Mathematical Statistics (Math 30) – Course Project for Spring 2011

Goals:

- To explore methods from class in a real life (historical) estimation setting
- To learn some basic statistical computational skills
- To explore a topic of interest via a short report on a selected paper

The course project has been divided into 2 parts. Each part will account for 5% of your grade, because the project accounts for 10%.

Part 1: German Tank Problem ***Due Date: April 8th***

Our Setting:

The Purple (Amherst) and Gold (Williams) armies are at war. You are a highly ranked intelligence officer in the Purple army and you are given the job of estimating the number of tanks in the Gold army's tank fleet. Luckily, the Gold army has decided to put sequential serial numbers on their tanks, so you can consider their tanks labeled as 1,2,3,…,N, and all you have to do is estimate N. Suppose you have access to a random sample of k serial numbers obtained by spies, found on abandoned/destroyed equipment, etc. Using your k observations, you need to estimate N. Note that we won't deal with issues of sampling without replacement, so you can consider each observation as an independent observation from a ***discrete*** uniform distribution on 1 to N. The derivations below should lead you to consider several possible estimators of N, from which you will be picking one to propose as your estimate of N.

Historical Setting:

This was a real problem encountered by statisticians/intelligence in WWII. The Allies wanted to know how many tanks the Germans were producing per month (the procedure was used for things other than tanks too). The Germans had serial numbers on their tanks, and the order of tank production ended up being highly correlated with the serial numbers, so you could treat them in this framework and get a reasonable estimate of N. How reasonable? Well, it is easy to compare the estimates of per month production (Allied statistical and intelligence estimates) to the actual German war records obtained after the war.

Month	Statistical Estimate	Intelligence Estimate	German Records
June 1940	169	1000	122
June 1941	244	1550	271
August 1942	327	1550	342

Accurately estimating the number of tanks led to accurate understanding of how many factories must be in use, the workforce involved, etc., and was a great help to the Allied forces. Moral of the story: If you don't want to have your tank production estimated by a statistician, use random serial numbers.

Derivations for the project:

- Find the pdf and cdf for X, a single observation from a discrete uniform distribution on 1 to N. (Note, you may need the floor function to make things easy to write.)
- Find the expected value and variance of X.
- Considering a random sample of k tanks, determine the method of moments estimator of N.
- Find the expected value and variance of the MOM estimator. Is it unbiased?

- Find the maximum likelihood estimator of N.
- Identify a sufficient statistic for N.
- Let's consider the maximum of our sample as an estimate for N. We want to determine if the maximum is unbiased. To help you determine this:
  - Find the cdf of the maximum, $X_{(k)}$. (Use our results on order statistics.)
  - Show that if Y is a random variable with non-negative integer values, then

    $$E(Y) = \sum_{k=1}^{\infty} P(Y \geq k)$$
  - Determine the series expression for $E(X_{(k)})$.
  - Show using Riemann sums that $\frac{E(X_{(k)})}{N} \approx \int_0^1 (1-x^k)dx.$ $\frac{E(X_{(k)})}{N}$ is a left Riemann sum.
  - Use the results above to find $E(X_{(k)})$, and determine if the maximum is unbiased.
- Not: it can be shown in a similar fashion to what was done above (just take this as fact, don't need to show) that $V(X_{(k)}) = \frac{(N+1)(N-k)k}{(k+1)^2(k+2)}$. Using all the information you have for the maximum, find a function of the maximum that is unbiased for N, and then determine the variance of that estimator.
- Identify the UMVU estimator for N (explain how you know it is UMVU).
- Determine if the UMVU estimator is consistent for N.
- Determine the relative efficiency of the MOM estimator to the UMVU estimator.
- Brainstorm at least 2 other semi-reasonable estimators for N (you don't have to use methods from class to derive them, just think about what else you might use). By semi-reasonable, I mean that you wouldn't use the minimum, mean, or median (something that might be smaller than the maximum value, when we know N>maximum most likely), but you could scale those. For example below, I try 3 times the mean as an example estimator (no, it's not very good).
- Finally, using your preferred estimate for N, give your estimated N if current intelligence has a sample of 15 Gold tanks labeled: 1548, 1707, 1998, 264, 1411, 157, 329, 835, 1734, 159, 1689, 1083, 660, 1921, and 1021.

Letter Write-up for Part I:

You should write/type up your derivations and frame them in a letter to your commanding officer (me) discussing our options for estimating the number of Gold tanks, and your best estimate based on the current intelligence. Your letter will also include some computational results (see below). It should be clear by the end of the letter what your final recommended estimator is. You can use any results from class without re-deriving them, but you should state results you are using (for example, you could state that an MLE that is sufficient is minimal sufficient).

Computing for the project:

As part of the evidence in favor of your chosen estimator, we will simulate the distributions of some of your chosen statistics in R as follows:

- Set N and k for your particular simulation.

- Take many samples of size k (after each sample is drawn, return those k tanks back to the population so they can be in the next sample).
- For each sample, compute the value of your estimator, $\hat{\theta}$.
- Save all computed values of $\hat{\theta}$ and compare to N to see how the estimator performs, as well as examine the distribution of $\hat{\theta}$ using basic graphs and descriptive statistics.

The code below demonstrates this procedure. It takes tanks labeled 1 to 350, and takes a random sample of 10 tanks, and does this 10000 times (runs). For each run, it computes the chosen statistic = 3*mean (note this is reasonable because 3*mean is usually larger than the maximum) and stores it in *statistic*[run number]. Then, using the saved values in *statistic*, you can make graphs of the statistics distribution, and get summary measures. For your statistics, you might need functions like mean, median, max, min, etc. In R, help(something) is the command to get help on whatever “something” is.

You can of course use other software, if you are comfortable with it. The basic idea is to get familiar with using a computer to do simulations. ***Computing help***: PLEASE ask for help if you can’t get something to run. I don’t want you to struggle with R when we are just using it as a tool for simulation.

Example R code looking at $\hat{\theta}$=3*mean of the sampled tanks:

```
N=350
k=10
runs=10000
statistic=NULL
samplevalues=NULL
for(i in 1:runs)
{samplevalues=sample(1:N,k);
statistic[i]=3*mean(samplevalues)}
hist(statistic)
boxplot(statistic)
summary(statistic)
sd(statistic)
IQR(statistic)
```

You can also save multiple statistics at once, just set up statistic1, statistic2, etc. and give each a different formula in the for loop.

In your letters for Part I:
Using code similar to this, examine the distributions of the MOM estimator for N, the UMVU estimator for N, and two other estimators of your choice, under different choices for N and k. You should include a discussion of these distributions and some supporting work (some graphs or descriptive stats) for some settings (not all please) in your letter for Part 1 of the project. Do your simulations provide evidence that the UMVU estimator is preferable to the others you considered? You should also see something fairly nice about the distribution of the MOM estimator for larger values of k.

Note that the code does NOT take long to run, even for 10000 runs (and you can run more or fewer runs), so you can investigate MANY settings. I expect you to investigate settings that you are NOT reporting in your letter, so please don’t try to report everything.

Part 2: Current Topics in the American Statistician (TAS) ***Due Date: April 25th***

In this half of the project, you will be briefly exploring a topic selected from a pre-screened recent TAS article. TAS is a general audience statistics journal, similar to the American Mathematical Monthly. I have tried to choose a variety of topics, and have chosen ones around the same difficulty level to tackle. Main themes in these articles are hypothesis testing, estimation, confidence intervals, and Bayesian inference.

You should read the TAS article you select, and then write a 3-5 page (double spaced) summary of what you learned, what methods were used, etc. You may have to look up some terms/distributions, etc. in order to write your summary, in which case you should cite additional sources. The target audience you are writing for should be someone else in the class. Therefore, you can assume knowledge of probability, and what we've covered in class, but they wouldn't necessarily know what a Cramer-Rao lower bound for variance or Fisher information or Jeffrey's prior is. You can compare/contrast methods in the paper with methods from class without re-explaining the methods from class. You should also try to verify calculations in the paper for yourself, and may include those in your paper as well (space permitting). For some papers, it may be beneficial to make an example to demonstrate the methods in use.

Example. Suppose you choose topic #13, which is a comparison of 6 CI methods. Besides introducing the setting, giving a summary of the methods and their advantages and disadvantages, you could make your own example and compute the different CIs for your example. (Feel free to ask for simulation help if you want to do one).

Paper list: (see me to look through the articles if you want to do that with hard copies, or the entire text of the journal is available online via the library) – All articles are from the American Statistician (TAS). Volume and number are given as (vol, num) after the article title.

1. A Quick, Compact, Two-Sample Dispersion Test: Count Five (59, 1) – hypothesis test for variances
2. A Sufficiency Paradox: An Insufficient Statistic Preserving Fisher Information (59, 1)
3. Testing Fisher, Neyman, Pearson, and Bayes (59, 2) – hypothesis testing discussion
4. Evaluation Criteria for Discrete Confidence Intervals: Beyond Coverage and Length (59, 2)
5. The Decline and Fall of Type II Error Rates (59, 4)
6. A Spatial Analysis of Basketball Shot Chart Data (60, 3) – Bayesian analysis
7. The Boxer, the Wrestler, and the Coin Flip: A Paradox of Robust Bayesian Inference and Belief Functions (60, 2)
8. Calibrated Bayes: A Bayes/Frequentist Roadmap (60, 3)
9. On the So-Called "Huber sandwich Estimator" and "Robust Standard Errors" (60, 4) – MLE topic
10. Characterizing Density Crossing Points (61, 1) – discussion of when pdfs (normal, t, etc.) cross
11. Likelihood Ratio, Score, and Wald Tests in a Constrained Parameter Space (61, 1)
12. Bayesian Inference on a Proportion Believed to be a Simple Fraction (61, 3)
13. Confidence Interval Estimation of a Normal Percentile (61, 4) – comparison of 6 CI methods
14. The Bagged Median and the Bragged Mean (61, 4) – estimation topics
15. The Role of Likelihood in Interval Estimation (62, 1)
16. Confidence Intervals for a Discrete Population Median (62, 1)
17. A Comparison of Bayes – Laplace, Jeffreys, and Other Priors: The Case of Zero Events (62, 1) – comparison of priors in the Binomial-Beta conjugate family
18. A Surprising MLE for Interval-Censored Binomial Data (62, 2)

19. A General Proof of Some Known Results of Independence Between Two Statistics (62, 2) – relates to Basu’s theorem and requires mgfs
20. Sufficiency in Finite Parameter and Sample Spaces (62, 3)
21. Parametric Nonparametric Statistics: An Introduction to Mixtures of Finite Polya Trees (62, 4)
22. Uniformly Hyper-Efficient Bayes Inference in a Class of Nonregular Problems (63, 3)
23. Three Examples of Accurate Likelihood Inference (64, 2)
24. A Geometric Comparison of Delta and Fieller Confidence Intervals (64, 3)
25. Fixed Width Sequential Confidence Intervals for p (64, 3)
26. Closed Form Prediction Intervals Applied for Disease Counts (64, 3)