Homework 7

Due Thursday, 3/25/21 on Canvas.

1. Let $p_1, \dots, p_m$ be p-values for null hypotheses $H_{0,1}, \dots, H_{0,m}$, where $p_j \sim U[0,1]$ if $H_{0,j}$ is true. Define $\pi_0 \in [0,1]$ to be the fraction of the m hypotheses that are true.

   (a) Show that the Bonferroni procedure controls the FWER regardless of the dependence between $p_1, \dots, p_m$. That is, show that for any $\alpha \in (0,1)$,

   $$P\left(\exists j = 1, \dots, m \text{ such that } H_{0,j} \text{ is true and } p_j \leq \alpha/m\right) \leq \pi_0 \alpha.$$

   (b) If $p_1, \dots, p_m$ are independent, show that the Bonferroni procedure provides exact control of the FWER for small α. That is, show that for any $m \geq 1$ and significance level $\alpha \in (0,1)$,

   $$P\left(\exists j = 1, \dots, m \text{ such that } H_{0,j} \text{ is true and } p_j \leq \alpha/m\right) = \pi_0 \alpha \left\{1 + o(1)\right\}$$

   as $\alpha \to 0$.

   (c) Is (b) necessarily true if $p_1, \dots, p_m$ are dependent? If yes, prove it. If not, find a counterexample.

2. (Preliminaries for estimation in random effects models) Let $\boldsymbol{Y} \in \mathbb{R}^n$ be a random vector with $\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{0}$ and $\text{Var}(\boldsymbol{Y}) = \boldsymbol{\Sigma}(\boldsymbol{\theta}) \in \mathbb{R}^{n \times n}$, where $\boldsymbol{\theta} \in \mathbb{R}^p$ is an unknown parameter. Assume that $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is continuously differentiable as a function of $\boldsymbol{\theta}$, i.e. there exist continuous matrix functions $\boldsymbol{M}_j(\boldsymbol{\theta}) \in \mathbb{R}^{n \times n}$ for $j = 1, \dots, p$ such that

   $$\boldsymbol{\Sigma}(\boldsymbol{\theta} + \boldsymbol{\delta}) - \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \sum_{j=1}^{p} \boldsymbol{\delta}_j \boldsymbol{M}_j(\boldsymbol{\theta}) + o\left(\|\boldsymbol{\delta}\|_2\right).$$

   Let

   $$\ell\left(\boldsymbol{\theta}\right) = -\frac{1}{2} \log\left[\det\left\{\boldsymbol{\Sigma}(\boldsymbol{\theta})\right\}\right] - \frac{1}{2} \boldsymbol{Y}^T \left\{\boldsymbol{\Sigma}(\boldsymbol{\theta})\right\}^{-1} \boldsymbol{Y}$$

   be the log-likelihood for the normal distribution (up to additive constants that do not depend on $\boldsymbol{\theta}$). Note that we are NOT assuming $\boldsymbol{Y}$ is normally distributed.

   (a) Let $\boldsymbol{V} \in \mathbb{R}^{n \times n}$ be a symmetric, positive definite matrix. For any symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, prove the following:

   $$\log\left\{\det\left(\boldsymbol{V} + \epsilon \boldsymbol{A}\right)\right\} - \log\left\{\det(\boldsymbol{V})\right\} = \epsilon \, \text{Tr}\left(\boldsymbol{A}\boldsymbol{V}^{-1}\right) + o(\epsilon)$$
   $$\boldsymbol{u}^T \left(\boldsymbol{V} + \epsilon \boldsymbol{A}\right)^{-1} \boldsymbol{u} - \boldsymbol{u}^T \boldsymbol{V}^{-1} \boldsymbol{u} = -\epsilon \boldsymbol{u}^T \boldsymbol{V}^{-1} \boldsymbol{A} \boldsymbol{V}^{-1} \boldsymbol{u} + o(\epsilon)$$

   for $\epsilon > 0$ and $\boldsymbol{u} \in \mathbb{R}^n$.

(b) Use part (a) to show that

$$[\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})]_j = -\frac{1}{2} \operatorname{Tr}\left[\boldsymbol{M}_j(\boldsymbol{\theta}) \{\Sigma(\boldsymbol{\theta})\}^{-1}\right] + \frac{1}{2} \boldsymbol{Y}^T \{\Sigma(\boldsymbol{\theta})\}^{-1} \boldsymbol{M}_j(\boldsymbol{\theta}) \{\Sigma(\boldsymbol{\theta})\}^{-1} \boldsymbol{Y}.$$

Conclude that a root of $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$ is a suitable estimator for $\boldsymbol{\theta}$. That is, show that

$$\mathbb{E}\{\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})\} = \mathbf{0}$$

regardless of whether or not $\boldsymbol{Y}$ is normally distributed.

3. A study was done to compare the yields of 56 varieties of wheat in a randomized complete block design (RCBD) with four blocks of size 56. The data for the experiment are in the file "wheat56.txt". The four blocks are observations 1-56, 57-112, 113-168 and 169-224. The varieties, yields, latitudes and longitudes of each plot (latitudes and longitudes in unstated units) are given. Although the units are unstated, keep in mind that agricultural field trials like this are carried out at a single farm so that the weather is essentially the same at all plots. The labeling of the varieties as 1-56 "in order" in Block 1 is for convenience; you may assume that in fact the variety assignments were properly randomized in all four blocks.

   (a) Estimate the variety effects using the standard model for an RCBD treating blocks and varieties as fixed effects. Using appropriate tables and/or plots, summarize your findings about the differences between varieties. As part of your analysis, include an F-test for the hypothesis of no variety effects.

   (b) Find a 95% confidence interval for the mean yield of varieties 1-20 minus the mean yield of varieties 21-56.

   (c) Plot the residuals as a function of the geographic coordinates of the plots. Discuss any patterns you see and comment on the reasonableness of the assumptions underlying the analyses in (a). Can you identify any varieties whose yields (relative to other varieties) might be over or underestimated because of the plots to which they were assigned? Comment.

   (d) Reanalyze the data including a linear function of the coordinates in your mean function. What effect does this change have on your inferences about variety effects? In particular, which estimated variety effects change the most from the analysis in (c)? Plot the residuals as a function of the geographic coordinates of the plots. To what extent are any problems you noted with the residual plot in (c) fixed?

   (e) Answer the same questions as in (d), but this time including a quadratic function of the coordinates (i.e., a second order polynomial in latitude and longitude) in your mean function.

   (f) Do you think the design used for this study was well-chosen? Discuss any problems you see and describe how the study might have been designed differently to avoid or reduce these problems.

4. **PhD problem**: Proof of Sheffé. Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ be a full rank design matrix and $\boldsymbol{Y} \sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I_n\right)$ for some $\boldsymbol{\beta} \in \mathbb{R}^p$. Let $\mathcal{S} = \left\{\boldsymbol{c} \in \mathbb{R}^p : \sum_{j=1}^{p} \boldsymbol{c}_j = 0\right\}$.

   (a) If $\hat{\boldsymbol{\beta}}$ is the the ordinary least squares estimate for $\boldsymbol{\beta}$, show that for any $\alpha \in (0,1)$,

$$P\left[\left\{\frac{\boldsymbol{c}^T\hat{\boldsymbol{\beta}} - \boldsymbol{c}^T\boldsymbol{\beta}}{se\left(\boldsymbol{c}^T\hat{\boldsymbol{\beta}}\right)}\right\}^2 \leq q_{1-\alpha} \forall \boldsymbol{c} \in \mathcal{S}\right] \leq \alpha$$

   where $se\left(\boldsymbol{c}^T\hat{\boldsymbol{\beta}}\right) = \sqrt{\hat{\sigma}^2\boldsymbol{c}^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{c}}$ and $q_{1-\alpha}$ is the $1-\alpha$ quantile of the $(p-1)F_{p-1,n-p}$ distribution.

   (b) Use (a) to derive simultaneous $1-\alpha$ confidence intervals for every point in $\left\{\boldsymbol{c}^T\boldsymbol{\beta} : \boldsymbol{c} \in \mathcal{S}\right\}$.