Binomial data: Testing $H_0 : p_X = p_Y$
Confidence Intervals for the two-sample problem

Bruce M. Boghosian

Tufts University | School of Arts and Sciences

Department of Mathematics

Tufts University

Outline

Bruce M. Boghosian

Statement of the problem

- The two-sample analysis that we did last time was valid only for pairs of normally distributed data.
- We can do the same thing with other distributions, including discrete distributions.
- We analyze next the situation with n Bernoulli trials.
- Suppose that n independent Bernoulli trials related to treatnent 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 $H_0 : \; p_X = p_Y (= p)$
- Alternative hypothesis $H_1 : p_X \neq p_Y$
- Two parameter spaces for GLRT:

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

- 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 $n = \sum_j^n x_j$ and $m = \sum_j^n y_j$.

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 $n = \sum_j^n x_j$ and $m = \sum_j^n y_j$.

- 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.”

Two-sample hypothesis testing with normal r.v.s

- Meaningful H_0 can always be defined for two-sample tests
- Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ be independent random samples drawn from normal distributions with means μ_X and μ_Y, respectively, and with the same standard deviation σ.
- Let s_p denote the pooled standard deviation.
- A $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ is given by

$$\left(\overline{x} - \overline{y} - t_{\alpha/2,n+m-2} s_p \sqrt{\frac{1}{n} + \frac{1}{m}}, \overline{x} - \overline{y} + t_{\alpha/2,n+m-2} s_p \sqrt{\frac{1}{n} + \frac{1}{m}}\right)$$

- **Pf.:**
  - We know $\frac{\overline{X}-\overline{Y}-(\mu_x-\mu_Y)}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}}$ is Student T distributed with $n+m-2$ df, so

$$P\left(-t_{\alpha/2,n+m-2} \leq \frac{\overline{X}-\overline{Y}-(\mu_x-\mu_Y)}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \leq +t_{\alpha/2,n+m-2}\right) = 1-\alpha$$

  - Rearrange inequality to isolate $\mu_X - \mu_Y$ to obtain confidence interval.

Confidence intervals for the variance ratio

- Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ be independent random samples drawn from normal distributions with standard deviations σ_X and σ_Y, respectively. A $100(1-\alpha)\%$ confidence interval for the variance ratio σ_X^2/σ_Y^2 is

$$\left(\frac{s_X^2}{s_Y^2} F_{\alpha/2,m-1,n-1}, \frac{s_X^2}{s_Y^2} F_{1-\alpha/2,m-1,n-1} \right)$$

- **Pf.:**
  - Note that $\frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2}$ is F distributed with $m-1$ and $n-1$ df.
  - Same strategy: Write probability

$$P\left(f_{1-\alpha/2,m-1,n-1} \leq \frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2} \leq f_{\alpha/2,m-1,n-1} \right) = 1-\alpha.$$

  - Isolate σ_X^2/σ_Y^2 in inequality.

Confidence intervals for two-sample Bernoulli trials

- Let x and y denote the number of successes in two independent sets of n and m Bernoulli trials, respectively.
- If p_X and p_Y denote the true success probabilities, an approximate $100(1-\alpha)\%$ confidence interval for $p_X - p_Y$ is given by

$$\left(\frac{x}{n} - \frac{y}{m} - z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n} + \frac{\frac{y}{m}\left(1-\frac{y}{m}\right)}{m}},\right.$$

$$\left.\frac{x}{n} - \frac{y}{m} + z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n} + \frac{\frac{y}{m}\left(1-\frac{y}{m}\right)}{m}}\right)$$

Comment

- We have studied confidence intervals for the two-sample problem.
- We have studied them for both Bernoulli trials and normally distributed data.
- We have studied them for both $\mu_X - \mu_Y$ and σ_X^2/σ_Y^2.