Homework 2 Solution

Math 445, Spring 2017

Exercise 4

Part a. The two-sided test that we are running compares the proportion of delays for the two airlines, so our hypotheses are:

$$H_0 : p_{AA} = p_{UA} \quad \text{vs.} \quad H_A : p_{AA} \neq p_{UA}$$

First, load in the data

```
library(resampledata)
data("FlightDelays")
```

and calculate the observed proportion of flights that are delayed for each airline, as well as the difference in proportions.

```
# Grab the delays from each airline
ua_delay <- subset(FlightDelays, select = Delay, subset = Carrier=="UA", drop = T)
aa_delay <- subset(FlightDelays, select = Delay, subset = Carrier=="AA", drop = T)

# Calculating the observed proportion for AA
mean(aa_delay > 20)
```

```
## [1] 0.1693049
```

```
# Calculating the observed proportion for UA
mean(ua_delay > 20)
```

```
## [1] 0.2128228
```

```
# Calculating the observed difference in proportions
observed <- mean(aa_delay > 20) - mean(ua_delay > 20)
observed
```

```
## [1] -0.04351791
```

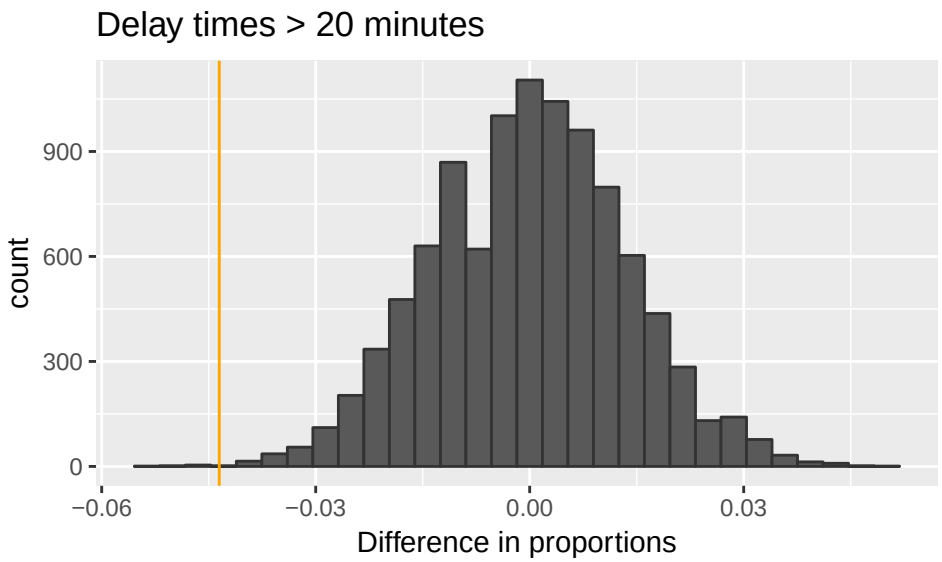
Next, we create the permutation distribution

```
# Extract the column of interest
delays <- FlightDelays$Delay

# Initialize everything
n <- length(delays)
m <- length(aa_delay)
N <- 10^4 - 1  #set number of times to repeat this process
result <- numeric(N)

# Create the permutation distributions
for(i in 1:N) {
  index <- sample(n, m, replace = FALSE)
  result[i] <- mean(delays[index]>20) - mean(delays[-index] > 20)
}
```

It's always a good idea to plot the permutation distribution (if it's not centered around 0, the value of $p_{AA} - p_{UA}$ specified by H_0, then you know there is an error in your code)

```
ggplot(data = data.frame(result)) +
  geom_histogram(mapping = aes(x = result), colour = "gray20") +
  geom_vline(xintercept = observed, colour = "orange") +
  xlab("Difference in proportions") +
  ggtitle("Delay times > 20 minutes")
```

```
## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
```



And, finally, we can calculate the two-sided p-value

```
2 * (sum(result <= observed) + 1) / (N + 1) # <= b/c of I did AA - UA
```

```
## [1] 0.0018
```

Based on an observed difference of proportions of -0.0435179 and associated p-value of 0.0018, there is very strong evidence that the proportion of delayed flights differs between American and United airlines for flights departing LaGuardia in 2009.

Part b. The two-sided test that we are running compares the variances of delays for the two airlines. A natural way to do this is to look at the ratio of the variances, so our hypotheses are:

$$H_0 : \sigma^2_{UA}/\sigma^2_{AA} = 1 \quad \text{vs.} \quad H_A : \sigma^2_{UA}/\sigma^2_{AA} > 1$$

The code that we can use is largely the same as in part a, with a few differences (such as calculating variances and ratios):

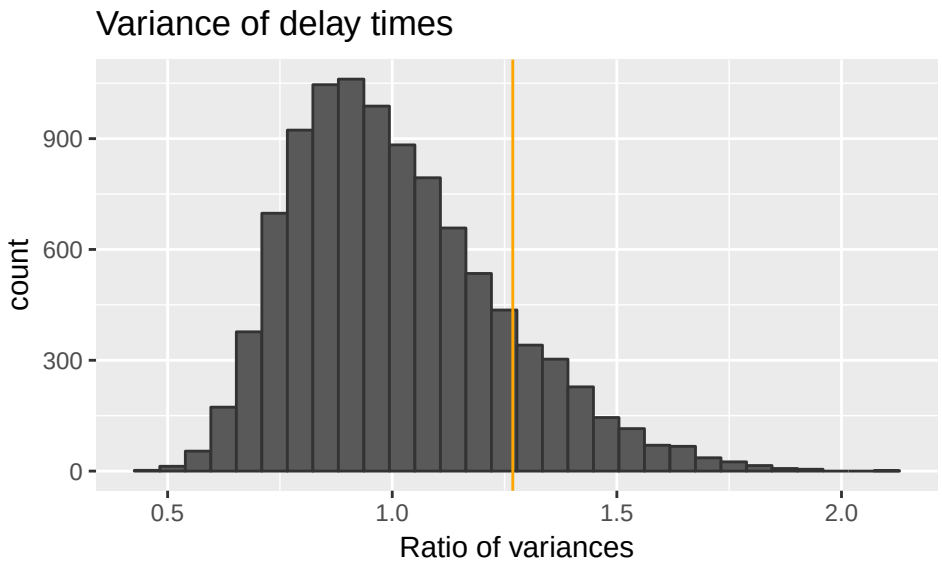
```
# Calculating the observed statistic
observedb <- var(ua_delay) / var(aa_delay)
observedb
```

```
## [1] 1.268334
```

```
# Initialize everything
n <- length(delays)
m <- length(aa_delay)
N <- 10^4 - 1  #set number of times to repeat this process
resultb <- numeric(N)

# Create the permutation distributions
for(i in 1:N) {
```

```
  index <- sample(n, m, replace = FALSE)
  resultb[i] <- var(delays[-index]) / var(delays[index]) # var(UA) / var(AA)
}

# Plot the permutation distribution
ggplot(data = data.frame(resultb)) +
  geom_histogram(mapping = aes(x = resultb), colour = "gray20") +
  geom_vline(xintercept = observedb, colour = "orange") +
  xlab("Ratio of variances") +
  ggtitle("Variance of delay times")
```

```
## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
```




```
# calculate the p-value
(sum(resultb >= observedb) + 1) / (N + 1)
```

```
## [1] 0.1424
```

Based on the permutation test with an observed ratio of variances (UA/AA) of 1.268 and an associated p-value of 0.1424, there is no evidence that the variance for United Airlines is greater than that of American Airlines.

Exercise 5

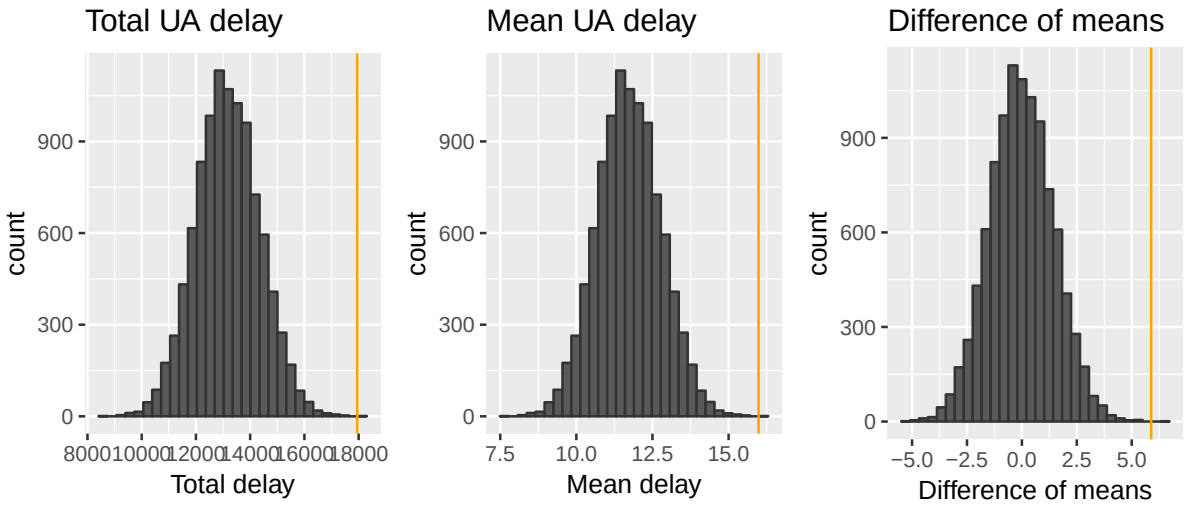
This problem investigates using different sample statistics with a permutation distribution to test whether the mean delays times differ between United and American airlines. Be sure to understand the hypotheses for each of the three tests (though I omit them in these solutions).

```
N <- 10^4 - 1

# Define delays, ua_delay, and aa_delay as in the previous problem

# Calculate the observed statistics
observed_sum_ua <- sum(ua_delay)
observed_mean_ua <- mean(ua_delay)
observed_mean_diff <- mean(ua_delay) - mean(aa_delay)

# Initialize everything for the permutations
```

```r
n <- length(delays)
m <- length(ua_delay)  #number of UA observations
sum_ua <- numeric(N)
mean_ua <- numeric(N)
mean_diff <- numeric(N)

# Run the permutations
for (i in 1:N) {
  index <- sample(n,  m, replace = FALSE)
  sum_ua[i] <- sum(delays[index])
  mean_ua[i] <- mean(delays[index])
  mean_diff[i] <- mean(delays[index]) - mean(delays[-index])
}
```

You need to be very careful when calculating the two-sided p-values here. Plotting the different permutation distributions will make sure you don't make an error by only considering the upper tail:

```r
p1 <- ggplot(data = data.frame(sum_ua)) +
  geom_histogram(mapping = aes(x = sum_ua), colour = "gray20") +
  geom_vline(xintercept = observed_sum_ua, colour = "orange") +
  xlab("Total delay") +
  ggtitle("Total UA delay")

p2 <- ggplot(data = data.frame(mean_ua)) +
  geom_histogram(mapping = aes(x = mean_ua), colour = "gray20") +
  geom_vline(xintercept = observed_mean_ua, colour = "orange") +
  xlab("Mean delay") +
  ggtitle("Mean UA delay")

p3 <- ggplot(data = data.frame(mean_diff)) +
  geom_histogram(mapping = aes(x = mean_diff), colour = "gray20") +
  geom_vline(xintercept = observed_mean_diff, colour = "orange") +
  xlab("Difference of means") +
  ggtitle("Difference of means")

library(gridExtra)
grid.arrange(p1, p2, p3, ncol = 3)
```



Now that you know where the observed stats fall with respect to the permutation distributions,

you can calculate the p-values without error:

```
# Calculate the two-sided p-values
2 * (sum(sum_ua >= observed_sum_ua) + 1) / (N + 1)

## [1] 4e-04

2 * (sum(mean_ua >= observed_mean_ua) + 1) / (N + 1)

## [1] 4e-04

2 * (sum(mean_diff >= observed_mean_diff) + 1) / (N + 1)

## [1] 4e-04
```

Notice that all of the p-values are identical in this situation. This is to be expected by Theorem 3.1 and the following remark. You should reread this section if you were surprised to help build your intuition.s

Exercise 12

Part a

The table can be made using the following R code:

```
data("Cereals")
with(Cereals, table(Age, Shelf))

##           Shelf
## Age        bottom middle top
##   adult         2      1  14
##   children      7     18   1
```

Part b

We are testing the following hypotheses:

H_0: Shelf placement is independent of the target consumer
H_A: Shelf placement is not independent of the target consumer

The test is carried out using the below commands

```
result_12b <- chisq.test(Cereals$Age, Cereals$Shelf)

## Warning in chisq.test(Cereals$Age, Cereals$Shelf):  Chi-squared approximation may be incorrect

result_12b

##
##  Pearson's Chi-squared test
##
## data:  Cereals$Age and Cereals$Shelf
## X-squared = 28.625, df = 2, p-value = 6.083e-07
```

Which results in a test statistic of $c = 28.625$ and associated p-value < 0.0001, indicating that there is strong evidence that shelf placement is not independent of target consumer.

Part c

R throws a warning message because there are cells with expected counts less than 5. If you store the results of the χ^2 test as shown above, you can retrieve the expected counts very simply, since the function returns a list

```
result_12b$expected
```

```
##            Cereals$Shelf
## Cereals$Age  bottom    middle      top
##    adult    3.55814  7.511628 5.930233
##    children 5.44186 11.488372 9.069767
```

Part d

There are two approaches to running this permutation test. The first approach is to write your own code, as shown below:

```
# Define the chisq function giveen on page 56 of the textbook
chisq <- function(obs) {
  expected <- outer(rowSums(obs), colSums(obs)) / sum(obs)
  RES <- sum((obs - expected)^2 / expected)
  return(RES)
}

# Calculate the observed counts
observed <- chisq(table(Cereals$age, Cereals$shelf))

# Initialize
N <- 10^4 - 1
result <- numeric(N)

# Permute
for (i in 1:N) {
   age.permuted <- sample(Cereals$age)
   cereal.table <- table(age.permuted, Cereals$shelf)
   result[i] <- chisq(cereal.table)
}

# Calculate the p-value
(sum(result > observed) + 1) / (N + 1)
```

```
## [1] 1e-04
```

The second approach is to use `chisq.test` and specify `simulate.p.value = TRUE`

```
chisq.test(Cereals$Age, Cereals$Shelf, simulate.p.value = TRUE, B = 10^4 - 1)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 9999
##  replicates)
##
## data:  Cereals$Age and Cereals$Shelf
## X-squared = 28.625, df = NA, p-value = 1e-04
```

Regardless of your approach, you p-value should be approximately 0.0001, providing very strong evidence that shelf placement is not independent of target consumer.

Exercise 16

Part a

```
data("GSS2002")
```

```
table(GSS2002$Gender, GSS2002$Pres00)
```

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

Part b

We are testing the following hypotheses:

H_0: a person's choice for president in the 2000 election was independent of gender
H_A: a person's choice for president in the 2000 election was not independent of gender

The test is carried out using the below commands

```
result_16b <- chisq.test(GSS2002$Gender, GSS2002$Pres00)
```

```
## Warning in chisq.test(GSS2002$Gender, GSS2002$Pres00): Chi-squared approximation may be incorrect
```

```
result_16b
```

```
##
##  Pearson's Chi-squared test
##
## data:  GSS2002$Gender and GSS2002$Pres00
## X-squared = 33.29, df = 4, p-value = 1.042e-06
```

which results in a test statistic of 33.2899273 with associated p-value of 1.0418493×10^{-6}, indicating that there is very strong evidence that gender and candidate choice were not independent in the 2000 presidential election. Notice, however, that R throws an error, since one of the expected counts is less that 5, so we should use a permutation test to be safe.

```
result_16b$expected
```

```
##                GSS2002$Pres00
## GSS2002$Gender     Bush Didnt vote     Gore   Nader    Other
##         Female 498.4134    5.63179 439.8428 32.1012 9.010863
##         Male   386.5866    4.36821 341.1572 24.8988 6.989137
```

Part c

```
# The observed test statistic
observed <- chisq(table(GSS2002$Gender, GSS2002$Pres00))

# Initializing
N <- 10^4 - 1
result <- numeric(N)

# Permuting
for (i in 1:N) {
  Pres00.permuted <- sample(GSS2002$Pres00)
  pres.table <- table(GSS2002$Gender, Pres00.permuted)
  result[i] <- chisq(pres.table)
}

# Calulcating a p-value
(sum(result > observed) + 1)/(N + 1)
```

```
## [1] 1e-04
```

From the permutation test, we find a p-value of 0.0001, so we can still conclude there is an association between gender and choice for president in the 2000 election.

Exercise 19

Part a. The value of C increases; the marginal probabilities do not change; the degrees of freedom do not change.

Part b. The p-value will be smaller, which makes it more likely you will conclude there is an association.

Exercise 20

Based on the information given in the problem, and integration to find the probabilities used to calculate the expected counts you can construct the following table:

	(0, 1.25]	(1.25, 1.75]	(1.75, 2.25]	(2.25, 2.75]	(2.75, 3]
observed counts	2	6	10	32	25
probability	0.072	0.126	0.223	0.348	0.23
expected counts	5.425	9.462	16.753	26.128	17.231

Combining this information using the χ^2 test we find an observed test statistic of 10.974, and calculate a p-value of 0.027 from a χ^2_4 distribution. There is strong enough evidence of disagreement, so it appears that the data are not drawn from this distribution.

Exercise 22

Part a

We can find the 0.2, 0.4, 0.6, and 0.8 using the below command:

```
qnorm(p = c(0.2, 0.4, 0.6, 0.8), mean = 22, sd = 7)
```

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

Part b

	$(-\infty, 16.11]$	(16.11, 20.23]	(20.23, 23.77]	(23.77, 27.89]	$(27.89, \infty]$
observed counts	16	13	9	9	3
expected counts	10	10	10	10	10

You can obtain these counts using the following R code:

```
# Enter in the data
values <- c(1.28, 4.53, 5.50, 7.91, 8.23, 9.67, 9.82, 10.28, 10.45, 11.91,
            12.57, 13.75, 13.80, 14.00, 14.05, 16.02, 16.18, 16.25, 16.58, 16.68,
            16.87, 17.61, 17.63, 17.71, 18.13, 18.42, 18.43, 18.44, 19.62, 20.401,
            20.73, 20.74, 21.29, 21.51, 21.66, 21.87, 22.67, 23.11, 24.40, 24.55,
            24.66, 25.30, 25.46, 25.91, 26.12, 26.61, 26.72, 29.28, 31.93, 36.94)

# Create a categorical variable for the intervals
bins <- cut(values, breaks = c(-Inf, qnorm(p = c(0.2, 0.4, 0.6, 0.8), mean = 22, sd = 7), Inf))

# Tabulate
```

```r
table(bins)
```

```
## bins
## (-Inf,16.1] (16.1,20.2] (20.2,23.8] (23.8,27.9] (27.9, Inf]
##          16          13           9           9           3
```

Part c

To finish the χ^2 test you can either manually calculate the value of the test statistic and obtain a p-value from a χ^2_4 distribution, or you can use the `chisq.test` function:

```r
chisq.test(c(16, 13, 9, 9, 3), p = c(.2, .2, .2, .2, .2))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(16, 13, 9, 9, 3)
## X-squared = 9.6, df = 4, p-value = 0.04773
```

Based on a p-value of 0.0477, there is weak evidence to suggest that the data do not come from a normal distribution. The evidence is weak enough that many people would not reject the null hypothesis of normality.

Exercise 26

Part a

The expected counts are found using $R_i C_j / n$:

	Disease	No Disease
Drug	3.5	6.5
Placebo	3.5	6.5

Part b

If every patient is equally likely to contract the disease, then this is a familiar counting problem from Math 240. You can solve this directly as a counting problem, or by noting that we have two sets of tags, so we're dealing with a Hypergeometric distribution.

$$\frac{\binom{10}{2}\binom{10}{5}}{\binom{20}{7}} = 0.1463$$

Part c

$$\sum_{k=0}^{2} \frac{\binom{10}{k}\binom{10}{7-k}}{\binom{20}{k}} = 0.1749$$

If the drug were not effective, then there is about a 17.5% chance of seeing an outcome this extreme or more extreme. This is not enough evidence to support the claim that the drug was effective.

Exercise 28

Part a

$X \sim Bin(N, p)$, hence $Var(X) = Np(1-p)$, and

$$Var\left(\frac{X+1}{N+1}\right) = \frac{Np(1-p)}{(N+1)^2}$$

The statement about the true P-value means that $P(T \geq t) = p$, and X is the sum of N trials with that probability.

Part b

The two-sided $\widehat{P}$ is 2 times the one-sided value; thus,

$$Var\left(2 \cdot \frac{X+1}{N+1}\right) = \frac{4Np(1-p)}{(N+1)^2}$$