STAT 500 Homework 1

Vu Thi-Hong-Ha; NetID: 851924086

August 28, 2020

1 Question 1

Experiment I:
- Experimental units: Pots of one-week-old seedlings.
- Observation units: The fresh weight of each seedling.
- Treatment: Low or high nitrogen treatment.
- Response variables: Biomass.
- Replication: yes. Six pots for each genotype; for each genotype, three pots for different treatments.
- Blocking: no. Treatment is done within genotypes, but genotypes are a factor of interest.
- Randomization: yes. For each genotype, the researchers randomly assigned three pots to a nitrogen level.

Experiment II:
- Experimental units: Introductory business statistics 226 classes.
- Observation units: Final exam scores.
- Treatment: Using clickers to respond to questions or not.
- Response variables: Student learning.
- Replication: no. There is only one class subjected to each treatment.
- Blocking: no. There is no grouping of similar experimental units other than using clickers and not using clickers (i.e. treatment and untreatment).
- Randomization: yes. The professor tossed a coin to choose the class to use clickers. Other sources of bias are reduced (same books, same assignments, same exams).

2 Question 2

Design 1:
- Experimental units: farms.
- Observation units: the ribeye area of steers in the farms at slaughter.
- Treatment: 5 different diets.
- Replication: no. There is only one farm subjected to each diet.
- Blocking: no.
- Randomization: yes.

Design 2:
- Experimental units: steers.
- Observation units: the ribeye area of steers in the farms at slaughter.
- Treatment: 5 different diets.
- Replication: yes. There are five steers subjected to each diet.
- Blocking: yes. Farms can be a blocking factor. This is also a balanced design where the number of replicates in each treatment is equal.
- Randomization: yes.

Therefore, design 2 is better.

3 Question 3

(a) This study is an experiment because the patients are subjected to treatment or not, then the effects of treatment are observed.
(b)
- Experimental units: patients.

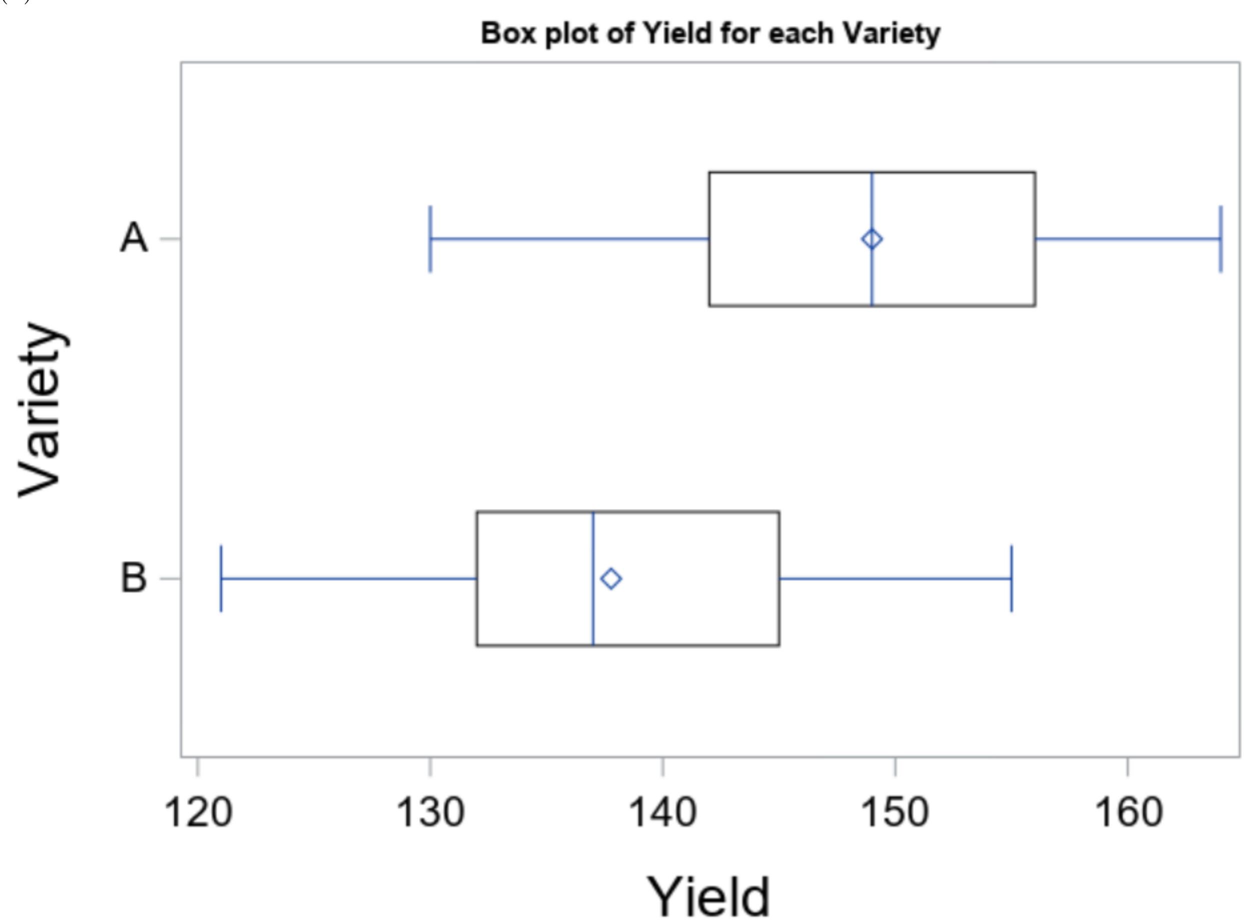
- Response variable: surviving time.
- Observation unit: the time each patient survive in three years following the second treatment.
- Replication: yes. There are 57 patients being treated the second time, and 43 patients not being treated.
- Blocking: no.
- Blinding: no. Patients are asked if they agree to be treated the second time, so they know that they are being treated or not.
- Control of extraneous variation: yes. Every patient is treated with the highest current standard of care, so the variation between patients' conditions might be minimized.
- The study compares treatment group with an untreatment group (control group).
- However, the conclusion from this study might not be used to generalize for every population because the sample is not random (only limited to recruited patients). The conclusions can be used to infer causal relationship related to the treatment.

4 Question 4

(a)

Statistics	Variety A	Variety B
Q1	142	132
Q3	156	145
IQR	14	13
Median	149	137
Sample Mean	149	137.7778
Sample Std. Dev.	9.22910996	9.13658

(b)



(c)
From (a) and (b), we can see that there is some variation but not much between individuals in each variety, which is expected. There is no extreme value or outlier in both of the varieties. The yield of plants in variety A tend to be higher, but this needs to be tested. The variety A seems to be symmetric (median = mean), and the variety B

is slightly right skewed.
(d)
Null hypothesis: The difference between mean yield of the two variety is 0.
Alternative hypothesis: The difference between mean yield of the two variety is not 0.
Observed test statistics: 11.2222.
By doing permutation 10000 times, we get the differences that look like this:

Differences in Means for the First 20 Permutations

Obs	_sample_	n1	mean1	n2	mean2	diff
1	1	18	144.889	18	141.889	3.00000
2	2	18	143.167	18	143.611	-0.44444
3	3	18	144.833	18	141.944	2.88889
4	4	18	140.944	18	145.833	-4.88889
5	5	18	145.778	18	141.000	4.77778
6	6	18	142.333	18	144.444	-2.11111
7	7	18	144.278	18	142.500	1.77778
8	8	18	143.889	18	142.889	1.00000
9	9	18	143.889	18	142.889	1.00000
10	10	18	144.333	18	142.444	1.88889
11	11	18	142.444	18	144.333	-1.88889
12	12	18	142.889	18	143.889	-1.00000
13	13	18	141.056	18	145.722	-4.66667
14	14	18	142.333	18	144.444	-2.11111
15	15	18	143.778	18	143.000	0.77778
16	16	18	143.778	18	143.000	0.77778
17	17	18	141.500	18	145.278	-3.77778
18	18	18	143.333	18	143.444	-0.11111
19	19	18	146.111	18	140.667	5.44444
20	20	18	145.222	18	141.556	3.66667

Out of 10000 differences obtained by permutation, there are 8 differences that are more extreme than the observed difference:

Differences as Extreme as the Observed Difference (11.2222)

Obs	_sample_	n1	mean1	n2	mean2	diff
1	799	18	137.722	18	149.056	-11.3333
2	1628	18	149.667	18	137.111	12.5556
3	2627	18	149.056	18	137.722	11.3333
4	3939	18	137.167	18	149.611	-12.4444
5	4766	18	137.389	18	149.389	-12.0000
6	8787	18	149.000	18	137.778	11.2222
7	9441	18	137.667	18	149.111	-11.4444
8	9579	18	137.667	18	149.111	-11.4444

Hence, the approximate p-value is $\frac{8}{10000} = 0.0008$. Therefore, we reject the null hypothesis at the significance level of 0.05.
In conclusion, the two varieties do not produce equal yields. If we carry out one tail t-test, we can also see that

variety A has a significantly higher yields than variety B.

5 Question 5

(a) The observed sample mean difference is $|\frac{4.8+5.2+5.0}{3} - \frac{7.7+8.2+8.1}{3}| = 3$

(b) The number of ways to assign treatment label is $\frac{6!}{3! \times 3!} = 20$.

(c) Permuted dataset:

All Permutations

Obs	_sample_	y1	y2
1	1	8.2	5.0
2	1	5.2	7.7
3	1	8.1	4.8
4	2	8.1	5.0
5	2	4.8	7.7
6	2	5.2	8.2
7	3	5.0	7.7
8	3	8.2	4.8
9	3	5.2	8.1
10	4	7.7	8.1
11	4	5.2	4.8
12	4	5.0	8.2
13	5	5.2	8.2
14	5	8.1	5.0
15	5	7.7	4.8
16	6	4.8	8.2
17	6	5.0	5.2
18	6	7.7	8.1
19	7	5.0	8.1
20	7	5.2	7.7
21	7	4.8	8.2
22	8	7.7	8.2
23	8	4.8	8.1
24	8	5.2	5.0
25	9	5.2	5.0
26	9	4.8	8.1
27	9	8.2	7.7
28	10	8.1	7.7
29	10	8.2	5.0
30	10	4.8	5.2
31	11	7.7	8.2
32	11	4.8	8.1
33	11	5.0	5.2
34	12	5.2	8.1
35	12	8.2	4.8
36	12	7.7	5.0
37	13	7.7	8.2
38	13	8.1	5.2
39	13	5.0	4.8
40	14	4.8	8.2
41	14	8.1	5.2
42	14	7.7	5.0
43	15	4.8	7.7
44	15	5.2	8.2
45	15	8.1	5.0
46	16	4.8	8.2
47	16	8.1	5.2
48	16	7.7	5.0
49	17	8.2	5.2
50	17	8.1	5.0
51	17	7.7	4.8
52	18	8.2	7.7
53	18	8.1	5.2
54	18	5.0	4.8
55	19	7.7	4.8
56	19	8.1	5.0
57	19	5.2	8.2
58	20	5.2	5.0
59	20	8.2	8.1
60	20	4.8	7.7

Difference between samples:

Differences in Means for the all Permutations

Obs	_sample_	n1	mean1	n2	mean2	diff
1	1	3	7.16667	3	5.83333	1.33333
2	2	3	6.03333	3	6.96667	-0.93333
3	3	3	6.13333	3	6.86667	-0.73333
4	4	3	5.96667	3	7.03333	-1.06667
5	5	3	7.00000	3	6.00000	1.00000
6	6	3	5.83333	3	7.16667	-1.33333
7	7	3	5.00000	3	8.00000	-3.00000
8	8	3	5.90000	3	7.10000	-1.20000
9	9	3	6.06667	3	6.93333	-0.86667
10	10	3	7.03333	3	5.96667	1.06667
11	11	3	5.83333	3	7.16667	-1.33333
12	12	3	7.03333	3	5.96667	1.06667
13	13	3	6.93333	3	6.06667	0.86667
14	14	3	6.86667	3	6.13333	0.73333
15	15	3	6.03333	3	6.96667	-0.93333
16	16	3	6.86667	3	6.13333	0.73333
17	17	3	8.00000	3	5.00000	3.00000
18	18	3	7.10000	3	5.90000	1.20000
19	19	3	7.00000	3	6.00000	1.00000
20	20	3	6.06667	3	6.93333	-0.86667

(d) There are two values of difference that are as extreme as the observed value. So p-value $= \frac{2}{20} = 0.1$. Hence, we fail to reject the null that the mean difference is 0 at the significance level of 0.05.

6 Question 6

(a)
$2E(2Y_1 + 3) = 2(2E(Y_1) + 3) = 2(2\mu_1 + 3) = 4\mu_1 + 6$
$E(Y_1 + Y_2 + Y_3) = \mu_1 + \mu_2 + \mu_3$
(b)
$E(c) = c$
$Var(c) = 0$
(c)
$Var(Y_1 - Y_2) = E[(Y_1 - Y_2)^2] - (E[Y_1 - Y_2])^2 = E(Y_1^2) - 2E(Y_1Y_2) + E(Y_2^2) - E(Y_1)^2 + 2E(Y_1)E(Y_2) - E(Y_2)^2 = Var(Y_1) + Var(Y_2) - 2(E(Y_1Y_2) - E(Y_1)E(Y_2))$
Y_1 and Y_2 are independent, so $E(Y_1Y_2) = E(Y_1)E(Y_2)$. Then $Var(Y_1 - Y_2) = Var(Y_1) + Var(Y_2) = 2\sigma^2$
$Var(\frac{Y_1 + Y_2}{2}) = \frac{1}{4}Var(Y_1 + Y_2) = \frac{1}{4} \times 2\sigma^2 = \frac{\sigma^2}{2}$
(d)

$$Var(\sum_{i=1}^{n} Y_i) = E[(\sum_{i=1}^{n} Y_i)^2] - (E[\sum_{i=1}^{n} Y_i])^2$$

$$E[(\sum_{i=1}^{n} Y_i)^2] = E[\sum_{i=1}^{n}\sum_{j=1}^{n} Y_iY_j] = \sum_{i=1}^{n}\sum_{j=1}^{n} E[Y_iY_j]$$

$$(E[\sum_{i=1}^{n} Y_i])^2 = (\sum_{i=1}^{n} E[Y_i])^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} E[Y_i]E[Y_j]$$

$$Var(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n}\sum_{j=1}^{n} E[Y_iY_j] - \sum_{i=1}^{n}\sum_{j=1}^{n} E[Y_i]E[Y_j] = \sum_{i=1}^{n}\sum_{j=1}^{n} cov(Y_i, Y_j) = \sum_{i=1}^{n} cov(Y_i, Y_i) = \sum_{i=1}^{n} Var(Y_i) = n \times \sigma^2$$

(for Y_i independent for all pairs $i \neq j$)

$$Var(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n^2} Var(\sum_{i=1}^{n} Y_i) = \frac{\sigma^2}{n}$$