MAST90105 Lab and Workshop 12 Solutions

The Lab and Workshop this week covers problems arising from Module 7.5 and 8.1.

[illegible]

1 Lab

m should

1. Let $X \sim U(0,1)$ and consider a random sample of size 11 from X. Recall that if m is the median and $Y_1, \ldots, Y_n$ are the order statistics then

$$P(Y_i < m < Y_j) = \sum_{k=i}^{j-1} \binom{n}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k}.$$

We will check this formula using R by computing some confidence intervals for the median of X.

a. Use the R command:

```
qbinom(c(0.025, 0.975), size = 11, prob = 0.5)

## [1] 2 9
```

to compute quantiles of the binomial(11,0.5) distribution. (Ie. in the first case we find $\pi_{0.975}$ so that $P(X \leq \pi_{0.975}) \approx 0.975$. It is approximate as the distribution is discrete. However, it gives a guide to the endpoints of the confidence interval.)

b. Being careful about the correct evaluation points, use the *pbinom* command in R determine $P(Y_2 < m < Y_9)$?

```
pbinom(8, 11, 0.5) - pbinom(1, 11, 0.5)

## [1] 0.9614258
```

c. Use the R command:

```
X <- runif(11)
```

to simulate 11 observations from X.

d. Use the *sort* command to compute the order statistics and store them in a new variable Y and hence compute Y_2 and Y_9.

e. Automate this in a function and check *f(11)* to see that it works:

```
f = function(n) {
    X = runif(n)
    Y = sort(X)
    c(Y[2], Y[9])
}
f(11)

## [1] 0.09459459 0.80220718
```

$$E(e^{tx}) = \int^{+\infty} e^{tx}\left(e^{-2x} + e^{-3x} + e^{-6x}\right) dx$$

$$= \int^{+\infty} e^{x(t-2)} + e^{x(t-3)} + e^{x(t-6)}\, dx$$

$= \int_0 \left(e^{x(t-2)} + e + e \right)$

$= \frac{1}{t-2} e^{x(t-2)} \Big|_0^{+\infty} + \frac{e^{x(t-3)}}{t-3} \Big|_0^{+\infty} + \frac{e^{x(t-6)}}{t-6} \Big|_0^{+\infty}$

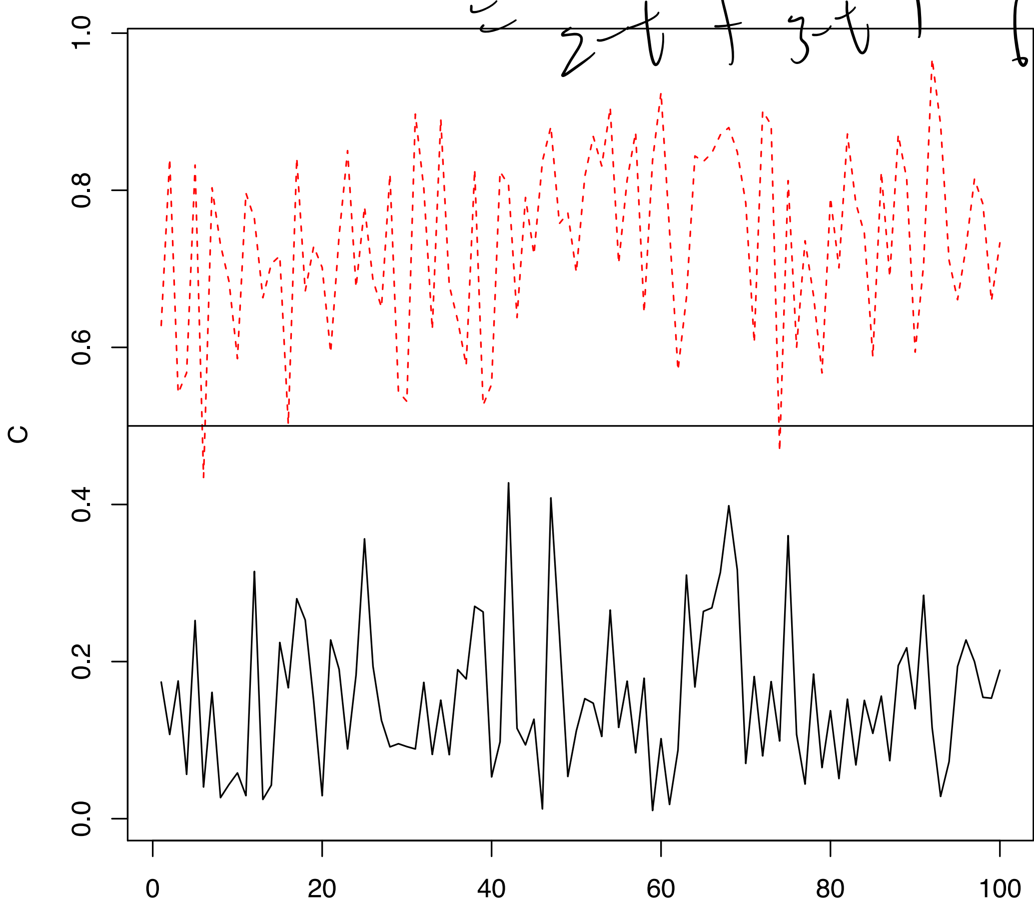
Enter *f(11)* to check it works.

f. Enter the following R commands:

```
t = as.matrix(rep(11, 100))  #t needs to be a matrix for apply to work.
C = t(apply(t, 1, f))  #this is a trick to avoid programming
matplot(C, type = "l")
abline(c(0.5, 0))
```

$= -\frac{1}{t-2} - \frac{1}{t-3} - \frac{1}{t-6}$

$= \frac{1}{2-t} + \frac{1}{3-t} + \frac{1}{6-t}$



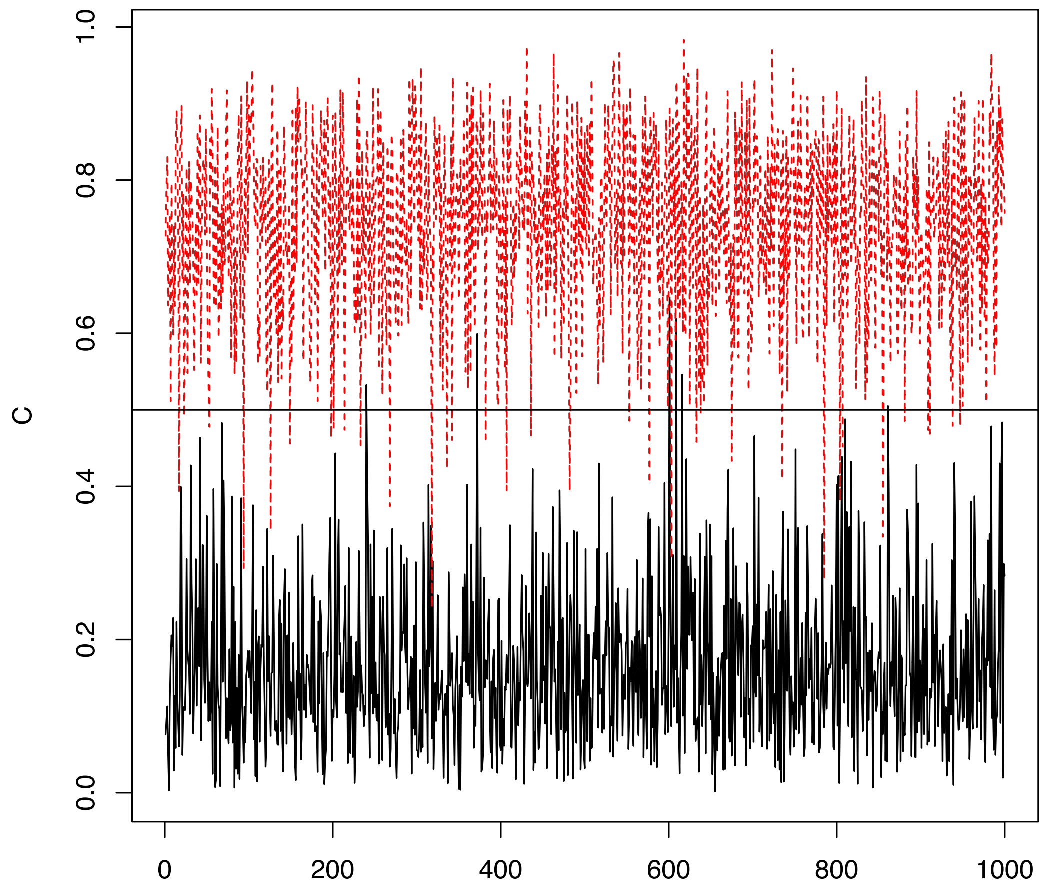

```
sum((C[, 1] < 0.5) & (C[, 2] > 0.5))/nrow(C)

## [1] 0.98
```

and hence compute the proportion of your simulated samples that contain the true mean value 1/2. Is this close to your answer in (b)? (The apply command applies the function f to each row in t and t(A) computes the transpose of the matrix A).

- *This is reasonably close to the true value.*

g. To get more precision, repeat with

```
t = as.matrix(rep(11, 1000))
C = t(apply(t, 1, f))  #this is a trick to avoid programming
matplot(C, type = "l")
abline(c(0.5, 0))
```

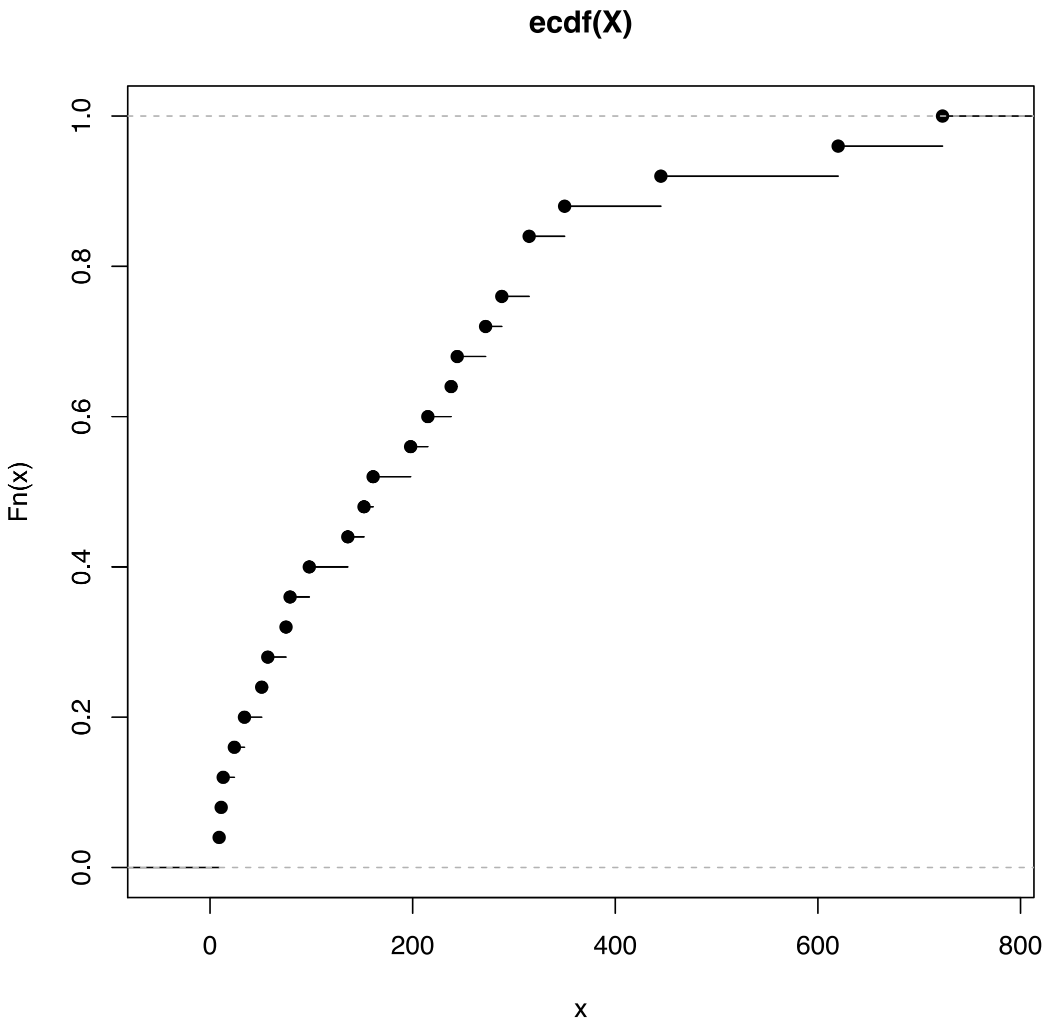



```
sum((C[, 1] < 0.5) & (C[, 2] > 0.5))/nrow(C)

## [1] 0.96
```

2. The following 25 observations give the time in seconds between submissions of computer programs to a printer queue.
79 315 445 350 136 723 198 75 161 13 215 24 57 152 238 288 272 9 315 11 51 98 620 244 34

a. The cumulative distribution function allows us to use graphical methods to approximate the percentiles. Store the above data a vector X in R, and use the command

```
X <- c(79, 315, 445, 350, 136, 723, 198, 75, 161, 13,
    215, 24, 57, 152, 238, 288, 272, 9, 315, 11, 51,
    98, 620, 244, 34)
plot(ecdf(X))
```

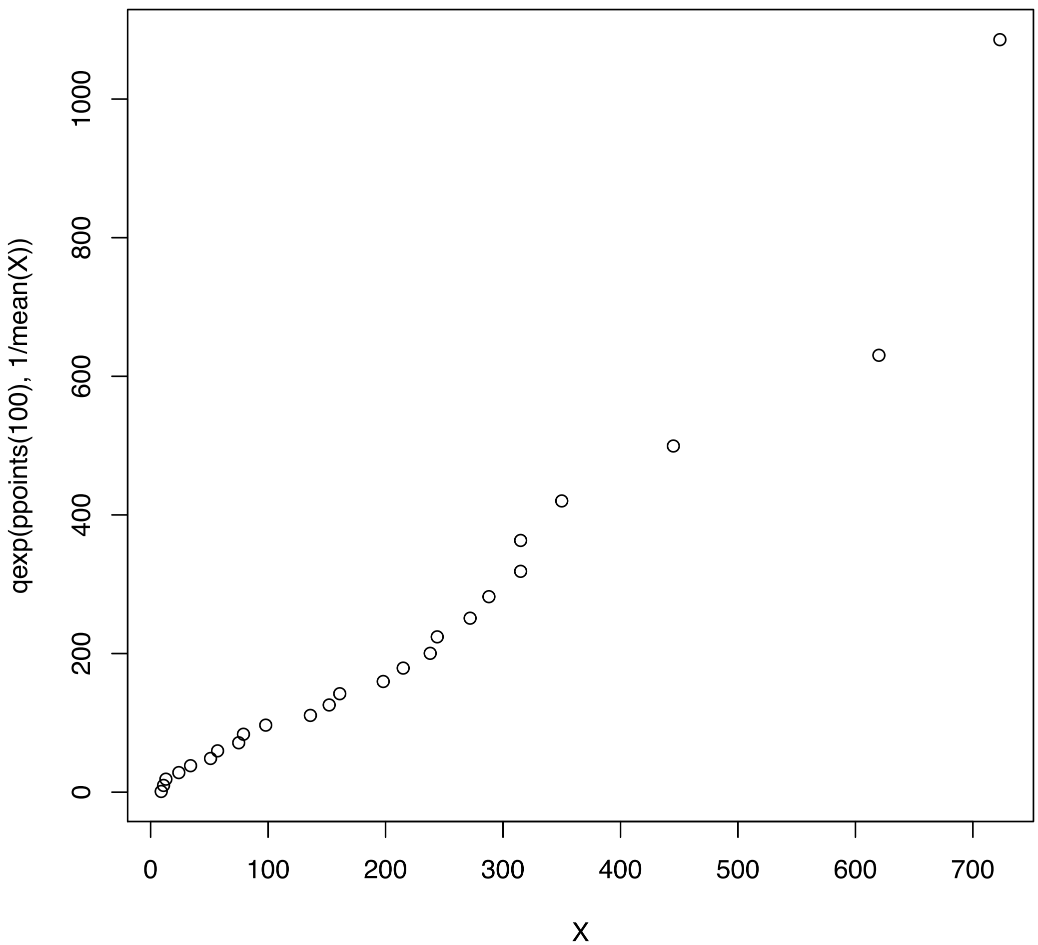


to plot the cumulative distribution function. Use the plot to give approximate point estimates of $\pi_{0.25}$, m and $\pi_{0.75}$.

- *The first quartile seems to be around 60, the median is around 160 and the third quartile around 300.*

b. Use the command

```
qqplot(X, qexp(ppoints(100), 1/mean(X)))
```

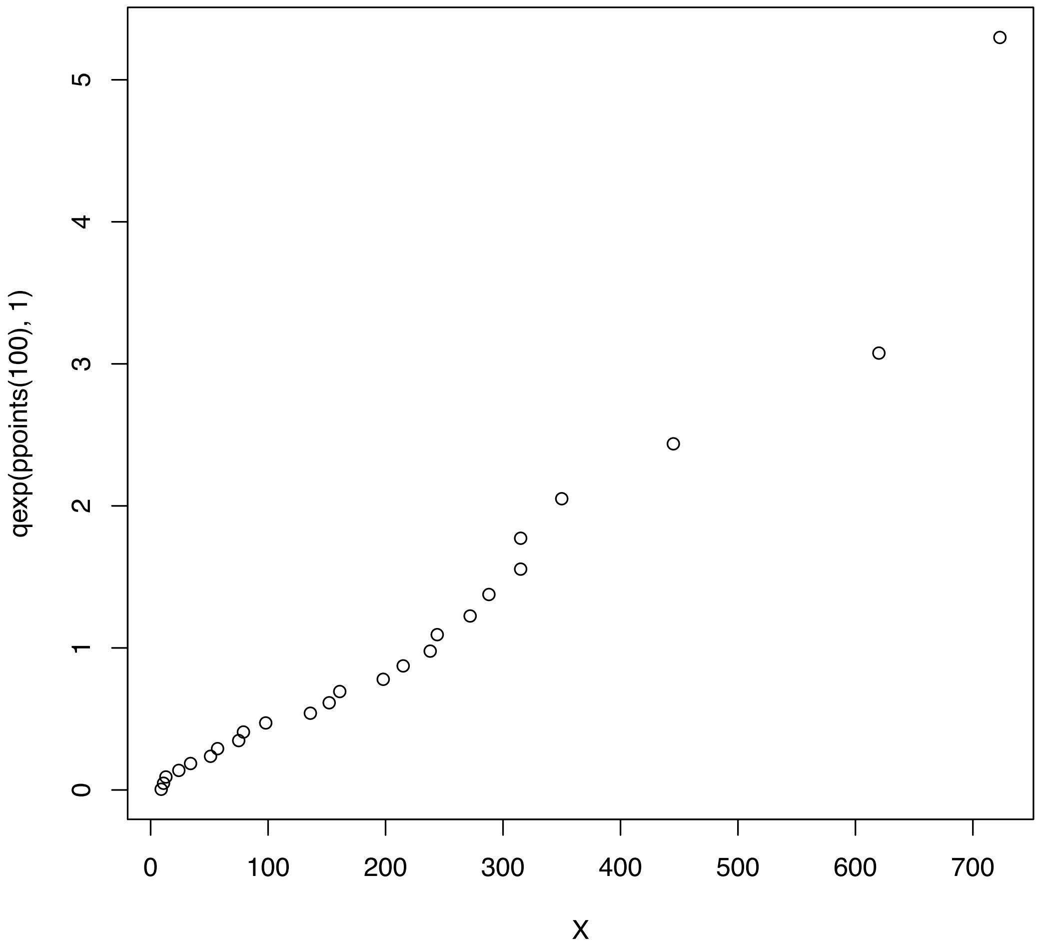


to obtain a quantile-quantile plot of X for the exponential distribution. What do you think?

- *There is a reasonable approximation to the line $y = x$ consistent with the exponential distribution with mean matching the sample, but the last points is an outlier.*

c. Use the command

```
qqplot(X, qexp(ppoints(100), 1))
```

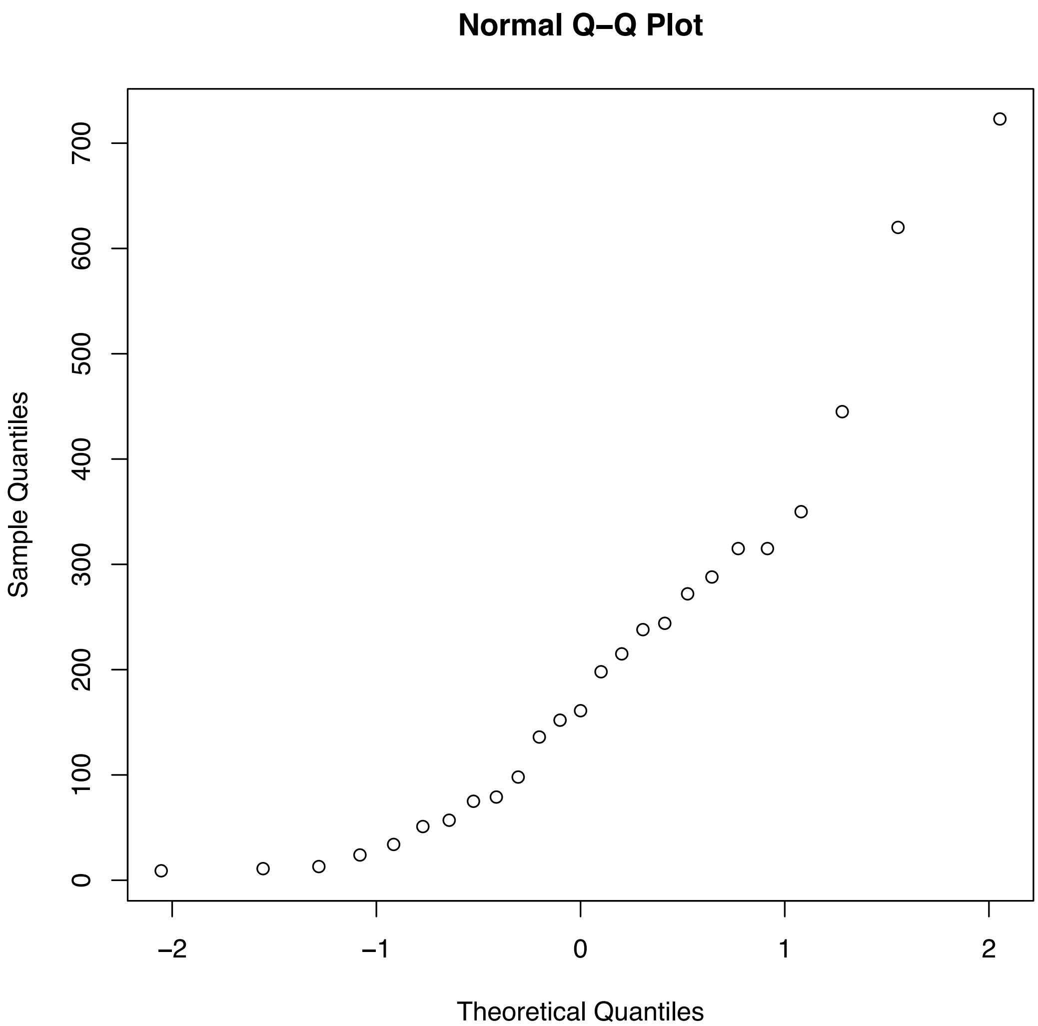


to obtain a quantile-quantile plot of X for the exponential distribution. How does this differ from your previous plot?

- *The plot looks similar - the only difference is the scale on the y-axis is changed by having mean 1.*

d. Use the command

```
qqnorm(X)
```



to obtain a normal quantile-quantile plot of X. What do you think?

- *The normal plot does not look linear at both ends indicating a poor fit to the normal distribution.*

e. Give point estimates of $\pi_{0.25}$, m and $\pi_{0.75}$. (Use the command:

```
quantile(X, c(0.25, 0.5, 0.75), type = 6)

##   25%   50%   75%
##  54.0 161.0 301.5
```

95%的CI区间　　求median

for the 25th percentile) `qbinom(c(0.025, 0.975), size = 11, prob = 0.5)` 25

f. Find the following confidence intervals and give the confidence level.

i. (y_3, y_{10}), a confidence interval for $\pi_{0.25}$.

可以用于估计

q-th quantile

$$P(Y_i < \pi_q < Y_j) = \sum_{k=i}^{j-1} \binom{n}{k} (q)^k (1-q)^{n-k}$$

```
Y <- sort(X)
c(Y[3], Y[10])
## [1] 13 98
pbinom(9, 25, 0.25) - pbinom(2, 25, 0.25)
## [1] 0.8965632
```

ii. (y_9, y_{17}), a confidence interval for the median m.

```
c(Y[9], Y[17])
## [1]  79 244
pbinom(16, 25, 0.5) - pbinom(8, 25, 0.5)
## [1] 0.8922479
```

iii. (y_{16}, y_{23}), a confidence interval for $\pi_{0.75}$.

```
c(Y[16], Y[23])
## [1] 238 445
pbinom(22, 25, 0.75) - pbinom(15, 25, 0.75)
## [1] 0.8965632
```

g. Find a t interval for the mean μ of the same confidence as that constructed for the median. Compare these two confidence intervals. Are the results surprising? (Your quantile plots and a histogram or stem and leaf plot may help).

```
t.test(X, conf.level = 0.892)$conf.int

## [1] 142.7571 267.0829
## attr(,"conf.level")
## [1] 0.892
```

- *The confidence interval is wider for the median and in a different position - the data does not appear to come from a normal distribution, so in part this may reflect the fact that the population median is not the same as the population but the t-confidence interval is probably not appropriate.*

3. *The data is in the file Lab11.RData in the LMS and Lab Folder.* Let p be the proportion of yellow lollies in a packet of mixed colours. It is claimed that $p = 0.2$.

a. Define a test statistic and critical region with a significance level of $\alpha = 0.05$ to test $H_0 : p = 0.2$ against $H_1 : p \neq 0.2$.

- *Let $\hat{p}$ be the proportion yellow lollies counted. Assuming that the number of yellow lollies counted is Binomial(n, p) where n is the total number of lollies counted, we would reject H_0 if*

assumption H_0 is true then try to reject

$$|z| = \frac{|\hat{p} - 0.2|}{\sqrt{0.2 \times 0.8/n}} > 1.96.$$

sample proportion

(不发生在小概率事件区)

拒绝原假设

$$\frac{\hat{p} - 0.2}{\sqrt{0.2 \times 0.8}} \approx N(0,1)$$

在一次实验中小概率事件发生了，就拒绝原假设　reject H_0 if Z is too small or too big.

b. To perform the test, each of 20 students counted the number of yellow lollies and the total number of lollies in a 48.1 gram packet. The results were:

95% rejection region $|Z| > 1.96$

两侧 2.5% $Z < Z_{0.025}$ 或 $Z > Z_{0.975}$

y	n	y	n
8.00	56.00	10.00	57.00
13.00	55.00	8.00	59.00
12.00	58.00	10.00	54.00
13.00	56.00	11.00	55.00
14.00	57.00	12.00	56.00
5.00	54.00	11.00	57.00
14.00	56.00	6.00	54.00
15.00	57.00	7.00	58.00
11.00	54.00	12.00	58.00
13.00	55.00	14.00	58.00

If each student made a test of $H_0 : p = 0.2$ at the 5% level of significance, what proportion of students rejected the null hypothesis?

- *0.05 - See R output.*

```
load("Lab11.RData")
phat = Data$y/Data$n
zabs = abs((phat - 0.2)/sqrt(0.8 * 0.2/Data$n))
sum(zabs > 1.96)/20

## [1] 0.05
```

c. If the null hypothesis were true, what proportion of students do you expect to reject the null hypothesis at the 5% level of significance?

- *0.05*

d. For each of the 20 ratios in part (b) an approximate 95% confidence interval can be constructed. What proportion of these intervals contains $p = 0.2$?

- *0.9 - see R output. There is no contradiction with part b, because the width of the confidence interval is determined by the observed proportion, $\hat{p}$, not the null hypothesis value 0.2.*

```
ci <- 1.96 * sqrt(phat * (1 - phat)/Data$n)
sum((phat - ci < 0.2) * (phat + ci > 0.2))/20

## [1] 0.9
```

e. If the 20 results are pooled do we reject $H_0 : p = 0.2$?

- *No - see R output. Note that this does not contradict the results in (b) because there were different total numbers counted each time.*

```
x = sum(Data$y)
N = sum(Data$n)
ptothat = x/N
(ptothat - 0.2)/sqrt(0.2 * 0.8/N)

## [1] -0.4324987
```

2 Workshop

4. Develop a function to simulate the distribution of the Wilcoxon two sample statistic based on sample sizes n, m by drawing random samples, using the R command `sample`.

```
f <- function(x) {
    sum(sample(x[1] + x[2], size = x[2]))
}
W <- function(x, r) {
    t <- matrix(rep(x, r), byrow = TRUE, nrow = r,
        ncol = 2)
    apply(t, 1, f)
}
# This can be tested out on the sample sizes of 8
# and 8, used in the cinammon packet filling
# example as follows.  The code produces 10,000
# values of the W statistic It checks the mean and
# variance of the simulated samples It then finds
# the emprical probability that W is at least 87
w <- W(c(8, 8), 10000)
# compare to theoretical values
c(mean(w), 8 * 17/2)

## [1] 68.0013 68.0000

c(sd(w), sqrt(8 * 8 * 17/12))
```

```
## [1] 9.490274 9.521905

# empirical p-value compared to normal
# approximation
c(sum(w >= 87)/10000, 1 - pnorm((87 - 4 * 17)/sqrt(64 *
    17/12)))

## [1] 0.02510000 0.02299968

# try again with a larger number of repititions
w <- W(c(8, 8), 1e+06)
# compare to theoretical values
c(mean(w), 8 * 17/2)

## [1] 68.00016 68.00000

c(sd(w), sqrt(8 * 8 * 17/12))

## [1] 9.523346 9.521905

# empirical p-value
c(sum(w >= 87)/1e+06, 1 - pnorm((87 - 4 * 17)/sqrt(64 *
    17/12)))

## [1] 0.02505300 0.02299968
```

5. Vitamin B_6 is one of the vitamins in a multiple vitamin pill manufactured by a pharmaceutical company. The pills are produced with a mean of 50 milligrams of vitamin B_6 per pill. The company believes there is a deterioration of 1 milligram per month, so that after 3 months they expect that $\mu = 47$. A consumer group suspects that $\mu < 47$ after 3 months.

   a. Define a critical region to test $H_0 : \mu = 47$ against $H_1 : \mu < 47$ at the $\alpha = 0.05$ significance level based on a random sample of of size $n = 20$. ($t_{0.05}(19) = 1.729$, $t_{0.05}(20) = 1.724$, $t_{0.025}(19) = 2.093$, $t_{0.025}(20) = 2.086$).
   b. If the 20 pills yielded a mean of $\bar{x} = 46.94$ with standard deviation of $s = 0.15$, what is your conclusion?
   c. What is the approximate p-value of this test?

- a. *Critical region is:*

$$t = \frac{\bar{x} - 47}{s/\sqrt{20}} < -t_{0.05}(19) = -1.729$$

$P(t < t_{0.05}(19) \mid H_0 \text{ is true}) = 0.05$

$t \overset{d}{=} t(19)$

or

The critical region is: $C = \{t : t < t_{0.05}(19)\}$

$= \{t : t < -1.729\}$

$p(t > t_{0.95}(19) \mid H_0 \text{ is true}) = 0.05$

$D = \{t : t > t_{0.95}(19)\} = \{t : t > 1.729\}$

D would be critical region if we were testing $\mu = 47$ vs $\mu > 47$

Q5b $\bar{x} = 46.94$ $s = 0.15$ $n = 20$

$\therefore t = \dfrac{46.94 - 47}{0.15/\sqrt{20}} = -1.789 \in \{t : t < -1.729\}$

Therefore we reject H_0

c) p value is $\underbrace{P(t < -1.789 \mid H_0 \text{ is true})}_{\overset{d}{=} t(19)}$ ☆

$\underbrace{P(t < -2.093 \mid H_0 \text{ is true})}_{0.025} < p < \underbrace{P(t < -1.729 \mid H_0 \text{ is true})}_{0.05}$

b. $t = (46.94 - 47)/0.15/\sqrt{20} = -1.789$. *This is less than -1.729 so we reject* H_0.
c. The p-value is between 0.025 and 0.05.

6. Let X be the forced vital capacity (FVC) in liters for a female college student. Assume that $X \sim N(\mu, \sigma^2)$ approximately. Suppose it is known that $\mu = 3.4$ litres. A volleyball coach claims the FVC of volleyball players is greater than 3.4. She plans to test this using a random sample of size $n = 9$.

   a. Define the null hypothesis.

   b. Define the alternative hypothesis.

   c. Define a critical region for which $\alpha = 0.05$. lllustrate this on a figure. ($t_{0.025}(8) = 2.306$, $t_{0.05}(8) = 1.859$, $t_{0.01}(8) = 2.896$)

   d. Calculate the value of the test statistic if $\bar{x} = 3.556$ and $s = 0.167$.

   e. What is your conclusion?

   f. What is the approximate p-value of this test?

   - *a.* $H_0 : \mu = 3.4$
     b. $H_1 : \mu > 3.4$
     c. $t = (\bar{x} - 3.4)/(s/3) > 1.859$
     d. $t = (3.556 - 3.4)/(0.167/3) = 2.802$
     e. Reject H_0.
     f. $2.306 < 2.802 < 2.896$ *so the p-value is between 0.01 and 0.025.*

7. Among the data collected for the World Health Organisation air quality monitoring project is a measure of suspended particles in $\mu g/m^3$. Let X and Y equal the concentration of suspended particles in the city centres of Melbourne and Houston. Using $n = 13$ observations of X and $m = 16$ observations of Y, we shall test $H_0 : \mu_X = \mu_Y$ against $H_1 : \mu_X < \mu_Y$.

   a. Define the test statistic and the critical region assuming the variances are equal. Let $\alpha = 0.05$

   b. If $\bar{x} = 72.9$, $s_x = 25.6$, $\bar{y} = 81.7$ and $s_y = 28.3$, calculate the value of the test statistic and state your conclusion.
   ($t_{0.025}(27) = 2.052$, $t_{0.05}(27) = 1.703$, ($t_{0.1}(27) = 1.314$, ($t_{0.25}(27) = 0.684$)

   c. Give limits for the p-value of this test.

   - *a. Critical region is given by*

$$T = \frac{\bar{x}-\bar{y}}{S_P\sqrt{\frac{1}{m}+\frac{1}{n}}} \overset{H_0}{\sim} t(m+n-2)$$

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{12s_x^2+15s_y^2}{27}\left(\frac{1}{13} + \frac{1}{16}\right)}} \leq t_{0.05}(27) = -1.703$$

-1.703

$y_1 \cdots y_n \sim N(\mu_y, \sigma^2)$

$X_1 \cdots X_{13} \sim N(\mu_x, \sigma^2)$

too small

a. recap 1) $H_0: \mu_x = \mu_y$ $H_1: \mu_x < \mu_y$

$\bar{x}-\bar{y}$ (13+16-2)

test statistic is $t = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{(13-1)S_x^2+(16-1)S_y^2}{27}}} \sim t(27)$

$\alpha = 0.05$

since $P(t < t_{0.05}(27) \mid H_0 \text{ is true}) = 0.05$

critical region is $C = \{t : t < t_{0.05}(27)\} = \{t : t < -1.703\}$

① if instead we were testing $\mu_X > \mu_Y$ $C^* = \{t : t > t_{0.95}(27)\}$

② $H_1 : \mu_X \neq \mu_Y$ $C^* = \{t : t < t_{0.025}(27) \text{ or } t > t_{0.975}(27)\}$

7.6 plugging given data do have have enough evidence

$t = -0.869 \notin C$ ~~to~~ reject H_0

---

8. wilcox-test (考虑大小)

a. $W = (\text{Rand} \times \text{sign})$ sum all $= -55$

if H_0 is true ($H_0 : M = 400 \cdot 0$)

$W \overset{d}{\approx} N(0, \underbrace{\frac{n(n+1)(2n+1)}{6}}_{819})$ (asymptotic)

test statistic is z-score

$Z = \frac{W-0}{\sqrt{819}} = -1.9218$ since $Z_{0.05} = -1.649$ and $Z = -1.9218 < 1.649$

$\Rightarrow$ reject H_0 at $\alpha = 0.05$ using wilcoxon rank sum test.

c. Sign test : $Y = \sum_{i=1}^{13} \mathbb{1}(X_i - 40000 < 0) = 9$ 小于的的个数

忽略大小 只考虑正负

under $H_0 : m = 400\ 00$ $\overset{d}{=} \text{Binomial}(13, 0.5)$ (一半一半)

$\therefore P(Y \geq 9 \mid H_0 \text{ is true}) = 0.1334$ is the p-value

p is more extreme

$1 - \text{pbinom}(8, 13, 0.5)$

$\parallel P(Y \geq 9)$

☆ 0.1334 > 0.05 : do not reject H_0 ☆

$W = 1\ 2\ 3\ 4 \ \ldots\ n$

$n = 2$

$1 \quad 2 = 3 \quad \frac{1}{4}$

$1 \quad -2 = -1 \quad \frac{1}{4}$

$-1 \quad 2 = 1 \quad \frac{1}{4}$

$-1 \quad -2 = -3 \quad \frac{1}{4}$

$1\ 2\ 3 = 6$

$1\ 2\ -3 = 0$

$1\ -2\ 3 = 2$

$-1\ 2\ 3 = 4$

$1\ -2\ -3 = -4$

$-1\ 2\ -3 = -2$

$-1\ -2\ 3 = 0$

$-1\ -2\ -3 = -6$

$-6, -4, -2, 0, 2, 4, 6$

$\frac{1}{8}\ \frac{1}{8}\ \frac{1}{8}\ \frac{2}{8}\ \frac{1}{8}\ \frac{1}{8}\ \frac{1}{8}$

10%都无法拒绝

~~[进~~ $0.1 < P < 0.25$

只知道区间 $Pr\{t<-1.314\} = 0.25$

和 稀有值 $P\{t<-0.684\} = 0.1$

b. *Observed value is*

$$t = \frac{72.9 - 81.7}{\sqrt{\frac{12\times 25.6^2 + 15\times 28.3^2}{27}\left(\frac{1}{13}+\frac{1}{16}\right)}} = -0.869 > -1.703$$

so we cannot reject H_0.

pvalue $= P(T < t = -0.869 \mid H_0$下成立$)$

$pt(-0.869, df=27)$

c. $0.10 <$ *p-value* < 0.25.

8. It is claimed that the median weight m of certain loads of candy is 40,000 pounds.

   a. Use the following data and the Wilcoxon test statistic at an approximate significance level of $\alpha = 0.05$ to test the null hypothesis $H_0 : m = 40,000$ against $H_1 : m < 40,000$.

   41195, 39485, 41229, 36840, 38050, 40890, 38345, 34930, 39245, 31031, 40780, 38050, 30906

   It may help to complete the following table. Ties are assigned the average rank.

按绝对值排名

signed Rank

3个相等值 排 8.3

sign x Rank 8.3 8.3

$-55 =$ Sum of (Signed Rank)

p-value

$Pr(Z < -1.9218)$

$= pnorm(-1.9218)$

$= 0.0273$

	X	$X-m$	Rank	Sign
1	41195	1	1	5.01
2	39485			
3	41229			
4	36840			
5	38050		8.5	
6	40890			
7	38345			
8	34930			
9	39245			
10	31031			
11	40780			
12	38050		8.5	
13	30906			

b. What is the approximate p-value?

```r
qnorm(seq(0.9, 0.975, 0.025))

## [1] 1.281552 1.439531 1.644854 1.959964
```

- *The table is*

	X	*X-m*	*Rank*	*Sign*
1	*41195.00*	*1195.00*	*5.00*	*1.00*
2	*39485.00*	*−515.00*	*1.00*	*−1.00*
3	*41229.00*	*1229.00*	*6.00*	*1.00*
4	*36840.00*	*−3160.00*	*10.00*	*−1.00*
5	*38050.00*	*−1950.00*	*8.50*	*−1.00*
6	*40890.00*	*890.00*	*4.00*	*1.00*
7	*38345.00*	*−1655.00*	*7.00*	*−1.00*
8	*34930.00*	*−5070.00*	*11.00*	*−1.00*
9	*39245.00*	*−755.00*	*2.00*	*−1.00*
10	*31031.00*	*−8969.00*	*12.00*	*−1.00*
11	*40780.00*	*780.00*	*3.00*	*1.00*
12	*38050.00*	*−1950.00*	*8.50*	*−1.00*
13	*30906.00*	*−9094.00*	*13.00*	*−1.00*

- *So*

$$W = 5.0-1.0+6.0-10.0-8.5+4.0-7.0-11.0-2.0-12.0+3.0-8.5-13.0 = -55.$$

- *Recall*

$$\mu_W = 0, \quad Var(W) = \frac{n(n+1)(2n+1)}{6} = 819.$$

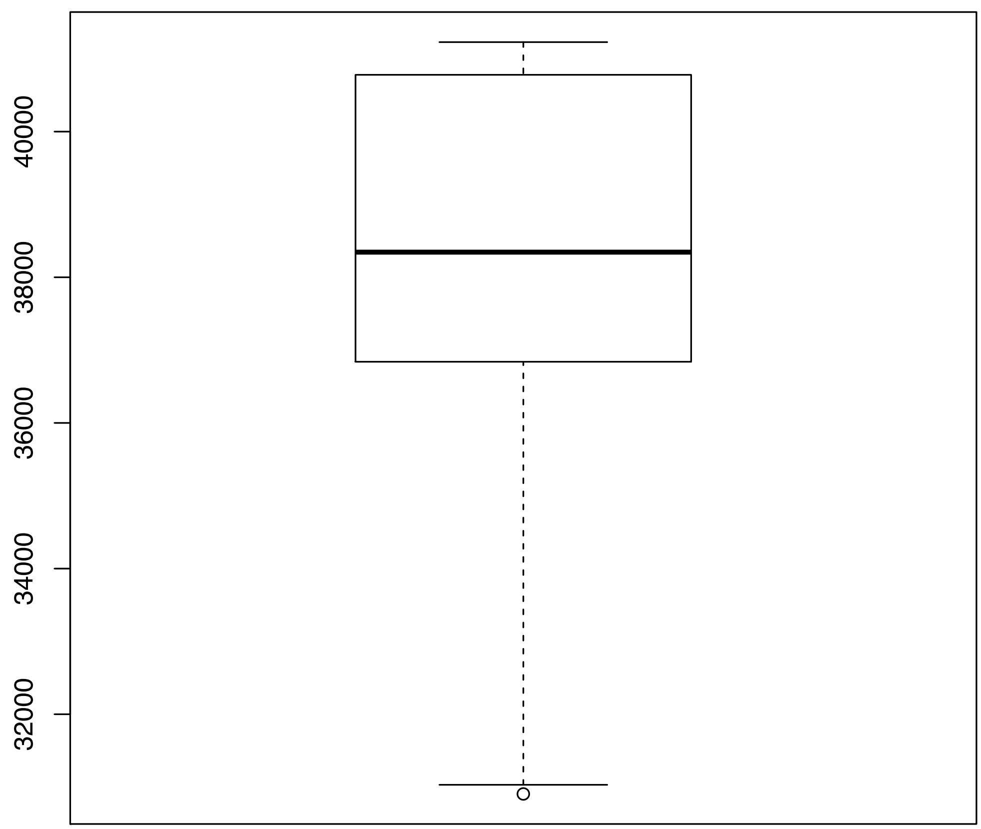
- *So*

$$Z = \frac{-55}{\sqrt{819}} = -1.9218.$$

- *This is less than -1.645 so at the 5% level of significance we reject H_0.*

- *The approximate p-value is between 0.025 and 0.025. The R output gives the approximate p-value as 0.0273.*
- *An exact evaluation using the Mathematica Lab12.nb shows the p-value is 0.0286865.*
- *There is some evidence of asymmetry of the distribution which is skewed to the left. This is seen from the boxplot and also from the different mean and median.*

```
x <- c(41195, 39485, 41229, 36840, 38050, 40890, 38345,
    34930, 39245, 31031, 40780, 38050, 30906)
pnorm(-55/sqrt(819))

## [1] 0.02731197

boxplot(x)
```



```
c(mean(x), median(x))

## [1] 37767.38 38345.00
```

c. Use the sign test to test the same hypothesis.

```
pbinom(6:12, 13, 0.5)

## [1] 0.5000000 0.7094727 0.8665771 0.9538574
## [5] 0.9887695 0.9982910 0.9998779
```

- *There are 9 negative signs.*

$$P(Y \geq 9 \mid p = 0.5) = 1 - 0.8666 = 0.1334$$

- *So we cannot reject the null hypothesis even at the 10% level.*

d. Compare the results of the two tests.

- *The null hypothesis is rejected using the rank sum test but cannot be rejected using the sign test.*
- *The difference could in part be rejection of the hypothesis of symmetry but the Wilcoxon signed rank test is more powerful than the sign test if the symmetry hypothesis holds, so it could also be due to the difference in power.*

9. A 1-pound bag of candy-coated chocolate covered peanuts contained 224 pieces of candy coloured brown, orange, green and yellow. Test the null hypothesis that the machine filling these bags treats the four colours of candy equally likely. That is test

$$H_0 : p_B = p_O = p_G = p_Y = \frac{1}{4}.$$

The observed values were 42 brown, 64 orange, 53 green, and 65 yellow. You may select the significance level or give an appropriate p-value.
($\chi^2_{0.025}(3) = 9.348$, $\chi^2_{0.05}(3) = 7.815$, $\chi^2_{0.10}(3) = 6.251$.

- *Observed and expected frequencies are:*

	B	*O*	*G*	*Y*	*Total*
O	*42*	*64*	*53*	*65*	*224*
E	*56*	*56*	*56*	*56*	*224*

- *So*

$$\chi^2 = \frac{(42-56)^2}{56} + \cdots + \frac{(65-56)^2}{56} = 6.25 < 7.815$$

- *Hence we cannot reject H_0 at the 5% level of significance.*

pearson statistic

10. In a biology laboratory the mating of two red eye fruit flies yielded $n = 432$ offspring among which 254 were red-eyed, 69 were brown-eyed, 87 were scarlet-eyed, and 22 were white-eyed. Use these data to test, with $\alpha = 0.05$, the hypothesis that the ratio among the offspring would be 9:3:3:1 respectively.
$(\chi^2_{0.025}(3) = 9.348,\ \chi^2_{0.05}(3) = 7.815,\ \chi^2_{0.10}(3) = 6.251)$.

H_0: 9 : 3 : 3 : 1 — red, brown, scarlet, white.

chi-square test

(若很小不拒绝)
Test statistic (太大拒绝)

- *Observed end expected frequencies are:*

(真实状态)

	red	brown	scarlet	white	total
O	254.00	69.00	87.00	22.00	432
Prob.	0.56	0.19	0.19	0.06	
E	243.00	81.00	81.00	27.00	

(under H_0 is true)

$432 \times \frac{9}{16}$, $\frac{3}{16} \times 432$, $\frac{3}{16} \times 432$, $= \frac{2}{16} \times 432$

$$\chi^2 = \frac{(254-243)^2}{243} + \frac{(69-81)^2}{81} + \frac{(87-81)^2}{81} + \frac{(22-27)^2}{27} = 3.646$$

$\chi^2 = \chi^2(4-1)$

$\chi^2 = 3.646$ *with 3 d.f. so we do not reject H_0 at the 5% level of significance.*

Since $\chi^2_{0.05}(3) = 7.815$ and $3.646 < 7.815$ (不够大) 拒绝 do not reject H_0 at 5%

11. We wish to determine if two groups of nurses distribute their time in six different categories about the same way. That is, the hypothesis under consideration is $H_0: p_{i1} = p_{i2}, i = 1, \ldots, 6$. To test this, nurses are observed at random throughout several days, each observation resulting in a mark in one of the six categories. The summary data is given in the following frequency table

	Category						
	1	2	3	4	5	6	Total
Group I	95	36	71	21	45	32	300
Group II	53	26	43	18	32	28	200

Use a chi-square test with $\alpha = 0.05$.

```
qchisq(seq(0.9, 0.975, 0.025), 5)

## [1]  9.236357 10.008315 11.070498 12.832502
```

- *R output below.*
- *Note that the appropriate hypotheses are that the two distributions for Groups 1 and 2 are the same.*
- *The chi-square test gives the right p-value since no extra parameters have been estimated.*

11, 12 not covered in lecture.

- *We cannot reject the null hypothesis at the 5% level and conclude there is not sufficient evidence that the two groups distribute their time differently.*

```
observed <- matrix(c(95, 36, 71, 21, 45, 32, 53, 26,
    43, 18, 32, 28), byrow = T, nrow = 2)
c1 <- chisq.test(observed, correct = F)
c1$observed

##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]   95   36   71   21   45   32
## [2,]   53   26   43   18   32   28

c1$expected

##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,] 88.8 37.2 68.4 23.4 46.2   36
## [2,] 59.2 24.8 45.6 15.6 30.8   24

c1$p.value

## [1] 0.6645015
```

12. A random sample of 1000 individuals from a rural area had 620 in favour of the election of a certain candidate, whilst a random sample of 1000 individuals from an urban area had 550 in favour of the same candidate. At the 5% level, test the hypothesis that area and opinion about the candidate are independent.

    - *R commands and output are below but this can be done by hand easily.*
    - *Since the p-value is smaller 5%, we reject the hypothesis that area and opinion are independent.*
    - *Note that the same p-value arises from a two-sample proportion test that the probabilities of favouring the candidates are the same for the two areas. This can be seen numerically from the following R commands and output.*
    - *The two tests are for* different *hypotheses but the square of the Z has the same value algebraically (it is a good exercise to check this out) and have the same p-values.*

```
observed <- array(c(620, 380, 550, 450), dim = c(2,
    2))
(c1 <- chisq.test(observed, correct = F))
```

```
##
##  Pearson's Chi-squared test
##
## data:  observed
## X-squared = 10.092, df = 1, p-value =
## 0.001489

# manual check of chisquare
chisqman <- (620 - 585)^2/585 + (550 - 585)^2/585 +
    (380 - 415)^2/415 + (450 - 415)^2/415
chisqman

## [1] 10.09165

p1 = 620/1000
p2 = 550/1000
p = (620 + 550)/2000
z = (p1 - p2)/(p * (1 - p) * (1/1000 + 1/1000))^0.5
z^2

## [1] 10.09165
```

Sp 计算

$K_{ij}\langle m\, y_j\rangle$

Wilcox.

Sign test

几种构的区间

几种假设 testing