

Due: in class, Tuesday Feb 26.

Remember, you are to do 3 of the following 4 problems. You can choose. My intent was to write 1 easier theory problem, 1 easier data analysis problem, 1 harder theory problem, and 1 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.

1. The data in AgMercury.txt are mercury concentrations in fish collected from agricultural watersheds across the US. Your charge is to estimate the mean and coefficient of variation from these data, then compare the agricultural data to similar data from urban watersheds.
 - (a) Is it reasonable to assume a normal distribution? Explain why or why not. If not, what distribution might be reasonable?
 - (b) Use maximum likelihood to estimate the mean and coefficient of variation.
 - (c) Use robust order statistics to estimate the mean and coefficient of variation.
 - (d) The data in UrbanMercury.txt are mercury concentrations in fish collected from urban watersheds. Construct a likelihood ratio test of whether the two types of watersheds have the same mean (or median, you choose, but be clear what you are testing).
 - (e) Use a robust or non-parametric test to test whether the two types of watersheds have the same mean (or median; again, justify your choice).
 - (f) You are told that your final report can only include results from one estimation method and one test method. Which results do you report? Justify your choice.

2. A trio of 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,
 - 2) substitution by dl, which may be different for every observation, always overestimates the mean, and
 - 3) substitution by dl/2 may be quite reasonable.If it helps to have a specific example, consider $\log Y \sim N(2, 0.5)$ 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, $e^{\bar{Y}+s^2/2}$, was an approximate estimator of the mean of log normal values. In fact, it is biased and overestimates $E X$, where $Y = \log X \sim N(\mu, \sigma^2)$. One suggestion, by no less than Sir David Cox, is the approximate estimator

$$\hat{\mu}X = e^{\bar{Y}+s_Y^2/2-s_Y^2/(2n)},$$

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_Y^2 is known, then $e^{\bar{Y}+\sigma_Y^2/2-\sigma_Y^2/2n}$ is an unbiased estimator of $E X$.

(c) Log transformations are rampant in environmental statistics, because skewed data are so common. It is easy to estimate means, variances and standard errors on the log scale (i.e. of $Y = \log X$). The problem is making conclusions about X , the untransformed data. Here are a few claims that you may see made in the literature. For each, show whether it is true or false. We will assume a log normal distribution and use the notation $Y = \log X \sim N(\mu, \sigma^2)$

1) $e^{\bar{Y}}$ is an unbiased estimator of the mean $E X$.

2) $e^{\bar{Y}}$ is an unbiased estimator of the median of X .

2a) for extra credit, is $e^{\bar{Y}}$ a consistent estimator of the mean of X , the median of X , or both?

3) Observations from two groups follow log normal distributions with different means but the same variance. I.e., $Y_1 = \log X_1 \sim N(\mu_1, \sigma^2)$ and $Y_2 = \log X_2 \sim N(\mu_2, \sigma^2)$. The claim is that $e^{\bar{Y}_2 - \bar{Y}_1}$ is an unbiased estimator of $E X_2 / E X_1$. This is sometimes described as the multiplicative effect between groups 1 and 2.

3a) For a bit of extra credit, is $e^{\bar{Y}_2 - \bar{Y}_1}$ a consistent estimator of $E X_2 / E X_1$?

3. The data in Cadmium.txt are measurements of Cadmium in fish collected from the Colorado Plateau and Southern Rocky Mountains, two regions of the state of Colorado. Your charge is to estimate the mean concentration in each region then test whether those means are the same. 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 don't like 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 indicate a single preferred answer.

4. An appealing feature of ml estimation is the straightforward estimation of the s.e. of an estimate, using the observed information matrix. When observations are independent, the observed information when estimating one parameter, θ , is

$$I_O = -H = - \sum_{i=1}^n \frac{\delta^2 \ln L(x_i)}{\delta \theta^2} \Big|_{\hat{\theta}}$$

Assume that $X \sim N(\mu, 1)$, i.e. that the variance is known and we only have to estimate the mean.

(a) Calculate the contribution to the observed information for an observed value of x . Does the information contribution vary with x ?

(b) Calculate the contribution to the observed information for a value reported as $< x$. Does the information contribution vary with x , the detection limit?

Remember that the variance of the mean is estimated as $1/I_O$ when the model has only one parameter (as this one does). You are attempting to measure concentrations of a contaminant that are sufficiently low that the analytical chemists struggle to measure those concentrations well. Your guess is that the mean concentration in the population is about 5.

(c) Your analytical chemist gives you a choice: he can use a "quick" procedure with a d.l. of 3 and measure five samples. Or, he can use a laborious procedure, essentially with no d.l. and measure two samples. Which is better, in the sense of providing the most (statistical) information and hence the smallest sampling variance.