Is language Spoken at Home Independent of Size of Household?

Peter Andrade, Paul Rozario, Ramseii Welton, Daniel Deneau

7/10/2021

```
load(file="/Users/Danny/Downloads/ACS_2017_MD.Rdata")
```

Section I Research Question:

Customs and beliefs can be shared and spread throughout members of a household, in order to preserve tradition and a culture's unique background and history. In the 2017 American Community Service Survey data for Maryland, social and demographic variables were recorded in order to gain certain political, cultural, financial, etc. standpoints of individuals within a household. The question we want to review is "Is language spoken at home independent of the size of the household". Numerous families and roommates need to be able to communicate with one another in order to understand each other's wants and needs so that they can live with one another. While some would argue that having a larger household size would influence the use of just one language that the family or roommates are all fluent in, this is not the case since the use of language really only depends on that person. As in, it is their choice to use or study a certain language. In this paper we would like to use both variables NP (Number of persons in the household) and LANP (Language spoken at home), to put together a list of subsets and graphs to determine if language can be independent of size of household.

Section II Data Exploration:

We would like to get an idea of how the number of people in a household might be related to the language spoken by that household. In order to do so, we want to clear up any missing observations in our data. This particular data set has many more recorded households speaking Spanish (in our data, LANP = 1200) than the other languages in the dataset. Therefore we also remove those observations as well in order to account for any bias they might cause.

```
load(file="/Users/Danny/Downloads/ACS_2017_MD.Rdata")
languages_full <- ACS_2017_MD$LANP # Full language set, no values removed
number_of_persons_full <-  ACS_2017_MD$NP # Full number of persons set,

for (row in 1:length(number_of_persons_full)) {
  a <- languages_full[row]
  if(is.na(a)){
    number_of_persons_full[row] <- NA
  } else if(languages_full[row] == 1200){
    languages_full[row] <- NA
    number_of_persons_full[row] <- NA
  }
}
languages <- na.omit(languages_full) # USE THIS. (Cleaned up language data)
```

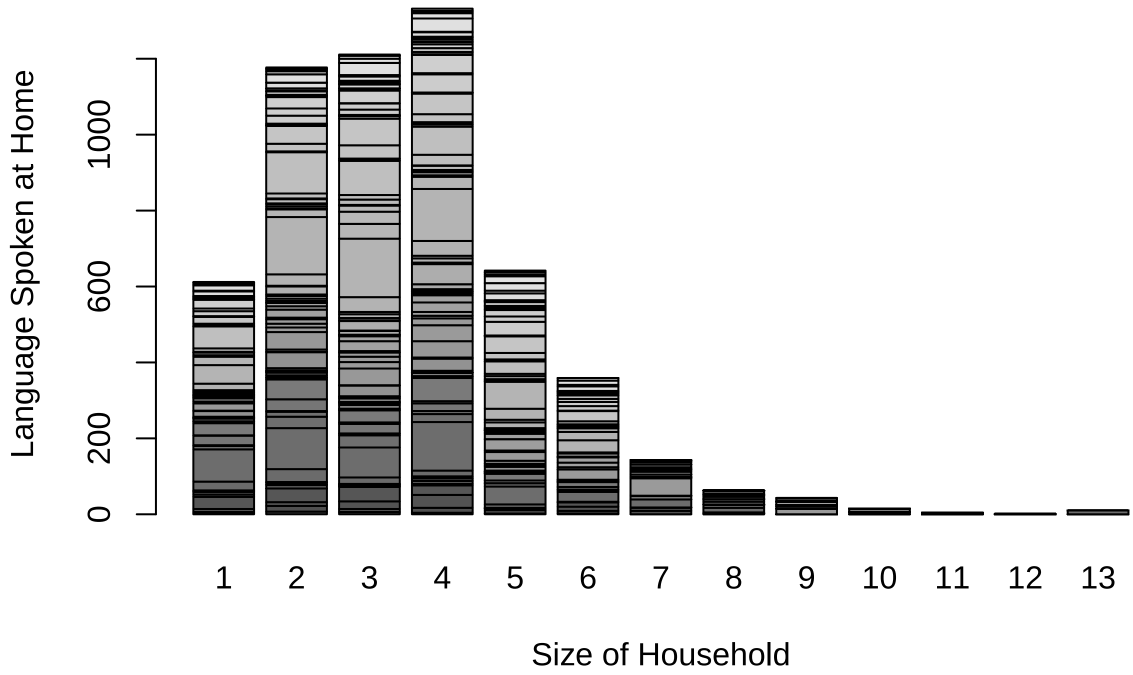
```
number_of_persons = na.omit(number_of_persons_full) # USE THIS (Cleaned up number of persons data)
np_lanp_data = data.frame(languages, number_of_persons) ## Use this as our data set instead
head(np_lanp_data)
```

```
##   languages number_of_persons
## 1      1250                 5
## 2      1250                 5
## 3      1250                 5
## 4      1000                 2
## 5      4590                 3
## 6      1292                 5
```

Section III ANOVA Analysis Technique:

Some output that might interest us would be a barplot showing a distribution of languages across household group sizes, and the related table and summary. We see that the data appears roughly normal, with a left skew. Indicating that many more smaller households were observed, with a median around a size of four people.

```
## Summary statistics and tables
tbl <- table(np_lanp_data$languages, np_lanp_data$number_of_persons)
barplot(tbl, xlab="Size of Household", ylab= "Language Spoken at Home")
```



```
head(table(np_lanp_data$languages, np_lanp_data$number_of_persons))
```

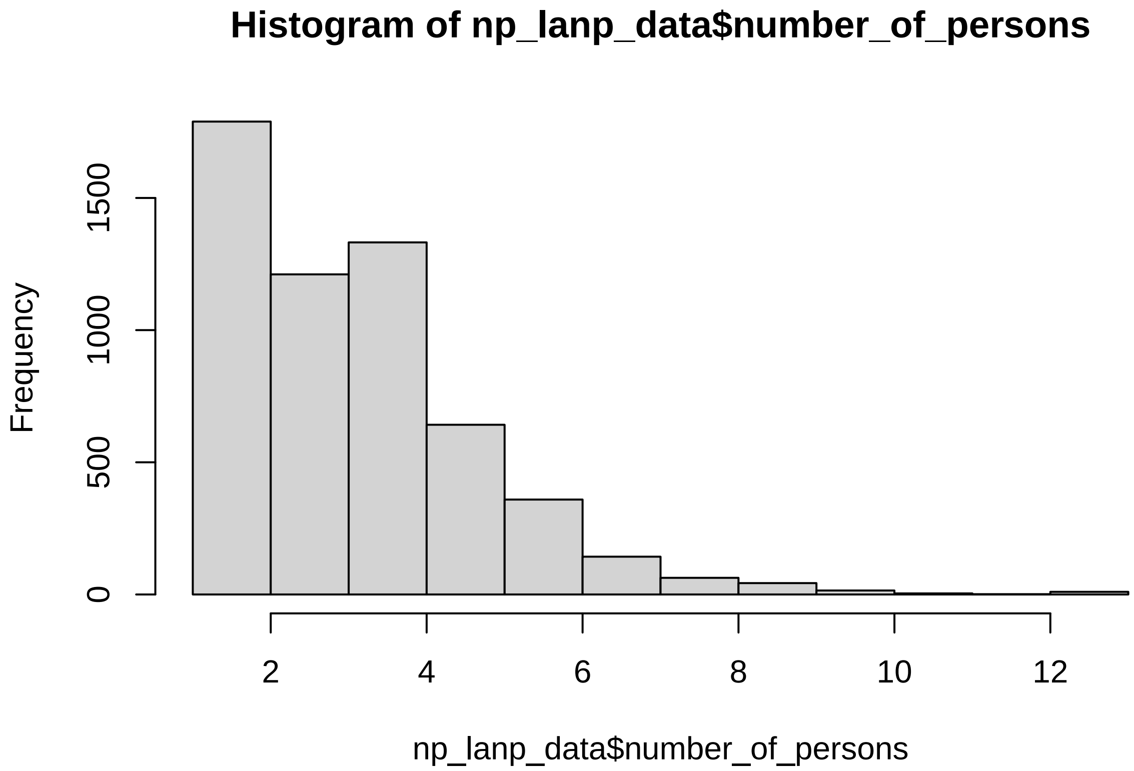
```
##
##         1  2  3  4  5  6  7  8  9 10 11 12 13
##    1000  3  7  6  4  1  2  0  0  0  0  0  0  0
##    1025  3 15  8 13  1  7  9  2  0  0  0  0  0
##   1055  8 10 20 34  9 11  8  0  0  0  0  0  0
##   1110 32 36 38 25  5 12  1  1  0  0  0  0  0
##   1125  0  0  2  4  0  0  0  0  0  6  0  0  0
##   1130  6  0  0  0  0  0  0  0  0  0  0  0  0
```

```
summary(table(np_lanp_data$languages, np_lanp_data$number_of_persons))
```

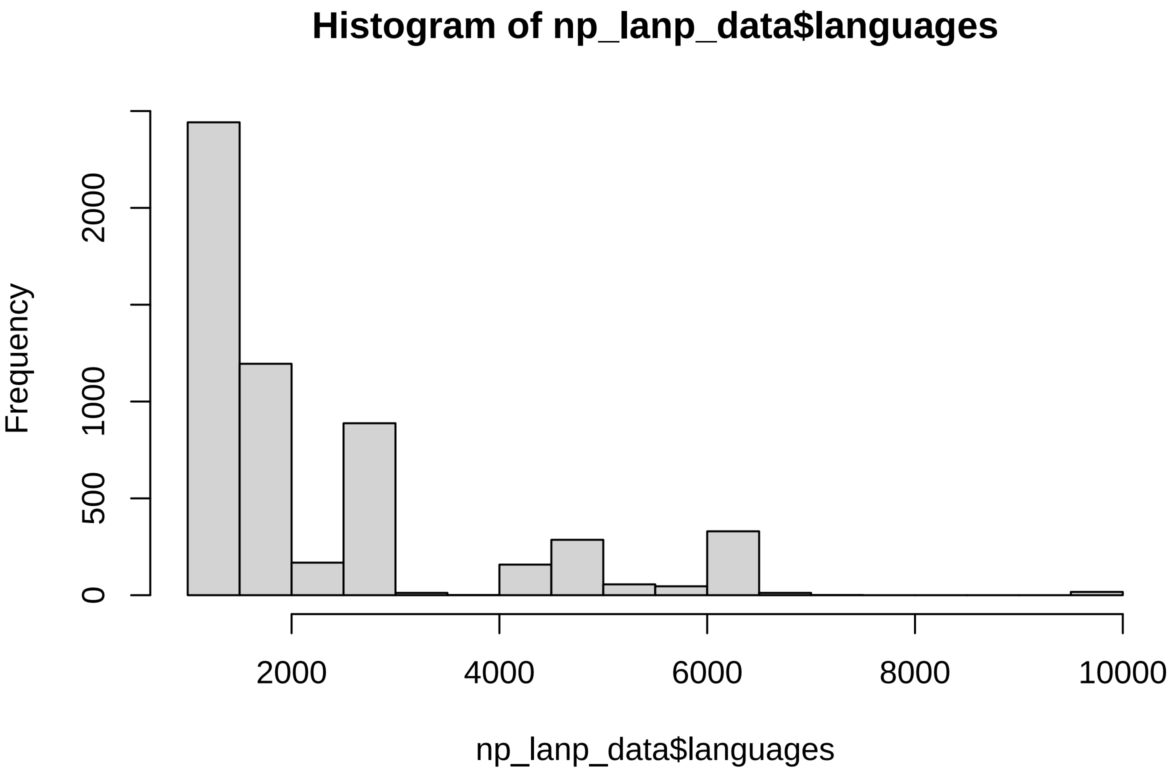
```
## Number of cases in table: 5612
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 4426, df = 1248, p-value = 0
##  Chi-squared approximation may be incorrect
```

It might be helpful to visualize histograms of our two variables in question independently. Both of them appear left skewed, with the number of people having less variance than the languages, which makes sense given that there are much more languages spoken than there are categories for number of people. When we randomly sample from each variable and plot the distribution, we see a normal distributions. The scatter plot does not give us much helpful information since we are comparing categories with categories, but it does show a trend towards less variability in languages as the household size increases. The QQ-Plot also almost follows a linear trend, which is weakened by many influential points. Finally, the box plot comparisons variance across household size groups shows that the variance is roughly the same across the board, with a few groups with larger language variances.

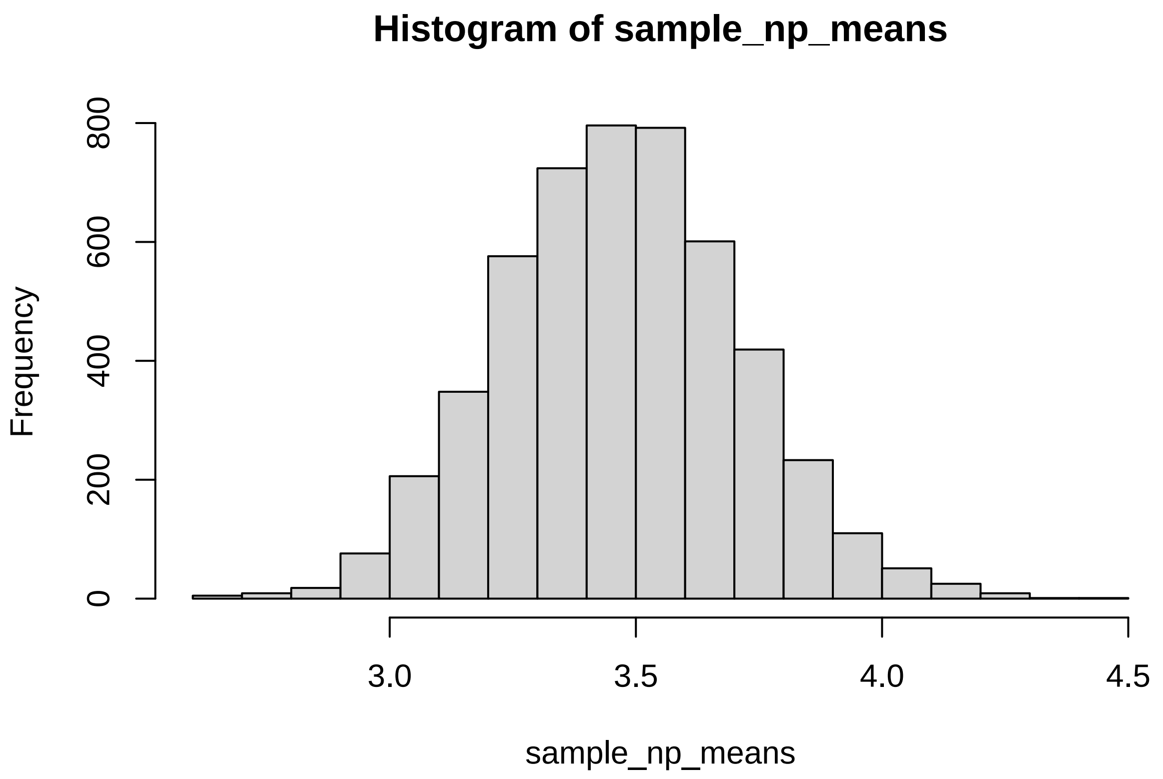
```
## Plots
hist(np_lanp_data$number_of_persons)
```



```
hist(np_lanp_data$languages)
```



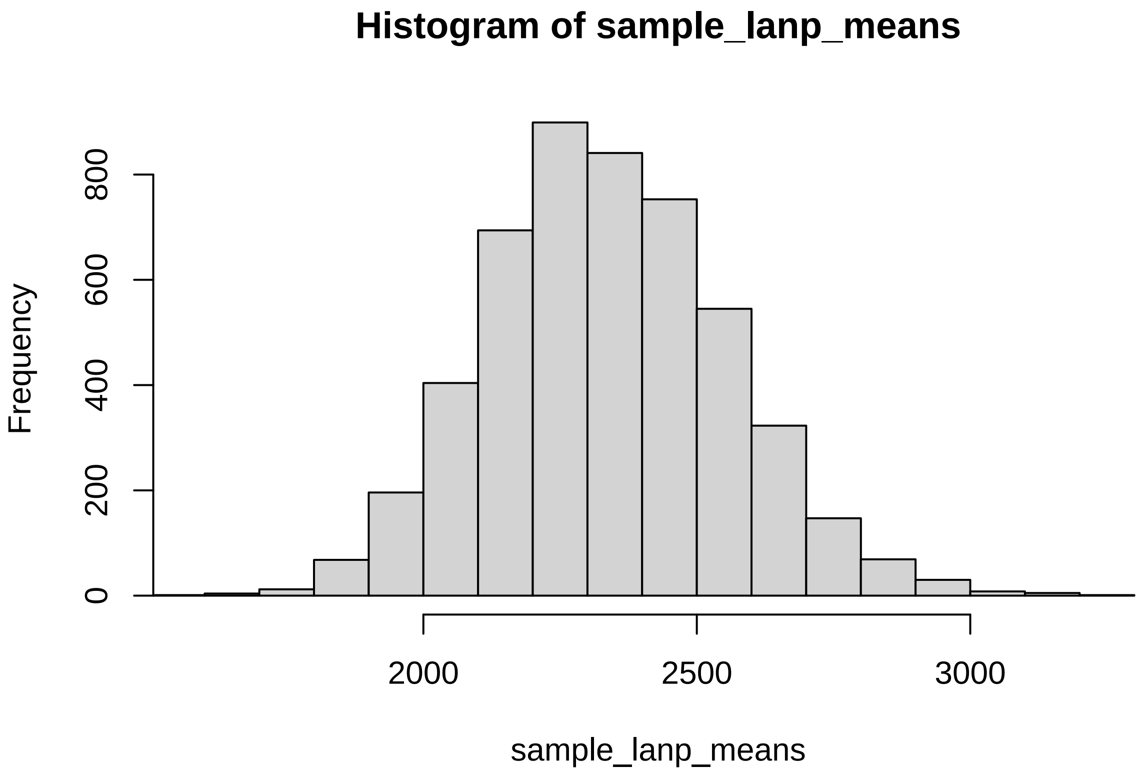
```
sample_np_means <- rep(NA, 5000)
for(i in 1:5000){
  samp = sample(np_lanp_data$number_of_persons, 50)
  sample_np_means[i] = mean(samp)
}
np_samp_hist <- hist(sample_np_means, breaks = 20, xlim = range(sample_np_means))
```



```r
np_samp_hist
```

```
## $breaks
##  [1] 2.6 2.7 2.8 2.9 3.0 3.1 3.2 3.3 3.4 3.5 3.6 3.7 3.8 3.9 4.0 4.1 4.2 4.3 4.4
## [20] 4.5
##
## $counts
##  [1]   5   9  18  76 206 348 576 724 796 792 601 419 233 110  51  25   9   1   1
##
## $density
##  [1] 0.010 0.018 0.036 0.152 0.412 0.696 1.152 1.448 1.592 1.584 1.202 0.838
## [13] 0.466 0.220 0.102 0.050 0.018 0.002 0.002
##
## $mids
##  [1] 2.65 2.75 2.85 2.95 3.05 3.15 3.25 3.35 3.45 3.55 3.65 3.75 3.85 3.95 4.05
## [16] 4.15 4.25 4.35 4.45
##
## $xname
## [1] "sample_np_means"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

```r
sample_lanp_means <- rep(NA, 5000)
for(i in 1:5000){
  samp = sample(np_lanp_data$languages, 50)
 sample_lanp_means[i] = mean(samp)
}
lanp_samp_hist <- hist(sample_lanp_means, breaks = 20, xlim = range(sample_lanp_means))
```

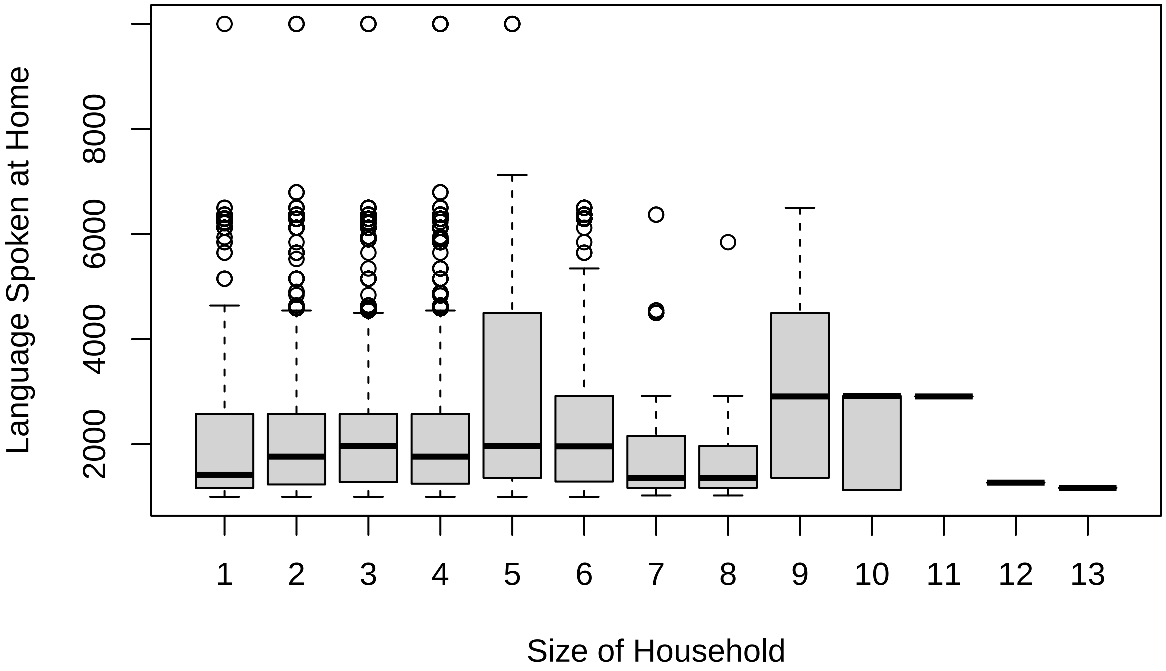


```
lanp_samp_hist
```

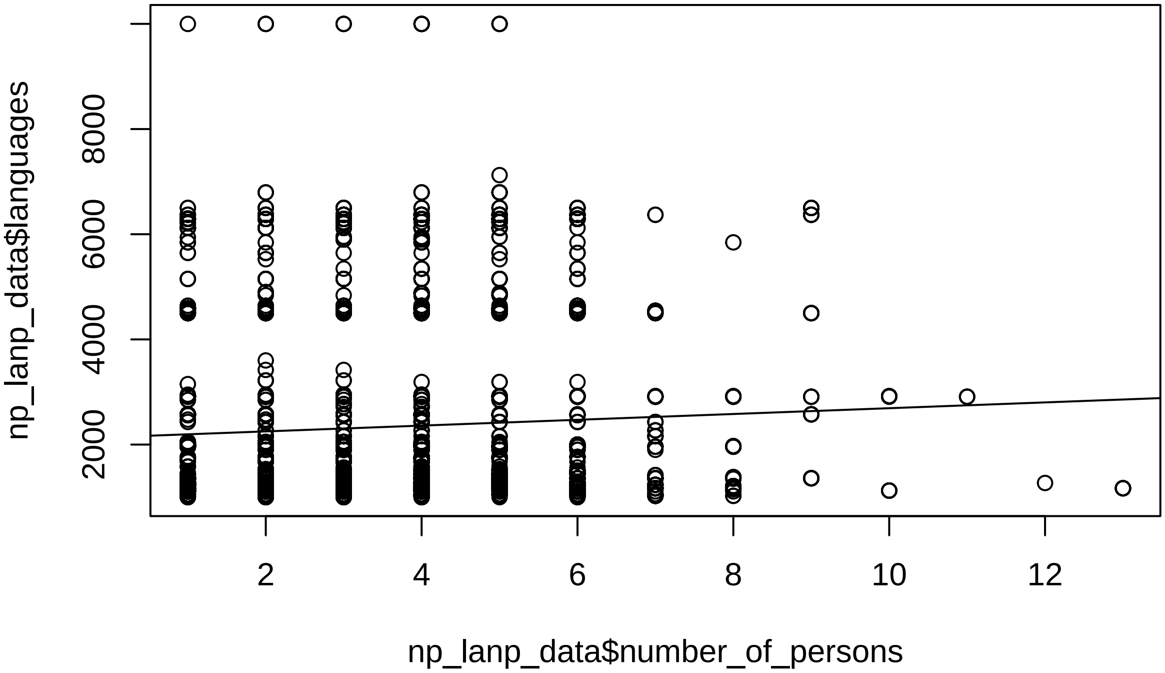
```
## $breaks
##  [1] 1500 1600 1700 1800 1900 2000 2100 2200 2300 2400 2500 2600 2700 2800 2900
## [16] 3000 3100 3200 3300
##
## $counts
##  [1]   1   4  12  68 196 404 694 899 841 753 545 323 147  69  30   8   5   1
##
## $density
##  [1] 0.000002 0.000008 0.000024 0.000136 0.000392 0.000808 0.001388 0.001798
##  [9] 0.001682 0.001506 0.001090 0.000646 0.000294 0.000138 0.000060 0.000016
## [17] 0.000010 0.000002
##
## $mids
##  [1] 1550 1650 1750 1850 1950 2050 2150 2250 2350 2450 2550 2650 2750 2850 2950
## [16] 3050 3150 3250
##
## $xname
## [1] "sample_lanp_means"
##
## $equidist
## [1] TRUE
```

```
##
## attr(,"class")
## [1] "histogram"
```

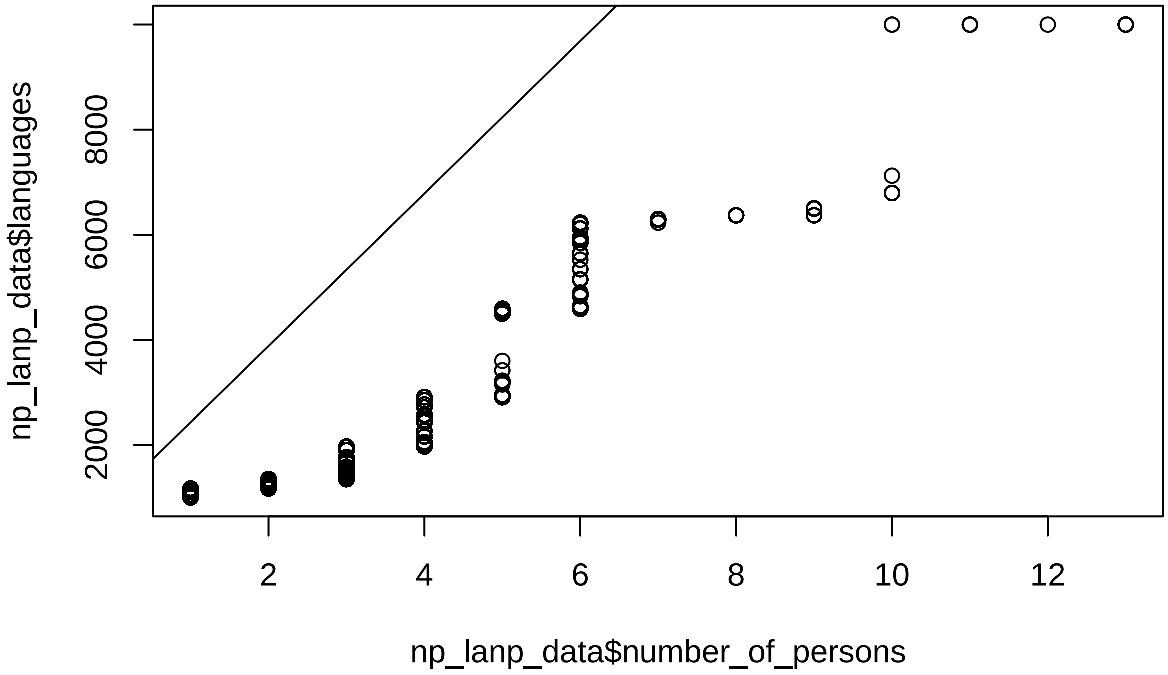
```
boxplot(np_lanp_data$languages ~ np_lanp_data$number_of_persons, xlab="Size of Household", ylab= "Langua
```



```
model <- lm(np_lanp_data$languages ~ np_lanp_data$number_of_persons, data = np_lanp_data)
plot(np_lanp_data$languages ~ np_lanp_data$number_of_persons)
abline(model)
```



```
qqplot(np_lanp_data$number_of_persons, np_lanp_data$languages)
qqline(np_lanp_data)
```



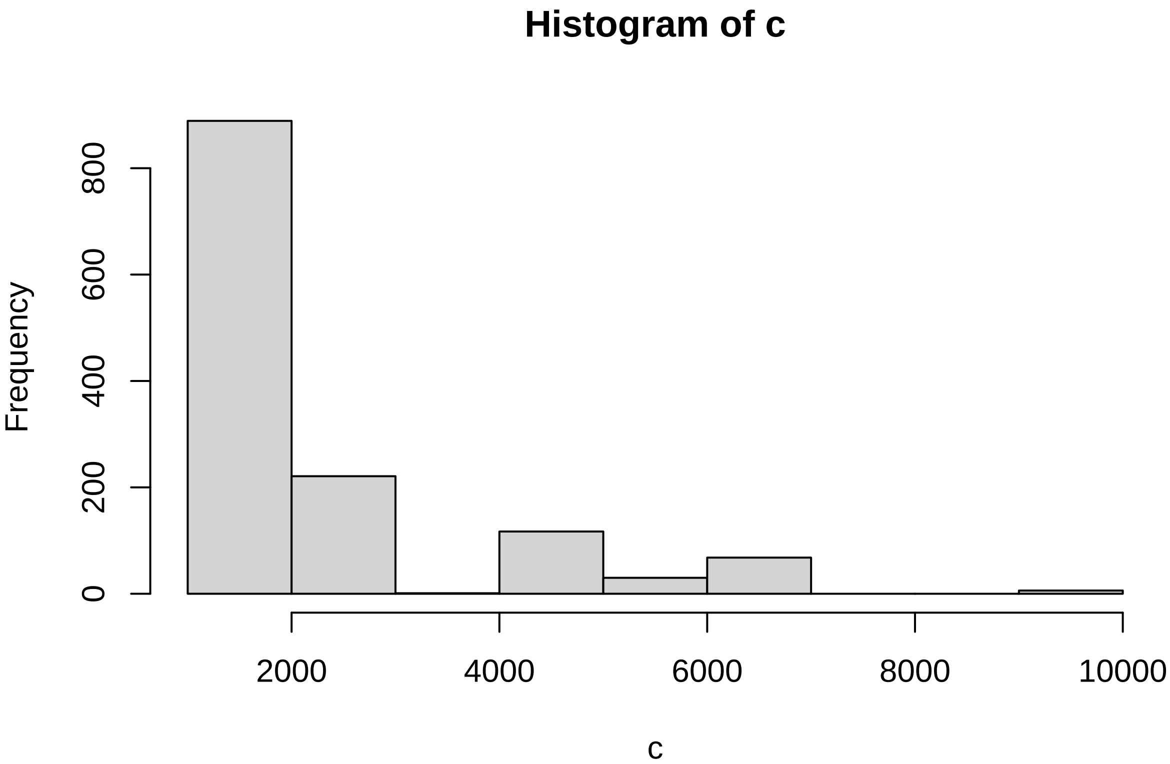
Section IV Interpretation and Discussion of Analysis Results:

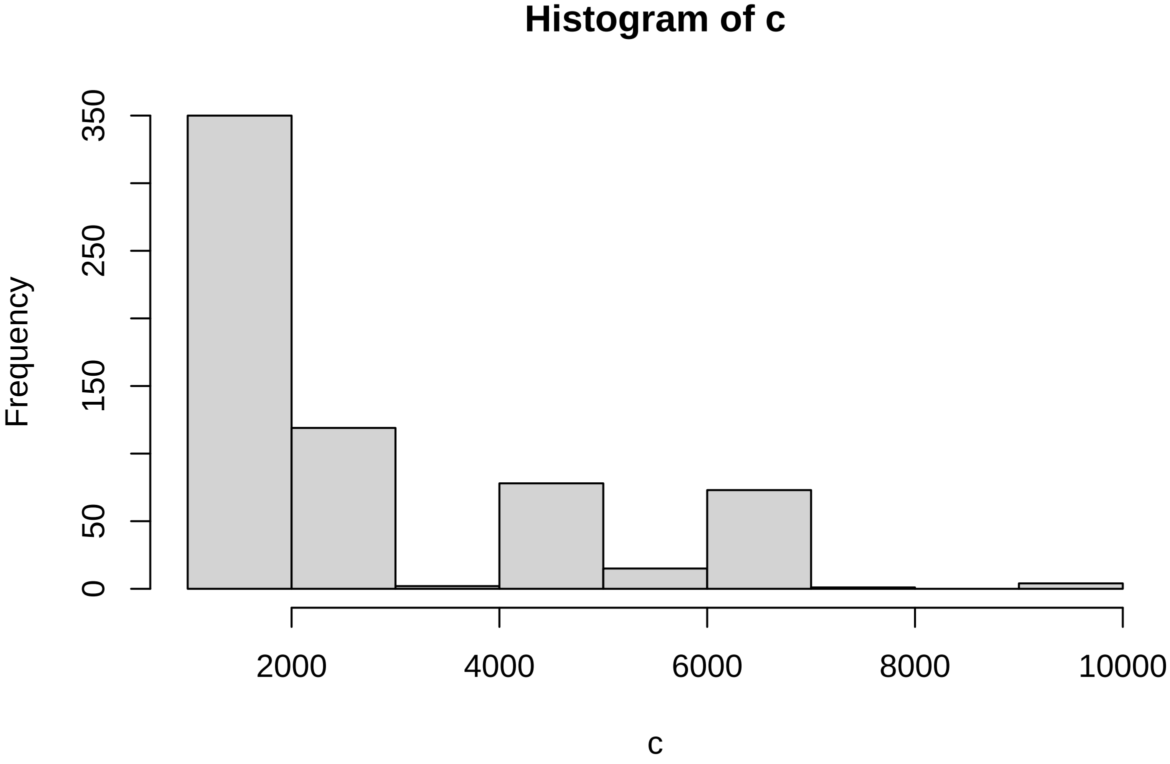
The three conditions we must meet in order to run this analysis is, first, independence. We know that our data comes from a large public database, free of biases in data observation, collected through random sampling.

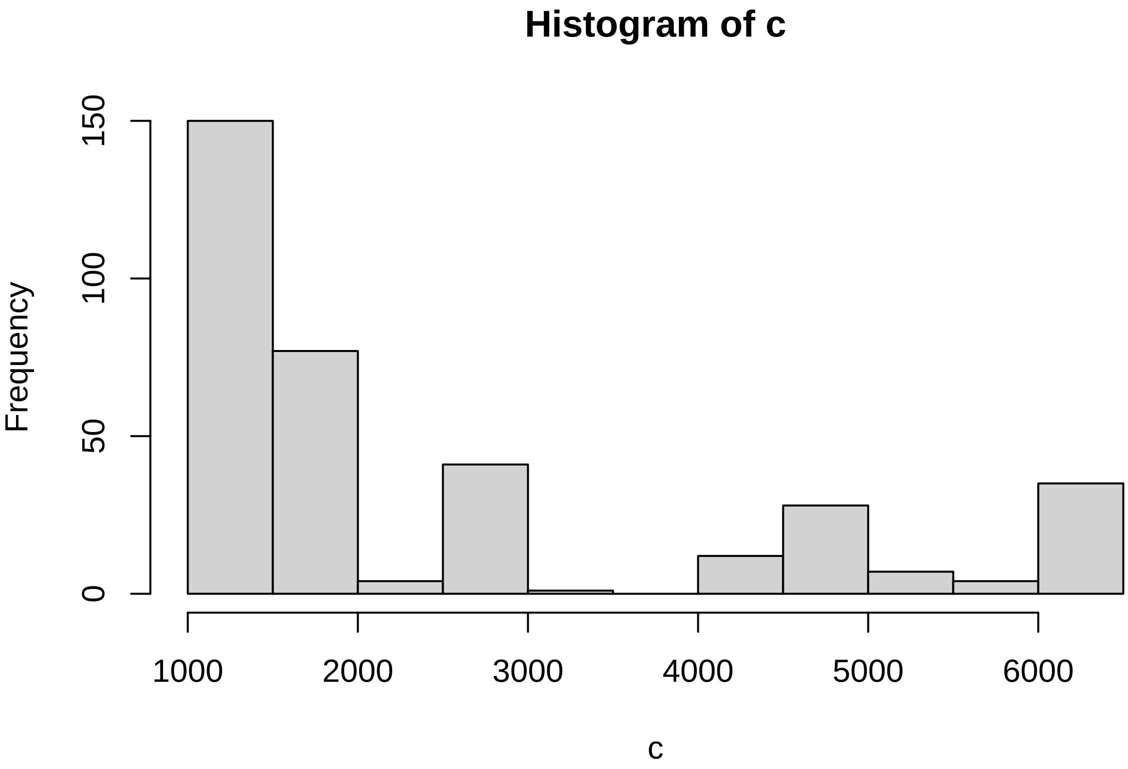
The second conditon is that each group has an approximately normal distribution, without any major outliers. To check this, we pick the 4th through 7th groups since they are most like the other data. Since our data set is relatively large, and due to the context of language groups being the bins, we see that each of our distributions is skewed to the left. Without any major outliers, we can be satisfied with this for the condition. Finally, we want to revist the side by side boxplots for each group. We see that the variance is roughly equal among the most common size groups from 1 to 8, with a few groups with higher or lower variance. Overall, the condition is satisfied.

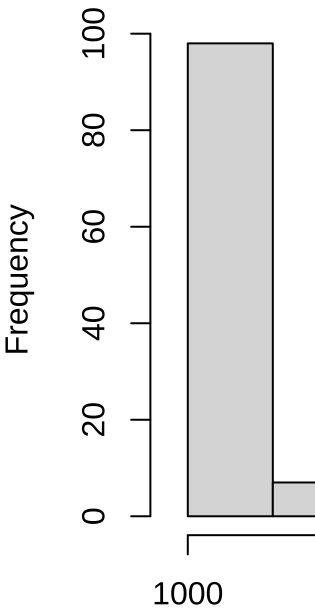
With all conditions satisfied, we may proceed with the ANOVA test.

```
for(o in 4:7) {
    c <- c(NA)
    for(i in 1:length(np_lanp_data$number_of_persons)) {
      if( np_lanp_data$number_of_persons[i] == o) {
        c[i] = np_lanp_data$languages[i]
      }
    }
    c <- na.omit(c)
    hist(c)
}
```

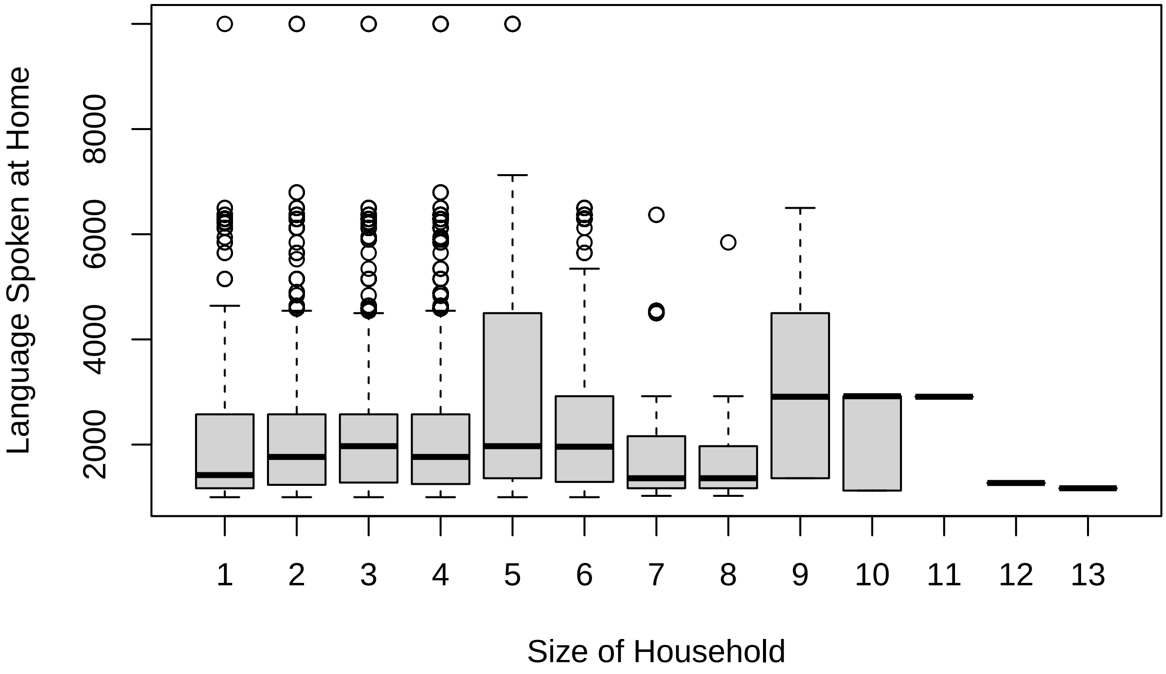

Histogram of c
Frequency
0
200
400
600
800
2000
4000
6000
8000
10000
c


Histogram of c
Frequency
0
50
150
250
350
2000
4000
6000
8000
10000
c





```
boxplot(np_lanp_data$languages ~ np_lanp_data$number_of_persons, xlab="Size of Household", ylab= "Langu
```



In order to determine if our observations have varying means due to chance, we use a summary command with the ANOVA analysis to see important values such as sum of squared error, mean squared error, and the F value 22.28, for the number of persons variable. The table also shows us the probability of observing data like this which is 2.41e-06 and highly unlikely. With these results, and a significance level of 0.05 (or 3.3 using the Bonferroni correction) , we can conclude that observed differences in means of language spoken are statistically significant.

```
analysis <- aov(formula = np_lanp_data$languages ~ np_lanp_data$number_of_persons, data = np_lanp_data)
summary(analysis)
```

```
##                                   Df    Sum Sq  Mean Sq F value   Pr(>F)    
## np_lanp_data$number_of_persons    1 5.153e+07 51534233   22.28 2.41e-06 ***
## Residuals                      5610 1.297e+10  2312595                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
a <- 0.05 * (12*11)/2
a
```

```
## [1] 3.3
```

Section V Conclusion:

Our results show us that the observed differences in the means for language spoken across the household size groups is likely not simply due to chance. This suggests that language spoken at home is not independent of household size. This is surprising, as we originally did expect language to be related to household size. It would be interesting to run the analysis other times with other major languages excluded to see how that affects the numbers. Since this was done without Spanish and English being considered, we could also try without Russian as well since that appears a lot in the data.