Greg Schaefer, et. al.
Stat8801: Presentation
27 January 2012

Forestry Survey Problem

Statistical problem?

We do have a statistical problem here if we assume that those doing the surveying will be happy if the assignments are fair, and that the assignments are fair if done *randomly*.

There are other ways of achieving happy crews and other approaches to fairness: e.g., one could give more pay to those doing the most strenuous work; or justify assignments based on the abilities of the crew members and emphasizing that the point of the survey is to get the best data possible (this reasoning is meaningful if all those doing the surveying are also interested in getting the best data possible).

But the experience of the client, we are told, is that happy crews can be achieved by mixing up the crews each week and by making sure that everyone gets to be a recorder at least once. Let's proceed assuming that the client is right about this.

Issues to Address

We want happy crews and good data. These desiderata might conflict.

For instance, some crew members might do a better job recording the data than other crew members. Some individuals might be better with the actual measuring. From the perspective of reducing measurement error, might crews actually perform better if they *stay together* throughout the entire 9 weeks (for example, if all six data collectors need to work as a team to get the measurement that they are interested in)?

Mixing the crews

Since it is too difficult to choose crews for each week from the 66,512,160 ways of partitioning 21 people into 3 unordered groups of 7 each, we need another algorithm for selecting the crews. I suggest the following if we have computing available to us:

(a) For each week, reset the seed for the random number generator. Assign (on paper) each of the crew members a number from 1 to 21. Then sample 7 from 1:21, then 7 from the remaining 14 numbers, then 7 from the remaining 7. Do this 9 times to construct the crews for each of the nine weeks.

(b) Sort the array of values by crew, so that we have 27 ordered lists of numbers, each of length 7. Let these be the rows of matrix M. Run 'anyDuplicated' on M to ensure that no crew is duplicated across the nine weeks. (I don't actually think this duplication can happen, given that we are changing the seed for the random number generator for each week's set of crews, but it is easy enough to run this test.)

(c) Then construct the set of data recorders for each of the nine weeks. If P = c(1:21), first sample 3 from P, then 3 from P[-first sample], then 3 from P[-(first & second samples)], and so forth. This gives us 21 distinct data recorders. Then sample 6 from P again to get the lucky few who get to be data recorders twice.

My suggestion would be to use the above algorithm to make up the schedule of crews and data recorders *before* any surveying begins. If the client expects that not all surveyors will survive the rigors of nine weeks in the field, rules should be in place about how to proceed with random assignment for the remaining weeks when individuals drop out. If the rules are agreed upon up front, before the surveying begins, there will be less cause for unhappiness.

Here is a suggestion for random assignment in the field (under the assumption that computing is not available to us):

Names of the 21 surveyors can be put on individual strips of paper; the strips of paper are then put into a hat. At the beginning of each week, before everyone goes out in the field, we can do the following:

(A) Draw three names without replacement. Each of these individuals will be the data recorder for one of the crews that week. If one of the persons drawn has already been a data recorder, and it is not yet weeks 8 or 9 in the field, draw again until we have three individuals who have yet to be data recorders. Once we have reached weeks 8 or 9 all crew members are again eligible to be data recorder one more time.

After the three data recorders have been selected, return to the hat those names of individuals who were drawn but who were not eligible to be data recorders.

(B) We now have 18 names to choose from. Draw six without replacement and assign to the first crew. Draw another six without replacement and assign to the second crew. The remaining six are assigned to the third crew.

The better we mix the names in the hat before drawing, the more likely we will have well-mixed crews across the nine weeks.

A schedule for all 9 weeks can be done up before any surveying starts, but then we have to worry about someone dropping out due to illness, injury, or being mauled by a bear.

Additional Questions for the client

1. Exactly what kind of data are you collecting? what are you measuring, and how?

2. Is there any concern that remixing the crews and data recorders will adversely affect the quality of the data collected?

3. Is it expected that all surveyors will remain on the job for the nine week period?

4. Should we expect measurement error to increase, decrease, or remain about the same over the nine week period?

Suggestions

1. If each person does their own measuring (as opposed to six individuals working as a unit to get a single measurement), assign an **id** to each data collector and make sure that this id is associated with each observation recorded by that crew member.

2. Similarly, the data recorder for each observation should be associated with that observation in the dataset.

The information from 1 and 2 above will allow us to see whether there are systematic differences between any of the data collectors and between any of the data recorders. If systematic differences do appear, we will know their magnitude and direction and hopefully be able to adjust accordingly when we make inferences from the data.

3. Will the crews be going over the same territory in a given week, or will they be working in entirely distinct areas? If different crews end up working the same territory for a given week, potential systematic differences in the measurements due to crew composition poses much less of a concern. (For we can do some sort of averaging across crews for the region in question.) If all crews are working the same territory across all nine weeks, this concern about systematic differences due to crew composition disappears entirely due to the remixing of the crews at the end of each week.

The concern raised here is more of an issue if crews must work as a unit to get each measurement and if the crews are working in distinct regions. Otherwise, if individuals are taking separate measurements, systematic differences in a region are reduced due to the differences in the six individuals working in that region.

4. If the crews are covering different regions, we may want to associate with each observation the region from which it came. For there may be variables such as tree species, soil quality and terrain, or proximity to water that can account for some of the variability in the measurements taken. Perhaps the variables just mentioned can be known to the extent that we need to know them simply by knowing from what region the measurements came. In any case, there may be systematic differences in the data collected due to the region from which the observations came.

5. If we expect measurement error to decrease over the nine week period as each surveyor gets better at what they are doing, or if we expect measurement error to increase over the nine week period due to fatigue setting in, it will be important to note what week the observation was recorded in.