Assignment 7 — due March 11, 2019

1. The point of this problem is to show what might happen with missing data.

   Note: An experimental design is said to be *unbalanced* if the numbers of observations for each treatment combination (or treatment and block combination) are unequal. A design is said to be *incomplete* if at least one treatment combination (or treatment and block combination) has zero observations, and otherwise it is *complete*. In that sense, a Randomized Complete Block Design is one where every treatment appears at least once in every block.

   We also use the word "complete" in a different sense. When we refer to a Completely Randomized Design, we use the word "complete" to mean that the randomization is as unconstrained as possible.

   Finally, note that a Latin Square Design is very incomplete, but in a very structured sort of way.

   An experiment was conducted to investigate the effects of combining 3 different fats with each of 3 different surfactants in the recipes for making bread. The outcome of interest was the specific volume of the bread. The experiment was conducted over 4 days; the days are thought to be blocking factors.

   The data are listed below, and are also in file `breadvolume.csv`. Note that many observations are missing — unfortunately there were problems with the ovens used to bake the bread and as a result, many loaves of bread had to be discarded.

		Day			
Fat	Surfactant	1	2	3	4
1	1	6.7	4.3	5.7	–
	2	7.1	–	5.9	5.6
	3	–	–	–	–
2	1	–	5.9	7.4	7.1
	2	–	–	–	–
	3	6.4	5.1	6.2	6.3
3	1	7.1	5.9	–	–
	2	7.3	6.6	8.1	6.8
	3	–	7.5	9.1	–

   (a) (Do this question by hand — do not use `R`.) Make an interaction plot. Based on this plot, does there seem to be an interaction between Fat and Surfactant? Explain your strategy. What interactions can be estimated?

   (b) (Do this question by hand again.) Make an ANOVA table for the analysis of these data, but with 2 columns only: source, and degrees of freedom.

   (c) (Use `R` for this question.) Perform an analysis of these data. In this question, use the default coefficient parametrization in `R`, which sets the first level (alphabetically) of each factor as the baseline level.

   Check your solution to (b).

   Is there evidence of differences between days? Do you get the same result using `anova` and `drop1`? Why? Which one is correct and why?

   What do you notice about coefficients for the interaction between Fat and Surfactant?

   Quantify the strength of evidence of an interaction between Fat and Surfactant.

   (d) (Use `R` again.) Repeat (c), but after changing the baseline level of Fat to level 3. What differs in your results, compared to (c)? Why might it be a good choice to set the baseline Fat level to 3?

   (e) (Use `R` again.) If the goal is to achieve the highest specific volume, what treatment combination(s) would you recommend?

2. Milk is tested after pasteurization to assure that pasteurization was effective. This experiment was conducted to determine variability in test results between laboratories, and to determine if the inter laboratory differences depend on the concentration of bacteria. Five contract laboratories are selected at random from those available in a large metropolitan area. Four levels of contamination are chosen at random by choosing four samples of milk from a collection of samples at various stages of spoilage. A batch of fresh milk from a dairy was obtained and split into 40 units. These 40 units are assigned at random to the twenty combinations of laboratory and contamination sample. Each unit is contaminated with 5 ml from its selected sample, marked with a numeric code, and sent to the selected laboratory. The laboratories count the bacteria in each sample by serial dilution plate counts without knowing that they received four pairs, rather than eight separate samples. The data (colony forming units per μl) are in file `milkcontamination.csv`.

   (a) (Do this question by hand.) Determine the most appropriate model to analyze these data: write a Hasse diagram, the model equation, and the formula that you might use in `R`. Also write the ANOVA table with source, degrees of freedom, and expected mean squares.

   What F test needs to be used to test each effect of interest? How does the answer depend on whether the terms associated with laboratories are fixed or random?

   (b) Use `R` to analyze these data. Determine if the effects of interest are present. If so, estimate them.
   If you find that the inter-laboratory differences depends on the contamination sample, is there a specific lab that we might blame?

3. To evaluate the possibility that the vitamin A content of baby food carrots may not be consistent, an experiment is planned. In one grocery store, four jars of carrots will be selected at random from each of the three brands of baby food that are sold in the region. From each jar, two samples (spoonfuls) will be taken and measured for their vitamin A content (for a total of 24 measurements).

   Determine the most appropriate model to analyze these data: write a Hasse diagram, the model equation, and the formula that you might use in `R`. Also write the ANOVA table with source, degrees of freedom, and expected mean squares.

   Which F test would need to be used to determine if vitamin A content varies across brands? across jars of the same brand?