HW1 - Class Poll Permutation

Code

2018-11-11

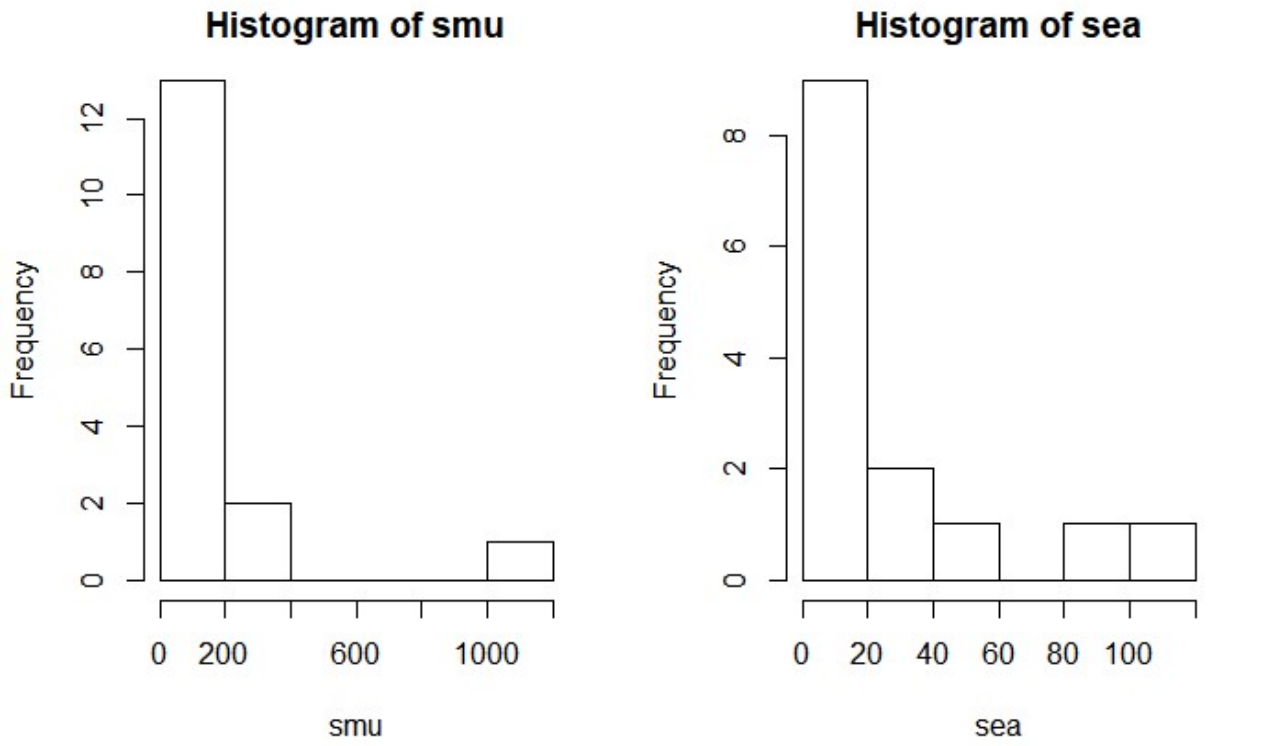
Homework 4 Questions on Class Poll Permutation test

A Business Stats class here at SMU was polled, and students were asked how much money (cash) they had in their pockets at that very moment. The idea was to see if there was evidence that those in charge of the vending machines should include the expensive bill / coin accept or or if the machines should just have the credit card reader. Also, a professor from Seattle University polled her class last year with the same question. Below are the results of the polls.

Andrew Rich—Getty Images

Hide

```
# SMU and Seattle University poll results
smu = c(34, 1200, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0)
sea = c(20, 10, 5, 0, 30, 50, 0, 100, 110, 0, 40, 10, 3, 0)
# plot histograms for distribution
par(mfrow = c(1, 2)) # split the plot
hist(smu)
hist(sea)
```



From the histogram on the provided polls from SMU and Seattle U, immediately we the distributions of both observations vary quite drastically in terms of scale. We can see SMU has 3 outliers (1200, 300, 400) that would then increase the mean in comparison to the Seattle U's population mean. The evidence that we gather to determine this observation was the spread of SMU's histogram bars in comparison to Seattle U's spread of histogram bars.

Post Permutation Test: we concluded there was not enough evidence to suggest that mean pocket amount of SMU students is equal to the mean pocket amount of Seattle U students.

Summary Statistics

Hide

```
cat("\n Summary Statistics for SMU: \n")
```

```

 Summary Statistics for SMU:
```

Hide

```
summary(smu)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    7.50   32.00  141.62   67.25 1200.00 
```

Hide

```
cat("\n Summary Statistics for Seattle University: \n")
```

```

 Summary Statistics for Seattle University: 
```

Hide

```
summary(sea)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    0.75   10.00   27.00   37.50  110.00 
```

Data Formatting

Currently we have two list of datasets, which we are tasked with performing a permutation test for our hypothesis. Prior to performing the test, lets create a dataframe for the dataset and indicate the two treatment groups `(1= SMU, 0 = Seattle)`

Hide

```
# create list of college name for our reference
college_name <- c(rep("smu", length(smu)), rep("sea", length(sea)))
# create binary dummy data to categorize the college_name string
treatment_group <- c(rep(0, length(smu)), rep(1, length(sea)))
# create list of polls results to be stored in one group
result <- c(smu, sea)
# create dataframe
poll <- data.frame(college_name, treatment_group, result)
head(poll)
```

	college_name <fctr>	treatment_group <dbl>	result <dbl>
1	smu	0	34
2	smu	0	1200
3	smu	0	23

	college_name <fctr>	treatment_group <dbl>	result <dbl>
4	smu	0	50
5	smu	0	60
6	smu	0	50

6 rows

Hide

```
# print summary statistics
summary(poll)
```

```
 college_name treatment_group      result
 sea:14       Min.   :0.0000   Min.   :   0.00
 smu:16       1st Qu.:0.0000   1st Qu.:   0.75
              Median :0.0000   Median :  21.50
              Mean   :0.4667   Mean   :  88.13
              3rd Qu.:1.0000   3rd Qu.:  50.00
              Max.   :1.0000   Max.   :1200.00
```

Permutation Test for the Difference between Two Means

Per homework assignment our objective is as follows: > Run a permutation test to test if the mean amount of pocket cash from students at SMU is different than that of students from Seattle University. Write up a statistical conclusion and scope of inference (similar to the one from the PowerPoint). (This should include identifying the Ho and Ha as well as the p-value.)

Step 1: State the Hypothesis

Null Hypothesis: Ho: mu(smu) - mu(sea) = 0 Alternative Hypothesis: Ha: mu(smu) - mu(sea) != 0

Step 2: Draw and Share and find the critical Value

alpha = 0.05

drew graph out by hand

Step 3: find the test statistic

Since we are interested in the difference in means, we will create a test statistic variable to indicate the differences in college student pocket dollar amount.

Hide

```
diff_mean <- mean(smu) - mean(sea)
diff_mean
```

```
[1] 114.625
```

Great! The interpretation of this result mean that college students have $114.63 more in their pocket if they attended SMU.

Permuting the treatment group and results

Hide

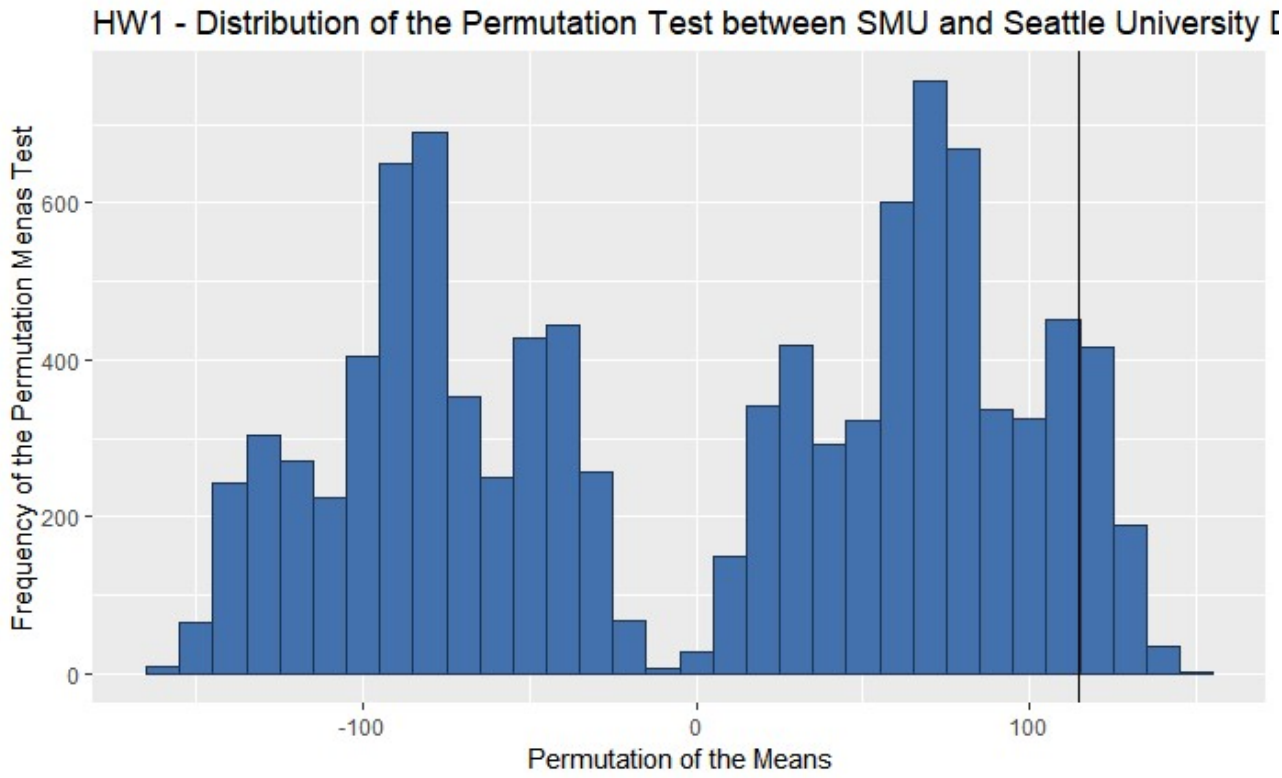
```
# set the seed to reproducability
set.seed(567)
# specify the number of permutations requred for the permutation test
nperm <- 10000
# create vector to hold average mean values during permutation test
perm_result <- numeric(0)
for(i in 1:nperm)
{
  scramble <- sample(poll$result,30); # indicate population of participants in poll
  smu_ <- scramble[1:16];             # randomly assign SMU group
  sea_ <- scramble[17:30];            # randomly assign SMU group
  diff <- mean(smu_)-mean(sea_);      # compute difference in sample means
  perm_result[i] <- diff;             # assign differnce in mean to the result vector

}
# Source: for loop code used from professors example on permutation
# .. test example for the creativity study
# .. which has been modifided for use of this study
cat("Number of values greater than the observed mean difference: ", sum(abs(perm_resul
t) > diff_mean))
```

```
Number of values greater than the observed mean difference:  1551
```

Hide

```
# create a dataframe for plotting purposes to include a title for mean differend
# ... we'll do this to indicate in the plot legend the value of the mean different tes
t statistic
observed_mean_diff_df <- data.frame(obs_mean="Observed Mean Difference", vals = diff_m
ean)
# create histogram to visulaize the distribution of the permutation test
library(ggplot2)
perm_plot <- ggplot(data=as.data.frame(perm_result), aes(perm_result)) +
        geom_histogram(binwidth = 10, fill = barfill, colour = barlines) +
        xlab("Permutation of the Means") +
        ylab("Frequency of the Permutation Menas Test") +
        ggtitle("HW1 - Distribution of the Permutation Test between SMU and Seattle Un
iversity Dollar Amount Means") +
        geom_vline(data = observed_mean_diff_df, aes(xintercept=diff_mean))
# plot the permutation histogram
perm_plot
```



The graph above provides a distribution of the permutation test performed in the section before. We can see the number of permuted mean differences from the permutation test. The `black vertical line` indicates the observed mean difference value, recall: we found interpretation of mean that college students had $114.63 more in their pocket if they attended SMU.

Step 4: Find the P-Value

Print the p-value which is the number of more extreme permutation result from the test (values > original observed difference `114.63`) over the total number of random permutations.

Hide

```
pvalue <- sum(abs(perm_result) > diff_mean) / nperm
cat("the pvalue from this permutation test is: ", pvalue)
```

```
the pvalue from this permutation test is:  0.1551
```

Of the 10,000 permutations, the significance level is simply the extreme permutation results over the total permutation test. Therefore the pvalue = 0.1551

Step 5: Reject / Fail To Reject Null Hypothesis

The result is not significant at $p < 0.05$, therefore, we **Fail to Reject the Null Hypothesis**

Step 6: conclusion

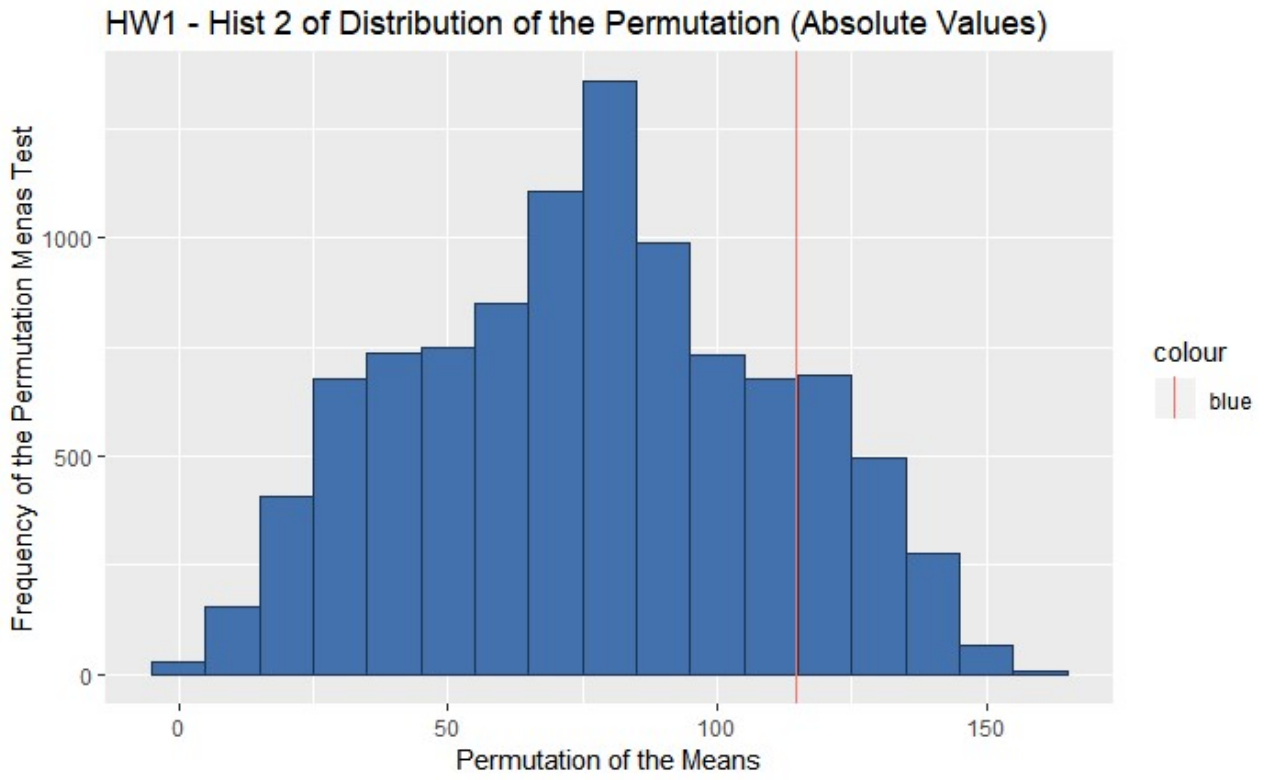
Conclusion: There is not enough evidence (p-value = 0.1551) to suggest that the mean pocket amount of SMU students is equal to the mean pocket amount of Seattle U students.

Appendix

Plot similar histogram just only accounting for absolute values.

Hide

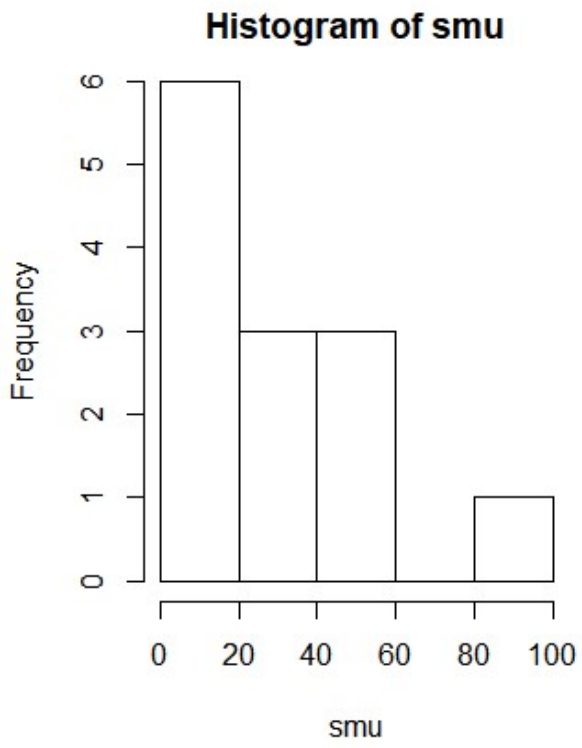
```
perm_plot_absolute <- ggplot(data=as.data.frame(abs(perm_result)), aes(abs(perm_resul
t))) +
        geom_histogram(binwidth = 10, fill = barfill, colour = barlines) +
        xlab("Permutation of the Means") +
        ylab("Frequency of the Permutation Menas Test") +
        ggtitle("HW1 - Hist 2 of Distribution of the Permutation (Absolute Values)") +
        geom_vline(data = observed_mean_diff_df, aes(xintercept=diff_mean, color = "bl
ue"))
# plot the permutation histogram (absolute)
perm_plot_absolute
```

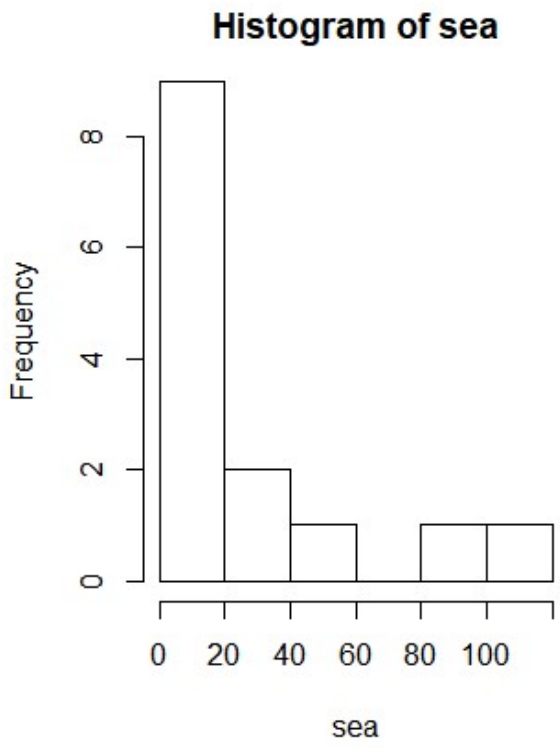


Same Test: What if We removed SMU outliers??

Hide

```
# SMU and Seattle University poll results
smu = c(34, 23, 50, 60, 50, 0, 0, 30, 89, 0, 20, 10, 0)
sea = c(20, 10, 5, 0, 30, 50, 0, 100, 110, 0, 40, 10, 3, 0)
# plot histograms for distribution
par(mfrow = c(1, 2)) # split the plot
hist(smu)
hist(sea)
```

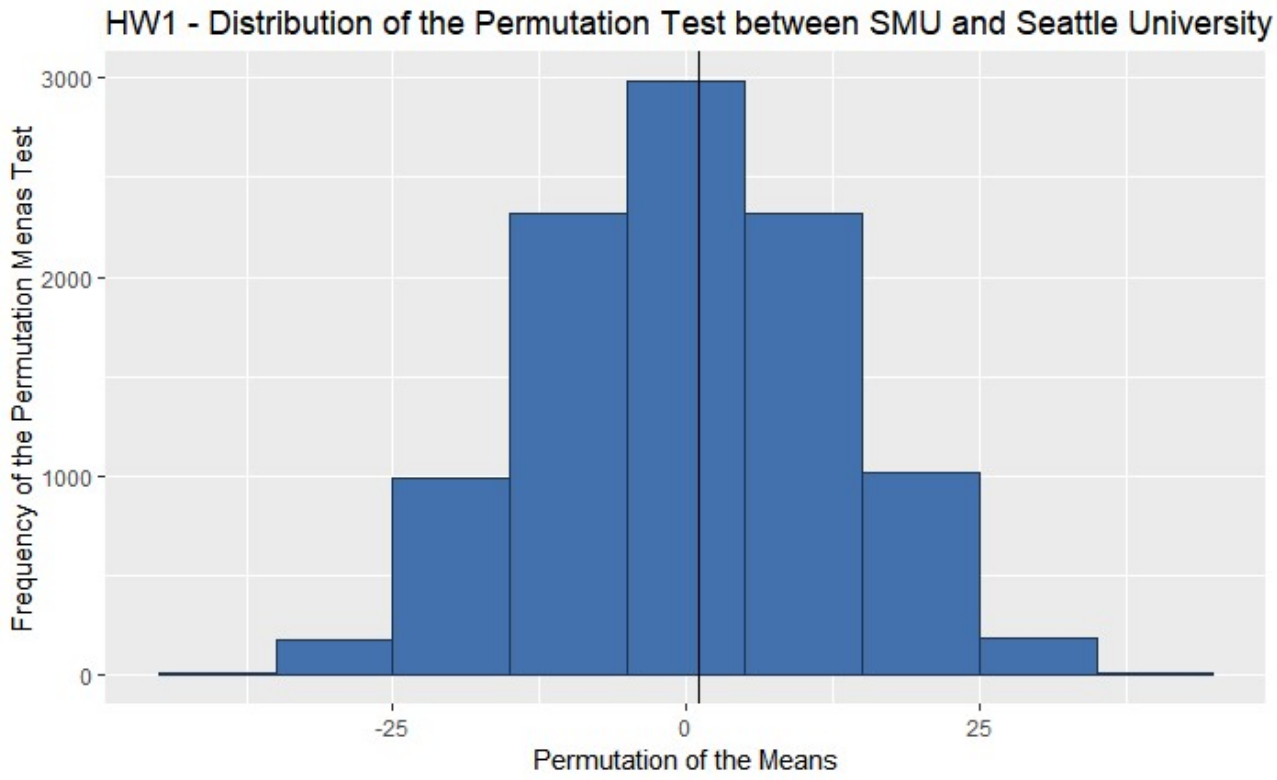

Histogram of smu
Frequency
0
1
2
3
4
5
6
0
20
40
60
80
100
smu


Histogram of sea
Frequency
0
2
4
6
8
0
20
40
60
80
100
sea

Hide

```
# create list of college name for our reference
college_name <- c(rep("smu", length(smu)), rep("sea", length(sea)))
# create binary dummy data to categorize the college_name string
treatment_group <- c(rep(0, length(smu)), rep(1, length(sea)))
# create list of polls results to be stored in one group
result <- c(smu, sea)
# create dataframe
poll <- data.frame(college_name, treatment_group, result)
diff_mean <- mean(smu) - mean(sea)
#cat("Differnce of observed mean: ",diff_mean)
# set the seed to reproducability
set.seed(567)
# specify the number of permutations requred for the permutation test
nperm <- 10000
# create vector to hold average mean values during permutation test
perm_result <- numeric(0)
for(i in 1:nperm)
{
  scramble <- sample(poll$result,27); # indicate population of participants in poll
  smu_ <- scramble[1:13];             # randomly assign SMU group
  sea_ <- scramble[14:27];            # randomly assign SMU group
  diff <- mean(smu_)-mean(sea_);      # compute difference in sample means
  perm_result[i] <- diff;             # assign differnce in mean to the result vector

}
# Source: for loop code used from professors example on permutation
# .. test example for the creativity study
# .. which has been modifided for use of this study
#cat("Number of values greater than the observed mean difference: ", sum(abs(perm_resu
lt) > diff_mean))
# create a dataframe for plotting purposes to include a title for mean differend
# ... we'll do this to indicate in the plot legend the value of the mean different tes
t statistic
observed_mean_diff_df <- data.frame(obs_mean="Observed Mean Difference", vals = diff_m
ean)
# create histogram to visulaize the distribution of the permutation test
library(ggplot2)
perm_plot <- ggplot(data=as.data.frame(perm_result), aes(perm_result)) +
        geom_histogram(binwidth = 10, fill = barfill, colour = barlines) +
        xlab("Permutation of the Means") +
        ylab("Frequency of the Permutation Menas Test") +
        ggtitle("HW1 - Distribution of the Permutation Test between SMU and Seattle Un
iversity Dollar Amount Means") +
        geom_vline(data = observed_mean_diff_df, aes(xintercept=diff_mean))
# plot the permutation histogram
perm_plot
```



Hide

```
cat("Differnce of observed mean: ",diff_mean)
```

```
Differnce of observed mean:  1.153846
```

Hide

```
cat("\n Number of values greater than the observed mean difference: ", sum(abs(perm_re
sult) > diff_mean))
```

```

 Number of values greater than the observed mean difference:  9267
```

Hide

```
pvalue <- sum(abs(perm_result) > diff_mean) / nperm
cat("\n the pvalue from this permutation test is: ", pvalue)
```

```

 the pvalue from this permutation test is:  0.9267
```