Homework Assignment #2

Math 437 - Modern Data Analysis

Due February 24, 2023

Instructions

You should submit either two or three files:

1. You should write your solutions to the Simulation and Applied Problems in this R Markdown file and submit the (.Rmd) file.
2. You should knit the final solution file to pdf and submit the pdf. If you are having trouble getting code chunks to run, add `eval = FALSE` to the chunks that do not run. If you are having trouble getting R Studio to play nice with your LaTeX distribution, I will begrudgingly accept an HTML file instead.
3. Solutions to the Key Terms and Conceptual Problems can be submitted in a separate Word or pdf file or included in the same files as your solutions to the Simulation and Applied Problems.

This homework assignment is worth a total of **40 points**.

Key Terms (5 pts)

Read Chapter 13 of Introduction to Statistical Learning, Second Edition. Based on your reading, answer the following questions.

1. What is a *p-value*? What is the difference between a one-sided and a two-sided p-value?

Answer: The p-value is the probability of observing a test statistic that is equal to or more extreme than the observed statistic under the assumption that H_0 is true. The difference between the one and two sided p-value is a two sided p-value is based on the absolute value of the test statistic, i.e. our t-score could be 1.65 or -1.65 but the one sided p-value is only looking at one of those values.

2. In traditional NHST-style significance testing, what are the two possible decisions? When do we make each decision?

Answer: The two possible decisions are to reject the null hypothesis, occurs when our p-value is less than our pre-determined alpha. The other is to fail to reject our null hypothesis which occurs when our p-value is larger than that alpha.

3. What is the difference between a *Type I Error* and a *Type II Error*?

Answer: The best case of Type I and Type II error I have learned is the story of the boy who cried wolf. At first the city makes a Type I error by rejecting H_0 (there is no wolf) when in fact there was no wolf, followed by them making a Type II error by not rejecting H_0 (again there is no wolf) when in reality there is. So a Type I Error occurs when we reject H_0 when H_0 was in fact true and a Type II Error is occurs when we do not reject H_0 when H_a was true.

4. Briefly explain why it is necessary to adjust the significance level (or equivalently, the p-values) when testing a large number of null hypotheses. Answer: The reason why we need to adjust the significance level when testing a large number of null hypotheses is to account for the chance of seeing a small p-value strictly by chance. When a large number of tests are conducted, the number of Type I Errors

we expect to make will increase so it becomes necessary to adjust the significance level to control for this.

5. Compare and contrast the *Family-Wise Error Rate* (FWER) and the *False Discovery Rate* (FDR). Answer: The FWER and FDR are similar in the sense that both are looking at the likelihood of Type I Error rates are but are very different in ideas. FWER looks at the probability of making at least one Type I Error across m tests whereas FDR is looking to minimize the proportion of type one error rates and total rejections of H_0.

6. Compare and contrast the *Bonferroni Method* and *Holm's Step-Down Method* for controlling the FWER. Answer: Bonferroni & Holm's Step-Down Method both aim to decrease (or control) the FWER based on the number of tests conducted without making assumptions about the form of H_0, the choice of test statistic, or the (in)dependence of p-values. Bonferroni's method adjusts the significance level based purely on the number of tests conducted, m, so that our new significance level is $\alpha^* = \frac{\alpha}{m}$. Holm's Step-Down Method ranks the p-values from every test in ascending order and begins comparing them beginning with the most conservative significance level of $\alpha^* = \frac{\alpha}{m}$ for the smallest p-value and if the p-value is significant, it adjusts the significance level for the subsequent p-value to be $\alpha^* = \frac{\alpha}{m-1}$ and so on until a non-significant p-value is found, resulting in a higher number of null hypotheses rejected.

7. Why do we prefer to use *Tukey's Method* or *Scheffe's Method* to control the FWER? In what conditions is it appropriate to use those methods instead of the Bonferroni or Holm methods?

Answer: We prefer to use Tukey's or Scheffe's Method because they allows us to further control the FWER while increasing power. It is more appropriate to use Tukeys methods when we have dependence between p-values caused by similar hypotheses tests so we can conduct pairwise comparisons and we can use Scheffe's to control FWER at a specific significance level α.

8. Briefly describe the *Benjamini-Hochberg* procedure for controlling the FDR. Answer: The Benjamini-Hochberg procedure begins by deciding on a specific q value to control the FDR at. Then finding and ordering our p-values such that the smallest p-value is first to the largest p-value. Next we define L as the maximum p-value for which the jth p-value is less than q(j/m). Lastly we reject all p-values for which p(j) is less than p(L).

9. What is/are the major assumption(s) of a *permutation test*? What is the general procedure for obtaining the null distribution of a test statistic using a permutation test? Answer: The major assumptions of a permutation test is that the distribution of X and Y are similar such that values can be interchanged when calculating our T-values. The process of obtaining the null distribution is through permuting n(x) + n(y) B times for a large number, B. Once that is done the distribution of B is that null distribution.

10. When is it useful/recommended to use a permutation testing approach as opposed to a traditional theory-based approach? Answer: We would want to use a permutation test when the distribution of our data does not follow a known shape. When our distribution follows that of a t-distribution, chi-squared distribution, F-distribution, etc. a traditional based approach works wonderful but when we are not following that we are inclined to go towards permutation methods.

Conceptual Problems

Conceptual Problem 1 (5 pts)

Textbook Exercise 13.7.6:

For each of the three panels in Figure 13.3, answer the following questions: (a) How many false positives, false negatives, true positives, true negatives, Type I errors, and Type II errors result from applying the Bonferroni procedure to control the FWER at level $\alpha = 0.05$?

1. Panel one has seven true positives, one false negative, and two true negatives. Thus it has One Type II Error and no Type I Errors.

2. Panel two has seven true positives, one false negative, and two true negatives. Thus it has one Type II Error and no Type I Errors.

3. Panel three has three true positives, five false negatives, and two true negatives. Thus it has five Type II Errors.

(b) How many false positives, false negatives, true positives, true negatives, Type I errors, and Type II errors result from applying the Holm procedure to control the FWER at level $\alpha = 0.05$?

1. Panel one has seven true positives, one false negative, and two true negatives. Thus it has One Type II Error and no Type I Errors.

2. Panel two has eight true positives, and two true negatives. Thus it has no Type I Errors or Type II Errors.

3. Panel two has eight true positives, and two true negatives. Thus it has no Type I Errors or Type II Errors.

(c) What is the false discovery rate associated with using the Bonferroni procedure to control the FWER at level $\alpha = 0.05$?

The False discovery rate associated with using the Bonferroni procedure to control the FWER at level $\alpha = 0.05$ is 0 for all three panels since there are no true null hypotheses being rejected.

(d) What is the false discovery rate associated with using the Holm procedure to control the FWER at level $\alpha = 0.05$?

The FDR associated with using the Holm procedure to control the FWER at level $\alpha = 0.05$ is 0 since there are no false positives.

(e) How would the answers to (a) and (c) change if we instead used the Bonferroni procedure to control the FWER at level $\alpha = 0.001$?

a) At $\alpha = 0.001$, both panel one and two would have three false negatives (Type II Errors), two true negatives, and five true positives. Panel three would have five false negatives (Type II Errors), two true negatives, and three true positives. No panel has false positives (Type I Errors).

b) The FDR would remain at 0 since there is still no false positives.

Conceptual Problem 2 (2 pts)

Suppose that we test $m = 1000$ independent null hypotheses, of which 10% are true, at significance level $\alpha = 0.05$ and achieve a false discovery rate of $q = 0.20$. Construct a table following Table 13.2 in the textbook, identifying the appropriate values of V, S, U, W, and R in this situation. Answer:

$m_0 = 100$

$V = \alpha * m_0 = 5$

$0.2 = V/(V + S)$ -> $S = 20$

	H_0 True	H_0 False	Total
Reject H_0	5	20	25
Do not reject H_0	95	880	975
Total	100	900	1000

Conceptual Problem 3 (3 pts)

Suppose that we test $m = 1000$ independent null hypotheses, of which an unknown number m_0 are true, at significance level $\alpha = 0.05$. Suppose that each test also has a power of 0.80. Find and plot the false discovery rate as a function of m_0. Answer:

Conceptual Problem 4 (2.5 pts)

Textbook Exercise 5.4.2 parts (a), (b), and (c).:

We will now derive the probability that a given observation is part of a bootstrap sample. Suppose that we obtain a bootstrap sample from a set of n observations. (a) What is the probability that the first bootstrap observation is not the jth observation from the original sample? Justify your answer. Answer: The probability of the first bootstrap observation not being the jth observation seems very likely. Thus logically the odds of any one observation being the jth is 1/n, so the odds of it not being the jth must then be 1 - 1/n.

(b) What is the probability that the second bootstrap observation is not the jth observation from the original sample? Answer: Using the same logic as the previous problem the probability of the second observation being the jth observation is also 1/n. Then the probability that it is not is also 1 - 1/n.

(c) Argue that the probability that the jth observation is not in the bootstrap sample is (1 - 1/n)n. Answer:

Simulation Problems

Simulation Problem 1 (Code: 1 pt; Explanation: 0.5 pts)

Textbook Exercise 5.4.2 parts (e), (g), and (h). For part (g), you should create a line plot (using either `plot` with argument `type = "l"` or `geom_line`). Then, to make clearer what you should be commenting on, find the limit as $n \to \infty$ of the probability that the j^{th} observation is in your bootstrap sample and add a horizontal red line (using `abline` or `geom_hline`) at that value. (Hint: the limit as $n \to \infty$ of the expression in part (c) is well-known and easily found on the Internet.)

Simulation Problem 2 (Code: 1.5 pts; Explanation: 3.5 pts)

Copy the *functions* you created in the Bootstrap Confidence Intervals class activity as well as Simulation Parts 3, 4, and 5.

Write a brief summary of what you learned from the activity. Make sure to address the following questions:

- Are theory-based methods *guaranteed* to achieve the appropriate coverage? What about bootstrap-based methods?
- Which of the four methods appear to be range-preserving even in a "worst-case scenario"?
- When and why would a bootstrap method be useful to obtain a confidence interval even if it doesn't achieve the appropriate coverage?

Applied Problems

Applied Problem 1 (Code: 4 pts; Explanation: 2 pts)

Using the `dplyr` package, subset the `mpg` dataset from the `ggplot2` package to include only the cars from 2008 that are minivans, pickups, or SUVs (`%in%` is a useful replacement for `==` when trying to match to more than one possibility). Using this new dataset, determine which of the following statements is/are true, using an $\alpha = 0.10$ significance level/family-wise error rate or a $q = 0.10$ false discovery rate:

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
filtered_data <- mpg %>%
  filter(year == 2008,
         class %in% c("minivan", "pickup", "suv"))
```

1. There is a significant difference in highway gas mileage between minivans and SUVs. H0: There is no significant difference between gas mileage between minivans and SUVs Ha: There is a significant difference between gas mileage between minivans and SUVs

```
lm1 <- aov(hwy ~ class, data = filtered_data)

anova(lm1)
```

```
## Analysis of Variance Table
##
## Response: hwy
##           Df Sum Sq Mean Sq F value   Pr(>F)
## class      2 110.15  55.075  5.4929 0.006851 **
## Residuals 52 521.38  10.026
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. There is a significant difference in highway gas mileage between pickups and SUVs.
3. There is a significant difference in highway gas mileage between minivans and pickups.

Use the following methods.

(a) Three two-sample t-tests with no adjustments for multiple testing. Store all three p-values in a single vector so that you can use the `p.adjust` function in later parts.
(b) Three two-sample t-tests followed by Bonferroni's method.
(c) Three two-sample t-tests followed by Holm's step-down method.
(d) A one-way ANOVA followed by Tukey's method.
(e) Three two-sample t-tests followed by the Benjamini-Hochberg (BH) method.

Compare and contrast your results.

Applied Problem 2 (Code: 1 pt; Explanation: 1 pt)

Use a one-way ANOVA followed by Scheffe's method (`ScheffeTest` in the DescTools package) to determine whether the following statement is true at the $\alpha = 0.10$ significance level:

There is a significant difference in highway gas mileage between pickups and non-pickups (SUVs and minivans).

Applied Problem 3 (Code: 5 pts; Explanation: 3 pts)

Textbook Exercise 5.4.9.

Conceptual problem: 2,3,4c,