Stat 505 - Environmental Statistics, Fall 2013 - HW 4

Due: 5 pm, Thursday, December 19.

Because the semester is almost over, the problems are shorter than those on previous HW. There are one data question, one 'practical' problem, and two theory questions. Please do three of the four.

1. The data in remediate.txt were collected as part of a study of revegetation of part of the Athabasca oil sands. This is an area in Alberta, Canada, along the Athabasca river, where the sand contains large amounts of heavy oil that is being extracted and converted to petroleum. After mining, the operators are required to revegetate the mined areas. The Alberta Provincial Government has established criteria for revegetation success that include many different biological characteristics. One of the criteria for tree growth is:

   > Measurement of the last 3 years annual height increments to determine if growth is comparable to that observed on sites of corresponding soil capability classes without disturbance.

   In general, these criteria are biologically specific and statistically vague. The above quote is typical. I think an equivalence test is one way to define comparable.

   Here are data on mean annual height increment of aspen trees on 15 pairs of plots. These pairs are on a range of soil classes, so the growth is expected to differ between the pairs. One of each pair is revegetated; the other is undisturbed.

   Your manager wants to know whether these data show that the annual height growth is 'comparable to that observed ... without disturbance'. Please use FDA criteria (i.e. equivalence region for revegetated mean is 80% to 125% of control mean) to define comparable.

   Please treat this as a data problem. That means I want you to write an executive summary (see HW quidelines if you don't remember the requirements for a data problem).

2. The data in airlead4.txt are 114 observations of the air concentration of lead in downtown Baltimore MD. The values are log transformed concentrations. I believe the original units are parts-per-billion (micrograms/liter). It is reasonable to treat these values as a simple random sample of human exposure to airborne lead in downtown Baltimore. For all parts of this problem, please assume a log normal distribution for concentration.

   (a) Calculate a one-sided 99.9% upper bound for the mean log-transformed concentration.

   (b) Calculate an upper bound that has probability of 99.9% of including a single newly measured concentration.

   (c) Calculate an upper bound that includes 99.9% of all airborne lead concentrations with 95% probability.

3. Equivalence testing has been proposed as a way to prove that a data fits predictions from a complex environmental model. The sort of model being considered here is a large computer code based on a-priori parameters. It is not a statistical model fit to data. The idea is to collect a set of pairs of observations (X=model prediction, Y=observed value), then regress Y on X, i.e. fit $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. If the model fits, the regression intercept, β_0, is 0, and regression slope, β_1 is 1. If the regression intercept is significantly equivalent to 0 and the regression slope is significantly equivalent to 1, this proves that the model fits the data.

   (a) You decide that the equivalence region for the intercept is $(-0.1,\ 0.1)$ and the equivalence region for the slope is $(0.9,\ 1.1)$. Derive a level α (or level $\leq \alpha$) test that the model fits the data.

   (b) An alternative approach is to calculate the difference between the model prediction and the observed value, then test whether the mean difference is equivalent to zero. You might decide that an appropriate equivalence region for the mean difference is $(-0.2,\ 0.2)$. Is this test of difference equivalent to 0 the same as the test in part a)? Explain why or why not.

   An example of two tests that are the same is the t-test of difference = 0 and the 1 df F test of difference = 0.

4. We discussed BACI designs as a way to control for both pre-existing differences between control and impact sites and for unanticipated differences between background and impacted periods. A BACI analysis does have some statistical consequences, which are explored in this problem.

   Consider the simplest BACI study:
   two sites: one control and one impacted,
   two periods: one background and one impacted.
   You have N independent replicate observations from each of the four combinations of sites and periods. The variance between replicate observations is σ^2, which is assumed to be the same variance for all four combinations of sites and periods. A cell-means model is especially easy to use to determine properties of interesting quantities:

$$\begin{aligned} Y_{ijk} &= \mu_{ij} + \varepsilon_{ijk} \\ \varepsilon_{ijk} &\sim \text{iid } N(0, \sigma^2), \end{aligned} \tag{1}$$

   where i indexes sites (C or I), j indexes periods (B or I), and k identifies observations within each combination of site and period.

   (a) What is the variance of $\hat{\mu}_{II} - \hat{\mu}_{CI}$, the mean difference between control and impacted sites during the impact period?

   (b) What is the variance of the BACI interaction: $(\hat{\mu}_{II} - \hat{\mu}_{IB}) - (\hat{\mu}_{CI} - \hat{\mu}_{CB})$, the difference of mean differences?

   (c) Is there a statistical cost to analyzing the data using a BACI design, instead of just looking at the difference between control and impact sites? Briefly explain.

   Now consider a multi-site random effects BACI, where there are multiple control and a single impact site. M is the number of control sites. Again, two periods and N replicate measurements per combination of site and period. A model for these data needs to account for the additional

variability between control sites. One such model is:

$$\begin{aligned} Y_{ijkl} &= \mu_{ij} + \gamma_{ik} + \varepsilon_{ijkl} && (2) \\ \gamma_{ik} &\sim \text{iid } N(0, \sigma^2_{site}) && (3) \\ \varepsilon_{ijkl} &\sim \text{iid } N(0, \sigma^2), \end{aligned}$$

where i indexes type of site (C or I), j indexes periods (B or I), k indexes sites identifies within site-type ($k = 1 \ldots M$ for control sites and $k = 1$ only for the impact site). l identifies observations within each combination of site and period. σ^2_{site} is the variance between control sites and σ^2 is the variance between replicate measurements in each combination of site and period.

Using this multi-site model:

(d) What is the variance of $\hat{\mu}_{II} - \hat{\mu}_{CI}$, the mean difference between control and impacted sites during the impact period?

(e) What is the variance of the BACI interaction: $(\hat{\mu}_{II} - \hat{\mu}_{IB}) - (\hat{\mu}_{CI} - \hat{\mu}_{CB})$, the difference of mean differences?

(f) Is there a statistical cost to analyzing the data using a BACI design with multiple sites instead of just looking at the difference between control and impact sites? Briefly explain.

The second model presumes that the mean difference between control and impact periods is the same for each site. This presumption can be relaxed by including a random site*period interaction effect in the model:

$$\begin{aligned} Y_{ijkl} &= \mu_{ij} + \gamma_{ik} + \tau_{ijk} + \varepsilon_{ijkl} && (4) \\ \gamma_{ik} &\sim \text{iid } N(0, \sigma^2_{site}) && (5) \\ \tau_{ijk} &\sim \text{iid } N(0, \sigma^2_{site*period}) \\ \varepsilon_{ijkl} &\sim \text{iid } N(0, \sigma^2), \end{aligned}$$

where all terms occuring in model (2) have the same definitions in model (3). τ_{ijk} is the site-period interaction.

(g) What is the variance of $\hat{\mu}_{II} - \hat{\mu}_{CI}$, the mean difference between control and impacted sites during the impact period?

(h) What is the variance of the BACI interaction: $(\hat{\mu}_{II} - \hat{\mu}_{IB}) - (\hat{\mu}_{CI} - \hat{\mu}_{CB})$, the difference of mean differences?

(i) When you allow a site*period interaction, is there a clear statistical justification choosing for one analysis or the other? Briefly explain.