Problem Set 2

David Hou

1. What happens when pilgrims attend the Hajj pilgrimage to Mecca?

On the one hand, participating in a common task with a diverse group of pilgrims might lead to increased mutual regard through processes identified in *Contact Theories*. On the other hand, media narritives have raised the spectre that this might be accompanied by "antipathy toward non-Muslims". Clingingsmith, Khwaja and Kremer (2009) investigates the question.

Using the data here, test the sharp null hypothesis that winning the visa lottery for the pilgrimage to Mecca had no effect on the views of Pakistani Muslims toward people from other countries. Assume that the Pakistani authorities assigned visas using complete random assignment. Use, as your primary outcome the `views` variable, and as your treatment feature `success`. If you're ambitious, write your fucntion generally so that you can also evaluate feeligns toward specific nationalities.

```r
rm(list = ls())
d <- data.table(read.csv("./data/Clingingsmith.2009.csv", stringsAsFactors = FALSE))
```

a. Using either `dplyr` or `data.table`, group the data by `success` and report whether views toward others are generally more positive among lottery winners or lottery non-winners.

```r
d[, .(mean_views = mean(views)), by = success]
```

```
##    success mean_views
## 1:       0   1.868304
## 2:       1   2.343137
```

```r
ate_actual = mean(d[success == 1, views]) - mean(d[success == 0, views])
```

Mean views are higher for lottery winners.

b. But is this a meaningful difference, or could it just be randomization noise? Conduct 10,000 simulated random assignments under the sharp null hypothesis to find out. (Don't just copy the code from the async, think about how to write this yourself.)

```r
N = nrow(d)
simulate = function() {
    # throw an extra person in treatment if N is odd
    treatment = sample(c(rep(0, N%/%2), rep(1, N%/%2 + N %% 2)))

    # assuming sharp null, d$views is the same for treatment and control
    # the ate is simply the difference in means of d$view indexed by our simulated treatment vector
    mean(d$views[treatment == 1]) - mean(d$views[treatment == 0])
}

ates_simulated = replicate(10000, simulate())
```

c. How many of the simulated random assignments generate an estimated ATE that is at least as large as the actual estimate of the ATE?

```r
sum(ates_simulated > ate_actual)
```

```
## [1] 24
```

d. What is the implied *one-tailed* p-value?

```
mean(ates_simulated > ate_actual)
```

```
## [1] 0.0024
```

e. How many of the simulated random assignments generate an estimated ATE that is at least as large *in absolute value* as the actual estimate of the ATE?

```
sum(abs(ates_simulated) > ate_actual)
```

```
## [1] 41
```

f. What is the implied two-tailed p-value?

```
mean(abs(ates_simulated) > ate_actual)
```

```
## [1] 0.0041
```

2. Term Limits Aren't Good.

Naturally occurring experiments sometimes involve what is, in effect, block random assignment. For example, Rocio Titiunik , in this paper studies the effect of lotteries that determine whether state senators in TX and AR serve two-year or four-year terms in the aftermath of decennial redistricting. These lotteries are conducted within each state, and so there are effectively two distinct experiments on the effects of term length.

The "thoery" in the news (such as it is), is that legislators who serve 4 year terms have more time to slack off and not produce legislation. If this were true, then it would stand to reason that making terms shorter would increase legislative production.

One way to measure legislative production is to count the number of bills (legislative proposals) that each senator introduces during a legislative session. The table below lists the number of bills introduced by senators in both states during 2003.

```
library(foreign)

rm(list = ls()) # clear the stuff from the last problem
d <- read.dta("./data/Titiunik.2010.dta")
head(d)
```

```
##   term2year bills_introduced texas0_arkansas1
## 1         0               18                0
## 2         0               29                0
## 3         0               41                0
## 4         0               53                0
## 5         0               60                0
## 6         0               67                0
```

a. Using either `dplyr` or `data.table`, group the data by state and report the mean number of bills introduced in each state. Does Texas or Arkansas seem to be more productive? Then, group by two- or four-year terms (ignoring states). Do two- or four-year terms seem to be more productive? **Which of these effects is causal, and which is not?** Finally, using `dplyr` or `data.table` to group by state and term-length. How, if at all, does this change what you learn?

```
# Let's do dplyr for this one.
d = as_tibble(d)
```

```r
d %>%
    group_by(texas0_arkansas1) %>%
    summarise(mean(bills_introduced))
```

```
## # A tibble: 2 x 2
##   texas0_arkansas1 `mean(bills_introduced)`
##              <int>                    <dbl>
## 1                0                     68.8
## 2                1                     25.5
```

Texas seems more productive.

```r
d %>%
    group_by(term2year) %>%
    summarise(mean(bills_introduced))
```

```
## # A tibble: 2 x 2
##   term2year `mean(bills_introduced)`
##       <int>                    <dbl>
## 1         0                     53.1
## 2         1                     38.6
```

4-year terms seem more productive.

```r
d %>%
    group_by(term2year, texas0_arkansas1) %>%
    summarise(mean(bills_introduced))
```

```
## # A tibble: 4 x 3
## # Groups:   term2year [?]
##   term2year texas0_arkansas1 `mean(bills_introduced)`
##       <int>            <int>                    <dbl>
## 1         0                0                     76.9
## 2         0                1                     30.7
## 3         1                0                     60.1
## 4         1                1                     20.6
```

State is a stronger predictor of introduced bills than term length. It makes sense to block on state. The data does not actually allow us to say which variable is the causal one. We do not have enough qualitative information to say.

b. For each state, estimate the standard error of the estimated ATE.

```r
# create summary tibble for use later
(d2 = d %>%
    group_by(term2year, texas0_arkansas1) %>%
    summarise(mean(bills_introduced), var(bills_introduced), n()))
```

```
## # A tibble: 4 x 5
## # Groups:   term2year [?]
##   term2year texas0_arkansas1 `mean(bills_introdu~ `var(bills_introd~ `n()`
##       <int>            <int>                <dbl>              <dbl> <int>
## 1         0                0                 76.9              956.     16
## 2         0                1                 30.7              149.     17
## 3         1                0                 60.1              414.     15
## 4         1                1                 20.6               50.3    18
```

```
mean_bills = d2 %>% pull('mean(bills_introduced)')
var_bills = d2 %>% pull('var(bills_introduced)')
n = d2 %>% pull('n()')
```

Using equation 3.4.

```
(se0 = sqrt(var_bills[1]/n[1] + var_bills[3]/n[3])) # SE for Texas
```

```
## [1] 9.345871
```

```
(se1 = sqrt(var_bills[2]/n[2] + var_bills[4]/n[4])) # SE for Arkansas
```

```
## [1] 3.395979
```

c. Use equation (3.10) to estimate the overall ATE for both states combined.

```
N = sum(n)
N0 = n[1] + n[3] # number of Texas senators
N1 = n[2] + n[4] # number of Arkansas senators

# Overall ATE
(ate = N0/N * (mean_bills[3]-mean_bills[1]) +
        N1/N * (mean_bills[4]-mean_bills[2]))
```

```
## [1] -13.2168
```

d. Explain why, in this study, simply pooling the data for the two states and comparing the average number of bills introduced by two-year senators to the average number of bills introduced by four-year senators leads to biased estimate of the overall ATE.

There are more senators from Arkansas than from Texas, so if all pooled together, Arkansas senators have a higher probability of being assigned to treatment. In addition, Arkansas senators tend to introduce far fewer bills than do Texas senator, leading to a negative bias on the estimate of the ATE.

e. Insert the estimated standard errors into equation (3.12) to estimate the stand error for the overall ATE.

```
# SE for overall ATE
sqrt(se0^2 * (N0/N)^2 + se1^2 * (N1/N)^2)
```

```
## [1] 4.74478
```

f. Use randomization inference to test the sharp null hypothesis that the treatment effect is zero for senators in both states.

```
bills = d %>%
    arrange(texas0_arkansas1) %>% # make sure d is sorted so block randomization is easier
    pull(bills_introduced)

simulate = function() {
    texas_treatment = sample(c(rep(0, n[1]), rep(1, n[3])))
    arkansas_treatment = sample(c(rep(0, n[2]), rep(1, n[4])))
    treatment = c(texas_treatment, arkansas_treatment)

    ate = mean(bills[treatment == 1]) - mean(bills[treatment == 0])
    bills_2year_texas = mean(bills[1:N0][texas_treatment == 1])
    bills_4year_texas = mean(bills[1:N0][texas_treatment == 0])
    bills_2year_arkansas = mean(bills[-(1:N0)][arkansas_treatment == 1])
    bills_4year_arkansas = mean(bills[-(1:N0)][arkansas_treatment == 0])
    c(ate,
      bills_2year_texas,
      bills_4year_texas,
      bills_2year_arkansas,
      bills_4year_arkansas)
}

simulation = as.data.frame(t(replicate(10000, simulate())))
names(simulation) = c('ate',
                      'bills_2year_texas',
                      'bills_4year_texas',
                      'bills_2year_arkansas',
                      'bills_4year_arkansas')
```

```
# One tailed p-value
mean(simulation$ate < ate)
```

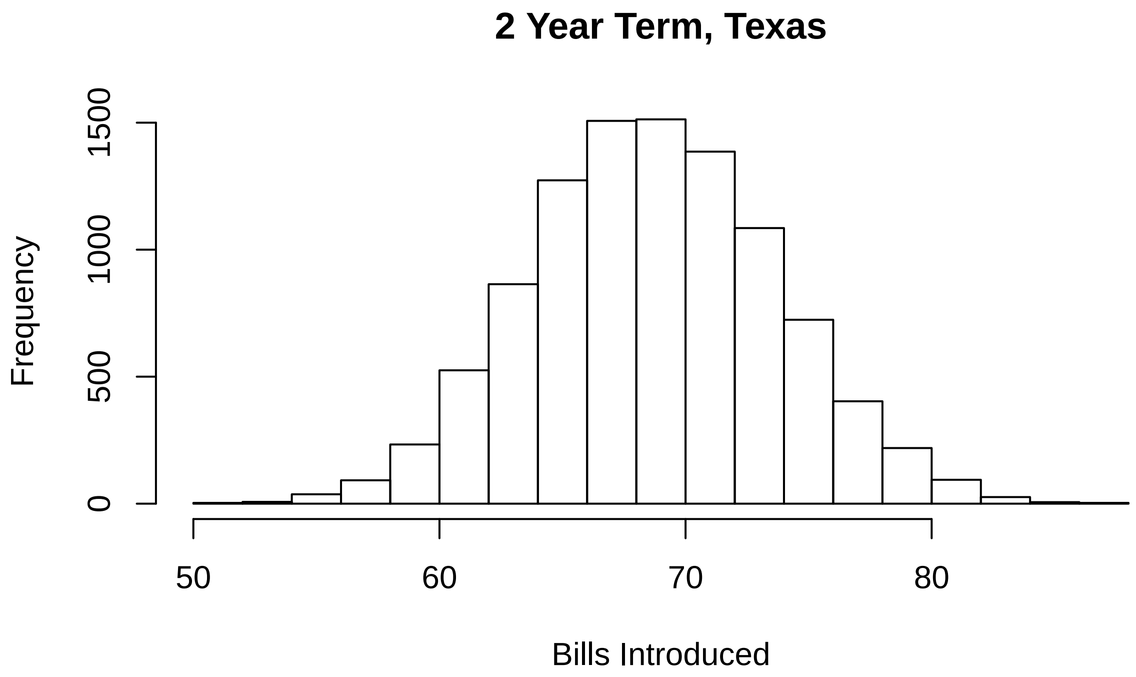
```
## [1] 0.0068
```

```
# Two tailed p-value
mean(abs(simulation$ate) > abs(ate))
```

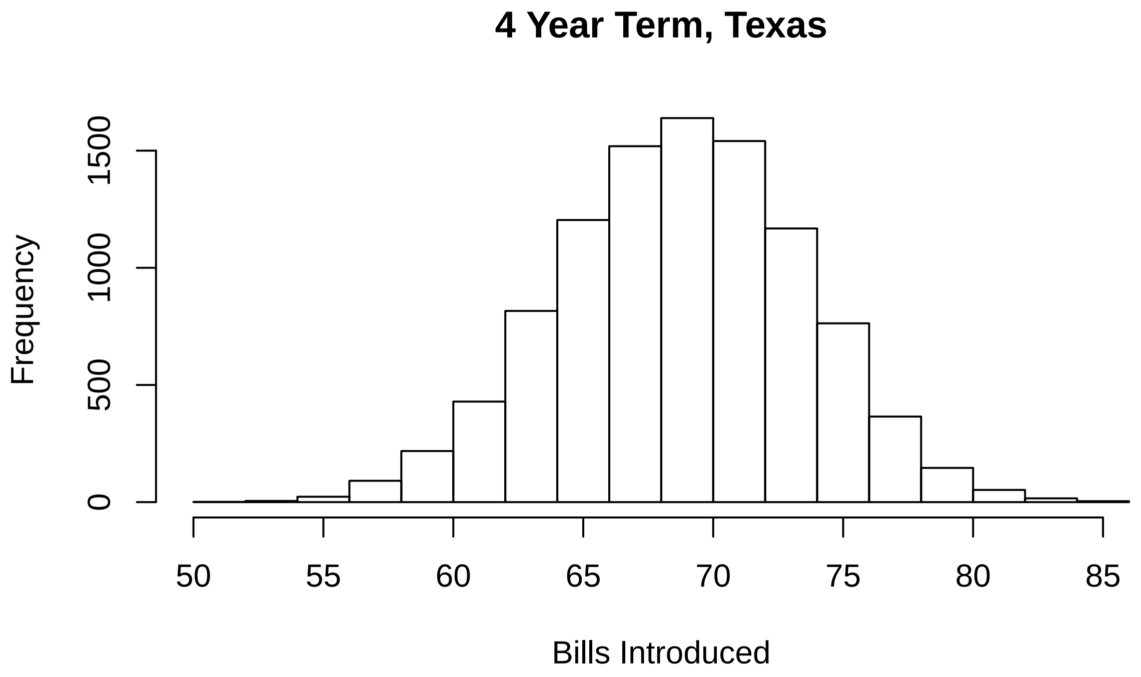
```
## [1] 0.0083
```

g. **IN Addition:** Plot histograms for both the treatment and control groups in each state (for 4 histograms in total).

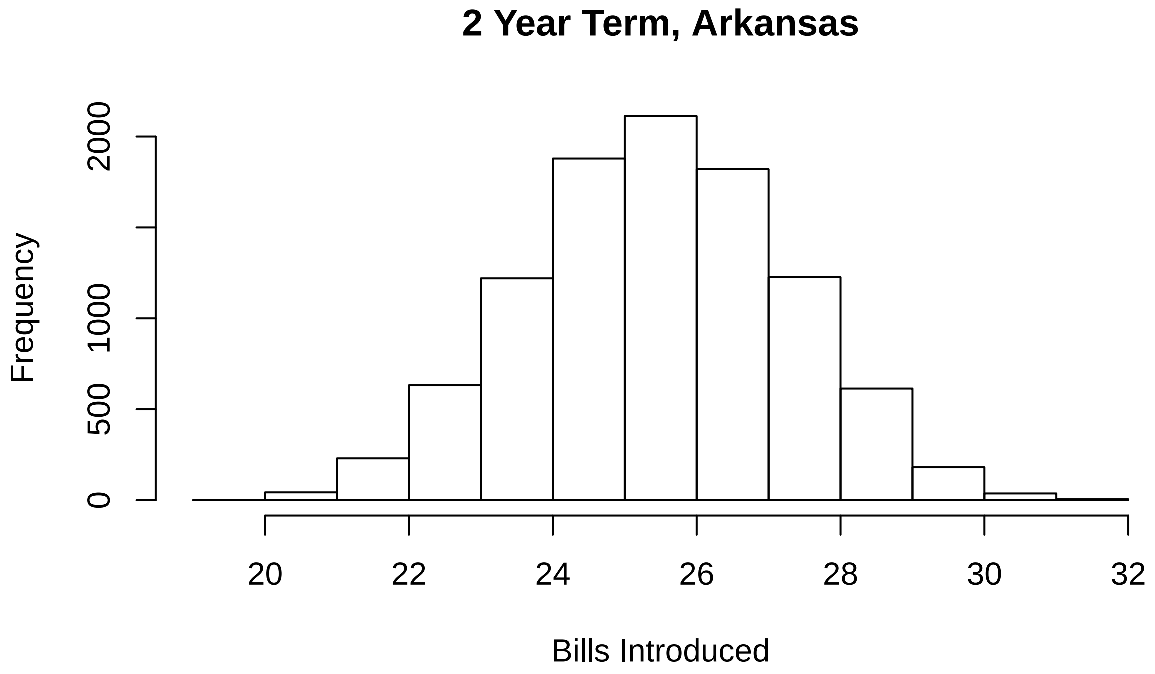
```
hist(simulation$bills_2year_texas, main = '2 Year Term, Texas', xlab = 'Bills Introduced')
```



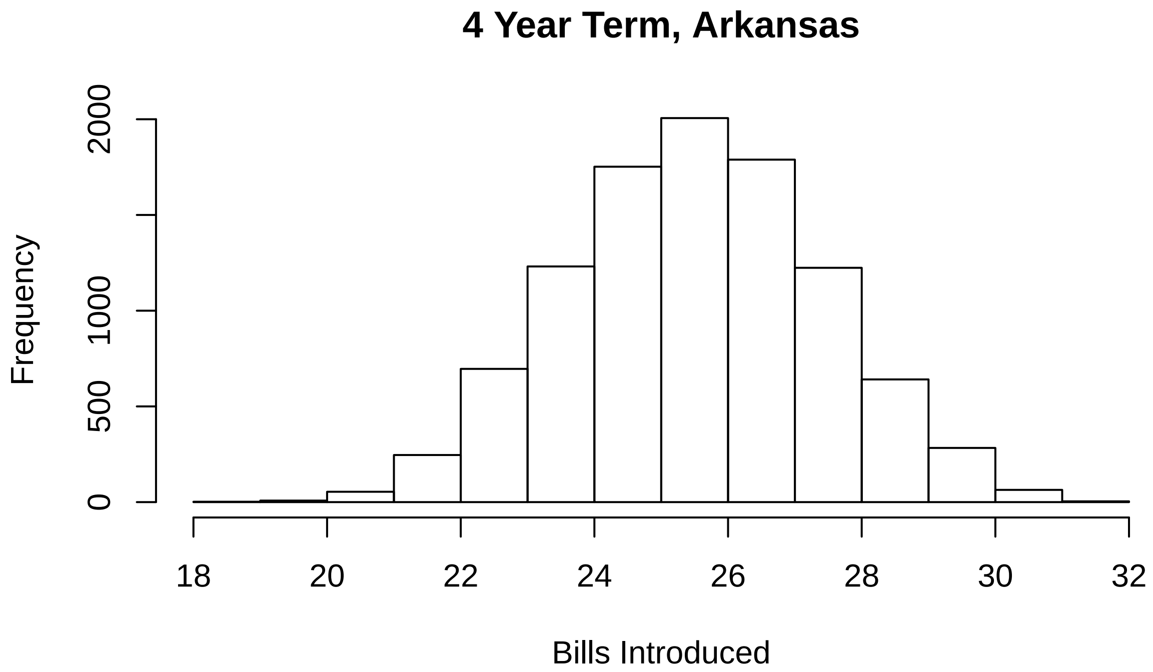
```
hist(simulation$bills_4year_texas, main = '4 Year Term, Texas', xlab = 'Bills Introduced')
```



```
hist(simulation$bills_2year_arkansas, main = '2 Year Term, Arkansas', xlab = 'Bills Introduced')
```



```
hist(simulation$bills_4year_arkansas, main = '4 Year Term, Arkansas', xlab = 'Bills Introduced')
```



One more for fun.

```
library(ggplot2)
library(reshape2)
```

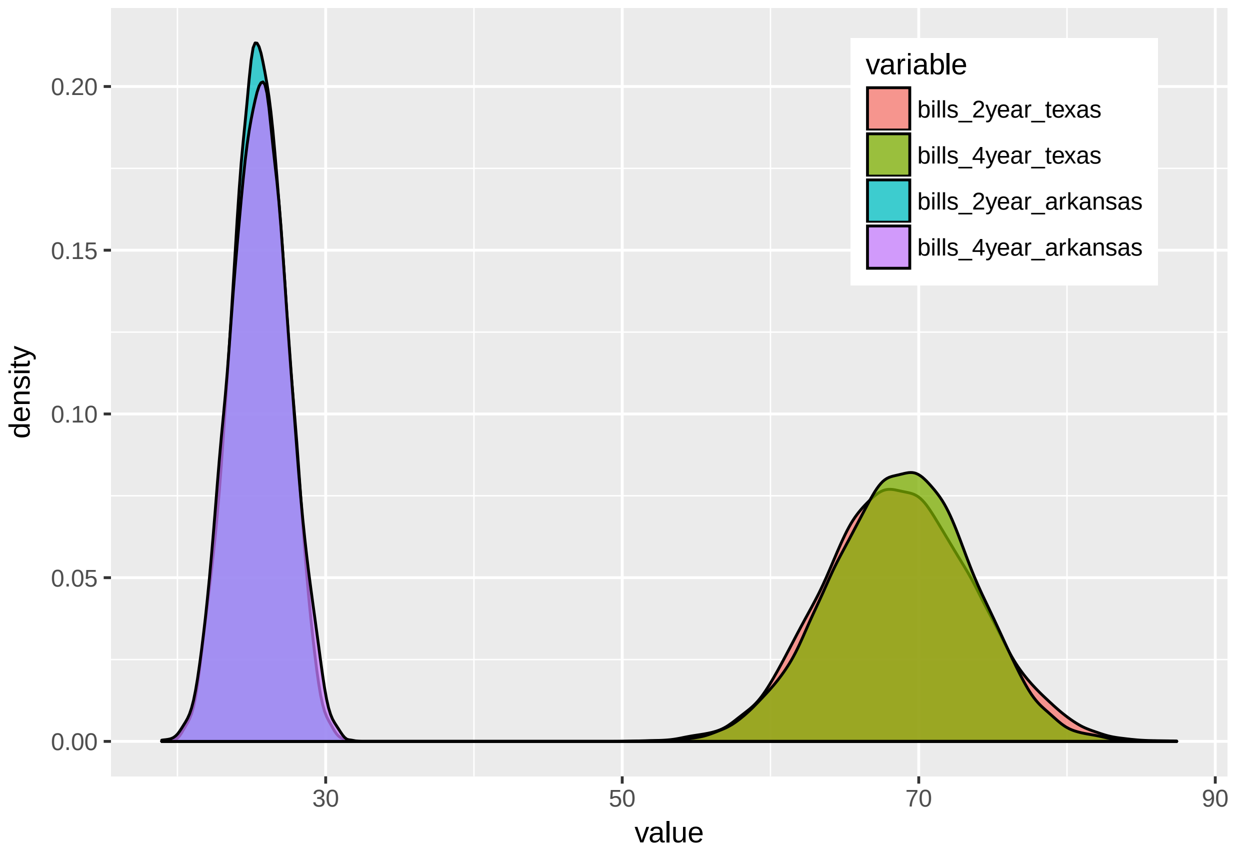
```
##
## Attaching package: 'reshape2'

## The following objects are masked from 'package:data.table':
##
##     dcast, melt
```

```
m = melt(simulation[2:5])
```

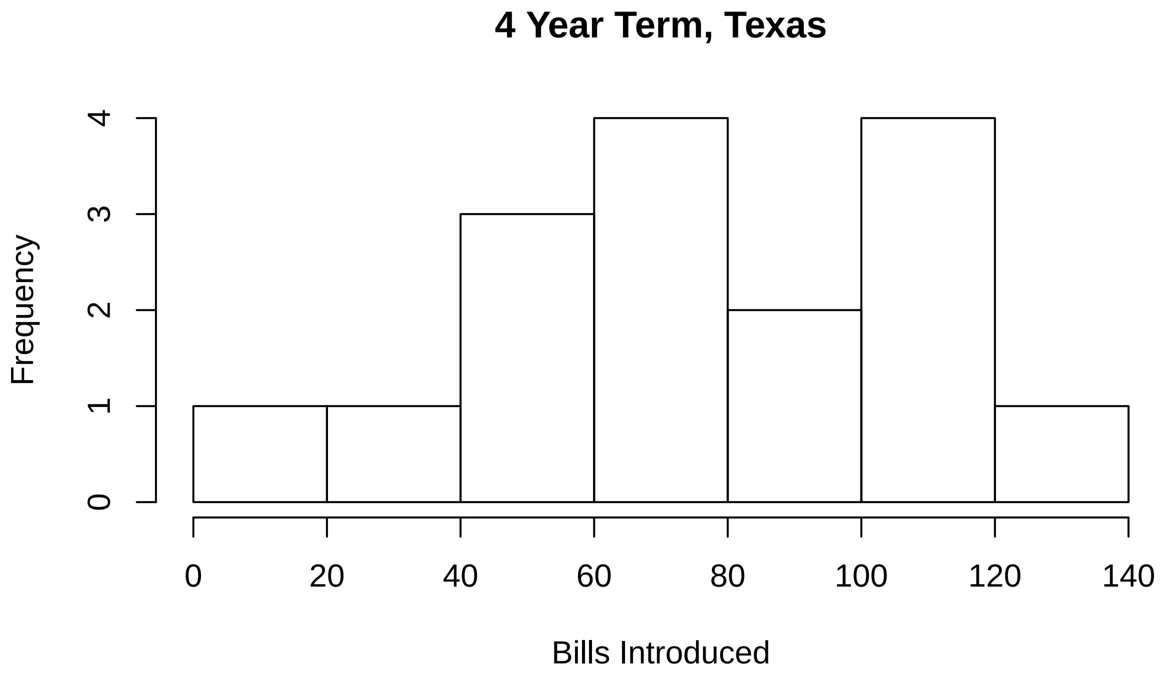
```
## No id variables; using all as measure variables
```

```
ggplot(m, aes(x=value, fill=variable)) +
    geom_density(alpha = .75) +
    theme(legend.position = c(.8, .8))
```

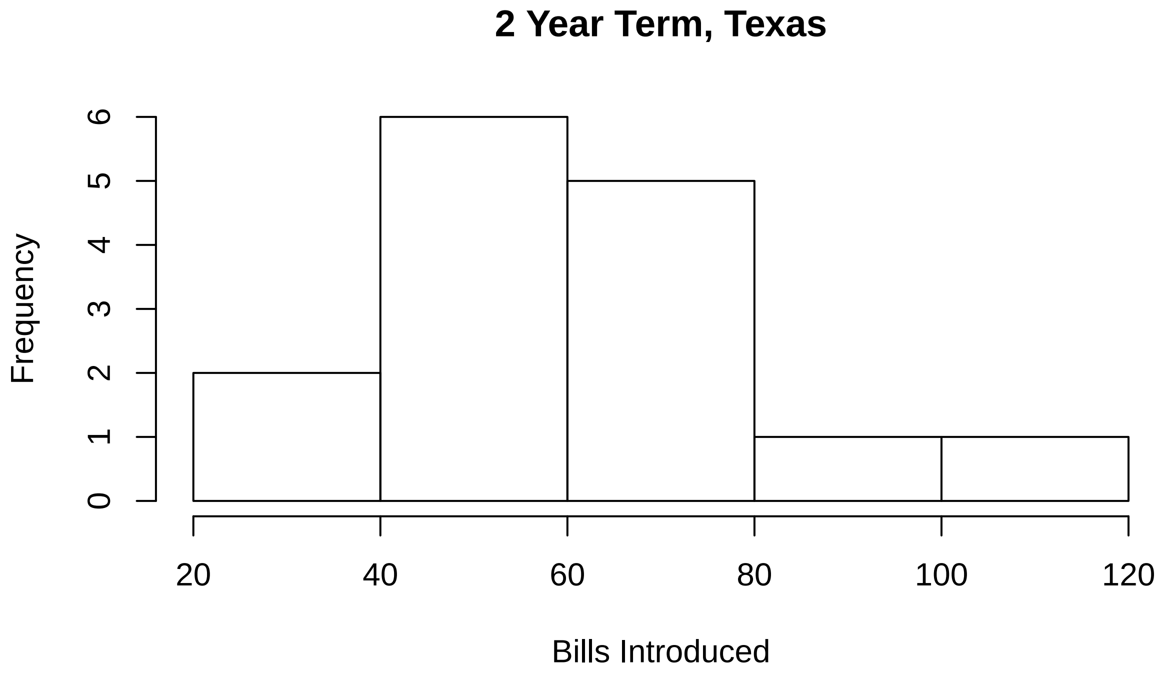


I wasn't sure if I was supposed to plot the simulation results or the results from the original data. But since the above histograms are all just demonstrations of the central limit theorem, here are histograms from the original data.

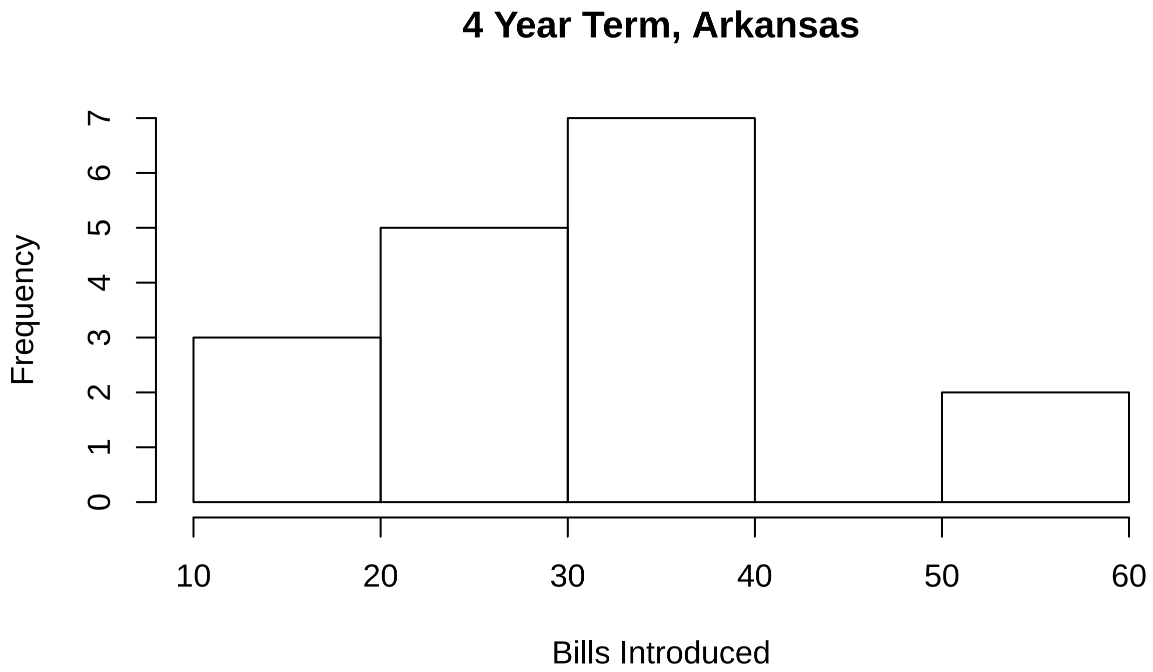
```
hist(d %>% filter(term2year == 0 & texas0_arkansas1 == 0) %>% pull(bills_introduced),
     main = '4 Year Term, Texas', xlab = 'Bills Introduced')
```



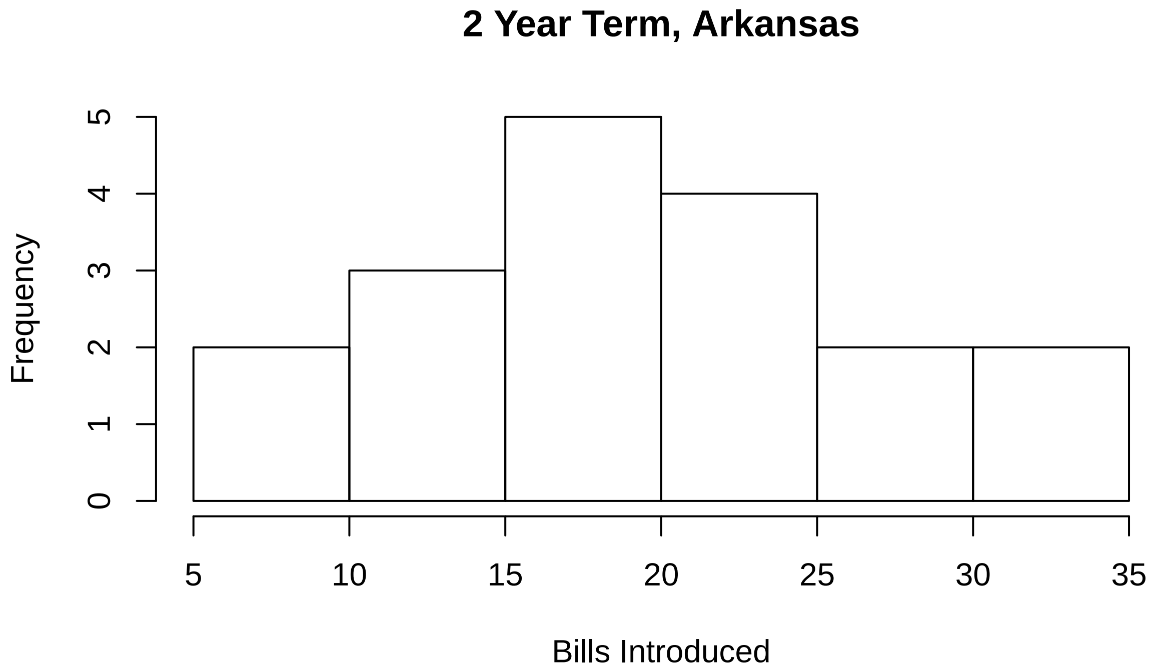
```r
hist(d %>% filter(term2year == 1 & texas0_arkansas1 == 0) %>% pull(bills_introduced),
     main = '2 Year Term, Texas', xlab = 'Bills Introduced')
```



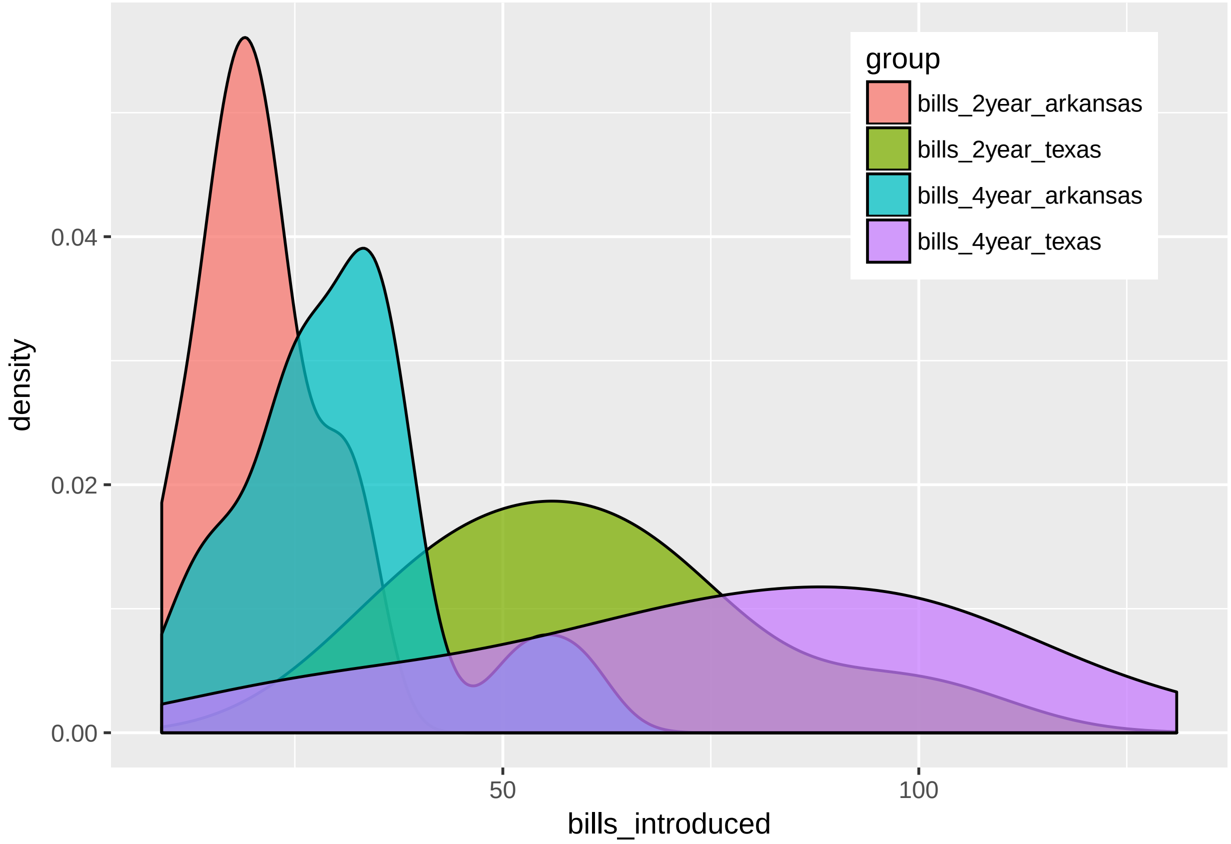
```r
hist(d %>% filter(term2year == 0 & texas0_arkansas1 == 1) %>% pull(bills_introduced),
     main = '4 Year Term, Arkansas', xlab = 'Bills Introduced')
```



```
hist(d %>% filter(term2year == 1 & texas0_arkansas1 == 1) %>% pull(bills_introduced),
     main = '2 Year Term, Arkansas', xlab = 'Bills Introduced')
```



```
d = d %>% mutate(group = c('bills_4year_texas',
                           'bills_2year_texas',
                          'bills_4year_arkansas',
                          'bills_2year_arkansas')[term2year + texas0_arkansas1*2 + 1]
                    )
ggplot(d, aes(x=bills_introduced, fill=group)) +
    geom_density(alpha = .75) +
    theme(legend.position = c(.8, .8))
```



3. Cluster Randomization

Use the data in *Field Experiments* Table 3.3 to simulate cluster randomized assignment. (*Notes: (a) Assume 3 clusters in treatment and 4 in control; and (b) When Gerber and Green say* `simulate'', they do not mean`*run simulations with R code’‘, but rather, in a casual sense “take a look at what happens if you do this this way.” There is no randomization inference necessary to complete this problem.*)

```
## load data
rm(list = ls())
d <- read.csv("./data/ggChapter3.csv", stringsAsFactors = FALSE)
```

a. Suppose the clusters are formed by grouping observations {1,2}, {3,4}, {5,6}, ... , {13,14}. Use equation (3.22) to calculate the standard error assuming half of the clusters are randomly assigned to treatment.

```
N = 14 # total units
m = 6 # number of treated units
k = 7 # number of clusters
```

```
Y0_a = sapply(seq(1,13,2), function(i) (d$Y[i] + d$Y[i+1]) / 2 ) # potential outcome for control
Y1_a = sapply(seq(1,13,2), function(i) (d$D[i] + d$D[i+1]) / 2 ) # potential outcome for treatment

# var in R returns sample variance, scale it to get population variance
varY0_a = var(Y0_a) * (k-1)/k
varY1_a = var(Y1_a) * (k-1)/k

# Standard Error
sqrt(1/(k-1) * (m*varY0_a/(N-m) + (N-m)*varY1_a/m) + 2*cov(Y0_a,Y1_a))
```

```
## [1] 9.044376
```

b. Suppose that clusters are instead formed by grouping observations {1,14}, {2,13}, {3,12}, … , {7,8}. Use equation (3.22) to calculate the standard error assuming half of the clusters are randomly assigned to treatment.

```
Y0_b = sapply(seq(1,7), function(i) (d$Y[i] + d$Y[15-i]) / 2 ) # potential outcome for control
Y1_b = sapply(seq(1,7), function(i) (d$D[i] + d$D[15-i]) / 2 ) # potential outcome for treatment

# var in R returns sample variance, we need to scale it to get population variance
varY0_b = var(Y0_b) * (k-1)/k
varY1_b = var(Y1_b) * (k-1)/k

# Standard Error
sqrt(1/(k-1) * (m*varY0_b/(N-m) + (N-m)*varY1_b/m) + 2*cov(Y0_b,Y1_b))
```

```
## [1] 1.604848
```

c. Why do the two methods of forming clusters lead to different standard errors? What are the implications for the design of cluster randomized experiments?

The data frame is sorted by potential outcome for control. Therefore, when we create clusters using villages that are adjacent in the list, we get a higher standard error than if we cluster using villages from opposite ends of the list. The implication is that we need to be careful with units that are clustered together. The intracluster variance in potential outcomes may be low, but the intercluster variance may be high, leading to a high standard error for our ATE estimate.

4. Sell Phones?

You are an employee of a newspaper and are planning an experiment to demonstrate to Apple that online advertising on your website causes people to buy iPhones. Each site visitor shown the ad campaign is exposed to $0.10 worth of advertising for iPhones. (Assume all users could see ads.) There are 1,000,000 users available to be shown ads on your newspaper's website during the one week campaign.

Apple indicates that they make a profit of $100 every time an iPhone sells and that 0.5% of visitors to your newspaper's website buy an iPhone in a given week in general, in the absence of any advertising.

a. By how much does the ad campaign need to increase the probability of purchase in order to be "worth it" and a positive ROI (supposing there are no long-run effects and all the effects are measured within that week)?

Let q be the percent increase that makes up the cost (I'm using q because there's a p in the next part).

$$100 \times 10^6 \times q = 0.10 \times 10^6$$

$$q = 0.001$$

b. Assume the measured effect is 0.2 percentage points. If users are split 50:50 between the treatment group (exposed to iPhone ads) and control group (exposed to unrelated advertising or nothing; something you can assume has no effect), what will be the confidence interval of your estimate on whether people purchase the phone?

```
N = 1e6

n1 = N * .5
n2 = N * .5
x1 = .007 * n1
x2 = .005 * n2
p = (x1+x2) / (n1+n2)

# Standard Error
(se = sqrt(p * (1-p) * (1/n1 + 1/n2)))
```

```
## [1] 0.0001544539
```

```
# Confidence Interval
c(.002 - 1.96*se, .002 + 1.96*se)
```

```
## [1] 0.00169727 0.00230273
```

- **Note:** The standard error for a two-sample proportion test is $\sqrt{p(1-p) * (\frac{1}{n_1} + \frac{1}{n_2})}$ where $p = \frac{x_1+x_2}{n_1+n_2}$, where x and n refer to the number of "successes" (here, purchases) over the number of "trials" (here, site visits). The length of each tail of a 95% confidence interval is calculated by multiplying the standard error by 1.96.

c. Is this confidence interval precise enough that you would recommend running this experiment? Why or why not?

This confidence interval is precise enough to run the experiment. The desired percentage rise is outside the CI, so we can be fairly confident that the 0.2% rise was unlikely due to chance.

d. Your boss at the newspaper, worried about potential loss of revenue, says he is not willing to hold back a control group any larger than 1% of users. What would be the width of the confidence interval for this experiment if only 1% of users were placed in the control group?

```
n1 = N * .01
n2 = N * .99
x1 = .007 * n1
x2 = .005 * n2
p = (x1+x2) / (n1+n2)

# Standard Error
(se = sqrt(p * (1-p) * (1/n1 + 1/n2)))
```

```
## [1] 0.0007102994
```

```
# Confidence Interval
c(.002 - 1.96*se, .002 + 1.96*se)
```

```
## [1] 0.0006078132 0.0033921868
```

Now, the desired percent increase lies in the CI. We cannot be confident that the measured effect was not due to pure chance.

5. Sports Cards

Here you will find a set of data from an auction experiment by John List and David Lucking-Reiley (2000).

```
d2 <- read.csv("./data/listData.csv", stringsAsFactors = FALSE)
head(d2)
```

```
##   bid uniform_price_auction
## 1   5                     1
## 2   5                     1
## 3  20                     0
## 4   0                     1
## 5  20                     1
## 6   0                     1
```

In this experiment, the experimenters invited consumers at a sports card trading show to bid against one other bidder for a pair trading cards. We abstract from the multi-unit-auction details here, and simply state that the treatment auction format was theoretically predicted to produce lower bids than the control auction format. We provide you a relevant subset of data from the experiment.

a. Compute a 95% confidence interval for the difference between the treatment mean and the control mean, using analytic formulas for a two-sample t-test from your earlier statistics course.

```
control = d2[d2$uniform_price_auction == 0, ]
treatment = d2[d2$uniform_price_auction == 1, ]
n0 = nrow(control)
n1 = nrow(treatment)
df = n0 + n1 - 2
cr = qt(.975, df) # 2 tailed critical value

# Standard Error
(se = sqrt(sd(control$bid)^2/n0 + sd(treatment$bid)^2/n1))
```

```
## [1] 4.326572
```

```
# Mean of Control
(x0bar = mean(control$bid))
```

```
## [1] 28.82353
```

```
# Mean of Treatment
(x1bar = mean(treatment$bid))
```

```
## [1] 16.61765
```

```
# ATE
(ate = x1bar - x0bar)
```

```
## [1] -12.20588
```

```
# Confidence Interval
(c(ate - cr*se, ate + cr*se))
```

```
## [1] -20.844162  -3.567603
```

b. In plain language, what does this confidence interval mean?

If we repeated the experiment several times, the true difference in means between the treatment and control groups would be in that interval 95% of the time.

c. Regression on a binary treatment variable turns out to give one the same answer as the standard analytic formula you just used. Demonstrate this by regressing the bid on a binary variable equal to 0 for the control auction and 1 for the treatment auction.

```
(model = lm(d2$bid ~ d2$uniform_price_auction))
```

```
##
## Call:
## lm(formula = d2$bid ~ d2$uniform_price_auction)
##
## Coefficients:
##              (Intercept)  d2$uniform_price_auction
##                    28.82                    -12.21
```

The coefficient on `uniform_price_auction` is the same as the previously calculated ATE.

d. Calculate the 95% confidence interval you get from the regression.

```
confint(model)
```

```
##                               2.5 %    97.5 %
## (Intercept)                22.71534 34.931716
## d2$uniform_price_auction -20.84416 -3.567603
```

The confidence interval on `uniform_price_auction` is the same confidence interval as from the analytical calculation.

e. On to p-values. What p-value does the regression report? Note: please use two-tailed tests for the entire problem.

```
summary(model)$coefficients[2,4]
```

```
## [1] 0.006314796
```

f. Now compute the same p-value using randomization inference.

```
simulate = function() {
    treatment = sample(c(rep(0, n0), rep(1, n1)))
    mean(d2$bid[treatment == 1]) - mean(d2$bid[treatment == 0])
}

simulated_ates = replicate(10000, simulate())

# p-value
mean(simulated_ates < ate)
```

```
## [1] 0.0028
```

g. Compute the same p-value again using analytic formulas for a two-sample t-test from your earlier statistics course. (Also see part (a).)

```
t.test(d2$bid ~ d2$uniform_price_auction)
```

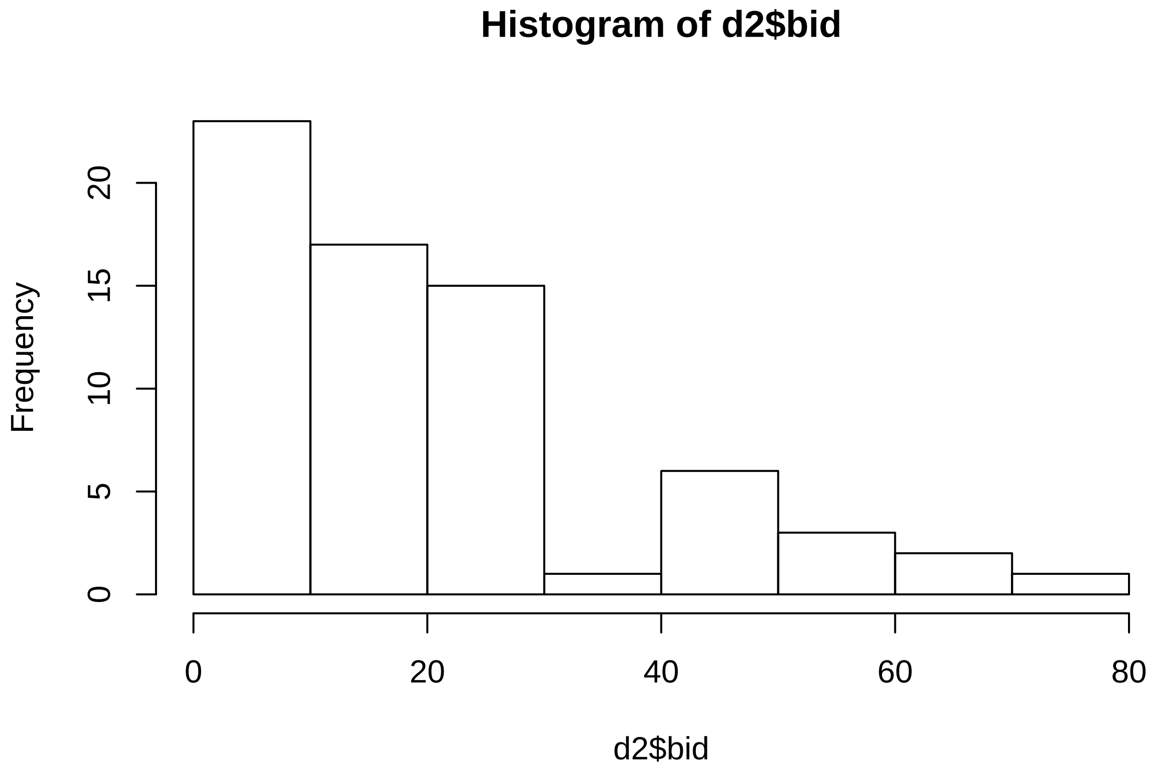
```
##
##  Welch Two Sample t-test
##
## data:  d2$bid by d2$uniform_price_auction
## t = 2.8211, df = 61.983, p-value = 0.006421
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   3.557141 20.854624
## sample estimates:
## mean in group 0 mean in group 1
##        28.82353        16.61765
```

The confidence interval flipped but it's because of how R interpretted which one is treatment and which is control.

h. Compare the two p-values in parts (e) and (f). Are they much different? Why or why not? How might your answer to this question change if the sample size were different?

The p-value from randomization inference is slightly different. I ran it a few times and the p-value seems to be a little bit lower from the simulation as opposed to the analytic solutions. This is probably due to the fact that the bid distribution is right skewed. Increasing the sample size should help smooth this out and make the central limit theorem more applicable.

```r
hist(d2$bid)
```


Histogram of d2$bid
Frequency
d2$bid