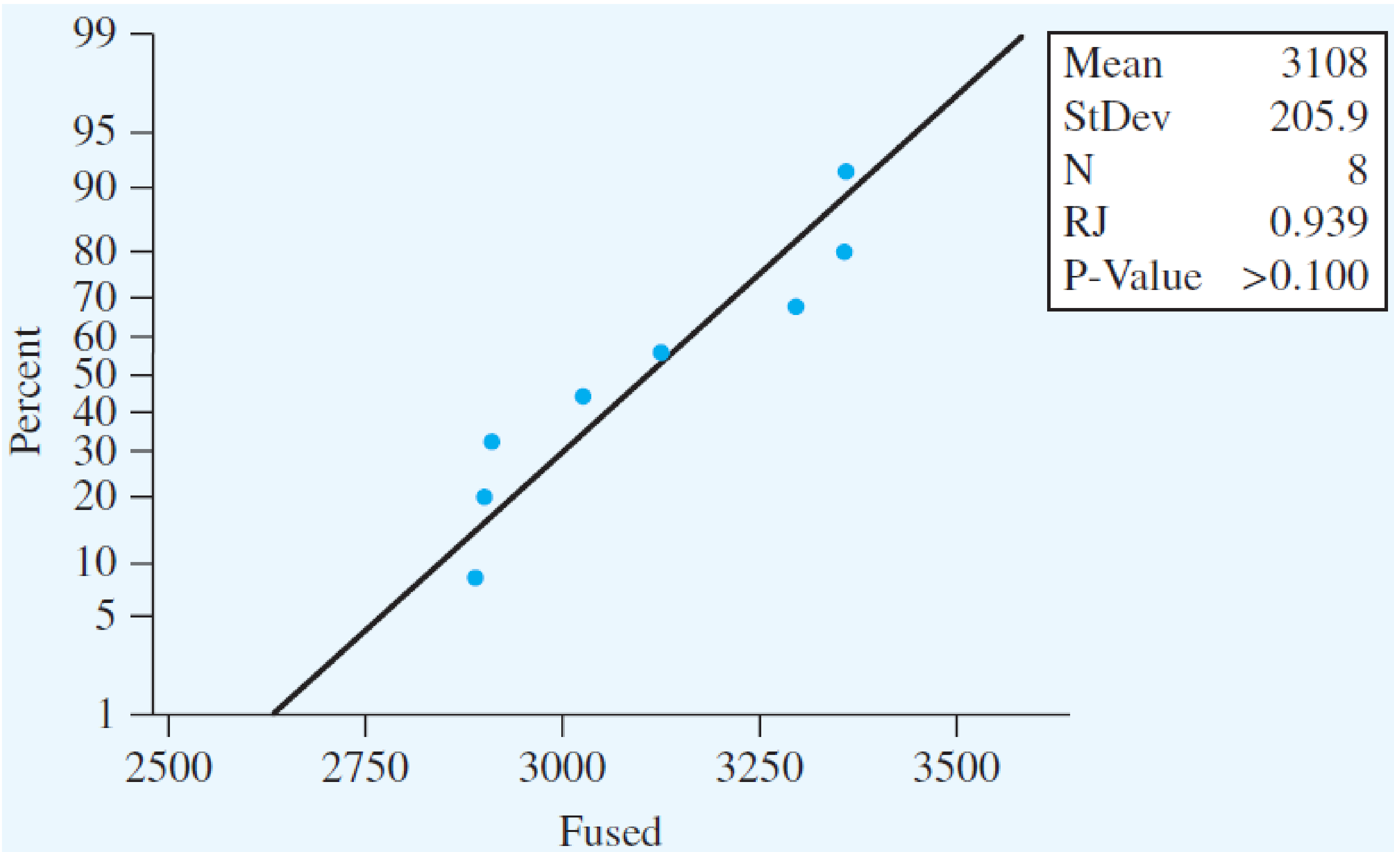
Now the question is: Is the pattern of plotted points reasonably close to linear? (Ryan-Joiner test)



Tests Concerning a Difference Between two Means: Independent Samples: An Example

Is the distribution Normal? 常態分佈檢定

Now the question is: Is the pattern of plotted points reasonably close to linear? (Ryan-Joiner test)



Tests Concerning a Difference Between two Means: Independent Samples: An Example

檢查樣本資料型態特性，並設計虛無假設H_0

1. Let μ_1 be the true average tensile strength of specimens when the no-fusion treatment is used and μ_2 denote the true average tensile strength when the fusion treatment is used.
2. $H_0: \mu_1 - \mu_2 = 0$ (no difference in the true average tensile strengths for the two treatments)
3. $H_a: \mu_1 - \mu_2 < 0$ (true average tensile strength for the no-fusion treatment is less than that for the fusion treatment, so the investigators' conclusion is correct)

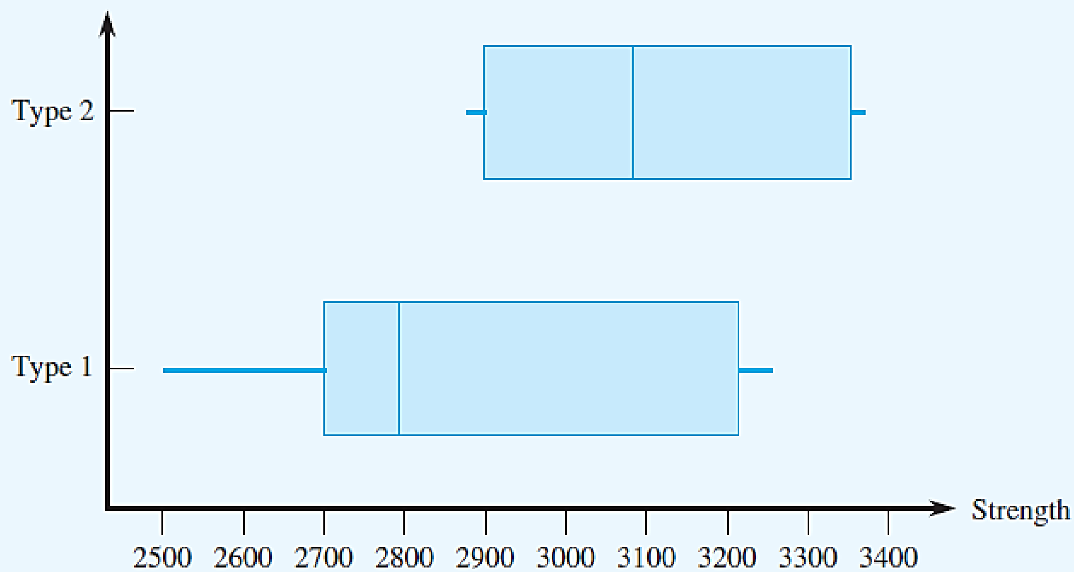


圖 8.6 強度數據的比較盒型圖。

Tests Concerning a Difference Between two Means: Independent Samples: An Example

4. The null value is $\Delta = 0$, so the test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

5. We now compute both the test statistic value and the df for the test:

$$t = \frac{2902.8 - 3108.1}{\sqrt{\frac{(277.3)^2}{10} + \frac{(205.9)^2}{8}}} = \frac{-205.3}{113.97} = -1.8$$

$$\text{df} = \frac{[(87.69)^2 + (72.80)^2]^2}{(87.69)^4/9 + (72.80)^4/7} = 15.94$$

so the test will be based on 15 df.

6. Appendix Table VI shows that the area under the 15 df t curve to the right of 1.8 is .046, so the P-value for a lower-tailed test is also .046. The following Minitab output summarizes all the computations:

```
Twosample T for nofusion vs fused

            N    Mean    StDev  SE Mean
nofusion   10    2903     277       88
fused       8    3108     206       73

95% C.I. for mu nofusion-mu fused: (-448, 38)
T-Test mu nofusion = mu fused (vs <): T= - 1.80 P = 0.046 DF=15
```

7. Using a significance level of .05, we can barely reject the null hypothesis in favor of the alternative hypothesis, confirming the conclusion stated in the article. However, someone demanding more compelling evidence might select $\alpha = .01$, a level for which H_0 cannot be rejected.

p-value = 0.046 > α = 0.01

無法拒絕H_0

p-value = 0.046 < α = 0.05

勉強拒絕H_0代表不同程序還是會造成強度上的差異

R: Shapiro-Wilk normality test

shapiro.test(x).

R: Test to compare Two Varances

var.test (x,y,ratio =1).

R: Test to compare Two Varances

t.Test (

x,y,alternative = "less"

R: Test to compare Two Varances

, var.equal =

, conf.level = 0

).

TRY it in R

R: Hypothesis Testing

R_testing_b.R

Tests Concerning a Difference Between two Means: Paired Data

The Paired *t* Test

- Null hypothesis: H_0: $\mu_d = \Delta$, μ_d denotes the population mean difference
- Test statistic: $$t = \frac{\bar{d} - \Delta}{s_d / \sqrt{n}}$$
- ***p*-value**: Calculated from the *t* curve with $n - 1$ df as described previously. The test is upper-tailed, lower-tailed, or two-tailed, depending on whether the inequality in H_a is >, <, or ≠, respectively.
- Assumptions: The sample differences $d_1, \ldots, d_n$ have been randomly selected from a difference population having a normal distribution. If *n* is large, the normality assumption is not necessary; the test statistic is labeled *z*, and the ***p*-value** is determined from the *z* curve.

Tests Concerning a Difference Between two Means: Paired Data: An Example

不同工作環境條件對於手臂的影響，抽樣樣本數量為16

Example 8.7

Musculoskeletal neck-and-shoulder disorders are all too common among office staff who perform repetitive tasks using visual display units. The article "Upper-Arm Elevation During Office Work" (*Ergonomics*, 1996: 1221–1230) reported on a study to determine whether more varied work conditions would have any impact on arm movement. The accompanying data was obtained from a sample of $n = 16$ subjects. Each observation is the amount of time, expressed as a proportion of total time observed, during which arm elevation was below 30°. The two measurements from each subject were obtained 18 months apart. During this period, work conditions were changed, and subjects were allowed to engage in a wider variety of work tasks. Does the data suggest that true average time during which elevation is below 30° differs after the change from what it was before the change?

Subject:	1	2	3	4	5	6	7	8
Before:	81	87	86	82	90	86	96	73
After:	78	91	78	78	84	67	92	70
Difference:	3	−4	8	4	6	19	4	3

Subject:	9	10	11	12	13	14	15	16
Before:	74	75	72	80	66	72	56	82
After:	58	62	70	58	66	60	65	73
Difference:	16	13	2	22	0	12	−9	9

Figure 8.7 shows a normal probability plot of the 16 differences; the pattern in the plot is quite straight, supporting the normality assumption. A boxplot of these differences appears in Figure 8.8; the boxplot is located considerably to the right of zero, suggesting that perhaps $\mu_d > 0$ (note also that 13 of the 16 differences are positive and only two are negative).

Tests Concerning a Difference Between two Means: Paired Data: An Example

Example 8.7: Determining the impact of varied work conditions on arm movement.

- The data was obtained from a sample of n = 16 subjects.

Subject:	1	2	3	4	5	6	7	8
Before:	81	87	86	82	90	86	96	73
After:	78	91	78	78	84	67	92	70
Difference:	3	−4	8	4	6	19	4	3

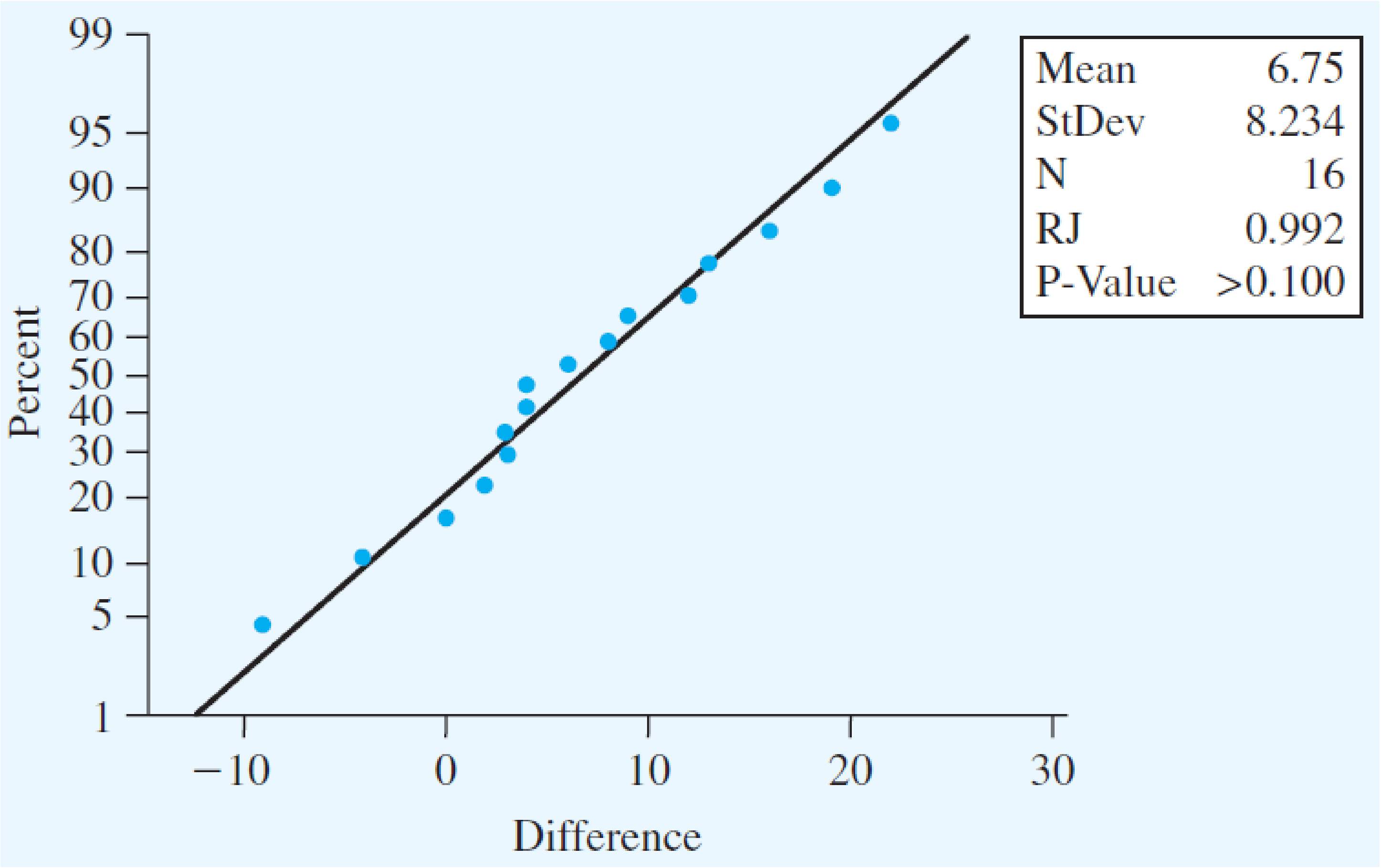
Subject:	9	10	11	12	13	14	15	16
Before:	74	75	72	80	66	72	56	82
After:	58	62	70	58	66	60	65	73
Difference:	16	13	2	22	0	12	−9	9

18個月後將手臂抬高30度的平均工作時間

- Each observation is the amount of time, expressed as a proportion of total time observed, during which arm elevation was below 30°.
- The two measurements from each subject were obtained <u>18</u> months apart.
- During this period, work conditions were changed, and subjects were allowed to engage in a wider variety of work tasks.

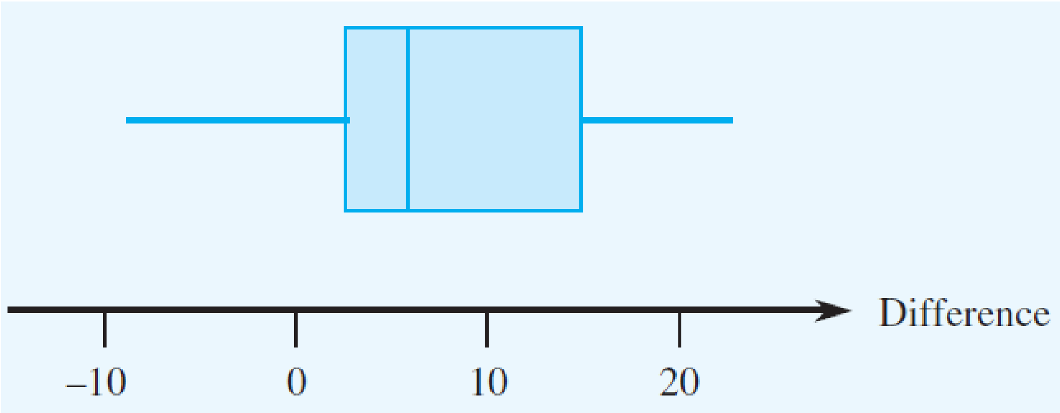
Tests Concerning a Difference Between two Means: Paired Data: An Example

Normality Test



Tests Concerning a Difference Between two Means: Paired Data: An Example

檢查樣本資料型態特性，並設計虛無假設H_0



- Does the data suggest that true average time during which elevation is below 30° differs after the change from what it was before the change? Supporting difference > 0
- We use the recommended sequence of steps to test the appropriate hypotheses.

Tests Concerning a Difference Between two Means: Paired Data: An Example

手部姿勢的不同，對於平均工作時數確實有影響

- Let μ_d denote the true average difference between elevation time before the change in work conditions and time after the change.
- H_0: $\mu_d = 0$
- H_a: $\mu_d \neq 0$

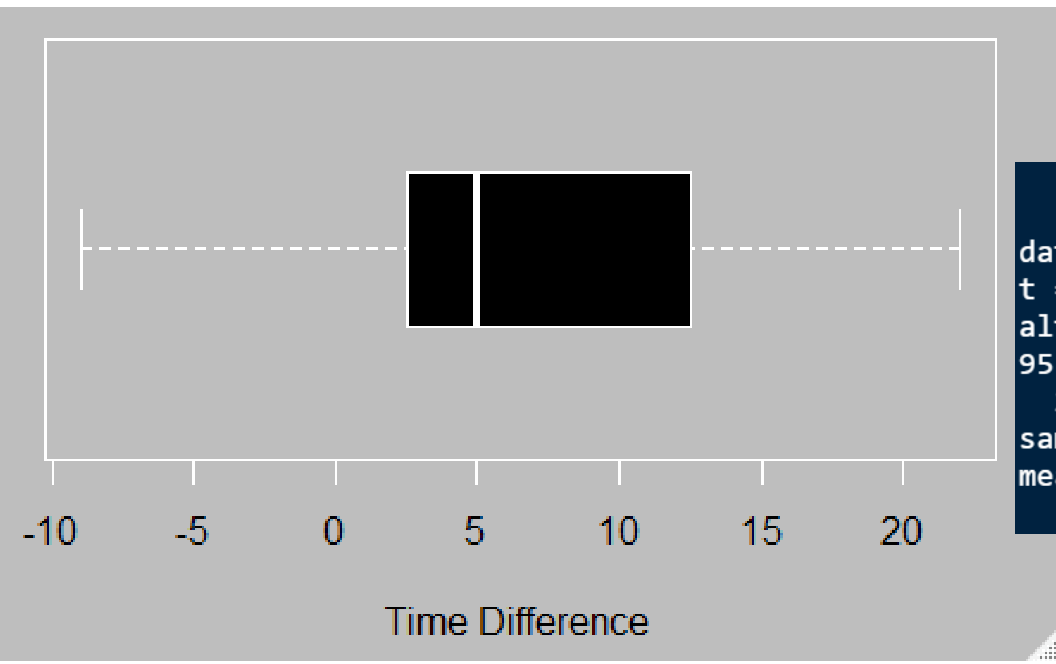
$$t = \frac{\bar{d} - 0}{s_d / \sqrt{n}} = \frac{\bar{d}}{s_d / \sqrt{n}}$$

$$n = 16, \sum d_i = 108, \sum d_i^2 = 1746, \bar{d} = 6.75, s_d = 8.234, t \approx 3.3$$

- The area to the right of 3.3 under the t curve with 15 df is .002 (see Appendix Table VI).
- The inequality in H_a implies that a two-tailed test is appropriate, so the P-value is ~.004 (Minitab gives .0051).
- <u>Since .004 < .01, the null hypothesis can be rejected at either significance level .05 or .01.</u>
- The true average difference between times is nonzero; so the true average time after the change is different from that before the change.

Tests Concerning a Difference Between two Means: Paired Data: An Example

```r
#7.
# --------paired data t-test------------------------
pair <- c(3,-4,8,4,6,19,4,3,16,13,2,22,0,12,-9,9)
# normmality test
shapiro.test(pair)
# boxplot
par(col.axis='black',bg ='gray',fg = 'white'
    , mai = c(0.9,0.9,0.7,0.3), cex = 0.5, cex.lab = 1.0 ,cex.main = 1)
par(mfrow = c(1,1))
boxplot(pair,xlab = 'Time Difference', col = 'black',
        horizontal = TRUE)
# paired t test
t.test(pair, alternative = 'two.sided')
```




```
        One Sample t-test

data:  pair
t = 3.2791, df = 15, p-value = 0.005072
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  2.362371 11.137629
sample estimates:
mean of x
     6.75
```

課堂練習: 學號-姓名-ch11-testing.R

data(iris)鳶尾花資料

種類setosa及versicolor的萼片(sepal)長度資料是否存在差異? 如果有差異其現象為何?區間估計為何?

(考慮顯著水平5%, H_0: $\mu_1 - \mu_2 = 0$; H_a: $\mu_1 - \mu_2 \neq 0$)

請參考以下步驟進行分析:

(0) data(iris) (讀入內建資料庫，若不知道如何整理資料，可以直接載入檔案 **iris_setosa_versicolor_SepalLength.csv**)

(1) shapiro.test() (Normality Test常態分佈檢定)

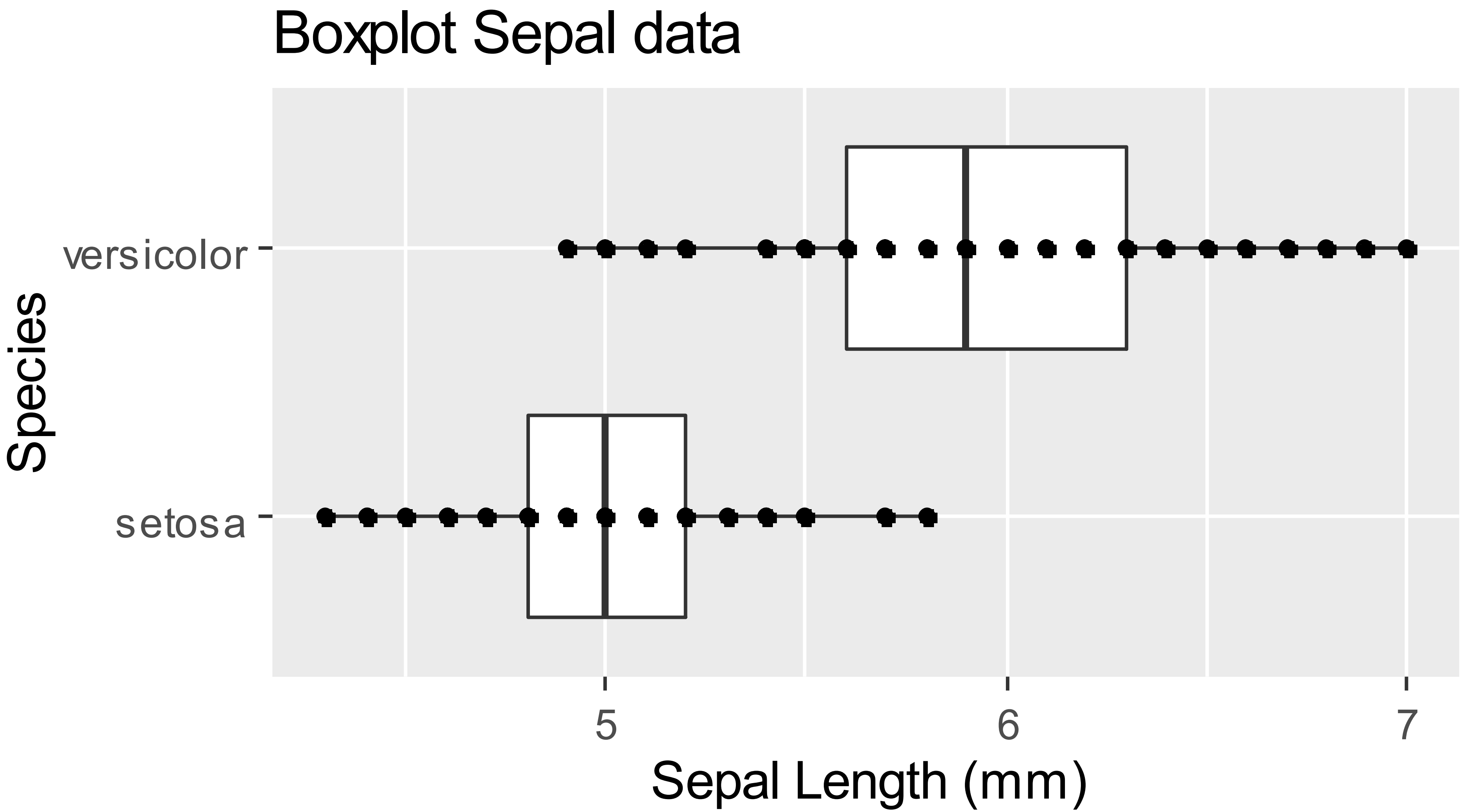
(2) geom_boxplot() (Box Plot盒鬚圖)

(3) Test Statistics (Two sample z test, two-sided)

R: Useful function

```
z.test(x,y,
alternative =
 "two.sided" ,
sigma.x =,
sigma.y =,
mu = 0, conf.level =).
```

課堂練習: 學號-姓名-ch10-estimation.R



課堂練習: 學號-姓名-ch10-estimation.R

```
        Two-sample z-Test

data:  d.versicolor and d.setosa
z = 10.521, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.7567495 1.1032505
sample estimates:
mean of x mean of y
    5.936     5.006
```