Tufts UNIVERSITY

Bruce M. Boghosian

Two-sample inferences

Binomial data: Testing $H_0 : p_X = p_Y$

Bruce M. Boghosian

Tufts UNIVERSITY | School of Arts and Sciences

Department of Mathematics

Tufts University

Outline

Motivation

- In a previous module, we studied the two-sample problem for pairs of normally distributed data.
- We can do the same thing for other distributions, including discrete distributions.
- In this module, we analyze the situation for n Bernoulli trials.

Statement of the problem

- Suppose that n independent Bernoulli trials related to treatment X have resulted in x successes.
- And that m independent Bernoulli trials related to treatment Y resulted in y successes.
- We want to know if p_X and p_Y, the true probability of success for the two treatments, are equal.

Applying the GLR criterion

- Null hypothesis $H_0 : \; p_X = p_Y \quad (=: p)$
- Alternative hypothesis $H_1 : p_X \neq p_Y$
- Two parameter spaces for GLRT:

$$\begin{aligned} \omega =& \{(p_X, p_Y) \mid 0 \leq p_X = p_Y \leq 1\} \\ \Omega =& \{(p_X, p_Y) \mid 0 \leq p_X \leq 1, 0 \leq p_Y \leq 1\} \end{aligned}$$

- Likelihood function

$$L(p_X, p_Y) = p_X^x (1 - p_X)^{n-x} \cdot p_Y^y (1 - p_Y)^{m-y},$$

where $x = \sum_j^n x_j$ and $y = \sum_j^n y_j$.

Maximum likelihood

- Likelihood function

$$L(p_X, p_Y) = p_X^x (1 - p_X)^{n-x} \cdot p_Y^y (1 - p_Y)^{m-y},$$

where $x = \sum_j^n x_j$ and $y = \sum_j^n y_j$.

- For ω, take derivative with respect to $p = p_X = p_Y$ and set to zero to obtain pooled success proportion

$$p_e = \frac{x + y}{n + m}$$

- For Ω, take derivatives separately with respect to p_X and p_Y, to obtain

$$p_{X_e} = \frac{x}{n} \quad \text{and} \quad p_{Y_e} = \frac{y}{m}$$

Generalized likelihood ratio

- We have

$$\lambda = \frac{\max_p L(p,p)}{\max_{p_X,p_Y} L(p_X,p_Y)} = \frac{L(p_e,p_e)}{L(p_{X\,e},p_{Y\,e})}$$

- Result is

$$\lambda = \frac{\left(\frac{x+y}{m+n}\right)^{x+y}\left(1-\frac{x+y}{m+n}\right)^{n+m-x-y}}{\left(\frac{x}{n}\right)^{x}\left(1-\frac{x}{n}\right)^{n-x}\left(\frac{y}{m}\right)^{y}\left(1-\frac{y}{m}\right)^{m-y}}$$

- Approximations to the above exist, e.g., $-2\ln\lambda$ has an asymptotic χ^2 distribution with one df. So approximate two-sided $\alpha = 0.05$ test is to reject H_0 if $-2\ln\lambda \geq \chi^2_{0.05,1} = 3.84$.

Approximation to GLR

- Observe that, by the CLT, the following is normally distributed

$$\frac{\frac{X}{n} - \frac{Y}{m} - E\left(\frac{X}{n} - \frac{Y}{m}\right)}{\sqrt{\mathrm{Var}\left(\frac{X}{n} - \frac{Y}{m}\right)}}$$

- Under H_0 we have $E\left(\frac{X}{n} - \frac{Y}{m}\right) = 0$ and

$$\mathrm{Var}\left(\frac{X}{n} - \frac{Y}{m}\right) = \frac{p(1-p)}{n} + \frac{p(1-p)}{m}.$$

- Replace p by $p_e = \frac{x+y}{n+m}$ to obtain a Z statistic.

Two-sample Bernoulli trial test

- Let x and y be the number of successes in two independent Bernoulli trials of n and m flips, respectively.
- Let p_X and p_Y denote the true success probabilities, let $p_e = \frac{x+y}{n+m}$ and define

$$z = \frac{\frac{X}{n} - \frac{Y}{m}}{\sqrt{\frac{p_e(1-p_e)}{n} + \frac{p_e(1-p_e)}{m}}}$$

- Tests are as follows
  - To test $H_0 : p_X = p_Y$ versus $H_1 : p_X > p_Y$ at α level of significance, reject H_0 if $z \geq +z_\alpha$.
  - To test $H_0 : p_X = p_Y$ versus $H_1 : p_X < p_Y$ at α level of significance, reject H_0 if $z \leq -z_\alpha$.
  - To test $H_0 : p_X = p_Y$ versus $H_1 : p_X \neq p_Y$ at α level of significance, reject H_0 if either $z \leq -z_{\alpha/2}$ or $z \geq +z_{\alpha/2}$.

Comment

- This test is more general than it seems.
- Any continuous variable can be dichotomized into a Bernoulli random variable.
- For example, blood pressure can be dichotomized into "normal" and "abnormal."

Summary

- We have studied two-sample confidence intervals for $\mu_X - \mu_Y$.
- We have studied them for both Bernoulli trials and normally distributed data.