

1. Consider the model

$$y_i = \beta x_i + |x_i| \varepsilon_i,$$

where for  $i = 1, \dots, n$ ;  $y_i$  is the response for observation  $i$ ;  $\beta$  is an unknown real-valued parameter;  $x_i$  is the  $i$ th known nonzero observation of an explanatory variable;  $\varepsilon_1, \dots, \varepsilon_n$  are independent and identically distributed as  $N(0, \sigma^2)$ ; and  $\sigma^2 > 0$  is an unknown variance component.

- (a) Provide an expression for the variance of the response vector  $\mathbf{y}$ .
- (b) Provide expression for the best linear unbiased estimator of  $\beta$ . Simplify your answer as much as possible.

2. Researchers were interested in studying how soil moisture level affects the ability of plants to respond to a virus infection. A total of 30 pots were assigned to three watering levels (1 = low, 2 = medium, 3 = high) using a balanced and completely randomized design. Each of the 30 pots contained four seedlings. Two randomly selected seedlings within each pot were injected with a virus. The remaining two seedlings in each pot were “mock infected” by injection with a harmless substance. Two weeks after treatment, each seedling was individually weighed, and these weights served as the response variable for subsequent analysis.

- (a) Name the experimental units in this experiment.
- (b) Name the observational units in this experiment.
- (c) Provide an ANOVA table for the data from this experiment. Include only the columns for source and degrees of freedom.

3. Suppose  $w_1 \sim N(0, 1)$ . Suppose the conditional distribution of  $w_2$  given  $w_1$  is  $N(w_1, 1)$ .

- (a) Determine the joint distribution of  $[w_1, w_2]'$ . (Hint: One simple way to do this is to introduce  $\varepsilon \sim N(0, 1)$  independent of  $w_1$ . Then write  $[w_1, w_2]'$  as a linear function of  $[w_1, \varepsilon]'$ .)
- (b) Suppose  $w_2 = 1.7$ . What is the best linear unbiased prediction for  $w_1$ ?

4. An experiment was conducted to compare the effectiveness of two sports drinks (denoted 1 and 2). The subjects included 60 males between the ages of 18 and 31. Each subject rode a stationary bicycle until his muscles were depleted of energy, rested for two hours, and biked again until exhaustion. During the rest period, each subject drank one of the two sports drinks as assigned by the researchers. Each subject’s performance on the second round of biking following the rest period was assigned a score between 0 and 100 based on the energy expended prior to exhaustion. Higher scores were indicative of better performance.

20 of the 60 subjects repeated the bike-rest-bike trial on a second occasion separated from the first by approximately three weeks. These subjects drank one sports drink during the first trial and the other during the second trial. The drink order was randomized for each subject by the researchers, even though previous research suggested no performance difference in repeated trials when three weeks passed between trials. The other 40 subjects performed the trial only a single time, drinking a randomly assigned sports drink during the rest period. 20 of these subjects received sports drink 1, and the other 20 received sports drink 2. A portion of the entire data set is provided below.

Subject	Drink 1	Drink 2
1	45	52
2	69	73
$\vdots$	$\vdots$	$\vdots$
20	29	46
21	35	-
22	81	-
$\vdots$	$\vdots$	$\vdots$
40	55	-
41	-	17
42	-	54
$\vdots$	$\vdots$	$\vdots$
60	-	61

Subjects 1 through 20 in the table above represent the 20 subjects who performed the trial separately for each of the sports drinks. Note that the data set contains no information about which drink was received in the first trial and which drink was received in the second trial. Throughout the remainder of this problem, please assume that this information is not important. In other words, you may assume that the subjects would have scored the same for drinks 1 and 2 regardless of the order the trials were performed.

Suppose the following model is appropriate for the data.

$$y_{ij} = \mu_i + u_j + e_{ij}, \quad (1)$$

where  $y_{ij}$  is the score for drink  $i$  and subject  $j$ ,  $\mu_i$  is the unknown mean score for drink  $i$ ,  $u_j$  is a random effect corresponding to subject  $j$ , and  $e_{ij}$  is a random error corresponding to the score for drink  $i$  and subject  $j$  ( $i = 1, 2$  and  $j = 1, \dots, 60$ ). Here  $u_1, \dots, u_{60}$  are assumed to be independent and identically distributed as  $N(0, \sigma_u^2)$  and independent of the  $e_{ij}$ 's, which are assumed to be independent and identically distributed as  $N(0, \sigma_e^2)$ .

- For each of the subjects that received both drinks, the difference between the scores (drink 1 score – drink 2 score) was computed. This yielded 20 score differences denoted  $d_1, \dots, d_{20}$ . Describe the distribution of these differences considering the assumptions about the distribution of the original scores in model (1).
- Suppose you were given only the differences  $d_1, \dots, d_{20}$  from part (a). Provide a formula for a test statistic (as a function of  $d_1, \dots, d_{20}$ ) that could be used to test  $H_0 : \mu_1 = \mu_2$ .
- Fully state the exact distribution of the test statistic provided in part (b).
- Let  $a_1, \dots, a_{20}$  be the scores of the subjects that received only drink 1. Let  $b_1, \dots, b_{20}$  be the scores of the subjects that received only drink 2. Suppose you were given only these 40 scores. Provide a formula for a 95% confidence interval for  $\mu_1 - \mu_2$  (as a function of  $a_1, \dots, a_{20}$  and  $b_1, \dots, b_{20}$ ).
- Suppose you were given  $d_1, \dots, d_{20}$  from part (a) and  $a_1, \dots, a_{20}$  and  $b_1, \dots, b_{20}$  from part (d). Provide method of moments estimates of  $\sigma_u^2$  and  $\sigma_e^2$  as a function of these observations.
- Suppose you were given  $\bar{d} = \sum_{i=1}^{20} d_i/20$ ,  $\bar{a} = \sum_{i=1}^{20} a_i/20$ , and  $\bar{b} = \sum_{i=1}^{20} b_i/20$ ; where  $d_1, \dots, d_{20}$  are from part (a) and  $a_1, \dots, a_{20}$  and  $b_1, \dots, b_{20}$  are from part (d). Provide a simplified expression for the best linear unbiased estimate of  $\mu_1 - \mu_2$  in terms of  $\bar{d}$ ,  $\bar{a}$ ,  $\bar{b}$ ,  $\sigma_u^2$ , and  $\sigma_e^2$ .