Inference for Proportions

David Gerard
Most slides from Yali Amit and Linda Collins

2017-11-08

Learning Objectives

- CI's and testing for proportions.
- CI's and testing for differences in proportions.
- χ^2 distribution.
- Inference in two-way tables.
- Sections 6.1, 6.2, 6.3, and 6.4 in DBC.

High-level Review of CI's and Testing

General Strategy

We've seen three cases (means (σ known), means (σ unknown), differences of means) of the following:

- Construct a test statistic, T, that has some known distribution (either N(0,1) or t_ν).
- T is a function of the parameter, say θ, we are interested in (either the mean μ or the difference of two means $\mu_1 - \mu_2$). Denote this dependence by $T(\theta)$.
- Find bounds t such that $Pr(T(\theta) \in [-t, t]) = 1 - \alpha$ for some predefined α (usually $\alpha = 0.05$).
- Solve for θ: $Pr(\theta \in [T^{-1}(-t), T^{-1}(t)]) = 1 - \alpha$.
- This usually ends up being something like $\hat{\theta} \pm t\hat{SD}(\hat{\theta})$, where $\hat{\theta}$ is some estimate of θ (e.g. $\bar{X}$) and $\hat{SD}(\hat{\theta})$ is an estimate of the standard deviation of $\hat{\theta}$ (e.g. $s/\sqrt{n}$).

Inference for Proportions

Set up

Suppose we want to estimate the proportion p of some characteristic of a population, and we undertake the following procedure:

1. Draw a SRS of size n.
2. Record the number X of "successes" (those individuals having the characteristic).
3. Estimate the unknown true population proportion p with the sample proportion of successes $\hat{p} = \frac{X}{n}$

p is the mean of the population, but in this case it determines the $SD = p(1-p)$. In the general case σ is not determined by μ.
What is the sampling distribution of $\hat{p}$?

Normal Approximation and CI

If n is sufficiently large – i.e. if

$$np \geq 10 \text{ and } n(1-p) \geq 10,$$

then

$$\hat{p} \dot{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Thus, an approximate $(1-\alpha)$ CI for the population proportion p is given by

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where z^* is chosen so that $P(Z > z^*) = \alpha/2$ for $Z \sim N(0,1)$.

Normal Approximation and Testing

What if we want to test whether $p = p_0$ for some fixed value p_0?
The null hypothesis is $p = p_0$, and under this hypothesis,

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \dot{\sim} N(0,1)$$

Notice that we are using a different value for the SD of $\hat{p}$ than was used for the CI. Since H_0 specifies a true value for p, the SD of $\hat{p}$ under H_0 is given by $\sqrt{\frac{p_0(1-p_0)}{n}}$
The p-values for this test are:

- $H_a : p > p_0 \qquad P(Z \geq z)$
- $H_a : p < p_0 \qquad P(Z \leq z)$
- $H_a : p \neq p_0 \qquad 2P(Z \geq |z|)$

for $Z \sim N(0,1)$.

Some care needs to be taken when p is very close to 0 or 1.

Example

A random sample of 2700 California lawyers revealed only 1107 who felt that the ethical standards of most lawyers are high (*AP, Nov. 12, 1994*).

1. Does this provide strong evidence for concluding that fewer than 50% of all California lawyers feel this way?
2. What is a 90% confidence interval for the true proportion of California lawyers who feel that ethical standards are high?

Using R

```
hp <- 1107 / 2700
Z <- (hp - 0.5) / sqrt(0.25 / 2700)
Z^2
[1] 87.48
prop.test(1107, 2700, conf.level = 0.90, correct = FALSE)

1-sample proportions test without continuity
correction

data:  1107 out of 2700, null probability 0.5
X-squared = 87, df = 1, p-value <2e-16
alternative hypothesis: true p is not equal to 0.5
90 percent confidence interval:
 0.3945 0.4257
sample estimates:
   p
0.41
```

Differences of Proportions

Comparing Two Proportions

Suppose we have two populations A and B with unknown proportions p_1 and p_2 respectively. A SRS of size n_1 from A yields $\hat{p}_1$, and an independent SRS of size n_2 from B yields $\hat{p}_2$. Then,

$$(\hat{p}_1 - \hat{p}_2) \dot{\sim} N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

when n_1 and n_2 are large.

An approximate 95% CI for $p_1 - p_2$ is then given by

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

Again, the estimates of the means/proportions determines the SD's.

Comparing Two Proportions

To test $H_0 : p_1 = p_2$, we compute the test statistic

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

where $\hat{p}$ is the *combined* proportion of successes in both samples

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

with X_1, X_2 denoting the number of successes in each sample. Under H_0, the Z-statistic has approximately a standard normal distribution (using the normal approximation to the binomial), and p-values are calculated as above.

Survey Questions Example

The ability of question wording to affect the outcome of a survey can be a serious issue. Consider the following two questions:

1. Would you favor or oppose a law that would require a person to obtain a police permit before purchasing a gun?
2. Would you favor or oppose a law that would require a person to obtain a police permit before purchasing a gun, or do you think such a law would interfere too much with the right of citizens to own guns?

Let n_i denote the number of people who were asked question i, and X_i denote the number of these who favor the permit law.

Ques.	X_i	n_i
1	463	615
2	403	585

Survey Questions Example

Is the true proportion of people favoring the permit law the same in both groups or not?

```
prop.test(c(463, 403), c(615, 585))

2-sample test for equality of proportions with
continuity correction

data:  c(463, 403) out of c(615, 585)
X-squared = 5.8, df = 1, p-value = 0.02
alternative hypothesis: two.sided
95 percent confidence interval:
 0.0116 0.1163
sample estimates:
prop 1 prop 2
0.7528 0.6889
```

Two-way Tables

Wine Example

In a study conducted in a Northern Ireland supermarket, researchers counted the number of bottles of French, Italian, and other wine purchased while shoppers were subject to one of three "treatments": no music, French accordion music, and Italian string music.

The following **two-way table** summarizes the data:

	Music			
Wine	None	French	Italian	Total
French	30	39	30	99
Italian	11	1	19	31
Other	43	35	35	113
Total	84	75	84	243

Wine Example

The table of counts looks like the joint distribution tables we studied earlier. Indeed, from these counts, we can ascertain the (empirical) joint distribution, marginal distributions, and conditional distributions of wine type and music type:

	Music			
Wine	None	French	Italian	Total
French	0.123	0.160	0.123	0.407
Italian	0.045	0.004	0.078	0.128
Other	0.177	0.144	0.144	0.465
Total	0.346	0.309	0.346	1.000

Wine Example

- We are interested in determining whether there is relationship between the row variable (wine type) and the column variable (music type).
- If this were the *true distribution*, then the answer would be clear: music and wine are not independent, so there *is* a relationship.
- However, this table is *random*, and we want to know whether or not music and wine are independent *under the true distribution*. This requires a statistical test.

Intuition of test

- H_0: the row and column variables are independent (i.e. there is no relationship between the two).
- H_a: the row and column variables are dependent.

Suppose H_0 is true, and the two variables are independent. What counts would we expect to observe?

Recall that under the independence assumption,

$P(x, y) = P(x)P(y)$, for all x, y.

We estimate the marginals from the data.

$\hat{P}(x) = \frac{\text{row total for x}}{\text{total count}}$, $\hat{P}(y) = \frac{\text{col total for y}}{\text{total count}}$

Intuition of test

Thus, for each cell, we have

$$\text{Expected Cell Count} = \text{total count} \cdot \hat{P}(x)\hat{P}(y) = \frac{\text{row total} \times \text{col total}}{\text{total count}}$$

Our test will be based on a measure of *how far the observed table is from the expected table.*

For the supermarket example, the expected counts are:

	Music			
Wine	None	French	Italian	Total
French	34.22	30.56	34.22	99
Italian	10.72	9.57	10.72	31
Other	39.06	34.88	39.06	113
Total	84	75	84	243

The χ^2 test statistic

	Observed			Expected			
	Music			Music			
Wine	None	Fr.	It.	None	Fr.	It.	Tot.
French	30	39	30	34.22	30.56	34.22	99
Italian	11	1	19	10.72	9.57	10.72	31
Other	43	35	35	39.06	34.88	39.06	113
Total	84	75	84	84	75	84	243

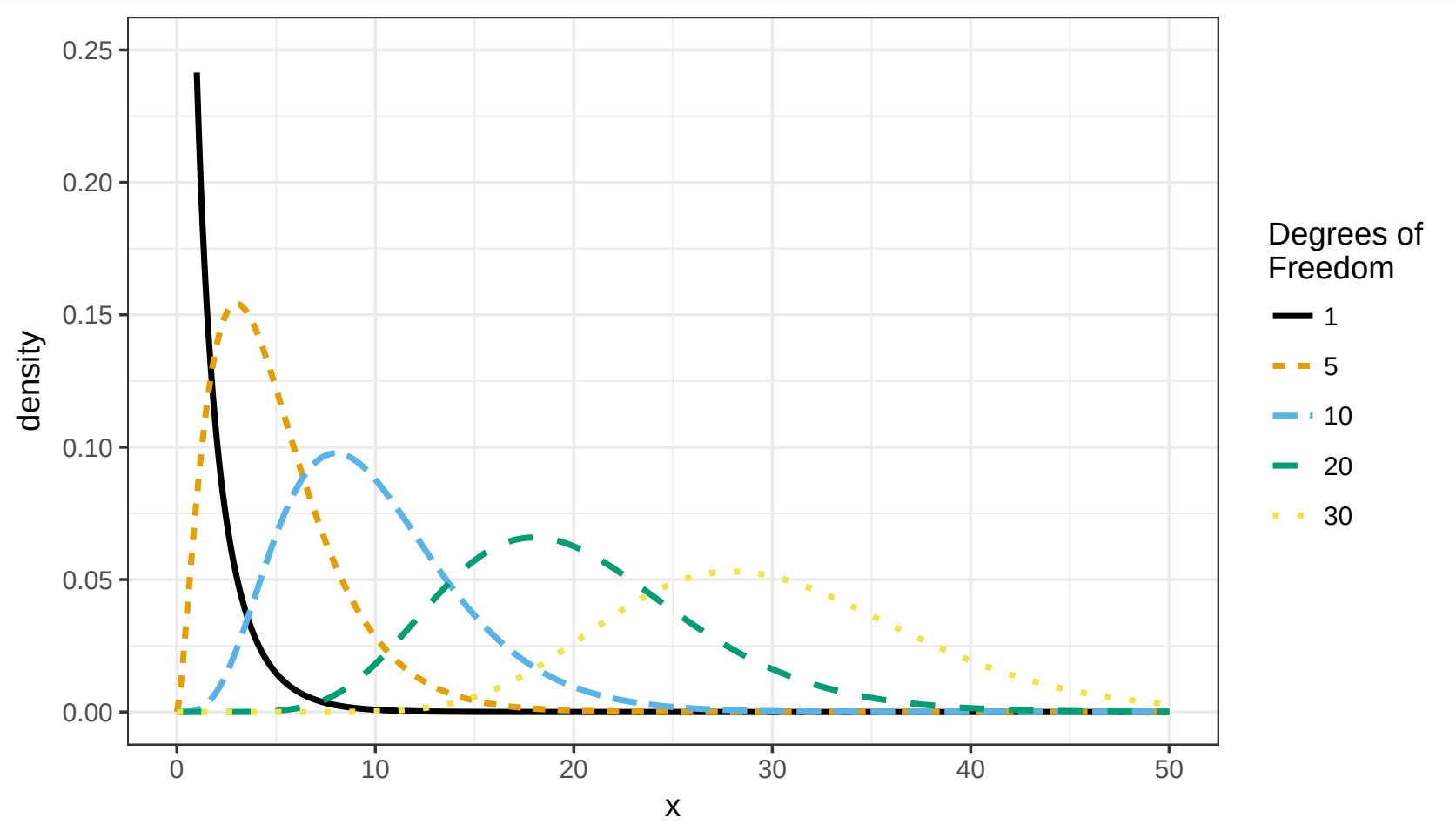
To measure how far the *expected* table is from the *observed* table, we will use the following test statistic:

$$X^2 = \sum_{\text{all cells}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

The χ^2 distribution

- Under H_0, the X^2 test statistic has an approximate χ^2 distribution with $(r-1)(c-1)$ degrees of freedom, denoted $\chi^2_{(r-1)(c-1)}$.
- Why $(r-1)(c-1)$?
- Recall that our "expected" table is based on some quantities estimated from the data: namely the row and column totals.
- Once these totals are known, filling in any $(r-1)(c-1)$ undetermined table entries actually gives us the whole table. Thus, there are only $(r-1)(c-1)$ freely varying quantities in the table.
- What does the χ^2 distribution look like?

The χ^2 distribution



More on χ^2

- Unlike the Normal or t distributions, the χ^2 distribution takes values in $(0, \infty)$.
- As with the t distribution, the exact shape of the χ^2 distribution depends on its degrees of freedom.

If the observed and expected counts are very different, X^2 will be large, indicating evidence against H_0. Thus, the p-value is always based on the right-hand tail of the distribution.

There is no notion of a two-tailed test in this context.

The p-value is therefore

$$P(\chi^2_{(r-1)(c-1)} \geq X^2)$$

Recall that X^2 has an *approximate* $\chi^2_{(r-1)(c-1)}$ distribution. When is the approximation valid?

Conditions for approximation to be valid

For any two-way table larger than 2×2, we require that the average expected cell count is at least 5 and each expected count is at least one.

For 2×2 tables, we require that each expected count be at least 5.

Let's get back to our example...

Recall the observed and expected counts:

	Observed			Expected			
	Music			Music			
Wine	None	Fr.	It.	None	Fr.	It.	Tot.
French	30	39	30	34.22	30.56	34.22	99
Italian	11	1	19	10.72	9.57	10.72	31
Other	43	35	35	39.06	34.88	39.06	113
Total	84	75	84	84	75	84	243

Calculating X^2

$$\begin{aligned} X^2 &= \frac{(30-34.22)^2}{34.22} + \frac{(39-30.56)^2}{30.56} + \frac{(30-34.22)^2}{34.22} \\ &\quad + \cdots + \frac{(35-34.88)^2}{34.88} + \frac{(35-39.06)^2}{39.06} \\ &= 18.28 \end{aligned}$$

The table is 3×3, so there are $(r-1)(c-1) = 2 \times 2 = 4$ degrees of freedom.

Finally, the p-value is found from the χ^2_4 table:

$$0.001 \leq P(\chi^2_4 \geq 18.28) \leq 0.0025$$

In R

```
# Enter the data
wine=c(30,11,43,39,1,35,30,19,35)
# Reshape it as a table
dim(wine)=c(3,3)
wine
     [,1] [,2] [,3]
[1,]   30   39   30
[2,]   11    1   19
[3,]   43   35   35
# Run the test
chisq.test(wine)

Pearson's Chi-squared test

data:  wine
X-squared = 18, df = 4, p-value = 0.001
```

The χ^2-test for the presence of a relationship between two directions in a two-way table is valid for data produced by several different study designs, although the exact null hypothesis varies.

1. Examining independence between variables

- Suppose we select an SRS of size n from a population and classify each individual according to 2 categorical variables. Then, a χ^2-test can be used to test
- H_0: The two variables are independent
- H_a: Not independent
- Suppose we collect an SRS of 114 college students, and categorize each by major and GPA (e.g. $(0, 0.5], (0.5, 1], \ldots, (3.5, 4]$). Then, we can use a χ^2 test to ascertain whether grades and major are independent.

2. Comparing several populations

- Suppose we select *independent* SRSs from each of c populations, of sizes $n_1, n_2, \ldots, n_c$. We then classify each individual according to a categorical response variable with r possible values (the same across populations). This yields a $r \times c$ table, and a χ^2-test can be used to test
- H_0: Distribution of the response variable is the same in all populations
- H_a: Distributions of response variables are not all the same.
- Suppose we select independent SRSs of Psychology, Biology and Math majors, of sizes 40, 39, 35, and classify each individual by GPA range. Then, we can use a χ^2 test to ascertain whether or not the distribution of grades is the same in all three populations.

Back to the survey example.

- There are two populations and for each there is a 2 category (Bernoulli) variable. So this is a 2×2 table.
- One variable is the population label and one variable is the response.
- Saying they have the same proportions is the same as saying the two variables are independent.
- The X^2 test statistic for independence is exactly the **square** of the z-statistic for equality of the proportions.
- The original data comes as two sample sizes and two counts of yes responses. It can be rewritten

Ques.	X_i	n_i
1	463	615
2	403	585

Ques.	Yes	No
1	463	152
2	403	182

R-code for the χ-squared test for independence

```
# Enter data as two way table.
X <- c(463, 403, 615 - 463, 585 - 403)
dim(X) <- c(2, 2)
chisq.test(X, correct = FALSE)

Pearson's Chi-squared test

data:  X
X-squared = 6.1, df = 1, p-value = 0.01
```

Explicit z-test for equal proportions

```
n1 <- X[1, 1] + X[1, 2]
n2 <- X[2, 1] + X[2, 2]
p1 <- X[1, 1] / n1
p2 <- X[2, 1] / n2
p  <- (X[1, 1] + X[2, 1]) / (n1 + n2)
z  <- (p1 - p2) / sqrt(p * (1 - p) * (1 / n1 + 1 / n2))
c(z, 2 * pnorm(-z))
[1] 2.47093 0.01348
```

R proportion test using the χ^2 statistic

```
prp <- prop.test(X[, 1], c(n1, n2), correct = FALSE)
c(prp$statistic, prp$p.value)
X-squared
  6.10547   0.01348
z^2
[1] 6.105
```

Note that X-squared $= z^2$