Categorical Data Analysis

STA 102: Introduction to Biostatistics

Yue Jiang

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.

Anonymous peer lab assessment

https://duke.qualtrics.com/jfe/form/SV_6m3jANQfco2clql

Please use this opportunity to provide honest, constructive feedback about what has worked well and what might need improvement (if anything!) regarding team participation. The survey is **completely anonymous**

The survey will take approximately 10-15 minutes; complete the peer evaluation by Friday, October 9th, at 11:59 PM.

Binomial distribution

If X is binomial with parameters n and p, then

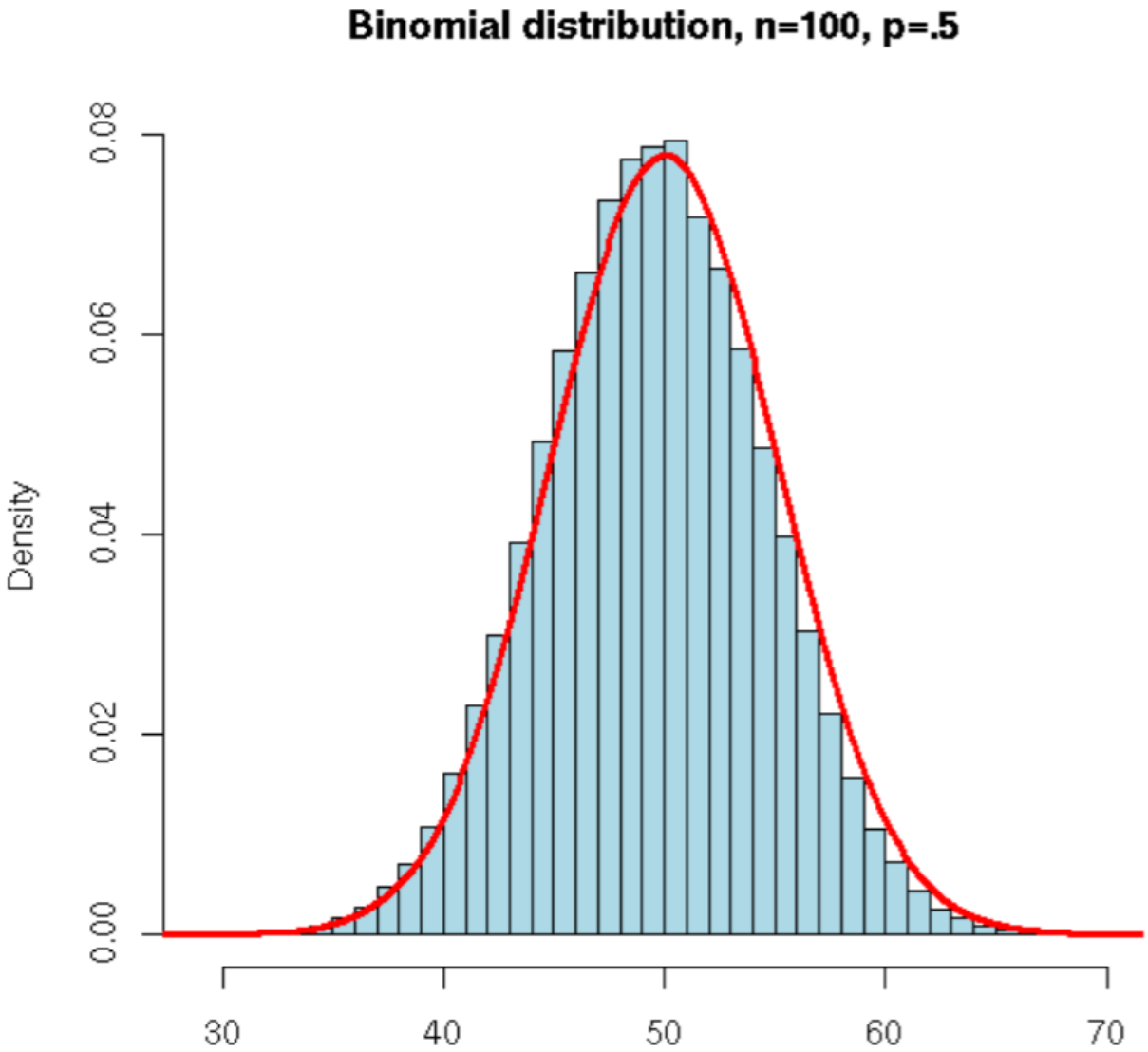
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

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

Estimating a single proportion

- Suppose we conduct a Bernoulli 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.

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.

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?

General Categorical Data

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.

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.

Chi-square test on the IPD data

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