# 9. 信賴區間估計
# Confidence Interval Estimation

鄒慶士 (Ching-Shih Tsou)
台北商業技術學院資訊與決策科學所
E-mail：cstsou@mail.ntcb.edu.tw

## Agenda

- Confidence Interval for Population Proportion
- Proportion Test
- The z-test
- The t-test
- Comparing p-values from t and z
- Confidence Interval for the median

# Agenda

- Confidence Interval for Population Proportion
- Proportion Test
- The z-test
- The t-test
- Comparing p-values from t and z
- Confidence Interval for the median



# Confidence Interval for Population Proportion

- 根據問卷或投票結果估計母體比例值的區間。
- 此為吾人在報章雜誌上，最常見之信賴區間的應用。
- 假設100位受訪者中，42位偏愛X品牌。
- 要估計的母體參數為何？

$$p = \frac{\text{Number who agree}}{\text{Size of population}}$$

- 合理的點估計式為何？

$$\hat{p} = \frac{\text{Number surveyed who agree}}{\text{size of survey}}$$

# Confidence Interval for Population Proportion (cont.)

- If the sampled answers are recorded as

$$X_i = \begin{cases} 1, \text{ if answer } i \text{ was "yes"} \\ 0, \text{ if answer } i \text{ was "no"} \end{cases}, \text{ for } i = 1,2,\ldots,n$$

- The point estimator of the population proportion, i.e. sample proportion, will be

$$\hat{p} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \overline{X}$$




# Sampling Distribution

- $\hat{p}$ has a known distribution, and if n is large enough we can say the following is approximately normal with mean 0 and variance 1:

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)}/\sqrt{n}} = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

- Z is in (-1,1) with probability approximately 0.68
- Z is in (-2,2) with probability approximately 0.95
- Z is in (-3,3) with probability approximately 0.998

# Std. Error and Probability Interval

$$SE=\sqrt{p(1-p)/n}$$

$$P(-1\le \frac{p-\hat{p}}{SE}\le 1)=0.68$$

$$P(-2\le \frac{P-\hat{P}}{SE}\le 2)=0.95$$

$$P(-3\le \frac{P-\hat{P}}{SE}\le 3)=0.998$$

$$P(\hat{P}-3SE\le P\le \hat{P}+3SE)=0.998$$



# qnorm

- p, mean=0, sd=1
- lower.tail=TRUE

# pnorm

- q, mean=0, sd=1
- lower.tail=TRUE

RGui
檔案 編輯 看 其它 程式套件 視窗 輔助

R Console

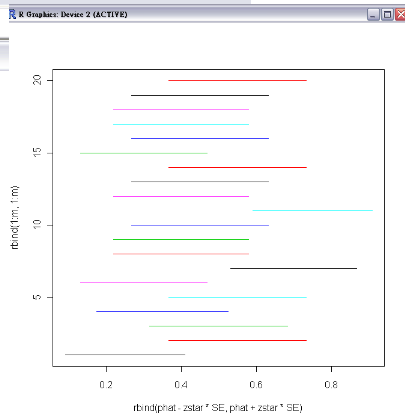
```
> 2*(1-pnorm(zstar))
[1] 0.200 0.100 0.050 0.001
>
```



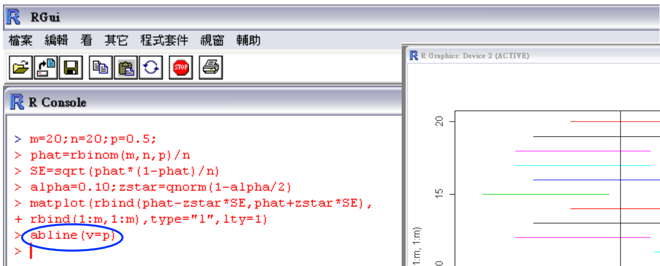
RGui
檔案 命令歷程 重設大小 視窗

R Console

```
> m=20;n=20;p=0.5;
> phat=rbinom(m,n,p)/n
> SE=sqrt(phat*(1-phat)/n)
> alpha=0.10;zstar=qnorm(1-alpha/2)
> matplot(rbind(phat-zstar*SE,phat+zstar*SE),
+ rbind(1:m,1:m),type="l",lty=1)
>
```




- 丟20枚硬幣二十次
- rbinom函數
  - no. of observations
  - no. of trials
  - prob. of success

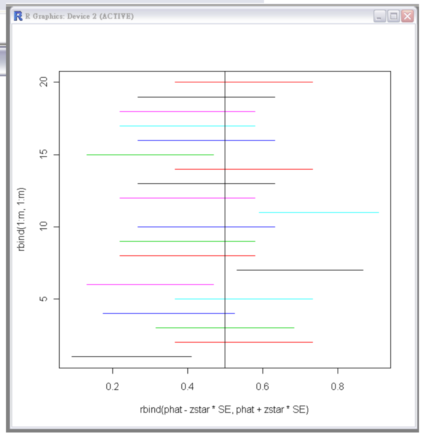



- 矩陣資料繪圖函數

  matplot()

  matpoints()

  matlines()






# Many other tests follow a similar procedure

- One finds a good statistic that involves the unknown parameter (called as a *pivotal quantity*).
- One uses the known (sampling) distribution of the statistic (i.e. pivotal quantity) to make a probabilistic statement.
- One unwraps things to form a confidence interval.
- It is often of the form the statistic (an unbiased point estimator) plus or minus a multiple of the standard error.

# Agenda

- Confidence Interval for Population Proportion
- Proportion Test
- The z-test
- The t-test
- Comparing p-values from t and z
- Confidence Interval for the median



# 母體比例值檢定

- 檢驗標的是比例值的型態，例如生育率、市場佔有率、留學生比例…等，比例值檢定是由抽樣所得到的樣本比例值去推斷『關於母體比例值的聲明』是否屬實。

# 藥廠例子

R Console

```
> prop.test(800,1000)

        1-sample proportions test with continuity correction

data:  800 out of 1000, null probability 0.5
X-squared = 358.801, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.7735599 0.8240979
sample estimates:
  p
0.8
```

 100年度教育部補助技專校院建立特色典範計畫 15

# 100人中有42人偏愛品牌X

R Console

```
> utils:::menuInstallLocal()
updating HTML package descriptions
> prop.test(42,100)

        1-sample proportions test with continuity correction

data:  42 out of 100, null probability 0.5
X-squared = 2.25, df = 1, p-value = 0.1336
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3233236 0.5228954
sample estimates:
   p
0.42
```

 100年度教育部補助技專校院建立特色典範計畫 16

# 改變信賴水準

R Console

```
> prop.test(42,100,conf.level=0.90)

        1-sample proportions test with continuity correction

data:  42 out of 100, null probability 0.5
X-squared = 2.25, df = 1, p-value = 0.1336
alternative hypothesis: true p is not equal to 0.5
90 percent confidence interval:
 0.3372368 0.5072341
sample estimates:
   p
0.42
```



# Agenda

- Confidence Interval for Population Proportion
- Proportion Test
- The z-test
- The t-test
- Comparing p-values from t and z
- Confidence Interval for the median

## z-test Statistic

- Statistic $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is normally distributed, if
  - σ is known, and the Xi's are normally distributed.
  - σ is known, and n is large enough to apply the CLT.

© Vince Tsou, IDS, NTCB 100年度教育部補助技專校院建立特色典範計畫 19

## CLT

- 若 $n \to \infty$ ，則抽樣分配會近似常態分配

(n > 30 → CLT 即成立)

1. $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ or $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

2. $\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$ or $\frac{\sum X_i - n\mu}{\sqrt{n\sigma^2}} \sim N(0,1)$

© Vince Tsou, IDS, NTCB 100年度教育部補助技專校院建立特色典範計畫 20

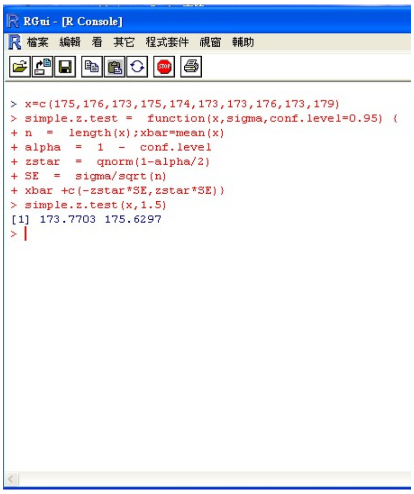
## 體重資料範例

- Suppose a person weighs himself on a regular basis and finds his weight to be
  175 176 173 175 174 173 173
  176 173 179
- Suppose that $\sigma = 1.5$ and the error in weighing is normally distributed.



## 自行撰寫 z 檢定函數

- 函數名simple.z.test
- 輸入
  – 資料向量
  – 母體標準差
  – 顯著水準
- 輸出信賴區間
- 注意!{ }



# Agenda

- Confidence Interval for Population Proportion
- Proportion Test
- The z-test
- The t-test
- Comparing p-values from t and z
- Confidence Interval for the median



# t-test Statistic

- More realistically, you may not know the standard deviation. To work around this we use the t-statistic, which is given by

$$t = \frac{\overline{X} - \mu}{s / \sqrt{n}}$$

where $s$, the sample standard deviation, replaces $\sigma$, the population standard deviation.

- The $X\_i$ are normal and $n$ is small then t has the t distribution with n-1 degrees of freedom.
- If n is large, then t is approximately normal.

# 左邊截略的常態分配

- x<-c(175,176,173,175,174,173,173,176,173,179)
- par(mfrow=c(1,3))
- hist(x)
- boxplot(x)
- qqnorm(x);qqline(x)

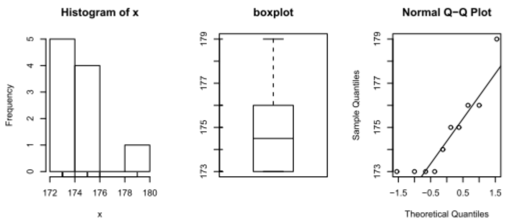



Figure 44: Plot of weights to assess normality



# 無母數檢定方法是否更適合？

**Step1**

R Console

```
> x=c(175,176,173,175,174,173,173,176,173,179)
```

**Step2**

R Console 複製和貼上

```
> x=c(175,176,173,175,174,173,173,176,173,179)
> t.test(x)

        One Sample t-test

data:  x
t = 283.8161, df = 9, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 173.3076 176.0924
sample estimates:
mean of x
    174.7
```

Notice we get a different confidence interval.

# Agenda

- Confidence Interval for Proportion
- Proportion Test
- The z-test
- The t-test
- Comparing p-values from t and z
- Confidence Interval for the median



# Some Extra Insight: Comparing p-values from t and z

- The confidence interval based on the t statistic would always be larger than that based on the z statistic as always $t^* > z^*$ . However, the standard error SE for the t also depends on s which is variable and can sometimes be small enough to offset the difference.
- Use side-by-side box plots of two random sets of data from t and z to compare t* and z*.

# t and z Comparison

```
x=rnorm(100);y=rt(100,9)
boxplot(x,y)
qqnorm(x);qqline(x)
qqnorm(y);qqline(y)
```

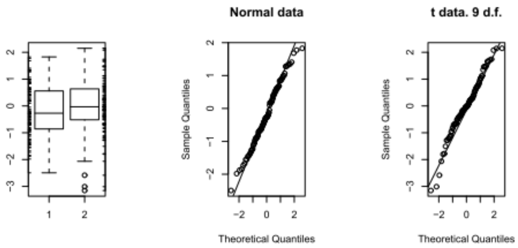



Figure 45: Plot of random normal data and random $t$-distributed data



# Normal Density and t-Density

- xvals=seq(-4,4,.01)
- plot(xvals,dnorm(xvals),type="l")
- for (i in c(2,5,10,20,50)) points(xvals, dt(xvals,df=i), type="l",lty=i)

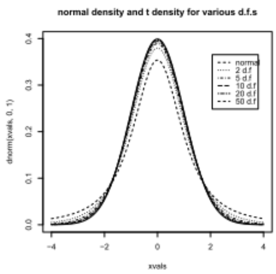



Figure 46: Normal density and the $t$-density for several degrees of freedom

# Agenda

- Confidence Interval for Population Proportion
- Proportion Test
- The z-test
- The t-test
- Comparing p-values from t and z
- Confidence Interval for the median

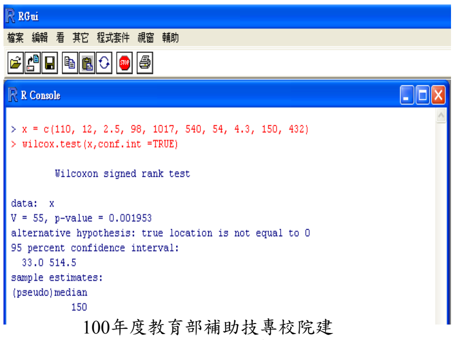


# Confidence interval for the median (1/3)

- Confidence intervals for the median are important too. They are mathematically different than the ones above, but in R these differences aren't noticed.
- The R function wilcox.test performs a non-parametric test for the median.

## Confidence interval for the median (2/3)

- Suppose the following data is the pay of CEO's in America in 2001, then the following creates a test for the median.






## Confidence interval for the median (3/3)

Notice a few things:

- Unlike prop.test and t.test, we needed to specify that we wanted a confidence interval computed. (須設定是否要C.I.)
- For this data, the confidence interval is enormous as the size of the sample is small and the range is huge. (因n小且range大，所以C.I.大)
- We couldn't have used a t-test as the data isn't even close to normal. (資料非常態，不建議做t檢定)

# The End