Analyzing Categorical Data

STA 198: Introduction to Health Data Science

Yue Jiang

June 29, 2020

The following material was used by Yue Jiang during a live lecture.

Without the accompanying oral comments, the text is incomplete as a record of the presentation.

Binomial distribution

If X is binomial with parameters n and p, then

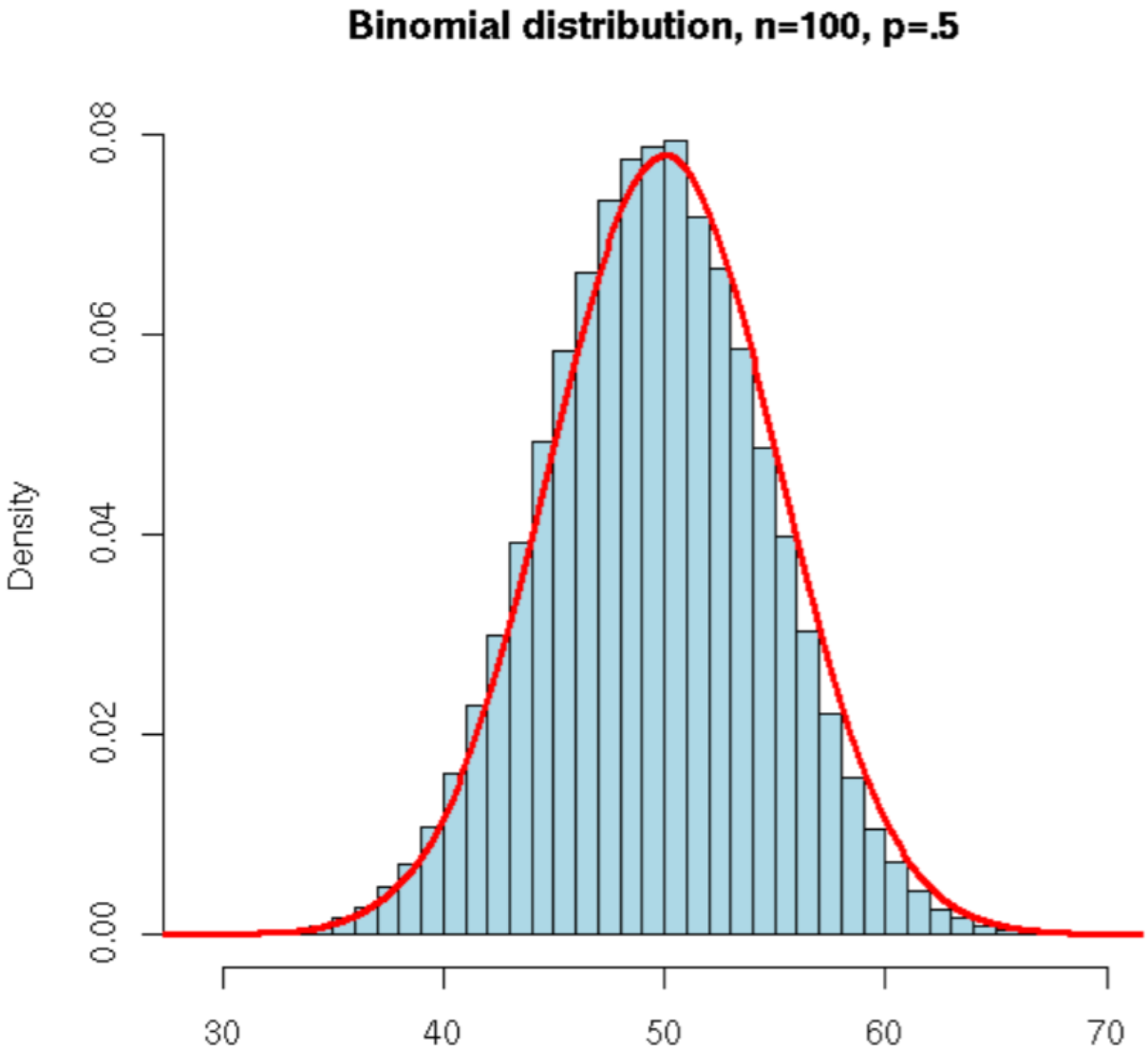
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

with mean np and standard deviation $\sqrt{np(1-p)}$.

Estimating a single proportion

- Suppose we conduct a binomial experiment n times, letting the ith event $h_i = 1$ if we get a success and $h_i = 0$ otherwise
- The number of successes is $k = \sum_{i=1}^{n} h_i$
- The sample proportion is $\hat{p} = \frac{k}{n} = \frac{1}{n} \sum_{i=1}^{n} h_i$, which is just a sample mean
- From the CLT, $\hat{p} \approx N\left(p, \frac{p(1-p)}{n}\right)$

Binomial distribution (normal superimposed)



Normal approximation to the binomial distribution

We see that for large enough n, the normal distribution provides a good approximation to the binomial. That is,

$$Z = \frac{k - np}{\sqrt{np(1-p)}} \approx N(0, 1).$$

n is "large enough" for both the approimation if both np and $n(1-p)$ are greater than or equal to 5 (some people say 10).

Sampling distribution of a proportion

Suppose we take repeated samples of size n from the population, and obtain estimates of the population proportion $\hat{p}_1$, $\hat{p}_2$, etc. According to the Central Limit Theorem, the distribution of the sample proportions has the following properties:

- Its mean is the population mean p
- Its standard deviation is given by $\sqrt{\frac{p(1-p)}{n}}$
- Its shape is approximately normal for n "large enough"

Then we know

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

is approximately $N(0, 1)$.

Using the normal approximation to get CIs

We can use this result to get confidence intervals. A $100(1-\alpha)\%$ CI would be given by

$$\hat{p} \pm z^{\star}_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}}.$$

However, p is an unknown parameter, and so we estimate it with $\hat{p}$ and use

$$\hat{p} \pm z^{\star}_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

to obtain an approximate CI.

R will provide these for you – more details on HW 06.

Hypothesis testing

How might we test $H_0 : p = p_0$ against $H_1 : p \neq p_0$?

We draw a random sample from the underlying population of interest, estimate p using $\hat{p}$, and find the probability of getting a sample proportion as extreme, or more extreme than $\hat{p}$ if the true population proportion is p_0. The statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

has a $N(0, 1)$ distribution if H_0 is true.

Again, we will calculate this using R.

Hypothesis testing

Because we use different standard error estimates for the CI and the test, the CI and hypothesis test results may not always agree as they did when we were estimating continuous means.

Why the different standard error estimate?

Comparison of two proportions

Now let's consider testing whether two proportions from *independent populations* are the same:

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2.$$

We estimate the two proportions using $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$ where x_i is the number of "successes" in sample i (drawn from population 1 or 2) and n_i is the corresponding sample size.

Comparison of two proportions

As before, the p-value of this test will represent the probability of obtaining a discrepancy $\hat{p}_1 - \hat{p}_2$ as large as or larger than what we see, if the two population proportions are indeed identical.

If H_0 is indeed true, then we can estimate the overall proportion simply by combining the samples to get

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} \text{ (Why?)}$$

If H_0 is true then the standard error of $\hat{p}_1 - \hat{p}_2$ can be estimated by

$$\widehat{SE} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Comparison of two proportions

Then the test statistic given by

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

is approximately $N(0, 1)$ when n_1 and n_2 are sufficiently large (if np_1, $n(1 - p_1)$, np_2, and $n(1 - p_2)$ are all at least 5).

CI for difference of two proportions

We can generate confidence intervals in R, given by

$$\hat{p}_1 - \hat{p}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Why don't we use the standard error estimate

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

from the hypothesis test?

Example: Coronavirus worry

Suppose we want to test the hypothesis that the proportions of people "very worried about COVID-19" are the same among those aged 13 – 18 and those 65 and older. We have

$$H_0 : p_1 = p_2 \text{ vs. } H_1 : p_1 \neq p_2.$$

Suppose we randomly sample 100 unrelated people in each age group, where 28 of the teenagers were worried about coronavirus while 72 of the older adults were.

Is there enough evidence to suggest that there is a difference in the proportion of worried individuals among the two age groups?

Example: Coronavirus worry

A contingency table is a display format for showing the relationship between two categorical variables.

	Worried	Not Worried	Total
Teenagers	28	72	100
Older Adults	72	28	100
Total	100	100	200

Using this table, we see $\hat{p}_1 = 0.28$ and $\hat{p}_2 = 0.72$. The z statistic is -6.2, corresponding to a p-value < 0.0001. We reject the null hypothesis; we have sufficient evidence to suggest that the two proportions are different at the $\alpha = 0.05$ level.

Calculate this using R.

Categorical Data

We just compared the two categorical variables

- Age: teenager vs. older adult
- Worried status: "very worried" vs. not

We created and analyzed a 2x2 table. However, we can also consider more general contingency tables.

Example: *Streptococcus pneumoniae*

The World Health Organization estimates that in developing countries 814,000 children under the age of five die annually from invasive pneumococcal disease (IPD), with an estimated 1.6 million deaths affecting all ages globally.

Several recent studies have identified associations between pneumococcal serotypes (species variations) and patient outcomes from IPD. We consider data from a study of pneumococcal serotypes and mortality (Inverarity et al. (2011), *Journal of Medical Microbiology*).

Contingency table for *S. pneumoniae* data

	Died	Survived	Total
Serotype 31	10	24	34
Serotype 10	7	37	44
Serotype 15	12	60	72
Serotype 20	9	97	106
Total	38	218	256

Typical questions of interest:

- Is there an association between the two variables?
- How strong is any association?

General hypothesis test for categorical variables

We phrase the general hypothesis test for two categorical variables based on whether there is an association between them:

- H_0: pneumococcal serotype is unrelated to mortality (there is no association between them)
- H_1: pneumococcal serotype is related to mortality (there IS an association between them)

Chi-square test

If we have large enough samples (> 10 in all cells for $\alpha = 0.05$), we will use a chi-square (χ^2) test.

This isn't quite the case for our IPD data, but we'll look the other way for now.

Chi-square test

The chi-square test has a nice motivation: it compares observed proportions to proportions we would expect if H_0 were true.

	Died	Survived	Total
Serotype 31	10	24	34
Serotype 10	7	37	44
Serotype 15	12	60	72
Serotype 20	9	97	106
Total	38	218	256

Suppose there is no association between serotype and mortality (H_0 true). $\frac{34}{256}$ of participants had serotype 31 disease and $\frac{38}{256}$ of our subjects died. Thus, the probability that they had serotype 31 disease AND died $= \frac{34}{256} \times \frac{38}{256}$ **if** H_0 is true (why?).

Observed vs. expected counts

In our study, 10 patients with serogroup 31 disease died. Under the null hypothesis, we would expect $256 \times \frac{34}{256} \times \frac{38}{256} \approx 5.05$ patients to have both serogroup 31 and have died.

More patients with serogroup 31 died than would be expected if serotype and mortality truly had no association.

Is this statistically significant?

The chi-square test statistic

- The chi-squared test compares the observed frequencies (O) in each cell of the table to the expected frequencies (E) if H_0 is true.
- If differences between what we observe and expect ($O - E$) are large enough, we reject H_0.
- To combine differences across table cells, we square them (to put more weight on larger deviations and also so extra high-risk serotypes are not cancelled out by fewer low-risk serotypes) before adding them up.
- Finally, we scale the differences by the expected count.

R will do this for us – details on HW 06.

The chi-square test statistic

The χ^2 test statistic is

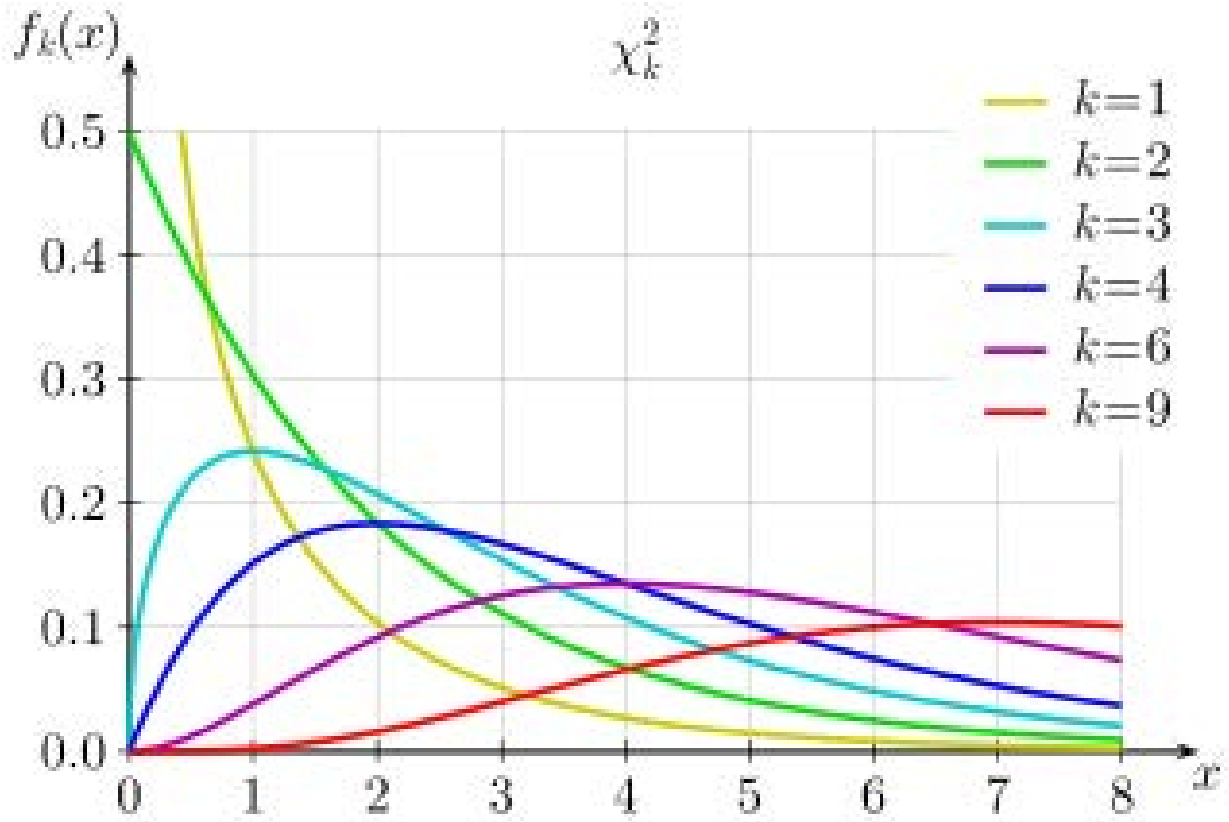
$$\chi^2 = \sum_{i=1}^{r \times c} \frac{(O_i - E_i)^2}{E_i},$$

where $r \times c$ is the number of cells in the table (rows times columns)

Under the null hypothesis, the distribution of this sum is approximated by a χ^2 distribution with $(r-1) \times (c-1)$ degrees of freedom.

There is a different χ^2 distribution for each degree of freedom, and we look at the area in the right tail only (why?).

Chi-square (χ^2) distributions



Chi-square test on the IPD data

The p-value for the χ^2 test is 0.025. What can we conclude?

As an aside...

Much like how ANOVA with only two categories is equivalent to the t-test, the chi-square test with only a 2x2 table is equivalent to the two-sample test of proportions using the z-test.

We can thus think of the chi-square test as "extending" what we have learned for proportions in some sense.