

Overview

- **Section 10.1** Proportions — Goodness-of-fit
- **Section 10.2** 2×2 tables — test of equality of population proportions
- **Section 10.3** 2×2 tables — test of independence of categorical variables
- **Section 10.4** 2×2 tables — Fisher's exact test
- **Section 10.5** $r \times k$ tables
- **Section 10.6** Applicability — when are the methods valid?
- **Section 10.7** Confidence intervals for differences in proportions
- **Section 10.8** 2×2 tables — paired data
- **Section 10.9** Relative risk and Odds ratios

χ^2 Test of Goodness-of-Fit

The method to test consistency of the genetic model follows that of other hypothesis tests seen in the course. Namely we will

- State hypotheses;
- Calculate a test statistic;
- Compare the test statistic to its null distribution;
- Compute a p-value; and
- Interpret the results in the context of the problem.

The logic is the same as all other hypothesis tests we have seen. Namely, small p-values are evidence against the null hypothesis while p-values that are not small indicate that the data is consistent with the null hypothesis.

The particular equations are different for the special setting we consider here.

Analysis of Categorical Data

Bret Larget

Department of Statistics

University of Wisconsin - Madison

November 10, 2003

Goodness-of-fit

Example:

Under a genetic model, a cross of white and yellow summer squash will yield progeny of colors white, yellow, and green with probabilities $12/16$, $3/16$ and $1/16$ respectively.

(It is more common in genetics to say that the expected ratios are 12:3:1.)

Suppose we observe the following data.

Color	Number of Offspring
white	155
yellow	40
green	10

Is this consistent with the genetic model?

The χ^2 Test Statistic

The χ^2 test statistic is a measure of discrepancy between the observed category counts and what is expected if the null hypothesis is true.

$$X^2 = \sum_{i \in \text{categories}} \frac{(O_i - E_i)^2}{E_i}$$

where the sum goes over the categories, O_i is the observed count in the i th category and E_i is the expected count in the i th category according to the null hypothesis.

Note that if the null hypothesis is true, each O_i has a binomial distribution with the same n but different p s. The binomial random variables are not independent because their sum must be n .

If p_i is the null probability of category i , then the expected value is just the mean of the corresponding binomial distribution, $E_i = np_i$.

Comments on the Equation

The denominator is np so the expected value of each term in the expression is $np(1-p)/(np) = 1-p$.

It follows that the expected value of the χ^2 test statistic is the sum of the expected values of each term.

$$\begin{aligned} E(X^2) &= \sum_i E\left(\frac{(O_i - E_i)^2}{E_i}\right) \\ &= \sum_i (1 - p_i) \\ &= (\text{number of categories}) - 1 \end{aligned}$$

Stating Hypotheses:

The data we observe is the category of each individual, summarized by a count of individuals in each category.

The null hypothesis specifies the probability that an individual is in each category.

$$H_0: p_{\text{white}} = \frac{12}{16}, \quad p_{\text{yellow}} = \frac{3}{16}, \quad p_{\text{green}} = \frac{1}{16}$$

The alternative hypothesis is that the probabilities are something different.

$$H_A: \text{the probabilities are different}$$

Comments on the Equation

$$X^2 = \sum_{i \in \text{categories}} \frac{(O_i - E_i)^2}{E_i}$$

Notice that X^2 cannot be negative. Larger values indicate larger discrepancy between what is observed and what is expected.

The numerator of each term is a squared difference between the observed value and the expected value. The denominator scales these squared differences by the size of the expected values.

The expected value of the numerator is the variance of a binomial random variable, namely $np(1-p)$ where p is the null probability of the category.

Sampling Distribution

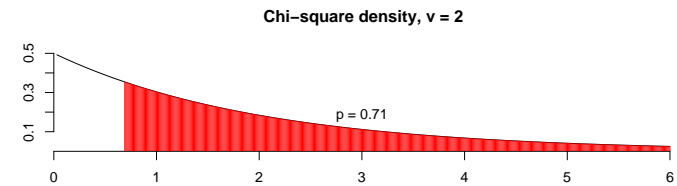
The χ^2 test statistic follows (approximately) a χ^2 distribution with v degrees of freedom under the null hypothesis where $v = (\text{number of categories}) - 1$.

- The χ^2 distribution with v degrees of freedom is the distribution of the **sum of v squared independent standard normal random variables**. Its mean is v .
- The density is not symmetric and is supported only on the non-negative real numbers.
- For $v = 1$, the density is strongly right-skewed and has a shape for which the density gets infinitely large for values close to 0 (but the area under the curve is 1).
- For $v = 2$, the density decreases exponentially.
- For $v > 2$, the density is 0 at 0 and is unimodal.
- As v increases, the mean of the density increases, the skewness decreases, and the shape becomes more symmetric and normal.

Example — Sampling Distribution

- In the example, there are three categories and so two degrees of freedom.
- The test statistic would need to be substantially larger than 2 to be statistically significant.
- Here is R code to compute the p-value.

```
> 1 - pchisq(0.691, 2)
[1] 0.7078663
```
- Here is a graphical depiction of the p-value.



Example — Calculation

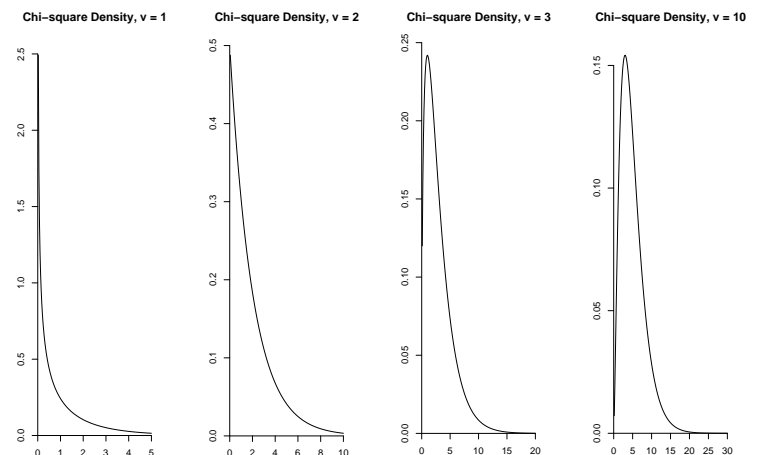
The following table contains the calculation of the test statistic.

Color	Observed	Expected	χ^2
white	155	$205 \times 12/16 = 153.75$	0.010
yellow	40	$205 \times 3/16 = 38.4375$	0.064
green	10	$205 \times 1/16 = 12.8125$	0.691
Total	205	205	0.691

It can be useful to keep track of the individual components of the χ^2 test statistic.

These values indicate which categories contribute the most to the discrepancy between what is observed and what is expected.

Graphs of χ^2 Distributions



2 × 2 tables, Test of Independence

Example:

Does heavy cell-phone use promote brain tumors?

If so, we might expect brain tumors to be more likely to develop on the side of the head where the cell phone is held.

Data is from a [retroactive study](#) on subject with brain tumors who used a cell-phone at least six months prior to detection of the tumor.

		Phone holding side		
		Left	Right	Total
Brain tumor side	Left	14	28	42
	Right	19	27	46
	Total	33	55	88

Example—Hypotheses

Let p_i be the population probability that a tumor develops on the left side of the brain where population 1 are those who hold the phone with the left hand and population 2 hold it with the right. (Note, we could just have easily used right instead of left.)

The null hypothesis is $H_0: p_1 = p_2$.

A directional alternative makes sense here if we think that the tumor is more likely to be close to the hand holding the cell phone. $H_A: p_1 > p_2$.

Example — Interpretation

The large p-value indicates that the data is consistent with the null hypothesis.

In the context of the problem,

There is no reason to doubt the genetic model for the color probabilities from the cross of white and yellow summer squash. The observed data is consistent with the expected 12:3:1 ratios. The difference between observed and expected values is easily explained by chance variation.

Example

The [explanatory variable](#) is the side of the head where the phone was held. The [response variable](#) is the side of the head where the tumor was.

Of the 33 subjects who held the phone on the left, 14 developed a tumor on the left. The sample proportion is $14/33 = 0.424$.

Of the 55 subjects who held the phone on the right, 28 developed a tumor on the left. The sample proportion is $28/55 = 0.509$.

Can the difference between these sample proportions be explained by chance variation?

Example—Test Statistic

so the expected number of individuals who use the cell phone on the left and get a brain tumor on the left is

$$88 \times \frac{33}{88} \times \frac{42}{88} = \frac{33 \times 42}{88} = 15.8$$

Notice the simple general formula.

$$(\text{Expected count in cell } ij) = \frac{(\text{sum of row } i) \times (\text{sum of column } j)}{(\text{table total})}$$

Here are the observed and (expected) counts.

		Observed		Total
		Phone holding side		
		Left	Right	
Brain tumor side	Left	14 (15.75)	28 (26.25)	42
	Right	19 (17.25)	27 (28.75)	46
Total		33	55	88

Plugging in to the χ^2 test statistic formula, we get $X^2 = 0.6$.

P-value—One-sided test

The p-value we computed is in fact the p-value for a two-sided test. For a one-sided test, we need to check to see if the sample proportions are in the correct direction. If so, the one-sided p-value would be half as big, as the two-sided p-value. However, in this case, a higher percentage of right ear cell phone users develop tumors on the left side of the brain than do those that use cell phones with the left hand. So the one-sided p-value is greater than 1/2.

Example—Test Statistic

We will again use the χ^2 test statistic, but for a 2×2 table we have a [different formula for finding the expected values for each cell](#) and a [different formula for the degrees of freedom](#).

If cell phone side and tumor side are independent, we would have

$$\Pr\{\text{cell phone left AND tumor left}\} = \Pr\{\text{cell phone left}\} \times \Pr\{\text{tumor left}\}$$

with a similar relationship for each cell in the table.

We can estimate the probabilities from marginal values of the table.

$$\Pr\{\text{cell phone left}\} \approx \frac{\# \text{ cell phone left}}{\text{total}} = \frac{33}{88},$$

and

$$\Pr\{\text{tumor left}\} \approx \frac{\# \text{ tumor left}}{\text{total}} = \frac{42}{88}$$

Example—Sampling Distribution

For a test of independence from a 2×2 table, the degrees of freedom is one.

In general, for arbitrary sized tables, the degrees of freedom is

$$\text{degrees of freedom} = (\# \text{ rows} - 1)(\# \text{ columns} - 1)$$

The area to the right of 0.6 under a χ^2 distribution with 1 degree of freedom is computed in R as follows.

```
> e11 = 33 * 42/88
> e12 = 55 * 42/88
> e21 = 33 * 46/88
> e22 = 55 * 46/88
> e = c(e11, e12, e21, e22)
> o = c(14, 28, 19, 27)
> x2 = sum((o - e)^2/e)
> pval = 1 - pchisq(x2, 1)
> pval
[1] 0.4404272
```

χ^2 Tests in R

Here is the built-in way to carry out a χ^2 test of independence in R. The first test is as in the textbook. The second uses a continuity correction that recognizes that the exact sampling distribution of the test statistic is only a discrete approximation of the continuous χ^2 distribution.

```
> x = matrix(c(14, 19, 28, 27), 2, 2)
> x
      [,1] [,2]
[1,]  14  28
[2,]  19  27
> chisq.test(x, correct = F)
      Pearson's Chi-squared test

data:  x
X-squared = 0.5952, df = 1, p-value = 0.4404
> chisq.test(x)
      Pearson's Chi-squared test with Yates' continuity correction

data:  x
X-squared = 0.3037, df = 1, p-value = 0.5816
```

Statistics 371, Fall 2003

19

Fisher's Exact Test

Fisher's exact test for 2×2 tables is a nonparametric test based on putting a probability distribution on the set of tables with the same marginal totals.

The p -value is then the probability of obtaining a table at least as extreme as that actually obtained. It is typically easiest to conceptualize this for directional hypotheses.

The probability distribution is the [hypergeometric](#) distribution. This is the distribution that arises by sampling without replacement from a bucket of colored balls and counting balls of a given color.

Fisher's Exact Test is actually a type of [permutation test](#) where we can think of permuting the category labels of one category.

Statistics 371, Fall 2003

20

Example—Interpretation

There is no evidence that the hand with which a cell phone is typically used is associated with the side where brain tumors occur. Specifically, there is no evidence that use of a cell phone on one side increases the risk of developing a brain tumor on that side of the brain.

Statistics 371, Fall 2003

18

χ^2 Tests in R

The `matrix` command creates a 2 by 2 matrix with the elements listed by columns. By default, R will use the correction for continuity in the `chisq.test` command. The option `correct=F` overrides the default behavior and computes the test statistic as in the textbook.

Statistics 371, Fall 2003

19

Example — More extreme tables

	No Shot	Flu Shot	Total
Flu? Yes	16	2	18
No	12	11	23
Total	28	13	41

	No Shot	Flu Shot	Total
Flu? Yes	17	1	18
No	11	12	23
Total	28	13	41

	No Shot	Flu Shot	Total
Flu? Yes	18	0	18
No	10	13	23
Total	28	13	41

Example — Colored Ball Model

For Fisher's Exact test applied to this example, consider a bucket of 41 balls, 18 of which are red and 23 of which are white. The 18 balls represent the people in the sample who will get the flu and the 23 represent the people who will not get the flu. In the actual data, the people are partitioned into groups of 28 people (those with no flu shot), and 13 (who took the flu shot).

If we drew 13 colored balls at random without replacement (and left 28 in the bucket), the probability of the actual table would be

$$\frac{\binom{18}{3}\binom{23}{10}}{\binom{41}{13}} = 0.052983$$

where $\binom{n}{j}$ is an alternative notation to the binomial coefficient ${}_n C_j$.

Example—Flu Shots

The following contingency table shows the relationship between flu incidence and whether or not students had a flu shot.

	No Shot	Flu Shot	Total
Flu? Yes	15	3	18
No	13	10	23
Total	28	13	41

For this data, 3 of 13 (0.23) subjects who had the flu shot got the flu as compared to 15 of 28 (0.54) subjects who did not take the shot.

There are a couple contingency tables with the same marginal totals that would have even a larger disparity between these sample proportions, namely those where the upper right entry was 2, 1, or 0.

Example — Hypotheses

The null hypothesis is that the flu shot has no effect so that the counts of people with and without the flu in the two groups is completely random.

Example — p -value

The p -value is the sum of these hypergeometric probabilities for the actual table and all of those that are more extreme.

$$P = \frac{\binom{18}{3}\binom{23}{10}}{\binom{41}{13}} + \frac{\binom{18}{2}\binom{23}{11}}{\binom{41}{13}} + \frac{\binom{18}{1}\binom{23}{12}}{\binom{41}{13}} + \frac{\binom{18}{0}\binom{23}{13}}{\binom{41}{13}} = 0.066169$$

Note. Had we let the column totals be the ball color counts and had we let the row totals be the sample sizes, the p -value would have been the same.

Comparison to χ^2 test.

Here is some R code to carry out the χ^2 test of independence.

```
> x = matrix(c(15, 13, 3, 10), 2, 2)
> chisq.test(x, correct = F)
Pearson's Chi-squared test
```

```
data: x
X-squared = 3.3522, df = 1, p-value = 0.06712
```

Notice that the p -value is about the same.

For tables with large enough counts, the two p -values will be very similar. If the total counts in each cell of the table are small, the p -values can differ.

In this case, Fisher's Exact Test is preferred because the χ^2 approximation may not be good.

Example — Colored Ball Model

Specifically, there are $\binom{18}{3}$ ways to choose which red balls to include and $\binom{23}{10}$ ways to choose which white balls we want. The number of ways to choose both is the product. Overall, there are $\binom{41}{13}$ ways to choose 13 balls from 41.

Example — Interpretation

The p -value is smallish, but not that small.

There is weak evidence that the flu shot resulted in fewer flu cases among the people who took it ($p = 0.066$, Fisher's Exact Test, directional alternative).

R for Fisher's Exact Test

```
> x = matrix(c(15, 13, 3, 10), 2, 2)
> x
      [,1] [,2]
[1,]  15   3
[2,]  13  10
> fisher.test(x, alternative = "greater")
      Fisher's Exact Test for Count Data
```

```
data: x
p-value = 0.06617
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.9114942      Inf
sample estimates:
odds ratio
 3.721944
```

We specify that `alternative="greater"` because we want to know the probability that the [upper left](#) cell is its value or greater.

Example — R

```
> x = matrix(c(438, 1387, 807, 288, 746, 189, 115, 946, 1768, 16,
+ 53, 47), 3, 4)
> x
      [,1] [,2] [,3] [,4]
[1,]  438  288  115   16
[2,] 1387  746  946   53
[3,]  807  189 1768   47
> chisq.test(x, correct = F)
      Pearson's Chi-squared test
```

```
data: x
X-squared = 1073.508, df = 6, p-value = < 2.2e-16
```

The null hypothesis is that hair color and eye color are independent of one another.

This test has $(4 - 1)(3 - 1) = 6$ degrees of freedom.

The test statistic is over 1000, rather unusual for the sum of six squared independent standard normal random variables. There is overwhelming evidence that hair color and eye color are associated with one another.

R for Fisher's Exact Test

The p-value of Fisher's Exact Test can be computed in R directly using the hypergeometric probability distribution.

```
> phyper(3, 18, 23, 13)
[1] 0.06616925
```

To understand this syntax, `phyper` computes the cumulative probability distribution of a hypergeometric distribution, 3 is the number of red balls in the sample, 18 is the number of red balls in the bucket, 23 is the number of other balls in the bucket, and 13 is the sample size.

Alternatively, we can use the function `fisher.test`.

Larger Tables

The ideas of the previous sections extend to larger contingency tables.

Consider the following example with eye and hair color among 6,800 German men.

		Hair Color			
		Brown	Black	Fair	Red
Eye Color	Brown	438	288	115	16
	Gray or Green	1387	746	946	53
	Blue	807	189	1768	47

In principle, we could compute the row and column totals, find all of the expected counts, find the χ^2 test statistic, and then compute a p-value.

R makes this task easier.

Applicability

- For a [goodness-of-fit](#) test as discussed in this chapter, the null hypothesis should specify the probabilities of each category.
- For a [test of independence](#), the null hypothesis is that the row and column variables are independent.

Confidence Intervals for $p_1 - p_2$

adjusted formula has a coverage probability that is closer to 95% than the unadjusted formula.

We will add four observations, two of each type, but we will only one of each type to each sample.

The adjusted proportions are $\tilde{p}_i = (y_i + 1)/(n_i + 2)$ for $i = 1, 2$.

The formula for the 95% confidence interval for a difference in population means $p_1 - p_2$ then becomes

$$(\tilde{p}_1 - \tilde{p}_2) \pm 1.96 \sqrt{\frac{\tilde{p}_1(1 - \tilde{p}_1)}{n_1 + 2} + \frac{\tilde{p}_2(1 - \tilde{p}_2)}{n_2 + 2}}.$$

Applicability

The chi-square test will be applicable if:

- For a [goodness-of-fit test](#), the data must be a random sample of categorical observations from a large population.
- For a [contingency table test of independence](#), the data must be either (1) two or more independent random samples for which each sampled individual is observed on one categorical variable, or (2) one random sample for which each individual is observed on two categorical variables. In both cases, observations within a single sample must be independent of each other.
- The sample sizes must be large enough. A rule of thumb is that all expected counts should be at least five. (If the expected counts are smaller, a permutation test or Fisher's Exact Test can be used instead of the χ^2 distribution to find p-values.)

Confidence Intervals for $p_1 - p_2$

The method recommended by the textbook is similar to that for single proportions, except that the adjusted sample proportions add a pair of observations per sample, one of each outcome, instead of two.

The theory is similar to the one sample case. If the sample sizes are n_1 and n_2 and the number of successes in each sample are y_1 and y_2 respectively, the sample proportions are $\hat{p}_i = y_i/n_i$ for $i = 1, 2$.

The exact expression for the SE of $\hat{p}_1 - \hat{p}_2$ is

$$\sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}.$$

Instead of estimating the unknown population parameters p_i with the sample proportions, the textbook advocates adjusting them by adding a couple observations. The justification is that the

The Odds Ratio

Odds ratios are another way to compare probabilities.

If the probability of an event E is $\Pr\{E\}$,

$$\text{the odds of event } E = \frac{\Pr\{E\}}{1 - \Pr\{E\}}.$$

The odds ratio of two events, often denoted θ , is the ratio of the odds. So, the odds ratio for events with probabilities p_1 and p_2 is

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

An advantage of the odds ratio

Even though the relative risk, p_1/p_2 and a difference in probabilities $p_1 - p_2$ are both easier to interpret than an odds ratio, there are situations where an odds ratio can be calculated from data where relative risk and differences in probabilities both cannot.

In a **case-control design**, a control group is found to compare to a group of cases, say individuals with the same medical condition. We may not know the population prevalence of the condition. In this case, we can estimate an odds ratio but not a relative risk.

Relative Risk

Relative risk is a ratio of two conditional probabilities, both of the same event, but under different conditions.

For example, if the probability of a low birthweight baby given that the mother is a smoker is twice as high as if the mother is a nonsmoker, the **relative risk of low birthweight** for smokers relative to nonsmokers is 2.

Comparing Relative Risk and Odds Ratios

Relative risk and odds ratios are not identical, but are similar to one another. The exact relationship is this.

$$\begin{aligned} \text{odds ratio} &= \frac{p_1}{p_2} \times \frac{1-p_2}{1-p_1} \\ &= \text{relative risk} \times \frac{1-p_2}{1-p_1} \end{aligned}$$

These will be very close when p_1 and p_2 are both small.

Example

See the textbook, section 10.9, for a detailed example on how to compute odds ratios from 2×2 tables and also on how to compute confidence intervals for odds ratios.