

Chapter 7: Categorical data

Previously we looked at comparing means and medians for quantitative variables from one or more groups.

We can think of these problems as having a quantitative response variable with a categorical predictor variable, which is the group or treatment variable (such as placebo vs. treatment A vs treatment B).

In this section, we consider when all variables are categorical. An example might be college major vs political affiliation. A typical null hypothesis for this type of data is that there is no association between the two variables. For this example, the null hypothesis might be that the proportions of students supporting the Democrat, Republican, Libertarian, and Green parties are the same for different majors: psychology, biology, statistics, history, etc.

Categorical data: a motivating example

An interesting historical example are the data of survivors from the Titanic shipwreck of 1912. Passengers from the event are classified as male or female, child or adult, and 1st, 2nd, or 3rd class. In addition, there were adult crew members. Here there are four variables: sex, age, and class (including crew members), and survival (yes or no). Sex, age, and survival have two levels each, and class has four levels (crew + three levels for ticket class).

```
library(datasets)
data(Titanic)
Titanic
```

Categorical data

, , 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
------	-----	---

Categorical data

, , 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
------	-----	----

Categorical data

The data can be reshaped into a single narrow format for the three variables, plus a frequency for the number of times each combination occurs.

```
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
```

Categorical data

We usually like to think of statistical inferences as being based on collecting data in order to test a certain scientific hypothesis. In this case, we have the data first. What kinds of questions can be asked about this data? One question is whether survival probability depended on age (child or non-child), passenger status and/or class (crew member vs passenger, 1st class vs 3rd class, etc.) or on sex.

Categorical data: inference for a proportion

We'll return to the Titanic example later. We'll start with making inferences for a single proportion. That is, we have a single yes/no or 0-1 variable, and we wish to know what proportion of the population is a yes. Examples include

- ▶ survival for patients receiving a certain treatment, such as a transplant
- ▶ proportion of college students graduating within five years
- ▶ proportion of high school students going to college
- ▶ proportion of likely voters planning to vote for candidate X
- ▶ proportion of registered voters who will actually vote
- ▶ proportion of products made at a factory that are defective (or will be returned)
- ▶ proportion of free throws made successfully by a basketball player

Categorical data: inference for a proportion

For a sample with n items, the best guess for the population proportion is:

$$\hat{p} = \frac{\# \text{ yes}}{n}$$

This is particularly convenient if you record “yes” as 1 and “no” as 0. Then the sample mean is also the sample proportion.

Categorical data: inference for a proportion

Often for proportions, we want a confidence interval (CI) for the proportion. The theory is based on the binomial distribution and the assumption that for large enough samples, the proportion is roughly normally distributed. Letting \hat{p} be the sample proportion and z_{crit} be the critical value for the standard normal distribution, the normal approximation CI is

$$\hat{p} \pm z_{crit} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Categorical data: inference for a proportion

For a 95% interval, z_{crit} is chosen so that the area in the middle of the standard normal distribution is 95% of the entire distribution. This means we use the value corresponding to the 97.5% quantile. This can be obtained in R as

```
qnorm(.975)  
#[1] 1.959964
```

which is often rounded to 1.96. Thus the 95% interval is usually computed as

$$\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Categorical data: inference for a proportion

Note that the CI follows the usual approach for CIs:

$$\text{best guess} \pm \text{critical value} \times SE$$

where here

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

Categorical data: inference for a proportion

A rule of thumb is the margin of error is close to $1/\sqrt{n}$ when $p \approx 0.5$. Why is this?

The function $f(p) = p(1 - p)$ is an upside-down parabola (like a frowny face) which has a maximum at $p = 0.5$. When $p = .5$, and using $z_{crit} \approx 2$, and $\hat{p} \approx p$, we have

$$z_{crit} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \approx 2\sqrt{(1/2)(1/2)/n} = 2(1/2)\sqrt{1/n} = \sqrt{1/n}$$

Some useful values of n to keep in mind are say, $n = 100, 400, 900$. Then the margin of error is about 10%, 5% and 3.3%, respectively.

Political opinion polls, or polls for election candidates often use sample sizes of around 900 to 1200 respondents. The margin of error (i.e., $z_{crit}SE$) of 3.5% or so is often reported for these polls, which corresponds to a bit more than 900 respondents. Margins of error can be a bit more complicated for more complex kinds of polling such as stratified polling, cluster sampling etc. See STAT572 for more on this topic...

Categorical data: inference for a proportion

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 95% CI is

$$0.7 \pm 1.96\sqrt{(0.7)(0.3)/270} = 0.7 \pm 0.055 = (0.645, 0.755)$$

This means that you are 95% confident that between 64.5% and 75.5% of consumers would be willing to pay extra for the packaging. The population here is consumers in 1983, so this proportion might have changed over time.

Categorical data: inference for a proportion

To get a confidence interval in R, you can use the function `prop.test()`. The default null hypothesis is $H_0 : p = 0.5$, which is not very meaningful for this problem. We are more interested in the confidence interval.

The syntax is `prop.test(x,n,p)`. Here x is the number of “successes”, n is the number of trials, and p is the probability of success under the null hypothesis, which is 0.5 by default, or you can change it.

```
prop.test(189,270,correct=F)
# 1-sample proportions test without continuity correction
#
#data:  189 out of 270, null probability 0.5
#X-squared = 43.2, df = 1, p-value = 4.942e-11
#alternative hypothesis: true p is not equal to 0.5
#95 percent confidence interval:
# 0.6428459 0.7515429
```

Categorical data: inference for a proportion

The function `prop.test()` actually uses a somewhat different formula than the usual one we presented, and it is supposed to have slightly better performance. There are quite a few variations on formulas for confidence intervals for proportions, as referenced in

Newcombe R.G. (1998) Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods. *Statistics in Medicine* **17**, 857–872.

The formula used in R (without continuity correction) is method 3 from the Newcombe reference, also called the Wilson score interval:

$$CI = \frac{2n\hat{p} + z^2}{2(n + z^2)} \pm \frac{z\sqrt{z^2 + 4n\hat{p}(1 - \hat{p})}}{2(n + z^2)}$$

where z is the critical value, or 1.96 for a 95% interval. The interval is centered at 0.5 for $\hat{p} = 0.5$ and otherwise is centered at a value in between 0.5 and \hat{p} , and closer to \hat{p} for larger n .

Categorical data: alternative CIs

Another formula for the CI for a proportion is named the Agresti-Coull estimator. It's formula is

$$\tilde{n} = n + z^2$$

$$\tilde{p} = \frac{1}{\tilde{n}} \left(X + \frac{1}{2} z^2 \right)$$

$$CI = \tilde{p} \pm z \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n}}$$

where z is a critical value such as 1.96 for a 95% interval. If $z = 2$ is used as the critical value, the interval has the interpretation that it is equivalent to the standard confidence interval using $z_{crit} \sqrt{p(1 - p)/n}$ but treating the data as having observed two successes and two failures. This biases the interval to be centered towards 0.5, but this bias is very small for large sample sizes.

Categorical data: alternative CIs

Why have statisticians come up with so many confidence intervals for proportions?

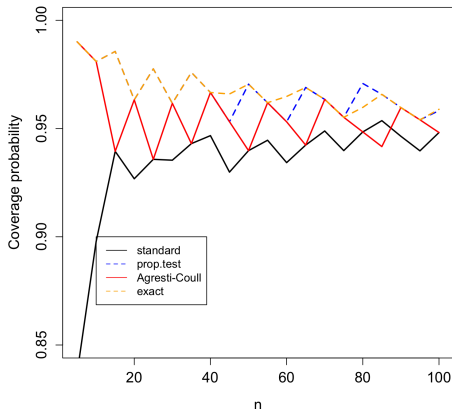
Categorical data: alternative CIs

The performance of a particular formula for a CI can be judged using simulations. How do you measure how good a CI is? Just like a good statistical method for hypothesis testing should have a low type I error rate (level α), a good confidence interval should capture the population parameter most of the time, ideally 95% of the time for a 95% CI.

The 95% CIs for proportions are usually based on the idea that for large n , the sample proportion \hat{p} is approximately normally distributed with mean p and variance $p(1 - p)/n$. However, this approximation is just an approximation, and works better for large sample sizes, i.e., large n . The relation to the normal distribution is based on the Central Limit Theorem, but it is not clear how good the approximation is for small samples. This leaves room for potentially finding better formulas for smaller values of n .

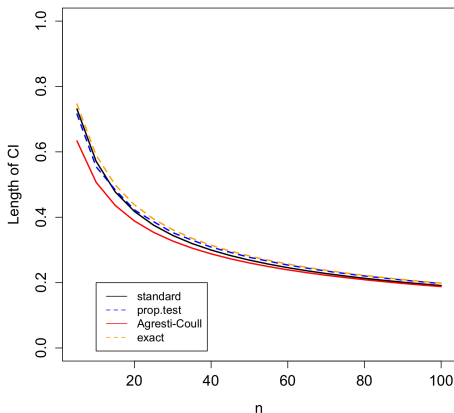
Categorical data: alternative CIs

We'll do a little simulation to look at how often the three confidence interval procedures capture the population proportion for different sample sizes. We see that the standard CI has the worst coverage probability.



Categorical data: alternative CIs

Another measure of performance for CIs is their width.



Categorical data: alternative CIs

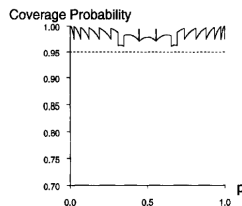
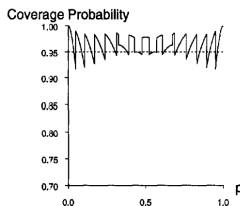
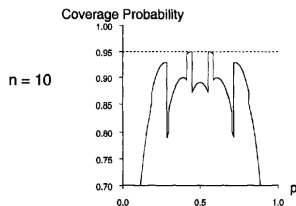
The Agresti-Coull interval is the narrowest, but as the sample size increases, all estimators getting increasingly precise (as you would expect), and all have about the same width.

Agresti and Coull wrote a widely cited paper in 1998 (in *The American Statistician*) investigating performance of estimators and made some pretty pictures using theory to get actual (instead of simulated) coverage probabilities for small sample sizes. In the next slide, “Wald” refers to the standard interval based on $z_{crit} \sqrt{\hat{p}(1 - \hat{p})/n}$. “Adjusted Wald” is the Agrestic-Coull interval, and exact is based on the exact binomial distribution, which leads to conservative confidence intervals.

A takeaway message from all of this is to take confidence levels, like p-values, with a grain of salt. A 95% confidence interval might not have exactly 95% coverage probability. And p-values are based on distributional assumptions that might just be approximate.

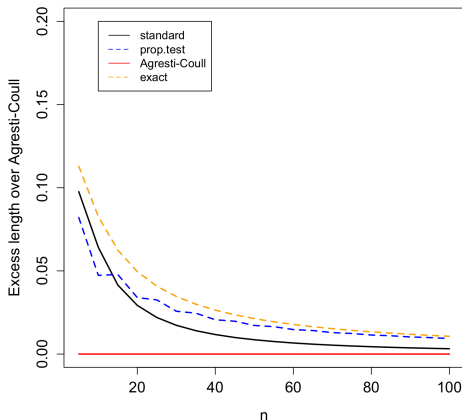
Categorical data: alternative CIs

Coverage probabilities for Wald (standard), Adjusted Wald (Agresti-Coull) and exact CIs for $n = 10$ from Agresti and Coull (1998).



Categorical data: alternative CIs

The graph might not look like much, but other methods can have larger CIs than Agresti-Coull:



Hypothesis testing for a proportion

To do hypothesis testing for a proportion, we can test

$$H_0 : p = p_0$$

against alternatives such as

$$H_A : p < p_0, \quad H_A : p \neq p_0, \text{ or } \quad H_A : p > p_0$$

The test statistic in this case is a z-score

$$z_{obs} = \frac{\hat{p} - p_0}{SE}$$

where

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

Note that the SE error used here is different from the standard (Wald) CI because we use p_0 from the null hypothesis instead of using \hat{p} . In fact, the Wilson score CI is based on finding a CI that includes p whenever you reject the null hypothesis. The standard (Wald) CI could occasionally fail to include p_0 even when you would not reject H_0 based on the test statistic. The Wilson CI corrects this issue.

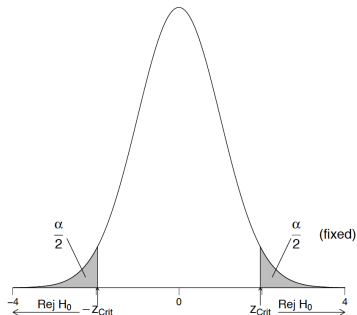
Hypothesis testing for a proportion

Hypothesis testing is then just based on comparing z_{obs} to a critical value, or comparing the p -value to α , as usual. As usual, you can just an R function such as `prop.test()` to do the hypothesis test.

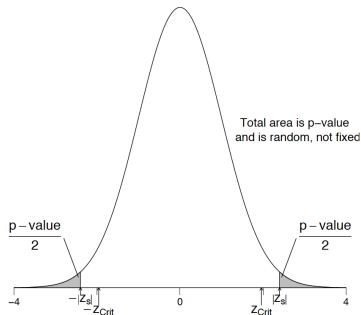
For small sample sizes, you might also consider using an exact binomial test, which uses probabilities from a binomial probability distribution instead of using a normal approximation. This is done using `binom.test()` in R, but works very similarly.

Hypothesis testing for a proportion

Z-distribution with two-sided size $\alpha = .05$ critical region



Z-distribution with two-sided p-value



Hypothesis testing for a proportion

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? Answer this question by computing a p-value for a one-sided test.

Hypothesis testing for a proportion

The null hypothesis is $H_0 : p = 0.25$ and the alternative is $H_A : p > 0.25$. In the formulas, $p_0 = 0.25$, the hypothesized value. We have $\hat{p} = 22/53 = 0.415$ and

$$z_{obs} = \frac{0.415 - 0.25}{\sqrt{(0.25)(0.75)/53}} = 2.78$$

For this one-sided test, the p-value is the area to the right of 2.78 under a standard normal curve. In R, this is

```
1-qnorm(2.78)
#[1] 0.002717945
```

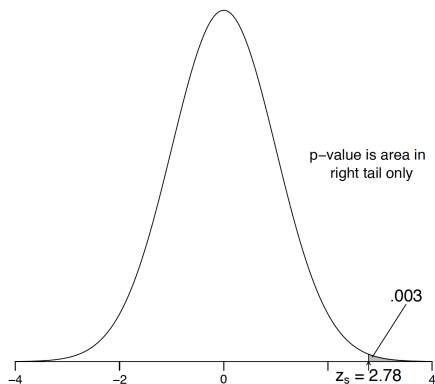
A p-value of 0.0027 means that there is sufficient evidence to reject the hypothesis that this population has the same proportion of right-side heavier brains than the general population.

Hypothesis testing for a proportion

```
#### 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
```

Hypothesis testing for a proportion

Right brain upper one-sided p-value



Comparing two proportions

Just like we can testing whether two populations have the same mean by comparing two independent samples, we can also test whether two populations have the same proportion by comparing two independent samples.

In addition, we can form confidence intervals for the difference in two proportions.

Comparing two proportions

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.

Comparing two proportions

To estimate the difference in proportions, we can use $\hat{p}_1 - \hat{p}_2$ as the best estimate for the difference in proportions. For a confidence interval, we follow the usual pattern of

$$\text{best estimate} \pm \text{critical value} \times \text{SE}$$

Here the critical value is the usual z_{crit} and the SE is

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

Comparing two proportions

To do a hypothesis test, we use a z-score with

$$z_{obs} = \frac{\hat{p}_1 - \hat{p}_2}{SE_{test}}$$

where

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

where

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

Here \bar{p} is a pooled estimate of the common proportion—i.e., assuming that both populations have the same proportion, you could pool all the data treating it as a single population to get a more precise estimate of the proportion $p = p_1 = p_2$.

Comparing two proportions

Because the two formulas for computing the SE are different, it is possible that the conclusions reached by hypothesis testing might not agree with whether or not the CI includes 0, but such cases would be unusual.

Comparing two proportions

To get back to the sample, we can perform both a CI and a hypothesis test. It is convenient to arrange the data in a 2×2 table.

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

Here $\hat{p}_1 = 17/139 = 0.122$, $\hat{p}_2 = 31/140 = 0.221$, and $\bar{p} = 48/279 = 0.172$, where the numerator is the total number of colds, and the denominator is the total sample size.

Comparing two proportions

The test statistic is

$$\hat{p}_1 - \hat{p}_2 = 0.122 - 0.221 = -0.99$$

The SE values are

$$SE_{CI} = \sqrt{\frac{0.221 \times (1 - 0.22)}{140} + \frac{0.122 \times (1 - 0.122)}{139}} = 0.04472$$

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

Comparing two proportions

The CI is

$$-0.99 \pm 1.96(0.04472) = -0.99 \pm 0.088 = (-0.187, -0.011)$$

The z-test gives

$$z_{obs} = \frac{0.122 - 0.221}{0.0452} = -2.19$$

The p-value is

```
2*pnorm(-2.19)
#[1] 0.02852424
```

The area is multiplied by two because it is a two-sided test.

Comparing two proportions

Both the CI and the hypothesis test suggest that there is evidence against the null hypothesis that the proportion of skiers getting colds was the same for those on placebo versus those on Vitamin C. Because this was a randomized study, this suggests that Vitamin C had a causal role in preventing in colds.

We might notice that the effect was moderate, with the CI meaning that a plausible effect for Vitamin C was that it made skiers between 1% and 18.7% less likely to develop a cold. At the time the study was controversial, and many studies since have not found an association between Vitamin C use and cold prevention.

Comparing two proportions

```
#### 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
```


Comparing two proportions within the same sample

Many surveys might have more than two categories. For example, during the 2016 election, registered voters might have indicating favoring Clinton, Trump, or another candidate. A typical poll before the election might have had Clinton at 48% and Trump at 44%. These numbers add up to 92%, not 100%, and the remaining 8% might have been due to undecided voters or due to voters supporting Gary Johnson (Libertarian), Jill Stein (Green) or other less well known candidates.

In this case, we might be interested in comparing whether 48% is “significantly” larger than 44%. The idea is whether Clinton could have felt reasonably confident that she had more support than Trump, or whether the sample was consistent with no difference in support in the population between the two candidates.

This is not a two sample problem because we are trying to compare two proportions from within the same sample. The two proportions are not independent. In particular, it would be impossible for both percentages to be larger than 50%.

Comparing two proportions within the same sample

Just to make the example simple, suppose the poll had 1000 respondents, with 480 supporting Clinton, 440 supporting Trump, 50 supported Johnson, and 30 supported Stein. How could we analyze the data to determine whether Clinton had a lead that was “statistically significant”?

Comparing two proportions within the same sample

Probably the simplest thing to do would be to think of the sample as having 920 respondents, so that you condition on respondents supporting either Clinton or Trump, and ignore everyone else. In this case, Clinton's proportion would be $\hat{p}_1 = 480/920 = 0.522$. In other words, 52% of people supporting either Clinton or Trump only supported Clinton in this poll. You then might test whether this proportion was significantly larger than 50%.

```
prop.test(480,920)
#
# 1-sample proportions test with continuity correction
#
#data: 480 out of 920, null probability 0.5
#X-squared = 1.6533, df = 1, p-value = 0.1985
#alternative hypothesis: true p is not equal to 0.5
#95 percent confidence interval:
# 0.4888946 0.5543999
```

Comparing two proportions within the same sample

The margin of error in this case is approximately

$$1.96\sqrt{0.522(1 - 0.522)/920} = 0.032$$

meaning 3.2%, which is very typical in these types of polls.

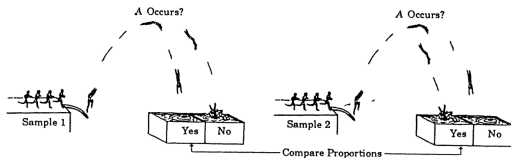
This margin of error is not for the 48%, but for the 52% conditional on only Trump or Clinton in the sample. It would also not apply to the other categories in the sample (Johnson and Stein).

Comparing two proportions within the same sample

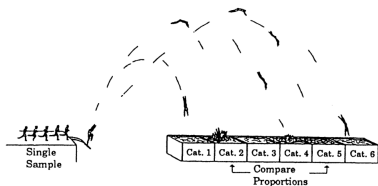
Another approach is to use something called a **multinomial distribution**. The multinomial distribution generalizes the binomial distribution to when there are more than two categories. In the binomial distribution, there are two categories: yes and no, or success and failure.

The multinomial distribution allows any finite number of categories and counts how many observation are in each category. The idea is that each observation is assumed to be independent, and belongs to category i with probability p_i , where $p_1 + p_2 + \cdots + p_k = 1$, and k is the number of categories.

Comparing two proportions within the same sample



Situation (a) *Two independent samples.*



Situation (b) *Single sample, several response categories.*

Comparing two proportions within the same sample

An example of a multinomial sample would be a bag of M&M candy (assuming that the pieces are independent. Here you have say, n pieces of candy, with n_1 brown, n_2 yellow, n_3 green, n_4 orange, and say n_5 blue (assuming there are only 5 colors. The sample proportions are

$$\hat{p}_1 = \frac{n_1}{n}, \hat{p}_2 = \frac{n_2}{n}, \hat{p}_3 = \frac{n_3}{n}, \hat{p}_4 = \frac{n_4}{n}, \hat{p}_5 = \frac{n_5}{n}$$

For the political candidate example, we have $\hat{p}_{Clinton} = \frac{n_{Clinton}}{n}$, $\hat{p}_{Trump} = \frac{n_{Trump}}{n}$, etc.

Comparing two proportions within the same sample

To get a CI, from the properties of the multinomial distribution, the idea is to compute (as usual)

$$(\hat{p}_1 - \hat{p}_2) \pm z_{crit} SE(\hat{p}_1 - \hat{p}_2)$$

What is different is the calculation of the standard error, which has to take into account the lack of independence between the two samples. This is done using a **covariance** term, which we won't go into. (For details, see “Difference of Proportions from the Same Survey” by Scott and Seber, *The American Statistician*, 1983.) The formula is (for a 95% interval):

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96 \sqrt{\frac{\hat{p}_1 + \hat{p}_2 - (\hat{p}_1 - \hat{p}_2)^2}{n}}$$

Comparing two proportions within the same sample

For the Clinton-Trump example. We get

$$(0.48 - 0.44) \pm 1.96 \sqrt{\frac{.48 + .44 - (.48 - .44)^2}{1000}} = 0.04 \pm 0.059 \approx (-.02, .10)$$

Note that if this sample has a reported margin of error of 3.2%, you might be misled into thinking that 48% is significantly larger than 44%. Either the conditional analysis or the multinomial analysis lead to the conclusion that Clinton's lead was actually within the margin of error.

Goodness of fit tests

Another use of the multinomial distribution is to test whether proportions in several categories are significantly different from expectation.

An example is to test whether jury selection is random with respect to age, where ages are grouped into discrete intervals. For example, if jury selection is random, then the proportion of jurors selected between ages 30 and 39 should reflect the proportion of adults who are between ages 30 and 39.

Goodness of fit tests

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.

Goodness of fit tests

Are the observed proportions reasonable given the census proportions?

Age	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

Goodness of fit tests

The null hypothesis can be described as being that the proportion for each category matches a null proportion for that category. In this case, we might have as a null hypothesis that the proportion of each age group in the jury pool match the population proportions for the same age categories. This could be written as

$$H_0 : p_1 = p_{01}, p_2 = p_{02}, \dots, p_7 = p_{07}$$

The alternative is that at least one of the proportions doesn't match the hypothesized proportion.

Goodness of fit tests

The idea for the test statistic is to compare the observed counts with the expected counts (observed versus expected number of individuals in each category): The test statistic for this problem is (with k categories):

$$\chi_{obs}^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

In terms of proportions, this can be written as

$$\chi_{obs}^2 = n \sum_{i=1}^k \frac{(\hat{p}_i - p_{0i})^2}{p_{0i}}$$

Goodness of fit tests

The test statistic looks different from others you have seen so far, but if you think of it this way

$$\chi^2_{obs} = \sum_{i=1}^k \left(\frac{O_i - E_i}{\sqrt{E_i}} \right)^2$$

Then the terms

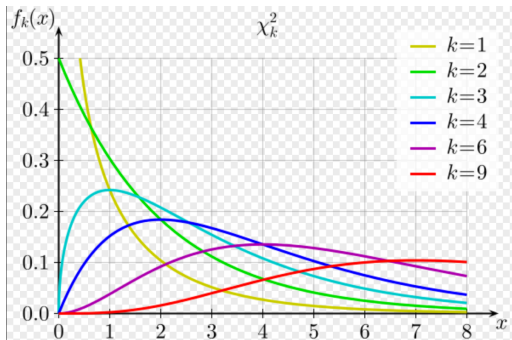
$$Z_i = \frac{O_i - E_i}{\sqrt{E_i}}$$

look like Z-scores and are also called category residuals:

$$Z = \frac{x - \mu}{\sqrt{\sigma^2}}$$

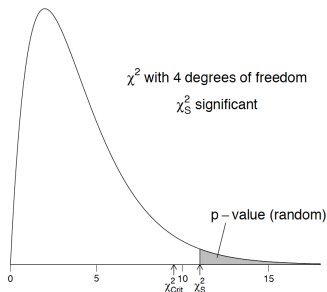
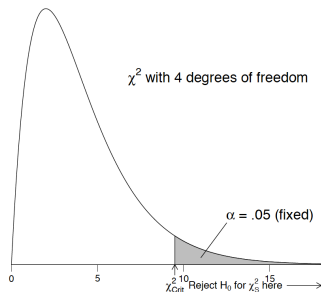
Goodness of fit tests

Under the null hypothesis, the statistic χ^2_{obs} has a chi-square distribution with $k - 1$ degrees of freedom.



Goodness of fit tests

Large values of χ^2_{obs} suggest that the proportions differ from expectation. We reject H_0 for sufficiently large values of χ^2_{obs} .



Goodness of fit tests

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 $H_A : \text{not } H_0$, using the sample of 1336 from the jury pool.

Goodness of fit tests

As an example of how to compute the expected count, we have

$$E_{18} = 1336 \times 0.061 = 81.5$$

The observed number from this category is the observed proportion times the sample size of 1336:

$$O_{18} = 1336 \times 0.017 \approx 23$$

The contribution to the χ^2_{obs} statistic from this category is therefore

$$(23 - 81.5)^2 / 81.5 = 41.99$$

The Pearson χ^2 statistic is

$$\chi^2_{obs} = (-6.48)^2 + (-7.37)^2 + (-3.45)^2 + 0.18^2 + 6.48^2 + 8.78^2 + (-1.99)^2 = 231.26$$

Goodness of fit tests

The critical value for the χ^2 test at the $\alpha = 0.05$ level can be obtained from R by

```
qchisq(1-.05,6)
#[1] 12.59159
```

Gernally, the expected value of a χ^2 random variable is it's degrees of freedom, and you will reject H_0 if the χ_{obs}^2 is much larger than the degrees of freedom. In this case, our observed test statistic is 231.26, which is much larger than the critical value of 12.59, so there is strong evidence that the ages in the jury pool have a different distribution than that of the general population.

Goodness of fit tests

To use the p-value approach instead of the critical value, you can use

```
1-pchisq(231.26,6)
#[1] 0
```

Here the p-value is so small that it is 0 to many decimal places.

Goodness of fit tests

Here is a way to implement this in R:

```
#### 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)
```

Goodness of fit tests

Here is a way to implement this in R:

```
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
```

Goodness of fit tests

To see how much individual categories deviate from the null hypothesis, it is helpful to look at the residuals. Here is a way to make a 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
```


Goodness of fit tests

Recall that the test statistic was $\chi^2_{obs} \approx 231$. The greatest contribution to this score was category 6, with a contribution of 77. This is for 50–64 year olds. It's residual was 8.77 ($8.77^2 \approx 77$), which is positive, meaning that it had an excess of jury members. It's expected number of jurors (out of 1336) was 243, but 380 were in that age category.

On the other hand, 18-19 year olds had an expected representation of about 81 jurors, but only 23 jurors were of that age. This resulted in a negative residual of -6.48, and a contribution of 41.99 to the χ^2_{obs} statistic. The categories of 30–39 year olds and those aged 65 and over were the most consistent with the null hypothesis.

In some cases, you might see that one category has much larger deviation from expectation than other categories. In this case, quite a few categories deviated quite strongly from expectation under the null hypothesis.

Goodness of fit tests

A follow up to the χ^2 test is to get confidence intervals for the proportions in each category. Here is an example of doing this for the age of jurors data. Here a confidence level of $1 - .05/7$ is used for Bonferroni adjustments.

```
b.sum1 <- prop.test(jury$Count[1], sum(jury$Count),  
p = jury$CensusProp[1], conf.level=1-.05/7)  
b.sum2 <- prop.test(jury$Count[2], sum(jury$Count),  
p = jury$CensusProp[2], conf.level=1-.05/7)  
b.sum3 <- prop.test(jury$Count[3], sum(jury$Count),  
p = jury$CensusProp[3], conf.level=1-.05/7)  
b.sum4 <- prop.test(jury$Count[4], sum(jury$Count),  
p = jury$CensusProp[4], conf.level=1-.05/7)  
b.sum5 <- prop.test(jury$Count[5], sum(jury$Count),  
p = jury$CensusProp[5], conf.level=1-.05/7)  
b.sum6 <- prop.test(jury$Count[6], sum(jury$Count),  
p = jury$CensusProp[6], conf.level=1-.05/7)  
b.sum7 <- prop.test(jury$Count[7], sum(jury$Count),  
p = jury$CensusProp[7], conf.level=1-.05/7)
```

Goodness of fit tests

```
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$CensusProp <- jury$CensusProp
```

Goodness of fit tests

```
> b.sum
```

	p.value	CI.lower	CI.upper	Age	CensusProp
1	3.362577e-11	0.009647384	0.03018145	18-19	0.061
2	1.709175e-15	0.054740455	0.09367530	20-24	0.150
3	2.410326e-04	0.079962727	0.12501521	25-29	0.135
4	8.636174e-01	0.190060970	0.25162323	30-39	0.217
5	2.579364e-12	0.192891624	0.25474777	40-49	0.153
6	4.126666e-22	0.252099261	0.31911186	50-64	0.182
7	3.953815e-02	0.065941074	0.10777784	65-99	0.102

Goodness of fit tests

Here is an example from genetics. In this case, a little bit of biology is needed to get the expected counts or proportions. (The example is from Falconer and Mackay, *Quantitative Genetics*).

For the M-N blood group, a sample of individuals from Iceland is given with the following genotype counts, both observed and expected:

	Genotypes			Total
	MM	MN	NN	
Observed	233	385	129	747
Expected	242.36	366.26	138.38	747
$(O_i - E_i)^2 / E_i$	0.362	0.959	0.634	1.956

Goodness of fit tests

To compute the expected values, we need a little bit of genetics here. The idea is that each person has two alleles, so there are 233 individuals with two copies of the M allele, 385 individuals with one M and one N, and 129 with two Ns. The total number of individuals is $233 + 385 + 129 = 747$, but the total number of allele copies is $747 * 2 = 1494$.

Goodness of fit tests

The total number of copies of M in the sample is $233 \times 2 + 385$, because each person with MM contributes two Ms to the sample, and each person with MN contributes one M. Thus the total number of Ms is $233 * 2 + 385 = 851$. The total number of N alleles in the sample is $385 + 2 \times 129 = 643$. The probability of a random allele being M or N is

$$P(M) = \frac{851}{851 + 643} = 0.5696, \quad P(N) = \frac{643}{851 + 643} = 0.4304$$

The expected proportion of individuals with genotype MM is $P(M) \times P(M) = 0.3245$ (assuming that alleles are random), with MN is $2P(M)P(N) = 0.4903$ and with NN is $P(N) \times P(N) = 0.1852$. The expected counts are these expected proportions multiplied by the sample size of 747 individuals. For example, $747 \times 0.3245 \approx 242.36$.

Goodness of fit tests

```
> o <- c(233, 385, 129)
m <- 233*2+385
n <- 385+2*129
pm <- m/(m+n)
pn <- n/(m+n)
eMM <- pm^2*sum(o)
e <- c(pm^2, 2*pm*pn, pn^2)*sum(o)
chisq <- sum((o-e)^2/e)
#[1] 1.955521
      (o-e)^2/e
#[1] 0.3622291 0.9588085 0.6344837
```


Goodness of fit tests

To use R to conduct the test, we need to specify the expected proportions. Because the p-value is greater than 0.05, we conclude that there is not sufficient evidence to reject the hypothesis that the proportions are consistent with their expected values.

```
a <- chisq.test(o,p=c(pm^2,2*pm*pn,pn^2))
a
#
# Chi-squared test for given probabilities
#
#data:  o
#X-squared = 1.9555, df = 2, p-value = 0.3762
```

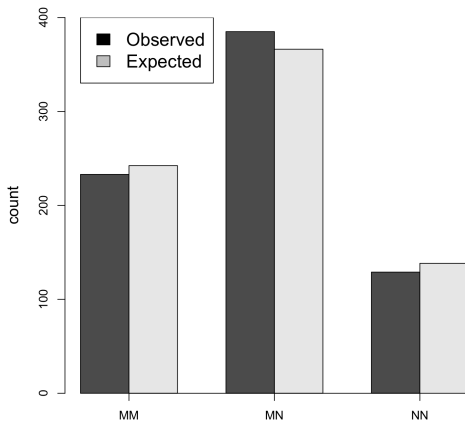
Goodness of fit tests

A nice way to visualize what is happening is to plot the observed versus expected counts. This can be done by making a matrix where each row is the frequency count. In this case, the rows are for observed and expected.

```
counts <- c(o,e)
counts <- matrix(counts,ncol=3,byrow=T)
barplot(counts,beside=TRUE,ylim=c(0,400),ylab="count",
cex.lab=1.3)
legend(1,400,legend=c("Observed","Expected"),
fill=c("black","grey"),cex=1.5)
axis(1,at=c(2,5,8),c("MM","MN","NN"),cex=1.5)
```

Goodness of fit tests 172.3

Comparison of observed versus expected counts.



Goodness of fit tests

A final thing to be aware of is that a chi-square test applied to a 2×2 table is equivalent to a two-sample test of proportions. Here we'll compare numbers of voters for Clinton and Trump in two polls, one for ABC and one for Fox. The ABC poll had 48% support for Clinton and 44% for Trump. The Fox poll had 44% for Clinton and 41% for Trump. The ABC poll was from 10-23-16 to 10-26-16 and had 1109 respondents, but only about 1020 for either Clinton or Trump. The Fox poll was from 10-22-16 to 10-25-16 and had 1221 respondents, but only about 1123 who were for Clinton or Trump. The polls can be summarized as follows

Candidate	ABC	Fox	Total
Clinton	532	537	1069
Trump	488	501	989
Total	1020	1038	2058

Goodness of fit tests

A χ^2_{obs} test answers the question of whether there is an association between the two variables: candidate supported versus poll. If there are differences, there are different possible explanations, including the agency conducting the poll (probably contracted by the media company), the slightly different dates of the polls, and uncertainty due to random sampling.

Goodness of fit tests

For a chi-square test on a 2×2 table, you can enter the table as a matrix. For `prop.test()`, give a vector of counts for one category, and a vector of sample sizes for independent samples.

```
x <- c(532,537,488,501)
x <- matrix(x,ncol=2,byrow=T)
chisq.test(x,correct=FALSE)
# Pearson's Chi-squared test
#X-squared = 0.036834, df = 1, p-value = 0.8478
#
prop.test(c(532,537),c(1020,1038),correct=FALSE)
#
2-sample test for equality of proportions without continuity
#
#data:  c(532, 537) out of c(1020, 1038)
#X-squared = 0.036834, df = 1, p-value = 0.8478
```

Goodness of fit tests

Both the independent proportions test and χ^2 test lead to exactly same conclusion, namely that the ABC and Fox polls did not have significantly different proportions of respondents supporting Clinton versus Trump conditional on respondents supporting either Clinton or Trump. The fact that the two analyses lead to the same conclusion (and p-value) is related to the fact that the square of a standard normal random variable has a χ^2 distribution with one degree of freedom.

Goodness of fit tests

The default for the χ^2 test in R is to use the continuity correction, sometimes called Yates' continuity correction. Here the formula is slightly different:

$$\chi^2_{Yates} = \sum_{i=1}^k \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

This formula is intended to make the χ^2 test more conservative when sample sizes are low. However, based on simulations, it seems that the χ^2 test without continuity correction already tends to be conservative when sample sizes are low, meaning that it is difficult to reject the null hypothesis for small sample sizes. Consequently, I tend not to use the continuity correction option.

Goodness of fit tests

A common rule of thumb is to say that the χ^2 test is reasonable when all expected cell counts are at least 5. Here is more specific advice: “No more than 20% of the expected counts are less than 5 and all individual expected counts are 1 or greater” (Yates, Moore, & McCabe, *The Practice of Statistics*, 1999, p. 734).

If your sample sizes are small for certain categories, R is likely to print out warnings, but will still compute results.

Goodness of fit tests

Generally for a χ^2 test of association, if you have two variables where variable 1 has r levels, and variable 2 has c levels, then you can make an $r \times c$ contingency table (think r for rows and c for columns). The χ^2 test then tests whether there is an association between the two variables. The degrees of freedom for this type of test is $(r - 1) \times (c - 1)$.

Goodness of fit tests

As an example, consider again the data for ABC versus Fox polls. This time, we can use all five categories: Clinton, Trump, Johnson, Stein, and Other. Our contingency table is then 5×2 (or you can arrange it to be 2×5).

Company	Clinton	Trump	Johnson	Stein	Other	Total
ABC	532	488	44	11	34	1109
Fox	537	501	85	37	61	1221
Total	1069	989	129	48	95	2330

Expected counts can be obtained by taking the row total times the column total divided by the overall total. For example, the expected count for Clinton and ABC is $1109 \times 1069 / 2330 = 508.8073$.

Goodness of fit tests

```
o <- c(532,488,44,11,34,537,501,85,37,61)
o <- matrix(o,byrow=T,ncol=5)
a <- chisq.test(o,correct=FALSE)
a
# Pearson's Chi-squared test
#X-squared = 29.667, df = 4, p-value = 5.721e-06
a$expected
#           [,1]    [,2]    [,3]    [,4]    [,5]
#[1,] 508.8073 470.73 61.39957 22.84635 45.21674
#[2,] 560.1927 518.27 67.60043 25.15365 49.78326
1109*1069/2330
#[1] 508.8073
```

Goodness of fit tests

Based on the χ^2 test, there is a significant association between the support for the different candidates and whether the poll was done by Fox or ABC. To get an idea of which categories deviate most from expectation, we can look at the contribution to the χ^2 test. The χ^2 test statistic is the sum of the squared residuals.

```
# also try (o-a$expected)^2/a$expected
a$residuals
#           [,1]      [,2]      [,3]      [,4]      [,5]
#[1,]  1.028193  0.7959858 -2.220526 -2.478427 -1.668080
#[2,] -0.979902 -0.7586007  2.116235  2.362023  1.589736
(a$residuals)^2
#           [,1]      [,2]      [,3]      [,4]      [,5]
#[1,]  1.057181  0.6335933  4.930736  6.142602  2.782492
#[2,]  0.960208  0.5754750  4.478449  5.579153  2.527259
```

Goodness of fit tests

The biggest contributions to the χ^2 statistic come from columns 3 and 4 in the residuals or squared residuals. Here the contribution to the χ^2 statistic is greater than 4 for each category. These columns are for the support for Johnson and Stein. The critical value for the χ^2 test here is based on $(2 - 1) \times (5 - 1) = 4$ degrees of freedom, and is

```
qchisq(.95,4)
#[1] 9.487729
```

2×2 tables and conditional probability

Contingency tables, particularly 2×2 tables, can be a good way to understand conditional probabilities.

Here is an example for a 2×2 table relating sex and migraine headaches, based on a survey of 50 women and 50 men:

	Migraines		Total
	Yes	No	
Women	12	38	50
Men	5	45	50
Total	17	83	100

Conditional probability

We can ask questions like:

- ▶ what is the probability that a person experienced a migraine given that they were a woman?
- ▶ What is the probability that someone was a man, given that they experienced a migraine?

From probability theory, we can use

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

when individuals are sampled randomly, we can think of this as

$$P(A|B) = \frac{\# \text{ in category A and category B}}{\# \text{ in category B}}$$

2×2 tables and conditional probability

	Migraines		Total
	Yes	No	
Women	12	38	50
Men	5	45	50
Total	17	83	100

- ▶ $P(\text{woman}|\text{migraine}) = \frac{12}{17} \approx 0.71$
- ▶ $P(\text{migraine}|\text{man}) = \frac{5}{50} = 0.1$

2×2 tables

For this example, if you wanted to test whether the proportion of men versus women experiencing migraines was different for the population that was sampled, you could use either a test of proportions or a χ^2 test. Using `prop.test()` has the advantage that it also gives a CI for the difference:

Note the differences between

```
prop.test(c(12,38),c(17,83),correct=FALSE)  
prop.test(c(12,5),c(50,50),correct=FALSE)
```

Both commands will result in the exact same p-value, but have different interpretations in terms of the CI. The first tests whether the proportion of women is the same among those experiencing migraines versus those not experiencing migraines. The second tests whether the proportion experiencing migraines is different for women versus men. The first uses the proportions 12/17 and 38/83, while the second compares 12/50 versus 5/50. But both are also equivalent to the χ^2 test of whether there is an association between sex and migraines.

```
prop.test(c(12,5),c(50,50),correct=FALSE)
```

2-sample test for equality of proportions without continuity
correction

data: c(12, 5) out of c(50, 50)

X-squared = 3.4727, df = 1, p-value = 0.06239

alternative hypothesis: two.sided

95 percent confidence interval:

-0.004666056 0.284666056

sample estimates:

prop 1 prop 2

0.24 0.10

2×2 tables

Here hypothesis test shows weak evidence of a difference, with a p-value greater than 0.05. At the $\alpha = .05$ level, there is insufficient evidence to conclude that there is a difference in proportions experiencing migraines at the population level. However, the p-value is close to .05. The confidence barely included 0, with difference ranging from women experiencing between -0.5% to 28% more migraines than men.

2×2 tables with correlated observations

Sometimes 2×2 tables are presented with correlated observations. For example, if a sample of people is first given treatment A, asked to record whether or not symptoms occur, and then later the same group is given treatment B, again asked to record whether the symptoms occur, then you have two sets of proportions but they are correlated rather than independent samples.

For a nonmedical example, you could track a sample of individual's response, for example to who they plan to vote for, over time to see if same individuals change their opinion. This is different from taking a second sample of an independent set of individuals.

2×2 tables with correlated observations

Here an example looks at presidential approval (example from Agresti, 1990, *Categorical Data Analysis*). The data is from a sample of 1600 voter-aged people who were asked whether they approved or disapproved of the president two times separated by one month. Of course, someone who approves is likely to continue to approve one month later, and someone who disapproves is likely to disapprove one month later. What is interesting is how many people change their minds. Here is the data:

First Survey	Second survey		total
	approve	disapprove	
approve	794	150	944
disapprove	86	570	656
total	880	720	1600

2×2 tables with correlated observations

The diagonals on the table represent people who's opinion of the president didn't change. The off-diagonals represent those who changed their mind. We see that 150 people changed from approval to disapproval, while 86 people changed from disapproval to approval. This means that the approval rating went down for this sample. The approval ratings started at $944/1600 = 59\%$ from the first survey to $880/1600 = 55\%$ in the second survey.

From independent samples each with a sample size of 1600, we would check whether 59% is significantly different from 55%. (This would give a p-value of .022). In this case, we have the same individuals changing their mind, and we can instead check, among those who changed their mind, are the 150 who went from approval to disapproval significantly larger than the 86.

2×2 tables with correlated observations

In McNemar's test, let

- ▶ \hat{p}_{AA} be the proportion that approved both times
- ▶ \hat{p}_{DD} be the proportion that disapproved both times
- ▶ \hat{p}_{AD} be the proportion that approved only the first time
- ▶ \hat{p}_{DA} be the proportion that approved only the second time
- ▶ \hat{p}_{A+} be the proportion that approved the first month (i.e., $944/1600 = 59\%$)
- ▶ \hat{p}_{+A} be the proportion that approved the second month (i.e., $880/1600 = 55\%$)

Then a CI for the difference is

$$\hat{p}_{A+} - \hat{p}_{+A} \pm z_{crit} SE$$

where

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

2×2 tables with correlated observations

Plugging in the numbers, we get for the SE

$$\sqrt{\frac{(.59)(.41) + (.55)(.45) - 2((.496)(.356) - (.094)(.054))}{1600}} = .0095$$

The 95% CI is

$$(0.59 - 0.55) \pm 0.019 = (0.021, 0.059)$$

Thus an estimate of the change in approval rating is that it decreased by between 2 and 6% for the population represented by this sample.

For a hypothesis test, we use

$$z_{obs} = \frac{\hat{p}_{A+} - \hat{p}_{+A}}{SE} = \frac{n_{AD} - n_{DA}}{\sqrt{n_{AD} + n_{DA}}}$$

where

$$SE = \sqrt{\frac{\hat{p}_{A+}\hat{p}_{+A} - 2\hat{p}_{AA}}{n}}$$

correlated 2×2 tables

The formula for the test statistic in terms of the counts is nice and easy to interpret. The numerator is the difference in the off-diagonals, and the denominator is the square root of the sum of the off-diagonals. An interesting thing about this test is that the numbers on the diagonals make no difference to the test. We are only interested in those people who changed their minds, and whether among those who changed their minds, the direction of the change was larger in one direction than the other.

The hypothesis test is implemented in R, but I'm not sure if the CI is:

```
x <- c(794,150,86,570) # try x <- c(1,150,85,10000) results are tl
x <- matrix(x,byrow=T,ncol=2)
mcnemar.test(x)
# McNemar's Chi-squared test with continuity correction
#
#data:  x
#McNemar's chi-squared = 16.818, df = 1, p-value = 4.115e-05
```

correlated 2×2 tables

It is interesting to compare the results if the data had been two independent samples, each of size 1600, with the observed proportions of 59% and 55%. In this case we can analyze the data this way in R

```
o <- c(944,656,880,720)
o <- matrix(o,ncol=2,byrow=T)
prop.test(o,correct=FALSE)
prop.test(o,correct=FALSE)

#X-squared = 5.2224, df = 1, p-value = 0.0223
#alternative hypothesis: two.sided
#95 percent confidence interval:
# 0.005721636 0.074278364
#sample estimates:
#prop 1 prop 2
#  0.59   0.55
```

correlated 2×2 tables

In medical examples, we often interpret the relative risk and odds ratio.

Outcome	Exposed population	non-exposed population
Diseased	p_1	p_2
Non-diseased	$1 - p_1$	$1 - p_2$

In this case

$$RR = p_1/p_2$$

is called the relative risk and is the probability of disease in the exposed population divided by the probability in the non-exposed population.

The odds of having the disease for the exposed population is $p_1/(1 - p_1)$.

The odds of having the disease for the non-exposed population is $p_2/(1 - p_2)$. The odds ratio is

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

You can also think of it as the ratio of relative risks.