Chapter 7

Categorical Data Analysis

Learning objectives

After completing this topic, you should be able to:

select the appropriate statistical method to compare summaries from categorical variables.
assess the assumptions of one-way and two-way tests of proportions and independence.
decide whether the proportions between populations are different, including in stratified and cross-sectional studies.
recommend action based on a hypothesis test.

Achieving these goals contributes to mastery in these course learning outcomes:

1. organize knowledge.
5. define parameters of interest and hypotheses in words and notation.
6. summarize data visually, numerically, and descriptively.
8. use statistical software.
12. make evidence-based decisions.

7.1 Categorical data

When the response variable is categorical, the interesting questions are often about the probability of one possible outcome versus another, and whether these probabilities depend on other variables (continuous or categorical).

Example: Titanic The sinking of the Titanic is a famous event, and new books are still being published about it. Many well-known facts — from the proportions of first-class passengers to the "women and children first" policy, and the fact that that policy was not entirely successful in saving the women and children in the third class — are reflected in the survival rates for various classes of passenger. The source provides a data set recording class, sex, age, and survival status for each person on board of the Titanic, and is based on data originally collected by the British Board of Trade[1].

```
# The Titanic dataset is a 4-dimensional table: Class, Sex, Age, Survived
library(datasets)
data(Titanic)

Titanic
```

```
## , , Age = Child, Survived = No
##
##       Sex
## Class  Male Female
##   1st     0      0
##   2nd     0      0
##   3rd    35     17
##   Crew    0      0
##
## , , Age = Adult, Survived = No
##
##       Sex
## Class  Male Female
##   1st   118      4
##   2nd   154     13
##   3rd   387     89
##   Crew  670      3
##
```

[1]British Board of Trade (1990), Report on the Loss of the "Titanic" (S.S.). British Board of Trade Inquiry Report (reprint). Gloucester, UK: Allan Sutton Publishing. Note that there is not complete agreement among primary sources as to the exact numbers on board, rescued, or lost.

```
## , , Age = Child, Survived = Yes
##
##       Sex
## Class  Male Female
##   1st     5      1
##   2nd    11     13
##   3rd    13     14
##   Crew    0      0
##
## , , Age = Adult, Survived = Yes
##
##       Sex
## Class  Male Female
##   1st    57    140
##   2nd    14     80
##   3rd    75     76
##   Crew  192     20
```

```
# reshape into long data.frame
library(reshape2)
df.titanic <- melt(Titanic, value.name = "Freq")
df.titanic
```

```
##    Class    Sex   Age Survived Freq
## 1    1st   Male Child       No    0
## 2    2nd   Male Child       No    0
## 3    3rd   Male Child       No   35
## 4   Crew   Male Child       No    0
## 5    1st Female Child       No    0
## 6    2nd Female Child       No    0
## 7    3rd Female Child       No   17
## 8   Crew Female Child       No    0
## 9    1st   Male Adult       No  118
## 10   2nd   Male Adult       No  154
## 11   3rd   Male Adult       No  387
## 12  Crew   Male Adult       No  670
## 13   1st Female Adult       No    4
## 14   2nd Female Adult       No   13
## 15   3rd Female Adult       No   89
## 16  Crew Female Adult       No    3
## 17   1st   Male Child      Yes    5
## 18   2nd   Male Child      Yes   11
## 19   3rd   Male Child      Yes   13
## 20  Crew   Male Child      Yes    0
## 21   1st Female Child      Yes    1
## 22   2nd Female Child      Yes   13
## 23   3rd Female Child      Yes   14
## 24  Crew Female Child      Yes    0
## 25   1st   Male Adult      Yes   57
## 26   2nd   Male Adult      Yes   14
```

```
## 27   3rd   Male Adult      Yes   75
## 28  Crew   Male Adult      Yes  192
## 29   1st Female Adult      Yes  140
## 30   2nd Female Adult      Yes   80
## 31   3rd Female Adult      Yes   76
## 32  Crew Female Adult      Yes   20
```

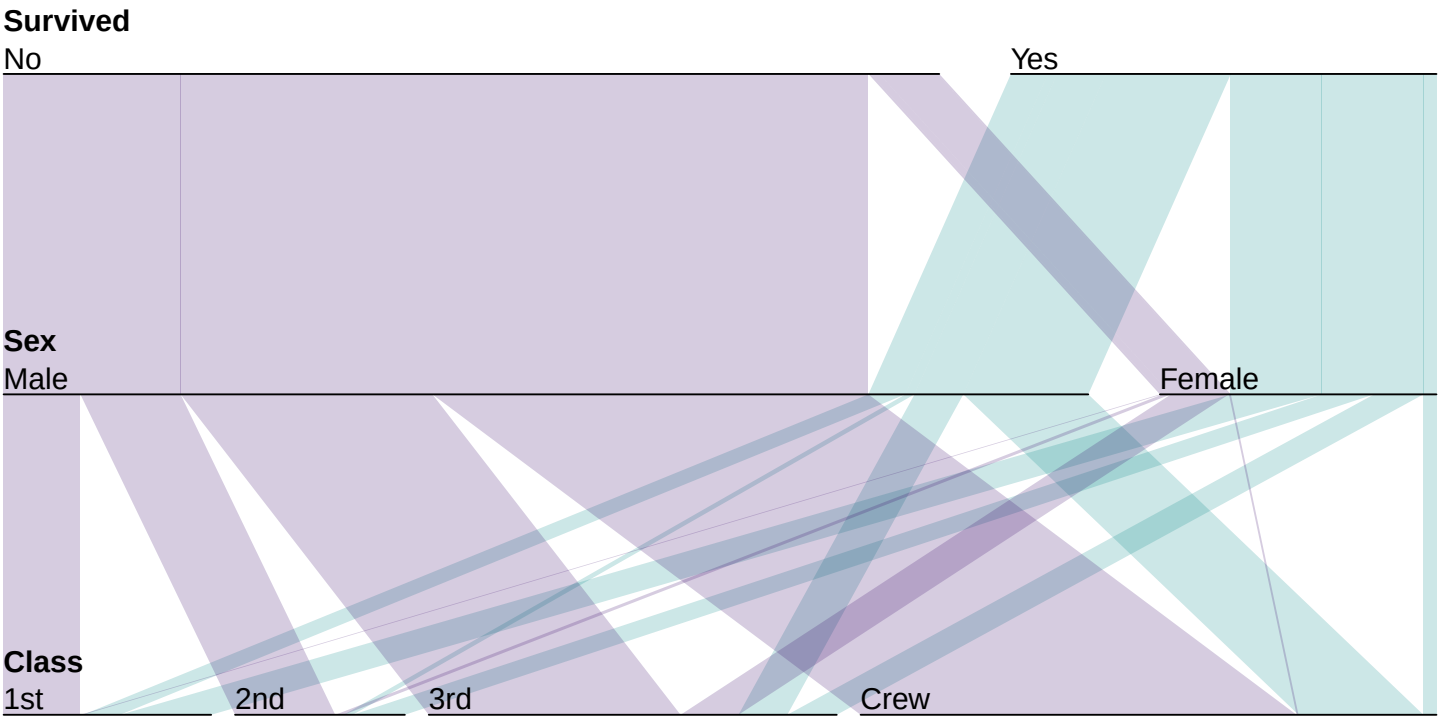
```
# Total number of people
sum(df.titanic$Freq)
```

```
## [1] 2201
```

```
# create colors based on survival
df.titanic$Color <- ifelse(df.titanic$Survived == "Yes", "#008888", "#330066")

# subset only the adults (since there were so few children)
df.titanic.adult <- subset(df.titanic, Age == "Adult")

# see R code on website for function parallelset()
# see help for with(), it allows temporary direct reference to columns in a data.frame
#   otherwise, we'd need to specify df.titanic.adult$Survived, ...
with(df.titanic.adult
   , parallelset(Survived, Sex, Class, freq = Freq, col = Color, alpha=0.2)
    )
```



There are many questions that can be asked of this dataset. How likely were people to survive such a ship sinking in cold water? Is the survival proportion dependent on sex, class, or age, or a combination of these? How different is the survival proportions for 1st class females versus 3rd class males?

7.2 Single Proportion Problems

Assume that you are interested in estimating the proportion p of individuals in a population with a certain characteristic or attribute based on a random or representative sample of size n from the population. The **sample proportion** $\hat{p}$ =(# with attribute in the sample)/n is the best guess for p based on the data.

This is the simplest **categorical data** problem. Each response falls into one of two exclusive and exhaustive categories, called "success" and "failure". Individuals with the attribute of interest are in the success category. The rest fall into the failure category. Knowledge of the **population proportion** p of successes characterizes the distribution across both categories because the population proportion of failures is $1 - p$.

As an aside, note that the probability that a randomly selected individual has the attribute of interest is the population proportion p with the attribute, so the terms population proportion and probability can be used interchangeably with random sampling.

7.2.1 A CI for p

A two-sided CI for p is a range of plausible values for the unknown population proportion p, based on the observed data. To compute a two-sided CI for p:

1. Specify the confidence level as the percent $100(1-\alpha)\%$ and solve for the error rate α of the CI.
2. Compute $z_{\text{crit}} = z_{0.5\alpha}$ (i.e., area under the standard normal curve to the left and to the right of z_{crit} are $1 - 0.5\alpha$ and 0.5α, respectively). `qnorm(1-0.05/2)=1.96`.
3. The $100(1-\alpha)\%$ CI for p has endpoints $L = \hat{p} - z_{\text{crit}}SE$ and $U = \hat{p} + z_{\text{crit}}SE$, respectively, where the "CI standard error" is

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

The CI is often written as $\hat{p} \pm z_{\text{crit}}SE$.

Reminder of CI interpretation. The CI is determined once the confidence level is specified and the data are collected. Prior to collecting data, the CI is unknown and can be viewed as random because it will depend on the actual sample selected. Different samples give different CIs. The "confidence" in, say, the 95% CI (which has a 0.05 or 5% error rate) can be interpreted as follows. If you repeatedly sample the population and construct 95% CIs for p, then 95% of the intervals will contain p, whereas 5% (the error rate) will not. The CI you get from your data either covers p, or it does not.

The length of the CI

$$U - L = 2z_{\text{crit}}SE$$

depends on the accuracy of the estimate $\hat{p}$, as measured by the standard error SE. For a given $\hat{p}$, this standard error decreases as the sample size n increases, yielding a narrower CI. For a fixed sample size, this standard error is maximized at $\hat{p} = 0.5$, and decreases as $\hat{p}$ moves towards either 0 or 1. In essence, sample proportions near 0 or 1 give narrower CIs for p. However, the normal approximation used in the CI construction is less reliable for extreme values of $\hat{p}$.

■ ClickerQs — CI for proportions STT.08.01.010 ■

Example: Tamper resistant packaging The 1983 Tylenol poisoning episode highlighted the desirability of using tamper-resistant packaging. The article "Tamper Resistant Packaging: Is it Really?" (Packaging Engineering, June 1983) reported the results of a survey on consumer attitudes towards tamper-resistant packaging. A sample of 270 consumers was asked the question: "Would you be willing to pay extra for tamper resistant packaging?" The number of yes respondents was 189. Construct a 95% CI for the proportion p of all consumers who were willing in 1983 to pay extra for such packaging.

Here $n = 270$ and $\hat{p} = 189/270 = 0.700$. The critical value for a 95% CI for

p is $z_{0.025} = 1.96$. The CI standard error is given by

$$SE = \sqrt{\frac{0.7 \times 0.3}{270}} = 0.028,$$

so $z_{\text{crit}} SE = 1.96 \times 0.028 = 0.055$. The 95% CI for p is 0.700 ± 0.055. You are 95% confident that the proportion of consumers willing to pay extra for better packaging is between 0.645 and 0.755. (Willing to pay how much extra?)

Appropriateness of the CI

The standard CI is based on a **large-sample** standard normal approximation to

$$z = \frac{\hat{p} - p}{SE}.$$

A simple rule of thumb requires $n\hat{p} \geq 5$ and $n(1 - \hat{p}) \geq 5$ for the method to be suitable. Given that $n\hat{p}$ and $n(1-\hat{p})$ are the observed numbers of successes and failures, you should have at least 5 of each to apply the large-sample CI.

In the packaging example, $n\hat{p} = 270 \times (0.700) = 189$ (the number who support the new packaging) and $n(1 - \hat{p}) = 270 \times (0.300) = 81$ (the number who oppose) both exceed 5. The normal approximation is appropriate here.

7.2.2 Hypothesis Tests on Proportions

The following example is typical of questions that can be answered using a hypothesis test for a population proportion.

Example Environmental problems associated with leaded gasolines are well-known. Many motorists have tampered with the emission control devices on their cars to save money by purchasing leaded rather than unleaded gasoline. A *Los Angeles Times* article on March 17, 1984 reported that 15% of all California motorists have engaged in emissions tampering. A random sample of 200 cars from L.A. county was obtained, and the emissions devices on 21 are

found to be tampered with. Does this suggest that the proportion of cars in L.A. county with tampered devices differs from the statewide proportion?

Two-Sided Hypothesis Test for p

Suppose you are interested in whether the population proportion p is equal to a prespecified value, say p_0. This question can be formulated as a two-sided test. To carry out the test:

1. Define the null hypothesis $H_0 : p = p_0$ and alternative hypothesis $H_A : p \neq p_0$.
2. Choose the size or significance level of the test, denoted by α.
3. Using the standard normal probability table, find the critical value z_{crit} such that the areas under the normal curve to the left and right of z_{crit} are $1 - 0.5\alpha$ and 0.5α, respectively. That is, $z_{\text{crit}} = z_{0.5\alpha}$.
4. Compute the test statistic (often to be labelled z_{obs})

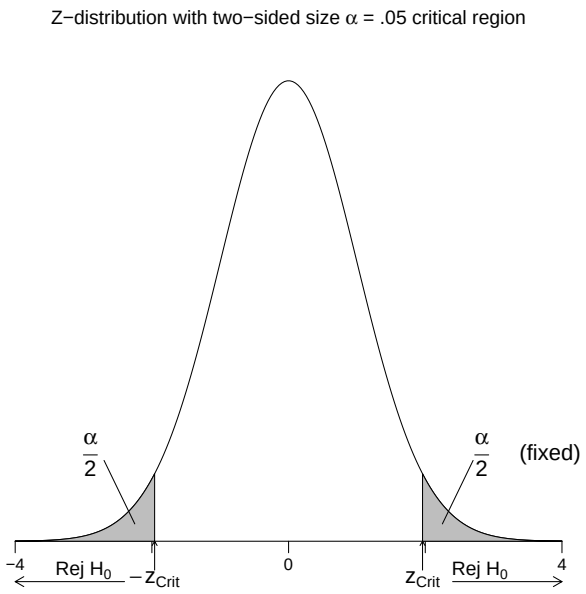
$$z_s = \frac{\hat{p} - p_0}{SE},$$

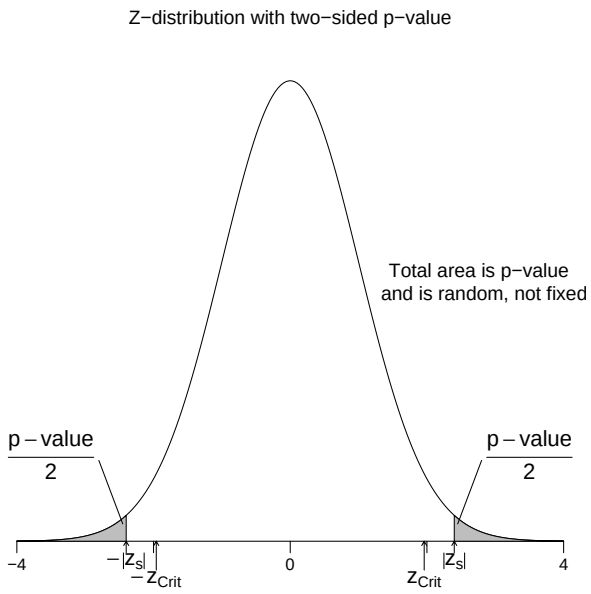
where the "test standard error" (based on the hypothesized value) is

$$SE = \sqrt{\frac{p_0(1-p_0)}{n}}.$$

5. Reject H_0 in favor of H_A if $|z_{\text{obs}}| \geq z_{\text{crit}}$. Otherwise, do not reject H_0.

The rejection rule is easily understood visually. The area under the normal curve outside $\pm z_{\text{crit}}$ is the size α of the test. One-half of α is the area in each tail. You reject H_0 in favor of H_A if the test statistic exceeds $\pm z_{\text{crit}}$. This occurs when $\hat{p}$ is significantly different from p_0, as measured by the standardized distance z_{obs} between $\hat{p}$ and p_0.





■ CLICKER Qs — Test statistic STT.07.01.057 ■

7.2.3 The p-value for a two-sided test

To compute the p-value (not to be confused with the value of the proportion p) for a two-sided test:

1. Compute the test statistic $z_s = z_{\text{obs}}$.
2. Evaluate the area under the normal probability curve outside $\pm|z_s|$.

Recall that the null hypothesis for a size α test is rejected if and only if the p-value is less than or equal to α.

Example: Emissions data Each car in the target population (L.A. county) either has been tampered with (a success) or has not been tampered with (a failure). Let $p =$ the proportion of cars in L.A. county with tampered emissions control devices. You want to test $H_0 : p = 0.15$ against $H_A : p \neq 0.15$ (here $p_0 = 0.15$). The critical value for a two-sided test of size $\alpha = 0.05$ is $z_{\text{crit}} = 1.96$.

The data are a sample of $n = 200$ cars. The sample proportion of cars that

have been tampered with is $\hat{p} = 21/200 = 0.105$. The test statistic is

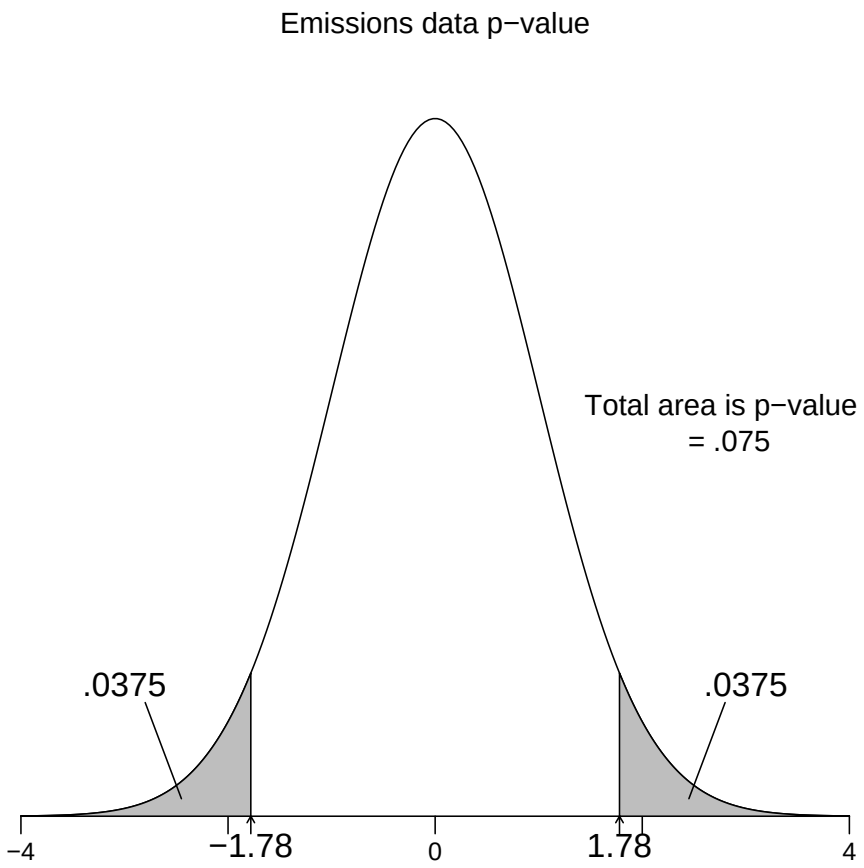
$$z_s = \frac{0.105 - 0.15}{0.02525} = -1.78,$$

where the test standard error satisfies

$$SE = \sqrt{\frac{0.15 \times 0.85}{200}} = 0.02525.$$

Given that $|z_s| = 1.78 < 1.96$, you have insufficient evidence to reject H_0 at the 5% level. That is, you have insufficient evidence to conclude that the proportion of cars in L.A. county that have been tampered with differs from the statewide proportion.

This decision is reinforced by the p-value calculation. The p-value is the area under the standard normal curve outside ± 1.78. This is $2 \times 0.0375 = 0.075$, which exceeds the test size of 0.05.



Remark The SE used in the test and CI are different. This implies that a hypothesis test and CI could potentially lead to different decisions. That is, a 95% CI for a population proportion might cover p_0 when the p-value for testing $H_0 : p = p_0$ is smaller than 0.05. This will happen, typically, only in cases where the decision is "borderline."

7.2.4 Appropriateness of Test

The z-test is based on a large-sample normal approximation, which works better for a given sample size when p_0 is closer to 0.5. The sample size needed for an accurate approximation increases dramatically the closer p_0 gets to 0 or to 1. A simple rule of thumb is that the test is appropriate when (the expected number

of successes) $np_0 \geq 5$ and (the expected number of failures) $n(1-p_0) \geq 5$.

In the emissions example, $np_0 = 200 \times (0.15) = 30$ and $n(1-p_0) = 200 \times (0.85) = 170$ exceed 5, so the normal approximation is appropriate.

7.2.5 R Implementation

```
#### Single Proportion Problems
# Approximate normal test for proportion, without Yates' continuity correction
prop.test(21, 200, p = 0.15, correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  21 out of 200, null probability 0.15
## X-squared = 3.1765, df = 1, p-value = 0.07471
## alternative hypothesis: true p is not equal to 0.15
## 95 percent confidence interval:
##  0.06970749 0.15518032
## sample estimates:
##     p
## 0.105
```

```
# Approximate normal test for proportion, with Yates' continuity correction
#prop.test(21, 200, p = 0.15)
```

■ ClickerQs — Parachute null hypothesis STT.08.01.040 ■

■ ClickerQs — Parachute conclusion STT.08.01.050 ■

■ ClickerQs — Parachute p-value STT.08.01.060 ■

7.2.6 One-Sided Tests and One-Sided Confidence Bounds

The mechanics of tests on proportions are similar to tests on means, except we use a different test statistic and a different probability distribution for critical values. This applies to one-sided and two-sided procedures. The example below illustrates a one-sided test and bound.

Example: brain hemispheres An article in the April 6, 1983 edition of *The Los Angeles Times* reported on a study of 53 learning-impaired youngsters at the Massachusetts General Hospital. The right side of the brain was found to be larger than the left side in 22 of the children. The proportion of the general population with brains having larger right sides is known to be 0.25. Does the data provide strong evidence for concluding, as the article claims, that the proportion of learning impaired youngsters with brains having larger right sides exceeds the proportion in the general population?

I will answer this question by computing a p-value for a one-sided test. Let p be the population proportion of learning disabled children with brains having larger right sides. I am interested in testing $H_0 : p = 0.25$ against $H_A : p > 0.25$ (here $p_0 = 0.25$).

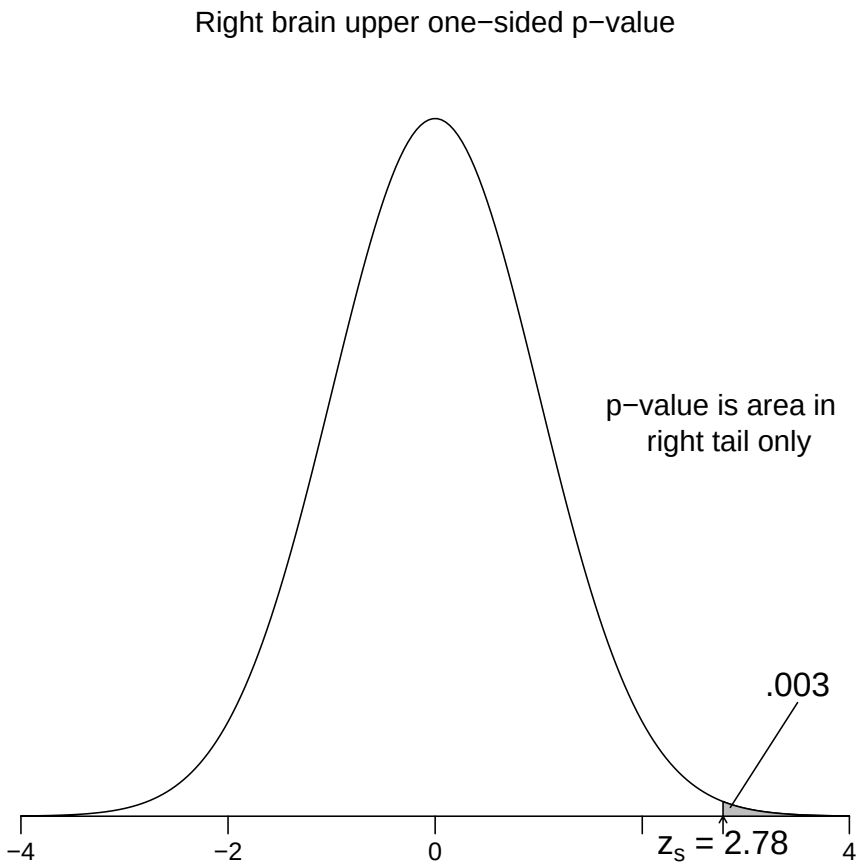
The proportion of children sampled with brains having larger right sides is $\hat{p} = 22/53 = 0.415$. The test statistic is

$$z_s = \frac{0.415 - 0.25}{0.0595} = 2.78,$$

where the test standard error satisfies

$$SE = \sqrt{\frac{0.25 \times 0.75}{53}} = 0.0595.$$

The p-value for an upper one-sided test is the area under the standard normal curve to the right of 2.78, which is approximately .003; see the picture below. I would reject H_0 in favor of H_A using any of the standard test levels, say 0.05 or 0.01. The newspaper's claim is reasonable.



A sensible next step in the analysis would be to compute a **lower confidence bound** $\hat{p} - z_{\text{crit}} SE$ for p. For illustration, consider a 95% bound. The CI standard error is

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.415 \times 0.585}{53}} = 0.0677.$$

The critical value for a one-sided 5% test is $z_{\text{crit}} = 1.645$, so a lower 95% bound on p is $0.415 - 1.645 \times 0.0677 = 0.304$. I am 95% confident that the population proportion of learning disabled children with brains having larger right sides is at least 0.304. Values of p smaller than 0.304 are not supported by the data.

You should verify that the sample size is sufficiently large to use the approximate methods in this example.

```r
#### Example: brain hemispheres
# Approximate normal test for proportion, without Yates' continuity correction
prop.test(22, 53, p = 0.25, alternative = "greater", correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  22 out of 53, null probability 0.25
## X-squared = 7.7044, df = 1, p-value = 0.002754
## alternative hypothesis: true p is greater than 0.25
## 95 percent confidence interval:
##  0.3105487 1.0000000
## sample estimates:
##         p
## 0.4150943
```

7.2.7 Small Sample Procedures

Large sample tests and CIs for p should be interpreted with caution in small sized samples because the true error rate usually exceeds the assumed (nominal) value. For example, an assumed 95% CI, with a nominal error rate of 5%, may be only an 80% CI, with a 20% error rate. The large-sample CIs are usually overly optimistic (i.e., too narrow) when the sample size is too small to use the normal approximation.

Alan Agresti suggests the following method for a 95% CI. The standard method computes the sample proportion as $\hat{p} = x/n$ where x is the number of successes in the sample and n is the sample size. Agresti suggested using the estimated proportion $\tilde{p} = (x+2)/(n+4)$ with the standard error

$$SE = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}},$$

in the "usual 95% interval" formula: $\tilde{p} \pm 1.96SE$. This appears odd, but amounts to adding two successes and two failures to the observed data, and then computing the standard CI.

This adjustment has little effect when n is large and $\hat{p}$ is not near either 0 or 1, as in the Tylenol example.

Example: swimming segregation This example is based on a case heard before the U.S. Supreme Court. A racially segregated swimming club was ordered to admit minority members. However, it is unclear whether the club has been following the spirit of the mandate. Historically, 85% of the white applicants were approved. Since the mandate, only 1 of 6 minority applicants has been approved. Is there evidence of continued discrimination?

I examine this issue by constructing a CI and a test for the probability p (or population proportion) that a minority applicant is approved. Before examining the question of primary interest, let me show that the two approximate CIs are very different, due to the small sample size. One minority applicant ($x = 1$) was approved out of $n = 6$ candidates, giving $\hat{p} = 1/6 = 0.167$.

A 95% large-sample CI for p is $(-0.14, 0.46)$. Since a negative proportion is not possible, the CI should be reported as $(0.00, 0.46)$. Agresti's 95% CI (based on 3 successes and 7 failures) is (0.02, 0.58). The big difference between the two intervals coupled with the negative lower endpoint on the standard CI suggests that the normal approximation used with the standard method is inappropriate. This view is reinforced by the rule-of-thumb calculation for using the standard interval. Agresti's CI is wider, which is consistent with my comment that the standard CI is too narrow in small samples. As a comparison, the exact 95% CI is $(0.004, 0.64)$, which agrees more closely with Agresti's interval.

I should emphasize that the exact CI is best to use, but is not available in all statistical packages, so methods based on approximations may be required, and if so, then Agresti's method is clearly better than the standard normal approximation in small sized samples.

Recall that the results of the asymptotic 95% CIs may disagree with the hypothesis test results. Exact methods will not contradict each other this way (neither do these asymptotic methods, usually).

```
#### Example: swimming segregation
## The prop.test() does an additional adjustment, so does not match precisely
##   the results in the above paragraphs

# Approximate normal test for proportion, without Yates' continuity correction
prop.test(1, 6, p = 0.85, correct = FALSE)$conf.int
```

```
## Warning in prop.test(1, 6, p = 0.85, correct = FALSE): Chi-squared approximation may be
incorrect
## [1] 0.03005337 0.56350282
## attr(,"conf.level")
## [1] 0.95
```

```
# Agresti's method
prop.test(1+2, 6+4, p = 0.85, correct = FALSE)$conf.int
```

```
## Warning in prop.test(1 + 2, 6 + 4, p = 0.85, correct = FALSE): Chi-squared approximation
may be incorrect
## [1] 0.1077913 0.6032219
## attr(,"conf.level")
## [1] 0.95
```

```
# Exact binomial test for proportion
binom.test(1, 6, p = 0.85)$conf.int
```

```
## [1] 0.004210745 0.641234579
## attr(,"conf.level")
## [1] 0.95
```

Returning to the problem, you might check for discrimination by testing $H_0 : p = 0.85$ against $H_A : p < 0.85$ using an **exact** test. The exact test p-value is 0.000 to three decimal places, and an exact upper bound for p is 0.582. What does this suggest to you?

```
# Exact binomial test for proportion
binom.test(1, 6, alternative = "less", p = 0.85)
```

```
##
##  Exact binomial test
##
## data:  1 and 6
## number of successes = 1, number of trials = 6, p-value =
## 0.0003987
## alternative hypothesis: true probability of success is less than 0.85
## 95 percent confidence interval:
##  0.0000000 0.5818034
## sample estimates:
## probability of success
##              0.1666667
```

7.3 Analyzing Raw Data

In most studies, your data will be stored in a spreadsheet with one observation per case or individual. For example, the data below give the individual responses to the applicants of the swim club.

```
#### Example: swimming segregation, raw data
## read.table options
# sep = default is any white space, but our strings contain a space,
#       so I changed this to a comma
# header = there are no column headers
# stringsAsFActors = default converts strings to factors, but I want them
#       to just be the plain character text
swim <- read.table(text="
not approved
not approved
not approved
approved
not approved
not approved
", sep = ",", header=FALSE, stringsAsFactors=FALSE)

# name the column
names(swim) <- c("application")
# show the structure of the data.frame
str(swim)
```

```
## 'data.frame': 6 obs. of  1 variable:
##  $ application: chr  "not approved" "not approved" "not approved" "approved" ...
```

```
# display the data.frame
swim
```

```
##    application
## 1 not approved
## 2 not approved
## 3 not approved
## 4     approved
## 5 not approved
## 6 not approved
```

The data were entered as alphabetic strings. We can use `table()` to count frequencies of categorical variables.

```
# count the frequency of each categorical variable
table(swim)
```

```
## swim
##     approved not approved
##            1            5
```

You can compute a CI and test for a proportion using raw data, provided the data column includes only two distinct values. The levels can be numerical or alphanumeric.

```
# use the counts from table() for input in binom.text()
# the help for binom.test() says x can be a vector of length 2
#   giving the numbers of successes and failures, respectively
#   that's exactly what table(swim) gave us
```

```
binom.test(table(swim), p = 0.85, alternative = "less")
```

```
##
##  Exact binomial test
##
## data:  table(swim)
## number of successes = 1, number of trials = 6, p-value =
## 0.0003987
## alternative hypothesis: true probability of success is less than 0.85
## 95 percent confidence interval:
##  0.0000000 0.5818034
## sample estimates:
## probability of success
##              0.1666667
```

It is possible that the order (alphabetically) is the wrong order, failures and successes, in which case we'd need to reorder the input to `binom.test()`.

In Chapter 6 we looked at the binomial distribution to obtain an exact Sign Test confidence interval for the median. Examine the following to see where the exact p-value for this test comes from.

```
n <- 6
x <- 0:n
p0 <- 0.85
bincdf <- pbinom(x, n, p0)
cdf <- data.frame(x, bincdf)
cdf
```

```
##   x       bincdf
## 1 0 1.139063e-05
## 2 1 3.986719e-04
## 3 2 5.885156e-03
## 4 3 4.733859e-02
## 5 4 2.235157e-01
## 6 5 6.228505e-01
## 7 6 1.000000e+00
```

■ CLICKER Qs — Excess successes STT.05.01.030 ■

7.4 Goodness-of-Fit Tests (Multinomial)

Example: jury pool The following data set was used as evidence in a court case. The data represent a sample of 1336 individuals from the jury pool

of a large municipal court district for the years 1975–1977. The fairness of the representation of various age groups on juries was being contested. The strategy for doing this was to challenge the representativeness of the pool of individuals from which the juries are drawn. This was done by comparing the age group distribution within the jury pool against the age distribution in the district as a whole, which was available from census figures.

Age group (yrs)	Obs. Counts	Obs. Prop.	Census Prop.
18-19	23	0.017	0.061
20-24	96	0.072	0.150
25-29	134	0.100	0.135
30-39	293	0.219	0.217
40-49	297	0.222	0.153
50-64	380	0.284	0.182
65-99	113	0.085	0.102
Total:	1336	1.000	1.000

A statistical question here is whether the jury pool population proportions are equal to the census proportions across the age categories. This comparison can be formulated as a **goodness-of-fit test**, which generalizes the large-sample test on a single proportion to a categorical variable (here age) with $r > 2$ levels. For $r = 2$ categories, the goodness-of-fit test and large-sample test on a single proportion are identical. Although this problem compares two populations, only one sample is involved because the census data is a population summary!

In general, suppose each individual in a population is categorized into one and only one of r levels or categories. Let p_1, p_2, ..., p_r, be the population proportions in the r categories, where each $p_i \geq 0$ and $p_1 + p_2 + \cdots + p_r = 1$. The hypotheses of interest in a goodness-of-fit problem are $H_0 : p_1 = p_{01}$, $p_2 = p_{02}$, ..., $p_r = p_{0r}$ and $H_A :$ not H_0, where p_{01}, p_{02}, ..., p_{0r} are given category proportions.

The plausibility of H_0 is evaluated by comparing the hypothesized category proportions to estimated (i.e., observed) category proportions $\hat{p}_1$, $\hat{p}_2$, ..., $\hat{p}_r$

from a random or representative sample of n individuals selected from the population. The discrepancy between the hypothesized and observed proportions is measured by the Pearson chi-squared statistic:

$$\chi^2_{\text{s}} = \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i},$$

where O_i is the **observed** number in the sample that fall into the i^{th} category ($O_i = n\hat{p}_i$), and $E_i = np_{0i}$ is the number of individuals **expected** to be in the i^{th} category when H_0 is true.

The Pearson statistic can also be computed as the sum of the squared residuals:

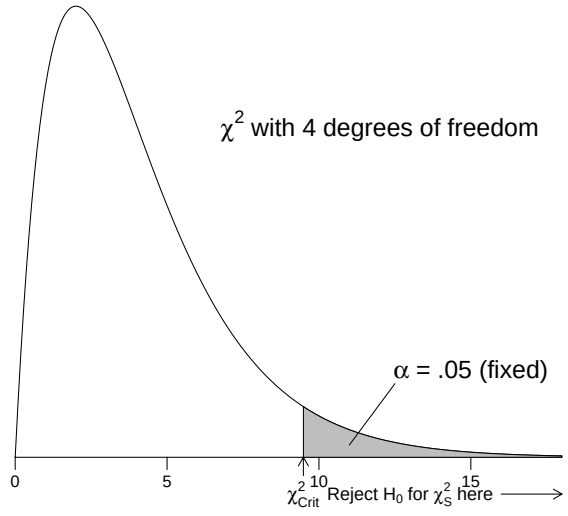
$$\chi^2_{\text{s}} = \sum_{i=1}^{r} Z_i^2,$$

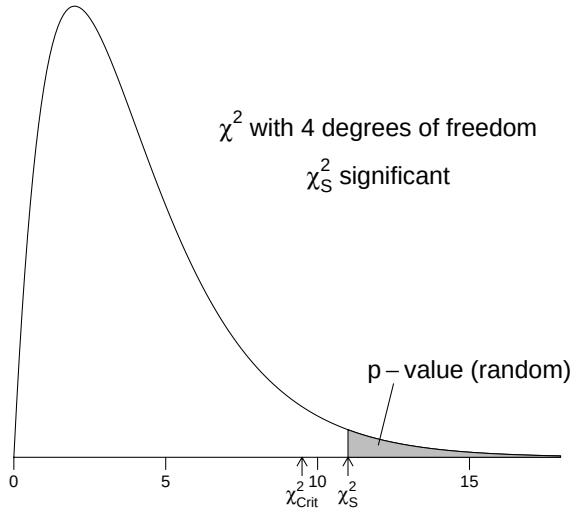
where $Z_i = (O_i - E_i)/\sqrt{E_i}$, or in terms of the observed and hypothesized category proportions

$$\chi^2_{\text{s}} = n \sum_{i=1}^{r} \frac{(\hat{p}_i - p_{0i})^2}{p_{0i}}.$$

The Pearson statistic χ^2_{s} is "small" when all of the observed counts (proportions) are close to the expected counts (proportions). The Pearson χ^2 is "large" when one or more observed counts (proportions) differs noticeably from what is expected when H_0 is true. Put another way, large values of χ^2_{s} suggest that H_0 is false.

The critical value χ^2_{crit} for the test is obtained from a chi-squared probability table with $r - 1$ degrees of freedom. The picture below shows the form of the rejection region. For example, if $r = 5$ and $\alpha = 0.05$, then you reject H_0 when $\chi^2_s \geq \chi^2_{\text{crit}} = 9.49$ (`qchisq(0.95, 5-1)`). The p-value for the test is the area under the chi-squared curve with $df = r - 1$ to the right of the observed χ^2_s value.





Example: jury pool Let p_{18} be the proportion in the jury pool population between ages 18 and 19. Define p_{20}, p_{25}, p_{30}, p_{40}, p_{50}, and p_{65} analogously. You are interested in testing that the true jury proportions equal the census proportions, $H_0 : p_{18} = 0.061$, $p_{20} = 0.150$, $p_{25} = 0.135$, $p_{30} = 0.217$, $p_{40} = 0.153$, $p_{50} = 0.182$, and $p_{65} = 0.102$ against H_A : not H_0, using the sample of 1336 from the jury pool.

The observed counts, the expected counts, and the category residuals are given in the table below. For example, $E_{18} = 1336 \times (0.061) = 81.5$ and $Z_{18} = (23 - 81.5)/\sqrt{81.5} = -6.48$ in the 18-19 year category.

The Pearson statistic is

$$\chi^2_{\rm s} = (-6.48)^2 + (-7.38)^2 + (-3.45)^2 + 0.18^2 + 6.48^2 + 8.78^2 + (-1.99)^2 = 231.26$$

on $r - 1 = 7 - 1 = 6$ degrees of freedom. Here $\chi^2_{\rm crit} = 12.59$ at $\alpha = 0.05$. The p-value for the goodness-of-fit test is less than 0.001, which suggests that H_0 is false.

Age group (yrs)	Obs. Counts	Exp. Counts	Residual
18-19	23	81.5	−6.48
20-24	96	200.4	−7.38
25-29	134	180.4	−3.45
30-39	293	289.9	0.18
40-49	297	204.4	6.48
50-64	380	243.2	8.78
65-99	113	136.3	−1.99

7.4.1 Adequacy of the Goodness-of-Fit Test

The chi-squared goodness-of-fit test is a large-sample test. A conservative rule of thumb is that the test is suitable when each **expected** count is at least five. This holds in the jury pool example. There is no widely available alternative method for testing goodness-of-fit with smaller sample sizes. There is evidence, however, that the chi-squared test is **slightly conservative** (the p-values are too large, on average) when the expected counts are smaller. Some statisticians recommend that the chi-squared approximation be used when the minimum expected count is at least one, provided the expected counts are not too variable.

7.4.2 R Implementation

```
#### Example: jury pool
jury <- read.table(text="
Age    Count  CensusProp
18-19     23       0.061
20-24     96       0.150
25-29    134       0.135
30-39    293       0.217
40-49    297       0.153
50-64    380       0.182
65-99    113       0.102
", header=TRUE)

# show the structure of the data.frame
str(jury)
```

```
## 'data.frame': 7 obs. of  3 variables:
```

```
##  $ Age       : Factor w/ 7 levels "18-19","20-24",..: 1 2 3 4 5 6 7
##  $ Count     : int  23 96 134 293 297 380 113
##  $ CensusProp: num  0.061 0.15 0.135 0.217 0.153 0.182 0.102
```

```
# display the data.frame
jury
```

```
##     Age Count CensusProp
## 1 18-19    23      0.061
## 2 20-24    96      0.150
## 3 25-29   134      0.135
## 4 30-39   293      0.217
## 5 40-49   297      0.153
## 6 50-64   380      0.182
## 7 65-99   113      0.102
```

```
# calculate chi-square goodness-of-fit test
x.summary <- chisq.test(jury$Count, correct = FALSE, p = jury$CensusProp)
# print result of test
x.summary
```

```
##
##  Chi-squared test for given probabilities
##
## data:  jury$Count
## X-squared = 231.26, df = 6, p-value < 2.2e-16
```

```
# use output in x.summary and create table
x.table <- data.frame(age = jury$Age
                      , obs = x.summary$observed
                      , exp = x.summary$expected
                      , res = x.summary$residuals
                      , chisq = x.summary$residuals^2
                      , stdres = x.summary$stdres)
x.table
```

```
##     age obs     exp        res       chisq     stdres
## 1 18-19  23  81.496 -6.4797466 41.98711613 -6.6869061
## 2 20-24  96 200.400 -7.3748237 54.38802395 -7.9991194
## 3 25-29 134 180.360 -3.4520201 11.91644267 -3.7116350
## 4 30-39 293 289.912  0.1813611  0.03289186  0.2049573
## 5 40-49 297 204.408  6.4762636 41.94199084  7.0369233
## 6 50-64 380 243.152  8.7760589 77.01921063  9.7033764
## 7 65-99 113 136.272 -1.9935650  3.97430128 -2.1037408
```

Plot observed vs expected values to help identify age groups that deviate the most. Plot contribution to chi-square values to help identify age groups that deviate the most. The term "Contribution to Chi-Square" (`chisq`) refers to the values of $\frac{(O-E)^2}{E}$ for each category. χ^2_s is the sum of those contributions.

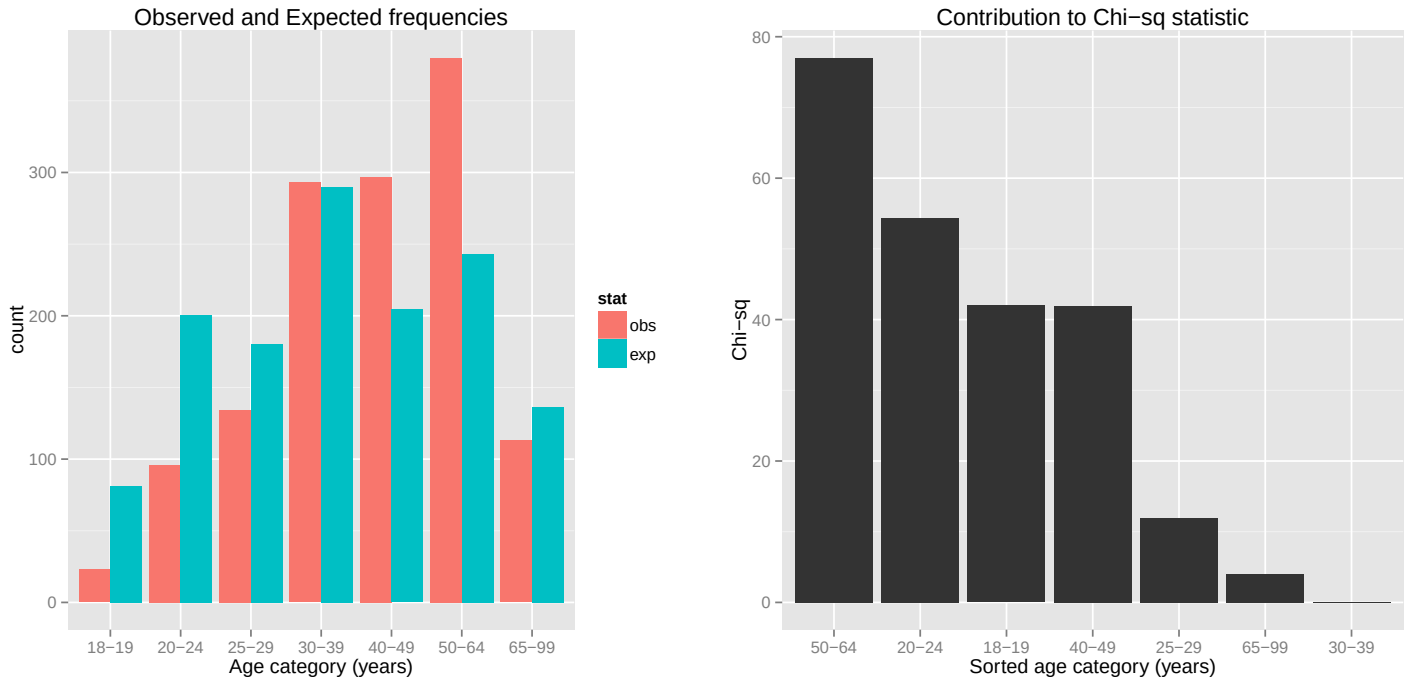
```r
library(reshape2)
x.table.obsexp <- melt(x.table,
              # id.vars: ID variables
              #   all variables to keep but not split apart on
              id.vars = c("age"),
              # measure.vars: The source columns
              #   (if unspecified then all other variables are measure.vars)
              measure.vars = c("obs", "exp"),
              # variable.name: Name of the destination column identifying each
              #   original column that the measurement came from
              variable.name = "stat",
              # value.name: column name for values in table
              value.name = "value"
            )

# naming variables manually, the variable.name and value.name not working 11/2012
names(x.table.obsexp) <- c("age", "stat", "value")

# Observed vs Expected counts
library(ggplot2)
p <- ggplot(x.table.obsexp, aes(x = age, fill = stat, weight=value))
p <- p + geom_bar(position="dodge")
p <- p + labs(title = "Observed and Expected frequencies")
p <- p + xlab("Age category (years)")
print(p)


# Contribution to chi-sq
# pull out only the age and chisq columns
x.table.chisq <- x.table[, c("age","chisq")]
# reorder the age categories to be descending relative to the chisq statistic
x.table.chisq$age <- with(x.table, reorder(age, -chisq))

p <- ggplot(x.table.chisq, aes(x = age, weight = chisq))
p <- p + geom_bar()
p <- p + labs(title = "Contribution to Chi-sq statistic")
p <- p + xlab("Sorted age category (years)")
p <- p + ylab("Chi-sq")
print(p)
```



7.4.3 Multiple Comparisons in a Goodness-of-Fit Problem

The goodness-of-fit test suggests that at least one of the age category proportions for the jury pool population differs from the census figures. A reasonable next step in the analysis would be to **separately** test the seven hypotheses: $H_0 : p_{18} = 0.061$, $H_0 : p_{20} = 0.150$, $H_0 : p_{25} = 0.135$, $H_0 : p_{30} = 0.217$, $H_0 : p_{40} = 0.153$, $H_0 : p_{50} = 0.182$, and $H_0 : p_{65} = 0.102$ to see which age categories led to this conclusion.

A Bonferroni comparison with a Family Error Rate ≤ 0.05 will be considered for this multiple comparisons problem. The error rates for the seven individual tests are set to $\alpha = 0.05/7 = 0.0071$, which corresponds to 99.29% two-sided CIs for the individual jury pool proportions. The area under the standard normal curve to the right of 2.69 is $0.05/2/7 = 0.00357$, about one-half the error rate for the individual CIs, so the critical value for the CIs, or for the tests, is $z_{\text{crit}} \approx 2.69$. The next table gives individual 99.29% CIs based on the large sample approximation. You can get the individual CIs in R using the `binom.test()` or `prop.test()` function. For example, the CI for Age Group

18-19 is obtained by specifying 23 successes in 1336 trials.

Below I perform exact binomial tests of proportion for each of the seven age categories at the Bonferroni-adjusted significance level. I save the p-values and confidence intervals in a table along with the observed and census proportions in order to display the table below.

```
b.sum1 <- binom.test(jury$Count[1], sum(jury$Count), p = jury$CensusProp[1], alternative = "two.sided", conf.level = 1-0.05/7)
b.sum2 <- binom.test(jury$Count[2], sum(jury$Count), p = jury$CensusProp[2], alternative = "two.sided", conf.level = 1-0.05/7)
b.sum3 <- binom.test(jury$Count[3], sum(jury$Count), p = jury$CensusProp[3], alternative = "two.sided", conf.level = 1-0.05/7)
b.sum4 <- binom.test(jury$Count[4], sum(jury$Count), p = jury$CensusProp[4], alternative = "two.sided", conf.level = 1-0.05/7)
b.sum5 <- binom.test(jury$Count[5], sum(jury$Count), p = jury$CensusProp[5], alternative = "two.sided", conf.level = 1-0.05/7)
b.sum6 <- binom.test(jury$Count[6], sum(jury$Count), p = jury$CensusProp[6], alternative = "two.sided", conf.level = 1-0.05/7)
b.sum7 <- binom.test(jury$Count[7], sum(jury$Count), p = jury$CensusProp[7], alternative = "two.sided", conf.level = 1-0.05/7)
# put the p-value and CI into a data.frame
b.sum  <- data.frame(
           rbind( c(b.sum1$p.value, b.sum1$conf.int)
                , c(b.sum2$p.value, b.sum2$conf.int)
                , c(b.sum3$p.value, b.sum3$conf.int)
                , c(b.sum4$p.value, b.sum4$conf.int)
                , c(b.sum5$p.value, b.sum5$conf.int)
                , c(b.sum6$p.value, b.sum6$conf.int)
                , c(b.sum7$p.value, b.sum7$conf.int)
           )
         )
names(b.sum) <- c("p.value", "CI.lower", "CI.upper")
b.sum$Age <- jury$Age
b.sum$Observed <- x.table$obs/sum(x.table$obs)
b.sum$CensusProp <- jury$CensusProp
b.sum
```

```
##        p.value   CI.lower   CI.upper   Age   Observed CensusProp
## 1 8.814860e-15 0.00913726 0.02920184 18-19 0.01721557      0.061
## 2 2.694633e-18 0.05415977 0.09294037 20-24 0.07185629      0.150
## 3 1.394274e-04 0.07939758 0.12435272 25-29 0.10029940      0.135
## 4 8.421685e-01 0.18962122 0.25120144 30-39 0.21931138      0.217
## 5 2.383058e-11 0.19245560 0.25433144 40-49 0.22230539      0.153
## 6 5.915839e-20 0.25174398 0.31880556 50-64 0.28443114      0.182
## 7 3.742335e-02 0.06536589 0.10707682 65-99 0.08458084      0.102
```

The CIs for the 30-39 and 65-99 year categories contain the census proportions. In the other five age categories, there are significant differences between the jury pool proportions and the census proportions. In general, young adults appear to be underrepresented in the jury pool whereas older age groups are overrepresented.

	Age	p.value	CI.lower	CI.upper	Observed	CensusProp
1	18-19	0.000	0.009	0.029	0.017	0.061
2	20-24	0.000	0.054	0.093	0.072	0.150
3	25-29	0.000	0.079	0.124	0.100	0.135
4	30-39	0.842	0.190	0.251	0.219	0.217
5	40-49	0.000	0.192	0.254	0.222	0.153
6	50-64	0.000	0.252	0.319	0.284	0.182
7	65-99	0.037	0.065	0.107	0.085	0.102

The residuals also highlight significant differences because the largest resid-

uals correspond to the categories that contribute most to the value of χ^2_s. Some researchers use the residuals for the multiple comparisons, treating the Z_is as standard normal variables. Following this approach, you would conclude that the jury pool proportions differ from the proportions in the general population in every age category where $|Z_i| \geq 2.70$ (using the same Bonferroni correction). This gives the same conclusion as before.

The two multiple comparison methods are similar, but not identical. The residuals

$$Z_i = \frac{O_i - E_i}{\sqrt{E_i}} = \frac{\hat{p}_i - p_{0i}}{\sqrt{\frac{p_{0i}}{n}}}$$

agree with the large-sample statistic for testing $H_0 : p_i = p_{0i}$, except that the divisor in Z_i omits a $1 - p_{0i}$ term. The Z_is are not standard normal random variables as assumed, and the value of Z_i underestimates the significance of the observed differences. Multiple comparisons using the Z_is will find, on average, fewer significant differences than the preferred method based on the large sample tests. However, the differences between the two methods are usually minor when all of the hypothesized proportions are small.

7.5 Comparing Two Proportions: Independent Samples

The New Mexico state legislature is interested in how the proportion of registered voters that support Indian gaming differs between New Mexico and Colorado. Assuming neither population proportion is known, the state's statistician might recommend that the state conduct a survey of registered voters sampled independently from the two states, followed by a comparison of the sample proportions in favor of Indian gaming.

Statistical methods for comparing two proportions using independent samples can be formulated as follows. Let p_1 and p_2 be the proportion of populations 1 and 2, respectively, with the attribute of interest. Let $\hat{p}_1$ and $\hat{p}_2$ be the corresponding sample proportions, based on independent random or representative

samples of size n_1 and n_2 from the two populations.

7.5.1 Large Sample CI and Tests for $p_1 - p_2$

A large-sample CI for $p_1 - p_2$ is $(\hat{p}_1 - \hat{p}_2) \pm z_{\text{crit}} SE_{CI}(\hat{p}_1 - \hat{p}_2)$, where z_{crit} is the standard normal critical value for the desired confidence level, and

$$SE_{CI}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

is the CI standard error.

A large-sample p-value for a test of the null hypothesis $H_0 : p_1 - p_2 = 0$ against the two-sided alternative $H_A : p_1 - p_2 \neq 0$ is evaluated using tail areas of the standard normal distribution (identical to one sample evaluation) in conjunction with the test statistic

$$z_s = \frac{\hat{p}_1 - \hat{p}_2}{SE_{test}(\hat{p}_1 - \hat{p}_2)},$$

where

$$SE_{test}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}} = \sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

is the test standard error for $\hat{p}_1 - \hat{p}_2$. The **pooled proportion**

$$\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

is the proportion of successes in the two samples combined. The test standard error has the same functional form as the CI standard error, with $\bar{p}$ replacing the individual sample proportions.

The pooled proportion is the best guess at the common population proportion when $H_0 : p_1 = p_2$ is true. The test standard error estimates the standard deviation of $\hat{p}_1 - \hat{p}_2$ assuming H_0 is true.

Remark: As in the one-sample proportion problem, the test and CI SE's are different. This *can* (but usually does not) lead to some contradiction between the test and CI.

Example, vitamin C Two hundred and seventy nine (279) French skiers were studied during two one-week periods in 1961. One group of 140 skiers receiving a placebo each day, and the other 139 receiving 1 gram of ascorbic acid (Vitamin C) per day. The study was double blind — neither the subjects nor the researchers knew who received which treatment. Let p_1 be the probability that a member of the ascorbic acid group contracts a cold during the study period, and p_2 be the corresponding probability for the placebo group. Linus Pauling (Chemistry and Peace Nobel prize winner) and I are interested in testing whether $H_0 : p_1 = p_2$. The data are summarized below as a two-by-two table of counts (a contingency table)

Outcome	Ascorbic Acid	Placebo
# with cold	17	31
# with no cold	122	109
Totals	139	140

The sample sizes are $n_1 = 139$ and $n_2 = 140$. The sample proportion of skiers developing colds in the placebo and treatment groups are $\hat{p}_2 = 31/140 = 0.221$ and $\hat{p}_1 = 17/139 = 0.122$, respectively. The difference is $\hat{p}_1 - \hat{p}_2 = 0.122 - 0.221 = -0.099$. The pooled proportion is the number of skiers that developed colds divided by the number of skiers in the study: $\bar{p} = 48/279 = 0.172$.

The test standard error is

$$SE_{test}(\hat{p}_1 - \hat{p}_2) = \sqrt{0.172 \times (1 - 0.172)\left(\frac{1}{139} + \frac{1}{140}\right)} = 0.0452.$$

The test statistic is

$$z_s = \frac{0.122 - 0.221}{0.0452} = -2.19.$$

The p-value for a two-sided test is twice the area under the standard normal curve to the right of 2.19 (or twice the area to the left of -2.19), which is $2 \times$

$(0.014) = 0.028$. At the 5% level, we reject the hypothesis that the probability of contracting a cold is the same whether you are given a placebo or Vitamin C.

A CI for $p_1 - p_2$ provides a measure of the size of the treatment effect. For a 95% CI

$$\begin{aligned} z_{\text{crit}} SE_{CI}(\hat{p}_1 - \hat{p}_2) &= 1.96\sqrt{\frac{0.221 \times (1-0.221)}{140} + \frac{0.122 \times (1-0.122)}{139}} \\ &= 1.96 \times (0.04472) = 0.088. \end{aligned}$$

The 95% CI for $p_1 - p_2$ is -0.099 ± 0.088, or $(-0.187, -0.011)$. We are 95% confident that p_2 exceeds p_1 by at least 0.011 but not by more than 0.187.

On the surface, we would conclude that a daily dose of Vitamin C decreases a French skier's chance of developing a cold by between 0.011 and 0.187 (with 95% confidence). This conclusion was somewhat controversial. Several reviews of the study felt that the experimenter's evaluations of cold symptoms were unreliable. Many other studies refute the benefit of Vitamin C as a treatment for the common cold.

```
#### Example, vitamin C
# Approximate normal test for two-proportions, without Yates' continuity correction
prop.test(c(17, 31), c(139, 140), correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(17, 31) out of c(139, 140)
## X-squared = 4.8114, df = 1, p-value = 0.02827
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.18685917 -0.01139366
## sample estimates:
##    prop 1    prop 2
## 0.1223022 0.2214286
```

Conditional probability

In probability theory, a conditional probability is the probability that an event will occur, when another event is known to occur or to have occurred. If

the events are A and B respectively, this is said to be "the probability of A given B". It is commonly denoted by $\Pr(A|B)$. $\Pr(A|B)$ may or may not be equal to $\Pr(A)$, the probability of A. If they are equal, A and B are said to be independent. For example, if a coin is flipped twice, "the outcome of the second flip" is independent of "the outcome of the first flip".

In the Vitamin C example above, the unconditional observed probability of contracting a cold is $\Pr(\text{cold}) = (17 + 31)/(139 + 140) = 0.172$. The conditional observed probabilities are $\Pr(\text{cold}|\text{ascorbic acid}) = 17/139 = 0.1223$ and $\Pr(\text{cold}|\text{placebo}) = 31/140 = 0.2214$. The two-sample test of $H_0 : p_1 = p_2$ where $p_1 = \Pr(\text{cold}|\text{ascorbic acid})$ and $p_2 = \Pr(\text{cold}|\text{placebo})$ is effectively testing whether $\Pr(\text{cold}) = \Pr(\text{cold}|\text{ascorbic acid}) = \Pr(\text{cold}|\text{placebo})$. This tests whether contracting a cold is independent of the vitamin C treatment.

Example, cervical dysplasia A case-control study was designed to examine risk factors for cervical dysplasia[2] (Becker et al. 1994). All the women in the study were patients at UNM clinics. The 175 cases were women, aged 18-40, who had cervical dysplasia. The 308 controls were women aged 18-40 who did not have cervical dysplasia. Each woman was classified as positive or negative, depending on the presence of HPV (human papilloma virus). The data are summarized below.

HPV Outcome	Cases	Controls
Positive	164	130
Negative	11	178
Sample size	175	308

We'll take a short detour to create this table in R, calculate the column proportions, and plot the frequencies and proportions.

We first create a table with labelled rows and columns.

```
# Create the labelled table
hpv <-
matrix(c(164, 130, 11, 178),
       nrow = 2, byrow = TRUE,
       dimnames = list("HPV.Outcome" = c("Positive", "Negative"),
```

[2] http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0002461/

```
                      "Group" = c("Cases", "Controls")))
hpv
```

```
##             Group
## HPV.Outcome Cases Controls
##    Positive   164      130
##    Negative    11      178
```

Next, we create column proportions.

```
# calculate the column (margin = 2) proportions
hpv.col.prop <- prop.table(hpv, margin = 2)
hpv.col.prop
```

```
##             Group
## HPV.Outcome      Cases  Controls
##    Positive 0.93714286 0.4220779
##    Negative 0.06285714 0.5779221
```

Here, we reshape the data from wide format to long format. This will allow us to make plots later, and also shows how to create these tables from a dataset in long format (which is the typical format).

```
# OR, convert to long format and use xtabs to produce the table
library(reshape2)
hpv.long <- melt(hpv, value.name = "Frequency")
hpv.long
```

```
##   HPV.Outcome    Group Frequency
## 1    Positive    Cases       164
## 2    Negative    Cases        11
## 3    Positive Controls       130
## 4    Negative Controls       178
```

```
T1 <- xtabs(Frequency ~ HPV.Outcome + Group, data = hpv.long)
T1
```

```
##             Group
## HPV.Outcome Cases Controls
##    Positive   164      130
##    Negative    11      178
```

```
hpv.col.prop <- prop.table(T1, margin = 2)
hpv.col.prop
```

```
##             Group
## HPV.Outcome      Cases   Controls
##    Positive 0.93714286 0.42207792
##    Negative 0.06285714 0.57792208
```

Add a column of proportions to our long-formatted data to plot these proportion values.

```
library(reshape2)
hpv.col.prop.long <- melt(hpv.col.prop, value.name = "Proportion")
hpv.col.prop.long
```

```
##   HPV.Outcome    Group Proportion
## 1    Positive    Cases 0.93714286
## 2    Negative    Cases 0.06285714
## 3    Positive Controls 0.42207792
## 4    Negative Controls 0.57792208
```

```
# join these two datasets to have both Freq and Prop columns
library(plyr)
hpv.long <- join(hpv.long, hpv.col.prop.long)
```

```
## Joining by:  HPV.Outcome, Group
```

```
hpv.long
```

```
##   HPV.Outcome    Group Frequency Proportion
## 1    Positive    Cases       164 0.93714286
## 2    Negative    Cases        11 0.06285714
## 3    Positive Controls       130 0.42207792
## 4    Negative Controls       178 0.57792208
```

Finally, plot the frequencies, and the proportions in three ways (the frequencies can obviously be plotted in many ways, too).

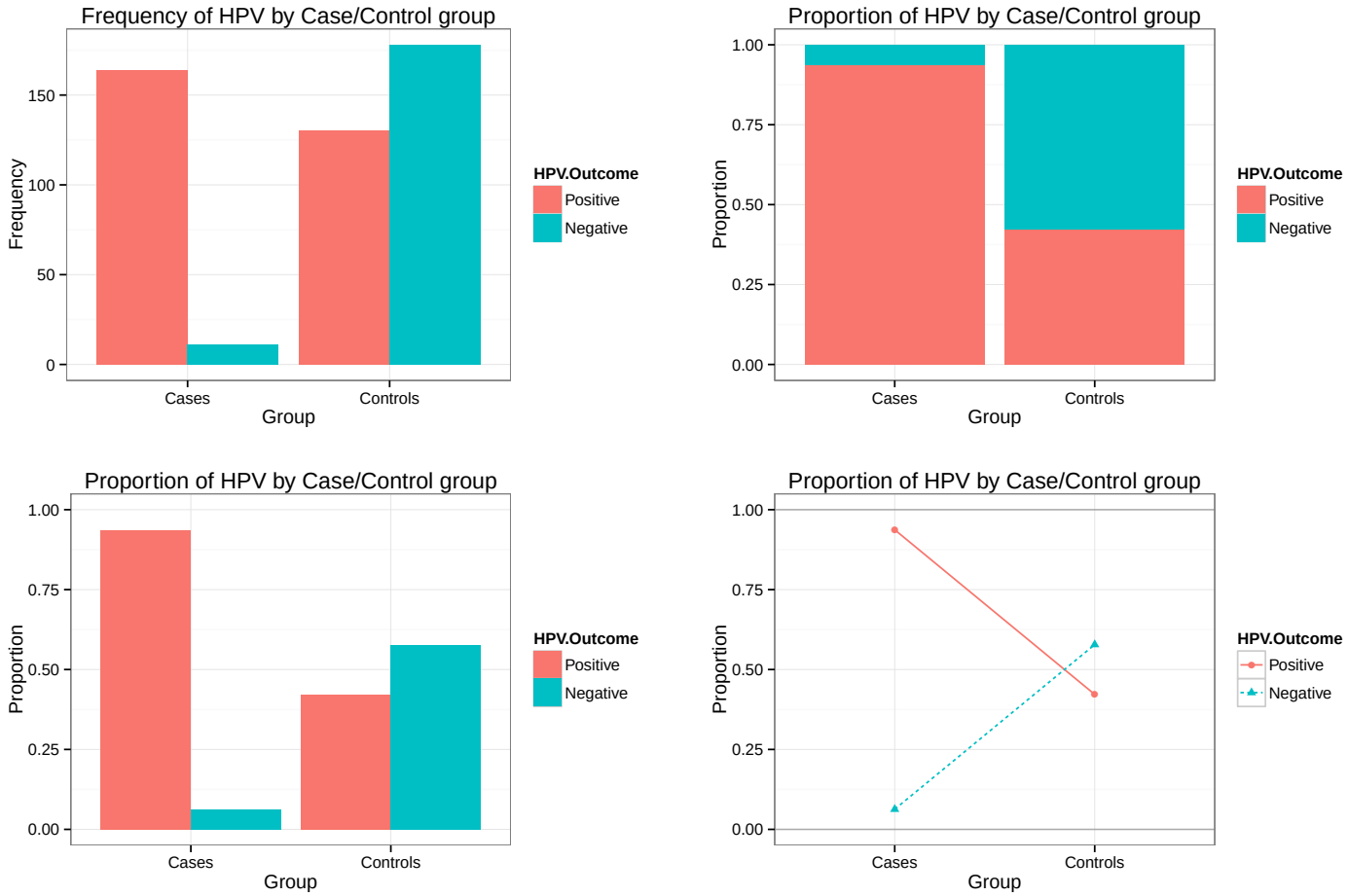
```
# plots are easier now that data are in long format.
library(ggplot2)
p <- ggplot(data = hpv.long, aes(x = Group, y = Frequency, fill = HPV.Outcome))
p <- p + geom_bar(stat="identity", position = "dodge")
p <- p + theme_bw()
p <- p + labs(title = "Frequency of HPV by Case/Control group")
print(p)

# bars, stacked
library(ggplot2)
p <- ggplot(data = hpv.long, aes(x = Group, y = Proportion, fill = HPV.Outcome))
p <- p + geom_bar(stat="identity")
p <- p + theme_bw()
p <- p + labs(title = "Proportion of HPV by Case/Control group")
print(p)

# bars, dodged
library(ggplot2)
p <- ggplot(data = hpv.long, aes(x = Group, y = Proportion, fill = HPV.Outcome))
p <- p + geom_bar(stat="identity", position = "dodge")
p <- p + theme_bw()
p <- p + labs(title = "Proportion of HPV by Case/Control group")
p <- p + scale_y_continuous(limits = c(0, 1))
print(p)

# lines are sometimes easier, especially when many categories along the x-axis
library(ggplot2)
p <- ggplot(data = hpv.long, aes(x = Group, y = Proportion, colour = HPV.Outcome))
p <- p + geom_hline(yintercept = c(0, 1), alpha = 1/4)
```

```
p <- p + geom_point(aes(shape = HPV.Outcome))
p <- p + geom_line(aes(linetype = HPV.Outcome, group = HPV.Outcome))
p <- p + theme_bw()
p <- p + labs(title = "Proportion of HPV by Case/Control group")
p <- p + scale_y_continuous(limits = c(0, 1))
print(p)
```



Returning to the hypothesis test, let p_1 be the probability that a case is HPV positive and let p_2 be the probability that a control is HPV positive. The sample sizes are $n_1 = 175$ and $n_2 = 308$. The sample proportions of positive cases and controls are $\hat{p}_1 = 164/175 = 0.937$ and $\hat{p}_2 = 130/308 = 0.422$.

For a 95% CI

$$\begin{aligned} z_{\text{crit}} SE_{CI}(\hat{p}_1 - \hat{p}_2) &= 1.96\sqrt{\frac{0.937 \times (1-0.937)}{175} + \frac{0.422 \times (1-0.422)}{308}} \\ &= 1.96 \times (0.03336) = 0.0659. \end{aligned}$$

A 95% CI for $p_1 - p_2$ is $(0.937-0.422) \pm 0.066$, or 0.515 ± 0.066, or $(0.449, 0.581)$. I am 95% confident that p_1 exceeds p_2 by at least 0.45 but not by more than

0.58.

```
# Approximate normal test for two-proportions, without Yates' continuity correction
prop.test(c(164, 130), c(175, 308), correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(164, 130) out of c(175, 308)
## X-squared = 124.29, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.4492212 0.5809087
## sample estimates:
##    prop 1    prop 2
## 0.9371429 0.4220779
```

Not surprisingly, a two-sided test at the 5% level would reject $H_0 : p_1 = p_2$. In this problem one might wish to do a one-sided test, instead of a two-sided test. Let us carry out this test, as a refresher on how to conduct one-sided tests.

```
# one-sided test, are cases more likely to be HPV positive?
prop.test(c(164, 130), c(175, 308), correct = FALSE, alternative = "greater")
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(164, 130) out of c(175, 308)
## X-squared = 124.29, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.4598071 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.9371429 0.4220779
```

Appropriateness of Large Sample Test and CI

The standard two-sample CI and test used above are appropriate when each sample is large. A rule of thumb suggests a minimum of at least five successes (i.e., observations with the characteristic of interest) and failures (i.e., observations without the characteristic of interest) in each sample before using these methods. This condition is satisfied in our two examples.

■ CLICKERQs — Comparing two proportions STT.08.02.010 ■

7.6 Effect Measures in Two-by-Two Tables

Consider a study of a particular disease, where each individual is either exposed or not-exposed to a risk factor. Let p_1 be the proportion diseased among the individuals in the exposed population, and p_2 be the proportion diseased among the non-exposed population. This population information can be summarized as a two-by-two table of population proportions:

Outcome	Exposed population	Non-Exposed population
Diseased	p_1	p_2
Non-Diseased	$1-p_1$	$1-p_2$

A standard measure of the difference between the exposed and non-exposed populations is the **absolute difference:** $p_1 - p_2$. We have discussed statistical methods for assessing this difference.

In many epidemiological and biostatistical settings, other measures of the difference between populations are considered. For example, the relative risk

$$\text{RR} = \frac{p_1}{p_2}$$

is commonly reported when the individual risks p_1 and p_2 are small. The odds ratio

$$\text{OR} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

is another standard measure. Here $p_1/(1-p_1)$ is the odds of being diseased in the exposed group, whereas $p_2/(1-p_2)$ is the odds of being diseased in the non-exposed group.

I mention these measures because you may see them or hear about them. Note that each of these measures can be easily estimated from data, using the sample proportions as estimates of the unknown population proportions. For example, in the vitamin C study:

Outcome	Ascorbic Acid	Placebo
# with cold	17	31
# with no cold	122	109
Totals	139	140

the proportion with colds in the placebo group is $\hat{p}_2 = 31/140 = 0.221$. The proportion with colds in the vitamin C group is $\hat{p}_1 = 17/139 = 0.122$.

The estimated absolute difference in risk is $\hat{p}_1 - \hat{p}_2 = 0.122 - 0.221 = -0.099$. The estimated risk ratio and odds ratio are

$$\widehat{RR} = \frac{0.122}{0.221} = 0.55$$

and

$$\widehat{OR} = \frac{0.122/(1-0.122)}{0.221/(1-0.221)} = 0.49,$$

respectively.

Interpretting odds ratios, two examples Let's begin with probability[3]. Let's say that the probability of success is 0.8, thus $p = 0.8$ Then the probability of failure is $q = 1 - p = 0.2$ The odds of success are defined as odds(success) $= p/q = 0.8/0.2 = 4$, that is, the odds of success are 4 to 1. The odds of failure would be odds(failure) $= q/p = 0.2/0.8 = 0.25$, that is, the odds of failure are 1 to 4. Next, let's compute the odds ratio by OR = odds(success)/odds(failure) $= 4/0.25 = 16$ The interpretation of this odds ratio would be that the odds of success are 16 times greater than for failure. Now if we had formed the odds ratio the other way around with odds of failure in the numerator, we would have gotten something like this, OR = odds(failure)/odds(success) $= 0.25/4 = 0.0625$ Interestingly enough, the interpretation of this odds ratio is nearly the same as the one above. Here the interpretation is that the odds of failure are one-sixteenth the odds of success. In fact, if you take the reciprocal of the first odds ratio you get $1/16 = 0.0625$.

[3]Borrowed graciously from UCLA Academic Technology Services at `http://www.ats.ucla.edu/stat/sas/faq/oratio.htm`

Another example This example is adapted from Pedhazur (1997). Suppose that seven out of 10 males are admitted to an engineering school while three of 10 females are admitted. The probabilities for admitting a male are, $p = 7/10 = 0.7$ and $q = 1 - 0.7 = 0.3$. Here are the same probabilities for females, $p = 3/10 = 0.3$ and $q = 1 - 0.3 = 0.7$. Now we can use the probabilities to compute the admission odds for both males and females, odds(male) $= 0.7/0.3 = 2.33333$ and odds(female) $= 0.3/0.7 = 0.42857$. Next, we compute the odds ratio for admission, OR $= 2.3333/0.42857 = 5.44$. Thus, the odds of a male being admitted are 5.44 times greater than for a female.

7.7 Analysis of Paired Samples: Dependent Proportions

Paired and more general **block analyses** are appropriate with longitudinal data collected over time and in medical studies where several treatments are given to the same patient over time. A key feature of these designs that invalidates the two-sample method discussed earlier is that repeated observations within a unit or individual are likely to be correlated, and not independent.

Example, President performance For example, in a random sample of $n = 1600$ voter-age Americans, 944 said that they approved of the President's performance. One month later, only 880 of the original 1600 sampled approved. The following two-by-two table gives the numbers of individuals with each of the four possible sequences of responses over time. Thus, 150 voter-age Americans approved of the President's performance when initially asked but then disapproved one month later. The row and column totals are the numbers of approvals and disapprovals for the two surveys (Agresti, 1990, p. 350).

(Obs Counts)	Second survey		
First Survey	Approve	Disapprove	Total
Approve	794	150	944
Disapprove	86	570	656
Total	880	720	1600

Let p_{AA}, p_{AD}, p_{DA}, p_{DD} be the population proportion of voter-age Americans that fall into the four categories, where the subscripts preserve the time ordering and indicate Approval or Disapproval. For example, p_{AD} is the population proportion that approved of the President's performance initially and disapproved one month later. The population proportion that initially approved is $p_{A+} = p_{AA} + p_{AD}$. The population proportion that approved at the time of the second survey is $p_{+A} = p_{AA} + p_{DA}$. The "+" sign used as a subscript means that the replaced subscript has been summed over.

Similarly, let $\hat{p}_{AA}$, $\hat{p}_{AD}$, $\hat{p}_{DA}$, $\hat{p}_{DD}$ be the sample proportion of voter-age Americans that fall into the four categories, and let $\hat{p}_{A+} = \hat{p}_{AA} + \hat{p}_{AD}$ and $\hat{p}_{+A} = \hat{p}_{AA} + \hat{p}_{DA}$ be the sample proportion that approves the first month and the sample proportion that approves the second month, respectively. The table below summarizes the observed proportions. For example, $\hat{p}_{AA} = 794/1600 = 0.496$ and $\hat{p}_{A+} = 944/1600 = 0.496 + 0.094 = 0.590$. The sample proportion of voting-age Americans that approve of the President's performance decreased from one month to the next.

(Obs Proportions)	Second survey		
First Survey	Approve	Disapprove	Total
Approve	0.496	0.094	0.590
Disapprove	0.054	0.356	0.410
Total	0.550	0.450	1.000

The difference in the population proportions from one month to the next can be assessed by a large-sample CI for $p_{A+} - p_{+A}$, given by $(\hat{p}_{A+} - \hat{p}_{+A}) \pm z_{\text{crit}} SE_{CI}(\hat{p}_{A+} - \hat{p}_{+A})$, where the CI standard error satisfies

$$SE_{CI}(\hat{p}_{A+} - \hat{p}_{+A}) = \sqrt{\frac{\hat{p}_{A+}(1-\hat{p}_{A+}) + \hat{p}_{+A}(1-\hat{p}_{+A}) - 2(\hat{p}_{AA}\hat{p}_{DD} - \hat{p}_{AD}\hat{p}_{DA})}{n}}$$

One-sided bounds are constructed in the usual way.

The $-2(\cdot)$ term in the standard error accounts for the dependence between the samples at the two time points. If independent samples of size n had been selected for the two surveys, then this term would be omitted from the standard error, giving the usual two-sample CI.

For a 95% CI in the Presidential survey,

$$\begin{aligned} z_{\text{crit}} SE_{CI}(\hat{p}_{A+} - \hat{p}_{+A}) &= 1.96\sqrt{\frac{0.590 \times 0.410 + 0.550 \times 0.450 - 2(0.496 \times 0.356 - 0.094 \times 0.054)}{1600}} \\ &= 1.96 \times (0.0095) \;=\; 0.0186. \end{aligned}$$

A 95% CI for $p_{A+} - p_{+A}$ is $(0.590 - 0.550) \pm 0.019$, or $(0.021,\ 0.059)$. You are 95% confident that the population proportion of voter-age Americans that approved of the President's performance the first month was between 0.021 and 0.059 larger than the proportion that approved one month later. This gives evidence of a decrease in the President's approval rating.

A test of $H_0 : p_{A+} = p_{+A}$ can be based on the CI for $p_{A+} - p_{+A}$, or on a standard normal approximation to the test statistic

$$z_s = \frac{\hat{p}_{A+} - \hat{p}_{+A}}{SE_{test}(\hat{p}_{A+} - \hat{p}_{+A})},$$

where the test standard error is given by

$$SE_{test}(\hat{p}_{A+} - \hat{p}_{+A}) = \sqrt{\frac{\hat{p}_{A+}\hat{p}_{+A} - 2\hat{p}_{AA}}{n}}.$$

The test statistic is often written in the simplified form

$$z_s = \frac{n_{AD} - n_{DA}}{\sqrt{n_{AD} + n_{DA}}},$$

where the n_{ij}s are the observed cell counts. An equivalent form of this test, based on comparing the square of z_s to a chi-squared distribution with 1 degree of freedom, is the well-known **McNemar's test** for marginal homogeneity (or symmetry) in the two-by-two table.

For example, in the Presidential survey

$$z_s = \frac{150 - 86}{\sqrt{150 + 86}} = 4.17.$$

The p-value for a two-sided test is, as usual, the area under the standard normal curve outside ± 4.17. The p-value is less than 0.001, suggesting that H_0 is false.

R can perform this test as McNemar's test.

```
#### Example, President performance
# McNemar's test needs data as a matrix

# Presidential Approval Ratings.
#  Approval of the President's performance in office in two surveys,
#  one month apart, for a random sample of 1600 voting-age Americans.
pres <-
matrix(c(794, 150, 86, 570),
       nrow = 2, byrow = TRUE,
       dimnames = list("1st Survey" = c("Approve", "Disapprove"),
                       "2nd Survey" = c("Approve", "Disapprove")))
pres
```

```
##             2nd Survey
## 1st Survey   Approve Disapprove
##   Approve        794        150
##   Disapprove      86        570
```

```
mcnemar.test(pres, correct=FALSE)
```

```
##
##  McNemar's Chi-squared test
##
## data:  pres
## McNemar's chi-squared = 17.356, df = 1, p-value = 3.099e-05
```

```
# => significant change (in fact, drop) in approval ratings
```

7.8 Testing for Homogeneity of Proportions

Example, cancer deaths The following two-way table of counts summarizes the location of death and age at death from a study of 2989 cancer deaths (Public Health Reports, 1983).

(Obs Counts)	Location of death			
Age	Home	Acute Care	Chronic care	Row Total
15-54	94	418	23	535
55-64	116	524	34	674
65-74	156	581	109	846
75+	138	558	238	934
Col Total	504	2081	404	2989

The researchers want to compare the age distributions across locations. A one-way ANOVA would be ideal if the actual ages were given. Because the ages are grouped, the data should be treated as categorical. Given the differences in numbers that died at the three types of facilities, a comparison of proportions or percentages in the age groups is appropriate. A comparison of counts is not.

The table below summarizes the proportion in the four age groups by location. For example, in the acute care facility $418/2081 = 0.201$ and $558/2081 = 0.268$. The **pooled proportions** are the Row Totals divided by the total sample size of 2989. The pooled summary gives the proportions in the four age categories, ignoring location of death.

The age distributions for home and for the acute care facilities are similar, but are very different from the age distribution at chronic care facilities.

To formally compare the observed proportions, one might view the data as representative sample of ages at death from the three locations. Assuming independent samples from the three locations (populations), a chi-squared statistic is used to test whether the population proportions of ages at death are identical (homogeneous) across locations. The **chi-squared test for homogeneity** of population proportions can be defined in terms of proportions, but is traditionally defined in terms of counts.

(Proportions)	Location of death			
Age	Home	Acute Care	Chronic care	Pooled
15-54	0.187	0.201	0.057	0.179
55-64	0.230	0.252	0.084	0.226
65-74	0.310	0.279	0.270	0.283
75+	0.273	0.268	0.589	0.312
Total	1.000	1.000	1.000	1.000

In general, assume that the data are independent samples from c populations (strata, groups, sub-populations), and that each individual is placed into one of r levels of a categorical variable. The raw data will be summarized as a $r \times c$ **contingency table** of counts, where the columns correspond to the samples, and the rows are the levels of the categorical variable. In the age distribution problem, $r = 4$ and $c = 3$.

To implement the test:

1. Compute the (estimated) **expected** count for each cell in the table as follows:
$$\text{E} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Total Sample Size}}.$$
2. Compute the Pearson test statistic
$$\chi^2_{\text{s}} = \sum_{\text{all cells}} \frac{(O-E)^2}{E},$$
where O is the **observed** count.
3. For a size α test, reject the hypothesis of homogeneity if $\chi^2_{\text{s}} \geq \chi^2_{\text{crit}}$, where χ^2_{crit} is the upper α critical value from the chi-squared distribution with $df = (r-1)(c-1)$.

The p-value for the chi-squared test of homogeneity is equal to the area under the chi-squared curve to the right of χ^2_{s}.

For a two-by-two table of counts, the chi-squared test of homogeneity of proportions is identical to the two-sample proportion test we discussed earlier.

The (estimated) expected counts for the (15-54, Home) cell and for the (75+, Acute Care) cell in the age distribution data are $E = 535 \times 504/2989 = 90.21$ and $E = 934 \times 2081/2989 = 650.27$, respectively. The other expected counts were computed similarly, and are summarized below. The row and column sums on the tables of observed and expected counts always agree.

(Exp Counts)	Location of death			
Age	Home	Acute Care	Chronic care	Row Total
15-54	90.21	372.48	72.31	535
55-64	113.65	469.25	91.10	674
65-74	142.65	589	114.35	846
75-	157.49	650.27	126.24	934
Col Total	504	2081	404	2989

Why is a comparison of the observed and expected counts relevant for testing homogeneity? To answer this question, first note that the expected cell count can be expressed

$$\text{E} = \text{Col Total} \times \text{Pooled proportion for category.}$$

For example, $E = 504 \times 0.179 = 90.21$ in the (15-54, Home) cell. A comparison of the observed and expected counts is a comparison of the observed category proportions in a location with the pooled proportions, taking the size of each sample into consideration. Thinking of the pooled proportion as a weighted average of the sample proportions for a category, the Pearson χ^2_{s} statistic is an aggregate measure of variation in the observed proportions across samples. If the category proportions are similar across samples then the category and pooled proportions are similar, resulting in a "small" value of χ^2_{s}. Large values of χ^2_{s} occur when there is substantial variation in the observed proportions across samples, in one or more categories. In this regard, the Pearson statistic is similar to the F-statistic in a one-way ANOVA (where large differences between groups result in a large F-statistic).

```
#### Example, cancer deaths
candeath <-
```

```
matrix(c(  94, 418,  23,
          116, 524,  34,
          156, 581, 109,
          138, 558, 238),
       nrow = 4, byrow = TRUE,
       dimnames = list("Age" = c("15-54", "55-64", "65-74", "75+"),
                  "Location of death" = c("Home", "Acute Care", "Chronic care")))
candeath
```

```
##        Location of death
## Age     Home Acute Care Chronic care
##   15-54   94        418           23
##   55-64  116        524           34
##   65-74  156        581          109
##   75+    138        558          238
```

```
chisq.summary <- chisq.test(candeath, correct=FALSE)
chisq.summary
```

```
##
##  Pearson's Chi-squared test
##
## data:  candeath
## X-squared = 197.62, df = 6, p-value < 2.2e-16
```

```
# The Pearson residuals
chisq.summary$residuals
```

```
##        Location of death
## Age           Home Acute Care Chronic care
##   15-54  0.3989527  2.3587229    -5.798909
##   55-64  0.2205584  2.5273526    -5.982375
##   65-74  1.1176594 -0.3297027    -0.500057
##   75+   -1.5530094 -3.6183388     9.946704
```

```
# The sum of the squared residuals is the chi-squared statistic:
chisq.summary$residuals^2
```

```
##        Location of death
## Age          Home Acute Care Chronic care
##   15-54 0.1591633  5.5635737    33.627351
##   55-64 0.0486460  6.3875111    35.788805
##   65-74 1.2491626  0.1087039     0.250057
##   75+   2.4118382 13.0923756    98.936922
```

```
sum(chisq.summary$residuals^2)
```

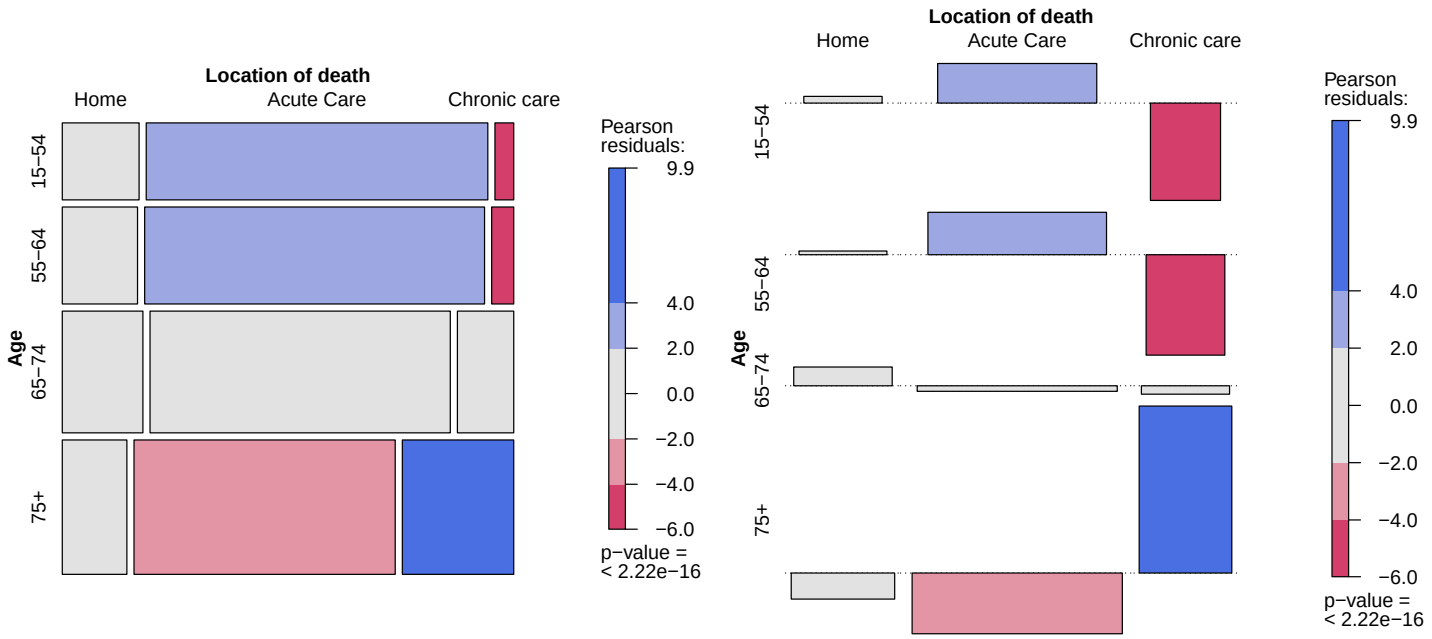
```
## [1] 197.6241
```

A visualization of the Pearson residuals is available with a `mosaic()` plot in the `vcd` package. Extended mosaic and association plots are each helpful methods of visualing complex data and evaluating deviations from a specified independence model. For extended mosaic plots, use `mosaic(x, condvar=, data=)`

where `x` is a table or formula, `condvar=` is an optional conditioning variable, and `data=` specifies a data frame or a table. Include `shade=TRUE` to color the figure, and `legend=TRUE` to display a legend for the Pearson residuals.

```
# mosaic plot
library(vcd)
mosaic(candeath, shade=TRUE, legend=TRUE)

# association plot
library(vcd)
assoc(candeath, shade=TRUE)
```



The `vcd` package provides a variety of methods for visualizing multivariate categorical data, inspired by Michael Friendly's wonderful "Visualizing Categorical Data". For more details, see The Strucplot Framework[4].

For example, a sieve plot for an n-way contingency table plots rectangles with areas proportional to the expected cell frequencies and filled with a number of squares equal to the observed frequencies. Thus, the densities visualize the deviations of the observed from the expected values.

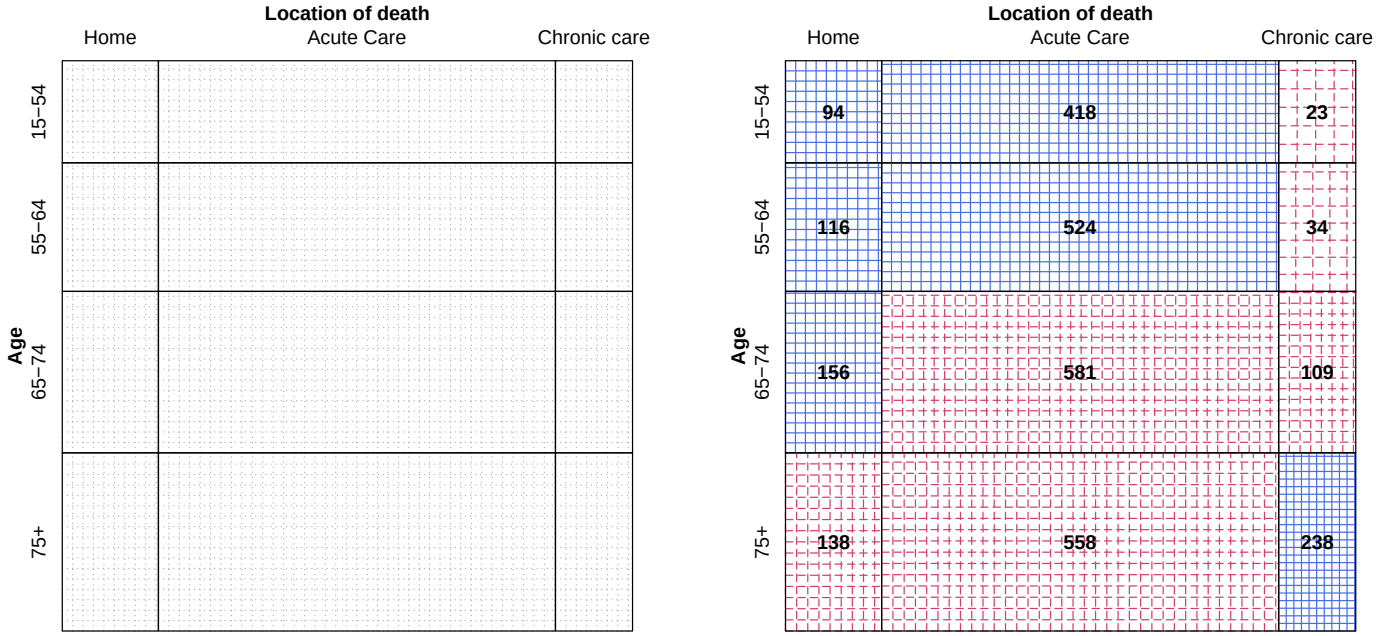
```
# sieve plot
library(vcd)

# plot expected values
```

[4] `http://cran.r-project.org/web/packages/vcd/vignettes/strucplot.pdf`

```
sieve(candeath, sievetype = "expected", shade = TRUE)

# plot observed table, then label cells with observed values in the cells
sieve(candeath, pop = FALSE, shade = TRUE)
labeling_cells(text = candeath, gp_text = gpar(fontface = 2))(as.table(candeath))
```



7.8.1 Adequacy of the Chi-Square Approximation

The chi-squared tests for homogeneity and independence are large-sample tests. As with the goodness-of-fit test, a simple rule-of-thumb is that the approximation is adequate when the expected (not observed) cell counts are 5 or more. This rule is conservative, and some statisticians argue that the approximation is valid for expected counts as small as one.

In practice, the chi-squared approximation to χ^2_s tends to be a bit conservative, meaning that statistically significant results would likely retain significance had a more accurate approximation been used.

R may not print out a warning message whenever a noticeable percentage of cells have expected counts less than 5. Ideally, one would use Fisher's exact test (`fisher.test()`) for tables with small counts, and can be used for larger than 2-by-2 tables when the frequencies are not too large.

7.9 Testing for Homogeneity in Cross-Sectional and Stratified Studies

Two-way tables of counts are often collected using either **stratified sampling** or **cross-sectional** sampling.

In a **stratified design**, distinct groups, strata, or sub-populations are identified. Independent samples are selected from each group, and the sampled individuals are classified into categories. The Indian gaming example is an illustration of a stratified design (where the two strata were NM and CO voters). Stratified designs provide estimates for the strata (population) proportion in each of the categories. A test for **homogeneity of proportions** is used to compare the strata.

In a **cross-sectional design**, individuals are randomly selected from a population and classified by the levels of **two** categorical variables. With cross-sectional samples you can test homogeneity of proportions by comparing either the row proportions or by comparing the column proportions.

Example, antismoking adverts The following data (*The Journal of Advertising*, 1983, pp. 34–42) are from a cross-sectional study that involved soliciting opinions on anti-smoking advertisements. Each subject was asked whether they smoked and their reaction (on a five-point ordinal scale) to the ad. The data are summarized as a two-way table of counts, given below:

	Str. Dislike	Dislike	Neutral	Like	Str. Like	Row Tot
Smoker	8	14	35	21	19	97
Non-smoker	31	42	78	61	69	281
Col Total	39	56	113	82	88	378

The row proportions (i.e., fix a row and compute the proportions for the column categories) are

(Row Prop)	Str. Dislike	Dislike	Neutral	Like	Str. Like	Row Tot
Smoker	0.082	0.144	0.361	0.216	0.196	1.000
Non-smoker	0.110	0.149	0.278	0.217	0.245	1.000

For example, the entry for the (Smoker, Str. Dislike) cell is: $8/97 = 0.082$. Similarly, the column proportions are

(Col Prop)	Str. Dislike	Dislike	Neutral	Like	Str. Like
Smoker	0.205	0.250	0.310	0.256	0.216
Non-smoker	0.795	0.750	0.690	0.744	0.784
Total	1.000	1.000	1.000	1.000	1.000

Although it may be more natural to compare the smoker and non-smoker row proportions, the column proportions can be compared across ad responses. There is no advantage to comparing "rows" instead of "columns" in a formal test of homogeneity of proportions with cross-sectional data. The Pearson chi-squared test treats the rows and columns interchangeably, so you get the same result regardless of how you view the comparison. However, one of the two comparisons may be more natural to interpret.

Note that checking for homogeneity of proportions is meaningful in stratified studies only when the comparison is between strata! Further, if the strata correspond to columns of the table, then the column proportions or percentages are meaningful whereas the row proportions are not.

7.9.1 Testing for Independence in a Two-Way Contingency Table

The row and column classifications for a population where each individual is cross-classified by two categorical variables are said to be **independent** if each **population** cell proportion in the two-way table is the product of the proportion in a given row and the proportion in a given column. One can show that independence is equivalent to homogeneity of proportions. In particular, the two-way table of population cell proportions satisfies independence if and only if the population column proportions are homogeneous. If the population column proportions are homogeneous then so are the population row proportions.

This suggests that a test for independence or **no association** between two variables based on a cross-sectional study can be implemented using the

chi-squared test for homogeneity of proportions. This suggestion is correct. If independence is not plausible, I interpret the dependence as a deviation from homogeneity, using the classification for which the interpretation is most natural.

The Pearson chi-squared test of independence is not significant (p-value = 0.56). The observed association between smoking status and the ad reaction is not significant. This suggests, for example, that the smoker's reactions to the ad were not statistically significantly different from the non-smoker's reactions, which is consistent with the smokers and non-smokers attitudes being fairly similar.

```
#### Example, antismoking adverts
antismokead <-
matrix(c(  8, 14,  35, 21, 19,
          31, 42,  78, 61, 69),
       nrow = 2, byrow = TRUE,
       dimnames = list(
         "Status" = c("Smoker", "Non-smoker"),
         "Reaction" = c("Str. Dislike", "Dislike", "Neutral", "Like", "Str. Like")))
antismokead
```

```
##             Reaction
## Status       Str. Dislike Dislike Neutral Like Str. Like
##   Smoker                8      14      35   21         19
##   Non-smoker           31      42      78   61         69
```

```
chisq.summary <- chisq.test(antismokead, correct=FALSE)
chisq.summary
```

```
##
##  Pearson's Chi-squared test
##
## data:  antismokead
## X-squared = 2.9907, df = 4, p-value = 0.5594
```

```
# All expected frequencies are at least 5
chisq.summary$expected
```

```
##             Reaction
## Status       Str. Dislike  Dislike  Neutral     Like Str. Like
##   Smoker         10.00794 14.37037 28.99735 21.04233  22.58201
##   Non-smoker     28.99206 41.62963 84.00265 60.95767  65.41799
```

```
# Contribution to chi-squared statistic
chisq.summary$residuals^2
```

```
##             Reaction
## Status       Str. Dislike     Dislike   Neutral         Like
##   Smoker        0.4028612 0.009545628 1.2425876 8.514567e-05
```

```
##   Non-smoker    0.1390660 0.003295110 0.4289359 2.939192e-05
##             Reaction
## Status       Str. Like
##   Smoker     0.5681868
##   Non-smoker 0.1961356
```

7.9.2 Further Analyses in Two-Way Tables

The χ^2_s statistic is a **summary measure** of independence or homogeneity. A careful look at the data usually reveals the nature of the **association** or **heterogeneity** when the test is significant. There are numerous meaningful ways to explore two-way tables to identify sources of association or heterogeneity. For example, in the comparison of age distributions across locations, you might consider the 4×2 tables comparing all possible pairs of locations. Another possibility would be to compare the proportion in the 75+ age category across locations. For the second comparison you need a 2×3 table of counts, where the two rows correspond to the individuals less than 75 years old and those 75+ years old, respectively (i.e., collapse the first three rows of the original 4×2 table). The possibilities are almost limitless in large tables. Of course, theoretically generated comparisons are preferred to data dredging.

Example: drugs and nausea, testing for homogeneity A randomized double-blind experiment compared the effectiveness of several drugs in reducing postoperative nausea. All patients were anesthetized with nitrous oxide and ether. The following table shows the incidence of nausea during the first four postoperative hours of four drugs and a placebo. Compare the drugs to each other and to the placebo.

Drug	# with Nausea	# without Nausea	Sample Size
Placebo	96	70	166
Chlorpromazine	52	100	152
Dimenhydrinate	52	33	85
Pentobarbitol (100mg)	35	32	67
Pentobarbitol (150mg)	37	48	85

Let p_{PL} be the probability of developing nausea given a placebo, and define p_{CH}, p_{DI}, p_{PE100}, and p_{PE150} analogously. A simple initial analysis would be to test homogeneity of proportions: $H_0 : p_{PL} = p_{CH} = p_{DI} = p_{PE100} = p_{PE150}$ against H_A : not H_0.

The data were entered as frequencies. The output shows that the proportion of patients exhibiting nausea (see the **column** percents — the cell and row percentages are not interpretable, so they are omitted) is noticeably different across drugs. In particular, Chlorpromazine is the most effective treatment with $\hat{p}_{CH} = 0.34$ and Dimenhydrinate is the least effective with $\hat{p}_{DI} = 0.61$.

The p-value for the chi-squared test is 0.00005, which leads to rejecting H_0 at the 0.05 or 0.01 levels. The data strongly suggest there are differences in the effectiveness of the various treatments for postoperative nausea.

```
#### Example: drugs and nausea, testing for homogeneity
nausea <-
matrix(c(96,  70,   52, 100,   52,  33,   35,  32,   37,  48),
       nrow = 5, byrow = TRUE,
       dimnames = list("Drug" = c("PL", "CH", "DI", "PE100", "PE150"),
                       "Result" = c("Nausea", "No Nausea")))
nausea
```

```
##        Result
## Drug    Nausea No Nausea
##   PL        96        70
##   CH        52       100
##   DI        52        33
##   PE100     35        32
##   PE150     37        48
```

```
# Sorted proportions of nausea by drug
nausea.prop <- sort(nausea[,1]/rowSums(nausea))
nausea.prop
```

```
##        CH     PE150     PE100        PL        DI
## 0.3421053 0.4352941 0.5223881 0.5783133 0.6117647
```

```
# chi-sq test of association
chisq.summary <- chisq.test(nausea, correct=FALSE)
chisq.summary
```

```
##
##  Pearson's Chi-squared test
##
## data:  nausea
## X-squared = 24.827, df = 4, p-value = 5.451e-05
```

```
# All expected frequencies are at least 5
chisq.summary$expected
```

```
##        Result
## Drug      Nausea No Nausea
##   PL    81.35495  84.64505
##   CH    74.49369  77.50631
##   DI    41.65766  43.34234
##   PE100 32.83604  34.16396
##   PE150 41.65766  43.34234
```

A sensible follow-up analysis is to identify which treatments were responsible for the significant differences. For example, the placebo and chlorpromazine can be compared using a test of $p_{PL} = p_{CH}$ or with a CI for $p_{PL} - p_{CH}$.

In certain experiments, specific comparisons are of interest, for example a comparison of the drugs with the placebo. Alternatively, all possible comparisons might be deemed relevant. The second case is suggested here based on the problem description. I will use a Bonferroni adjustment to account for the multiple comparisons. The Bonferroni adjustment accounts for data dredging, but at a cost of less sensitive comparisons.

There are 10 possible comparisons here. The Bonferroni analysis with an overall Family Error Rate of 0.05 (or less) tests the 10 individual hypotheses at the 0.05/10=0.005 level.

```
nausea.table <- data.frame(Interval = rep(NA,10)
                         , CI.lower = rep(NA,10)
                         , CI.upper = rep(NA,10)
                         , Z        = rep(NA,10)
                         , p.value  = rep(NA,10)
                         , sig.temp = rep(NA,10)
                         , sig      = rep(NA,10))
# row names for table
nausea.table[,1] <- c("p_PL - p_CH"
                    , "p_PL - p_DI"
                    , "p_PL - p_PE100"
                    , "p_PL - p_PE150"
                    , "p_CH - p_DI"
                    , "p_CH - p_PE100"
                    , "p_CH - p_PE150"
                    , "p_DI - p_PE100"
                    , "p_DI - p_PE150"
                    , "p_PE100 - p_PE150")
# test results together in a table
i.tab <- 0
for (i in 1:4) {
  for (j in (i+1):5) {
    i.tab <- i.tab + 1
    nausea.summary <- prop.test(nausea[c(i,j),], correct = FALSE, conf.level = 1-0.05/10)
    nausea.table[i.tab, 2:6] <- c(nausea.summary$conf.int[1]
                                , nausea.summary$conf.int[2]
                                , sign(-diff(nausea.summary$estimate)) * nausea.summary$statistic^0.5
                                , nausea.summary$p.value
                                , (nausea.summary$p.value < 0.05/10))
    if (nausea.table$sig.temp[i.tab] == 1) { nausea.table$sig[i.tab] <- "*" }
                                      else { nausea.table$sig[i.tab] <- " " }
  }
}
```

```
# remove temporary sig indicator
nausea.table <- subset(nausea.table, select = -sig.temp)
#nausea.table
```

The following table gives two-sample tests of proportions with nausea and 99.5% CIs for the differences between the ten pairs of proportions. The only two p-values are less than 0.005 corresponding to $p_{PL} - p_{CH}$ and $p_{CH} - p_{DI}$. I am 99.5% confident that p_{CH} is between 0.084 and 0.389 less than p_{PL}, and I am 99.5% confident that p_{CH} is between 0.086 and 0.453 less than p_{DI}. The other differences are not significant.

	Interval	CI.lower	CI.upper	Z	p.value	sig
1	p_PL - p_CH	0.0838	0.3887	4.2182	0.0000	*
2	p_PL - p_DI	−0.2167	0.1498	−0.5099	0.6101	
3	p_PL - p_PE100	−0.1464	0.2582	0.7788	0.4361	
4	p_PL - p_PE150	−0.0424	0.3284	2.1485	0.0317	
5	p_CH - p_DI	−0.4532	−0.0861	−4.0122	0.0001	*
6	p_CH - p_PE100	−0.3828	0.0222	−2.5124	0.0120	
7	p_CH - p_PE150	−0.2788	0.0924	−1.4208	0.1554	
8	p_DI - p_PE100	−0.1372	0.3160	1.1058	0.2688	
9	p_DI - p_PE150	−0.0352	0.3881	2.3034	0.0213	
10	p_PE100 - p_PE150	−0.1412	0.3154	1.0677	0.2857	

Using ANOVA-type groupings, and arranging the treatments from most to least effective (low proportions to high), we get:

```
CH (0.34)   PE150 (0.44)   PE100 (0.52)   PL (0.58)   DI (0.61)
-----------------------------------
            ---------------------------------------------------
```