Homework4

Megan Jasek, Rohan Thakur, Charles Kekeh

Saturday, February 20, 2016

```
# Load the dataset and print descriptions
load("athletics.RData")
desc
```

```
##    variable                         label
## 1      year                  1992 or 1993
## 2      apps        # applics for admission
## 3     top25  perc frsh class in 25 hs perc
## 4    ver500 perc frsh >= 500 on verbal SAT
## 5    mth500   perc frsh >= 500 on math SAT
## 6    stufac          student-faculty ratio
## 7      bowl    = 1 if bowl game in prev yr
## 8    btitle  = 1 if men's cnf chmps prv yr
## 9   finfour    = 1 if men's final 4 prv yr
## 10    lapps                     log(apps)
## 11   avg500               (ver500+mth500)/2
## 12   school             name of university
## 13    bball         =1 if btitle or finfour
```

Question1

How many observations and variables are in the dataset?

There are 116 observations of 14 variables in the data set.

Data Analysis

- **apps** - The histogram shows a data distribution that's positevely skewed with most universities showing between 3000 and 10000 applications for the years 1992 and 1993.
- **bowl** - The bowl variable is a binary categorical variable. We note that over the 2 year period considered there are fewer universities that appear in bowl games than universities that do.
- **btitle** - The btitle variable is also a binary categorical variable. The histogram, (as one would expect) shows that there are significantly fewer universities that have won a title over the 2 year period considered than universities that have.
- **finfour** - The finfour variable is also a binary categorical variable. The histogram, (as one would expect) shows that there are significantly fewer universities that have participated in men's final four games over the 2 year period considered than universities that have.

```
# How many observations and variables are in the dataset
str(data)
```

```
## 'data.frame':    116 obs. of  14 variables:
##  $ year   : int  1992 1993 1992 1993 1992 1993 1992 1993 1992 1993 ...
```

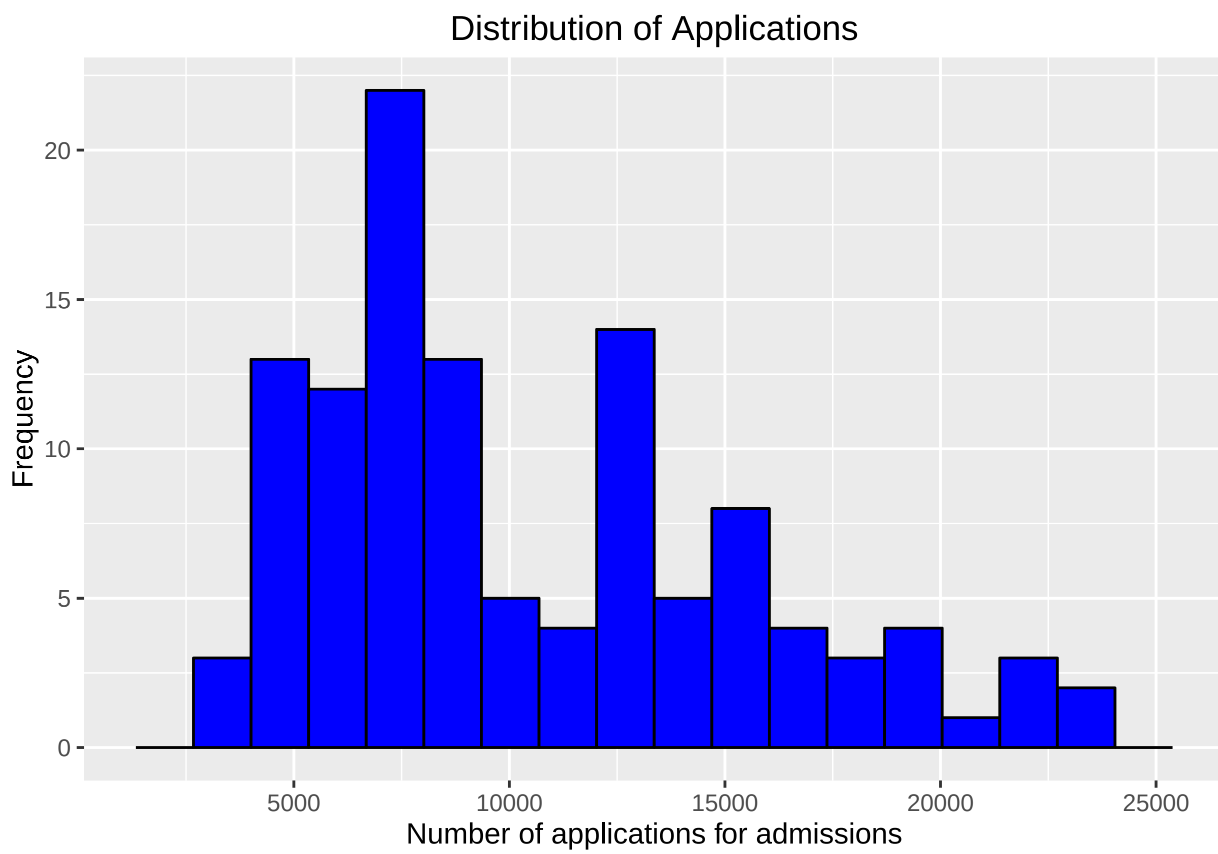
```
##  $ apps   : int  6245 7677 13327 19860 10422 12809 4103 3303 8661 7548 ...
##  $ top25  : int  49 58 57 57 37 49 60 67 54 54 ...
##  $ ver500 : int  NA NA 36 36 28 31 NA NA 46 51 ...
##  $ mth500 : int  NA NA 58 58 58 62 NA NA 86 83 ...
##  $ stufac : int  20 15 16 16 20 14 16 18 16 16 ...
##  $ bowl   : int  1 1 0 1 0 0 1 0 0 0 ...
##  $ btitle : int  0 0 0 1 0 0 1 0 0 0 ...
##  $ finfour: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ lapps  : num  8.74 8.95 9.5 9.9 9.25 ...
##  $ avg500 : num  NA NA 47 47 43 46.5 NA NA 66 67 ...
##  $ school : chr  "alabama" "alabama" "arizona" "arizona" ...
##  $ bball  : int  0 0 0 1 0 0 1 0 0 0 ...
##  $ perf   : int  1 1 0 2 0 0 2 0 0 0 ...
```

```r
# apps variable
print(quantile(data$apps, probs = c(0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9,
    0.95, 0.99, 1)))
```

```
##       1%       5%      10%      25%      50%      75%      90%      95%
##  3565.20  4218.50  4801.00  6896.75  8646.00 13423.50 18116.00 19975.00
##      99%     100%
## 22815.25 23342.00
```

```r
# Plot the histogram of apps at 15 bins
apps.hist <- ggplot(data, aes(apps)) + theme(legend.position = "none") +
    geom_histogram(fill = "Blue", colour = "Black", binwidth = (range(data$apps)[2] -
        range(data$apps)[1])/15) + labs(title = "Distribution of Applications",
    x = "Number of applications for admissions", y = "Frequency")

plot(apps.hist)
```

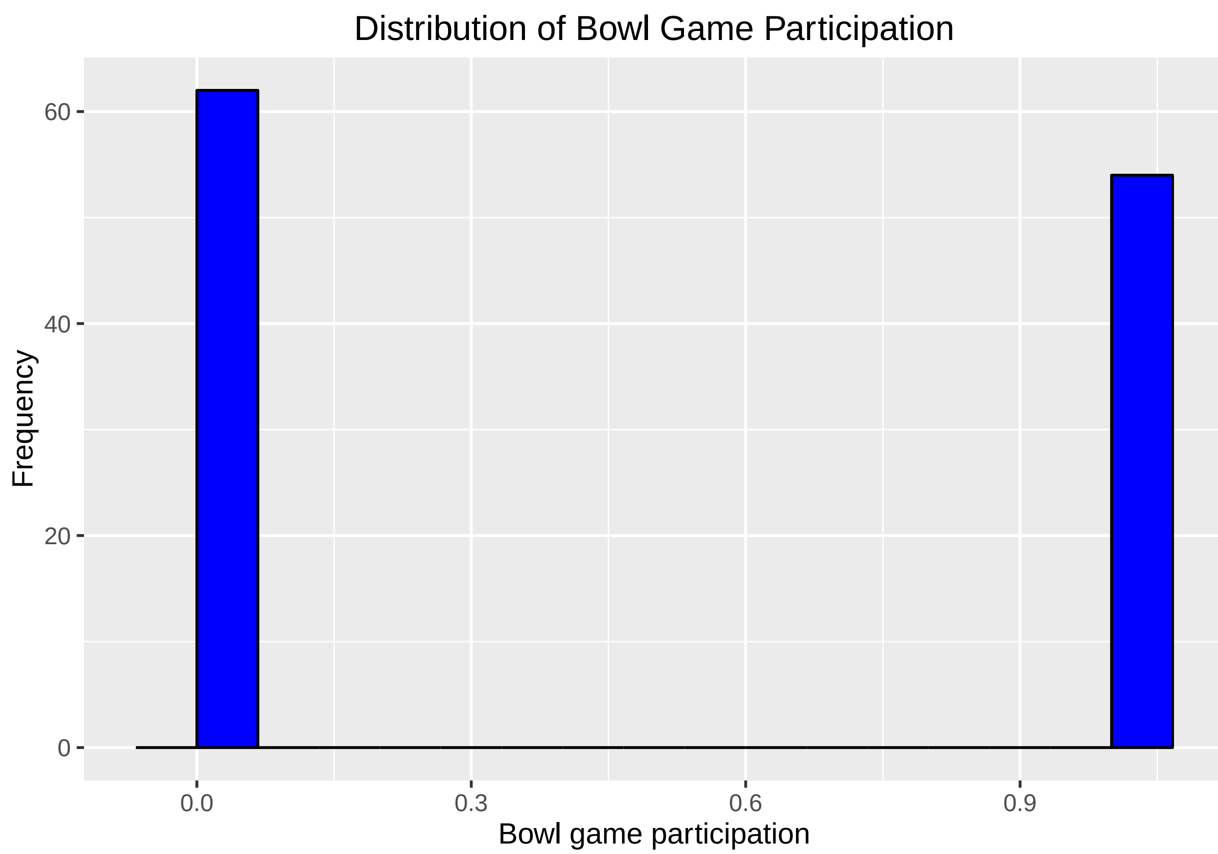



```r
# bowl variable
print(quantile(data$bowl, probs = c(0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9,
    0.95, 0.99, 1)))
```

```
##   1%   5%  10%  25%  50%  75%  90%  95%  99% 100%
##    0    0    0    0    0    1    1    1    1    1
```

```r
# Plot the histogram of bowl at 15 bins
bowl.hist <- ggplot(data, aes(bowl)) + theme(legend.position = "none") +
    geom_histogram(fill = "Blue", colour = "Black", binwidth = (range(data$bowl)[2] -
        range(data$bowl)[1])/15) + labs(title = "Distribution of Bowl Game Participation",
    x = "Bowl game participation", y = "Frequency")

plot(bowl.hist)
```

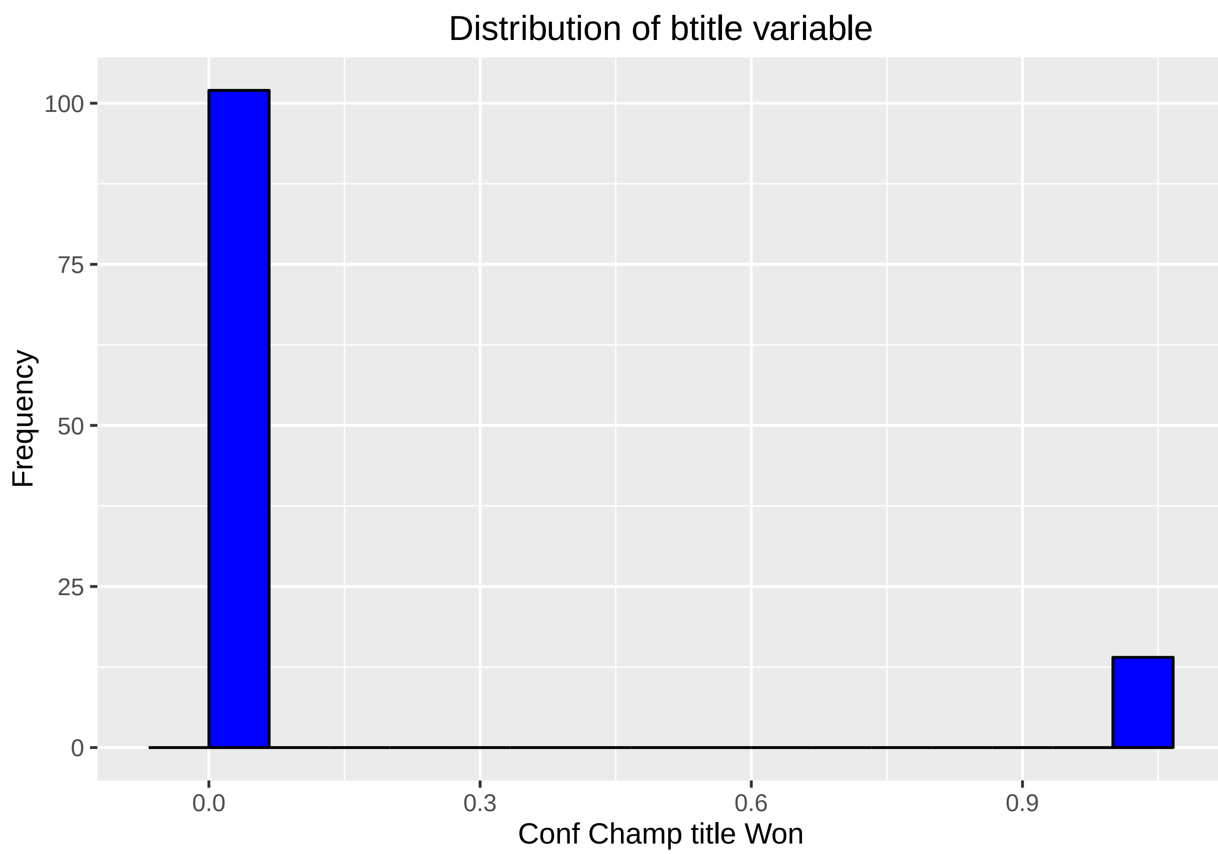



```r
# btitle variable
print(quantile(data$btitle, probs = c(0.01, 0.05, 0.1, 0.25, 0.5, 0.75,
    0.9, 0.95, 0.99, 1)))
```

```
##   1%   5%  10%  25%  50%  75%  90%  95%  99% 100%
##    0    0    0    0    0    0    1    1    1    1
```

```r
# Plot the histogram of btitle at 15 bins
btitle.hist <- ggplot(data, aes(btitle)) + theme(legend.position = "none") +
    geom_histogram(fill = "Blue", colour = "Black", binwidth = (range(data$btitle)[2] -
        range(data$btitle)[1])/15) + labs(title = "Distribution of btitle variable",
    x = "Conf Champ title Won", y = "Frequency")

plot(btitle.hist)
```

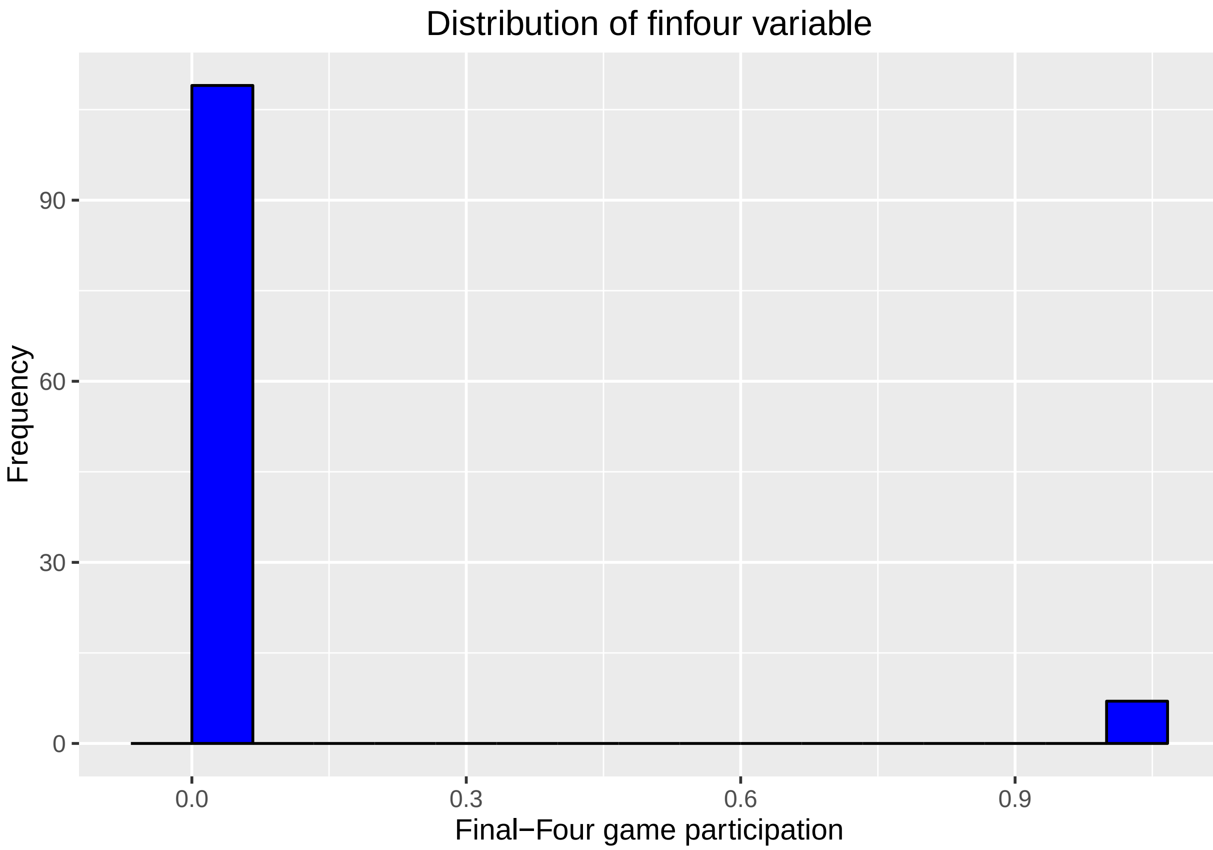



```
# finfour variable
print(quantile(data$finfour, probs = c(0.01, 0.05, 0.1, 0.25, 0.5, 0.75,
    0.9, 0.95, 0.99, 1)))
```

```
##   1%   5%  10%  25%  50%  75%  90%  95%  99% 100%
##    0    0    0    0    0    0    0    1    1    1
```

```
# Plot the histogram of finfour at 15 bins
finfour.hist <- ggplot(data, aes(finfour)) + theme(legend.position = "none") +
    geom_histogram(fill = "Blue", colour = "Black", binwidth = (range(data$finfour)[2] -
        range(data$finfour)[1])/15) + labs(title = "Distribution of finfour variable",
    x = "Final-Four game participation", y = "Frequency")

plot(finfour.hist)
```



Question2

The distribution of the change of log of application number has the appearance of a normal(ish) distribution. There are 2 outlier points with a change of -.2 and +.4 that correspond to Arizona University and Arkansas University. There are however no indications that these outliers would affect the regression at this point or that they should be removed.

Which schools had the greatest increase in number of log applications?

Arizona, Alabama and Arizona State (see table below).

Which schools had the greatest decrease in number of log applications?

Arkansas, Oklahoma State and Penn State (see table below).

```
reshaped.data <- reshape(data, v.names = c("apps", "top25", "ver500", "mth500",
    "stufac", "bowl", "btitle", "finfour", "lapps", "avg500", "bball",
    "perf"), timevar = "year", idvar = "school", direction = "wide")
# Check the layout of the reshaped data
str(reshaped.data)
```

```
## 'data.frame':    58 obs. of  25 variables:
```

```
## $ school      : chr  "alabama" "arizona" "arizona state" "arkansas" ...
## $ apps.1992   : int  6245 13327 10422 4103 8661 12283 20281 8037 13761 12420 ...
## $ top25.1992  : int  49 57 37 60 54 96 NA 63 66 97 ...
## $ ver500.1992 : int  NA 36 28 NA 46 86 75 38 NA 94 ...
## $ mth500.1992 : int  NA 58 58 NA 86 98 93 78 NA 98 ...
## $ stufac.1992 : int  20 16 20 16 16 15 20 19 19 12 ...
## $ bowl.1992   : int  1 0 0 1 0 0 1 1 1 0 ...
## $ btitle.1992 : int  0 0 0 1 0 0 0 0 0 1 ...
## $ finfour.1992: int  0 0 0 0 0 0 0 0 0 1 ...
## $ lapps.1992  : num  8.74 9.5 9.25 8.32 9.07 ...
## $ avg500.1992 : num  NA 47 43 NA 66 92 84 58 NA 96 ...
## $ bball.1992  : int  0 0 0 1 0 0 0 0 0 1 ...
## $ perf.1992   : int  1 0 0 2 0 0 1 1 1 2 ...
## $ apps.1993   : int  7677 19860 12809 3303 7548 13112 19873 8065 14063 13789 ...
## $ top25.1993  : int  58 57 49 67 54 NA NA 65 57 97 ...
## $ ver500.1993 : int  NA 36 31 NA 51 82 NA 44 52 93 ...
## $ mth500.1993 : int  NA 58 62 NA 83 98 NA 81 81 98 ...
## $ stufac.1993 : int  15 16 14 18 16 15 18 17 22 12 ...
## $ bowl.1993   : int  1 1 0 0 0 1 0 0 1 0 ...
## $ btitle.1993 : int  0 1 0 0 0 0 0 0 0 0 ...
## $ finfour.1993: int  0 0 0 0 0 0 0 0 0 0 ...
## $ lapps.1993  : num  8.95 9.9 9.46 8.1 8.93 ...
## $ avg500.1993 : num  NA 47 46.5 NA 67 90 NA 62.5 66.5 95.5 ...
## $ bball.1993  : int  0 1 0 0 0 0 0 0 0 0 ...
## $ perf.1993   : int  1 2 0 0 0 1 0 0 1 0 ...
## - attr(*, "reshapeWide")=List of 5
##  ..$ v.names: chr  "apps" "top25" "ver500" "mth500" ...
##  ..$ timevar: chr "year"
##  ..$ idvar  : chr "school"
##  ..$ times  : int  1992 1993
##  ..$ varying: chr [1:12, 1:2] "apps.1992" "top25.1992" "ver500.1992" "mth500.1992" ...
```

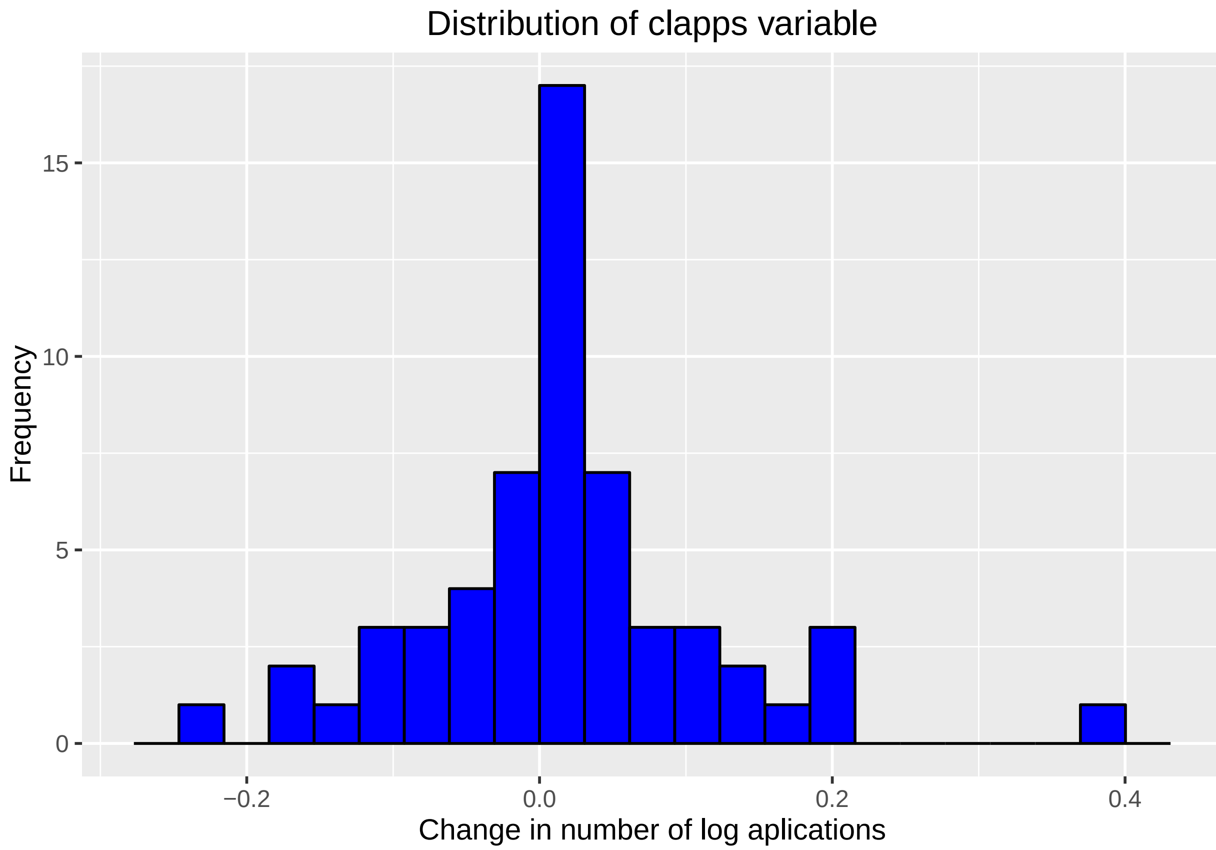
```r
# Create the new variable for the change in the log of the number of
# applications
reshaped.data$clapps <- reshaped.data$lapps.1993 - reshaped.data$lapps.1992

# examine the new variable
print(quantile(reshaped.data$clapps, probs = c(0.01, 0.05, 0.1, 0.25, 0.5,
    0.75, 0.9, 0.95, 0.99, 1)))
```

```
##           1%           5%          10%          25%          50%
## -0.193653297 -0.142649984 -0.099300671 -0.026628017  0.006773949
##          75%          90%          95%          99%         100%
##  0.049490690  0.133792496  0.189876366  0.289209118  0.398916245
```

```r
# Plot the histogram of at 20 bins
clapps.hist <- ggplot(reshaped.data, aes(clapps)) + theme(legend.position = "none") +
    geom_histogram(fill = "Blue", colour = "Black", binwidth = (range(reshaped.data$clapps)[2] -
        range(reshaped.data$clapps)[1])/20) + labs(title = "Distribution of clapps variable",
    x = "Change in number of log aplications", y = "Frequency")

plot(clapps.hist)
```




```r
# Which schools had the greatest increase in number of log
# applications?
head(reshaped.data[order(reshaped.data$clapps, decreasing = TRUE), c("school",
    "clapps")])
```

```
##              school    clapps
## 3           arizona 0.3989162
## 1           alabama 0.2064476
## 5     arizona state 0.2062283
## 77           oregon 0.1869907
## 107       villanova 0.1601181
## 61       n.c. state 0.1362371
```

```r
# Which schools had the greatest decrease in number of log
# applications?
head(reshaped.data[order(reshaped.data$clapps, decreasing = FALSE), c("school",
    "clapps")])
```

```
##               school     clapps
## 7           arkansas -0.2168865
## 75    oklahoma state -0.1761265
## 81        penn state -0.1715641
## 9             auburn -0.1375475
## 45   louisiana state -0.1113930
## 23     florida state -0.1036940
```

Question3

Which of these new variables has the highest variance?

The cperf variable has the highest variance, with a value of 0.82425892

```
# Create the additional variables
reshaped.data$cperf <- reshaped.data$perf.1993 - reshaped.data$perf.1992
print(quantile(reshaped.data$cperf, probs = c(0.01, 0.05, 0.1, 0.25, 0.5,
    0.75, 0.9, 0.95, 0.99, 1)))
```

```
##    1%    5%   10%   25%   50%   75%   90%   95%   99%  100%
## -2.00 -1.15 -1.00  0.00  0.00  0.00  1.00  1.15  2.43  3.00
```

```
print(stat.desc(reshaped.data$cperf))
```

```
##      nbr.val     nbr.null       nbr.na          min          max
##  58.00000000  34.00000000   0.00000000  -2.00000000   3.00000000
##        range          sum       median         mean      SE.mean
##   5.00000000  -1.00000000   0.00000000  -0.01724138   0.11921141
## CI.mean.0.95          var      std.dev     coef.var
##   0.23871673   0.82425892   0.90788707 -52.65744978
```

```
reshaped.data$cbball <- reshaped.data$bball.1993 - reshaped.data$bball.1992
print(quantile(reshaped.data$cbball, probs = c(0.01, 0.05, 0.1, 0.25, 0.5,
    0.75, 0.9, 0.95, 0.99, 1)))
```

```
##   1%   5%  10%  25%  50%  75%  90%  95%  99% 100%
## -1.0 -1.0 -0.3  0.0  0.0  0.0  0.0  1.0  1.0  1.0
```

```
print(stat.desc(reshaped.data$cbball))
```

```
##      nbr.val     nbr.null       nbr.na          min          max
##  58.00000000  48.00000000   0.00000000  -1.00000000   1.00000000
##        range          sum       median         mean      SE.mean
##   2.00000000  -2.00000000   0.00000000  -0.03448276   0.05480824
## CI.mean.0.95          var      std.dev     coef.var
##   0.10975160   0.17422868   0.41740709 -12.10480548
```

```
reshaped.data$cbowl <- reshaped.data$bowl.1993 - reshaped.data$bowl.1992
print(quantile(reshaped.data$cbowl, probs = c(0.01, 0.05, 0.1, 0.25, 0.5,
    0.75, 0.9, 0.95, 0.99, 1)))
```

```
##   1%   5%  10%  25%  50%  75%  90%  95%  99% 100%
##   -1   -1   -1    0    0    0    1    1    1    1
```

```
print(stat.desc(reshaped.data$cbowl))
```

```
##      nbr.val     nbr.null       nbr.na          min          max
##  58.00000000  40.00000000   0.00000000  -1.00000000   1.00000000
##        range          sum       median         mean      SE.mean
##   2.00000000   0.00000000   0.00000000   0.00000000   0.07378785
## CI.mean.0.95          var      std.dev     coef.var
##   0.14775761   0.31578947   0.56195149          Inf
```

```
reshaped.data$cbtitle <- reshaped.data$btitle.1993 - reshaped.data$btitle.1992
print(quantile(reshaped.data$cbtitle, probs = c(0.01, 0.05, 0.1, 0.25,
    0.5, 0.75, 0.9, 0.95, 0.99, 1)))
```

```
##   1%   5%  10%  25%  50%  75%  90%  95%  99% 100%
## -1.0 -1.0 -0.3  0.0  0.0  0.0  0.0  1.0  1.0  1.0
```

```
print(stat.desc(reshaped.data$cbtitle))
```

```
##      nbr.val     nbr.null       nbr.na          min          max
##  58.00000000  48.00000000   0.00000000  -1.00000000   1.00000000
##        range          sum       median         mean      SE.mean
##   2.00000000  -2.00000000   0.00000000  -0.03448276   0.05480824
## CI.mean.0.95          var      std.dev     coef.var
##   0.10975160   0.17422868   0.41740709 -12.10480548
```

```
reshaped.data$cfinfour <- reshaped.data$finfour.1993 - reshaped.data$finfour.1992
print(quantile(reshaped.data$cfinfour, probs = c(0.01, 0.05, 0.1, 0.25,
    0.5, 0.75, 0.9, 0.95, 0.99, 1)))
```

```
##    1%    5%   10%   25%   50%   75%   90%   95%   99%  100%
## -1.00  0.00  0.00  0.00  0.00  0.00  0.00  0.15  1.00  1.00
```

```
print(stat.desc(reshaped.data$cfinfour))
```

```
##      nbr.val     nbr.null       nbr.na          min          max
##  58.00000000  53.00000000   0.00000000  -1.00000000   1.00000000
##        range          sum       median         mean      SE.mean
##   2.00000000   1.00000000   0.00000000   0.01724138   0.03882250
## CI.mean.0.95          var      std.dev     coef.var
##   0.07774072   0.08741682   0.29566335  17.14847443
```

Question4

a. Additional assumptions needed for the population model to be causal

There are 2 additional assumptions that are needed for the proposed population model to be causal:

Assumption 1: Exogeneity of the dependent variables in the model with the error term in the model. Exogeneity can be defined as the requirement that the dependent variables are not related to the error term. The presence of omitted variables that are correlated with any of the variables in the model creates a bias in the estimation of the coefficients of the variables selected in the model, hence this assumption is important.

Assuming the population model presented:

$$lapps_i = \gamma_0 + \beta_0 I_{1993} + \beta_1 bowl_{it} + \beta_2 btitle_{it} + \beta_3 finfour_{it} + a_i + u_{it}$$

Exogeneity of the variables in the model can be formulated as:

$$Cov(bowl_{it}, a_i + u_{it}) = Cov(btitle_{it}, a_i + u_{it}) = Cov(finfour_{it}, a_i + u_{it}) = 0$$

Assumption 2: Manipulation. We must assume that we have the ability to make manipulations to at least one of the x terms (the independent variables, but only one at a time) in order to observe changes in the y term (the dependent variable), without affecting the error term, so that we can estalish that there is no bias introduced. In mathematical terms we can say that if we define our changes in x as δx and our error term as ϵ, then we have:

$$\delta x = f(x), and\ Cov(\delta x, \epsilon) = 0$$

Without this assumption, then we are talking about correlation and not causality. However, the assumption regarding manipulaation does not seem plausible as it is hard to think of a situation where we can make a school appear in the final four / win a title / be champions in the previous year without affecting other variables included in the error term.

b. Additional assumption needed for OLS to consistently estimate the first-difference model

For OLS to consistently estimate the first difference model, we need to make two additonal assumptions:

Assumption 1: Exogeneity as explained in part a.

Assuming the difference model presented:

$$clapps_i = \beta_0 + \beta_1 cbowl_i + \beta_2 cbtitle_i + \beta_3 cfinfour_i + a_i + u_i$$

Exogeneity of the variables in the model can be formulated as:

$$Cov(cbowl_i, a_i + u_i) = Cov(cbtitle_i, a_i + u_i) = Cov(cfinfour_i, a_i + u_i) = 0$$

Assumption 2: As long as the effects of the omitted variables from the population model are constant, they don't affect the differences between the independent x variables from time 1 to time 2, so they will not affect the coefficients in the difference model. If the effects of the omitted variables are not constant, then there will still be omitted-variable bias in the difference model. So the assumption that we have to make for the difference model is that the effects of the omitted variables are constant over time.

Question5

Model Analysis

- **F-statistic** - The F-statistic for the model has a p-value of 0.03855, which is significant at the 0.05 level.
- **intercept** - The coefficient for the intercept is 0.01684 indicating that the year over year incease in application is 1.6% from 1992 to 1993. However, the t-statistic for that coefficient has a value of 0.1932 and is not significant at the 0.05 level.
- **cbowl** - The coefficient for the cbowl variable is .057, indicating that a win in a bowl the year prior, contributes to a 5.7% increase in applications the subsequent year. The t-statistic for the coefficient is significant at the 0.05 level, with a value of 0.0236.

- **cbtitle** - The coefficient for the cbtitle variable is .041, indicating that a win in the men's conference championship in the previous year, translates into an increase of 4.1% year over year from 1992 to 1993, however we cannot be sure of this relationship given the lack of statistical significance. The t-statistic for the coefficient is 0.1950 and is not significant at the 0.05 level.
- **cfinfour** - The coefficient for the cfinfour variable is -0.06961 and seems to indicate that an appearance in the men's final four the year prior is related to a decrease of 6.9% of applications, however we cannot be sure of this relationship given the lack of statistical significance. The t-statistic for the coefficient is 0.1348 and is not significant at the 0.05 level.

```
# Create the model.
change.lapps.model <- lm(clapps ~ cbowl + cbtitle + cfinfour, data = reshaped.data)
# Estimate the model
print(summary(change.lapps.model))
```

```
##
## Call:
## lm(formula = clapps ~ cbowl + cbtitle + cfinfour, data = reshaped.data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.192965 -0.042868 -0.006367  0.040005  0.283578
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.01684    0.01278   1.318   0.1932
## cbowl        0.05702    0.02448   2.329   0.0236 *
## cbtitle      0.04148    0.03161   1.312   0.1950
## cfinfour    -0.06961    0.04585  -1.518   0.1348
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09674 on 54 degrees of freedom
## Multiple R-squared:  0.1428, Adjusted R-squared:  0.09513
## F-statistic: 2.998 on 3 and 54 DF,  p-value: 0.03855
```

Question6

The F-statistic obtained from the model is the test of overall significance of the model. We have already establish that it has a p-value of 0.03855, which is significant at the 0.05 level. We can obtain the same statistic with the linearHypothesis function.

Though the model is significant, we do not want to read too much into the t-statistics of each of the coefficients of the model. We had noted that 3 out of 4 coefficients had p-values that were not significant at the 0.05 level. It appears that the combined explanatory power of the model is still relevant, as evidenced by the F-statistic.

Alternate second para (Rohan): RT: Since the model is significant, with an R squared value of 14%, we can infer that the independent variables do have some explanatory power over the dependent variable. However, we cannot be confident of relationships between the variables due to high standard errors and a lack of statistical significance of the coefficients.

```
linearHypothesis(change.lapps.model, c("cbowl", "cbtitle", "cfinfour"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## cbowl = 0
## cbtitle = 0
## cfinfour = 0
##
## Model 1: restricted model
## Model 2: clapps ~ cbowl + cbtitle + cfinfour
##
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1     57 0.58953
## 2     54 0.50537  3   0.08416 2.9976 0.03855 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```