Statistics 5114: Odds Ratios: Delta Method, Bootstrap and Bayes

For each homework assignment, turn in at the beginning of class on the indicated due date. Late assignments will only be accepted with special permission. Write each problem up *very* neatly (LaTeX is preferred). Show all of your work.

Problem 1

Consider the Bernoulli sampling model, defined by:

$$Pr(x_i = 1) = p, \text{ and } Pr(x_i = 0) = 1 - p.$$

A good estimator of p is the sample average $\hat{p} = \bar{X} = \sum_i x_i/N$. Notice that $E[\bar{X}] = p$, and $Var(\bar{X}) = \frac{p(1-p)}{N}$. Additionally, $\bar{X} \to N(p, p(1-p)/N)$. Say we wish to estimate the odds ratio $g(p) = p/(1-p)$.

Part a

The Delta method says:

$$\sqrt{N}(g(\bar{X}) - g(p)) \to N(0, [g'(p)]^2 Var(x_i)). \tag{1}$$

For this problem, sample $x_i \sim Bernoulli(0.5), i = 1, \ldots, N$, with N=1,000. Compute $g(\bar{X})$. Repeat this 5,000 times, so that you'll have 5,000 realizations of $g(\bar{X})$. Plot a histogram of these samples, and compare to the result implied by equation (1) (i.e. plot the correct normal curve over your histogram. How does it look?).

Part b

Repeat for $N = \{10, 30, 50, 100, 500\}$.

Problem 2

For each of the samples sizes $N = \{10, 30, 100, 1000\}$, and true parameters $p = \{0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 0.99\}$ (there are 28 simulation configurations), sample N

Bernoulli realizations. Based on these N sample points, using the delta method, approximate both the expected value and variance of $g(\bar{X})$. Recall that you're pretending that you don't know p here, so you need to use:

$$\hat{E} = \hat{p}/(1-\hat{p})$$

and

$$\hat{V} = \left([g'(p)]^2 Var(\bar{X})\right)_{p=\hat{p}}.$$

For each configurations, repeat the exercise 100 times (You will have 100 random estimates of $E[g(\bar{X})]$ and $Var(g(\bar{X}))$. Save the results, as you will compare these to results obtained from Bootstrapping, and a Bayesian analysis.

Problem 2

Using the samples from Problem 2, estimate both the expected value and variance of $g(\bar{X})$ using the bootstrap. That is, for each sample realization, create a bootstrap distribution of $g(\bar{X})$ and estimate both the mean and the variance. For each random sample, you will create B bootstrap realizations: $\{g(\bar{X}_{(b)})|b = 1, \ldots, B\}$, and compute both:

$$\bar{g} = \sum_b g(\bar{X}_{(b)})/B$$

and

$$\sigma_g^2 = \sum_b (g(\bar{X}_{(b)}) - \bar{g})^2/(B-1).$$

For each configuration, repeat the exercise 100 times.

Problem 3

In this problem, you will perform a Bayesian analysis of $p/(1-p)$. Keep in mind, that in classical statistics, we use plug in estimators of the data to form $\hat{p}/(1-\hat{p})$; however, in Bayesland, we form the distribution of $z = p/(1-p)$, conditioned on the observed data. That is, we will form the posterior distribution (distribution seen a-poseriori to (after) seeing the data)

$$p(z|x_1, \ldots, x_N) = \frac{L(z|x_1, \ldots, x_N)p(z)}{\int L(z|x_1, \ldots, x_N)p(z)dz},$$

where $L(z|x_1, \ldots, x_N)$ is the likelihood function, and $p(z)$ is the prior on z (distribution known a-priori (before) seeing the data). Now, one of the nice features of a Bayesian analysis is that it's completely probabilistic. That is, our inferences are on z, which follows a probability distribution. We need to simply find this probability distribution. First, we'll find the probability distribution of p and then do a simple

transformation to find the distribution on z. Using the $Beta(1/2, 1/2)$ prior, form the posterior for p. Why $p \sim Beta(1/2, 1/2)$ is a good question. We know that a $Beta(1/2, 1/2)$ has some nice properties: 1) it makes finding the posterior on p very easy (i.e. it's the conjugate choice), 2) it has optimal frequentist properties, and 3) it is invariant to transformations. We have seen that the first property is true, but the other 2 reasons are reserved for classes like STAT 5444.

Part a

Given $x_1, \ldots, x_N$, find the likelihood function for p.

Part b

Given $p \sim Beta(1/2, 1/2)$, find $\pi(p|x_1, \ldots, x_N)$.

Part c

Using $z = p/(1-p)$, transform your result found in Part b to find $\pi(z|x_1, \ldots, x_N)$.

Part d

Using the samples from Problem 2, estimate both the expected value and variance of $z|x_1, \ldots, x_N$. That is compute the expectation and variance for the distribution of the posterior for $p/(1-p)$ based on $x_1, \ldots, x_N$. I don't care how you do this. You can either estimate these quantities using Monte Carlo, or work them out analytically.

Problem 4

Provide side-by-side box plots showing your simulation results (both expectations and variances across each of the three methods). Also, give a brief discussion or your results and conclusions.

Comments and Potential Pitfalls

While I've explained quite a lot in this exercise, you are going to discover "real-world" issues as you perform this study. Here's one such issue: when the sample size is small, some of these estimates might not exist (why?). For instance, in the Bayesian analysis, the transformed distribution is called a Beta-Prime distribution (look it up on wiki). Under small sample sizes, the posterior expectation and variance won't exist (although the posterior distribution is proper (say what?)). So instead of reporting the expected value and variance, it might make more sense to report the *mode* of the posterior. This is referred to as the MAP ('Max A-Posteriori') estimator (how do you

find this?) and a 95% credible interval (what's that? what would you report?). So, you may use your own good judgement in reporting statistics in the various simulation studies.

WWSD?: For the Bayesian analysis, he'd report the MAP and the coverage rate of the simulated intervals.