Stat 505 - Environmental Statistics, Fall 2013 - HW 1

Due: in class, Thursday, 3 October.

Remember, you are to do 3 of the following 4 problems. You can choose. My intent was to write one easier theory problem, one easier data analysis problem, one harder theory problem, and one harder data analysis problem. Your perception may differ; even so, choose 3 problems. If you do all 4, I'll grade them all and drop the lowest. Further information and suggestions for organizing answers to open ended data analysis problems (like problem 3) are on the 'homework information' part of the class web site.

Note: sometimes the data questions will be longer than the theory questions. Sometimes the other way around.

Note 2: I am very willing to help with R. If you tell me, "I want to do XXXXX", I'll tell you what functions do that, or that you need to write your own function and help you figure out how to write that function. If something isn't working, I'm very willing to help you debug. I believe you learn by trying, which is the purpose of the HW.

1. The data in Atra.csv are measurements of Atrazine in a series of Nebraska wells. Atrazine is the most widely used herbicide by US farmers and is often found in groundwater. If you want to see a map of atrazine application, go to:
   `http://water.usgs.gov/nawqa/pnsp/usage/maps/compound_listing.php` and click on Atrazine.

   The values in the data set are in parts per billion (ppb) = μg/L. The detection limit is 0.01 ppb. The file has data from the 24 wells measured in June and again in September. The variable June has the value (observed or detection limit); the variable JuneCen is true if the value is a <dl value.

   For the first set of questions, consider only the data collected in June.

   (a) Is it reasonable to assume a normal distribution? Explain why or why not. If not, what distribution might be reasonable? Include with your answer a plot supporting your choice of distribution.

   (b) Use maximum likelihood and your chosen distribution to estimate the arithmetic mean and coefficient of variation.

   (c) If you assume that the data are logNormally distributed, calculate 95% confidence interval for the arithmetic mean using 1) a normal approximation and 2) profile likelihood.

   (d) Which confidence interval computed in question 1c you do believe is more appropriate to report? Support your choice, e.g., by including a plot. (As far as I know, this requires writing some non-EnvStats functions).

   (e) Define μ as the arithmetic mean. You want to test the one-sided null hypothesis H_o: $\mu \geq 0.10$, i.e. you want to demonstrate the alternative hypothesis: $\mu < 0.10$. This hypothesis is environmentally interesting because a level of 0.10 ppm has been claimed to result in adverse effects in some animals. Based on what you found in question 1d, is it more appropriate to use a Wald (normal distribution) test or a likelihood ratio test?

(f) Using the mle's and associated information, test H_o: $\mu \geq 0.10$. Report your test statistic (z or χ^2) and p-value.

(g) Use robust order statistics to estimate the arithmetic mean and coefficient of variation.

(h) You are told that your final report can only include results from one estimation method (ROS or MLE). Which results do you report? Justify your choice.

These data were collected to see whether the concentration of Atrazine in wells increased between June and September. Atrazine is heavily applied to Nebraska corn fields during the summer.

(i) Use a non-parametric test to test the null hypothesis of no change in median Atrazine concentration between June and September.
Hint: consider the non-EnvStats function wilcox.test()

2. Two unrelated theoretical issues

(a) I used conditional expectations to explore the bias of substitution estimators for data with a single detection limit. The approach can also be used when there are multiple detection limits. Generalize our earlier approach to multiple detection limits to show:
1) substitution by 0 always underestimates the mean, and
2) substitution by dl, which may be different for every observation, always overestimates the mean.
If it helps to have a specific example, consider $\ln Y \sim N(2, 0.5)$ with some observations reported as < 1, some as < 5, some as < 10, and a few as < 15.

(b) I argued in class that the plug-in estimator, $e^{\overline{X}+s^2/2}$, was an approximate estimator of the mean of log normal values. In fact, it is biased and overestimates E Y, where $X = \log Y \sim N(\mu, \sigma^2)$. One suggestion, by the amazing British statistician Sir David Cox, is the approximate estimator

$$\hat{\mu}_Y = e^{\overline{X}+s_X^2/2-s_X^2/(2n)},$$

where n is the number of observations in a sample.

Show why this may be a reasonable estimator. It may help your discussion to show that if σ_X^2 is known, then $e^{\overline{X}+\sigma_X^2/2-\sigma_X^2/2n}$ is an unbiased estimator of E Y.

3. The data in AgMercury.txt and UrbanMercury.txt are measurements of Mercury (Hg), in parts per million (ppm), from fish collected from random samples of streams in agricultural watersheds and urban watersheds. There is one observation per stream. Currently, the 'do-not-sell' limit for Hg in fish is 0.5 ppm in Canada and 1.0 ppm in the US. Your department head needs to know:
the mean concentration in each type of stream,
whether those means are the same, and
the probability that fish Hg concentration in a randomly chosen agricultural stream exceeds 0.5 ppm.

Examine the data, decide how best to analyze it, do that analysis (or analyses) and report your conclusion(s) and justify your choice of analysis. Your report will be read by managers who:
1) need an executive summary (see the homework guidelines)
2) really hate answers that say something like 'well if I assume this, I get this, but if I assume that, I get something different'. You may (and should) consider multiple analyses, but in the end, you have to report a single prefered answer.

4. This problem uses simulation to explore efficiency and precision of estimators of the arithmetic mean of a log-normal distribution when the observations include < dl values. I would do this in R, but other languages are certainly possible. I will provide code on the class web site that illustrates one possible way to organize a simulation. Again, there are lots of ways to do this.

   We'll consider samples of 25 observations drawn from a logN(3,1) population. That is $\ln Y \sim N(3,1)$. The observations are concentrations of a chemical, measured in parts-per-million (ppm). We are interested in estimating the mean of this population. I suggest you consider simulating 10,000 samples for each estimator.

   The first few parts assume there is no detection limit. All 25 observations are completely observed.

   Note: many questions ask whether something calculated from a simulation is "similar" to or "approximately" equal to a theoretical quantity. The simulation value will never be identical because of Monte-Carlo uncertainty in the simulated value. If you know how to calculate the Monte-Carlo standard error in the simulated quantity, please use that to define "similar"; otherwise, use your judgement.

   (a) For warmup, a simple estimator is the sample average. Use simulation to estimate its variance when the sample size = 25. Calculate the theoretical value for the variance of the sample average. Are the two similar?

   (b) Calculate the usual unbiased estimates of μ and σ^2 from log transformed observations, then use the plug-in estimator of the lognormal mean, $\exp(\hat{\mu}+\hat{\sigma}^2/2)$ to estimate the sample mean. Is the estimator approximately unbiased? What is the its variance? What is its efficiency compared to the sample average?
   Statistical Efficiency for one estimator relative to the other is the inverse ratio of their variances.

   (c) Calculate the usual unbiased estimates of μ and σ^2 from log transformed observations, then use Finney's UMVUE to estimate the sample mean. Is the estimator unbiased? What is the its variance? What is its efficiency compared to the sample average?

   (d) Which of the above three estimators has the smallest Mean Squared Error?
   Note: MSE for an estimator is bias^2 + variance. It measures the accuracy of an estimator, in the sense of how far the estimates are from the true value.

   Now, let's assume that the measurement process has a single detection limit of 15ppm.

   (e) Estimate the sample mean for each of your 10,000 samples using ROS, ML/plug-in, and ML/Finney. For each, calculate the bias and variance. For each estimator, what is the loss of efficiency introduced by reporting some values as < dl?
   Note: The efficiency for ROS would be based on a comparison to the variance of the sample average from all observations.

   (f) Now let's apply your results to answer a practical question. Your analytical chemist friend gives you a choice. He has available two different methods to measure the concentration of this chemical. One is a slow and expensive method. This method is capable of measuring extremely small concentrations and has essentially no detection limit. The second method is quick and much less expensive, but has a detection limit of 15 ppm. The instrument used in the second method will only report <15; it will not give you the actual value if that is below the dl.

You have enough money to measure 5 samples using the slow method or 25 samples using the quick method. Which will give you the most precise estimate of the population mean concentration?