Code ▾

Unit 2 - Homework 2 with R code and Question Responses

by Phillip Hale

2018-11-16

Note to reviewer, this notebook contains the R code and analysis if the homework assignment. Please refer to the SAS homework submitted documents as appropriate. This document include: bumblebee problem (1), Samoan discrimination problem (2&4), SMU v Seattle U Problem (3)

Here is the code used for shading:

Hide

```
#' Credit: Volodymyr Orlov
#' modified by MSDS SMU
#' https://github.com/VolodymyrOrlov/MSDS6371/blob/master/shade.r
#' Draws a t-distribution curve and shades rejection regions
#'
#' @param df degrees of freedom.
#' @param alpha significance level
#' @param h0 null hypothesis value
#' @param sides one of: both, left, right
#' @param t_calc calculated test statistics
#' @examples
#' shade(49, 0.05, 0, t_calc=1.1)
#' shade(91, 0.05, 0, t_calc=NULL, sides = 'right')
#' shade(7, 0.05, 0, t_calc=1.5, sides = 'left')
#' shade(7, 0.05, 0, t_calc=1.5, sides = 'both')
shade <- function(df, alpha, h0 = 0, sides='both', t_calc=NULL) {
  e_alpha = alpha
  if(sides == 'both'){
    e_alpha = alpha / 2
  }
  cv = abs(qt(e_alpha, df))
  curve(dt(x, df), from = -4, to = 4, ylab='P(x)', xaxt='n')
  abline(v = 0, col = "black", lwd = 0.5)
  labels = h0
  at = 0
  if(sides == 'both' | sides == 'left'){
    x <- seq(-4, -abs(cv), len = 100)
    y <- dt(x, df)
    polygon(c(x, -abs(cv)), c(y, min(y)), col = "blue", border = NA)
    lines(c(-cv, -cv), c(0, dt(-cv, df)), col = "black", lwd = 1)
    text(-cv - (4 - cv) / 2, 0.05, e_alpha)
    labels = c(round(-cv, 3), labels)
    at = c(-cv, at)
  }
  if(sides == 'both' | sides == 'right'){
    x <- seq(abs(cv), 4, len = 100)
    y <- dt(x, df)
    polygon(c(abs(cv), x), c(min(y), y), col = "blue", border = NA)
    lines(c(cv, cv), c(0, dt(cv, df)), col = "black", lwd = 1)
    text(cv + (4 - cv) / 2, 0.05, e_alpha)
    labels = c(labels, round(cv, 3))
    at = c(at, cv)
  }
  if(is.numeric(t_calc)){
    abline(v = t_calc, col = "red", lwd = 2)
    text(t_calc + 0.5, 0.2, t_calc, col = "red")
  }
  axis(1, at=at, labels=labels)
}
#The above defines the function shade. To use it, you must call it. More examples are in the com
ments above.
#shade(49, 0.05, 0, t_calc=1.1)
```

Problem 1 Part B

> The world’s smallest mammal is the bumblebee bat, also known as the Kitti’s hog nosed bat. Such bats are roughly the size of a large bumblebee! Listed below are weights (in grams) from a sample of these bats. Test the claim that these bats come from the same population having a mean weight equal to 1.8 g.

Hide

```
sample = c(1.7, 1.6, 1.5, 2.0, 2.3, 1.6, 1.6, 1.8, 1.5, 1.7, 1.2, 1.4, 1.6, 1.6, 1.6)
t.test(x=sample, mu = 1.8, conf.int = "TRUE", alternative = "two.sided")
```

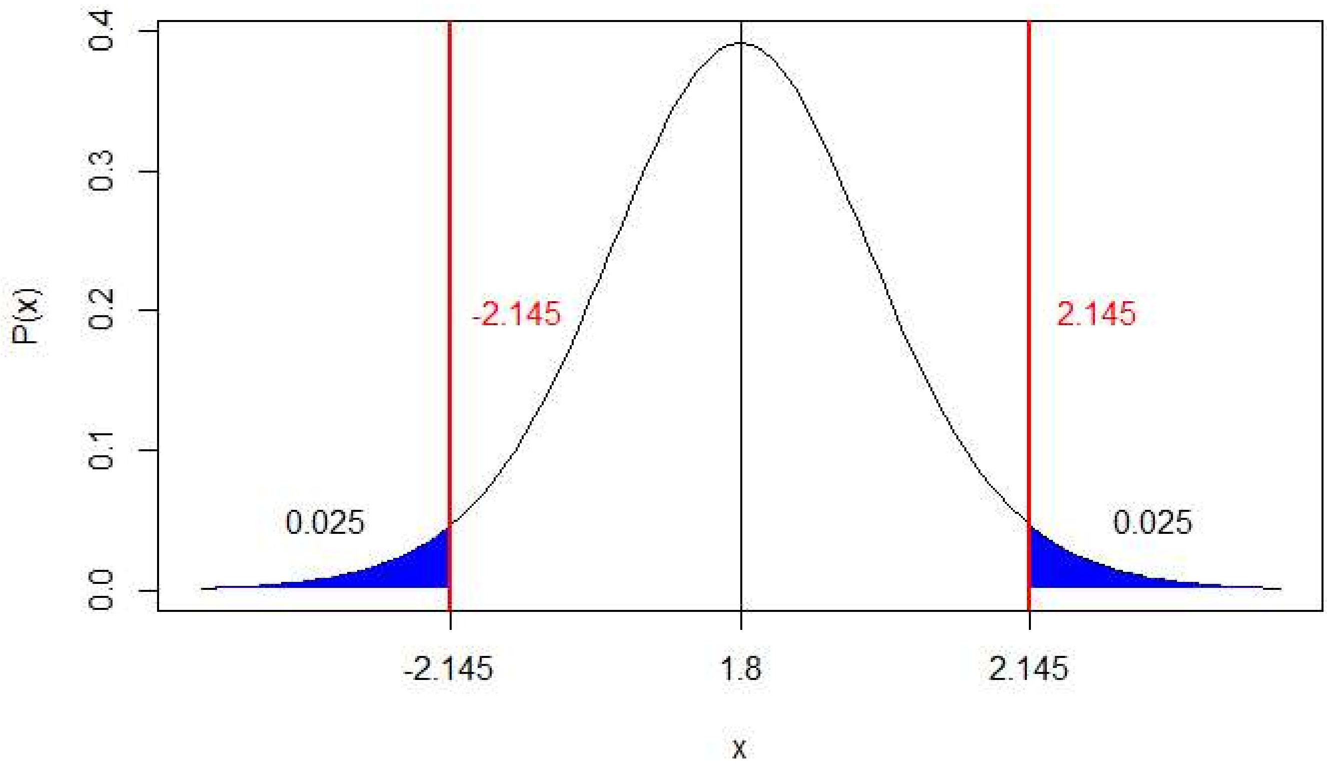
```
    One Sample t-test

data:  sample
t = -2.3457, df = 14, p-value = 0.03424
alternative hypothesis: true mean is not equal to 1.8
95 percent confidence interval:
 1.506466 1.786868
sample estimates:
mean of x 
 1.646667 
```

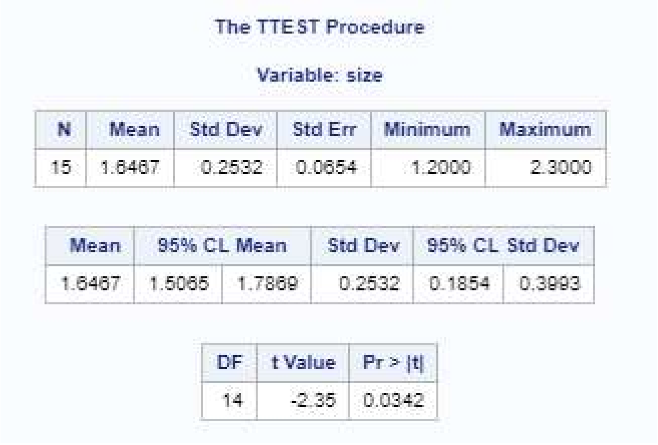
Below is the graph of the shaded region

Hide

```
#The above defines the function shade. Using function by Volodymr
shade(df = 14, alpha = 0.05, h0 = 1.8, t_calc=c(-2.145,2.145))
```



Below is a screenshot from the SAS t-test performed in SAS:



The TTEST Procedure

Variable: size

N	Mean	Std Dev	Std Err	Minimum	Maximum
15	1.6467	0.2532	0.0654	1.2000	2.3000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1.6467	1.5065	1.7869	0.2532	0.1854	0.3993

DF	t Value	Pr > |t|
14	-2.35	0.0342

The results found in this R code in comparison to the SAS code provided equivalent results. Therefore we can conclude the use of either R or SAS provides essentially the same results (R appeared to round the results related to the t-value, p-value and mean more so than SAS)

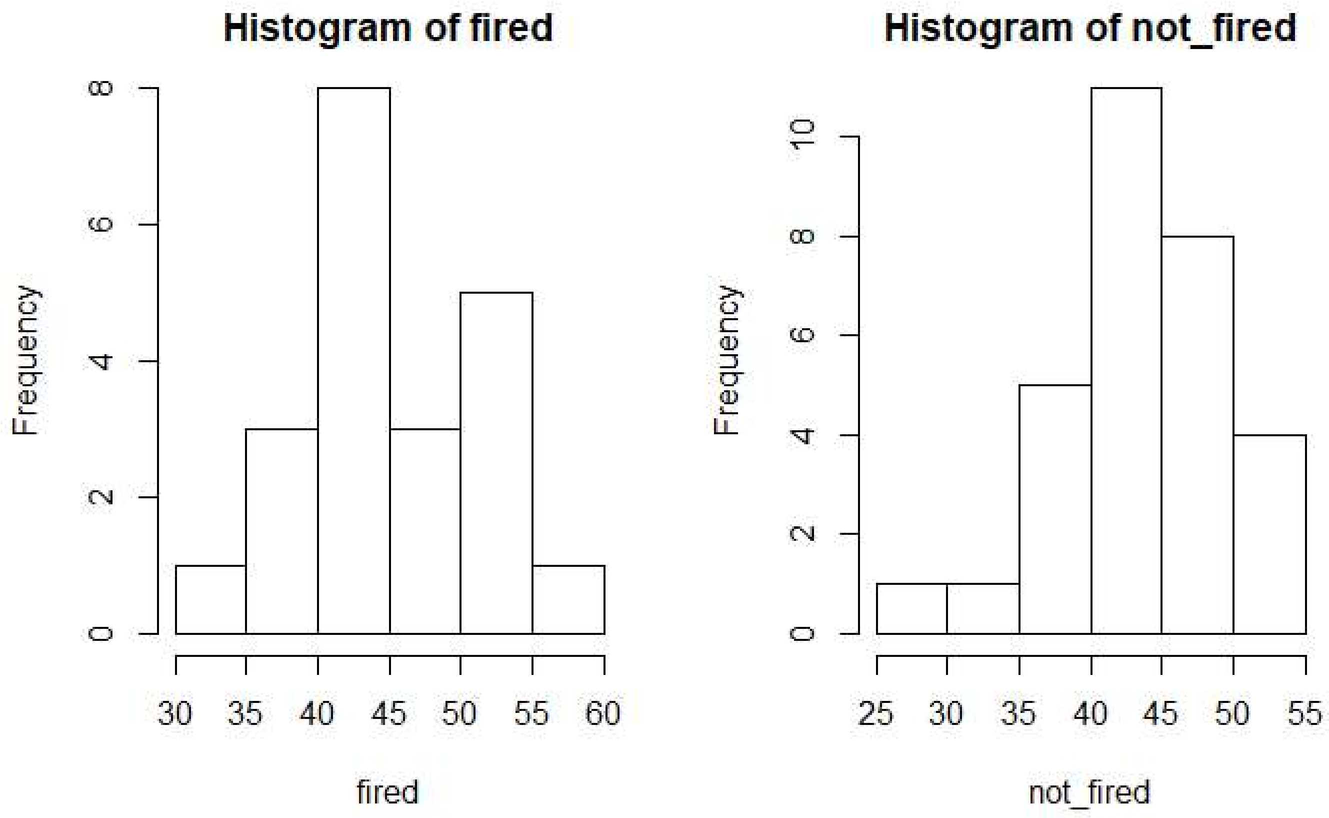
Problem 2: Age Discrimiation Test

Copyright : Ioulia Bolchakova

> In the United States, it is illegal to discriminate against people based on various attributes. One example is age. An active lawsuit, filed August 30, 2011, in the Los Angeles District Office is a case against the American Samoa Government for systematic age discrimination by preferentially firing older workers. Though the data and details are currently sealed, **suppose that a random sample of the ages of fired and not fired people in the American Samoa Government are listed below:**

Hide

```
fired = c(34, 37, 37, 38, 41, 42, 43, 44, 44, 45, 45, 45, 46, 48, 49, 53, 53, 54, 54, 55, 56)
not_fired = c(27, 33, 36, 37, 38, 38, 39, 42, 42, 43, 43, 44, 44, 44, 45, 45, 45, 45, 46, 46, 47
, 47, 48, 48, 49, 49, 51, 51, 52, 54)
# plot histograms for distribution
par(mfrow = c(1, 2)) # split the plot
hist(fired)
hist(not_fired)
```



Data Preparation and Analysis

Summary Statistics

Hide

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

```

 Summary Statistics for Fired:
```

Hide

```
summary(fired)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  34.00   42.00   45.00   45.86   53.00   56.00 
```

Hide

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

```


 Summary Statistics for Not_Fired:
```

Hide

```
summary(not_fired)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  27.00   42.00   45.00   43.93   47.75   54.00 
```

Data Formatting

Currently we have two list of data sets, which we are tasked with performing a permutation test for our hypothesis. Prior to performing the test, lets create a data frame for the data set and indicate the two treatment groups `(1= fired, 0 = not_fired)`

Hide

```
# create list of fired/not_fired result for our reference
decision <- c(rep("fired", length(fired)), rep("not_fired", length(not_fired)))
# create binary dummy data to categorize the fired/not_fired result string
treatment_group <- c(rep(0, length(fired)), rep(1, length(not_fired)))
# create list of Fired / Not_Fired decision age to be stored in one group
age <- c(fired, not_fired)
# create dataframe
data <- data.frame(decision, treatment_group, age)
head(data)
```

	decision	treatment_group	age
	<fctr>	<dbl>	<dbl>
1	fired	0	34
2	fired	0	37
3	fired	0	37
4	fired	0	38
5	fired	0	41
6	fired	0	42

6 rows

Hide

```
summary(data)
```

```
     decision  treatment_group       age
 fired    :21   Min.   :0.0000   Min.   :27.00
 not_fired:30   1st Qu.:0.0000   1st Qu.:42.00
                Median :1.0000   Median :45.00
                Mean   :0.5882   Mean   :44.73
                3rd Qu.:1.0000   3rd Qu.:48.50
                Max.   :1.0000   Max.   :56.00
```

Problem 2 Part A

> a. Perform a permutation test to test the claim that there is age discrimination.

Step 1: Identify the null H0 and alternative Ha hypothesis and scope of inference

Hypotehsis Statements: Ho: mu_fired - mu_not_fired = 0 Ha: mu_fired - mu_not_fired != 0

Scope of Inference: * Experimental Design: Observational Study as there is not chance model used in control and treatment assignment * Random Sampling: yes, random sampling was involved in the selection sampling methodology * Random Assignment: not needed, random selected population for fired/not fired * Casual Inference: causal inference relationship can be determined as randomization was involved in selecting participants.

Draw and Share and Find the Critical Value

- alpha = 0.05
- critical value (t_score) = 2.009575

Hide

```
alpha_ = 0.05
H0_ = 0 #null value
# determine degrees of freedom
df_ = length(data$decision)-2
# find the critical t score
t_score = abs(qt(p = alpha_/2, df = df_ ))
cat("Null Value : ", H0_, "\n")
```

```
Null Value :  0
```

Hide

```
cat("Significance Level : ", alpha_, "\n")
```

```
Significance Level :  0.05
```

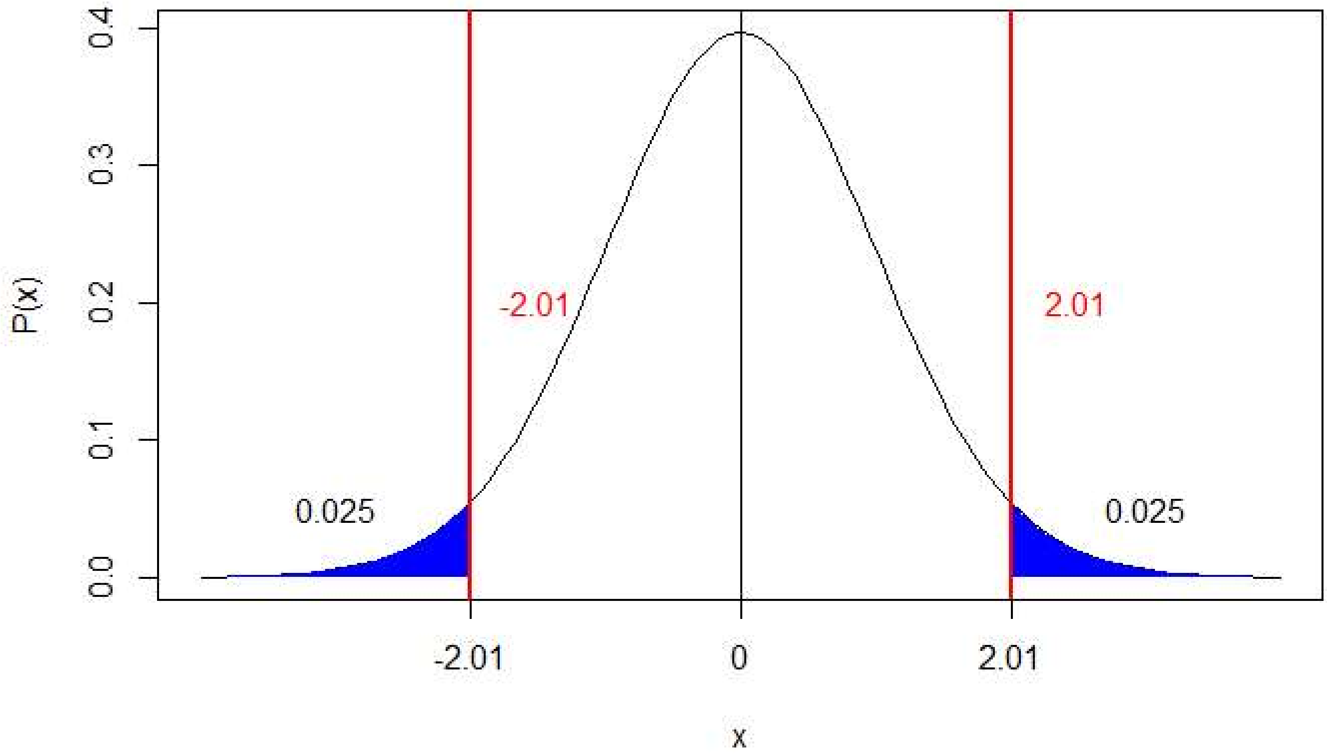
Hide

```
cat("Critical T_score: ", t_score)
```

```
Critical T_score:  2.009575
```

Hide

```
#The above defines the function shade. Using function by Volodymr
shade(df = df_, alpha = alpha_, h0 = H0_, t_calc=c(-round(t_score, 3),round(t_score, 3)))
```



Step 3: find the test statistic

Since we are interested in the difference in ages to identify age discrimination, we will create a test statistic variable to indicate the differences in yeas of age from those who were fired from those who were not fired.

Hide

```
diff_mean <- mean(fired) - mean(not_fired)
diff_mean
```

```
[1] 1.92381
```

Great! The interpretation of this result means that there is a 1.9 years of age difference between those that were fired and those employees not fired.

Permuting the treatment group and results

Hide

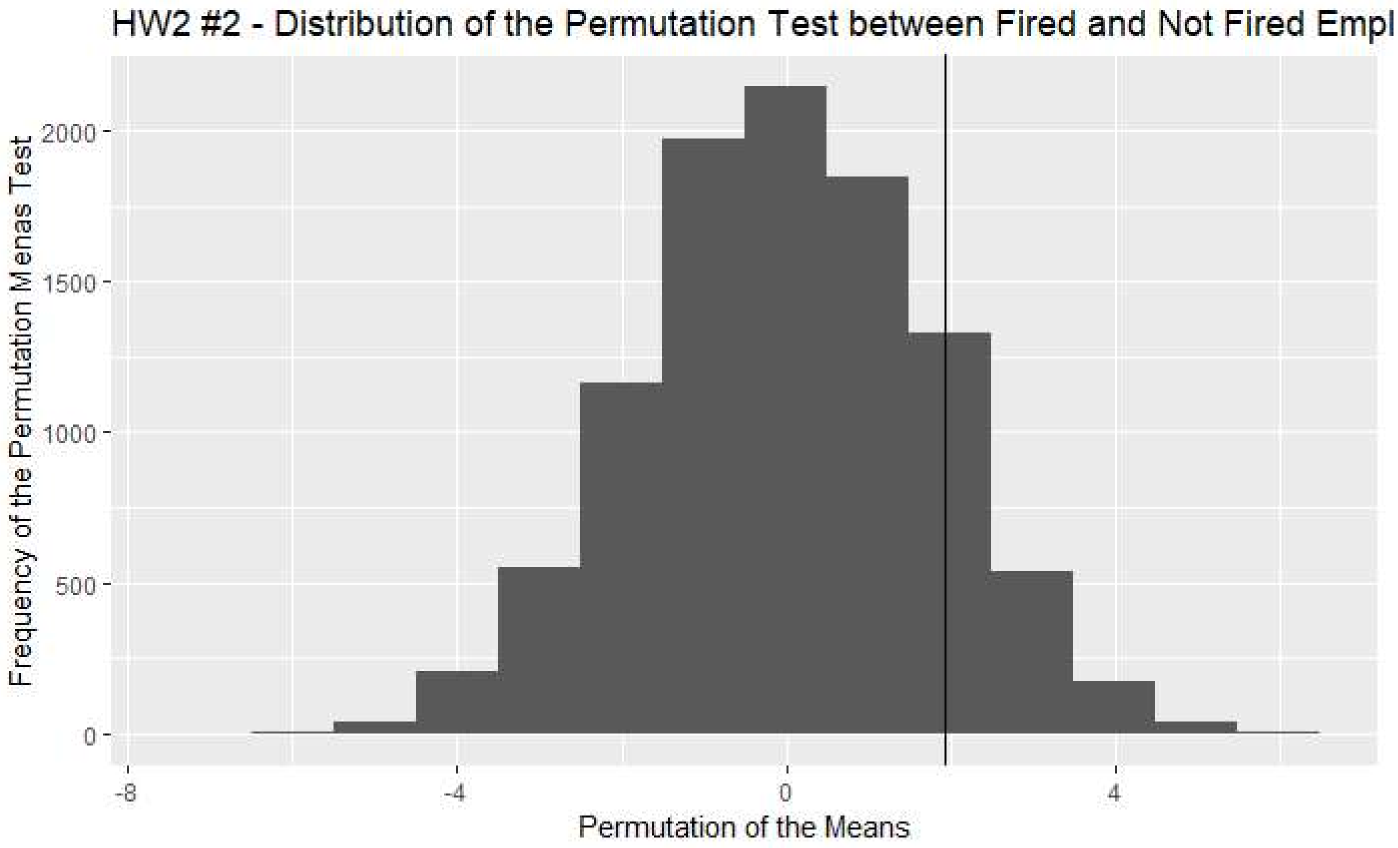
```
# set the seed to reproducability
set.seed(567)
# specify the number of permutations requred for the permutation test
nperm <- 10000
# create vector to hold average mean values during permutation test
perm_result <- numeric(0)
for(i in 1:nperm)
{
  scramble <- sample(data$age,51); # indicate population of participants in poll
  fired_sample <- scramble[1:21];             # randomly assign fired to a sample group
  not_fired_sample <- scramble[22:51];            # randomly assign not fired to a sample group
  diff <- mean(fired_sample)-mean(not_fired_sample);      # 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_result) > diff_
mean))
```

```
Number of values greater than the observed mean difference:  2764
```

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 test statisti
c
observed_mean_diff_df <- data.frame(obs_mean="Observed Mean Difference", vals = diff_mean)
# 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 = 1) +
        xlab("Permutation of the Means") +
        ylab("Frequency of the Permutation Menas Test") +
        ggtitle("HW2 #2 - Distribution of the Permutation Test between Fired and Not Fired Emplo
yees") +
        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 difference of 1.9 years of age difference between those that were fired and those employees not fired.

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 `1.92381`) 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, "\n")
```

```
the pvalue from this permutation test is:  0.2764
```

Hide

```
pvalue < alpha_
```

```
[1] FALSE
```

Of the 10,000 permutations, the significance level is simply the extreme permutation results over the total permutation test. Therefore the value = 0.2764

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: we found the p-value of approximately 27.6%, which is more than the standard significance level of 5%, therefore we fail to reject the null hypothesis, and conclude that these is not enough evidence to suggest that fired employees ages are different than those that were not fired. Additionally, the confidence interval for the average difference in age was -1.59 to 5.44 which further supports the conclusion of to fail to reject the null hypothesis as the confidence interval between the two means equal to 0 and 0 is included in the CI.

Problem 2 Part B.

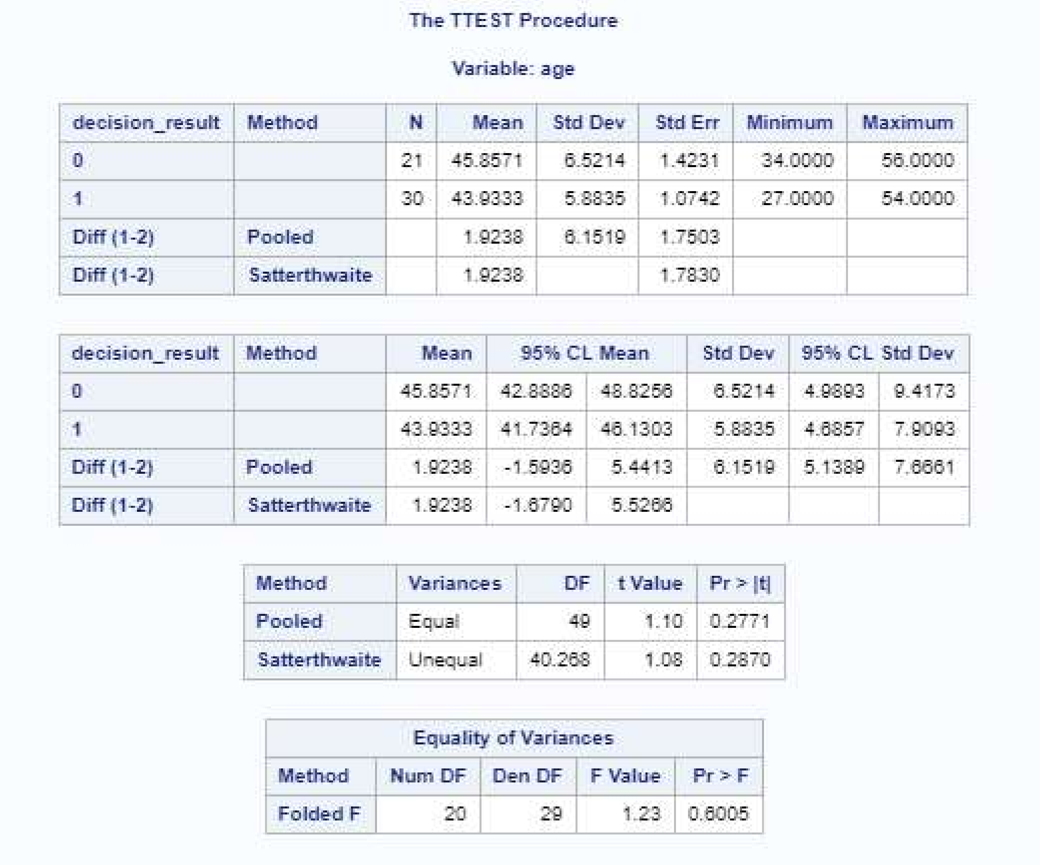
> Now run a two sample t-test appropriate for this scientific problem. (Use SAS.)

I have attached the source code for this t-test performed in SAS. Refer back to Homework 2_PH documents.

Problem 2 Part C.

> Compare this p-value to the randomized p-value found in the previous sub-question.

Below is a screenshot of the `SAS TTEST PROCEDURE` for the age discrimination test.



The TTEST Procedure

Variable: age

decision_result	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		21	45.8571	6.5214	1.4231	34.0000	56.0000
1		30	43.9333	5.8835	1.0742	27.0000	54.0000
Diff (1-2)	Pooled		1.9238	6.1519	1.7503		
Diff (1-2)	Satterthwaite		1.9238		1.7830		

decision_result	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		45.8571	42.8886	48.8256	6.5214	4.9893	9.4173
1		43.9333	41.7364	46.1303	5.8835	4.6857	7.9093
Diff (1-2)	Pooled	1.9238	-1.5936	5.4413	6.1519	5.1389	7.6661
Diff (1-2)	Satterthwaite	1.9238	-1.6790	5.5266			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	49	1.10	0.2771
Satterthwaite	Unequal	40.268	1.08	0.2870

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	20	29	1.23	0.6005

- Recall the p-value from the permutation example above was: `permudation p-value: 0.2764` , and;
- The p-value from the SAS test procedure was: `SAS ttest p-value: 0.2771`

In the permutation randomize experiment, we performed 10,000 permutations, in which the p-value from the permutation test was smaller than the SAS t-test pvalue. Perhaps one explanation in the differences is that it appears that when performing the permutation test places more reliance on the approximation on a normal distribution.

Problem 2 Part D (Confidence Interval)

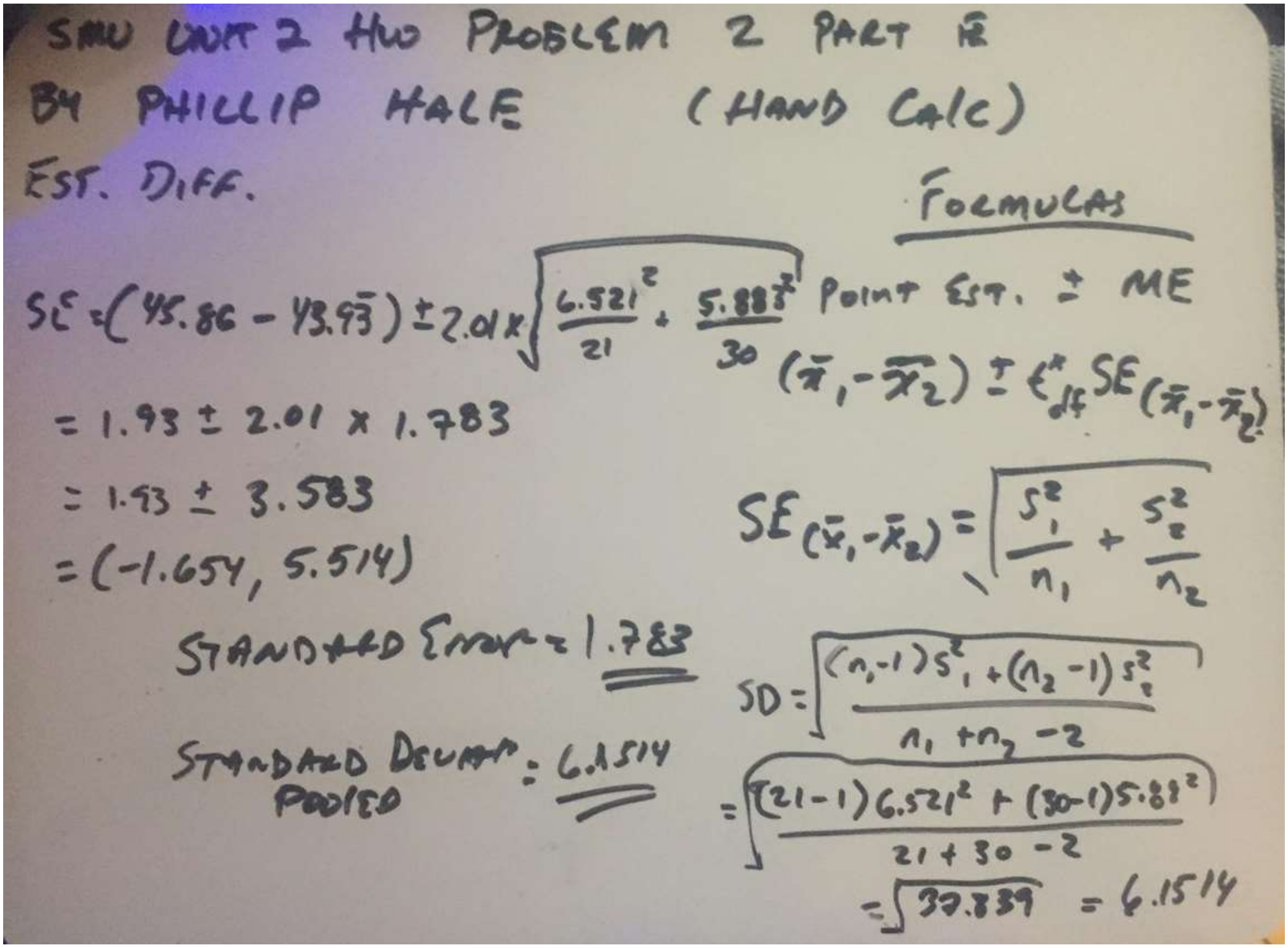
> The jury wants to see a range of plausible values for the difference in means between the fired and not fired groups. Provide them with a confidence interval for the difference of means and an interpretation.

Using the `SAS TTEST Pocedure` from above and also the `two sample t-test (R-code)` in `Problem 2 Part F (T-Test)` we are 95% confident the ages of those fired were -1.59 and 5.44 older than those who were not fired, on average, which further supports the conclusion of to fail to reject the null hypothesis as the confidence interval between the two means equal to 0 and 0 is included in the CI.

Problem 2 Part E

> Given the sample standard deviations from SAS, calculate by hand - i. Pooled standard deviation (sp) - ii. The standard error of (X¯_FIRED−X¯_(Not Fired))

See below for hand calculation:



SMU UNIT 2 HW PROBLEM 2 PART E

BY PHILLIP HALE (HAND CALC)

EST. DIFF.

$SE = (45.86 - 43.9\bar{3}) \pm 2.01 \times \sqrt{\frac{6.521^2}{21} + \frac{5.883^2}{30}}$

$= 1.93 \pm 2.01 \times 1.783$

$= 1.93 \pm 3.583$

$= (-1.654, 5.514)$

STANDARD ERROR = $\underline{\underline{1.783}}$

STANDARD DEVIATION POOLED = $\underline{\underline{6.1514}}$

FORMULAS

POINT EST. ± ME

$(\bar{x}_1 - \bar{x}_2) \pm t^*_{df} SE_{(\bar{x}_1 - \bar{x}_2)}$

$SE_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

$SD = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$

$= \sqrt{\frac{(21-1)6.521^2 + (30-1)5.883^2}{21 + 30 - 2}}$

$= \sqrt{37.839} = 6.1514$

Result:] Standard Deviation Pooled = 6.1514 Standard Error = 1.783

Problem 2 Part F (T-Test)

> Inspect and run this R Code and compare the results (t statistic, p-value, and confidence interval) to those you found in SAS. To run the code, simply copy and paste the code below into R.

Hide

```
t.test(x = fired, y = not_fired, conf.int = .95, var.equal = TRUE, alternative = "two.sided")
```

```

	Two Sample t-test

data:  fired and not_fired
t = 1.0991, df = 49, p-value = 0.2771
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.593635  5.441254
sample estimates:
mean of x mean of y 
 45.85714  43.93333 
```

The result for both SAS and R-code (above) provided relatively the same t-statistics, p-value and confidence intervals. If the functions called in both software are accurate, we would expect these scores to output the same results.

Problem 3 (SMU v Seattle University)

> In the last homework, it was mentioned that a Business Stats professor here at SMU polled his class and asked students them how much money (cash) they had in their pockets at that very moment. The idea was that we wanted to see if there was evidence that those in charge of the vending machines should include the expensive bill / coin acceptor or if it should just have the credit card reader. However, another professor from Seattle University was asked to poll her class with the same question. Below are the results of our polls.

Problem 3 Part A

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)
```

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

Hide

```
alpha_ = 0.05
H0_ = 0 #null value
# determine degrees of freedom
df_ = length(c(smu,sea))-2
# find the critical t score
t_score = abs(qt(p = alpha_/2, df = df_ ))
cat("Null Value : ", H0_, "\n")
```

```
Null Value :  0
```

Hide

```
cat("Significance Level : ", alpha_, "\n")
```

```
Significance Level :  0.05
```

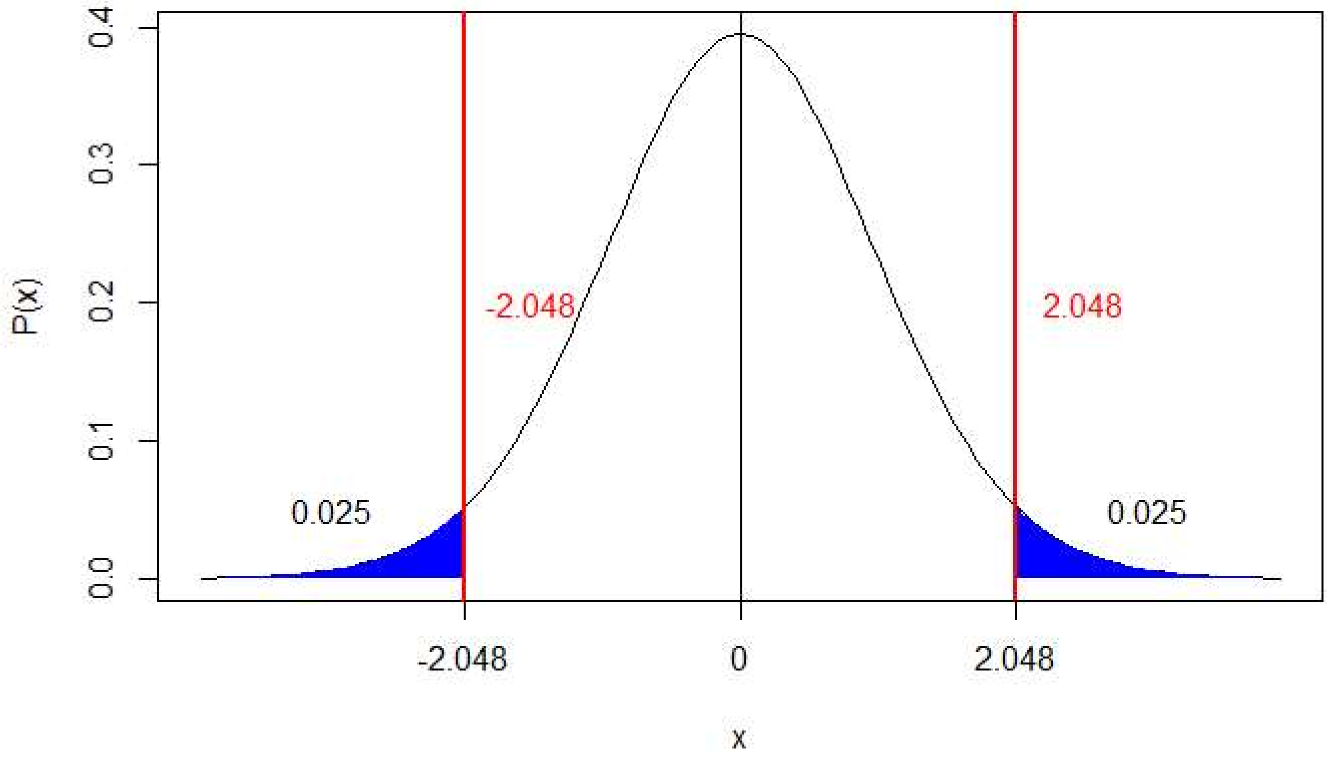
Hide

```
cat("Critical T_score: ", t_score)
```

```
Critical T_score:  2.048407
```

Hide

```
#The above defines the function shade. Using function by Volodymr
shade(df = df_, alpha = alpha_, h0 = H0_, t_calc=c(-round(t_score, 3),round(t_score, 3)))
```



Steps 3 & 4 Find the test Statistics and P-Value

Hide

```
t.test(x = smu, y = sea, conf.int = .95, var.equal = TRUE, alternative = "two.sided")
```

```

    Two Sample t-test

data:  smu and sea
t = 1.3976, df = 28, p-value = 0.1732
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -53.37112 282.62112
sample estimates:
mean of x mean of y 
  141.625    27.000 
```

Hide

```
p_value = pt(1.3976, df = df_, lower.tail = FALSE) * 2
cat("T value: ", 1.3976, "\n")
```

```
T value:  1.3976
```

Hide

```
cat("P value: ", p_value, "\n")
```

```
P value:  0.1732094
```

Hide

```
cat("Should we reject H0? (T/F): ", p_value < alpha_, " *(if 'FALSE' FTR)*")
```

```
Should we reject H0? (T/F):  FALSE  *(if 'FALSE' FTR)*
```

Analysis & Interpration: As observed from the two sample t test performed above, the test statistic is = `t = 1.3976` and the `p-value = 0.1732`.

Step 5: Make a Decision

As a result, we will `Fail to reject` the null hypothesis, the data does not provide convincing evidence of a difference between mean pocket amount of SMU students is equal to the mean pocket amount of Seattle U students.

Step 6: conclusion

Conclusion: The confidence interval for the average difference the student’s pocket amount was `$-53 to $282`. As a result of the hypothesis testing for the differences in mean resulted in a p value of about 17%, leading to the conclusion of failing to reject the null hypothesis in which there is not enough evidence (p-value = 0.1732) to suggest that the mean pocket amount of SMU students is equal to the mean pocket amount of Seattle U students.

Problem 3 Part B

> Compare the p-value from this test with the one you found from the permutation test from last week. Provide a short 2 to 3 sentence discussion on your thoughts as to why they are the same or different.

Last week (unit 1) during the permutation test for difference in mean, we found a `pvalue = 0.1551` and concluded *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*. The sample statistic in both test suggested there was some difference in mean, however performing the two sample t-test along with the 95% confidence, there was not any statistically significance with the relationship in population mean. Moreover, in last weeks homework, in the appendix section, I also removed three of the extreme outliers, where the p_values was increased quite significantly, inferring that when performing the permutation test, there is an increased likelihood of have a non-normal distribution for the sample that could effect the p-value. However, in this comparison it appears that both p-values are reasonably equivalent.

Problem 4 (See SAS Descrimination Problem)