STAT616 - HW #1

Elliot Smith

1/25/2018

Problem 1

The answer is (ii) Drinking is associated with smoking, and alcohol causes liver cancer. The premise behind a confounding variable is, in this scenario, that drinkers tend to smoke more, and as a result, while the actual cause of liver cancer is alcohol, the act of smoking is being inappropriately labeled as the cause of liver cancer. In summary, while alcohol is the cause of liver cancer, because so many drinkers also smoke, smoking is being attributed as the cause of the liver cancer, when the alcohol is the true culprit.

Problem 2

I believe that this can be explained by the fact that the design of the experiment was relayed to the particpants. The question wording "... sometimes treated with zinc sulfate" leads me to believe that the solution is not always successful at treating the malady, otherwise it would always be used. I will identify the following groups:

- Group 1 — 1st Half - Placebo / 2nd Half - Placebo
- Group 2 — 1st Half - Placebo / 2nd Half - Zinc
- Group 3 — 1st Half - Zinc / 2nd Half - Placebo
- Group 4 — 1st Half - Zinc / 2nd Half - Zinc

Assuming that all group sizes are equal, if I am a participant in the experiment and I am given the Placebo in the first half of the study, then I know there is approximately a 50% chance that I will be given Zinc for the second half of the study. I will now further break the groups down:

- Group A — 1st Half - given Zinc Sulfate and the treatment was successful
- Group B — 1st Half - given Zinc Sulfate and the treatment was unsuccessful
- Group C — 1st Half - given the Placebo

Now, before the first half of the experiment, the participants have no information on whether they are in any of the four original groups. However, after the first half has concluded, the participants will now fall into one of the three new groups. A new element is inserted, the patients in Group A will be fairly certain that they were treated successfully, and as such, will most certainly be able to tell whether or not they are in Group 3 or Group 4 by if their symptoms return (Group 3) or if they symptoms remain treated (Group 4). However, those in Group B and Group C, which make up more than half of the total participants, having seen no improvement in symptoms will infer that they received a Placebo in the first half of the experiment. At this point, those who received a Placebo in the first half of the experiment have a 50% chance of receiving the Zinc Sulfate treatment in the second half. However, since Group B and Group C combined are greater than the total of those who received a Placebo in the first half, a greater number of patients than should believe they will be receiving the treatment. The Placebo Effect kicks in and more patients than should believe that they are now on the Zinc Treatment and will thus "feel better" when its actually the result of perception and not the treatment of Zinc Sulfate.

Problem 3

Problem a

I would choose to test the new participants at the beginning of the second year. The reason I would choose to test them during registration (as opposed to not testing them) is because the goal of the fitness program is to measure the effects it has on the participants. The assumption that the investigators are making (by questioning the necessity of testing at the beginning of the year) is that the group that will sign up for the second year will generally be the same as the group that signed up for the first year; we cannot make this assertion. For example, the first year group that signed up may have all been in poor, physical shape and thus were eager to sign up. However, those that signed up for the second year, hearing of the great improvements made by the previous cohort, may be comprised of individuals in even worse shape than the previous year. In this scenario, while the second year cohort may make the same incremental gains in their physical fitness, it would appear as though they did not perform as well as the previous year because we never measured where their level of fitness was beforehand.

Problem b

The method of analysis that I would like to perform is a 2-sample t-test. By collecting the physical fitness levels of the registrants at the beginning of the trial and at its conclusion, I would ascertain the mean fitness level from each sample and perform a t-test to conclude whether or not the difference in means is significant. This would allow me to conclude whether or not the physical fitness program has a positive effect on those that partake in it.

Problem 4

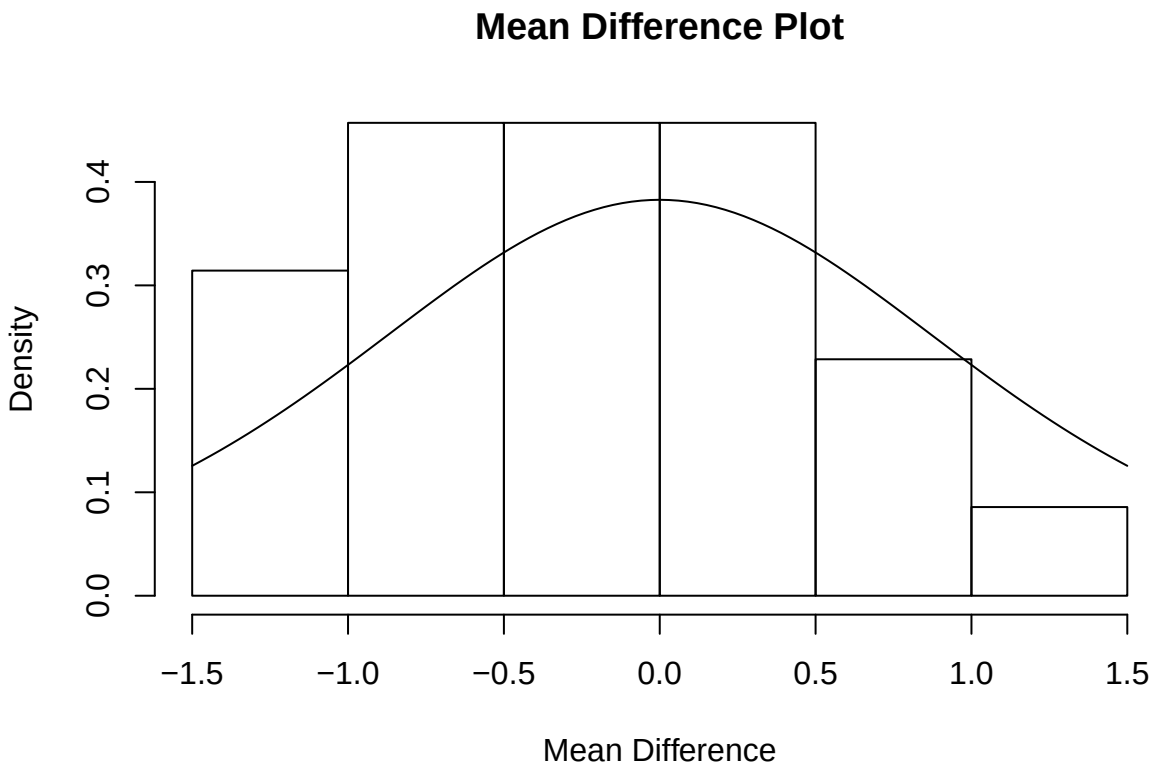
Problem a

It looks like the the groups are not homogeneous aside from the medication based on the fraction of each group that adhered to the treatment protocol. While only 53% of the Nicotinic Acid group adhered to the treatment regiment, 67% of the Placebo group adhered to the treatment; this sizeable difference implies that there is a strong probability that there are other differences between the groups aside from the treatment. For example, if a larger fraction of a group adheres to a treatment, this may imply that they are more concerned about taking care of themselves, and as a result, may have a decreased chance of death than those that are less concerned about their health.

Problem b

I do not believe that a statement can be made regarding the efficacy of the Nicotinic Acid treatment. The only valid comparison in this experiment that could be made would be the comparision between the two group total; however, in light of the above, I do not think that it is an acceptable comparison to make at this time due to the fact that, as detailed in the previous answer, the groups appear to not be homogenous. Due to the self-selection that occurred when patients decided whether or not they would adhere to the protocol, any other comparisions between the sub-groups are not relevant. In general, one should never compare to groups that differ in some systematic way other than the treatment (in this case, the adherence rate) and comparing any of the sub-groups associated with a self-seletion process is biased and not satisfactory.

Problem 5



- $\bar{y}_B - \bar{y}_A = 31 - 29.5 = 1.5$
- $n_A = 4$
- $n_B = 4$
- $s_A^2 = 0.3333$
- $s_B^2 = 0.6667$
- $s_p^2 = \frac{(n_A-1)s_A^2+(n_B-1)s_B^2}{(n_A-1)+(n_B-1)} = \frac{3(0.3333)+3(0.6667)}{6} = 0.5$
- $SE = \sqrt{s_p^2(\frac{1}{n_A} + \frac{1}{n_B})} = \sqrt{0.5(\frac{1}{2})} = \sqrt{\frac{1}{4}} = 0.5$
- $t_0 = \frac{1.5}{SE} = \frac{1.5}{0.5} = 3$
- I am clearly doing something wrong here as the answer that I am getting for the p-value is zero. I will define how am calculating my p-value and then define it. Essentially, to calculate the p-value following the permutation test, you want to compare the number of mean differences that are greater than your t-statistic against the total number of mean differences. In this way, the p-value illuminates the fraction of mean differences that are greater than your t-statistic.
- $p - value = countof \frac{(\bar{y}_B-\bar{y}_A)>t_0}{totalpermutations} = 0$
- To clarify, the p-value is calculated by, first performing the permutation, then taking the count of mean differences that are greater than our statistic and dividing by the total number of mean differences.

Problem 6

Problem a

- $s = \frac{(tn)!}{(n!)^t}$
- I will derive this using the multinomial distribution:
- $\binom{N}{n_1 n_2 ... n_k} p_1^{x_1} p_2^{x_2} ... p_k^{N-x_1-x_2-...} = \frac{N!}{n_1! n_2! ... n_k!} = \frac{(tn)!}{n! n! ... n!} = \frac{(tn)!}{(n!)^t}$

Problem b

- $s^* = \frac{s}{(t!)}$
- The reason why there are s^* different values of an arbitrary statistics C is because since there are s possible assignments (to groups), then the groups have $t!$ ways of distributing the treatments. So, for example, if we had 3 possible groups assignments and 3 treatments, we can interpret s^* as: "Once we have assigned all of our patients to one of the three groups, we now have 6 ways that we may choose which group to receive which treatment."

Problem c

- $p - value = countof \frac{s > C_{obs}}{s}$
- To clarify, the p-value is calculated by, first performing the permutation, then taking the count of mean differences that are greater than our statistic and dividing by the total number of mean differences. Also, here I have reasoned that s is the number of possible permutations, if this is not the case, know that I at least know what the numerator and denominator should be.

Problem 7

Problem a

$s = \frac{(tn)!}{(n!)^t} = \frac{(3 \cdot 5)!}{(5!)^3} = 756756$ $s^* = \frac{s}{(t!)} = \frac{756756}{(3!)} = 126126$

Problem b

Variable Definitions

For this test to evaluate the difference (or lack thereof) between means, I will use the One-Way ANOVA Test.

- Lab 1 = X
- Lab 2 = Y
- Lab 3 = Z

Hypotheses

- H_0: $\mu_x = \mu_y = \mu_z$
- H_1: $\mu_i \neq \mu_j$ for any $i \neq j$

Test Statistic

- The test statistics that I will be using is the F-statistic:
- $MS_W = s_p^2 = \frac{(n_x-1)s_x^2+(n_y-1)s_y^2+(n_x-1)s_z^2}{(n_x-1)+(n_y-1)+(n_z-1)} = \frac{0.0147+0.0153+0.037}{12} = 0.0056$
- $MS_B = \frac{1}{k-1}\sum_{i=1}^{k} n(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = s_{\bar{y}}^2 = 0.0001$
- $F = \frac{MS_B}{MS_W} = \frac{0.0001}{0.0056} = 0.0179$

Null Distribution and Test Statistic

- Unfortunately, the script that I have designed to randomize the data crashes every time I try to run it, so I am unable to perform the randomization. In hopes of gaining at least a few points, I will discuss my methodology for the randomization.
  - Firstly: I needed to figure out how many permutations of the data there could be: $\binom{15}{5}\cdot\binom{10}{5} = 756756$
  - Then to apply the randomization I applied the code in my Code Appendix section
  - Then, to find the p-value, I will take the count of values greater than my test statistic and divide by the count of total values

p-Value Interpretation

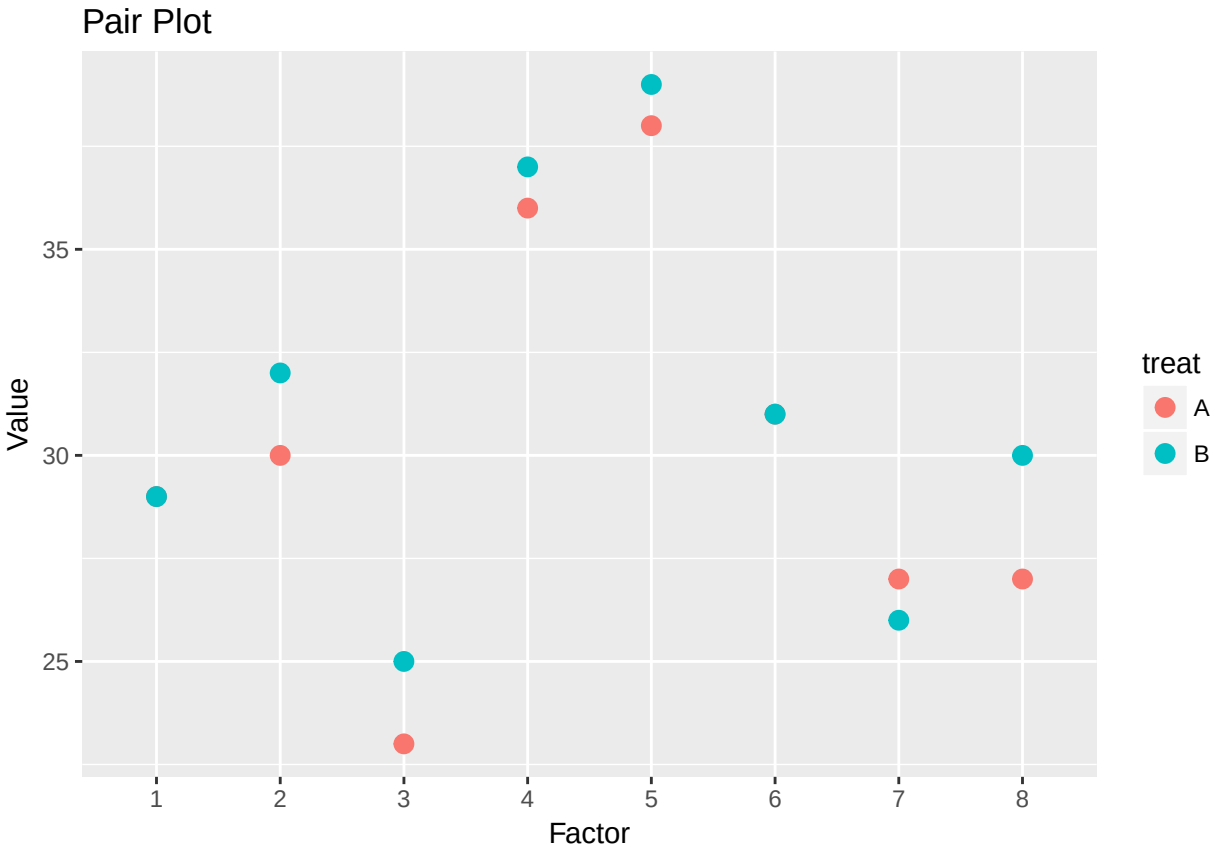
- Again, since I could not complete the analysis, I will make predictions based on the case where my p-value is a randomly selected value: 0.12. In this case, I would not reject my null hypothesis that the means of all of the groups are equal. In layman's terms, the p-value tells me that, given all of the treatments are equally effective, if we repeated this experiment 100 times, we would expect to see an observed difference in expected results or greater at least 12 times.

Problem 8

Problem a

- $s_p^2 = \frac{(n_A-1)s_A^2+(n_B-1)s_B^2}{(n_A-1)+(n_B-1)} = \frac{(7)(24.1250)+(7)(23.8393)}{14} = 23.9822$
- $SE = s_p\sqrt{\frac{1}{n_A}+\frac{1}{n_B}} = \sqrt{23.9822}\cdot\sqrt{\frac{1}{4}} = 2.4486$
- $t_0 = \frac{\bar{y}_B-\bar{y}_A}{SE(\bar{y}_B-\bar{y}_A)} = \frac{31.125-30.125}{2.4486} = 0.4084$
- $p-value = 0.6892$
- Based on this test, I conclude that I will not reject the null hypothesis that the means between Group A and Group B are equal.

Problem b



Based on this plot, there is certainly evidence supporting the blocking effect. First of all, if we look at the following factors: 2, 3, 4, 5 and 8, we can see that the are all separated by a roughly similar distance, with Treatment B being above Treatment A; this would provide evidence that the block effect is additive, because regardless of the observed value of the factor, the distance between A and B is roughly constant. The other factors (1, 6 and 7) that do not follow this paradigm are very similar; factors 1 and 6 have the same value while factor 7 has its Treatment A above Treatment B. While this goes against the typical behavior, based on this plot, it does not appear to be outside the realm of reasonable occurrence.

Problem c

I do not think that the unpaired t test is valid. My reasoning for this is that we cannot compare, for example, observed values of A across blocks because the value is directly affected by the block in which it resides. However, I do believe that a permutation test woudld help to account for the pairing, though to what extend is hard to say. My preference would be to treat the difference of the observed values of the blocks and see perform a t-test on whether or not the average difference is significantly different than zero. In this cas our null hypothesis would be that the difference is equal to 0 and we would run our typical t-test at this point.

Problem d

I was unable to prepare these results in R.

Problem 3

In order to get some credit (given that I could not answer the previous question) I will attempt to answer it generally. My general thought would be that it is preferable to use the results from the scaled t-distribution, given the fact that the randomization distribution's goal as per the Box, Hunter, Hunter text is to approvimate the scaled t-distribution. The premise being described here shows that using the scaled t is favorable when we are able to compute it. Because of my inability to perform the analysis, any final conclusion I would make on the null hypothesis would be almost purely speculation. On that note, I would infer that we would not reject the null hypothesis, the values appear too similar to me (to reiterate, this is based on very minimal information and my assertion should be paired with that fact).

Code Appendix

Problem 5

```
labels <- c("B","A","B","A","A","A","B","B")
values <- c(32,30,31,29,30,29,31,30)

num_perms <- 1
total_perms <- choose(8,4)
perm_list <- list()
perm_list[[1]] <- labels

while (num_perms != total_perms) {

    tests <- 0
    curr_perm <- sample(labels)

    for (i in 1:length(perm_list)) {

        if (!(identical(curr_perm, perm_list[[i]])) || length(perm_list) == 0) {

            tests <- tests + 1

        }

    }

    if (tests == length(perm_list)) {

        perm_list[[length(perm_list) + 1]] <- curr_perm
        num_perms <- num_perms + 1

    }

}

df_list <- list()

for (i in 1:length(perm_list)) {

    df_list[[i]] <- data.frame(label = perm_list[[i]], value = values)
```

```
}

mean_diffs <- numeric()

for (i in 1:length(df_list)) {

    df <- tapply(df_list[[i]]$value, df_list[[i]]$label, mean)
    mean_diffs[i] <- df[["B"]] - df[["A"]]

}

hist(mean_diffs, prob = T, main = "Mean Difference Plot", xlab = "Mean Difference")
curve(dt(x, df = 6), add = TRUE)
```

Problem 7

```
labels <- c("Lab1", "Lab1", "Lab1", "Lab1", "Lab1", "Lab2", "Lab2", "Lab2", "Lab2", "Lab2", "Lab3", "Lab
values <- c(4.02, 3.95, 4.02, 3.89, 3.91, 4.02, 3.86, 3.96, 3.97, 4.00, 4.00, 4.02, 4.03, 4.04, 3.81)

num_perms <- 1
total_perms <- 5000
perm_list <- list()
perm_list[[1]] <- labels

while (num_perms != total_perms) {

    tests <- 0
    curr_perm <- sample(labels)

    for (i in 1:length(perm_list)) {

        if (!(identical(curr_perm, perm_list[[i]])) || length(perm_list) == 0) {

            tests <- tests + 1

        }

    }

    if (tests == length(perm_list)) {

        perm_list[[length(perm_list) + 1]] <- curr_perm
        num_perms <- num_perms + 1

    }

}

df_list <- list()

for (i in 1:length(perm_list)) {
```

```r
    df_list[[i]] <- data.frame(label = perm_list[[i]], value = values)

}

mean_diffs <- numeric()

for (i in 1:length(df_list)) {

    df <- tapply(df_list[[i]]$value, df_list[[i]]$label, mean)
    mean_diffs[i] <- df[["Lab1"]] - df[["Lab2"]]

}

hist(mean_diffs, prob = T, main = "Mean Difference Plot", xlab = "Mean Difference")
curve(dt(x, df = 6), add = TRUE)
```