STAT 510 Homework 10
Due Date: 11:00 A.M., Wednesday, April 17

1. For each of the following special cases, derive the REML estimator of σ^2.

   (a) Suppose $y_1, y_2 \overset{iid}{\sim} N(\mu, \sigma^2)$.

   (b) Suppose
   $$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \sigma^2/2 & 0 \\ \sigma^2/2 & \sigma^2 & \sigma^2/2 \\ 0 & \sigma^2/2 & \sigma^2 \end{bmatrix} \right).$$

2. Suppose 100 maize genotypes were assigned to 304 plots in a field using an unbalanced completely randomized design in which some genotypes were assigned to only one plot while others were assigned to as many as six plots. Plots were planted with seed from their assigned genotypes, and yield in bushels per acre was recorded for each plot at the end of the growing season. The dataset is available at

   dnett.github.io/S510/hw10GenotypeYield.txt.

   Consider the model
   $$y_{ij} = \mu + g_i + e_{ij},$$
   where $\mu + g_i$ is the mean yield for the ith genotype, and $e_{ij} \sim N(0, \sigma_e^2)$ for all i and j, with independence among all e_{ij} terms.

   (a) Find the BLUE of $\mu + g_i$ for each $i = 1, \ldots, 100$.

   (b) For this and all subsequent parts of this problem, assume $g_1, \ldots, g_{100} \overset{iid}{\sim} N(0, \sigma_g^2)$ and independent of all the e_{ij} terms. Find the REML estimates of σ_g^2 and σ_e^2.

   (c) Find the BLUP of $\mu + g_i$ for each $i = 1, \ldots, 100$.

   (d) Make a plot of the BLUPs (vertical axis) vs. the BLUEs from part (a) (horizontal axis) with one point for each genotype. Add the $y = x$ line to your plot. Explain why the plot looks the way it does.

   (e) According to the BLUEs from part (a), list the top five highest yielding genotypes.

   (f) According to the BLUPs, list the top five highest yielding genotypes.

   (g) Why is the top-yielding genotype according to the BLUEs from part (a) not so highly rated according to the BLUPs?

3. Suppose a total of 14 plants, 7 of genotype 1 and 7 of genotype 2, are randomly positioned in a growth chamber. Fungal spores are showered over all 14 plants in the growth chamber. After 5 days, the number of fungus-induced lesions appearing on the top leaf of each plant is determined. For $i = 1, 2$ and $j = 1, \ldots, 7$, let y_{ij} be the number of fungus-induced lesions on the top leaf of plant j for genotype i. Suppose all y_{ij} counts are independent and that
   $$y_{ij} \sim \text{Poisson}(\theta_i) \tag{1}$$
   for some unknown positive parameters θ_1 and θ_2. The observed counts $\{y_{ij} : i = 1, 2; j = 1, \ldots, 7\}$ are provided in the following table.

Genotype	Numbers of Fungus-Induced Top-Leaf Lesions						
1	14	9	10	5	18	9	9
2	17	10	17	18	13	17	16

(a) Compute AIC for model (1).

(b) Compute BIC for model (1).

(c) Compute AIC for a simplified version of model (1) in which $\theta_1 = \theta_2$.

(d) Compute BIC for a simplified version of model (1) in which $\theta_1 = \theta_2$.

(e) Which of the two models is preferred according to AIC?

(f) Which of the two models is preferred according to BIC?

(g) Compute the likelihood ratio test statistic $-2 \ln \Lambda$ for testing $H_0 : \theta_1 = \theta_2$.

(h) Find the p-value corresponding to the likelihood ratio statistic in part (g).

(i) Compute a Wald statistic for testing $H_0 : \theta_1 = \theta_2$.

(j) Find the p-value corresponding to the Wald statistic in part (i).

4. For $\theta \in [0, 1]$, a random variable has a Bernoulli distribution with success probability θ if it takes the value 1 with probability θ and the value 0 with probability $1 - \theta$. We use $y \sim$ Bernoulli(θ) to indicate that the random variable y has a Bernoulli distribution with success probability θ.

   If $y \sim$ Bernoulli(θ), it follows that the probability mass function of y is

$$P(y = k) = \begin{cases} \theta^k (1 - \theta)^{(1-k)} & \text{for } k \in \{0, 1\} \\ 0 & \text{otherwise} \end{cases} .$$

   Suppose $y_1, \ldots, y_n \overset{iid}{\sim}$ Bernoulli(θ) and complete the following problems for this special case.

   (a) Provide an expression for the likelihood function.

   (b) Write down the score equation.

   (c) Solve the score equation.

   (d) Verify that the solution to the score equation maximizes the likelihood function; i.e., verify that the solution to the score equation is an MLE for θ.

   (e) Find the Fisher information matrix. (This is just a 1×1 matrix in this case.)

   (f) Find the inverse of the Fisher information matrix.

   (g) Verify that the inverse of the Fisher information matrix gives the exact variance of the MLE in this special case.

   (h) Provide an expression for an estimator of the variance of the MLE by replacing the unknown parameter θ with the MLE of θ in the inverse Fisher information matrix.

5. Suppose researchers wish to estimate the probability that a certain type of plant will die when infected with a virus. They randomly select 100 seeds from a large collection of seeds for the plant type of interest. They grow the seeds into plants, infect the plants, and discover that 17 of the plants die as a result of the infection.

   (a) Using a Wald-based approach (slide 17 of slide set 22), find an approximate 95% confidence interval for the proportion of all plants of this type that would die as a result of infection with the virus.

   (b) Using a likelihood-based approach (slide 31 of slide set 22), find an approximate 95% confidence interval for the proportion of all plants of this type that would die as a result of infection with the virus.

6. Scientists studied 700 adults who typically smoked at least 15 cigarettes a day but wished to quit smoking. The subjects were randomly divided into two treatment groups using a balanced and completely randomized design. In treatment group 1, subjects were allowed to smoke as usual for two weeks and then asked to suddenly quit smoking altogether for as long as possible. In treatment group 2, subjects gradually reduced cigarette smoking for two weeks according to a specific smoking reduction plan and then quit smoking altogether for as long as possible. Four weeks after the quit date, 172 of the 350 people assigned to treatment group 1 were still not smoking, and 137 of the 350 people assigned to treatment group 2 were still not smoking. All other subjects had resumed smoking within four weeks of the quite date. Is one treatment more effective than the other at enabling people to quit smoking for at least four weeks? Construct one confidence interval that will help you answer this question.