107-1 Statistics
2017年考古題講解

助教：廖晧宇、吳家禎、賴冠宇

2018/11/23

1.

設台大學生中，外籍學生的比例為 12%。若隨機調查 400 位台大學生，令 p 表 此 400 人為外籍學生的比例。求 p 的抽樣誤差不超過 3%的近似機率。(15%)

Sample error is equal to or smaller than 3% means that $\hat{p}$ is between 0.09~0.15.

```
p <- 0.12
n <- 400
s.e. <- sqrt(p*(1-p)/n)
pLeft <- pnorm(0.09, mean=p, sd=s.e.)
probability <- 2*(0.5-pLeft)
probability
```

```
## [1] 0.9351618
```

The probability is about 93.52%.

2.

台大學生會發起最新的政見規劃，力求能夠在校內設置公共自行車租借站。讓 同學不用每人都購買一台腳踏車，減輕經濟負擔，同時也可以舒緩校內停車空 間不足的問題。根據過往的經驗已知，台大學生贊成於校內設立公共自行車租 借站的比例是 62%。請在顯著水準為 0.05 的情況下，回答下列問題:

```
p <- 0.62
alpha <- 0.05
```

2 (a)

假設學生意見無改變，若現在隨機抽樣調查 400 人，則在此 400 人中贊成的比例在 64.5%以上的機率是多少?(10%)

```
n <- 400
s.e. <- sqrt(p*(1-p)/n)
p_hat <- 0.645
z <- (p_hat-p)/s.e.
probability <- 1 - pnorm(z)
probability
```

$$s.e. = \sqrt{\frac{0.62 \times (1 - 0.62)}{400}}$$

$$z = \frac{0.645 - 0.62}{s.e.}$$

```
## [1] 0.1514799
```

The probability is about 15.15%.

2 (b)

假設實際進行抽樣調查所獲得的結果為贊成比例恰巧等於 64.5%。而台大校 方宣稱因為調查的贊成比例與過往差不多，顯見學生意願並未提高，所以暫 緩此議題的討論。請問校方是否能夠這樣宣稱?(5%)

```r
p <- 0.62
alpha <- 0.05
```

5 steps for hypotheses test:

Step1

H0: p = 0.62 ; H1: p > 0.62

Step2

The z-statistics has been calculated in (a).

```r
z
```

$$z = \frac{0.645-0.62}{s.e.}$$

```
## [1] 1.030107
```

Step3

$$p\ value = 1\text{-pnorm}(z) = 0.15148$$

The p-value is the answer of (a).

Step4

Because the p.value is larger than 0.05, we cannot reject H0

Step5

Conclusion: we cannot reject the claim from NTU.

2 (c)

為了能讓議題獲得充分的討論，學生會決定要在第二次抽樣調查中增加調查 的樣本數，好讓統計檢定的結果為顯著。請問在贊成比例維持 64.5% 的假設 下，學生會至少要調查多少的樣本數才能達到目的?(5%)

```
z_alpha <- qnorm(1-alpha)
n_new <- (p*(1-p))/(((p_hat-p)/z_alpha)^2)
n_new
```

$$\sqrt{\frac{p(1-p)}{n}} = \frac{\hat{p}-p}{z_\alpha}$$

```
## [1] 1019.882
```

$$n = \frac{p(1-p)}{\left(\frac{\hat{p}-p}{z_\alpha}\right)^2}$$

The sample number must be at least 1020.

3.

隨著核電廠停用以及氣候變遷等因素，空氣汙染近年來逐漸成為國人的重要健康議題。根據過去的氣象觀測資料，表1及表2分別為台北市和高雄市近2年來（100週）的每週紫爆日數紀錄。請根據此資料在顯著水準為0.05的情況下，回答下列問題：

表 1 台北市每週紫爆日數紀錄

每週紫爆日數	0	1	2	3	4	5	6	7
週數	20	36	27	11	3	1	1	1

表 2 高雄市每週紫爆日數紀錄

每週紫爆日數	0	1	2	3	4	5	6	7
週數	2	13	26	29	19	8	2	1

3 (b)

假設中央氣象局定義一週內出現4天以上（包含4天）的紫爆日，該週即為空汙週。~~請問在近2年中，台北市和高雄市的空汙週比例是否存在顯著差異？~~

以近2年的資料為樣本來推論，常年中台北市及高雄市的空汙比例是否存在差異？

```
p_hat1 <- 6/100 #taipei
p_hat2 <- 30/100 #kaohsiung
n1 <- 100
n2 <- 100
```

5 steps for hypotheses test:

1. $H_0: p_1 - p_2 = 0$
   $H_1: p_1 - p_2 \neq 0$
2. **Calculate z-statistics**

```
p_hat <- (n1*p_hat1+n2*p_hat2)/(n1+n2)
s.e. <- sqrt(p_hat*(1-p_hat)*(1/n1 + 1/n2))
z <- (p_hat1-p_hat2)/s.e.
z
```

```
## [1] -4.417261
```

3. **Calculate p-value**

```
p_value <- 2*(1-pnorm(abs(z)))
p_value
```

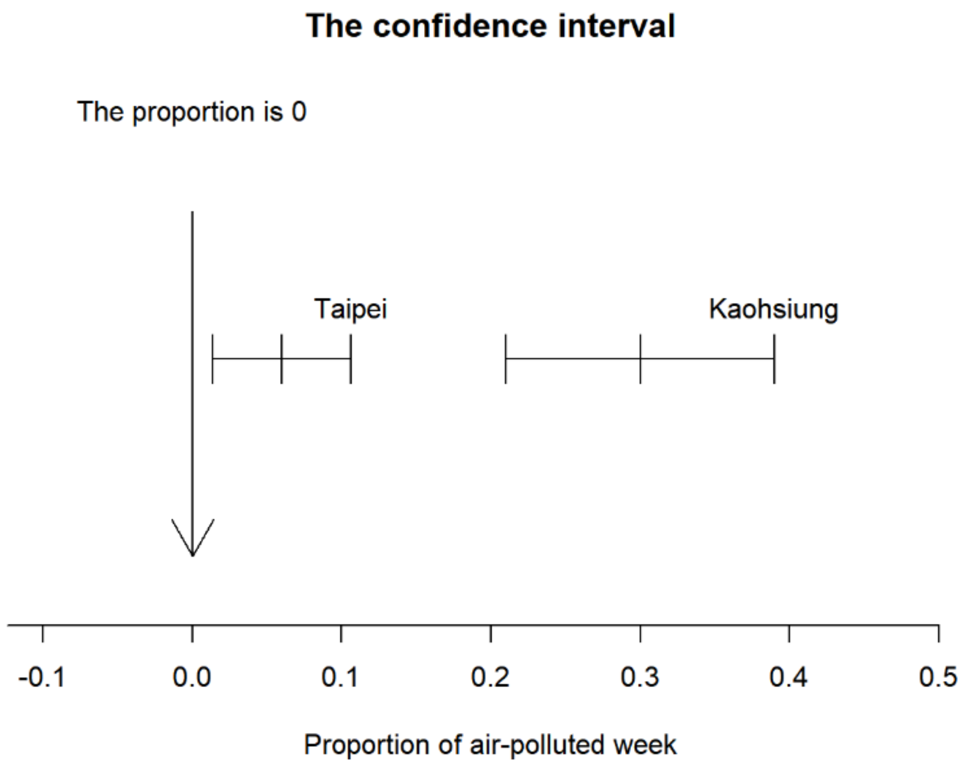
```
## [1] 9.995948e-06
```

4. **Because the p.value is smaller than 0.05, we can reject H_0**
5. **Conclusion: the proprotion of air-polluted week is different between Taipei and Kaohsiung.**

3(b)的Ans: 台北市及高雄市的空汙週比例存在差異。

3 (c)

呈上題，中央氣象局局長希望可以有圖表作為向外界解說時的輔助素材。請利用繪製信賴區間的方式，繪製台北市和高雄市的空汙週比例信賴區間於同一張圖表，並與(b)小題的答案做比較和說明。



According to the plot, the difference of the proportion of air-polluted week between Taipei and Kaohsiung may exists, which is consistent with the concludion of (b).

3 (c) cont.

```
s.e.1 <- sqrt(p_hat1*(1-p_hat1)/n1) #Taipei
s.e.2 <- sqrt(p_hat2*(1-p_hat2)/n2) #Kaohsiung
z <- qnorm(0.975)
up1 <- p_hat1 + z*s.e.1
low1 <- p_hat1 - z*s.e.1
up2 <- p_hat2 + z*s.e.2
low2 <- p_hat2 - z*s.e.2

plot(c(0,1),type='n',xlim=c(-0.1,0.5),xlab='Proportion of air-polluted week', ylab='',axes=FALSE,main='The confidence interv
al')
lines(c(low1,up1),c(0.5,0.5))
lines(c(low1,low1),c(0.45,0.55))
lines(c(p_hat1,p_hat1),c(0.45,0.55))
lines(c(up1,up1),c(0.45,0.55))
text(up1,0.6,label='Taipei')

lines(c(low2,up2),c(0.5,0.5))
lines(c(low2,low2),c(0.45,0.55))
lines(c(p_hat2,p_hat2),c(0.45,0.55))
lines(c(up2,up2),c(0.45,0.55))
text(up2,0.6,label='Kaohsiung')

arrows(x0=0,y0=0.8,x1=0,y1=0.1)
text(0,1 ,label='The proportion is 0')


axis(1,seq(-0.5,0.5,by=0.1))
```

4.

台北市政府衛生局發言人指出，台北市的抽菸人口比例不超過三成。某公衛團隊針對此議題進行台北市抽菸人口比例的抽樣調查。抽樣調查400人，其中吸菸者有131人。請問：

p = 台北市的抽菸人口比例(母體比例)

p.hat = 抽樣調查中的台北市抽菸人口比例(樣本比例)

H0: p <= 0.3

H1: p > 0.3

右尾檢定

4 (a)

在0.05的顯著水準下，利用統計檢定檢驗台北市政府衛生局發言人的言論是否恰當？

```r
alpha = 0.05

n = 400
p.hat = 131/n
p = 0.3

se = sqrt(p*(1-p)/n)
z = (p.hat - p)/se; print(paste0("z = ", toString(z)))
```

```
## [1] "z = 1.20019839629796"
```

```r
p.value = 1-pnorm(z); print(paste0("p.value = ", toString(p.value))) # don't reject H0
```

```
## [1] "p.value = 0.115031149013198"
```

```r
if (p.value < alpha) {
  print("Reject H0.")
}else {
  print("Don't reject H0.")
  }
```

4(a)的Ans: 該發言人的言論應為恰當。

```
## [1] "Don't reject H0."
```

Ans: 根據統計檢定結果判斷，該發言人的言論應為恰當。

4 (b)

若台北市實際的抽菸人口比例為35%，~~則該團隊的推論犯錯機率為何？~~

1. 台北市的實際抽菸人口比例: p = 0.35 –> **虛無假設為假，對立假設為真**
2. (a)小題的檢定結果 –> **無拒絕虛無假設**

當該團隊下了(a)小題的結論時，其犯Type II error的機率為何？

根據1.及2.，該團隊的犯錯機率為犯下type II error的機率。

在0.05信心水準下的統計檢定，不拒絕域為: z.thres = 1.644854 –> x.thres = 0.3275

```
z.thres = qnorm(1-0.05); print(paste0("z.thres = ", toString(z.thres)))
```

```
## [1] "z.thres = 1.64485362695147"
```

```
# 從z = 1.644854 回推未標準化的值(x.thres):
x.thres = z.thres*se + p; print(paste0("x.thres = ", toString(x.thres)))
```

```
## [1] "x.thres = 0.337688331263139"
```

在實際母體下的樣本分布:

- mean = 0.35
- se = sqrt(0.35*0.65/400)

4 (b) cont.

在實際母體下的樣本分布:

- mean = 0.35
- se = sqrt(0.35*0.65/400)

```
#H1 sample distribution
h1.mean = 0.35
h1.se = sqrt(h1.mean*(1-h1.mean)/400)

type2error = pnorm(x.thres, mean = h1.mean, sd = h1.se); type2error
```

```
## [1] 0.3028415
```

Ans: 該團隊的犯錯機率為0.3028415。

4(b)的Ans: 發生type II error的機率為0.3028。

4 (c)

該團隊進一步對新北市做調查，抽樣600人當中吸菸者有225人。根據該研究團隊在兩個城市的抽樣調查結果，台北市與新北市的抽菸人口比例是否相同？

台北市與新北市的抽菸人口比例是否相同? –> **兩個母體比例的雙尾檢定**

p1 = 台北市的抽菸人口比例(母體比例)
p.hat1 = 抽樣調查中的台北市抽菸人口比例(樣本比例)
n1 = 台北市的抽樣人數

p2 = 新北市的抽菸人口比例(母體比例)
p.hat2 = 抽樣調查中的新北市抽菸人口比例(樣本比例)
n2 = 新北市的抽樣人數

H0: p1 - p2 = 0
H1: p1 - p2 != 0

雙尾檢定

4 (c) cont.

```r
n1 = 400; p.hat1 = 131/400
n2 = 600; p.hat2 = 225/600

# 綜合樣本比例pp
pp = (n1*p.hat1+n2*p.hat2)/(n1+n2); print(paste0("pp = ", toString(pp)))
```

```
## [1] "pp = 0.356"
```

```r
se = sqrt(pp*(1-pp)*(1/n1 + 1/n2))

z = ((p.hat1 - p.hat2)-0) / se; print(paste0("z = ", toString(z)))
```

```
## [1] "z = -1.53684935014695"
```

```r
p.value = pnorm(z); print(paste0("p.value = ", toString(p.value)))
```

```
## [1] "p.value = 0.0621651025950764"
```

$\mathbf{p-value = 0.0622 \times 2 = 0.1244}$
(與$\boldsymbol{\alpha} = \mathbf{0.05}$比較)

4 (c) cont.

```
half.alpha = 0.05/2

if (p.value < half.alpha) {
  print("Reject H0.")
}else {
  print("Don't reject H0.")
  }
```

$$\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$$

(與 $\frac{p - value}{2} = 0.0622$比較)

```
## [1] "Don't reject H0."
```

Ans: 根據統計檢定結果判斷，台北市與新北市的抽菸人口比例相同。

4(c)的Ans: 台北市與新北市的抽菸人口比例相同。

106-1期末考第一大題

1. 某研究員要採購實驗儀器的電池 。電池有原廠與副兩種選擇，其中的價格較高，故研究員想要以統計檢定評估是否購買原廠電池 。若原廠電池的續航力可超過副廠電池的150小時以上，則購買原廠電池較為划算。根據過去抽樣調查資料，原廠與副電池的續航力整理如下表 1：

(a) 請問在 85%的信心水準下，原廠及副電池各自的信賴區間為何？ (10%)

(b) 請以 0.05的顯著水準進行統計檢定，並推論該研究員應購買原廠或副電池較為划算？ (15%)

表 1.

	樣本數	平均續航力	標準差
原廠電池	20	1435	94
副廠電池	10	1209	68

(A) 請問在 85%的信心水準下，原廠及副電池各自的信賴區間為何？(10%)

1. 根據情境，選擇適當的方式：

   Calculating a Confidence Interval for a Population Mean

$$\bar{x} \pm t^{*} \times s.e.(\bar{x}) \Rightarrow \bar{x} \pm t^{*} \times \frac{s}{\sqrt{n}}$$

2. 計算需要的參數：

   t*、standard error

```{r}
n1 = 20; x1 = 1435; sd1 = 94
n2 = 10; x2 = 1209; sd2 = 68

alpha = 1-0.85
half.alpha = alpha/2

t.star1 = qt(half.alpha, df = n1-1, lower.tail = F)
t.star2 = qt(half.alpha, df = n2-1, lower.tail = F)
```

原廠電池

```
CI.upper1 = x1 + t.star1*(sd1/sqrt(n1)); CI.upper1
CI.lower1 = x1 - t.star1*(sd1/sqrt(n1)); CI.lower1
```

副廠電池

```
CI.upper2 = x2 + t.star2*(sd2/sqrt(n2)); CI.upper2
CI.lower2 = x2 - t.star2*(sd2/sqrt(n2)); CI.lower2
```

Ans:
原廠電池CI = [1403.467, 1466.533]；
副廠電池CI = [1175.159, 1242.841]

(B) 請以 0.05的顯著水準進行統計檢定，並推論該研究員應購買原廠或副電池較為划算？

(15%)

根據情境，選擇適當的方式：

Testing Hypotheses about Difference between Two Means

$$t=\frac{\text{sample mean}-\text{null value}}{\text{standard error}}=\frac{(\bar{x}_1-\bar{x}_2)-150}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

Step1: 設立假說

H_0: μ1 − μ2 ≤ 150

H_1: μ1 − μ2 > 150

α=0.05 –> 右尾檢定

Step2: 計算t-statistics

```
se = sqrt((sd1^2)/n1 + (sd2^2)/n2); se
t = (x1-x2-150)/se; t
```

Step3: 計算P值

```
p.value = pt(t, df = 10-1, lower.tail = F); p.value
```

Step4: 檢定結果

p.value < α ，拒絕虛無假設。

Step5: 結論

根據0.05顯著水準下的統計檢定，研究員購買原廠電池較為划算

p.value = 0.01618545

106-1期末考第二大題

2. 美國家海洋暨大氣總署 (NOAA) 想根據歷史氣候資料了解國內氣候是否有暖化現象，隨機抽取國內 50 座氣象測站 1968及 2016的年均溫資料 （ US_temperature.csv ）。
假設 50座氣象測站資料彼此獨立， 請以 0.05的顯著水準：

(a) 請以假設檢定檢驗分別在 1968及 2016年所觀測到的氣溫是否有顯著差異。 (10%)

(b)在兩個年份均沒有氣候異常的情況下若溫差大於 1.6度C，則可宣稱美國境內有暖化現象。請計算 95%的信賴區間並繪製成圖，依據結果判斷 NOAA是否能如此宣稱。 (15%)

(A) 請以假設檢定檢驗分別在 1968及 2016年所觀測到的氣溫是否有顯著差異。(10%)

根據情境，選擇適當的方式：

Testing Hypotheses about Mean of Paired Differences

$$t = \frac{\text{sample mean} - \text{null value}}{\text{standard error}} = \frac{\bar{d} - 0}{s_d / \sqrt{n}}$$

輸入資料與資類分類:

```
```{r}
setwd("路徑")
data.q2 = read.csv(file="US_temperature.csv", header=TRUE)

x1968 = data.q2$year_1968
x2016 = data.q2$year_2016

xd = x2016 - x1968
mean.xd = mean(xd)
sd.xd = sd(xd)
n = 50
```

Step1: 設立假說

H0: μd = 0

Ha: μd ≠ 0

α= 0.05 α2=0.025 –> 雙尾檢定

Step2: 計算t-statistics

```
t = (mean.xd-0)/(sd.xd/sqrt(n)); t
```

Step3: 計算P值

```
p.value = pt(t, df = n-1, lower.tail = F); p.value
p.value*2
```

Step4: 檢定結果

p.value*2 < α ，拒絕虛無假設。

Step5: 結論

根據0.05顯著水準下的統計檢定，1968及2016觀測到的年均溫有顯著差異。

p.value*2 = 0.03265
```

(B)在兩個年份均沒有氣候異常的情況下若溫差大於 1.6度C，則可宣稱美國境內有暖化現象。請計算 95%的信賴區間並繪製成圖，依據結果判斷 NOAA是否能如此宣稱。 (15%)

1. 根據情境，選擇適當的方式：

   CI for Population Mean of Paired Differences

$$\bar{d} \pm t^* \times s.e.\left(\bar{d}\right) \Rightarrow \bar{d} \pm t^* \times \frac{s_d}{\sqrt{n}}$$

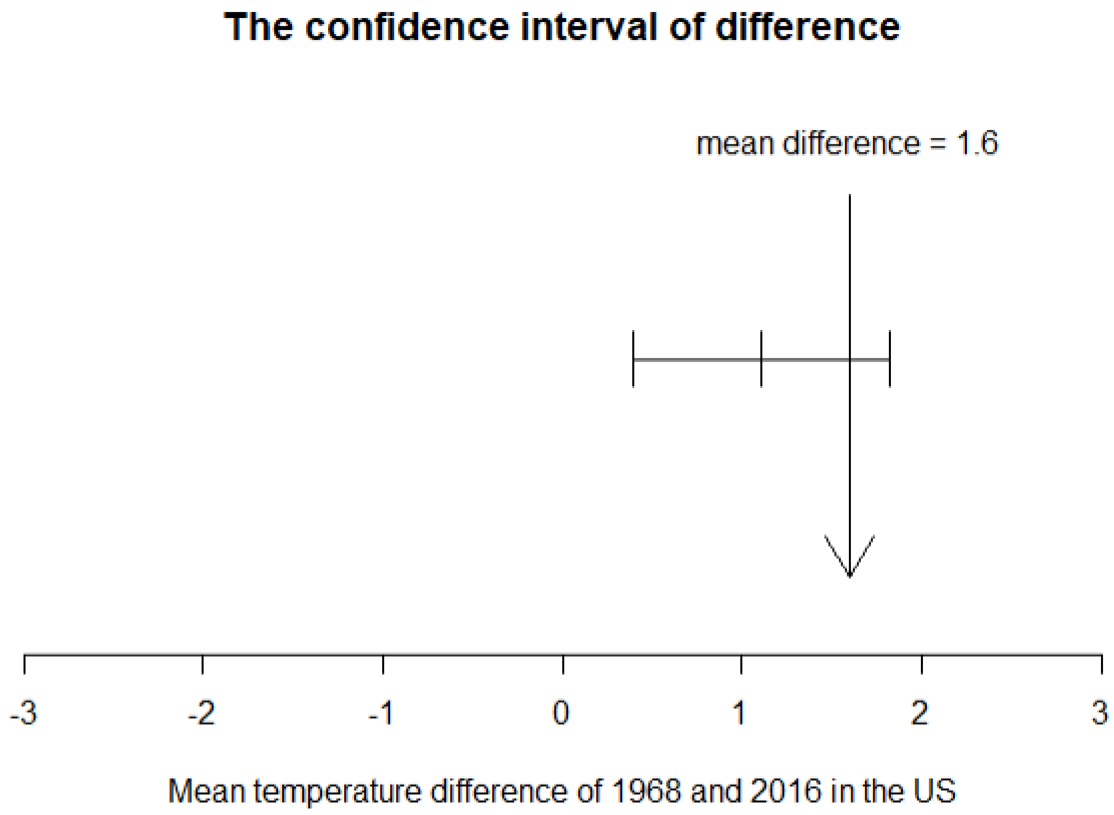
2. 計算需要的參數：

   t* 、standard error

```{r}
alpha = 0.05
half.alpha = alpha/2
t.star = qt(half.alpha, df = n-1, lower.tail = F)

CI.upper = mean.xd+t.star*(sd.xd/sqrt(n)); CI.upper
CI.lower = mean.xd-t.star*(sd.xd/sqrt(n)); CI.lower
```

```
[1] 1.823412
[1] 0.3881885
```

```{r}
#畫圖
plot(c(0,1),type='n',xlim=c(-3,3),xlab='Mean temperature difference of 1968 and 2016 in the US',ylab='',axes=FALSE,main='The confidence interval of difference')
lines(c(0.3881885,1.823412),c(0.5,0.5))
arrows(x0=1.6,y0=0.8,x1=1.6,y1=0.1)

text(1.6,0.9,label='mean difference = 1.6')
lines(c(0.3881885,0.3881885),c(0.45,0.55))
lines(c(1.1058,1.1058),c(0.45,0.55))
lines(c(1.823412,1.823412),c(0.45,0.55))

axis(1,seq(-3,3,by=1))
```



Ans: 兩個年份氣溫差的信賴區間為[0.3881885, 1.823412]。信賴區間包含1.6度C，代表1968及2016的年溫差可能大於或小於1.6oc，故沒有足夠的信心宣稱美國境內有暖化現象。