P8130: Biostatistical Methods I

Lecture 9: Methods of Inference for Proportions

Cody Chiuzan, PhD
Department of Biostatistics
Mailman School of Public Health (MSPH)

Lecture 6-8: Recap

- Inferences for one- and two-samples means
  - Confidence intervals
  - Hypothesis testing
  - Power and sample size calculation
- Inferences for more than two means (ANOVA)
  - Multiple comparison adjustments

Lectures 9: Outline

- One and two-sample proportions (Normal Approximation)
  - Point estimates
  - Confidence intervals
  - Hypothesis testing

Estimation of Binomial Distribution

Suppose we are conducting a 'yes/no' survey of a group of randomly sampled people. The overall goal is to determine the proportion of people who said 'yes' and draw conclusions for the overall population.

Let n be the total number of samples, independent and identically distributed (i.i.d.):

$X_i, i = 1, 2, \dots, n$; where the probability that each X_i is 1 is given by p;

p, probability of answering 'yes', represents the parameter of interest

p is considered fixed, but unknown.

It follows that each $X_i \sim Bernoulli(1, p)$, and that $\sum_{i=1}^{n} X_i = X \sim Binomial\ (n, p)$.

Estimation of Binomial Distribution

Our estimate of *p (population proportion),* is called $\hat{p}$ and is given by:

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X}{n}$$

Based on the fact that $X \sim Binomial\ (n, p)$, we can compute the following:

$$E(\hat{p}) = \frac{1}{n} E(X) = \frac{1}{n} np = p$$

$$var(\hat{p}) = \frac{1}{n^2} var(X) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$$

Because $E(\hat{p}) = p$, we can say that $\hat{p}$ is an unbiased estimator for p.

Also, notice that the variance of $\hat{p}$ decreases as the number of samples increases.

Normal Approximation of Binomial

We showed that $\hat{p}$ can be represented as an average of Bernoulli trials, each with mean p and variance $p(1-p)$.

Thus, for large n, the Central-Limit Theorem applies and we can see that the sample proportion $\hat{p}$ is normally distributed:

$$\hat{p} = \bar{X} \sim N(p, \frac{p(1-p)}{n})$$

Large *n*, rule of thumb: $np(1-p) \geq 5$

Notes:

- If p is close to 0.5, the Binomial distribution will look almost Normal for only $n = 10$
- If p is close to 0.1 or 0.9, the Binomial distribution will look almost Normal for $n = 50$
- If p is much closer to 0 or 1, then a normal approximation might not work well.

Interval Estimation: One-Sample Proportion

Based on $\hat{p} \sim N(p, \frac{p(1-p)}{n})$, we derive the 95% confidence interval for a population proportion:

$$P(-1.96 < z < 1.96) = 0.95 \rightarrow P\left(-1.96 < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < 1.96\right) = 0.95$$

Substitute p with $\hat{p}$ in the denominator (why?) :

$$P\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 0.95$$

In general, a $100(1-\alpha)\%$ confidence interval for one population proportion is given by:

$$\left(\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Exact Interval Estimation: One-Sample Proportion

What if the normal approximation is not valid, i.e., $np(1-p) \not\geq 5$?

In this case, a small-sample method needs to be applied.

Clopper–Pearson Method for obtaining an EXACT $100(1-\alpha)\%$ confidence interval for the binomial parameter p is given by (p_1, p_2), where:

$$P(X \geq x | p = p_1) = \frac{\alpha}{2} = \sum_{k=x}^{n} \binom{n}{k} p_1^k (1-p_1)^{n-k}$$

$$P(X \leq x | p = p_2) = \frac{\alpha}{2} = \sum_{k=0}^{x} \binom{n}{k} p_2^k (1-p_2)^{n-k}$$

- The limits of this Exact CI can be found in Tables or computed with software.
- Note that the Exact CI is not symmetric about the estimated sample proportion.

One-Sample Test for Binomial Proportion

Tests for One-Population Proportion, Normal Theory Methods

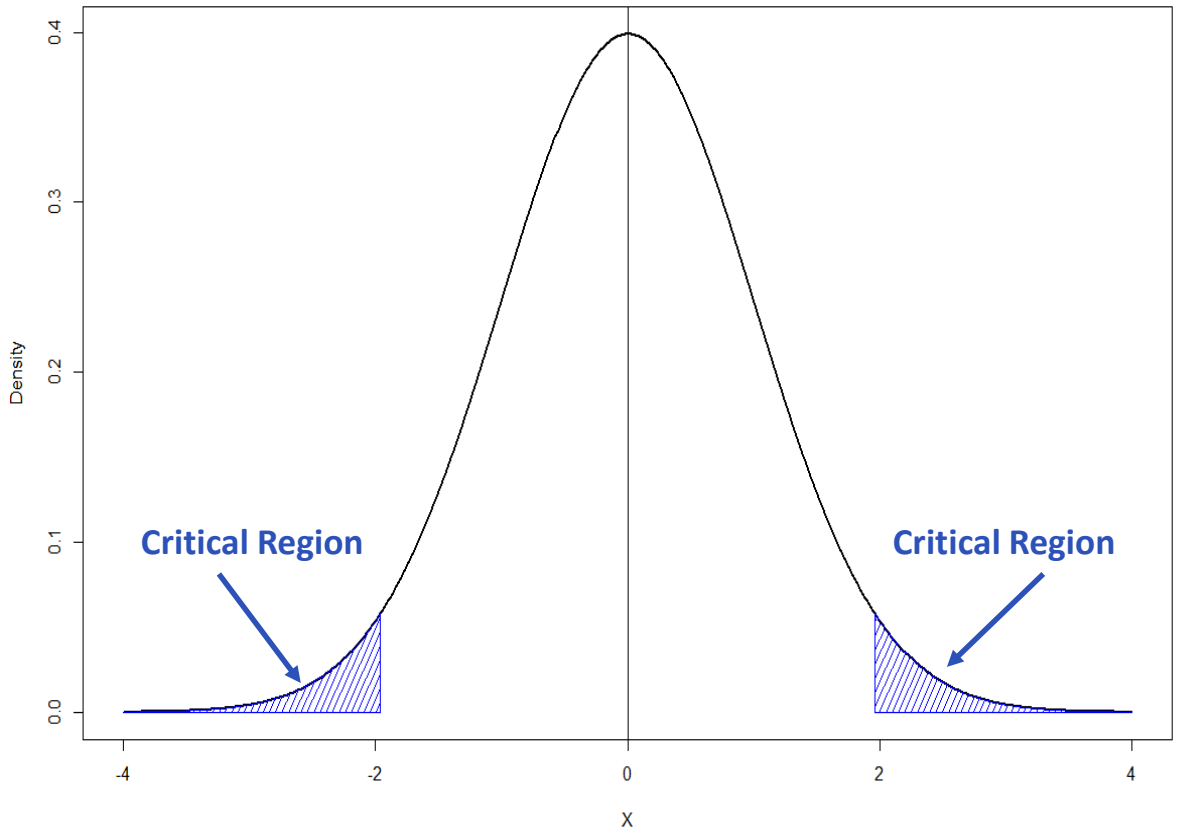
$H_0: p = p_0$ vs $H_1: p \neq p_0$

With significance level α pre-specified, compute the test statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim \mathrm{N}(0,1)$$

Reject H_0: if $|z| > z_{1-\alpha/2}$

Fail to reject H_0: if $|z| \leq z_{1-\alpha/2}$



One-Sample Exact Test for Binomial Proportion

If the assumption of normal approximation is not valid, then we use exact probabilities to calculate the p-value.

If $\hat{p} \leq p_0$, the p-value = $2 \times P(\leq x \; successes \; in \; n \; trials | H_0 \; is \; true)$

$$= \sum_{k=0}^{x} \binom{n}{k} p_0^k (1 - p_0)^{n-k}$$

If $\hat{p} > p_0$, the p-value = $2 \times P(\geq x \; successes \; in \; n \; trials | H_0 \; is \; true)$

$$= \sum_{k=x}^{n} \binom{n}{k} p_0^k (1 - p_0)^{n-k}$$

One-Sample Binomial Proportion

Example: In a survey of 300 randomly selected drivers, 125 claimed that they regularly wear seat belts.

Can we conclude from these data that the population proportion who regularly wear seat belts is 0.50?

Perform a hypothesis test and construct a 95% confidence interval for the population proportion.

In class practice.

Two-Sample Binomial Test for Proportions

- Data can be classified into two mutually exclusive categories.
- Let sample 1 have an estimated proportion $\widehat{p_1} \sim N(p_1, p_1(1-p_1)/n_1)$
- Let sample 2 have an estimated proportion $\widehat{p_2} \sim N(p_2, p_2(1-p_2)/n_2)$

Assuming that these are two independent random samples and that Normal Theory holds

$$n_1 p_1(1-p_1) \geq 5 \text{ and } n_2 p_2(1-p_2) \geq 5,$$

it follows that

$$\widehat{p_1} - \widehat{p_2} \sim N(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}).$$

Two-Sample Binomial Test for Proportions

Suppose we want to test the null hypothesis $H_0: p_1 = p_2 = p$.

If the null hypothesis is true then

$$\widehat{p_1} - \widehat{p_2} \sim N\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

However, p and $q = 1 - p$ are unknown, but can be estimated as:

$\hat{p} = \frac{n_1\widehat{p_1} + n_2\widehat{p_2}}{n_1 + n_2}$, an weighted average of the two sample proportions (*).

It we use this weighted average to estimate the common population proportion, then under the null:

$$z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1).$$

Two-Sample Binomial Test for Proportions

Tests for Two-Population Proportions, Normal Theory Methods:

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

The test statistic with continuity correction is given by:

$$z = \frac{|\widehat{p_1} - \widehat{p_2}| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ where } \hat{p} = \frac{n_1\widehat{p_1} + n_2\widehat{p_2}}{n_1 + n_2}$$

Reject H_0: if $|z| > z_{1-\alpha/2}$

Fail to reject H_0: if $|z| \leq z_{1-\alpha/2}$

Confidence Intervals for the Difference in Two Population Proportions

Two-sided $100(1-\alpha)\%$ confidence interval for the difference between two population proportions is given by:

$$(\widehat{p_1} - \widehat{p_2} - z_{1-\alpha/2}\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}},\ \widehat{p_1} - \widehat{p_2} + z_{1-\alpha/2}\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}})$$

Interpretation of a $100(1-\alpha)\%$ confidence interval:

We are $100(1-\alpha)\%$ confident that the true difference between the population proportions lies between the lower limit of the CI and the upper limit of the CI.

Note: the underlined portions should be substituted with specific numerical values.

Two-Sample Test for Binomial Proportions

Example: A total of 400 households were randomly selected from two communities and one respondent from each household was asked if he/she was bothered by the air pollution. Use the table below to perform an hypothesis test for comparing the proportions of respondents that were bothered by pollution in the two communities.

	Response to Air Pollution		
	Yes	**No**	Total
Community 1	43	157	200
Community 2	81	119	200
Total	124	276	400

Two-Sample Test for Binomial Proportions

Step 1: Extract the data info.

Community 1: $n_1 = 100, \widehat{p_1}$ = 43/200 = 0.215

Community 2: $n_2 = 100, \widehat{p_2}$ = 81/200 = 0.405

$\widehat{p_1}$ and $\widehat{p_2}$ represent the sample proportions of respondents that were bothered by pollution

Step 2: Check normal approximation validity:

$200 \cdot 0.215 \cdot (1 - 0.215) \geq 5$ and $200 \cdot 0.405 \cdot (1 - 0.405) \geq 5$

Step 3: Set up the hypotheses:

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

Two-Sample Test for Binomial Proportions

Step 4: Calculate the test-statistic

$$z = \frac{|0.215-0.405|-\left(\frac{1}{2\cdot 200}+\frac{1}{2\cdot 200}\right)}{\sqrt{0.31(1-0.31)\left(\frac{1}{200}+\frac{1}{200}\right)}}, \text{ where } \hat{p} = \frac{(200\cdot 0.215)+(200\cdot 0.405)}{400} = \frac{124}{400} = 0.31$$

$$z = \frac{|0.215 - 0.405| - \left(\frac{1}{2 \cdot 200} + \frac{1}{2 \cdot 200}\right)}{0.046} = 4.0$$

Step 5: At 0.05 significance level, z-stat > $z_{1-\alpha/2}$=1.96; thus we reject the null and conclude that there is a significant difference between the proportions of respondents bothered by pollution in the two communities.

Assuming that these communities were selected from the NY Metro area and Queens.

What can we conclude about the air pollution opinion over the entire NY State?

Readings

Rosner, *Fundamentals of Biostatistics 8th Edition*

- Chapter 6, Section 6.8
- Chapter 7, Section 7.10
- Chapter 10, Section 10.2