Lab 008 - Contingency Tables and Difference of two Proportions Solutions

EPIB607 - Inferential Statistics[a]

[a]Fall 2020, McGill University

This version was compiled on November 17, 2020

1. Inference for the difference of two proportions

1. The use of screening mammograms for breast cancer has been controversial for decades because the overall benefit on breast cancer mortality is uncertain. A 30-year study to investigate the effectiveness of mammograms versus a standard non-mammogram breast cancer exam was conducted in Canada with 89,835 female participants.[1] Each woman was randomized to receive either annual mammograms or standard physical exams for breast cancer over a 5-year screening period.

   By the end of the 25 year follow-up period, 1,005 women died from breast cancer. The results are summarized in the following table.[2]

	Death from breast cancer?	
	Yes	No
Mammogram Group	500	44,425
Control Group	505	44,405

   a) Calculate $\hat{p}_1$ and $\hat{p}_2$, the two sample proportions of interest.

      The two sample proportions of interest are the proportion of breast cancer deaths in the mammogram group and the proportion of breast cancer deaths in the control group: $\hat{p}_M = 500/(500+44425) = 0.0111$, $\hat{p}_C = 505/(505+44405) = 0.0112$.

```
#use r as a calculator
x = c(500, 505)
n = c(500 + 44425, 505 + 44405)
p.hat.vector = x/n
p.hat.vector
```

```
#  [1] 0.01112966 0.01124471
```

[1] Miller AB. 2014. *Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. BMJ* 348 (2014): g366.

[2] During the 25 years following the screening period, each woman was screened for breast cancer according to the standard of care at her health care center.

b) Analyze the results; do the data suggest that annual mammography results in a reduction in breast cancer mortality relative to standard exams? Be sure to check the assumptions for using the normal approximation.

Since the participants were randomly assigned to each group, the groups can be treated as independent, and it is reasonable to assume independence of patients within each group.

The pooled proportion $\hat{p}$ is

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = 0.0112$$

The success-failure condition is met; the expected number of successes is about 503 in both groups (and expected number of failures is naturally much larger, given that both $\hat{p}$ is less than 0.50 and n is very large).

Test $H_0 : p_M = p_C$ against $H_A : p_M \neq p_C$. Let $\alpha = 0.05$.

The two-sided p-value is 0.895. Since $p > \alpha$, there is insufficient evidence to reject the null hypothesis; the observed difference in proportions of breast cancer deaths is reasonably explained by chance. The results do not suggest that annual mammography results in a reduction in breast cancer mortality relative to standard exams.

```
#use r as a calculator
p.hat.pooled = sum(x)/sum(n)
p.hat.pooled
```

```
#  [1] 0.01118718
```

```
#check success-failure
n*p.hat.pooled
```

```
#  [1] 502.5839 502.4161
```

```
n*(1 - p.hat.pooled)
```

```
#  [1] 44422.42 44407.58
```

```
#conduct inference
prop.test(x = x, n = n)$p.val
```

```
#  [1] 0.8948174
```

c) Calculate and interpret a 95% confidence interval for the difference in proportions of deaths from breast cancer. Be sure to check the assumptions for using the normal approximation.

The success-failure condition should be checked for each sample: from the data, the number of successes and failures are both well over 10 in each group.

The 95% confidence interval is (-0.0015, 0.0013). With 95% confidence, the difference in probability of death is within the interval (-0.15%, 0.13%); i.e., 0.15% lower in the mammogram group to 0.13% higher in the mammogram group. As expected from the large p-value, the confidence interval contains the null value 0.

```
#conduct inference
prop.test(x = x, n = n)$conf.int
```

```
#  [1] -0.001512853  0.001282751
#  attr(,"conf.level")
#  [1] 0.95
```

2. Remdesivir is an antiviral drug previously tested in animal models infected with coronaviruses like SARS and MERS. As of May 2020, remdesivir had temporary approval from the FDA for use in severely ill COVID-19 patients and was the subject of numerous ongoing studies.

   A randomized controlled trial conducted in China enrolled 236 patients with severe COVID-19; 158 were assigned to receive remdesivir and 78 to receive a placebo. In the remdesivir group, 103 patients showed clinical improvement; in the placebo group, 45 patients showed clinical improvement.[3]

   a) Calculate $\hat{p}_1$ and $\hat{p}_2$, the two sample proportions of interest.

      The two sample proportions of interest are 0.652 and 0.577, the proportion of individuals in each treatment group that showed clinical improvement.

```r
#use r as a calculator
x = c(103, 45)
n = c(158, 78)
p.hat.vector = x/n
p.hat.vector
```

```
#  [1] 0.6518987 0.5769231
```

   b) Conduct a formal comparison of the clinical improvement rates and summarize your findings. Be sure to check the assumptions for using the normal approximation.

      Since the participants were randomly assigned to each group, the groups can be treated as independent, and it is reasonable to assume independence of patients within each group.

      The pooled proportion $\hat{p}$ is

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = 0.627$$

      The success-failure condition is met; the expected number of successes and failures are all larger than 10.

      Test $H_0 : p_1 = p_2$ against $H_A : p_1 \neq p_2$, where p_1 represents the population proportion of clinical improvement in COVID-19 patients treated with remdesivir and p_2 represents the population proportion of clinical improvement in COVID-19 patients treated with placebo. Let $\alpha = 0.05$. The p-value is 0.328, which is greater than α; there is insufficient evidence to reject the null hypothesis of no difference. Even though the proportion of patients who experienced clinical improvement about 7% higher in the remdesivir group, this difference is not extreme enough to represent sufficient evidence that remdesivir is more effective than placebo.

[3] Wang, Y, et al. Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multi-centre trial. *Lancet* 395(10236). 16 May 2020.

```r
#use r as a calculator
p.hat.pooled = sum(x)/sum(n)
p.hat.pooled
```

```
#  [1] 0.6271186
```

```r
#check success-failure
n*p.hat.pooled
```

```
#  [1] 99.08475 48.91525
```

```r
n*(1 - p.hat.pooled)
```

```
#  [1] 58.91525 29.08475
```

```r
#conduct inference
prop.test(x = x, n = n)$p.val
```

```
#  [1] 0.3284038
```

c) Report and interpret an appropriate interval estimate. Be sure to check the assumptions for using the normal approximation.

The success-failure condition should be checked for each sample: from the data, the number of successes and failures are both well over 10 in each group. The 95% confidence interval is (-0.067, 0.217); with 95% confidence, this interval captures the difference in population proportion of clinical mortality between COVID-19 patients treated with remdesivir and those treated with placebo. The interval contains 0, which is consistent with no statistically significant evidence of a difference. The interval reflects the lack of precision around the effect estimate that is characteristic of an insufficiently large sample size.

```r
prop.test(x = x, n = n)$conf.int
```

```
#  [1] -0.06703113  0.21698245
#  attr(,"conf.level")
#  [1] 0.95
```

2. Contingency Tables

The χ^2 test of independence

In the χ^2 test of independence, the observed number of cell counts are compared to the number of **expected** cell counts, where the expected counts are calculated under the null hypothesis.

- H_0: the row and column variables are not associated
- H_A: the row and column variables are associated

The expected count for the i^{th} row and j^{th} column is

$$E_{i,j} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{n},$$

where n is the total number of observations.

Assumptions for the χ^2 test:

- *Independence*. Each case that contributes a count to the table must be independent of all other cases in the table.
- *Sample size*. Each expected cell count must be greater than or equal to 10. For tables larger than 2×2, it is appropriate to use the test if no more than 1/5 of the expected counts are less than 5, and all expected counts are greater than 1.

The χ^2 **test statistic** is calculated as

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}},$$

and is approximately distributed χ^2 with degrees of freedom $(r-1)(c-1)$, where r is the number of rows and c is the number of columns. $O_{i,j}$ represents the observed count in row i, column j.

For each cell in a table, the **residual** equals

$$\frac{O_{i,j} - E_{i,j}}{\sqrt{E_{i,j}}}.$$

Residuals with a large magnitude contribute the most to the χ^2 statistic. If a residual is positive, the observed value is greater than the expected value; if a residual is negative, the observed value is less than the expected.

Fisher's exact test

When the expected counts in a two-way table are less than 10, Fisher's exact test is used to compute a p-value without relying on the normal approximation. In this course, only the logic behind Fisher's exact test for a 2×2 table is discussed. In the 2×2 table case, the hypotheses for Fisher's exact test can be expressed in the same way as for a two-sample test of proportions; the null hypothesis is $H_0 : p_1 = p_2$.

The p-value is the probability of observing results as or more extreme than those observed under the assumption that the null hypothesis is true.

- Thus, the p-value is calculated by adding together the individual conditional probabilities of obtaining each table that is as or more extreme than the one observed, under the null hypothesis and given that the marginal totals are considered fixed.
- When the marginal totals are held constant, the value of any one cell in the table determines the rest of entries. When marginal totals are considered fixed, each table represents a unique set of results.
- Extreme tables are those which contradict $H_0 : p_1 = p_2$.
- A two-sided p-value can be calculated by doubling the smaller of the possible one-sided p-values; this method is typically used when calculating p-values by hand. Another common method is to classify "more extreme" tables as all tables with probabilities less than that of the observed table, in both directions; the p-value is the sum of probabilities for the qualifying table.

The probability of a particular table (i.e., set of results) can be calculated with the **hypergeometric distribution**.

Let X represent the number of successes in a series of repeated Bernoulli trials, where sampling is done without replacement. Suppose that in the population of size N, there are m total successes. What is the probability of observing exactly k successes when drawing a sample of size n?

For example, imagine an urn with m white balls and $N-m$ black balls (thus, there are N total balls). Draw n balls without replacement (i.e., a sample of n balls). What is the probability of observing k white balls in the sample?

The possible results of a sample can be organized in a 2×2 table:

	White Ball	Black Ball	Total
Sampled	k	$n-k$	n
Not Sampled	$m-k$	$N-n-(m-k)$	$N-n$
Total	m	$N-m$	N

The probability of observing exactly k successses in a sample of size n (i.e., n dependent trials) is given by

$$P(X=k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}.$$

Hypergeometric probabilities are calculated in R with the use of `dhyper()` and `phyper()`. The following code shows how to calculate $P(X=5)$, $P(X \leq 5)$, and $P(X > 5)$ for $X \sim \text{HGeom}(10, 15, 8)$, where $m = 10$, $N-m = 15$, and $n = 8$.

```
#probability X equals 5
dhyper(5, 10, 15, 8)
```

```
#  [1] 0.1060121
```

```
#probability X is less than or equal to 5
phyper(5, 10, 15, 8)
```

```
#  [1] 0.9779072
```

```
#probability X is greater than 5
phyper(5, 10, 15, 8, lower.tail = FALSE)
```

```
#  [1] 0.02209278
```

Measures of association in two-by-two tables

Chapter 1 introduced the **relative risk (RR)**, a measure of the risk of a certain event occurring in one group relative to the risk of the event occurring in another group, as a numerical summary for two-by-two (2×2) tables. The relative risk can also be thought of as a measure of association.

Consider the following hypothetical two-by-two table. The relative risk of Outcome A can be calculated by using either Group 1 or Group 2 as the reference group:

	Outcome A	Outcome B	Sum
Group 1	a	b	$a+b$
Group 2	c	d	$c+d$
Sum	$a+c$	$b+d$	$a+b+c+d=n$

Table 1. A hypothetical two-by-two table of outcome by group.

$$RR_{\text{A, comparing Group 1 to Group 2}} = \frac{a/(a+b)}{c/(c+d)}$$

$$RR_{\text{A, comparing Group 2 to Group 1}} = \frac{c/(c+d)}{a/(a+b)}$$

The relative risk is only valid for tables where the proportions $a/(a+b)$ and $c/(c+d)$ represent the incidence of Outcome A within the populations from which Groups 1 and 2 are sampled.

The **odds ratio (OR)** is a measure of association that remains applicable even when it is not possible to estimate incidence of an outcome from the sample data. The **odds** of Outcome A in Group 1 are a/b, while the odds of Outcome A in Group 2 are c/d.

$$OR_{\text{A, comparing Group 1 to Group 2}} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

$$OR_{\text{A, comparing Group 2 to Group 1}} = \frac{c/d}{a/b} = \frac{bc}{ad}$$

3. The χ^2 test of independence

3. In resource-limited settings, single-dose nevirapine (NVP) is given to an HIV-positive woman during birth to prevent mother-to-child transmission of the virus. Exposure of the infant to NVP may foster the growth of more virulent strains of the virus in the child.

   If a child is HIV-positive, should they be treated with NVP or a more expensive drug, lopinavir (LPV)? In this setting, success means preventing a growth of the virus in the child (i.e., preventing

virologic failure). The following table contains data from a 2012 study conducted in six African countries and India.[4]

	NVP	LPV	Total
Virologic Failure	60	27	87
Stable Disease	87	113	200
Total	147	140	287

a) State the null and alternative hypotheses.

The null hypothesis is that there is no association between treatment and outcome; i.e., treatment and outcome are independent.

The alternative hypothesis is that there is an association between treatment and outcome; i.e., treatment and outcome are not independent.

b) Calculate the expected cell counts.

The expected cell counts are shown in parentheses next to the observed cell counts.

	NVP	LPV	Total
Virologic Failure	60 (44.56)	27 (42.44)	87
Stable Disease	87 (102.44)	113 (97.56)	200
Total	147	140	287

```
#use r as a calculator

#set parameters
n = 287
row.1.total = 87
row.2.total = 200
col.1.total = 147
col.2.total = 140

#calculate expected values
exp.1.1 = (row.1.total * col.1.total)/n
exp.1.1
```

```
  #  [1] 44.56098
```

```
exp.1.2 = (row.1.total * col.2.total)/n
exp.1.2
```

[4] A. Violari, et al. "Nevirapine versus ritonavir-boosted lopinavir for HIV-infected children." *NEJM* 366: 2380-2389.

```
  #  [1] 42.43902
```

```
exp.2.1 = (row.2.total * col.1.total)/n
exp.2.1
```

```
  #  [1] 102.439
```

```
exp.2.2 = (row.2.total * col.2.total)/n
exp.2.2
```

```
  #  [1] 97.56098
```

c) Check the assumptions for using the χ^2 test.

Independence holds, since this is a randomized study. All expected counts are greater than 10.

d) Calculate the χ^2 test statistic.

$$\chi^2 = \sum \frac{(\text{obs - exp})^2}{\text{exp}} = \frac{(60-44.56)^2}{44.56} + \frac{(27-42.44)^2}{42.44} + \frac{(87-102.44)^2}{102.44} + \frac{(113-97.56)^2}{97.56} = 15.74$$

```
#use r as a calculator
obs.1.1 = 60
chi.sq.1.1 = ((obs.1.1 - exp.1.1)^2)/exp.1.1

obs.1.2 = 27
chi.sq.1.2 = ((obs.1.2 - exp.1.2)^2)/exp.1.2

obs.2.1 = 87
chi.sq.2.1 = ((obs.2.1 - exp.2.1)^2)/exp.2.1

obs.2.2 = 113
chi.sq.2.2 = ((obs.2.2 - exp.2.2)^2)/exp.2.2

chi.sq = chi.sq.1.1 + chi.sq.1.2 + chi.sq.2.1 + chi.sq.2.2
chi.sq
```

```
#  [1] 15.73587
```

e) Calculate the p-value for the test statistic using `pchisq()`. The p-value represents the probability of observing a result as or more extreme than the sample data.

The p-value for the test statistic is 7.28×10^{-5}.

```
#use pchisq()
pchisq(chi.sq, df = (2 - 1)*(2 - 1), lower.tail = FALSE)
```

f) Confirm the results from parts c) and d) using `chisq.test()`. Note that the value of the test statistic will be slightly different because R is applying a 'continuity correction'.

From `chisq.test()`, the χ^2 statistic is 14.73 and the associated p-value is 0.0001.

```
#enter the data as a table
hiv.table = matrix(c(60, 27, 87, 113),
              nrow = 2, ncol = 2, byrow = T)

#add labels and confirm the table was entered correctly
dimnames(hiv.table) = list("Outcome" = c("Virologic Failure", "Stable Disease"),
                           "Drug" = c("NVP", "LPV"))
hiv.table
```

```
#                     Drug
#  Outcome             NVP LPV
#    Virologic Failure  60  27
#    Stable Disease     87 113
```

```
#use chisq.test()
chisq.test(hiv.table)
```

```
#
#   Pearson's Chi-squared test with Yates' continuity correction
#
#  data:  hiv.table
#  X-squared = 14.733, df = 1, p-value = 0.0001238
```

g) Summarize the conclusions; be sure to include which drug is recommended for treatment, based on the data.

There is sufficient evidence at the $\alpha = 0.05$ significance level to reject the null hypothesis and accept the alternative hypothesis that treatment and outcome are associated.

From comparing the expected and observed cell counts (or looking at the residuals), it is possible to determine the direction of the association. When treated with lopinarvir, fewer children than expected experience virologic failure (27 observed versus ~42 expected), and more than expected experience stable disease (113 observed versus ~98 expected). In contrast, when treated with nevirapine, more children than expected experience virologic failure (60 observed versus ~45 expected), and fewer children than expected experience stable disease (87 observed versus ~102 expected).

The data suggest that HIV-positive children should be treated with lopinarvir.

```
#look at residuals
chisq.test(hiv.table)$resid
```

```
#                     Drug
#  Outcome                   NVP       LPV
#    Virologic Failure  2.312824 -2.369939
#    Stable Disease    -1.525412  1.563082
```

```
#to view the expected values, use $expected
chisq.test(hiv.table)$expected
```

```
#                     Drug
#  Outcome                   NVP      LPV
#    Virologic Failure  44.56098 42.43902
#    Stable Disease    102.43902 97.56098
```

h) Repeat the analysis using inference for the difference of two proportions and confirm that the results are the same.

The proportion of successes on nevirapine is 0.59 and the proportion of successes on lopinarvir is 0.81. The p-value is 0.0012; there is sufficient evidence to reject the null of no difference and conclude that stable disease is associated with lopinarvir.

```
#use prop.test( )
prop.test(x = c(87, 113), n = c(147, 140))
```

```
#
#   2-sample test for equality of proportions with continuity correction
#
#  data:  c(87, 113) out of c(147, 140)
#  X-squared = 14.733, df = 1, p-value = 0.0001238
#  alternative hypothesis: two.sided
#  95 percent confidence interval:
#   -0.3251572 -0.1054551
#  sample estimates:
#     prop 1    prop 2
#  0.5918367 0.8071429
```

4. In the PREVEND study introduced in Chapter 6, researchers measured various features of study participants, including data on statin use and highest level of education attained. From the data in `prevend.samp`, is there evidence of an association between statin use and educational level? Summarize the results.

   Test the null hypothesis that statin use and education level are not associated against the alternative hypothesis that statin use and education level are associated. Let $\alpha = 0.05$.

   The p-value of the χ^2 statistic is 0.0003. The results are highly significant, and there is evidence to support accepting the alternative hypothesis that statin use and education level are associated.

   The largest deviations from independence occur in the primary school group and university group. There are more statin users than expected in the primary school group and fewer statin users than expected in the university group. There is an observable overall trend; as highest educational level attained increases, the proportion of statin users goes from higher than expected to lower than expected.

```r
#load the data
library(oibiostat)
data("prevend.samp")

#convert variables to factors
prevend.samp$Statin = factor(prevend.samp$Statin, levels = c(0, 1),
                         labels = c("NonUser", "User"))

prevend.samp$Education = factor(prevend.samp$Education, levels = 0:3,
                            labels = c("Primary", "LowerSec",
                                       "UpperSec", "Univ"))

#create a table
statin.edu.table = table(prevend.samp$Statin, prevend.samp$Education)

#run chi-squared test
chisq.test(statin.edu.table)
```

```
#
#   Pearson's Chi-squared test
#
#  data:  statin.edu.table
#  X-squared = 19.054, df = 3, p-value = 0.0002665
```

```
chisq.test(statin.edu.table)$residuals
```

```
#
#               Primary   LowerSec   UpperSec       Univ
#    NonUser -1.3196995 -0.8994999  0.3760673  1.3000955
#    User     2.4146629  1.6458208 -0.6880929 -2.3787932
```

4. Fisher's exact test

5. *Clostridium difficile* is a bacterium that causes inflammation of the colon. Antibiotic treatment is typically not effective, particularly for patients who experience multiple recurrences of infection. Infusion of feces from healthy donors has been reported as an effective treatment for recurrent infection. A randomized trial was conducted to compare the efficacy of donor-feces infusion versus vancomycin, the antibiotic typically prescribed to treat *C. difficile* infection. The results of the trial are shown in the following table.

	Cured	Uncured	Sum
Fecal Infusion	13	3	16
Vancomycin	4	9	13
Sum	17	12	29

a) Can a χ^2 test be used to analyze these results?

A χ^2 test is not advisable since there are expected counts less than 10.

```
infusion.table = matrix(c(13, 3, 4, 9), nrow = 2, ncol = 2, byrow = T)
dimnames(infusion.table) = list("Outcome" = c("Cured", "Uncured"),
                                "Treatment" = c("Fecal Infusion",
                                                "Vancomycin"))

chisq.test(infusion.table)$expected
```

```
#           Treatment
#  Outcome   Fecal Infusion Vancomycin
#    Cured          9.37931    6.62069
#    Uncured        7.62069    5.37931
```

b) Researchers are interested in understanding whether fecal infusion is a more effective treatment than vancomycin. Write the null hypothesis and appropriate one-sided alternative hypothesis.

Let p_1 represent the population proportion of individuals cured on the fecal infusion treatment and p_2 represent the population proportion of individuals cured on the vancomycin treatment. Under the null hypothesis, the proportion cured is equal between the two treatment groups: $H_0 : p_1 = p_2$. The appropriate alternative hypothesis of interest is $H_0 : p_1 > p_2$.

c) Under the assumption that the marginal totals are fixed, enumerate all possible sets of results that are more extreme than what was observed, in the same direction.

Results more extreme than what was observed are those that constitute stronger evidence in favor of the fecal treatment group being more effective than vancomyin; i.e., results where either 14, 15, or 16 of the cured patients were in the fecal infusion group.

	Cured	Uncured	Sum
Fecal Infusion	14	2	16
Vancomycin	3	10	13
Sum	17	12	29

	Cured	Uncured	Sum
Fecal Infusion	15	1	16
Vancomycin	2	11	13
Sum	17	12	29

d) Calculate the probability of the observed results.

Use the hypergeometric distribution with parameters $N = 29$, $m = 17$, and $n = 16$; calculate $P(X = 13)$. Consider the "successes" to be the individuals cured and the "sample size" to be the number of individuals in the fecal infusion group. The probability that 13 of the cured individuals were in the fecal infusion group, given the table margins are fixed, is 0.0077.

```
#probability of the observed results
dhyper(13, 17, 29 - 17, 16)
```

```
#  [1] 0.007715441
```

e) Calculate the probability of each set of results enumerated in part c).

The probabilities of observing 14, 15, or 16 cured individuals in the fecal infusion group, respectively, are 6.61×10^{-4}, 2.41×10^{-5}, and 2.51×10^{-7}.

```
#probability of results more extreme than observed
dhyper(14, 17, 29 - 17, 16)
```

```
#  [1] 0.0006613235
```

```
dhyper(15, 17, 29 - 17, 16)
```

```
#  [1] 2.404813e-05
```

```
dhyper(16, 17, 29 - 17, 16)
```

	Cured	Uncured	Sum
Fecal Infusion	16	0	16
Vancomycin	1	12	13
Sum	17	12	29

```
#   [1] 2.505013e-07
```

f) Based on the answers in parts d) and e), compute the one-sided p-value and interpret the results.

The probability of observing results as or more extreme than the observed table is 0.0084. Since $p < 0.05$, there is sufficient evidence to reject the null hypothesis at significance level $\alpha = 0.05$; the data support fecal infusion as a more effective treatment for *C. difficile* infection than vancomycin.

```
#summing previous probabilities
p.observed = dhyper(13, 17, 29 - 17, 16)
p.more.extreme = dhyper(14, 17, 29 - 17, 16) + dhyper(15, 17, 29 - 17, 16) +
  dhyper(16, 17, 29 - 17, 16)
p.observed + p.more.extreme
```

```
#   [1] 0.008401063
```

```
#using phyper
phyper(12, 17, 29 - 17, 16, lower.tail = FALSE)
```

```
#   [1] 0.008401063
```

g) Use `fisher.test( )` to confirm the calculations in part f) and to calculate the two-sided p-value.

The two-sided p-value is 0.0095.

```
#one-sided p-value
fisher.test(infusion.table, alternative = "greater")$p.val
```

```
#   [1] 0.008401063
```

```r
#two-sided p-value
fisher.test(infusion.table, alternative = "two.sided")$p.val
```

```
#  [1] 0.009530323
```

6. Psychologists conducted an experiment to investigate the effect of anxiety on a person's desire to be alone or in the company of others (Schacter 1959; Lehmann 1975). A group of 30 individuals were randomly assigned into two groups; one group was designated the "high anxiety" group and the other the "low anxiety" group. Those in the high-anxiety group were told that in the "upcoming experiment", they would be subjected to painful electric shocks, while those in the low-anxiety group were told that the shocks would be mild and painless.[5]. All individuals were informed that there would be a 10 minute wait before the experiment began, and that they could choose whether to wait alone or with other participants.

   The following table summarizes the results:

	Wait Together	Wait Alone	Sum
High-Anxiety	12	5	17
Low-Anxiety	4	9	13
Sum	16	14	30

   a) Under the null hypothesis of no association, what are the expected cell counts?

      Under the null hypothesis of no association, the expected cell counts are 9.07 and 7.93 in the wait together and wait alone groups, respectively, for those considered "high anxiety" and 6.93 and 6.07 in the wait together and wait alone groups, respectively, for those considered "low anxiety".

```r
#enter the data
anxiety.table = matrix(c(12, 5, 4, 9), nrow = 2, ncol = 2, byrow = T)
dimnames(anxiety.table) = list("Treatment" = c("High Anxiety", "Low Anxiety"),
                               "Outcome" = c("Wait Together", "Wait Alone"))

chisq.test(anxiety.table)$expected
```

```
#                Outcome
#  Treatment      Wait Together Wait Alone
#    High Anxiety      9.066667   7.933333
#    Low Anxiety       6.933333   6.066667
```

[5] Individuals were not actually subjected to electric shocks of any kind

b) Under the assumption that the marginal totals are fixed and the null hypothesis is true, what is the probability of the observed set of results?

Use the hypergeometric distribution with parameters $N = 30$, $m = 16$, and $n = 17$; calculate $P(X = 12)$. Consider the "successes" to be the individuals who wait together, and the "number sampled" to be the people randomized to the high-anxiety group. The probability of the observed set of results, assuming the marginal totals are fixed and the null hypothesis is true, is 0.0304.

```
dhyper(12, 16, 30 - 16, 17)
```

```
#  [1] 0.03042455
```

c) Enumerate the tables that are more extreme than what was observed, in the same direction.

More individuals than expected in the high-anxiety group were observed to wait together; thus, tables that are more extreme in the same direction also consist of those where more people in the high-anxiety group wait together than observed. These are tables in which 13, 14, 15, or 16 individuals in the high-anxiety group wait together.

	Wait Together	Wait Alone	Sum
High-Anxiety	13	4	17
Low-Anxiety	3	10	13
Sum	16	14	30

	Wait Together	Wait Alone	Sum
High-Anxiety	14	3	17
Low-Anxiety	2	11	13
Sum	16	14	30

	Wait Together	Wait Alone	Sum
High-Anxiety	15	2	17
Low-Anxiety	1	12	13
Sum	16	14	30

	Wait Together	Wait Alone	Sum
High-Anxiety	16	1	17
Low-Anxiety	0	13	13
Sum	16	14	30

d) Conduct a formal test of association for the results and summarize your findings. Let $\alpha = 0.05$.

Let p_1 represent the population proportion of individuals waiting together in the high-anxiety group and p_2 represent the population proportion of individuals waiting together in the low-anxiety group. Test $H_0 : p_1 = p_2$ against $H_A : p_1 \neq p_2$. Let $\alpha = 0.05$. The two-sided p-value is 0.063. There is insufficient evidence to reject the null hypothesis; the data do not suggest there is an association between high anxiety and a person's desire to be in the company of others.

Note that the results are borderline; a one-sided p-value rejects the null hypothesis with $p = 0.036$. The choice of two-sided hypothesis is more impartial.

```
#two-sided test
fisher.test(anxiety.table)$p.val
```

```
#  [1] 0.06335838
```

```
#one-sided test
fisher.test(anxiety.table, alternative = "greater")$p.val
```

```
#  [1] 0.03548226
```

5. Measures of association in two-by-two tables

7. Suppose a study is conducted to assess the association between smoking and cardiovascular disease (CVD). Researchers recruited a group of 231 study participants then categorized them according to smoking and disease status: 111 are smokers, while 40 smokers and 32 non-smokers have CVD. Calculate and interpret the relative risk of CVD.

   The relative risk of CVD comparing smokers to non-smokers is 1.35. Smoking is associated with a 35% increase in the probability of CVD. In other words, the risk of CVD is 35% greater in smokers compared to non-smokers.

```
#use r as a calculator
risk.smokers = 40/111
risk.nonsmokers = 32/(231-111)

risk.smokers / risk.nonsmokers
```

```
#  [1] 1.351351
```

8. Suppose another study is conducted to assess the association between smoking and CVD, but researchers use a different design: 90 individuals with CVD and 110 individuals without CVD are recruited. 40 of the individuals with CVD are smokers, and 80 of the individuals without CVD are non-smokers.

   a) Is relative risk an appropriate measure of association for these data? Explain your answer.

      No, relative risk should not be calculated for these observations. Since the number of individuals with and without CVD is fixed by the study design, the proportion of individuals with CVD within a certain group (smokers or non-smokers) as calculated from the data is not a measure of CVD risk for that population.

   b) Calculate the odds of CVD among smokers and the odds of CVD among non-smokers.

      Since there are 110 individuals without CVD and 80 of those are non-smokers, there are 30 individuals without CVD who smoke. Thus, there are $30 + 40 = 70$ individuals who smoke, and $(110 + 90) - 70 = 130$ individuals who do smoke. Of the 130 non-smokers, 80 do not have CVD; thus, 50 non-smokers have CVD.

      The odds of CVD among smokers is the number of smokers with CVD divided by the number of smokers without CVD: $40/30 = 1.33$. The odds of CVD among non-smokers is the number of non-smokers with CVD divided by the number of non-smokers without CVD: $50/80 = 0.625$.

```
#use r as a calculator
odds.smokers = 40/30
odds.nonsmokers = 50/80
```

```
odds.smokers
```

```
#  [1] 1.333333
```

```
odds.nonsmokers
```

```
#  [1] 0.625
```

c) Calculate and interpret the odds ratio of CVD, comparing smokers to non-smokers.

The odds ratio of CVD, comparing smokers to non-smokers is 2.13. The odds of CVD in smokers are approximately twice as large as the odds of CVD in smokers. The data suggest that smoking is associated with CVD.

```
#use r as a calculator
odds.smokers / odds.nonsmokers
```

```
#  [1] 2.133333
```

d) What would an odds ratio of CVD (comparing smokers to non-smokers) equal to 1 represent, in terms of the association between smoking and CVD? What would an odds ratio of CVD less than 1 represent?

An odds ratio equal to 1 would represent no association between smoking and CVD. An odds ratio less than 1 would represent an association between not smoking and CVD; i.e., that the odds of CVD in non-smokers were higher than the odds of CVD in smokers.