Assignment 5 — due February 25, 2019

1. This question is an example of a question that could appear on the in-class midterm. You are to answer it without using the computer.
   An experiment was conducted in the greenhouse as follows. 25 pots were arranged in a square 5 pots long and 5 pots wide. 5 different cultivars of potato (RB, YG, O, C, and W) were planted such that there was one potato plant per pot, and the cultivars were randomly arranged in a Latin Square design as shown below:

RB	YG	W	O	C
C	O	RB	YG	W
W	C	O	RB	YG
YG	RB	C	W	O
O	W	YG	C	RB

   After the plants had grown to maturity, for each pot three tubers were sampled from the pot. The tubers were then coated with a bacterium thought to cause bacterial ring rot disease: the amount was 10^6 colony forming units per tuber. The response is a measurement of bacterial ring rot severity after 25 days.

   Write down the ANOVA table for this experiment, providing source and degrees of freedom only.

2. Recall problem 4 from assignment 4, which said: "Note that the quantity 'branch total' refers to the total weight of the 10 apples for a given branch. The last two columns of the table give the branch total for the first and second branches sampled from a tree. For this analysis you should analyze the sum of the weights for the two branches as the dependent variable."

   Now repeat problem 4 from assignment 4, but do not analyze the sum of the weights for the two branches. Instead, analyze the full data set as provided in assignment 4.

3. This experiment concerns the success with which different species of plants establish themselves in a prairie. The prairie was established about 75 years ago, and is referred to as a "mature" prairie.

   For the purposes of this question, it is not too important to know the specific species of plants; we will simply refer to them as species A, B, C, and D. It is worth knowing that species A and D are thought to establish well in mature prairies, while species B and C are thought to establish well in other forms of prairies. For that reason, species A and D are sometimes called "mature" species, while species B and C are called "transitional" species. Note that none of the species studied in this experiment were present in the prairie before the experiment began.

   The experiment in the mature prairie was conducted as follows. 20 plots measuring 2 m by 2 m were located near the center of the mature prairie. The species were randomly assigned to the plots such that there were 5 plots for each of the species listed above. For a given plot and a given species, 800 seeds of the species were uniformly spread across the plot. Whatever existing plants that might be in the plot were allowed to remain. A map of the treatment assignment is given below.

T3 P1 15 13 14 16	T1 P2 10 8 17 22	T2 P3 11 6 7 10	T1 P4 10 11 12 18	T4 P5 18 18 17 16
T4 P6 17 17 18 21	T3 P7 5 12 7 10	T4 P8 15 18 20 20	T3 P9 6 3 3 4	T3 P10 6 7 7 6
T3 P11 10 3 4 4	T2 P12 7 10 8 8	T1 P13 10 12 14 8	T1 P14 6 6 12 5	T2 P15 6 10 6 8
T2 P16 11 12 8 10	T2 P17 0 5 3 4	T4 P18 20 19 17 18	T4 P19 15 17 10 8	T1 P20 10 11 6 9

   In this map, each rectangle represents a plot. The map reflects the shape of the experimental region: it was 5 plots wide and 4 plots long. In each rectangle of the map, the first value, starting with a "T," is the

treatment applied to that plot. The next value, starting with a "P," represents the plot number. Note that T1 corresponds to species A, T2 to B, etc.

The next four values in the square represent the observed data: One year after the species were planted, the experimenter went to each plot and randomly located, within the plot, four small, non-overlapping squares, each measuring 0.3 m by 0.3 m. Within each square the experimenter counted the number of seedlings of the species that was applied to that plot. So, for example, for Plot 1, the experimenter counted the number of seedlings corresponding to species C within each small square and obtained values 15, 13, 14, and 16. The higher this number, the better the species is said to establish.

(a) The data are available in file `prairiespecies.csv`. However, some typos were made (as is typical of real data, such as could appear on the Stat MS exam). Check for expected patterns to detect and fix the typos. Do not check consistency between the table displayed above and the data file *manually*: use `R`, like you would for a real data set.

(b) Using `R`, conduct an analysis of the data for the mature prairie, including any appropriate residual analyses. Do some species establish better than others? Taken as a group, do the mature species tend to establish better on average than the transitional species, taken as a group?

(c) The experimenter is considering repeating the experiment in a portion of the mature prairie that is expected to be similar to the portion already observed. When the experiment is repeated, the experimenter will use the same 4 species. However, she will use 8 plots for each species and only two 0.3 m by 0.3 m squares will be located within each plot. Provide an estimate of the variance of the sample mean for treatment 1 (species A).

4. Consider the balanced one-way random effects model in the context of subsampling: $Y_{ij} = \mu + \varepsilon_i + \delta_{ij}$ where $i = 1, \ldots, k$ indexes samples; $j = 1, \ldots, n$ indexes subsamples; $\varepsilon_i \sim \mathcal{N}(0, \sigma^2_\varepsilon)$ and $\delta_{ij} \sim \mathcal{N}(0, \sigma^2_\delta)$, all independent.

(a) In this model, two observations on the same sample, i.e. Y_{ij} and $Y_{ij'}$ are correlated. This correlation is sometimes called the "intraclass correlation". Determine this correlation in terms of σ^2_ε and σ^2_δ.

(b) Suppose $\sigma^2_\varepsilon = 15$ and $\sigma^2_\delta = 5$ and suppose that, for a given choice of k and n, you must pay $\text{var}(\bar{Y}_{\cdot\cdot})$ in dollars to estimate μ. In addition, it costs \$0.3 for each sample and it costs \$0.1 for each subsample. For example, if we take $k = 2$ samples and $n = 5$ subsamples per sample, then it costs $0.3 \times 2 + 0.1 \times 5 \times 2 =$ \$1.6 to take obtain such a data set. The total cost, then, is the sum of the cost due to variance, plus the cost of sampling. What choice of n and k minimizes the total cost?
Hint: there is little harm, and much benefit, to assuming that n and k need not be integers.

(c) Repeat part (c) if instead, $\sigma^2_\varepsilon = 5$ and $\sigma^2_\delta = 15$.

(d) Suppose that we build an ANOVA table for this model. Prove that $\mathbb{E}(\text{MSTrt}) = \sigma^2_\delta + n\sigma^2_\varepsilon$.

(e) Consider 2 ways to get a confidence interval for μ, based on two different methods to calculate the standard error for $\hat{\mu} = \bar{Y}_{\cdot\cdot}$:

i. average the subsamples to get a single value per sample and use the traditional one-sample standard error of $\hat{\mu}$;

ii. build the ANOVA table as in (d) and use $\sqrt{\dfrac{\text{MSTrt}}{nk}}$ for the standard error of $\hat{\mu}$.

Prove that these two methods are equivalent (when the design is balanced, as is assumed here).