STATS 200: Homework 6

Due Wednesday, November 16, at 5PM

1. **A discrete model (based on Rice 8.4).** Suppose that X is a discrete random variable with

$$\mathbb{P}[X=0] = \frac{2}{3}\theta$$
$$\mathbb{P}[X=1] = \frac{1}{3}\theta$$
$$\mathbb{P}[X=2] = \frac{2}{3}(1-\theta)$$
$$\mathbb{P}[X=3] = \frac{1}{3}(1-\theta)$$

where $0 \leq \theta \leq 1$ is a parameter. The following 10 independent observations were taken from such a distribution: $\{3,0,2,1,3,2,1,0,2,1\}$. (For parts (a) and (b), feel free to use any asymptotic approximations you wish, even though $n = 10$ here is rather small.)

(a) Find the method of moments estimate of θ, and compute an approximate standard error of your estimate using asymptotic theory.

(b) Find the maximum likelihood estimate of θ, and compute an approximate standard error of your estimate using asymptotic theory. (Hint: Your formula for the log-likelihood based on n observations $X_1, \ldots, X_n$ should depend on the numbers of 0's, 1's, 2's, and 3's in this sample.)

(c) Compute, instead, an approximate standard error for your MLE in part (b) using the nonparametric bootstrap and $B = 10000$ bootstrap simulations. (Provide both your code and the standard error estimate.)

(In R, `sample(X,n,replace=TRUE)` returns a vector containing n samples with replacement from a vector `X` of length n.)

2. **Confidence intervals for a binomial proportion.**
Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(p)$ be n tosses of a biased coin, and let $\hat{p} = \bar{X}$. In this problem we will explore two different ways to construct a 95% confidence interval for p, both based on the Central Limit Theorem result

$$\sqrt{n}(\hat{p} - p) \to \mathcal{N}(0, p(1-p)). \tag{1}$$

(a) Use the plugin estimate $\hat{p}(1-\hat{p})$ for the variance $p(1-p)$ to obtain a 95% confidence interval for p. (This is the procedure discussed in Lecture 19, yielding the Wald interval for p.)

(b) Instead of using the plugin estimate $\hat{p}(1-\hat{p})$, note that equation (1) implies, for large n,

$$\mathbb{P}_p\left[-\sqrt{p(1-p)}z(\alpha/2) \leq \sqrt{n}(\hat{p}-p) \leq \sqrt{p(1-p)}z(\alpha/2)\right] \approx 1-\alpha.$$

Solve the equation $\sqrt{n}(\hat{p}-p) = \sqrt{p(1-p)}z(\alpha/2)$ for p in terms of $\hat{p}$, and solve the equation $\sqrt{n}(\hat{p}-p) = -\sqrt{p(1-p)}z(\alpha/2)$ for p in terms of $\hat{p}$, to obtain a different 95% confidence interval for p.

(c) Perform a simulation study to determine the true coverage of the confidence intervals in parts (a) and (b), for the 9 combinations of sample size $n = 10$, 40, 100 and true parameter $p = 0.1$, 0.3, 0.5. (For each combination, perform at least $B = 100{,}000$ simulations. In each simulation, you may simulate $\hat{p}$ directly from $\frac{1}{n}\,\text{Binomial}(n,p)$ instead of simulating $X_1,\ldots,X_n$. E.g. in R, `phat = rbinom(1,n,p)/n`.) Report the simulated coverage levels in two tables. Which interval yields true coverage closer to 95% for small values of n?

3. **MLE in a misspecified model.** Suppose you fit the model Exponential(λ) to data $X_1,\ldots,X_n$ by computing the MLE $\hat{\lambda} = 1/\bar{X}$, but the true distribution of the data is $X_1,\ldots,X_n \overset{IID}{\sim} \text{Gamma}(2,1)$.

(a) Let $f(x|\lambda)$ be the PDF of the Exponential(λ) distribution, and let $g(x)$ be the PDF of the Gamma(2, 1) distribution. Compute an explicit formula for the the KL-divergence $D_{\text{KL}}(g(x)\|f(x|\lambda))$ in terms of λ, and find the value λ^* that minimizes this KL-divergence.

(You may use the fact that if $X \sim \text{Gamma}(\alpha,\beta)$, then $\mathbb{E}[X] = \alpha/\beta$ and $\mathbb{E}[\log X] = \psi(\alpha) - \log\beta$ where ψ is the digamma function.)

(b) Show directly, using the Law of Large Numbers and the Continuous Mapping Theorem, that the MLE $\hat{\lambda}$ converges in probability to λ^* as $n \to \infty$.

(c) Perform a simulation study for the standard error of $\hat{\lambda}$ with sample size $n = 500$, as follows: In each of $B = 10000$ simulations, sample $X_1,\ldots,X_n \overset{IID}{\sim} \text{Gamma}(2,1)$, compute the MLE $\hat{\lambda} = 1/\bar{X}$ for the exponential model, compute an estimate of the standard error of $\hat{\lambda}$ using the Fisher information $I(\hat{\lambda})$, and compute also the sandwich estimate of the standard error, $S_X/(\bar{X}^2\sqrt{n})$, derived in Lecture 16.

Report the true mean and standard deviation of $\hat{\lambda}$ that you observe across your 10000 simulations. Is the mean close to λ^* from part (a)? Plot a histogram of the 10000 estimated standard errors using the Fisher information, and also plot a histogram of the 10000 estimated standard errors using the sandwich estimate. Do either of these methods for estimating the standard error of $\hat{\lambda}$ seem accurate in this setting?

4. **The delta method for two samples.** Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(p)$, and let $Y_1, \ldots, Y_m \overset{IID}{\sim} \text{Bernoulli}(q)$, where the X_i's and Y_i's are independent. For example, $X_1, \ldots, X_n$ may represent, among n individuals exposed to a certain risk factor for a disease, which individuals have this disease, and $Y_1, \ldots, Y_m$ may represent, among m individuals not exposed to this risk factor, which individuals have this disease. The **odds-ratio**

$$\frac{p}{1-p}\Big/\frac{q}{1-q}$$

provides a quantitative measure of the association between this risk factor and this disease. (For more details, see Rice Section 13.6.) The log-odds-ratio is the (natural) logarithm of this quantity,

$$\log\left(\frac{p}{1-p}\Big/\frac{q}{1-q}\right).$$

(a) Suggest reasonable estimators $\hat{p}$ and $\hat{q}$ for p and q, and suggest a plugin estimator for the log-odds-ratio.

(b) Using the first-order Taylor expansion

$$g(\hat{p}, \hat{q}) \approx g(p, q) + (\hat{p} - p)\frac{\partial g}{\partial p}(p, q) + (\hat{q} - q)\frac{\partial g}{\partial q}(p, q)$$

as well as the Central Limit Theorem and independence of the X_i's and Y_i's, derive an asymptotic normal approximation to the sampling distribution of your plugin estimator in part (a). (Hint: Recall the proof of the delta method from Lecture 18.)

(c) Give an approximate 95% confidence interval for the log-odds-ratio $\log \frac{p}{1-p}/\frac{q}{1-q}$. Translate this into an approximate 95% confidence interval for the odds-ratio $\frac{p}{1-p}/\frac{q}{1-q}$. (You may use a plugin estimate for the variance of the normal distribution that you derived in part (b).)