Prof. David Draper
Department of Statistics
University of California, Santa Cruz

STAT 131: Take-Home Test 3, part 2 (extra credit) *[250 total points]*

Due date: upload to `canvas.ucsc.edu` by **11.59pm Sun 14 Jun 2020**

3. *[130 total points]* (binomial and negative binomial sampling) You and I are both getting ready to sample from a Bernoulli process with unknown success probability $0 < \theta < 1$. You decide to use *binomial sampling*: you propose to

(1) set a fixed known number $n \geq 1$ of Bernoulli trials in advance,

(2) observe that many trials, and

(3) record the random number S of successes you see.

I instead propose to use *negative binomial sampling*: I'll watch the same process that you do, but I'll

(1′) set a fixed known number $s \geq 1$ of successes in advance,

(2′) observe the Bernoulli trials until I've seen s successes, and

(3′) record the random number N of trials that were needed to get that many successes.

Question, to be answered by parts (a–c) of this problem below: if your S equals my s and my N equals your n, should you and I draw essentially the same conclusions about θ?

(a) Briefly explain why your probability model for S should be Binomial(n, θ), so that your S has PMF

$$f_S(s \,|\, n, \theta) = \left(\begin{array}{c} n \\ s \end{array} \right) \theta^s (1-\theta)^{n-s} \, I_{\{0,1,\ldots,n\}}(s) \,, \tag{1}$$

and why a natural estimator of θ for you to use is therefore $\hat{\theta}_B = \frac{S}{n}$. Show that $E\left(\hat{\theta}_B\right) = \theta$, so that $\hat{\theta}_B$ is unbiased; show further that $SE\left(\hat{\theta}_B\right) \triangleq \sqrt{V\left(\hat{\theta}_B\right)} = \sqrt{\frac{\theta(1-\theta)}{n}}$; and briefly explain under what conditions the distribution of $\hat{\theta}_B$ should be approximately Normal. *[50 points]*

(b) Recall that if X records the number of failures before the sth success, then $X \sim$ Negative Binomial(s, θ), with PMF

$$f_X(x \,|\, s, \theta) = \left(\begin{array}{c} s+x-1 \\ x \end{array} \right) \theta^s (1-\theta)^x \, I_{\{0,1,\ldots\}}(x) \,. \tag{2}$$

(i) Briefly explain why the random N I'll observe with my sampling method is related to X via the simple expression $N = X + s$. *[10 points]*

(ii) Show that the PMF of N is

$$f_N(n \mid s, \theta) = \left(\begin{array}{c} n-1 \\ s-1 \end{array} \right) \theta^s (1-\theta)^{n-s} \, I_{\{s, s+1, \ldots\}}(n) \tag{3}$$

(*Hint:* use Theorem 1.8.3 of DS (page 34): for all integers $n \geq 1$ and all integers $k = 0, 1, \ldots, n$, $\left(\begin{array}{c} n \\ k \end{array} \right) = \left(\begin{array}{c} n \\ n-k \end{array} \right)$) *[10 points].*

Notice how similar equations (1) and (3) are; this encourages the idea that you and I will get more or less the same answers about θ if I use the estimator $\hat{\theta}_{NB} = \frac{s}{N}$.

(iii) Use results from class or DS about $E(X)$ and $V(X)$ to show that $E(N) = \frac{s}{\theta}$ and $V(N) = \frac{s(1-\theta)}{\theta^2}$ *[10 points].* Then use the Delta Method with your results about N to show that $E\left(\hat{\theta}_{NB}\right) \doteq \theta$, so that $\hat{\theta}_{NB}$ is approximately unbiased, and that $SE\left(\hat{\theta}_{NB}\right) \triangleq \sqrt{V\left(\hat{\theta}_{NB}\right)} \doteq \sqrt{\frac{\theta(1-\theta)}{E(N)}}$ *[20 points].*

(iv) Use Jensen's Inequality to show that — in a refinement to the Delta Method — $E\left(\hat{\theta}_{NB}\right) > \theta$, so that $\hat{\theta}_{NB}$ is actually biased on the high side. It can be shown (you're not asked to show this) that $E\left(\frac{s-1}{N-1}\right) = \theta$ (call this fact $(*)$); for a fixed observed value n of N, use $(*)$ to show that the bias of $\hat{\theta}_{NB}$ goes to 0 like $\frac{1}{n}$, so that — for large N — $\hat{\theta}_{NB}$ is indeed approximately unbiased. *[20 points]*

(c) Looking at the expressions for the means and standard errors (SEs) of $\hat{\theta}_B$ and $\hat{\theta}_{NB}$, is it true that you and I will come to pretty much the same conclusions about θ with our different but related sampling methods? Explain briefly. *[10 points]*

4. *[120 total points]* (public health) In one of the largest human experiments ever conducted, in 1954 a randomized controlled trial was run to see whether a vaccine developed by a doctor named Jonas Salk was effective in preventing paralytic polio. A total of 401,974 children (ages 6–9), chosen to be representative of those who might be susceptible to the disease, were randomized to two groups: 200,745 children (the control group C) were injected with a harmless saline solution (a placebo) and the other 201,229 children (the treatment group T) were injected with Salk's vaccine.

(a) What was the point of giving saline solution to the children who didn't get the vaccine? Explain briefly. *[10 points]*

(b) In experimental design, *double-blinding* is the process by which neither the subjects nor the people running the experiment know the treatment-control status of the subjects at the time the outcome of interest is measured for each subject. Would it have been possible to run this experiment in a double-blinded fashion? Would it have been a good idea to do so? Explain briefly. *[10 points]*

(c) The results of the trial were as follows: 33 of the 201,229 children who got the vaccine later developed paralytic polio, whereas 115 of the 200,745 saline children suffered this fate. Let $\hat{\theta}_T = \frac{33}{201229} \doteq 0.0001640$ and $\hat{\theta}_C = \frac{115}{200745} \doteq 0.0005729$ be the observed polio incidences in the T and C groups, respectively. Does the difference between these rates seem large to you in practical terms? Build a probability model for this situation, being explicit about

all assumptions you make and why they're reasonable, and use your model to construct a 99.9% confidence interval for the population mean difference in rates of polio between the two groups. Sketch your confidence interval with $\left(\hat{\theta}_C - \hat{\theta}_T\right)$ as the center, locating the left and right endpoints, the center and the reference point of 0. Is the observed difference statistically significant at the 99.9% confidence level? What do you conclude about the effectiveness of the Salk vaccine? Explain briefly. *[70 points]*

(d) Your confidence interval sketch in (c) should have revealed that there was quite a bit of distance between the left endpoint and 0, which means that — in retrospect, after the experiment had finished — the designers of the trial had chosen T and C sample sizes that were quite a bit bigger than necessary. In the rest of this problem, let's roll the clock back to the period in which the trial was designed, and reconsider the sample size issue.

Let $n = (n_C + n_T)$ be the total sample size planned for the experiment, and for simplicity suppose that exactly $\frac{n}{2}$ children are randomized to each of the T and C groups. If the polio incidences turned out to precisely match the rates in the actual trial, what value of n would have been necessary to make the left edge of the 99.9% confidence interval be just barely positive? Show your work. (This method is one way to perform *sample size determination* at design time.) Do you think the designers of the Salk trial were stupid, or is there some other explanation for their retrospectively-unnecessarily-large sample sizes? Explain briefly. *[30 points]*