Problem Set 2 – MATH 392

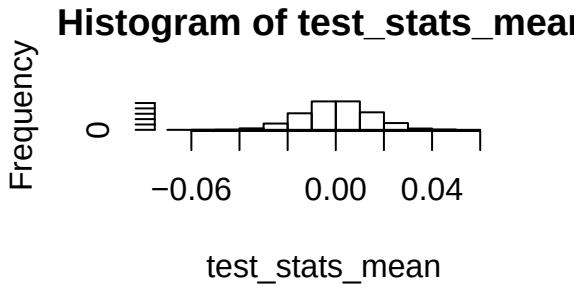
Theodore Dounias

2/1/2018

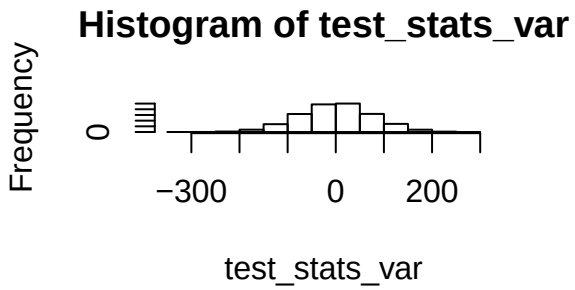
3.4

I will conduct two separate permutation tests.

```
n <- 10000
test_stats_mean <- rep(0, n)
test_stats_var <- rep(0, n)

for(i in 1:n){
  index <- sample(fdata[ ,1], replace = FALSE)
  test <- fdata %>%
    mutate(index = index) %>%
    group_by(index) %>%
    summarize(xbar = mean(Delay20), var = var(FlightLength))

  t_mean <- as.numeric(test[1, 2] - test[2, 2])
  t_var <- as.numeric(test[1, 3] - test[2, 3])

  test_stats_mean[i] <- t_mean
  test_stats_var[i] <- t_var
}

observed_mean <- fdata %>%
  group_by(Carrier) %>%
  summarize(xbar = mean(Delay20))

observed_var <- fdata %>%
  group_by(Carrier) %>%
  summarize(var = var(FlightLength))

obs_mean <- observed_mean[1, 2] - observed_mean[2, 2]

obs_var <- observed_var[1, 2] - observed_var[2, 2]

hist(test_stats_mean)
```




```
hist(test_stats_var)
```




```
p_value_mean <- 2*(sum(test_stats_mean <= obs_mean) + 1)/(n + 1)
p_value_var <- (sum(test_stats_var <= obs_var) + 1)/(n + 1)
p_value_mean
```

```
## [1] 0.00019998
```

```
p_value_var
```

```
## [1] 9.999e-05
```

Therefore in both cases we reject the null hypothesis of no difference. In the case of the variance, we can also assume that the variance in flight delay time for United is strictly larger from the one-sided test.

3.16

```
O <- table(pdata)
E <- chisq.test(O)$expected

O
```

```
##         Pres00
## Gender   Bush Didnt vote Gore Nader Other
##   Female  459          5  492    26     3
##   Male    426          5  289    31    13
```

```
chisq.test(O)
```

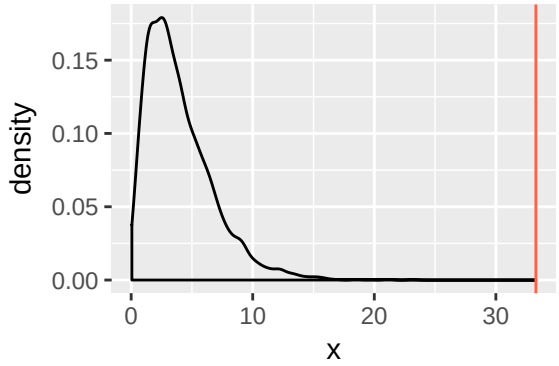
```
##
##  Pearson's Chi-squared test
##
## data:  O
## X-squared = 33.29, df = 4, p-value = 1.042e-06
```

```
obs_stat <- chisq.test(O)$stat
obs_stat
```

```
## X-squared
##  33.28993
```

```
it <- 5000
chisqs <- rep(NA, it)

for (i in 1:it) {
  perm <- sample(pdata$Gender, replace = FALSE)
  tab <- table(perm, pdata$Pres00)
  chisqs[i] <- chisq.test(tab)$stat
}
```

```
df <- data.frame(x = chisqs)
ggplot(df, aes(x = x)) +
  geom_density() +
  geom_vline(xintercept = obs_stat, col = "tomato")
```




```
p_chisq <- (sum(chisqs >= obs_stat) + 1)/(it + 1)
p_chisq
```

```
## [1] 0.00019996
```

The conclusion from part b would be that there is significant association between gender and vote, based on the p-value. The permuted test agrees by showing how, if we assume no difference, the chi-squared value we observed would be an extreme outlier.

3.22

```
qnorm((1:4)*.2, 22, 7)
```

```
## [1] 16.10865 20.22657 23.77343 27.89135
```

```
#The values that fall in each interval are 16, 13, 9, 9, 3

observed <- c(16, 13, 9, 9, 3)
expected <- c(rep(.2, 5))

chisq.test(x = observed, p = expected)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 9.6, df = 4, p-value = 0.04773
```

We would normally reject the null hypothesis at the .05 level, meaning that we would not accept the given distribution as a good fit for the data.

3.31

a) $P(T(X) \geq t) = 1 - F_{T(X)|H_o}(t)$. So let's examine what sort of distribution $F_{T(X)|H_o}(T(X))$ takes the form of. Let's assume $A = F_{T(X)|H_o}(T(X))$, then we have:

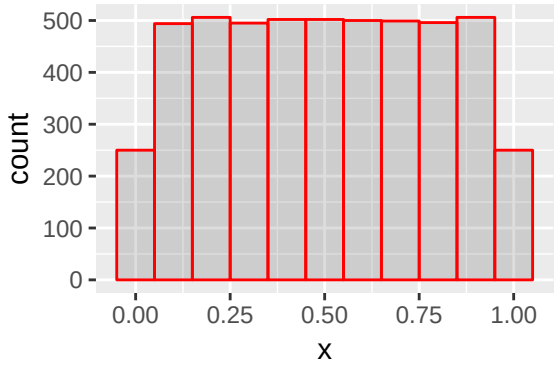
$$F_A(a) = P(A < a) = P(F_{T(X)}(T(X)) < a) = P(T(X) < F^{-1}_{T(X)}(a)) = F_{T(X)}(F^{-1}_{T(X)}(a)) = a$$

This is possible because the cdf is an invertible, strictly increasing function.

This means that A is uniformly distributed, and the p-values themselves will also be uniformly distrbuted as their pdf is the sum of a constant and a uniformly distributed rv. Since the cdf of T(X) can only take values between 1 and 0, the p-values follow a Uniform(0, 1).

b) Using the data for even and odd students, we can conduct a simulation as follows:

```
#Create Permuted T(X)
even <- grades$even
n <- 5000
test_stats <- rep(0, n)
for(i in 1:n){
  index <- sample(even, 30, replace = FALSE)
  test <- grades %>%
    mutate(index = index) %>%
    group_by(index) %>%
    summarize(xbar = mean(final))

  t <- as.numeric(test[1, 2] - test[2, 2])

  test_stats[i] <- t
  }

#Create Permuted p-values
p_value <- rep(0, n)
for(i in 1:n){
  p_value[i] <- sum(test_stats >= test_stats[i])/n
}

#Plot Histogram
data <- data.frame(x = p_value)
ggplot(data, aes()) +
  geom_histogram(aes(x = x), binwidth = .1, col = "red", alpha = .2)
```



Here we calculate each p-value assuming one of our permutations is the observed statistic. We find that the p-values are approximatelly uniformly distributed.

3.32

I will first find the cdf of X in terms of the cdf of the standard normal Z. Then, since the pdf is the derivative of the pdf, I will use this relationship to find the pdf of X, using the derivation rule for composite functions.

$$F_X(t) = P(X \leq t) = P(-\sqrt{t} \leq Z \leq \sqrt{t}) = 2F_Z(\sqrt{t})$$

Therefore:

$$f_X(t) = \frac{2}{\sqrt{(t)}} f_Z(\sqrt{t}) = \frac{1}{\sqrt{(t)}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t}$$

Which is the pdf of a chi-squared distribution with two degrees of freedom.