HW7_Yiyuan_Zhang

Yiyuan Zhang

December 12, 2018

Problem 1

a

According to the problem, the following table is generated.

	First Male	**First Female**	**Total**
Second Male	218(x)	278	496(n_{sm})
Second Female	227	277(y)	504(n_{sf})
Total	445(n_{fm})	555(n_{ff})	1000(n)

Let $p_1 = \frac{x}{n}$, $p_2 = \frac{n_{sm}-x}{n}$,$p_3 = \frac{n_{sf}-y}{n}$,and $p_4 = \frac{y}{n}$.
According to the hypothesis, the gender of offspring within a famility are independent and identically distributed with males and females being equally likely. Thus, $p_{male} = p_{female} = 0.5$, so, theoretically speaking, $p_1 = p_2 = p_3 = p_4 = 0.5 \times 0.5 = 0.25$, which will be our null hypothesis. The alternative hypothesis is that the four probablities are not equal to 0.25. Using Chi-squared test, the test statistic is:

$$TS = \frac{(observed - expected)^2}{expected}$$

In this case, $E_{11} = E_{12} = E_{21} = E_{22} = 250$, and the degree of freedom is 3, using the following R code:

```
ts <- (218-250)^2/250+(278-250)^2/250+(227-250)^2/250+(277-250)^2/250
pchisq(ts,3,lower.tail = FALSE)
```

```
## [1] 0.006531413
```

According to above output, p = 0.0065, which is much smaller than $\alpha = 0.05$, thus, we reject the null hypothesis. So, the data does not support the hypothesis stated in the problem.

b

We use Chi-squared test to test the independence of the gender of the first child to the second. Thus, the null and alternative hypothesis is:
H_0:Gender of the first child is independent of the gender of the second child.
H_a:Gender of the first child is not independent of the gender of the second child.
Under the above null hypothesis,

P(First Male & First Female) = P(First Male) $\times$ P(First Female) = $\frac{445}{1000} \times \frac{496}{1000}$, thus E_{11} =P(First Male & First Female)$\times n$.
Similarly, we get the expected value for every cell:

$$E_{11} = \frac{445}{1000} \times \frac{496}{1000} \times 1000 = 220.72$$
$$E_{12} = \frac{555}{1000} \times \frac{496}{1000} \times 1000 = 275.28$$
$$E_{21} = \frac{445}{1000} \times \frac{504}{1000} \times 1000 = 224.28$$
$$E_{22} = \frac{504}{1000} \times \frac{555}{1000} \times 1000 = 279.72$$

```
ts <- (218-220.72)^2/220.72+(278-275.28)^2/275.28+(227-224.28)^2/224.28+(277-279.72)^2/279.72
pchisq(ts,1,lower.tail = FALSE)
```

```
## [1] 0.7292168
```

According to the above output, p=0.73, which is larger than $\alpha = 0.05$, so we failed to reject the null hypothesis. So, the data support the claim that gender of the first child is independent of the gender of the second child.

Problem 2

According to the problem, we derived the following table:

$n_{11} = x$	$n_{12} = n_1 - x$	$n_1 = n_{1+}$
$n_{21} = y$	$n_{22} = n_2 - y$	$n_{11} = n_{2+}$
n_{+1}	n_{+2}	

Let $\hat{p_1} = \frac{n_{11}}{n_1}$, $\hat{p_2} = \frac{n_{21}}{n_2}$, and $\hat{p} = \frac{(n_{11}+n_{21})}{n_1+n_2}$

Thus, the square of Z statisitic is:

$$(\frac{\hat{p_1} - \hat{p_2}}{SE_{\hat{p_1}-\hat{p_2}}})^2 = \frac{(\hat{p_1} - \hat{p_2})^2}{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}$$

The χ^2 statistic is:

$$\sum \frac{(observed - expected)^2}{expected}$$

can be written as

$$\chi^2 = \frac{(n_1\hat{p_1} - n_1\hat{p})^2}{n_1\hat{p}} + \frac{(n_1(1-\hat{p_1}) - n_1(1-\hat{p}))^2}{n_1(1-\hat{p})} + \frac{(n_2\hat{p_2} - n_2\hat{p})^2}{n_2\hat{p}} + \frac{(n_2(1-\hat{p_2}) - n_2(1-\hat{p}))^2}{n_2(1-\hat{p})}$$

$$= \frac{n_1(\hat{p_1} - \hat{p})^2}{\hat{p}} + \frac{n_1(1 - \hat{p_1} - 1 + \hat{p})^2}{1 - \hat{p}} + \frac{n_2(\hat{p_2} - \hat{p})^2}{\hat{p}} + \frac{n_2(1 - \hat{p_2} - 1 + \hat{p})^2}{1 - \hat{p}}$$

$$= \frac{n_1(\hat{p_1} - \hat{p})^2}{\hat{p}} + \frac{n_1(\hat{p_1} - \hat{p})^2}{1 - \hat{p}} + \frac{n_2(\hat{p_2} - \hat{p})^2}{\hat{p}} + \frac{n_2(\hat{p_2} - \hat{p})^2}{1 - \hat{p}}$$

$$= \frac{n_1(\hat{p_1} - \hat{p})^2 + n_2(\hat{p_2} - \hat{p})^2}{\hat{p}(1 - \hat{p})}$$

$$= \frac{n_1\hat{p_1}^2 + n_2\hat{p_2}^2 - 2\hat{p}(n_1\hat{p_1} + n_2\hat{p_2}) + (n_1 + n_2)\hat{p}^2}{\hat{p}(1 - \hat{p})}$$

For

$$\hat{p} = \frac{n_1\hat{p_1} + n_2\hat{p_2}}{n_1 + n_2}$$

Thus,

$$n_1\hat{p_1} + n_2\hat{p_2} = \hat{p}(n_1 + n_2)$$

So,

$$\chi^2 = \frac{n_1\hat{p_1}^2 + n_2\hat{p_2}^2 - (n_1 + n_2)\hat{p}^2}{p\hat{(}1 - \hat{p})}$$

$$= \frac{\frac{(n_1\hat{p_1}^2 + n_2\hat{p_2}^2)(n_1 + n_2) - (n_1\hat{p_1} + n_2\hat{p_2})^2}{n_1 + n_2}}{\hat{(}p)(1 - \hat{p})}$$

$$= \frac{\frac{n_1n_2(\hat{p_1} - \hat{p_2})^2}{n_1 + n_2}}{\hat{p}(1 - \hat{p})} = (\frac{\hat{p_1} - \hat{p_2}}{SE_{\hat{p_1} - \hat{p_2}}})^2$$

Problem 3

a

According to the problem, the following data is generated:

	Died within 12 months	**Did not die within 12 months**	**Total**
Streptokinase Treatment	2(x)	15	17(n_t)
Control	4(y)	19	23(n_c)
Total	6	34	40

We use Fisher's exact test to test if strptokinase is effective in the treatment of myocardial infarction. Let $p_t = \frac{x}{n_t}$ and $p_c = \frac{y}{n_c}$ denote the probablity of death within 12 months for treatment and control groups.The null and alternative hypothesis are:

$$H_0 : p_t = p_c$$
$$H_a : p_t \neq p_c$$

All the possible tables that are as extreme or more extreme than observed tables are as followed:

	Died within 12 months	**Did not die within 12 months**	**Total**
Streptokinase Treatment	2(x)	13	15(n_t)
Control	4(y)	15	19(n_c)
Total	6	28	34

	Died within 12 months	**Did not die within 12 months**	**Total**
Streptokinase Treatment	1(x)	14	15(n_t)
Control	5(y)	14	19(n_c)
Total	6	28	34

	Died within 12 months	**Did not die within 12 months**	**Total**
Streptokinase Treatment	0(x)	15	15(n_t)
Control	6(y)	13	19(n_c)
Total	6	28	34

	Died within 12 months	**Did not die within 12 months**	**Total**
Streptokinase Treatment	4(x)	11	15(n_t)
Control	2(y)	17	19(n_c)
Total	6	28	34

	Died within 12 months	**Did not die within 12 months**	**Total**
Streptokinase Treatment	5(x)	10	15(n_t)
Control	1(y)	18	19(n_c)
Total	6	28	34

	Died within 12 months	**Did not die within 12 months**	**Total**
Streptokinase Treatment	6(x)	9	15(n_t)
Control	0(y)	19	19(n_c)

	Died within 12 months	Did not die within 12 months	Total
Total	6	28	34

Thus,

$$p = \frac{\binom{15}{2}\binom{19}{4}}{\binom{34}{6}} + \frac{\binom{15}{1}\binom{19}{5}}{\binom{34}{6}} + \frac{\binom{15}{0}\binom{19}{6}}{\binom{34}{6}} + \frac{\binom{15}{4}\binom{19}{2}}{\binom{34}{6}} + \frac{\binom{15}{5}\binom{19}{1}}{\binom{34}{6}} + \frac{\binom{15}{6}\binom{19}{0}}{\binom{34}{6}} = 0.6722$$

Thus, p = 0.6722, which is larger than $\alpha = 0.05$, thus, we failed to reject the null hypothesis. In conclusion, according to this data, the treatment of streptokinase not effective because its 12 month mortality rate is equal to the mortality rate for control.

The result is further confirmed by the output of the following code:

```
q3a <- matrix(c(2,4,13,15),2)
fisher.test(q3a,alternative = "two.sided")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  q3a
## p-value = 0.6722
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.04590206 4.89390008
## sample estimates:
## odds ratio
##   0.586089
```

b

We use score test to test if streptokinase treatment leads to less mortality rate in 12 months
The null and alternative hypothesis are:

$$H_0 : p_t = p_c$$
$$H_a : p_t \neq p_c$$

The test statistic is

$$TS = \frac{\hat{p_t} - \hat{p_c}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_t} + \frac{1}{n_c}))}}$$

Where $\hat{p} = \frac{x+y}{n_t+n_c}$.

```
pt <- 2/15
pc <- 4/19
p <- 6/34
TS <- (pt-pc)/sqrt(p*(1-p)*(1/15+1/19))
TS
```

```
## [1] -0.586253
```

```
pnorm(-abs(TS))*2
```

```
## [1] 0.5577055
```

Thus, p = 0.58 in this case. At level α = 0.05, we failed to rejcet the null hypothesis that the motality rate in 12 months is equal for treatment and control groups. Thus, this data does nof support that streptokinase is effective in treatment of myocardial infarction.
Also, using Chi-squared test with the same hypothesis:

```
chisq.test(q3a,correct= FALSE)
```

```
## Warning in chisq.test(q3a, correct = FALSE): Chi-squared approximation may
## be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  q3a
## X-squared = 0.34369, df = 1, p-value = 0.5577
```

In this case, TS=0.24, and p = 0.6223, thus, we also failed to reject the null.

Problem 4

a

We perform Chi-squared test to test if number 1-10 appear to be equally likely. First, we count the number of 1-10 in the data and generate a vector that contains the number of 1-10. Then perform Chi-squared test on the vector using the following null and alternative hypothesis:

$$H_0 : p_1 = p_2 = p_3 = \ldots = p_{10}$$

H_a :at leaset two of the probablities are not equal

Using the following code:

```
dat <- read.csv("task1_copy-1.csv",header=FALSE)
dat2 <- dat[,1:10]
dat2 <- dat2[complete.cases(dat2),]
vec1<- as.vector(unlist(dat2))
obs <- c(1:10)
for (i in 1:10)
{
  obs[i] <- length(vec1[vec1==i])
  i = i+1
}
chisq.test(obs)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  obs
## X-squared = 34.93, df = 9, p-value = 6.13e-05
```

According to the output, p = 6.13×10^{-5}, which is much smaller than $\alpha = 0.001$.Thus, we reject the null hypothesis. So, in this data, number 1-10 appear to be not equally likely.

b

Using the following code:

```
simsize <- 1000
obStats<-chisq.test(obs)$statistic
simdat <- t(rmultinom(simsize,size = 430, p = rep(.1,10)))
chsqStats <- apply(simdat,1,function(x) chisq.test(x)$statistic)
compare <- chsqStats >= obStats
length(compare[compare==TRUE])/simsize
```

```
## [1] 0
```

According to the output,the percentage of time that the simulated statistics are greater than the observed statistic is 0, with 1000 simulation. Because observed p value is 6.13×10^{-5}, which is approximately 0, we could say that the percentae of time that the simulated statisitcs are greater than the observed statistic is the p value for Chi-squared time for the observed sample. Because the simulated statistics (when Monte Carlo sample is large) is the null distribution that follows χ^2distribution with the same degree of freedom for the Chi-squared test of the observed data. Then according to the definition of p value, the probablity that the observed statistic is samller (or within the null distribution) is the p value. The following code, with 200000 simulations, generates a probability that much closer to the observed p value, which further confirms this point:

```
simsize <- 1000##change to 20000 before submission
obStats<-chisq.test(obs)$statistic
simdat <- t(rmultinom(simsize,size = 430, p = rep(.1,10)))
chsqStats <- apply(simdat,1,function(x) chisq.test(x)$statistic)
compare <- chsqStats >= obStats
length(compare[compare==TRUE])/simsize
```

```
## [1] 0
```

Problem 5

Assuming a constant odd ratio across age-strata, we use Mantel/Haenszel test to test if the true odd ratio is 1. Let θ denot the true odd sample of the population, $\theta_1, \theta_2, \theta_3$ are the odd ratios for Age 35-44, Age 45-54, and Age 55-64, repectively.Then the null and alternative hypothesis are as followed:

$$H_0 : \theta_1 = \theta_2 = \theta_3 = 1$$
$$H_a : \theta_1 = \theta_2 = \theta_3 \neq 1$$

Using the following code:

```
q5 <- array(c(8,52,5,164,25,29,21,138,50,27,61,208),c(2,2,3))
mantelhaen.test(q5,correct = FALSE)
```

```
##
##  Mantel-Haenszel chi-squared test without continuity correction
##
## data:  q5
## Mantel-Haenszel X-squared = 83.725, df = 1, p-value < 2.2e-16
## alternative hypothesis: true common odds ratio is not equal to 1
## 95 percent confidence interval:
##  3.959672 8.904458
## sample estimates:
## common odds ratio
##          5.937907
```

According to the above output, p = 2.2×10^{-16}, which is smaller than 0.001. Thus, we reject the null hypothesis. Thus, the true odd ratio is not 1. Then, using the below code, we use MH estimator to estimate the odd ratio, which equals to 5.94. The same result is also in the output above.

```
(8*164/229+25*138/213+50*208/346)/(5*52/229+21*29/213+61*27/346)
```

```
## [1] 5.937907
```

Problem 6

Accoding to the problem, the following table is generated.

	Sex-linked	**Recessive**	**Dominant**	**Total**
English Cases	46	25	54	125(n_e)
Swiss Cases	1	99	10	110(n_s)
Total	47	124	64	235(n)

To test the association between ethnic origin and genetic type, we perform a Chi-squared test for independence. Let n_{ij} be the value of row i and column j and E_{ij} be the expected value of row i and column j. The null and alternative hypothesis are:
H_0:Ethnic origin is independent of the genetic types.
H_a:Ethnic origin is not independent of the genetic types.
The expected value are calculated as followed:

$$
\begin{aligned}
E_{11} &= \frac{47}{235} \times \frac{125}{235} \times 235 \\
E_{12} &= \frac{124}{235} \times \frac{125}{235} \times 235 \\
E_{13} &= \frac{64}{235} \times \frac{125}{235} \times 235 \\
E_{21} &= \frac{47}{235} \times \frac{110}{235} \times 235 \\
E_{22} &= \frac{124}{235} \times \frac{110}{235} \times 235 \\
E_{23} &= \frac{64}{235} \times \frac{110}{235} \times 235
\end{aligned}
$$

Test statistic is

$$
\sum_{ij}^{i=1,j=1} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}
$$

And the degree of freedom is $(3-1) \times (2-1) = 2$, then using the following code to perform the test:

```
nij <-c(46,25,54,1,99,10)
eij <- c(47*125/235,124*125/235,64*125/235,47*110/235,124*110/235,64*110/235)
ts <- sum((nij-eij)^2/eij)
ts
```

```
## [1] 117.0157
```

```
pchisq(ts,2,lower.tail = FALSE)
```

```
## [1] 3.89371e-26
```

Accoridng to the output, $p = 3.89 \times 10^{-26}$, which is much smaller than 0.001. Thus, we reject the null hypothesis. So, there is statistical evidence that the ethnic origin is not independent to the genetic type and there's an association between the two.

Problem 7

Using the following code, the odd ratio, log odd ratio, 95% CI for OR, and CI graph are generated. With 1-5 indicates group 1-4,5-9,10-24,25-49, 50+ cigarette daily consumption, respectively.

```
n11 <- c(7,7,7,7,7)
n12 <- c(49,516,445,299,41)
n21 <- c(61,61,61,61,61)
n22 <- c(91,615,408,162,20)
OR <- n11*n22/(n12*n21)
logOR <- log(OR)
SE_logOR <- sqrt(1/n11+1/n12+1/n21+1/n22)
CI_OR_upper <-exp(logOR+SE_logOR*qnorm(0.975))
CI_OR_lower <-exp(logOR-SE_logOR*qnorm(0.975))
OR
```

```
## [1] 0.21311475 0.13677087 0.10521275 0.06217446 0.05597761
```

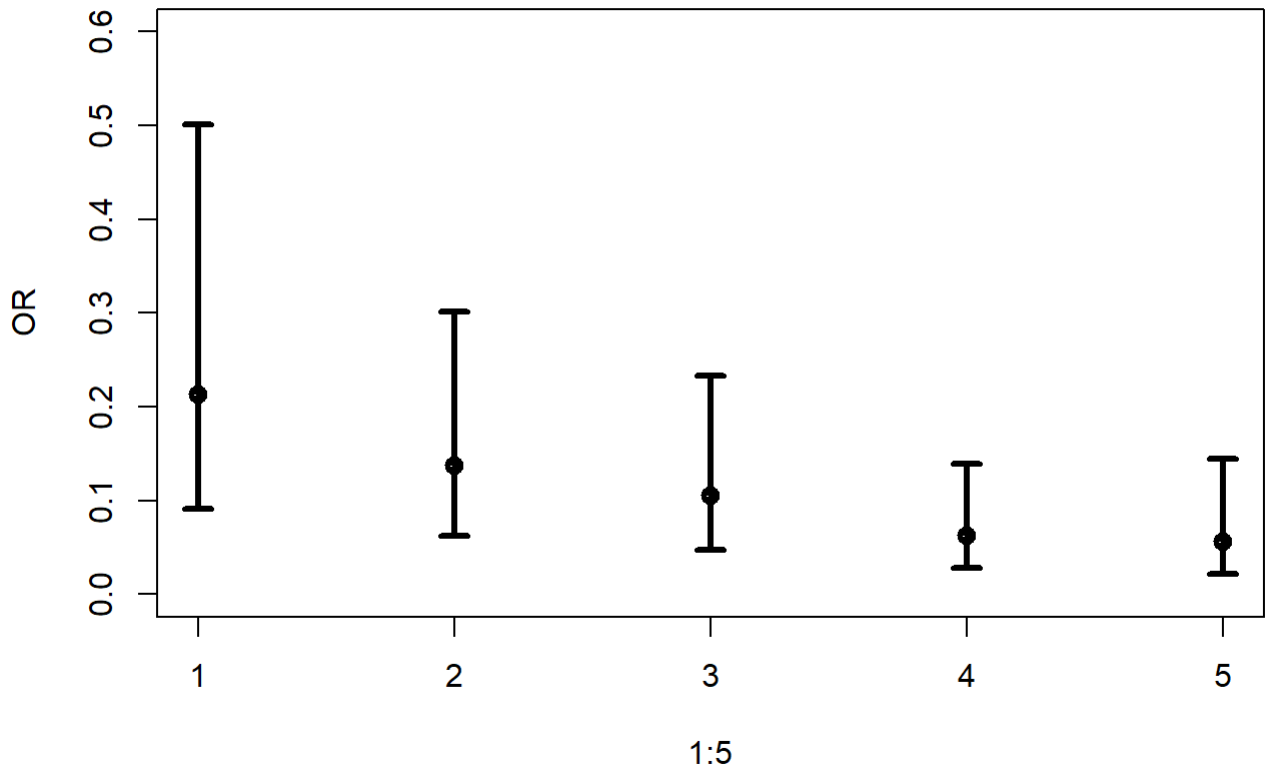
```
logOR
```

```
## [1] -1.545925 -1.989448 -2.251771 -2.777811 -2.882804
```

```
CI <- cbind(CI_OR_lower,CI_OR_upper)
CI
```

```
##      CI_OR_lower CI_OR_upper
## [1,]  0.09056313   0.5015054
## [2,]  0.06201994   0.3016171
## [3,]  0.04757862   0.2326617
## [4,]  0.02779245   0.1390904
## [5,]  0.02170572   0.1443626
```

```
library(plotrix)
plotCI(x=1:5,y=OR,li=CI_OR_lower,ui=CI_OR_upper,lwd=3,ylim=c(0,0.6))
```



According to the output, we can see that the OR are all smaller than 1 and all CIs for OR do not contain 1, which indicates that there's an association between lung cancer and smoking. And with more cigarette smoked daily, the smaller the OR, meaning that there's an higher chance to get lung cancer.

For the relative risk can be estimated from OR, because $OR = RR \times \frac{1-P(C|\bar{S})}{1-P(C|S)}$. Thus, when lung cancer is rare in both smokers and non-smokers, or have the same frequency in somkers and non-smokers, the relative risks can be estimated from OR. So, in this data, the relative risks cannot be estimated, because lung cancer is frequent in smokers and relatively rare in non-smokers.

Problem 8

a

Accoding to the problem, following three tables are generated:

London	Group 0	Group A	Total
Peptic Ulcer	911	579	1490
Control	4578	4219	8797
Total	5489	4798	10287

```
n11 <- 911
n12 <- 579
n21 <- 4578
n22 <- 4219
OR <- n11*n22/(n12*n21)
logOR <- log(OR)
SE_logOR <- sqrt(1/n11+1/n12+1/n21+1/n22)
CI_OR <-exp(logOR+c(-1,1)*qnorm(0.975)*SE_logOR)
OR
```

```
## [1] 1.450019
```

```
CI_OR
```

```
## [1] 1.296051 1.622277
```

As we see in the above output, the estimated OR is 1.45 and the 95% CI is [1.30,1.62], which indicates that the odd of peptic ulcer for group 0 is 1.45 times that of the odds for group A in London.With repeated experiment, in 95% of the times the true value of OR is within the range between 1.30 and 1.62. All these results may indicate that odd of ulcer for group 0 is greater than group A in London.

Manchester	**Group 0**	**Group A**	**Total**
Peptic Ulcer	361	246	607
Control	4532	3775	8307
Total	4893	4021	8914

```
n11 <- 361
n12 <- 246
n21 <- 4532
n22 <- 3775
OR <- n11*n22/(n12*n21)
logOR <- log(OR)
SE_logOR <- sqrt(1/n11+1/n12+1/n21+1/n22)
CI_OR <-exp(logOR+c(-1,1)*qnorm(0.975)*SE_logOR)
OR
```

```
## [1] 1.22236
```

```
CI_OR
```

```
## [1] 1.033641 1.445536
```

As we see in the above output, the estimated OR is 1.22 and the 95% CI is [1.03,1.45], which indicates that the odd of peptic ulcer for group 0 is 1.22 times that of the odds for group A in Manchester. With repeated experiment, in 95% of the times the true value of OR is within the range between 1.03 and 1.45. All these results may indicate

that odd of ulcer for group 0 is greater than group A in Manchester.

Newcastle	Group 0	Group A	Total
Peptic Ulcer	396	219	615
Control	6598	5261	11859
Total	6994	5480	12474

```
n11 <- 396
n12 <- 219
n21 <- 6598
n22 <- 5261
OR <- n11*n22/(n12*n21)
logOR <- log(OR)
SE_logOR <- sqrt(1/n11+1/n12+1/n21+1/n22)
CI_OR <-exp(logOR+c(-1,1)*qnorm(0.975)*SE_logOR)
OR
```

```
## [1] 1.441807
```

```
CI_OR
```

```
## [1] 1.217644 1.707237
```

As we see in the above output, the estimated OR is 1.44 and the 95% CI is [1.22,1.71], which indicates that the odd of peptic ulcer for group 0 is 1.44 times that of the odds for group A in Newcastle With repeated experiment, in 95% of the times the true value of OR is within the range between 1.22 and 1.71. All these results may indicate that odd of ulcer for group 0 is greater than group A in Newcastle.

b

To estimate P(ulcer|A)-P(ulcer|0), we have to combine the data from all three cities. Thus, we have to test if there's a difference between the data from all three cities. Also, we need to know that if the subjects matched for all three cities.

Problem 9

According to the problem, the following data is generated:

		Treatment S		
		Lived at least 5 years	Died within 5 years	Total
Treatment R	Lived at least 5 years	10	1	11
	Died within 5 years	8	1	9
	Total	18	2	20

As we can observe in the data, the the probability when treatment S is effective while treatment R is not is larger than the probablity when treatment S is not effective while treatment R is effective. To test this more rigorously, we perform Mcnemar test to test if the proportion of treatment S is effective while treatment R is not equals to the proportion of treatment S is not effective while treatment R is effective. Let n_{ij} equals to the value in row i and column j and $n = 20$. $\hat{\pi}_{ij} = \frac{n_{ij}}{n}$. The null and alternative hypothesis are:

$$H_0 : \pi_{12} = \pi_{21}$$
$$H_a : \pi_{12} \neq \pi_{21}$$

And the test statistic is computed by:

$$\frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

Using the following code, we can perform the test by hand or by a single function:

```
n12 <- 1
n21 <- 8
ts<-(n12-n21)^2/(n12+n21)
pchisq(ts,1,lower.tail = FALSE)
```

```
## [1] 0.01963066
```

```
mcnemar.test(matrix(c(10,8,1,1),2),correct = FALSE)
```

```
##
##  McNemar's Chi-squared test
##
## data:  matrix(c(10, 8, 1, 1), 2)
## McNemar's chi-squared = 5.4444, df = 1, p-value = 0.01963
```

According to the output, p=0.020, which is smaller than 0.05. Thus, we reject the null hypothesis. So, given this data, there's statistical evidence that there's a difference in the effectiveness between simple mastectomy and radical mastecomy in treating breast cancer. More specifically, simple mastectomy has better effect than radical mastectomy.

Problem 10

The following data is generated according the problem:

		Antibiotic A		
		Effective	Not Effective	Total
Antibiotic B	Effective	40	16	56
	Not Effective	20	3	23
	Total	60	19	79

To compare the effectiveness of tow different antibiotics, A and B, in treating gonorrhea, we perform Mcnemar test to test if the proportion of A is effective while B is not equals to the proportion of A is not effective while B is . Let n_{ij} equals to the value in row i and column j and $n = 20$. $\hat{\pi}_{ij} = \frac{n_{ij}}{n}$. The null and alternative hypothesis are:

$$H_0 : \pi_{12} = \pi_{21}$$
$$H_a : \pi_{12} \neq \pi_{21}$$

Using the following code:

```
mcnemar.test(matrix(c(40,20,16,3),2),correct = FALSE)
```

```
##
##  McNemar's Chi-squared test
##
## data:  matrix(c(40, 20, 16, 3), 2)
## McNemar's chi-squared = 0.44444, df = 1, p-value = 0.505
```

According to the output, p=0.51, which is larger than 0.05. Thus, we failed to reject the null hypothesis. So, given this data, there's statistical evidence that there's no difference in the effectiveness between antibiotic A and B in treating gonorrhea.

Problem 11

Theoretically speaking, McNemar's test is equivalent to the CMH test where subjects is the stratifying variable and each 2x2 table is the observed zero-one table for each subject. Thus, we generate the following tables to check if the test statistics are equal for two cases.

First, we make the table for McNemar's test:

		First Survey		
		Yes	No	Total
Second Survey	Yes	a	b	a+b
	No	c	d	c+d
	Total	a+c	b+d	a+b+c+d

The test statistic for McNemar's test is

$$\frac{(b-c)^2}{b+c}$$

.

Then, the following 4 possible tables are generated for CMH test. According to the theory, the four table each have a, b, c, and d subjects, respectively.So, there are a, b, c, and d times of each table from top to bottom. The corresponding number is marked on the top-left cell of each table.
For CMH test statistic is

$$\frac{[\sum_k (n_{11k} - E(n_{11k}))]^2}{\sum_k Var(n_{11k})}$$

$$E(n_{11k}) = \frac{n_{1+k} n_{+1k}}{n_{++k}}$$

$$Var(n_{11k}) = \frac{n_{1+k} n_{2+k} n_{+1k} n_{+2k}}{n_{++k}^2 (n_{++k} - 1)}$$

a		**Response**		
		Yes	No	Total
Time	First	1	0	1
	Second	1	0	1
	Total	2	0	2

b		**Response**		
		Yes	No	Total
Time	First	1	0	1
	Second	0	1	1
	Total	1	1	2

c		**Response**		
		Yes	No	Total
Time	First	0	1	1
	Second	1	0	1
	Total	1	1	2

d		**Response**		
		Yes	No	Total
Time	First	0	1	1
	Second	0	1	1
	Total	0	2	2

According to above, the numerator and denumerator for CMH test statistic for table corresponding to a is

$$E(n_{11a} = \frac{1 \times 2}{2}) = 1$$
$$Var(n_{11a}) = \frac{0}{4} = 0$$
$$n_{11a} - E(n_{11a}) = 1 - 1 = 0$$

Similarly, we can find that for b,c,d:

$$Var(n_{11b}) = \frac{1}{2}$$
$$n_{11b} - E(n_{11b}) = \frac{1}{4}$$
$$Var(n_{11c}) = -\frac{1}{2}$$
$$n_{11c} - E(n_{11c}) = \frac{1}{4}$$
$$Var(n_{11d}) = 0$$
$$n_{11d} - E(n_{11d}) = 0$$

Combining all the above, we find the test statistic for CMH test in this case is:

$$\frac{(\frac{b}{2} - \frac{c}{2})^2}{\frac{b}{4} + \frac{c}{4}}$$
$$= \frac{\frac{1}{4}(b-c)^2}{\frac{1}{4}(b+c)}$$
$$= \frac{(b-c)^2}{b+c}$$

Thus, we show that the test statistics for McNemar's test is equivalent to the test statistic when performing a MH test for all 2x2 tables.
Using the same condition in question 9 and the following code, we can further confirm the conclusion.

```
q5 <- array(c(1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,0,1,
0,0,1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,0,1,0,1,1,0,0,0,1,1),c(2,2,20))
mantelhaen.test(q5,correct = FALSE)
```

```
##
##  Mantel-Haenszel chi-squared test without continuity correction
##
## data:  q5
## Mantel-Haenszel X-squared = 5.4444, df = 1, p-value = 0.01963
## alternative hypothesis: true common odds ratio is not equal to 1
## 95 percent confidence interval:
##   1.000586 63.962519
## sample estimates:
## common odds ratio
##                 8
```

```
mcnemar.test(matrix(c(10,8,1,1),2),correct = FALSE)
```

```
##
##  McNemar's Chi-squared test
##
## data:  matrix(c(10, 8, 1, 1), 2)
## McNemar's chi-squared = 5.4444, df = 1, p-value = 0.01963
```

Problem 12

a

Let n be the total number of subject and n_{ij} be the value in row i and column j. $\hat{\pi}_{ij} = \frac{n_{ij}}{n}$. Thus, the null and alternative hypothesis are:

$$H_0 : \pi_{12} = \pi_{21}, \pi_{13} = \pi_{31}, \pi_{23} = \pi_{32}$$

H_a: At least one of the above pair is not equal

b

Uder null hypothesis,

$$\begin{aligned} \pi_{12} &= \pi_{21} \\ \frac{n_{12}}{n} &= \frac{n_{21}}{n} \\ n_{12} &= n_{21} \end{aligned}$$

Thus,

$$E_{12} = E_{21} = \frac{n_{12} + n_{21}}{2} = \frac{267 + 135}{2} = 201$$

Similarly,

$$\begin{aligned} E_{13} = E_{31} &= \frac{n_{13} + n_{31}}{2} = \frac{255 + 240}{2} = 247.5 \\ E_{23} = E_{32} &= \frac{n_{23} + n_{32}}{2} = \frac{139 + 234}{2} = 186.5 \end{aligned}$$

c

Accoding to part a and b, we perform Chi-squared test using the following code:

```
n1 <- c(267,255,139)
n2 <- c(135,240,234)
e <- (n1+n2)/2
ts <- sum((n1-e)^2/e + (n2-e)^2/e)
ts
```

```
## [1] 67.99354
```

```
pchisq(ts,3,lower.tail = FALSE)
```

```
## [1] 1.147668e-14
```

Accoding to the above output, p = 1.14×10^{-14}, which is much smaller than 0.001. Thus, we reject the null hypothesis. In conclsion, the data support that there's a difference between the probablity of moving from place a to be and moving from place b to a.