140.652 Problem Set 7 Solutions

Problem 1

The table below gives the children's genders in a random sample of 1,000 two children families.

	First child				
	Male		Female		
Second child	Male	Female	Male	Female	Total
Count	218	227	278	277	1,000

For this problem, assume that the families with two children have been randomly sampled. We can re-format the table in our usual 2×2 format:

		First child		
		Male	Female	
Second child	Male	218	278	496
	Female	227	277	504
		445	555	1000

a. It is typically thought that the gender of offspring within a family are independent and identically distributed with males and females being equally likely. Is this hypothesis supported by the data above?

Let p_F be the probability that an offspring is female and p_M the probability that an offspring is male. Note that since $p_F = 1 - p_M$, one way of testing whether the gender of offspring within a family are independent and identically distributing is by using the score test to test the hypothesis,

$$H_0 : p_F = 0.5$$
$$H_A : p_A \neq 0.5$$

Note that we have 2000 children total, so our estimate for $\hat{p}_F$ is given by,

$$\hat{p}_F = \frac{227 + 278 + 2 * 277}{1000} = 0.5295$$

The score test statistic is,

$$TS = \frac{\hat{p}_F - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.5295 - 0.5}{\sqrt{0.5(0.5)/2000}} \approx 2.639$$

resulting in a p-value of

```
pF <- (227 + 278 + 2*277)/2000
TS <- (pF - 0.5)/sqrt(0.5*0.5/2000)
2*pnorm(TS, lower.tail = FALSE)
```

```
[1] 0.008325891
```

For $\alpha = 0.05$ and a p-value of 0.008, we reject the null and conclude that the probability that an offspring is male may not be the same as the probability that an offspring is female. In other words, it seems that the data does not support the hypothesis.

Note: We can also test this hypothesis using a χ^2 goodness of fit test – both should lead you to the same answer.

b. Specifically test independence of the gender of the first child to the second.

Under the null hypothesis, we assume that gender of the first child is independent of the second. Let $p_{M\cdot}$ denote the probability that child one is male. Then, we have

$$\hat{p}_{M\cdot} = \frac{445}{1000}$$
$$\hat{p}_{F\cdot} = \frac{555}{1000}$$
$$\hat{p}_{\cdot M} = \frac{596}{1000}$$
$$\hat{p}_{\cdot F} = \frac{504}{1000}$$

Let E_{FM} denote the expected number of families in which the first child is female and the second child is male. Under the assumption of independence, we have,

$$E_{MF} = p_{M\cdot} \times p_{\cdot F} \times 1000 = \frac{445 \times 504}{1000}$$
$$E_{FM} = p_{F\cdot} \times p_{\cdot M} \times 1000 = \frac{555 \times 596}{1000}$$
$$E_{MM} = p_{M\cdot} \times p_{\cdot M} \times 1000 = \frac{445 \times 596}{1000}$$
$$E_{FF} = p_{F\cdot} \times p_{\cdot F} \times 1000 = \frac{555 \times 504}{1000}$$

The χ^2 test statistic is

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})}{\text{Expected}}$$

```
O_FF <- 277
O_MM <- 218
O_FM <- 278
O_MF <- 227

E_MF <- 445 * 504/1000
E_FM <- 555 * 496/1000
E_MM <- 445 * 496/1000
E_FF <- 555 * 504/1000

# Calculate ChiSq Test Statistic
TS <- (O_FF - E_FF)^2/E_FF + (O_MM - E_MM)^2/E_MM +
  (O_FM - E_FM)^2/E_FM + (O_MF - E_MF)^2/E_MF

pchisq(TS, df = 1, lower.tail = FALSE)
```

```
[1] 0.7292168
```

With $\alpha = 0.05$ and a p-value of 0.73, we fail to reject the null and conclude that the gender of the first child is not independent of the gender of the second gender.

Problem 2

Consider the hypothesis testing problem of comparing two binomial probabilities $H_0 : p_1 = p_2$. Show that the square of statistic $(\hat{p}_1 - \hat{p}_2)/\text{SE}_{\hat{p}_1 - \hat{p}_2}$ is the same as the χ^2 statistic. Here, the standard error in the denominator is calculated under the null hypothesis. (Clearly define any notation you introduce.)

See Homework 6, Problem 1 solutions

Problem 3

A study of the effectiveness of *streptokinase* in the treatment of patients who have been hospitalized after myocardial infarction involves a treated and control group. In the streptokinase group, 2 of 15 patients died within 12 months. In the control group, 4 of 19 died with 12 months.

a. Use Fisher's exact test to test for a difference in mortality rates. Do this by hand by writing down all possible tables with fixed marginal totals. You may confirm your results with a computer.

Let p_T and p_C be the probability of death in the treatment and control groups, respectively. We observe,

	Dead	Not Dead	N
Streptokinase	2	13	15
Control	4	15	19
Total	6	28	34

We want to test the hypothesis,

$$H_0 : p_T = p_C$$
$$H_A : p_T \neq p_C$$

There are two approaches we can use to test this hypothesis and get a p-value: find the one-sided p-value and double it, or use the `alternative = "two.sided"` argument in `fisher.test()` function in `R`.

Approach 1: doubling the one-sided p-value

To get the one-sided p-value, we are interested in all cases where p_T is less than or equal to our observed p_T. These cases and their probabilities are given below.

	Dead	Not Dead	N
Streptokinase	0	15	15
Control	6	13	19
Total	6	28	34

$$P_0 = \frac{\binom{15}{0}\binom{19}{6}}{\binom{34}{6}} = 0.02017$$

	Dead	Not Dead	N
Streptokinase	1	14	15
Control	5	14	19
Total	6	28	34

$$P_1 = \frac{\binom{15}{1}\binom{19}{5}}{\binom{34}{6}} = 0.1297$$

	Dead	Not Dead	N
Streptokinase	2	13	15
Control	4	15	19
Total	6	28	34

$$P_2 = \frac{\binom{15}{2}\binom{19}{4}}{\binom{34}{6}} = 0.3026$$

Thus, the one-sided p-value is given by $P_0 + P_1 + P_2 = 0.45247$, and the two-sided p-value is 0.905. Verifying in R,

```
data <- matrix(c(2,13,4,15), byrow = TRUE, ncol = 2)

# One-sided fisher's exact test
fisher_test <- fisher.test(data, alternative="less")
fisher_test
```

```
    Fisher's Exact Test for Count Data

data:  data
p-value = 0.4525
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.000000 3.702048
sample estimates:
odds ratio
  0.586089
```

```
# Double p-value
2 * fisher_test$p.value
```

```
[1] 0.9049449
```

Approach 2: calculating the two-sided alternative

In the Fisher's exact test, the two-sided p-value is found by adding the probabilities of all outcomes with probability less than or equal to the observed outcomes. In addition to the three tables above, the other possible outcomes (for fixed row and column margins) and their probabilities are,

	Dead	Not Dead	N
Streptokinase	3	12	15
Control	3	16	19
Total	6	28	34

$$P_3 = \frac{\binom{15}{3}\binom{19}{3}}{\binom{34}{6}} = 0.3278$$

	Dead	Not Dead	N
Streptokinase	4	11	15
Control	2	17	19
Total	6	28	34

$$P_4 = \frac{\binom{15}{4}\binom{19}{2}}{\binom{34}{6}} = 0.1736$$

	Dead	Not Dead	N
Streptokinase	5	10	15
Control	1	18	19
Total	6	28	34

$$P_5 = \frac{\binom{15}{5}\binom{19}{1}}{\binom{34}{6}} = 0.04242$$

	Dead	Not Dead	N
Streptokinase	6	9	15
Control	0	19	19
Total	6	28	34

$$P_6 = \frac{\binom{15}{6}\binom{19}{0}}{\binom{34}{6}} = 0.003721$$

The two-sided p-value is given by, $P_0 + P_1 + P_2 + P_4 + P_5 + P_6 = 0.6722$. Using the `alternative = two.sided` parameter in `fisher.test()` in `R` to verify, we have

```
fisher.test(data, alternative = "two.sided")
```

```
    Fisher's Exact Test for Count Data

data:  data
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. Compare your results using the test statistics based on the normal and χ^2 approximations.

Normal approximation:

Rather than using Fisher's exact test, we can test our hypothesis using the score test which uses a normal approximation. The score test statistic is,

$$TS = \frac{\hat{p}_T - \hat{p}_C}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_T} + \frac{1}{n_C}\right)}}$$

where $\hat{p} = \frac{\hat{p}_T n_T + \hat{p}_C n_C}{n_T + n_C}$.

```
pT <- 2/15
pC <- 4/19
nT <- 15
nC <- 19

# Calculate test statistic
p <- (2 + 4)/(15 + 19)
TS_norm <- (pT - pC)/sqrt(p*(1-p)*(1/nT + 1/nC))

# calculate p-value
2*pnorm(TS_norm)
```

```
[1] 0.5577055
```

χ^2 approximation:

The χ^2 test statistic is given by,

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

$$O_{11} = 2 \qquad E_{11} = \frac{15}{34} \cdot \frac{6}{34} \times 34$$

$$O_{12} = 13 \qquad E_{12} = \frac{15}{34} \cdot \frac{28}{34} \times 34$$

$$O_{21} = 4 \qquad E_{21} = \frac{19}{34} \cdot \frac{6}{34} \times 34$$

$$O_{22} = 15 \qquad E_{22} = \frac{19}{34} \cdot \frac{28}{34} \times 34$$

```
O_11 <- 2
O_12 <- 13
O_21 <- 4
O_22 <- 15

E_11 <- 15/34 * 6/34 * 34
E_12 <- 15/34 * 28/34 * 34
E_21 <- 19/34 * 6/34 * 34
E_22 <- 19/34 * 28/34 * 34

# Chi-sq T-statistic
TS_Chi <- (O_11 - E_11)^2/E_11 + (O_12 - E_12)^2/E_12 +
  (O_21 - E_21)^2/E_21 + (O_22 - E_22)^2/E_22

# p-value
pchisq(TS_Chi, 1, lower.tail = FALSE)
```

```
[1] 0.5577055
```

Summarizing our results, we have

Test	p-value (two-sided)
Fisher's Exact	0.6722
Normal approx	0.5577
χ^2 approx	0.5577

With $\alpha = 0.05$ for all three tests we fail to reject the null and conclude that there may be a difference in mortality by treatment groups. Due to the small sample size, the exact test provides the most appropriate and conservative estimate for the p-value in comparison to the normal and χ^2 approximation.

Problem 4

Download the class simulation data set "task1.csv" from the course web site. Here's the commands that I used to read it in

```
dat <- read.csv("task1.csv", header = FALSE)
dat2 <- dat[,1 : 10]
dat2 <- dat2[complete.cases(dat2),]
vec1 <- as.vector(unlist(dat2))
```

Dat is the original data. Dat2 contains only the data, removing any subjects containing errors. Vec1 is the data disregarding subject level information.

a. Do the numbers 1-10 appear to be equally likely? Perform the appropriate Chi-squared test.

Suppose the data are randomly selected from the population of origin. To test whether the number appear to be equally likely, we can perform a χ^2 test on the table of the numbers. That is, the test statistic is given by,

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})}{\text{Expected}}$$

Since we assume that all numbers are equally likely, the expected counts of each number is 0.1×430 where 430 is the length of `vec1`.

```
# Read in the data
dat <- read.csv("./data/task1.csv", header = FALSE)
dat2 <- dat[,1 : 10]
dat2 <- dat2[complete.cases(dat2),]
vec1 <- as.vector(unlist(dat2))

# Use chisq.test
chisq.test(table(vec1))
```

```
    Chi-squared test for given probabilities

data:  table(vec1)
X-squared = 34.93, df = 9, p-value = 6.13e-05
```

```
# Calculate chisq manually
chisq_stat <- sum((table(vec1) - 0.1 * length(vec1))^2/(0.1*length(vec1)))
chisq_stat
```

```
[1] 34.93023
```

With $\alpha = 0.05$, we reject the null and conclude that the numbers 1-10 are not equally likely.

b. Approximate an exact Chi-squared test by doing the following. Simulate 1,000 random nultinomials under the null hypothesis with the command

```
simdat <- t(rmultinom(1000, size = length(vec1), p = rep(.1, 10)))
```

Obtain the chi-squared statistics for each with the command

```
chsqStats <- apply(simdat, 1, function(x) chisq.test(x)$statistic)
```

Calculate the percentage of time that these statistics are greater than the observed statistic. Explain how, provided the Monte Carlo sample is large, this is a P-value.

The p-value is the probability of observing a value as or more extreme than the test statistic under the null. Here, we have simulated data, then generated χ^2 test statistics. The percentage of the time that the resulting test statistics from samples from the null distribution are greater than the observed statistic is by definition the p-value. Note that while the p-value generated here is 0, the true p-value is not zero; we just did not simulate enough times.

```
set.seed(1)
simdat <- t(rmultinom(1000, size = length(vec1), p = rep(.1, 10)))
chsqStats <- apply(simdat, 1, function(x) chisq.test(x)$statistic)
mean(chsqStats >= chisq_stat)
```

```
[1] 0
```

Problem 5

A case-control study of esophageal cancer was performed. Daily alcohol consumption was ascertained (80+ gm = high, 0 − 79 gm = low). The data was stratified by 3 age groups.

	Alcohol			Alcohol			Alcohol	
	H	L		H	L		H	L
case	8	5	case	25	21	case	50	61
control	52	164	control	29	138	control	27	208
	Age 35-44			Age 45-54			Age 55-64	

Assuming a constant odds ratio across age-strata, test to see if the odds ratio is 1. If not, estimate it.

Let θ_1, θ_2, and θ_3 denote the odds ratio for individuals aged 35-44, aged 45-54, and aged 55-65, respectively. We want to test the hypothesis,

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

To test this hypothesis, we will use the Mantel/Haenszel estimator/test. Recall that the Mantel/Hanzel estimator is given by,

$$\hat{\theta}_{MH} = \frac{\sum_k \frac{n_{11k} n_{22k}}{n_{++k}}}{\sum_k \frac{n_{21k} n_{12k}}{n_{++k}}}$$

In the context of the problem, this is given by,

$$\hat{\theta}_{MH} = \frac{\frac{8 \times 164}{229} + \frac{25 \times 138}{213} + \frac{50 \times 208}{346}}{\frac{52 \times 5}{229} + \frac{29 \times 21}{213} + \frac{27 \times 61}{346}} = 5.938$$

Verifying our result in R, we have,

```
a35 <- matrix(c(8,5,52,164), ncol=2, byrow=TRUE)
a45 <- matrix(c(25,21,29,138), ncol=2, byrow=TRUE)
a55 <- matrix(c(50,61,27,208), ncol=2, byrow=TRUE)

mantelhaen.test(array(c(a35,a45, a55), c(2, 2, 3)),correct=FALSE)
```

```
    Mantel-Haenszel chi-squared test without continuity correction

data:  array(c(a35, a45, a55), c(2, 2, 3))
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
```

With $\alpha = 0.05$, the estimated common odds ratio is 5.94 with a p-value of 2.2×10^{-16} and 95% confidence interval of $(3.96, 8.90)$ suggesting that we reject the null and conclude that the common odds ratio is not 1.

Problem 6

Retinitis pigmentosa is a disease which manifests itself via different genetic modes of inheritance. Cases have been documented with a dominant, recessive, and sex-linked form of inheritance. It has been conjectured that the form of inheritance is related to the ethnic origin of the individual. Cases of the disease have been surveyed in an English and Swiss population with the following results: out of 125 English cases, 46 had sex-linked disease, 25 had recessive disease and 54 had dominant disease; out of the 110 Swiss cases, one had sex-linked disease, 99 had recessive disease, and 10 had dominant disease. Based on these data is there a significant association between ethnic origin and genetic type? Analyze and interpret (in words) this data.

We can summarize the data using the following table,

		Disease			
		Dominant	Recessive	Sex-Linked	
Ethnic Origin	English	54	25	46	125
	Swiss	10	99	1	110
		64	124	47	235

We want to test whether ethnic origin and genetic type are independent. To test this hypothesis, we can used the χ^2 test. Recall that the χ^2 test statistic is given by,

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

For our data set, we have,

$$O_{11} = 54 \qquad E_{11} = \frac{125}{235} \cdot \frac{64}{235} \times 235$$

$$O_{12} = 25 \qquad E_{12} = \frac{125}{235} \cdot \frac{124}{235} \times 235$$

$$O_{13} = 46 \qquad E_{12} = \frac{125}{235} \cdot \frac{47}{235} \times 235$$

$$O_{21} = 10 \qquad E_{12} = \frac{110}{235} \cdot \frac{64}{235} \times 235$$

$$O_{22} = 99 \qquad E_{12} = \frac{110}{235} \cdot \frac{124}{235} \times 235$$

$$O_{23} = 1 \qquad E_{12} = \frac{110}{235} \cdot \frac{47}{235} \times 235$$

```
O_11 <- 54
O_12 <- 25
O_13 <- 46
O_21 <- 10
O_22 <- 99
O_23 <- 1

E_11 <- 125/235*64/235*235
E_12 <- 125/235*124/235*235
E_13 <- 125/235*47/235*235
E_21 <- 110/235*64/235*235
E_22 <- 110/235*124/235*235
E_23 <- 110/235*47/235*235

chistat <- (O_11 - E_11)^2/E_11 + (O_12 - E_12)^2/E_12 + (O_13 - E_13)^2/E_13 +
  (O_21 - E_21)^2/E_21 + (O_22 - E_22)^2/E_22 + (O_23 - E_23)^2/E_23
chistat
```

```
[1] 117.0157
```

```
pchisq(chistat, df = 2, lower.tail = FALSE)
```

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

With $\alpha = 0.05$ and a p-value of 3.89×10^{-26} we reject the null and conclude that there may be an association between ethic origin and genetic type.

Problem 7

In a study of the association between cigarette smoking and lung cancer, 1,357 male lung cancer patients were compared with 1,357 controls in terms of their cigarette consumption as follows:

	Cigarette Consumption Daily						
	0	1−	5−	15−	25−	50+	Total
Lung cancer patients	7	49	516	445	299	41	1,357
Controls	61	91	615	408	162	20	1,357

Compute the odds ratio and log odds ratio in each of the 5 smoking groups compared with non-smokers. Find confidence intervals and graphically display. Comment and interpret. Can relative risks be estimated. Why or why not.

Let $p_0, p_{1-}, p_{5-}, p_{15-}, p_{25-}$ and p_{50+} be the probability of lung cancer in individuals who consume $0, 1-, 5-, 15-, 25-$ and $50-$ cigarettes daily. We can estimate the odds ratio and log-odds ratio using,

$$\hat{OR}_k = \frac{\hat{p}_k/(1-\hat{p}_k)}{\hat{p}_0/(1-\hat{p}_0)}$$

for $k = 1-, 5-, 15-, 25-, 50+$. Recall that from the delta method, the 95% confidence interval for the log-odds ratio is:

$$\hat{LOR}_k \pm z_{1-0.052}\sqrt{\frac{1}{\hat{p}_k n_k} + \frac{1}{(1-\hat{p}_k)n_k} + \frac{1}{p_0 n_0} + \frac{1}{(1-p_0)n_0}}$$

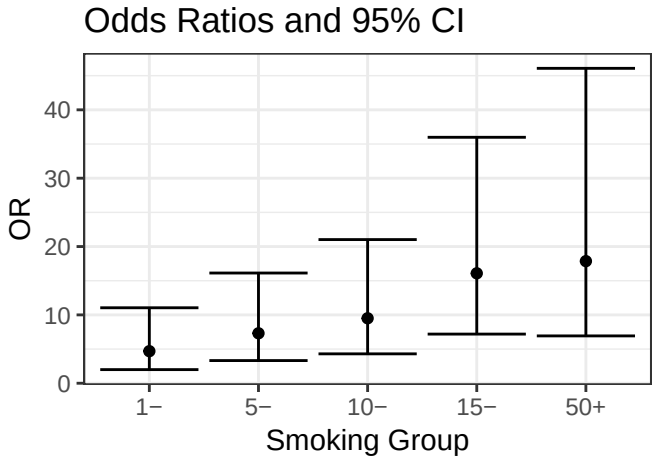
To get the 95% CI for the odds ratio, we can exponentiate the endpoints for the 95% CI for the log-odds ratio.

```
cancerS <- c(49, 516, 445, 299, 41)
controlS <- c(91, 615, 408, 162, 20)

# Calculate OR and LOR
OR <- (cancerS/controlS)/(7/61)
LOR <- log(OR)

# Calculate 95% CI for LOR and OR
LOR_SE <- sqrt(1/cancerS + 1/controlS + 1/7 + 1/61)
CI_LOR_low  <- LOR - 1.96*LOR_SE
CI_LOR_high <- LOR + 1.96*LOR_SE
CI_OR_low   <- exp(CI_LOR_low)
CI_OR_high  <- exp(CI_LOR_high)

# Arrange in a neat data frame
OR_data <- data.frame(cat = c("1-", "5-", "10-", "15-", "50+"),
```

```
  LOR, CI_LOR_low, CI_LOR_high, OR, CI_OR_low, CI_OR_high)
OR_data
```

```
  cat      LOR CI_LOR_low CI_LOR_high        OR CI_OR_low CI_OR_high
1  1- 1.545925  0.6901252    2.401724  4.692308  1.993965   11.04219
2  5- 1.989448  1.1985825    2.780314  7.311498  3.315414   16.12408
3 10- 2.251771  1.4581553    3.045386  9.504552  4.298024   21.01815
4 15- 2.777811  1.9726161    3.583006 16.083774  7.189460   35.98153
5 50+ 2.882804  1.9354099    3.830197 17.864286  6.926883   46.07162
```

```
# Plot results for OR only
OR_data %>% ggplot(aes(x = cat)) +
  geom_point(aes(y = OR)) +
  geom_errorbar(aes(ymin = CI_OR_low, ymax = CI_OR_high)) +
  labs(title = "Odds Ratios and 95% CI", x = "Smoking Group", y = "OR") +
  scale_x_discrete(limits = c("1-", "5-", "10-", "15-", "50+"))+
  theme_bw()
```



From the plot above, we see that the greater cigarette consumption, the higher odds of lung cancer as compared to non-smokers. In addition, none of the confidence intervals overlap with 1, indicating a significant difference risk of lung cancer in smokers and non-smokers for all categories of cigarette smoking considered here.

Note that we cannot use the odds ratio to estimate the relative risk because we do not know the prevalence of lung cancer in the population.

Problem 8

In a retrospective study of the possible effect of blood group on the incidence of peptic ulcers, Woolf (1955) obtained data from three cities. The table gives for each city data for blood groups 0 and A only. In each city, blood group is recorded for peptic ulcer subjects and for a control series of individuals not having peptic ulcer.

	Peptic Ulcer		Control	
	Group 0	Group A	Group 0	Group A
London	911	579	4578	4219
Manchester	361	246	4532	3775
Newcastle	396	219	6598	5261

a. Compute the odds ratio for each city with a confidence interval. Interpret.

Let $p_{L,A}, p_{M,A}, p_{N,A}$ be the probability of peptic ulcers in blood group A for individuals in London, Manchester, and Newcastle, respectively. Additionally, let $p_{L,0}, p_{M,0}, p_{N,0}$ be the probability of peptic ulcers in blood group 0 for individuals in London, Manchester, and Newcastle, respectively. We can estimate the odds ratio and log-odds ratio using,

$$\hat{OR}_k = \frac{\hat{p}_{k,A}/(1-\hat{p}_{k,A})}{\hat{p}_{k,0}/(1-\hat{p}_{k,0})}$$

for $k = L, M, N$. Recall that from the delta method, the 95% confidence interval for the log-odds ratio is:

$$\hat{LOR}_k \pm z_{1-0.052}\sqrt{\frac{1}{\hat{p}_{k,A}n_{k,A}} + \frac{1}{(1-\hat{p}_{k,A})n_{k,A}} + \frac{1}{p_{k,0}n_{k,0}} + \frac{1}{(1-p_{k,0})n_{k,0}}}$$

To get the 95% CI for the odds ratio, we can exponentiate the endpoints for the 95% CI for the log-odds ratio.

```
# Get OR and LOR
OR <- c((579/911)/(4219/4578),
        (246/361)/(3775/4532),
        (219/396)/(5261/6598))
LOR <- log(OR)

# Calculate CI
SE_LOR <- c(sqrt(1/579 + 1/911 + 1/4219 + 1/4578),
            sqrt(1/246 + 1/361 + 1/3775 + 1/532),
            sqrt(1/219 + 1/396 + 1/5261 + 1/6598))
CI_LOR_low <- LOR - 1.96 * SE_LOR
CI_LOR_high <- LOR + 1.96 * SE_LOR
CI_OR_low <- exp(CI_LOR_low)
CI_OR_high <- exp(CI_LOR_high)

# Summarize in a table
data.frame(location = c("London", "Manchester", "Newcastle"),
           OR, CI_OR_low, CI_OR_high)
```

```
    location        OR CI_OR_low CI_OR_high
1     London 0.6896464 0.6164163  0.7715761
2 Manchester 0.8180896 0.6794200  0.9850616
3  Newcastle 0.6935742 0.5857400  0.8212606
```

Note that the odds ratio for peptic ulcers in each location is less than 1 and 1 is not contained in the 95% CI. Thus, with $\alpha = 0.05$, we conclude that individuals with blood type A have a lower risk of developing peptic ulcers in these cities than individuals with blood type 0.

b. Suppose that it is required to estimate P(ulcer|A) − P(ulcer|0). What further information is needed to do this from the current data?

To estimate risk from odds, we would need to know the prevalence of peptic ulcer in each population.

Problem 9

Suppose we wish to compare two treatments for breast cancer, viz., simple mastectomy (S) and radical mastectomy (R). We form matched pairs of women who are within the same decade of age and with the same clinical condition to receive the two treatments and measure their 5-year survival. The results are given

(L=lived at least 5 years, D=died within 5 years) below. Perform an analysis of this data, and interpret your results.

Pair	Treatment S Person	Treatment R Person	Pair	Treatment S Person	Treatment R Person
1	L	L	11	D	D
2	L	D	12	L	D
3	L	L	13	L	L
4	L	L	14	L	L
5	L	L	15	L	D
6	D	L	16	L	L
7	L	L	17	L	D
8	L	D	18	L	D
9	L	D	19	L	L
10	L	L	20	L	D

Since we have paired data, we will use McNemar's test to marginal homogeneity. We can represent this data as,

		Treatment R		
		D	L	
Treatment S	D	1	1	2
	L	8	10	18
		9	11	20

The test statistic is

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

```
TS <- c("L","L","L","L","L",
        "D","L","L","L","L",
        "D","L","L","L","L",
        "L","L","L","L","L")
TR <- c("L","D","L","L","L",
        "L","L","D","D","L",
        "D","D","L","L","D",
        "L","D","D","L","D")
data <- table(TS, TR)

# Calculate McNemar's Test Statistic and get p-value
M_TS <- (data[1,2] - data[2,1])^2/(data[1,2] + data[2,1])
pchisq(M_TS, df = 1, lower.tail = FALSE)
```

```
[1] 0.01963066
```

```
# Check with R function
mcnemar.test(data,correct=FALSE)
```

```
    McNemar's Chi-squared test

data:  data
McNemar's chi-squared = 5.4444, df = 1, p-value = 0.01963
```

For $\alpha = 0.05$ and a p-value of 0.020, we reject the null and conclude that that the probability of 5 year mortality is likely different for a simple mastectomy and radical mastectomy in the target population of women.

Problem 10

Suppose we are interested in comparing the effectiveness of 2 different antibiotics A and B in treating gonorrhea. We match each person receiving antibiotic A with an equivalent person (age within 5 years, same sex) to whom we give antibiotic B and we ask that these persons return to the clinic within 1 week to see if the gonorrhea has been eliminated.

Suppose the results are as follows:

For 40 pairs of people both antibiotics are successful.

For 20 pairs of people antibiotic A is effective while antibiotic B is not.

For 16 pairs of people antibiotic B is effective while antibiotic A is not.

For 3 pairs of people neither antibiotic is effective.

Perform an analysis to compare the relative effectiveness of the two antibiotics. Interpret your results.

Let S denote successful and NS denote not successful. Our data is as follows:

		Antibiotic B		
		S	NS	
Antibiotic A	S	40	20	60
	NS	16	3	19
		56	23	79

We want to test whether the efficacy of antibody A is equal to the efficacy of antibody B. Since the data is paired, we can us McNemar's test.

The test statistic is

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

which follows a $\chi^2(1)$ distribution.

```
# Manually perform McNemar's test
data <- matrix(c(40,20,16,3),ncol=2,byrow=TRUE)
M_TS <- (data[1,2] - data[2,1])^2/(data[1,2] + data[2,1])
pchisq(M_TS, df = 1, lower.tail = FALSE)
```

```
[1] 0.5049851
```

```
# Check with R function
mcnemar.test(data,correct=FALSE)
```

```
    McNemar's Chi-squared test

data:  data
McNemar's chi-squared = 0.44444, df = 1, p-value = 0.505
```

For $\alpha = 0.05$ and a p-value of 0.505, we fail to reject the null and conclude that the effectiveness of both antibiotics may be the same.

Problem 11

Consider a retrospective study with matched pairs. Show that McNemar's test statistic is equivalent to performing a Mantel-Haenszel test for all 2×2 tables (with one table for each pair).

Suppose we have K matched pairs. We can represent our data in the following 2×2 table:

		Cases		
		Exposed	Not Exposed	
Controls	Exposed	N_{11}	N_{12}	N_{1+}
	Not Exposed	N_{21}	N_{22}	N_{2+}
		N_{+1}	N_{+2}	$N_{++} = K$

Alternatively, we can express this data using K 2×2 tables with one table for each pair. Each table can be one of the following four tables:

	Case	Control		Case	Control		Case	Control		Case	Control
E	1	1	E	1	0	E	0	1	E	0	0
NE	0	0	NE	0	1	NE	1	0	NE	1	1

for the kth table, the entries are denoted as follows:

	Case	Control	
Exposed	n_{11k}	n_{12k}	n_{1+k}
Not Exposed	n_{21k}	n_{22k}	n_{2+}
	n_{+1k}	n_{+2k}	n_{++k}

A few notes about these K 2×2 tables:

1. In a retrospective case-control study, each of the K tables has one case and one control, hence the column totals are 1, i.e. $n_{+1k} = n_{+2k} = 1$ and $n_{++k} = 2$.
2. $n_{1+k}n_{2+k} = 1$ if and only if (1) the case is exposed, but the control is not exposed, or (2) the control is exposed, but the case is not exposed. Otherwise, $n_{1+k}n_{2+k} = 1$.

Additionally, recall from lecture 23, slide 17 that,

3. $E(n_{11k}) = \frac{n_{1+k}n_{+1k}}{n_{++k}}$
4. $Var(n_{11k}) = n_{1+k}n_{2+k}n_{+1k}n_{+2k}/n_{++k}^2(n_{++k} - 1)$

Suppose we want to test the null hypothesis that the sample odds ratio for all K pairs is 1. We want to show that the CMH test statistic is equivalent to McNemar's test statistic. The CMH test statistic is given by,

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

And McNemar's test statistic is,

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

Now observe that,

$$\begin{aligned}
\sum_k \{n_{11k} - E(n_{11k})\} &= \sum_k \left\{ n_{11k} - \frac{n_{1+k}n_{+1k}}{n_{++k}} \right\} && \text{by (3)} \\
&= \sum_k (n_{11k} - n_{1+k}/2) && \text{by (1)} \\
&= \sum_k \left(n_{11k} - \frac{n_{11k} + n_{12k}}{2} \right) \\
&= \sum_k \frac{n_{11k} - n_{12k}}{2} \\
&\overset{(*)}{=} \frac{(N_{11} + N_{21}) - (N_{11} - N_{12})}{2} \\
&= \frac{N_{21} - N_{12}}{2}
\end{aligned}$$

where $(*)$ follows from noting that $\sum_k n_{11k}$ is the total number of cases that are exposed which is given by $N_{11} + N_{21}$. Similarly, $\sum_k n_{12k}$ is the total number of controls that are exposed which is given by $N_{11} + N_{12}$. Additionally,

$$\begin{aligned}
\sum_k Var(n_{11k}) &= \sum_k n_{1+k}n_{2+k}n_{+1k}n_{+2k}/n_{++k}^2(n_{++k} - 1) && \text{by (4)} \\
&= \sum_k \frac{n_{1+k}n_{2+k}}{4} && \text{by (1)} \\
&= \frac{N_{12} + N_{21}}{4} && \text{by (2)}
\end{aligned}$$

Combining the two equation above, we find that the CMH test statistic is equal to McNemar's test statistic:

$$\frac{[\sum_k \{n_{11k} - E(n_{11k})\}]^2}{\sum_k \text{Var}(n_{11k})} = \frac{[(N_{21} - N_{12})/2]^2}{(N_{12} + N_{21})/4} = \frac{(N_{12} - N_{21})^2}{N_{12} + N_{21}}$$

Problem 12

A researcher is studying migration patterns. She collected the location of the current and previous homes for subjects who moved across regions. She recorded the following:

	Previous home		
Current home	Northeast	Southeast	West
Northeast	-	267	255
Southeast	135	-	139
West	240	234	-

Here the diagonals are not included since she only studied subjects who moved between regions. She would like to know if the probability of moving from region a to b is the same as the probability of moving from region b to a for all regions a and b.

a. Mathematically state her null and alternative hypotheses defining any notation you use.

Assume individuals in this population were randomly sampled. We are interested in testing a null hypothesis of symmetry. More precisely,

$$\begin{aligned}
H_0 &: \pi_{ij} = \pi_{ji}, \qquad i, j \in \{\text{NE, SE, W}\} \\
H_A &: \text{At least one } \pi_{ij} \neq \pi_{ji}
\end{aligned}$$

b. Calculate the expected counts under the null hypothesis.

Under the null hypothesis, the expected counts for both n_{12} and n_{21} is $(n_{12} + n_{21})/2$. Thus,

$$\begin{aligned}
E[n_{NE,SE}] = E[n_{SE,NE}] &= \frac{135 + 267}{2} \\
E[n_{NE,W}] = E[n_{W,NE}] &= \frac{255 + 240}{2} \\
E[n_{SE,W}] = E[n_{W,SE}] &= \frac{139 + 234}{2}
\end{aligned}$$

```
E_NESE = E_SENE = (135 + 267)/2
E_NEW = E_WNE = (255 + 240)/2
E_SEW = E_WSE = (149 + 234)/2
E <- c(E_NESE, E_SENE, E_NEW, E_WNE, E_SEW, E_WSE)
```

c. Perform the Chi-squared test and state your conclusions in the language of the problem. (Hint the df is 3.)

Recall that χ^2 statistic is given by,

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

```
O <- c(267, 135, 255, 240, 139, 234)

# Calculate chi stat and get p-value
chi_stat <- sum((O - E)^2/E)
pchisq(chi_stat, df = 3, lower.tail = FALSE)
```

```
[1] 1.377678e-14
```

For $\alpha = 0.05$ and a p-value of 1.38×10^{-14}, we reject the null and conclude that the probability of moving from each of these regions is not symmetric across the three regions.