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 \( y = 0, 1, 2, \ldots, n \) 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 \( \pi \) (unknown). Want to make inferences about \( \pi \).

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: \( Y = I_1 + I_2 + \cdots + I_n \), where each indicator variable takes the values 0 (failure) or 1 (success)).
Estimation of $\pi$ (data: $y, n$)

Likelihood:

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

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

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

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

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

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

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

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

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

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)

$$\hat{\text{V}}(\hat{\pi}) = \frac{\hat{\pi}(1-\hat{\pi})}{n+4}$$

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

Text: adjusted estimators of, & CIs for, $\pi$ when $y$ is 0 or $n$:

$$\hat{\pi}_{\text{adj}} = \frac{\frac{3}{8}}{n+\frac{1}{4}}, \ y = 0; \ \hat{\pi}_{\text{adj}} = \frac{n+\frac{3}{8}}{n+\frac{3}{4}}, \ y = n$$

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

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

$H_0$: $\pi = \pi_0$ (known constant)

$H_a$: $\begin{cases} \pi > \pi_0 \\ \pi < \pi_0 \\ \pi \neq \pi_0 \end{cases}$

Test statistic:

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

Rejection region:

$$\text{reject } H_0 \text{ if } \begin{cases} z \geq z_\alpha \\ z \leq -z_\alpha \\ |z| \geq z_{\alpha/2} \end{cases}$$

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

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

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

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

$$\text{reject } H_0 \text{ if } G^2 \geq \chi^2_{\alpha/2}$$

LR statistic can be used to build CI's; for a 95% CI, take set of all values of $\pi_0$ for which $G^2 < 3.84 \ (= \chi^2_{0.05})$: 
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 \( \pi_1 \) and \( \pi_2. \)

Data: \( y_1, n_1, y_2, n_2 \)

ML estimates:

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

\[
\hat{\pi}_2 = \frac{y_2}{n_2} \quad \sqrt{V(\hat{\pi}_2)} = \sqrt{\frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}
\]
Wilson estimates (add two successes and two failures, just like before, only split them between both samples)

\[ \tilde{\pi}_1 = \frac{Y_1 + 1}{n_1 + 2} \quad \sqrt{V(\tilde{\pi}_1)} = \sqrt{\frac{\tilde{\pi}_1(1 - \tilde{\pi}_1)}{n_1 + 2}} \]

\[ \tilde{\pi}_2 = \frac{Y_2 + 1}{n_2 + 2} \quad \sqrt{V(\tilde{\pi}_2)} = \sqrt{\frac{\tilde{\pi}_2(1 - \tilde{\pi}_2)}{n_2 + 2}} \]

100(1 - \alpha)\% 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 \( \pi_1 - \pi_2 \)

\[ H_0: \pi_1 - \pi_2 = 0 \quad (\pi_1 = \pi_2 = \pi; \quad \hat{\pi} = \frac{Y_1 + Y_2}{n_1 + n_2}) \]

\[ H_a: \pi_1 - \pi_2 \left\{ \begin{array}{ll} > 0 \\ < 0 \\ \neq 0 \end{array} \right\} \]

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:

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

**Likelihood ratio approach for comparing two binomial proportions (PROC GENMOD)**

Data: $y_1, n_1$; $y_2, n_2$.

Maximized likelihoods under $H_0$ and $H_a$:

$$\hat{L}_0 = \left( \begin{array}{c} n_1 \\ y_1 \end{array} \right) \hat{\pi}^{y_1} (1 - \hat{\pi})^{n_1-y_1} \left( \begin{array}{c} n_2 \\ y_2 \end{array} \right) \hat{\pi}^{y_2} (1 - \hat{\pi})^{n_2-y_2}$$

where $\hat{\pi} = \frac{y_1 + y_2}{n_1 + n_2}$

$$\hat{L}_a = \left( \begin{array}{c} n_1 \\ y_1 \end{array} \right) \hat{\pi}_1^{y_1} (1 - \hat{\pi}_1)^{n_1-y_1} \left( \begin{array}{c} n_2 \\ y_2 \end{array} \right) \hat{\pi}_2^{y_2} (1 - \hat{\pi}_2)^{n_2-y_2}$$

where $\hat{\pi}_1 = \frac{y_1}{n_1}$, $\hat{\pi}_2 = \frac{y_2}{n_2}$

$$G^2 = -2 \log \left( \frac{\hat{L}_0}{\hat{L}_a} \right)^{\text{approx}} \chi^2(v)$$

$$\text{df} \quad v = 2 - 1 = 1$$
Logit “link” function used in PROC GENMOD (re-parameterizes in terms of real-valued parameters instead of parameters constrained between 0 and 1):

\[
\log\left(\frac{\pi_1}{1-\pi_1}\right) = \lambda + \alpha \\
\pi_1 = \frac{1}{1 + \exp[-(\lambda + \alpha)]},
\]

or

\[
\log\left(\frac{\pi_2}{1-\pi_2}\right) = \lambda \\
\pi_2 = \frac{1}{1 + \exp(-\lambda)},
\]

Here \(\alpha\) is an adjustment for treatment 1.

\[
H_0: \pi_1 = \pi_2 \iff H_0: \alpha = 0 \\
H_a: \pi_1 \neq \pi_2 \iff H_a: \alpha \neq 0
\]

Aspirin data, SAS PROC GENMOD output:

\[
G^2 = 24.27 \quad P < .0001
\]

\[
\hat{\lambda} = -4.0466 \quad \hat{\alpha} = -0.6052 \quad \text{ML estimates}
\]

\[
\hat{\pi}_1 = \frac{1}{1 + \exp[-(\hat{\lambda} + \hat{\alpha})]} = .0094542 \\
( \frac{y_1}{n_1} = \frac{104}{11000} = .0094542)
\]

\[
\hat{\pi}_2 = \text{etc.}
\]