Exam 1 – Take home

Steven Chiou

Due Tuesday February 19

Instructions:

- All exams are due to me in hard copy in class on Tuesday, February 19.
- Late exams will be considered late and will be penalized.
- You are not allowed to collaborate with classmates and/or people outside of this class (including on-line forum).
- You must write/type legibly and make sure your work is neat.
- Please indicate by number which questions you are answering.
- Please circle or highlight your final answer.
- Start a problem in a new line.
- Total possible points is 70.

Addendum:

- Please remove the 0's when calculating class quiz average (#4) and test average (#6) but preserve the 0's when calculating individual averages (#7, #8, #9).
- You can solve # 10 part c with 0 removed or without. If you decide to remove the 0, you should have $n_1 = 7$, $n_2 = 15$, and $\binom{7+15}{7} = 170,544$ permutations to consider.

Please read the following statement before you proceed: I understand that violation of this agreement will result in an **F** on this exam and it cannot be replaced, it will be averaged in (as a 0%) with my other scores. Violations of this agreement as it relates to the UTD Student Code of Conduct will be turned in to the Dean of Students Office.

Student name (print): _____ Signature: _____

Permutation test

The idea of a permutation test is simple. The general procedure can be summarized into the following steps:

Step 1 Compute the desired statistic based on the observed data; we will call this the observed statistic.

Step 2 Permute the data under the null.

Step 3 Compute the statistics for each possible permutation in Step 2.; we will call these permutation statistics.

Step 4 Draw conclusion based on where the observed statistic stands among the permutation statistics.

We will use the following example to illustrate an application in testing the equality of means of two independent samples.

Notations and the hypothesis test:

Suppose we have two independent samples,

$$\begin{array}{ll} \text{Group 1:} & x_1, x_2, x_3, \dots x_{n_1}, \\ \text{Group 2:} & y_1, y_2, y_3, \dots y_{n_2}, \end{array}$$

where $x_1, x_2, x_3, \dots x_{n_1}, y_1, y_2, y_3, \dots y_{n_2}$ are scalars and n_1 and n_2 are the sample sizes for the two groups. In this case, $x_1, x_2, x_3, \dots x_{n_1}$ represents a sample from a population (say Population 1), which we will assume to have a population mean of μ_1. Similarly, $y_1, y_2, y_3, \dots y_{n_2}$ represents a sample from Population 2, and we assume Population 2 has a mean of μ_2. Our goal is to use a permutation test to see if μ_1 is *significantly higher* than μ_2. The *null hypothesis* in this problem is $H_o : \mu_1 = \mu_2$ and the *alternative hypothesis* is $H_a : \mu_1 > \mu_2$.

Step 1 It is natural to compare the two sample means (averages) directly. Let the corresponding sample means be

$$\bar{x} = \frac{\sum_{i=1}^{n_1} x_i}{n_1} \text{ and } \bar{y} = \frac{\sum_{i=1}^{n_2} y_i}{n_2}.$$

We will use $d = \bar{x} - \bar{y}$ as the observed statistic.

Step 2 We assume $\mu_1 = \mu_2$ under the null hypothesis. If this is true, then $\bar{x}$ should be close to $\bar{y}$. In fact, if the null hypothesis is true, we can change the group labels and the two sample means should be close. For example,

$$\begin{array}{lllll} \textbf{Permutation \#1:} & \text{Group 1:} & y_1, x_2, x_3, \dots x_{n_1}, & \text{Group 2:} & x_1, y_2, y_3, \dots y_{n_2}, \\ \textbf{Permutation \#2:} & \text{Group 1:} & x_1, y_2, x_3, \dots x_{n_1}, & \text{Group 2:} & y_1, x_2, y_3, \dots y_{n_2}, \\ \textbf{Permutation \#3:} & \text{Group 1:} & x_1, x_2, y_3, \dots x_{n_1}, & \text{Group 2:} & y_1, y_2, x_3, \dots y_{n_2}, \\ & & \vdots & & \end{array}$$

For all possible group assignments (permutation), the two sample means should be close. We will assume the permutation $\{y_1, x_2, x_3, \dots x_{n_1}\}$ is the same as $\{x_2, y_1, x_3, \dots x_{n_1}\}$, so the order of x's and y's does not matter.

Step 3 For each permutation, we will compute the difference in sample means, $d^* = \bar{x} - \bar{y}$. We will call these d^*'s permutation statistics. We expect to have a total of $\binom{n_1+n_2}{n_1}$ possible permutation statistics.

Step 4 We will compare d with d_i^*'s. If μ_1 is significantly greater than μ_2 (in favor of H_a), d should be greater than the majority of d_i^*'s. For now, let's say μ_1 is significantly greater than μ_2 if d falls in the top 5% of d_i^*'s.

Problems:

We will perform data analysis based on a real data from the STAT2332 class that I taught in Fall 2017. In order to load the data set, please download the data set - `grade.RData` - and place it in the same folder as this `RMarkdown` document. The data set then can be loaded with the following codes:

```
> load("grade.RData")
> dim(grade)
[1] 76 17
> head(grade, 5)
  ID Q1 Q2 Q3 Q4 Q5 Q6 T1 V1 T2 V2 T3 V3  T4 V4  CT yr
1  1 70 80 80 70 70 80 88  a 88  a 92  b  92  a  92  4
2  2 70 65  0 80 70 80 76  b 96  a 92  b  80  c  80  2
3  3 50 55 70 80 80 80 72  a 88  b 84  a 100  a 103  2
4  4 70 65 80 70 80 80 64  b 72  b 88  b  76  a  76  1
5  5 70 65 80 70 80 80 64  a 72  b 88  b  72  c  84  1
```

The data set records the grades of 76 randomly selected students (thus 76 rows). The variables (columns) are

- **ID**: Student's id (marked).
- **Q1**–**Q6**: Quiz grades.
- **T1**–**T4**: Test grades.
- **V1**–**V4**: Test versions for the four tests, respectively.
- **CT**: Clicker grades. We use this as a participation grade.
- **yr**: Student's academic standing; 1 = freshman, 2 = sophomore, 3 = junior, 4 = senior.

Please use this data set to answer the following questions:

Preliminary study:

1. (5 points) How many upper-classmen (junior and senior) are there in the class?
2. (5 points) How many different test versions are there in Test #4?
3. (5 points) A grade of 0 indicates a missed assignment. Which quiz has the least attendance?
4. (5 points) After removing 0's, which quiz has the highest average? lowest average?
5. (5 points) On average, what is the number of missed quizzes per student?
6. (5 points) Calculate the test 1 average for the two versions. Repeat for tests 2 – 4.
7. (5 points) The lowest quiz grade is dropped when calculating the quiz average. Calculate the quiz average for all students and save the result in a new column with the name `Qavg` (This will increase the dimension of `grade` by one column). Print `summary(grade$Qavg)`.
8. (5 points) The lowest test grade of Tests 1 – 3 is dropped when calculating the test average (Test 4 is the final and cannot be dropped). Calculate the test average for all students and save the result in a new column with the name `Tavg` (This will increase the dimension of `grade` by one more column). Print `summary(grade$Tavg)`.
9. (5 points) The 5 quizzes (after dropping the lowest quiz grade) account for 25% of the overall course grade, the 3 tests (two of Tests 1–3 & Test 4) account for 70% of the overall course grade, and the clicker grade (`CT`) account for 5%. Calculate the final grade for all students and save the results in a new column with the name `final` (This will bring the number of columns to 20). Print `table(cut(grade$final, c(6:9*10, 0, 200)))`.

Permutation test:

10. Use a permutation test to test if the Version "a" is easier than Version "b" in Test 2, which shows the largest difference in test average between the different versions. For the ease of computing, we will focus on the *freshman subset* for this part.
    a. (5 points) Create a subset of `grade` that contains only the freshman (`yr = 1`). Name this subset `grade1`. What is the Test 2 average for this subset, under Version "a" and Version "b"?
    b. (5 points) Let Version "a" be group 1 and Version "b" be group 2, what is the observed statistics (d)?
    c. (5 points) in this subset, there are 7 students who took Version "a" ($n_1 = 7$) and 16 students who took Version "b" ($n_2 = 16$). This results in a total of $\binom{7+16}{7} = 245,157$ different permutations. Compute the permutation statistics, d^* for each permutation and save these d^*'s in a vector named `perm`. The length of this vector should be 245,157, that is, `length(perm)` should return `245,157`. Print the output of `summary(perm)`.
    d. (5 points) How many of these d^*'s are smaller than d?
    e. (5 points) Based on the 5% rule, what is your conclusion from #d?

Solution

1. There are 18 juniors and seniors.

```
> sum(grade$yr > 2)
[1] 18
```

2. There are 3 different versions in Test #4.

```
> length(unique(grade$V4))
[1] 3
```

3. Quiz #4 has the least attendance with 4 students with grade 0.

```
> quiz <- grade[,grep("Q", names(grade))]
> colSums(quiz == 0)
Q1 Q2 Q3 Q4 Q5 Q6
 1  2  2  4  3  3
```

4. Quiz #6 has the highest average while Quiz #2 has the lowest average.

```
> apply(quiz, 2, function(x) mean(x[x > 0]))
      Q1       Q2       Q3       Q4       Q5       Q6
65.00000 58.58108 60.81081 68.33333 66.98630 70.27397
```

5. The average number of missed quizzes per student is 0.1973684.

```
> mean(rowSums(quiz == 0))
[1] 0.1973684
```

6. The test averages are below:

```
> aggregate(T1 ~ V1, data = grade, function(x) mean(x[x > 0]))
  V1       T1
1  a 79.45946
2  b 78.75676
> aggregate(T2 ~ V2, data = grade, function(x) mean(x[x > 0]))
  V2       T2
1  a 87.15789
2  b 81.08108
> aggregate(T3 ~ V3, data = grade, function(x) mean(x[x > 0]))
  V3       T3
1  a 85.62162
2  b 88.34286
> aggregate(T4 ~ V4, data = grade, function(x) mean(x[x > 0]))
  V4       T4
1  a 87.53846
2  b 88.32000
3  c 85.28000
```

7. The summary is printed below:

```
> grade$Qavg <- apply(quiz, 1, function(x) mean(sort(x)[-1]))
> summary(grade$Qavg)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  36.00   63.00   68.00   66.99   71.25   80.00
```

8. The summary is printed below:

```
> test <- grade[,grep("T1|T2|T3|T4", names(grade))]
> grade$Tavg <- (rowSums(test) - apply(test[,-4], 1, min)) / 3
> summary(grade$Tavg)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  60.00   80.00   87.33   86.96   94.67  105.33
```

9. The final grade distribution is printed below:

```
> grade$final <- with(grade, Qavg * .25 / .8 + Tavg * .7 + CT * .05)
> table(cut(grade$final, c(6:9*10, 0, 200)))

  (0,60]  (60,70]  (70,80]  (80,90] (90,200]
       1        4       13       35       23
```

Permutation test

10. a. The test #2 averages can be computed as follow:

```
> grade1 <- subset(grade, yr == 1 & T2 > 0, select = c(T2, V2))
> aggregate(T2 ~ V2, data = grade1, mean)
  V2       T2
1  a 92.00000
2  b 80.26667
```

b. The observed statistic, d, is 11.7333333.

```
> (d <- diff(rev(aggregate(T2 ~ V2, data = grade1, mean)$T2)))
[1] 11.73333
```

c. The summary of the permutation statistic is

```
> grp1 <- t(combn(grade1$T2, sum(grade1$V2 == "a")))
> perm <- rowMeans(grp1) - (sum(grade1$T2) - rowSums(grp1)) / sum(grade1$V2 == "b")
> summary(perm)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-19.276  -3.352   0.000   0.000   3.352  17.600
```

Check if `grp1` has the right dimension.

```
> dim(grp1)
[1] 170544      7
> choose(nrow(grade1), sum(grade1$V2 == "a"))
[1] 170544
```

d. There are 168641 d^*'s that are smaller than d.

```
> sum(perm < d)
[1] 168641
```

e. The observed statistic d is among the top 0.0077048%. So we will reject H_0 and conclude $\mu_1 > \mu_2$.

```
> mean(perm < d)
[1] 0.9888416
```

Re-do question 10 without removing the 0:

```
> grade1 <- subset(grade, yr == 1, select = c(T2, V2))
> aggregate(T2 ~ V2, data = grade1, mean)
  V2    T2
1  a 92.00
2  b 75.25
> d <- diff(rev(aggregate(T2 ~ V2, data = grade1, mean)$T2))
> grp1 <- t(combn(grade1$T2, sum(grade1$V2 == "a")))
> perm <- rowMeans(grp1) - (sum(grade1$T2) - rowSums(grp1)) / sum(grade1$V2 == "b")
> summary(perm)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-29.250  -7.893   2.375   0.000   7.304  22.500
> sum(perm < d)
[1] 243254
> mean(perm < d)
[1] 0.9922376
```

Solution with `tidyverse`

1.

```
> grade %>% count(yr)
```

2.

```
> grade %>% count(V4)
> grade %>% select(V4) %>% n_distinct()
```

3.

```
> grade %>% select(starts_with("Q")) %>% map(~sum(.==0))
```

4.

```
> grade %>% select(starts_with("Q")) %>% map(~mean(.[.>0]))
> grade %>% select(starts_with("Q")) %>% summarize_all(~mean(.[.>0]))
```

5.

```
> grade %>% select(starts_with("Q")) %>% transmute(rowSums(. == 0)) %>% map(mean)
> grade %>% select(starts_with("Q")) %>% transmute(rowSums(. == 0)) %>% summarize_all(mean)
```

6. ...