Homework 1

Brian Cervantes Alvarez
April 10, 2024
ST553 Statistical Methods

Problem 1

[2 points] Oehlert gives the "separate means model" on page 37. For this question, assume there are n_i observations in the i^{th} group. Under this model, what is the distribution of the i^{th} sample mean $\bar{y}_{i\cdot}$? Give an expression for an estimate of $\text{var}(\bar{y}_{i\cdot})$. The term standard error refers to an estimate of standard deviation. Give an expression for $SE(\bar{y}_{i\cdot})$.

Solution

We know that the i^{th} sample mean $\bar{y}_{i\cdot}$ follows a normal distribution such that $\bar{y}_{i\cdot} \sim N\left(\mu_i, \frac{\sigma_i^2}{n_i}\right)$, where μ_i and σ_i^2 represent the population mean and variance for the i^{th} group, respectively, and n_i is the count of observations in that group. The estimate of variance for $\bar{y}_{i\cdot}$ is given by $\hat{\text{var}}(\bar{y}_{i\cdot}) = \frac{s_i^2}{n_i}$, which leads to a standard error: $SE(\bar{y}_{i\cdot}) = \sqrt{\frac{s_i^2}{n_i}}$. which quantifies the variability in the average outcome estimation for the i^{th} group. Therefore,

$$\hat{\text{var}}(\bar{y}_{i\cdot}) = \frac{s_i^2}{n_i},$$

$$SE(\bar{y}_{i\cdot}) = \sqrt{\frac{s_i^2}{n_i}}$$

Problem 2

Given Display 3.2 on page 44 of Oehlert which provides the point estimator and standard error for treatment effect α_i, justify the expression for the standard error of $\hat{\alpha}_1$ by showing that its variance is $\sigma^2\left(\frac{1}{n_1}-\frac{1}{N}\right)$. Follow the suggested approach below.

Solution

$$
\begin{aligned}
\hat{\alpha}_1 &= \bar{y}_{1\cdot} - \bar{y}_{\cdot\cdot} \\
&= \frac{1}{n_1}\sum_{j=1}^{n_1} y_{1j} - \frac{1}{N}\sum_{i=1}^{g}\sum_{j=1}^{n_i} y_{ij} \\
&= \left(\frac{1}{n_1}-\frac{1}{N}\right)\sum_{j=1}^{n_1} y_{1j} + \frac{1}{N}\sum_{i=2}^{g}\sum_{j=1}^{n_i} y_{ij} \\
\operatorname{var}(\hat{\alpha}_1) &= \operatorname{var}\left(\left(\frac{1}{n_1}-\frac{1}{N}\right)\sum_{j=1}^{n_1} y_{1j}\right) + \operatorname{var}\left(\frac{1}{N}\sum_{i=2}^{g}\sum_{j=1}^{n_i} y_{ij}\right) \\
&= \left(\frac{1}{n_1}-\frac{1}{N}\right)^2 \operatorname{var}\left(\sum_{j=1}^{n_1} y_{1j}\right) + \frac{1}{N^2}\sum_{i=2}^{g} n_i\sigma^2 \\
&= \left(\frac{1}{n_1}-\frac{1}{N}\right)^2 n_1\sigma^2 + \frac{N-n_1}{N^2}\sigma^2 \\
&= \left(\frac{1}{n_1}-\frac{1}{N}\right)\sigma^2 \\
SE(\hat{\alpha}_1) &= \sqrt{\operatorname{var}(\hat{\alpha}_1)} \\
&= \sigma\sqrt{\left(\frac{1}{n_1}-\frac{1}{N}\right)}
\end{aligned}
$$

Problem 3

The file `fats.csv` contains data from an experiment to compare absorption of five different fats when preparing French fries. Fifteen batches of fries were randomly allocated to the five types of fat in a balanced completely randomized design. One batch is fried at a time. The column fat contains codes (A, B, C, D, and E) denoting fats. The column Y gives the grams of fat absorbed by the batch of fries. Of interest is whether fat absorption differs among the fat types

3.1

[1 point] What is the experimental unit in this study? Explain briefly.

The experimental unit is a single batch of fries, since each batch being the smallest entity that independently receives a specific treatment.

3.2

[2 points] Read the data into SAS. Compute the ANOVA table and least squares means with their standard errors using SAS. Include SAS code, ANOVA table, and a table showing the least squares means and standard errors

SAS CODE

```
PROC ANOVA DATA=fats;
    CLASS fat;
    MODEL Y = fat;
RUN;

PROC GLM DATA=fats;
    CLASS fat;
    MODEL Y = fat;
    LSMEANS fat / STDERR CL;
RUN;
```

ANOVA TABLE & LEAST SQAURES MEAN

The ANOVA Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	764.666667	191.166667	3.27	0.0586
Error	10	584.666667	58.466667		
Corrected Total	14	1349.333333			

R-Square	Coeff Var	Root MSE	Y Mean
0.566700	18.49923	7.646350	41.33333

Source	DF	Anova SS	Mean Square	F Value	Pr > F
fat	4	764.6666667	191.1666667	3.27	0.0586

The GLM Procedure
Least Squares Means

fat	Y LSMEAN	Standard Error	Pr > |t|
A	51.6666667	4.4146222	<.0001
B	47.6666667	4.4146222	<.0001
C	38.0000000	4.4146222	<.0001
D	32.6666667	4.4146222	<.0001
E	36.6666667	4.4146222	<.0001

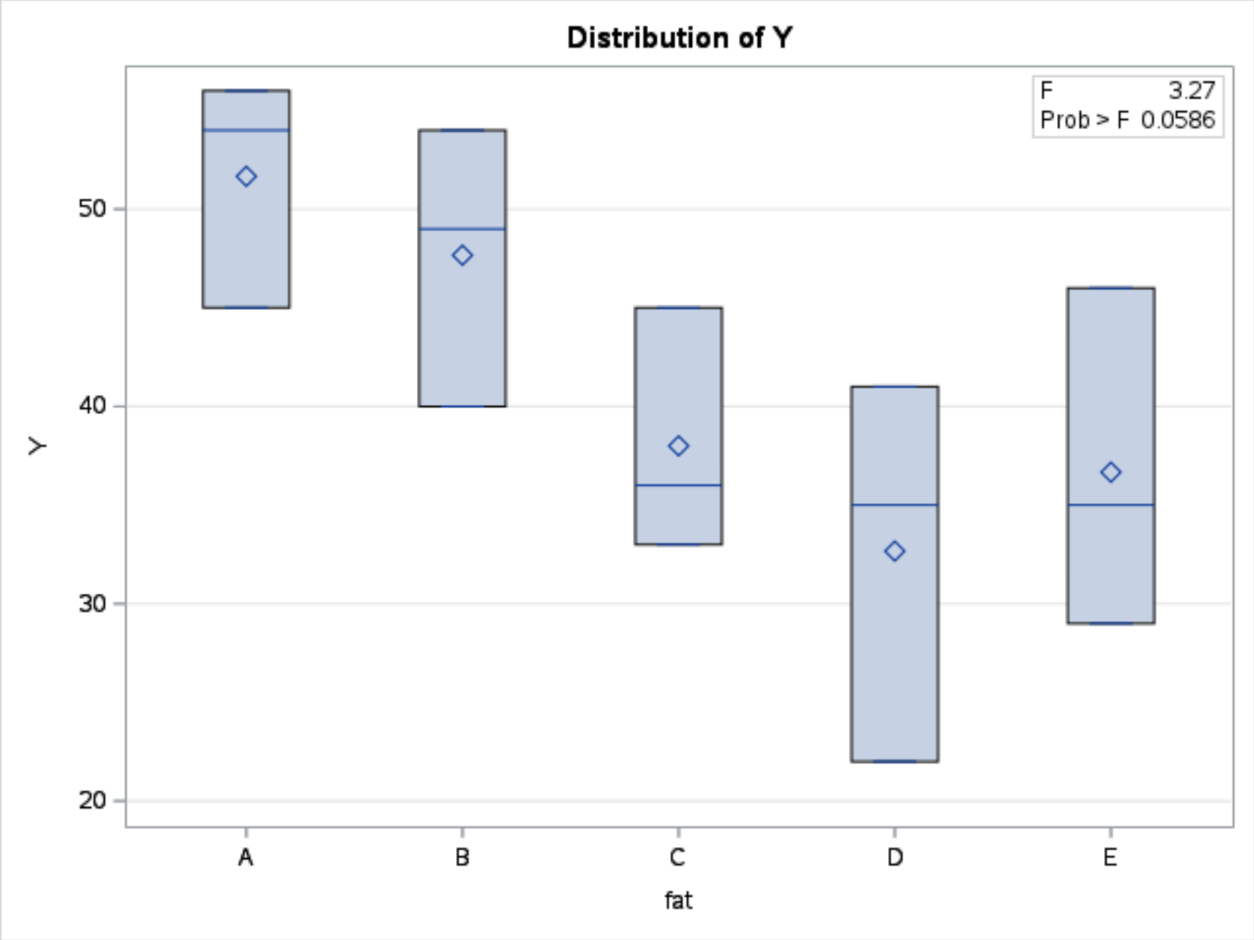
fat	Y LSMEAN	95% Confidence Limits	
A	51.666667	41.830275	61.503058
B	47.666667	37.830275	57.503058
C	38.000000	28.163609	47.836391
D	32.666667	22.830275	42.503058
E	36.666667	26.830275	46.503058

3.3

[3 points] Write a short report (at least one sentence and at most one paragraph) answering the research question. Include side-by-side boxplots. Please see this Canvas page for guidelines about reporting results of a statistical analysis

Our ANOVA test showed a p-value of 0.0586, which means we're not quite sure (95% confidence) if some fats absorb more than others. Despite visual indications from the boxplot suggesting variability, the evidence is insufficient to refute the null hypothesis of no absorption rate differences across fats.

BOXPLOT



Problem 4

Using the fat absorption data, do the following. Use `R` or `Matlab` to do the matrix calculations. `SAS` can perform matrix calculations, but it's rather clunky. You can copy and paste the needed portions of the `R` or `Matlab` transcript into your homework document

4.1

[2 points] Using the cell means parametrization, write down a model and assump- tions for the data and analysis. Define all notation (e.g. what does μ_1 represent in the context of the study?)

The model under the cell means parametrization can be expressed as:

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

where:

- Y_{ij} is the response variable (fat absorption) for the j^{th} observation in the i^{th} treatment group.
- μ_i is the mean per i^{th} treatment group.
- ϵ_{ij} is the error term for the j^{th} observation in the i^{th} treatment, assumed to be $N(0, \sigma^2)$ and independent.

In this context, μ_1 would represent the mean response for the first treatment group, adjusted for the overall mean.

4.2

[3 points] Give the elements of the response vector Y and the design matrix X (cell means parametrization). Give the elements of the parameter vector β in terms of your model in part 1.

```r
library(readr)
ds <- read_csv("fats.csv")
fatModel <- lm(Y ~ fat - 1, data = ds)
# Model Matrix X
X <- model.matrix(fatModel)
# Vector Y
Y <- ds$Y
```

$$X = \begin{bmatrix} \text{fatA} & \text{fatB} & \text{fatC} & \text{fatD} & \text{fatE} \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, Y = \begin{bmatrix} 45 \\ 56 \\ 54 \\ 40 \\ 54 \\ 49 \\ 45 \\ 36 \\ 33 \\ 35 \\ 22 \\ 41 \\ 35 \\ 29 \\ 46 \end{bmatrix}$$

4.3

[4 points] Calculate X^TX, $(X^TX)^{-1}$, X^TY and $(X^TX)^{-1}X^TY$, and the least squares estimate $\hat{\beta} = (X'X)^{-1}X'Y$. The elements of the least squares estimate vector should be the least square means in your SAS output in question 3 part 2.

R-Code

```
# X^TX
XTX <- t(X) %*% X
# (X^TX)^-1
inverseXTX <- solve(t(X) %*% X)
# X^TY
XTY <- t(X) %*% Y
# Least squares estimator
betaHat <- solve(t(X) %*% X) %*% t(X) %*% Y
```

Matrix Solutions

$$X^TX = \begin{bmatrix} 3 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{bmatrix}$$

$$(X^TX)^{-1} = \begin{bmatrix} 0.3333333 & 0 & 0 & 0 & 0 \\ 0 & 0.3333333 & 0 & 0 & 0 \\ 0 & 0 & 0.3333333 & 0 & 0 \\ 0 & 0 & 0 & 0.3333333 & 0 \\ 0 & 0 & 0 & 0 & 0.3333333 \end{bmatrix}$$

$$X^TY = \begin{bmatrix} 155 \\ 143 \\ 114 \\ 98 \\ 1109 \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} fatA & 51.66667 \\ fatB & 47.66667 \\ fatC & 38.00000 \\ fatD & 32.66667 \\ fatE & 36.66667 \end{bmatrix}$$

4.4

[1 point] Using R, calculate the sample means for each of the five fats. Confirm that these are also the same as the least squares means in question 3 part

```
library(dplyr)
sampMeans <- ds %>%
  group_by(fat) %>%
  summarize(MeanAbsorption = mean(Y))
print(sampMeans)
```

```
# A tibble: 5 x 2
  fat   MeanAbsorption
  <chr>          <dbl>
1 A               51.7
2 B               47.7
3 C               38
4 D               32.7
5 E               36.7
```

4.5

[2 points] Give the design matrix X, and the parameter vector β for the factor effects parametrization of the model.

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}$$

This configuration allows the model to estimate and compare the mean fat absorption across different treatment groups directly, with each β coefficient indicating the mean absorption for its corresponding fat treatment.

Problem 5

[2 points] In setting up the French fry experiment, the researchers had to randomly allocate the batches to the fat types. Use SAS to allocate 15 batches to five treatments in a balanced completely randomized design. Include SAS code and explicit instructions that a research assistant could use to assign the batches to the fats.

```
proc setinit; run;
%let seed = 12345;
/* Generate the randomized allocation */
proc plan seed=&seed;
   /* Specify the number of factors and their levels */
   factors batches=15 /* 15 batches */
           fats=5    /* 5 types of fats, numerically coded */
           cyclic;
   /* Assign batches to fats in a balanced way */
   output out=RandomAllocation;
run;
/* Map numeric codes back to fat labels */
data AllocatedBatches;
   set RandomAllocation;
   length FatType $ 1;
   if fats = 1 then FatType = 'A';
   else if fats = 2 then FatType = 'B';
   else if fats = 3 then FatType = 'C';
   else if fats = 4 then FatType = 'D';
   else if fats = 5 then FatType = 'E';
run;
/* Display the allocation with fat labels */
proc print data=AllocatedBatches;
   var batches FatType;
   title "Random Allocation of Batches to Fat Types";
run;
```

The PLAN Procedure

Factor	Select	Levels	Order	Initial Block / Increment
batches	15	15	Random	
fats	5	5	Cyclic	(1 2 3 4 5) / 1

batches	fats				
6	1	2	3	4	5
12	2	3	4	5	1
13	3	4	5	1	2
7	4	5	1	2	3
4	5	1	2	3	4
3	1	2	3	4	5
5	2	3	4	5	1
1	3	4	5	1	2
8	4	5	1	2	3
14	5	1	2	3	4
9	1	2	3	4	5
2	2	3	4	5	1
11	3	4	5	1	2
10	4	5	1	2	3
15	5	1	2	3	4

1. **Open SAS** and prepare for scripting.
2. **Copy and paste the provided SAS code** into the script editor. This code will create a balanced, randomized allocation of batches to fats, converting numeric codes to fat labels A through E.
3. **Execute the script** to generate and map the random allocation of fats.
4. **Review the output** titled "Random Allocation of Batches to Fat Types", and use it to accurately allocate batches in your experiment.
5. **Document the allocation** as per the output for consistent reference.

Random Allocation of Batches to Fat Types

Obs	batches	FatType
1	6	A
2	6	B
3	6	C
4	6	D
5	6	E
6	12	B
7	12	C
8	12	D
9	12	E
10	12	A
11	13	C
12	13	D
13	13	E
14	13	A
15	13	B
16	7	D
17	7	E
18	7	A
19	7	B
20	7	C
21	4	E
22	4	A
23	4	B
24	4	C
25	4	D
26	3	A
27	3	B
28	3	C
29	3	D
30	3	E
31	5	B
32	5	C
33	5	D
34	5	E
35	5	A
36	1	C
37	1	D
38	1	E
39	1	A
40	1	B
41	8	D
42	8	E
43	8	A
44	8	B
45	8	C
46	14	E
47	14	A
48	14	B
49	14	C
50	14	D
51	9	A
52	9	B
53	9	C
54	9	D
55	9	E
56	2	B
57	2	C
58	2	D
59	2	E
60	2	A
61	11	C
62	11	D
63	11	E
64	11	A
65	11	B
66	10	D
67	10	E
68	10	A
69	10	B
70	10	C
71	15	E
72	15	A
73	15	B
74	15	C
75	15	D