Test of Homogeneity ($\chi^2$):
Are different American races/ethnicities educated differently in our present society? Possibly a chi-squared test of homogeneity will shed some light.

The data file `race_educ.csv` was developed to investigate this question. This rather extensive data set contains information on the amount of formal education Americans of various racial origin obtained around the year 2000. A table of this information is displayed below.

<table>
<thead>
<tr>
<th></th>
<th>Advanced Degree</th>
<th>College Grad</th>
<th>H.S. Grad</th>
<th>Not an H.S. Grad</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>117</td>
<td>549</td>
<td>1598</td>
<td>263</td>
<td>2527</td>
</tr>
<tr>
<td>Hispanic</td>
<td>99</td>
<td>412</td>
<td>1269</td>
<td>1031</td>
<td>2811</td>
</tr>
<tr>
<td>Other</td>
<td>197</td>
<td>305</td>
<td>341</td>
<td>66</td>
<td>909</td>
</tr>
<tr>
<td>White</td>
<td>1127</td>
<td>4725</td>
<td>6429</td>
<td>810</td>
<td>13091</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1540</td>
<td>5991</td>
<td>9637</td>
<td>2170</td>
<td>19338</td>
</tr>
</tbody>
</table>

Our research question is: Did all races have equitable formal educational experiences, given natural variability? We will use the $\chi^2$ test of homogeneity to come up with an initial answer to this question. The Chi-squared statistic formula is shown below.

$$\chi^2 = \sum \frac{\text{observed counts} - \text{expected counts}}{\text{expected counts}}$$

I usually structure my 2-way table `.csv` file with first column being y (response) variable categories, second column being x (explanatory) variable categories, with last column being the cell counts. See below.

Let us use the following code?
# test of homogeneity

data1 <- read.csv("race_educ.csv")
table1 <- xtabs(count ~ race+school, data=data1)
addmargins(table1)
spineplot(table1, main="RACE vs SCHOOL")
chisq.test(table1)

I used the `spineplot()` command for the graph. Results are shown below.

Below is the relevant output, showing that there is a strong relationship (because of the low p-value in the test) between RACE and EDUCATION.

```
Chi-Squared Test of Independence:
Students in a large Statistics class at a university in the Northeast were asked to describe their political position and whether they had engaged in binge drinking.
```

![Spine plot of RACE vs SCHOOL](image)

Chi-Squared Test of Independence:
Students in a large Statistics class at a university in the Northeast were asked to describe their political position and whether they had engaged in binge drinking.
drinking (5 drinks at a sitting for a man; 4 for a woman). Below is a 2-way table of the results of this study.

We will do a Chi-Squared test of independence. This is a test of independence because we have only one group we are testing under 2 variables —so, we are testing the independence of the two variables. The last example we did was a test of homogeneity (of races with educational opportunities) because we were comparing many groups (i.e., the different races) for uniformity of education.

### Table: Binge Drinking vs. Political Position

<table>
<thead>
<tr>
<th></th>
<th>Conservative</th>
<th>Moderate</th>
<th>Liberal</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>In past 3 days</td>
<td>7</td>
<td>19</td>
<td>14</td>
<td>40</td>
</tr>
<tr>
<td>In past week</td>
<td>9</td>
<td>19</td>
<td>28</td>
<td>56</td>
</tr>
<tr>
<td>In past month</td>
<td>3</td>
<td>8</td>
<td>23</td>
<td>34</td>
</tr>
<tr>
<td>Some other time</td>
<td>8</td>
<td>15</td>
<td>22</td>
<td>45</td>
</tr>
<tr>
<td>Never</td>
<td>9</td>
<td>22</td>
<td>20</td>
<td>51</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>36</strong></td>
<td><strong>83</strong></td>
<td><strong>107</strong></td>
<td><strong>226</strong></td>
</tr>
</tbody>
</table>

Let us use the script below to make a test of independence for BINGE DRINKING vs POLITICS

```r
# test of independence
data2 <- read.csv("binge.csv")
table2 <- xtabs(count ~ binge+politics, data=data2)
addmargins(table2)
spineplot(table2, main="RACE vs SCHOOL")
chisq.test(table2)
```
HW#1: What do college graduates do after graduating? The 2-way-table below lists the response variable CHOICE (categories- Employed, Grad, and Other) and explanatory variable MAJOR (categories- Agric, ArtsScience, Engin, SocSci), where the data was collected in 2006 at respective universities in the United States. Perform a chi-squared test of homogeneity on this information, to see how Homogenous CHOICE and MAJOR are. In your results show tables with marginal row/column, segmented bar graphs, and chi-squared statistics, and give about a 50 word conclusion, using proper protocol and English sentences, backing up your conclusion.

<table>
<thead>
<tr>
<th>CHOICE</th>
<th>MAJOR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Agriculture</td>
</tr>
<tr>
<td>Employed</td>
<td>379</td>
</tr>
<tr>
<td>Grad School</td>
<td>186</td>
</tr>
<tr>
<td>Other</td>
<td>104</td>
</tr>
</tbody>
</table>

HW#2: A survey was done at the University of Montana a few years back which was used as a major text book homework problem (STATS, DATA and MODELS, 3rd ED, DeVeaux, Velleman, and Bock, ISBN 13:978-0-321-69260-3, p661), which asked students from various parts of Montana what their political affiliation was. The response variable is REGION (categories-West, Northeast, Southeast) and the explanatory variable is PARTY (categories-Dem, Repub, Ind). Perform a chi-squared test of independence on this information, to see how Independent REGION and PARTY are. In your results show tables with marginal row/column, segmented bar graphs, and chi-squared statistics, and give about a 50 word conclusion, using proper protocol and English sentences, backing up your conclusion.
<table>
<thead>
<tr>
<th>REGION</th>
<th>Democrat</th>
<th>Republican</th>
<th>Independent</th>
</tr>
</thead>
<tbody>
<tr>
<td>West</td>
<td>39</td>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>Northeast</td>
<td>15</td>
<td>30</td>
<td>12</td>
</tr>
<tr>
<td>Southeast</td>
<td>30</td>
<td>31</td>
<td>16</td>
</tr>
</tbody>
</table>

**Introduction To Linear Regression Analysis:**

We will now begin investigation, using R, of 2 quantitative variables—a technique called **linear regression**. Basically, we first construct a scatter plot, located usually in the first quadrant only (because real quantitative variables usually have values greater than 0), of points \((x, y)\) = (explanatory variable values, response variable values). Notice that the response is **always** on the y axis. Then we see if a **line-of-best-fit** can appropriately approximate the progression of points. If so, then the resulting **linear model** we end up with is of the form

\[
\hat{y} = \beta_0 + \beta_1 \cdot x + \epsilon
\]

where \(\hat{y}\) = estimated or predicted y value

\(\beta_0\) = the intercept of the linear model,

\(\beta_1\) = the slope of the linear model, and

\(\epsilon\) = the value of error term in our model approximation

The resulting line is the least squares regression line, the one where the sum of the y errors from our model at every point = 0. See below.

Notice that the error distances (under predictions) are positive values and error distances below the line (over predictions) are negative values.

Once we have a linear model (assuming that linear fit is appropriate) we can predict what the y value will be for a given x value, for any value within the x range of values where the model “works”. We have to be careful about predicting using our model for values too far above or below the range of x values used to create the model (called **extrapolation**) because we have no information that our model will be relevant at those “unexplored” x values.
Linear regression is a much used and valuable statistical tool. If we have a model that fits the points well (as echoed by a correlation coefficient close to either 1 or -1), where there is not much spread between the line and the points it fits, we can predict (or explain) \( y \) accurately, given just \( x \). A close fitting line shows a strong relationship between the 2 variables, and allows us to determine a value of \( y \) (which may be hard to measure directly) from a value of \( x \) (which may be much easier to measure). A line-of-best-fit which is not a good fit (resulting in a correlation coefficient close to 0) is less useful as a predictor.

The 2 parameters of slope and intercept (\( \beta_1 \) and \( \beta_0 \), respectively), are used by statisticians a bit differently. The slope (or rate of change of \( y \) with respect to \( x \)) is of much greater value to users of statistics than the intercept, usually. For example, I am much more interested in how the IBM stock is changing (rate of increase/decrease) than I am about its static value at this instant of time.

When we conduct the linear regression procedure to model the relationship between \( x \) and \( y \), we are doing a parametric test. Specifically, we assume that the errors \( \varepsilon_i \) are considered independent of each other, are approximately normally distributed, and have about the same variance across the range of \( x \). This is shown below visually.

---

Scatterplot and Line of “best fit”:

Put the .csv file exam_scores.csv on your Desktop.
This contains scores by 25 students in consecutive tests, from a past statistics class. Let us make a scatter plot, line of best fit, and do a linear regression of the information from scores.csv, using the script shown below.

```r
# Linear Regression Intro
# ------------------------------
data3 <- read.csv("exam_scores.csv")
linmodell <- lm(data3$Exam.2 ~ data3$Exam.1)
linmodell
summary(linmodell)
plot(data3$Exam.1, data3$Exam.2,
     main="2 Course Exam Scores",
     xlab="exam 1", ylab="exam 2",
     sub="exam2 = -4.09 + 0.95 exam1")
abline(linmodell)
```

The model and linear regression output are shown below.

![R Console Output](image)

Below is the resulting graph with line of best fit superimposed on it.

![Graph](image)
Let us analyze what we did a step at a time, since the \texttt{lm()} and \texttt{summary()} commands of R do a whole lot of things.

We first read in the data, 2 columns of Exam.1 and Exam.2 results. Then we use the \texttt{lm()} command to generate a linear model of Exam.2 vs Exam.1.

The \texttt{lm()} command, which we stored in label \texttt{linmodel1} generates all kinds of useful information for our linear model. When we just type the label, intercept $\beta_0$ (of -4.09) and slope $\beta_1$ (of 0.95) are produced. Notice that the \texttt{lm()} function worked on the formula $y \sim x$, which for us was Exam.2 $\sim$ Exam.1. If we type \texttt{summary()} of that label, we get the slope, intercept, 5 number summary of residuals, residual standard error, $R^2$ value, adjusted $R^2$ value, and significance statements of the $\beta$ parameters, based on $t$ and $F$ statistics. There is more information generated by the \texttt{lm()} command, but you have to specifically ask for it, using “extractor” functions. A partial list of these functions are shown below (the ones most useful for us now). We will elaborate on these things later on.

<table>
<thead>
<tr>
<th>Table 10.1 Extractor functions for the result of \texttt{lm()}</th>
</tr>
</thead>
<tbody>
<tr>
<td>\texttt{summary()}</td>
</tr>
<tr>
<td>\texttt{plot()}</td>
</tr>
<tr>
<td>\texttt{coef()}</td>
</tr>
<tr>
<td>\texttt{residuals()}</td>
</tr>
<tr>
<td>\texttt{fitted()}</td>
</tr>
<tr>
<td>\texttt{deviance()}</td>
</tr>
<tr>
<td>\texttt{predict()}</td>
</tr>
<tr>
<td>\texttt{anova()}</td>
</tr>
<tr>
<td>\texttt{AIC()}</td>
</tr>
</tbody>
</table>

Within the parentheses we would use argument \texttt{linmodel1} for our study.

Let us use \texttt{predict()} to predict Exam.2 scores from our 24 listed Exam.1 scores, using our model. See below.

```
R> predict(linmodel1)  
1     2     3     4     5     6     7     8
89.01046 68.11037 81.41043 86.16045 88.06045 80.46042 69.06038 85.21044
 9    10    11    12    13    14    15    16
79.51042 86.16045 78.56042 75.71040 82.36043 74.76040 74.76040 84.26044
 17    18    19    20    21    22    23    24
87.11045 89.01046 75.71040 68.11037 88.06045 64.31036 71.91039 66.21036
R> |
```

\textbf{Residual plot:}

A good visual way to check for linearity as well as relative uniformity of error variation is to construct a residual plot of our exam score data. See below for code used and results.
# Residual plot
# -----------------
plot(data3$Exam.1, resid(linmod1), main=
"residuals vs Exam.1 Scores")
abline(h=0,lty=2)

plot(predict(linmod1), resid(linmod1),
main="residuals vs Predicted Values")
abline(h=0,lty=3)

The 2 residual plots, with different line type horizontal axes, are shown below.

I produced both residual plots to show you their similarity. Our residual plot shows no stark pattern, and variability seems sort of consistent across the explanatory data range, so I would say that our linear fit model is appropriate here.

**Conclusion:**
Our residual plot shows that our linear model

\[
\text{Exam.2} = -4.09 + 0.95 \times \text{Exam.1} + \epsilon
\]

is appropriate, and our correlation coefficient \( r(R^2) = r(0.3601) = 0.6 \) (note that our line slope is positive, so our \( r \) is also positive). So we have a mildly moderate fit (with \( r = 0.6 \)), with only about 36\% (\( R^2 = 0.3601 \)) of our variability in Exam.2 scores accounted for with our linear model of Exam.2 on Exam.1. Our model predicts a 0.95 point increase in Exam.2 score for every 1 point increase in exam.1.

We will do some more analysis of regression characteristics in the next lab.

**HW#3:** Using the data set `fan_cost06.csv`, generated in 2006, construct a scatter plot, residual plot, and linear regression for the response variable BEER (beer price at ball parks) vs explanatory variable SODA (soda price at ball parks), where numbers are in units of dollars.
Report the summary information of the linear regression (using `summary()` in your lab report, and list your R code you used to generate the plots. Make about a 50 word proper conclusion (stating if a linear model is appropriate, and if so, how appropriate, with r, r^2, etc., evidence), using proper English sentences.

HW#4: Using the same data set `fan_cost06.csv`, perform the same regression analysis as HW#3 above for the response variable PARKING ($ to park per car) vs response variable ADULT (ticket price, $ each adult). Is a linear model appropriate?