Assignment 8 — due March 25, 2019

1. Five patients were selected for a study of the variability in serum cholesterol measurements over runs, for a hospital spectrophotometer. Eight blood sample tubes were prepared from each patient and analyzed in each of four spectrophotometer runs (two samples per patient and per run). The data is available in `cholestrol.csv`. Denote the measurement for patient i, run j, sample k.

   (a) Write a Hasse diagram and an appropriate linear model for Y_{ijk}. State the distributional assumptions being made.

   (b) Fit the model and report the ANOVA table. Provide the estimated variance components. Which factor explains most of the variance in the data?

   (c) Do the effects of patient and run appear to be additive? Explain.

   (d) The manufacturer of the equipment claims that results from this spectrophotometer are highly consistent from one run to the next, and that measurements from different runs will be no more variable than measurements from the same run. Under the model in part (a), write down the variance of the difference between two measurements from the same patient and same run

   $$\text{var}(Y_{ijk} - Y_{ijk'}) \quad \text{for} \quad k \neq k' \; ;$$

   and between two measurements from the same patient in different runs

   $$\text{var}(Y_{ijk} - Y_{ij'k'}) \quad \text{for} \quad j \neq j' \; .$$

   Under what conditions are these two variances the same? Test the null hypothesis that they are the same.

2. A study will be conducted to evaluate whether a 12-week community-level activity intervention increases the physical activity levels of rural communities. The experimenters plan to sample v rural villages. Before the 12-week intervention, n adult participants will be randomly sampled from each village to answer a questionnaire about their physical activity. The outcome of interest, Y, is the number of metabolic-equivalent minutes spent in physical activity of moderate intensity. Twelve months after the intervention, a new set of n adult participants will be randomly sampled from each village to answer the questionnaire.

   Based on other studies, it is thought that Y will need to be log-transformed before analysis to better follow a normal distribution within groups; that $\log(Y)$ has variance 0.77 between adults of the same village, and that Y variation between villages (that is, any variance component including village) is much smaller, around 0.02. The experimenters would like to detect an intervention effect if the intervention increases Y by 10% or more.

   What combination of v and n do you recommend, given that implementing the intervention in a village is somewhat costly?

3. The goal of this problem is determine which distributions of treatment effects maximize and minimize the power of the F test, given a fixed maximum difference between any pair of treatments. This maximum difference is what an experimenter would want to detect with 80% power, say, without further specification of all pairwise differences. A treatment effect will be detected if the F test results in a p-value below 0.05.

   In this problem, we consider the constraint that $|\mu_i| \leq 1$ for all i. This is to impose a maximum difference of at most 2 between any pair of treatments. We also assume that the analysis will involve an F test with denominator degree of freedom of 24, and with non-centrality parameter

$$\lambda = 3\sum_{i=1}^{a} \alpha_i^2 \quad \text{where } \alpha_i = \mu_i - \bar{\mu} \quad \text{and } \bar{\mu} = \frac{1}{a}\sum_{i=1}^{a} \mu_i.$$

   (a) What configuration(s) of treatment means $\boldsymbol{\mu}$ give the maximum power of the F test, given the constraint that $|\mu_i| \leq 1$ for all i? What is this maximum power when $a = 4$ and when $a = 5$?

   Hint: first prove that, for a configuration $\boldsymbol{\mu}$ with maximum power, $\mu_i = 1$ or $\mu_i = -1$ for every i. Then prove that exactly $\lfloor a/2 \rfloor$ or $\lceil a/2 \rceil$ of the μ_i are equal to 1 (the others being equal to -1). To prove that $|\mu_i| = 1$ for all i, consider a configuration $\boldsymbol{\mu}$ with maximum power and assume that $-1 < \mu_i < 1$ for some i. If $\mu_i > \bar{\mu}$, consider the modified configuration where $\mu_i' = \mu_i + \epsilon \leq 1$ for some small $\epsilon > 0$, all other μ_j remaining unchanged, and prove that power is strictly higher for the modified configuration (leading to a contradiction). If $\mu_i < \bar{\mu}$ consider the modification $\mu_i' = \mu_i - \epsilon \geq -1$ instead.

   (b) Assume that $\mu_1 = 1$, $\mu_2 = -1$ and $|\mu_i| \leq 1$ for all i. Under these constraints, what configuration(s) of treatment means $\boldsymbol{\mu}$ lead to the minimum power of the F test? What is this minimum power when $a = 4$ and $a = 5$? Compare with your answers in (a).

   Hint: first prove that $\sum_{i=1}^{a} \alpha_i^2 = 2 + \sum_{i=3}^{a} \alpha_i^{*2} + \frac{2(a-2)}{a}(\bar{\mu}^*)^2$ under the constraint in this question, where $\bar{\mu}^*$ is the average of $\mu_3, \ldots, \mu_a$ only, and $\alpha_i^* = \mu_i - \bar{\mu}^*$ for $i \geq 3$. Then prove that the configuration with minimum power is with $\mu_i = 0$ for all $i \geq 3$.

4. This problem aims to find optimal sample sizes when the goal of an experiment is to compare various treatments to a control. Assume a CRD with p treatment groups (named groups 1 through p) and one control group (named group 0). The data will be analyzed with p T-tests, each one comparing one treatment group to the control group. Traditionally, balanced designs are chosen with an equal sample size in each group. Is this a good choice in terms of power?

   Consider a design with n samples in each treatment group, and n_0 samples in the control group, for a total of $N = n_0 + pn$ samples. Assume that N is fixed by the total cost that the experimenters can afford. Also assume that p is fixed. Prove that the optimal choice satisfies $n_0 \approx \sqrt{p}\, n$. Conclude.