The Big Idea:
We are often interested in estimating the proportion $p$ of a population that has some feature (e.g. “supports Obama”, “is diabetic”, or “responds to a drug treatment”). To estimate $p$ we take an SRS of size $n$ and form $\hat{p}$, the proportion of our sample that has the feature. To give a confidence interval for $\hat{p}$, we use the fact that if our sample was large, $\hat{p}$ is distributed approximately $N(p, \sqrt{p(1-p)/n})$. Our capacity to give confidence intervals and calculate $P$-values then emerges from the fact that we can use a $Z$-table to calculate probabilities for normal random variables.

Formulas

- We can use the normal approximation when $np \geq 10$ and $n(1-p) \geq 10$.
- The standard error (i.e. our approximate standard deviation) for $\hat{p}$ is

$$S.E. = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- A an approximate $C$ level confidence interval for $p$ is

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

where we choose $z^*$ so that a $N(0,1)$ curve has area $C$ between $z^*$ and $-z^*$.
- To calculate $P$-values for a null hypothesis of the form $H_0 : p = p_0$, first form the $z$-statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

Then
- if $H_a : p < p_0$, the approximate $P$-value is the area under a $N(0,1)$ curve to the left of $z$. (Assuming $z$ is negative; if not, do not report a $P$-value.)
- if $H_a : p > p_0$, the approximate $P$-value is the area under a $N(0,1)$ curve to the right of $z$. (Assuming $z$ is positive; if not, do not report a $P$-value.)
- if $H_a : p \neq p_0$, the approximate $P$-value is the area under a $N(0,1)$ curve to the right of $|z|$ and to the left of $-|z|$.