

Inferences about population proportions

Single population model: $Y \sim \text{binomial}(n, \pi)$

$$P(Y = y) = \frac{n!}{y!(n - y)!} \pi^y (1 - \pi)^{n - y}$$

with $y = 0, 1, 2, \dots, n$ and $0 \leq \pi \leq 1$. Also

$$E(Y) = n\pi$$

$$V(Y) = n\pi(1 - \pi)$$

Data: y, n . The observation y is assumed to have been generated by a binomial distribution with parameters n (known) and π (unknown). Want to make inferences about π .

- ex.'s**
- random sample of n voters, observe y Democrats
 - radio collar n adult deer; observe y alive after 1 yr
 - n insects each given a dose of pesticide; observe y alive

(Note: response on each trial is failure or success. The response is categorical. Y is a count. Also, Y is a sum: $Y = I_1 + I_2 + \dots + I_n$, where each indicator variable takes the values 0 (failure) or 1 (success)).

Estimation of π (data: y, n)

Likelihood:

$$L = \frac{n!}{y!(n-y)!} (\pi)^y (1-\pi)^{n-y}$$

$$\hat{\pi} = \frac{y}{n} \quad \text{ML estimate}$$

$$\hat{L} = \frac{n!}{y!(n-y)!} \left(\frac{y}{n}\right)^y \left(1 - \frac{y}{n}\right)^{n-y} \quad \text{maximized likelihood}$$

Recall: Y is a sum; CLT $\Rightarrow Y_{\text{approx}} \sim \text{normal}(n\pi, n\pi(1-\pi))$

$\hat{\pi} = \frac{1}{n}Y$ is a constant $\times Y$ (& is also a sum); CLT \Rightarrow

$$\hat{\pi}_{\text{approx}} \sim \text{normal}\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

Approximate $100(1-\alpha)\%$ CI for π

$$\hat{\pi} \pm z_{\alpha/2} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

Computer studies have revealed that the CI based on the ML estimate does not converge very fast to the stated coverage (95%, etc.). The CI is too narrow; a *large n* is required to attain “asymptopia”

$$\tilde{\pi} = \frac{y+2}{n+4} \quad \text{Wilson estimate (much better)}$$

Slight adjustment to ML adds a little bias, but improves convergence of CLT & produces CIs with good coverage properties (Agresti, A. & Coull, B. A. 1998. *The American Statistician* 52:119-126)

$$\widehat{V}(\tilde{\pi}) = \frac{\tilde{\pi}(1-\tilde{\pi})}{n+4}$$

$$\text{CI: } \tilde{\pi} \pm z_{\alpha/2} \sqrt{\frac{\tilde{\pi}(1-\tilde{\pi})}{n+4}}$$

Text: adjusted estimators of, & CIs for, π when y is 0 or 1:

$$\hat{\pi}_{\text{adj}} = \frac{\frac{3}{8}}{n+\frac{3}{4}}, \quad y = 0; \quad \hat{\pi}_{\text{adj}} = \frac{n+\frac{3}{8}}{n+\frac{3}{4}}, \quad y = n$$

$$y = 0; \quad \text{CI: } \left(0, 1 - \left(\frac{\alpha}{2}\right)^{\frac{1}{n}}\right) \quad (\text{asymmetric})$$

$$y = 1; \quad \text{CI: } \left(\left(\frac{\alpha}{2}\right)^{\frac{1}{n}}, 1\right) \quad (\text{asymmetric})$$

Hypothesis tests for π

$$H_0: \pi = \pi_0 \text{ (known constant)}$$

$$H_a: \pi \left\{ \begin{array}{l} > \\ < \\ \neq \end{array} \right\} \pi_0$$

Test statistic:

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \text{ (approx. normal}(0, 1) \text{ under } H_0)$$

Rejection region:

$$\text{reject } H_0 \text{ if } \left\{ \begin{array}{l} z \geq z_\alpha \\ z \leq -z_\alpha \\ |z| \geq z_{\alpha/2} \end{array} \right\}$$

Likelihood ratio approach for testing $H_a: \pi \neq \pi_0$

$$\hat{L}_0 = \frac{n!}{y!(n-y)!} (\pi_0)^y (1 - \pi_0)^{n-y}$$

$$\hat{L}_a = \frac{n!}{y!(n-y)!} \left(\frac{y}{n}\right)^y \left(1 - \frac{y}{n}\right)^{n-y}$$

$$G^2 = -2 \log_e \left(\frac{\hat{L}_0}{\hat{L}_a} \right) \underset{\text{approx}}{\sim} \text{chi-square}(1) \text{ under } H_0$$

$$\text{reject } H_0 \text{ if } G^2 \geq \chi_\alpha^2$$

LR statistic can be used to build CI's; for a 95% CI, take set of all values of π_0 for which $G^2 < 3.84$ ($= \chi_{0.05}^2$):

This gives an asymmetric interval which has fairly good coverage properties (called the **profile likelihood CI**)

Comparing two proportions

Model: $Y_1 \sim \text{binomial}(n_1, \pi_1)$, $Y_2 \sim \text{binomial}(n_2, \pi_2)$;
interest is in comparing π_1 and π_2 .

Data: y_1, n_1, y_2, n_2

ML estimates:

$$\hat{\pi}_1 = \frac{y_1}{n_1} \quad \sqrt{\hat{V}(\hat{\pi}_1)} = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1}}$$

$$\hat{\pi}_2 = \frac{y_2}{n_2} \quad \sqrt{\hat{V}(\hat{\pi}_2)} = \sqrt{\frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

Wilson estimates (add two successes and two failures, just like before, only split them between both samples)

$$\tilde{\pi}_1 = \frac{y_1+1}{n_1+2} \quad \sqrt{\widehat{V}(\tilde{\pi}_1)} = \sqrt{\frac{\tilde{\pi}_1(1-\tilde{\pi}_1)}{n_1+2}}$$

$$\tilde{\pi}_2 = \frac{y_2+1}{n_2+2} \quad \sqrt{\widehat{V}(\tilde{\pi}_2)} = \sqrt{\frac{\tilde{\pi}_2(1-\tilde{\pi}_2)}{n_2+2}}$$

100(1 - α)% CI for $\pi_1 - \pi_2$ (using Wilson)

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

Hypothesis tests for $\pi_1 - \pi_2$

$$H_0: \pi_1 - \pi_2 = 0 \quad (\pi_1 = \pi_2 = \pi; \quad \hat{\pi} = \frac{y_1+y_2}{n_1+n_2})$$

$$H_a: \pi_1 - \pi_2 \left\{ \begin{array}{l} > \\ < \\ \neq \end{array} \right\} 0$$

Test statistic:

$$z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n_1} + \frac{\hat{\pi}(1-\hat{\pi})}{n_2}}}$$

Rejection region:

$$\text{reject } H_0 \text{ if } \left\{ \begin{array}{l} z \geq z_\alpha \\ z \leq -z_\alpha \\ |z| \geq z_{\alpha/2} \end{array} \right\}$$