Stat 549
Fall 2014 Homework 3: due Fri, Oct. 31

Note: please use R for all calculations. Include your R scripts with output.

1. In the Exxon-Valdez oil spill trial, one important issue was the amount of lost harvest of fish, shellfish, seals, etc. suffered by native subsistence harvesters in the region. In particular, it was desired to estimate the per capita harvest loss and the total loss. Some data had been collected over the years on such harvests, both prespill and postspill. You will analyze data from just one community for one year (1989, the spill year) and estimate total harvest and per capita harvest for this community in 1989. A random sample of 42 households out of 67 was selected from the community. The number of individuals living in each household and the total household harvest (in pounds, estimated through detailed questioning about harvests of many individual species) were recorded.

   The data are available on the website in the file `harvest.csv` which can be read into R using the `read.csv` command. The resulting data frame has two variables: `size` (number of people in household) and `harvest` (in pounds).

   (a) Estimate the per capita harvest for this community. Attach a standard error to your estimate and calculate a 95% confidence interval.

   (b) Estimate the total harvest for the community. Attach a standard error to your estimate and calculate a 95% confidence interval.

   (c) Now suppose the total number of individuals in the community is known to be 190 (do not use this information in part b). Estimate total harvest using both ratio and regression estimation and calculate SE's. Compare these results with your result in (b); is there much of an improvement? Which seems more appropriate based on a plot of the data: ratio or regression estimation?

2. Kruuk et al. (1989) used a stratified sample to estimate the number of otter dens (called holts) along the coastline of Shetland, UK. The coastline (except for parts that were predominantly buildings) was divided into 237 5-km sections and each section was assigned to one of four terrain types. A random sample of sections within each stratum were chosen for counting. In each section chosen, researchers counted the number of otter dens in a 110-m-wide strip along the coast. The data are in the file `otters.csv` which can be read into R using `read.csv`. The population and sample sizes for the strata are given in the table below. Estimate the total number of otters along the coast in Shetland, along with a standard error and a 95% confidence interval. Note: use the `tapply` command in R to compute the variance (or any other function) of the observations by stratum. If `df` is a data frame with a quantitative variable `y` and a categorical variable `x`, then `tapply(df$y,df$x,var)` will give the variance of `y` for each of the categories of `x`.

Stratum	Total Sections	Sections Counted
1 Cliffs over 10 m	89	19
2 Agriculture	61	20
3 Not 1 or 2, peat	40	22
4 Not 1 or 2, nonpeat	47	21

3. Using aerial photographs, a forester stratifies a 400-acre forest into three strata, and plans to sample each stratum using 0.1-acre plots to estimate the total cubic foot volume of timber. From past surveys, he has information on the range of volumes (largest minus smallest volumes likely to occur per plot) and cost in dollars for surveying a plot in each stratum. These data can be used in planning the survey and are summarized below.

Stratum	Acreage	Range	Cost (c_i)
1. Small Pines (trees 50-70 feet tall)	180	80	20
2. Large Pines (trees 70-90 feet tall)	70	120	25
3. Mixed Pine - hardwood swamp	150	200	35

Suppose he wants to estimate the total cubic foot volume on the tract with an allowable error of $20,000$ cubic feet at the 95% confidence level. (On a per plot basis, the allowable error would be 5 cubic feet.) Compute equal, proportional, optimal equal cost (Neyman), and optimal unequal cost allocations and compute the total cost of each allocation. Briefly discuss.

4. A nursery manager wants to estimate the average height of seedlings in a large field that is divided into 25 plots that vary slightly in size. She believes the heights are fairly constant throughout each plot, but may vary considerably from plot to plot. Although this seems to indicate the need for a stratified random sample, for time and logistical reasons, she decides to use a two-stage sample, where with the low variability within a plot, she decides to sample 10% of the trees within each of 5 plots. The data are given in the table below. Estimate the average height of seedlings in the field, give the standard error, and find an approximate 95% confidence interval for the mean. [Problem taken from page 304, Scheaffer, Mendenhall, & Ott.]

	Number of	Number of	Heights of Seedlings					
Plot	Seedlings	Seedlings Sampled	(in inches)					
1	52	5	12	11	11	10	13	
2	56	6	10	9	7	8	8	10
3	60	6	6	5	7	5	6	4
4	46	5	7	8	6	7	6	
5	49	5	10	11	13	12	12	

5. An auditor wishes to sample sick-leave records of a large firm in order to estimate the average number of days of sick leave per employee over the past quarter. The firm has eight divisions, with varying numbers of employees per division. Since the number of days of sick leave used within each division should be highly correlated with the number of employees, the auditor decides to sample $n = 3$ divisions with probabilities proportional to the number of employees. The data are given in the table below.

Division	1	2	3	4	5	6	7	8
# of Employees	1200	550	2240	860	2800	1910	390	3200

   (a) Explain how, using a random number table or generator, such a sample could be taken. You do not need to do it - just explain it.

   (b) Suppose now that such a sample is taken where divisions 1, 3, and 8 are selected, and that the total number of sick days used by the three sampled divisions during the past quarter are, respectively,

$$y_1 = 2410, \quad y_2 = 4320, \quad y_3 = 5790.$$

   Estimate the average number of sick days used per person for the entire firm, and give the corresponding standard error.

6. Show that in two-stage random sampling with equal-sized primary units, that if m secondary units will be sampled within each primary unit, then the number of primary units n required to estimate the population mean per secondary unit to within d with probability $100(1-\alpha)\%$ is as given by

$$n = \frac{\sigma_b^2 + \left(\frac{\overline{M} - m}{\overline{M}}\right)\frac{\sigma_w^2}{m}}{\frac{d^2}{z^2} + \frac{\sigma_b^2}{N}}.$$

7. MATH ONLY Use conditional expectations to solve the following problems.

   (a) Derive the variance of a geometric random variable (we derived the expected value by conditioning in class).

   (b) What is the expected number of rolls of a fair die until you have obtained two consecutive 6's?