Bruce M. Boghosian

Interval Estimation

Confidence Intervals for the Binomial Parameter p

Bruce M. Boghosian

Tufts University | School of Arts and Sciences

Department of Mathematics

Tufts University

Outline

Confidence intervals for the binomial parameter p

- We conduct n Bernoulli trials with heads probability p.
- For one trial, mean is p, standard deviation is $\sqrt{p(1-p)}$
- For n trials, we have a binomial probability distribution with mean p and standard deviation $\sqrt{p(1-p)/n}$
- Using MLE or MM, we have $p_e = \frac{1}{n}\sum_{j=1}^{n} k_j$, so for large n

$$Z = \frac{p_e - p}{\sqrt{p_e(1-p_e)/n}}$$

will be distributed like a standard normal, by the CLT.

Margin of error

- *Margin of error* is half max width of confidence interval.
- Let k be the number of successes in n Bernoulli trials.
- Estimate is $p_e = k/n$.
- Confidence interval is $\left[p_e - z_{\alpha/2}\frac{\sigma_e}{\sqrt{n}},\ p_e + z_{\alpha/2}\frac{\sigma_e}{\sqrt{n}}\right]$
- Width of confidence interval is $2z_{\alpha/2}\frac{\sigma_e}{\sqrt{n}} = \frac{z_{\alpha/2}(4\sigma_e)}{2\sqrt{n}}$
- Estimate of standard deviation is $\sigma_e = p_e(1-p_e)$
- Problematic because we don't know p_e in advance.
- Note that $4\sigma_e = 4p_e(1-p_e)$ has upper bound of one.
- Margin of error is $100d\%$ where $d = \frac{z_{\alpha/2}}{2\sqrt{n}}$
- Usually $\alpha = 0.05$, but other values of α are possible.

Choosing sample sizes

Bruce M. Boghosian

- Largest interval half width possible is $d = \frac{z_{\alpha/2}}{2\sqrt{n}}$.
- We have in general

$$\text{Prob}\left(-d \le \frac{1}{n}\sum_{j=1}^{n} x_j - p \le +d\right) = 1 - \alpha.$$

- This can be regarded as an equation for the minimum value of n needed to attain the confidence α, and margin of error $100d\%$.
- For fixed n, you can have more confidence in estimates with larger margins of error
- Likewise, you can have smaller margins of error, but you may have less confidence in those.

Comparing normal and binomial interval estimation

- In all examples with normal data, we *specified* variance σ_0.
- We found confidence intervals for the estimate of the mean

$$\mu_e = \frac{1}{n}\sum_{j=1}^{n} x_j$$

- Rather than insist on a priori knowledge of σ_0, why not use

$$\sigma_e = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(x_j - \mu_e\right)^2} \quad ?$$

- After all, for the binomial distribution, we had no hesitation about using both

$$p_e = \frac{1}{n}\sum_{j=1}^{n} k_j \quad \text{and} \quad \sigma_e = \sqrt{p_e(1-p_e)}.$$

An important distinction

- For the binomial distribution, the mean is p and the standard deviation is $\sqrt{p(1-p)}$. The latter is completely determined by the former.
- For the normal distribution, mean μ and standard deviation σ are two separately specifiable parameters, each with its own estimator.
- When we use an estimator to find μ_e from our n data points, we effectively "use up" a data point.
- When we use μ_e in the calculation of an average to obtain σ_e, our average is effectively over only $n-1$ points.
- For this reason, the *sample standard deviation* used for interval estimation for normally distributed data is not that given by the MLE (or MM) estimator.

Summary

- We constructed confidence intervals for Bernoulli trials.
- We have defined margin of error.
- We have shown how to estimate needed sample sizes.
- We have contrasted interval estimation with normal and binomial data.