Confidence Interval for Distribution Variance and Standard Deviation:
In STAT 451 Jon discussed how to compute confidence intervals of a single variance and standard deviation Graham ($\sigma^2$ and $\sigma$), as well as the confidence interval of the ratio of two variances ($\frac{\sigma_1^2}{\sigma_2^2}$ along with $\frac{\sigma_1}{\sigma_2}$) of 2 samples, using the chi-squared distribution and F statistic.

The formula for a 95% CI of a single sample $\sigma^2$ is:

$$\left( \frac{(n-1)s^2}{X_{low}^{95}}, \frac{(n-1)s^2}{X_{hi}^{95}} \right)$$

and that 95% CI of a double sample $\frac{\sigma_1^2}{\sigma_2^2}$ is:

$$\left( \frac{Flow_{95} \cdot s_1^2}{s_2^2} \frac{s_2^2}{s_1^2} \right)$$

where: $n$ = single sample size, $df = n - 1$ = degrees of freedom, $X_{low}^{95} = X_{0.025,df}$

$X_{hi} = X_{0.975,df}$, $s$ = single sample standard deviation, $s_1$ = the first of the 2 sample standard deviations, $s_2$ = the second of the 2 sample standard deviations, $Flow_{95} = F_{0.025,df1,df2}$ and $F_{hi} = F_{0.975,df1,df2}$, $df1 = n_1-1$, and $df2 = n_2-1$.

R Commands Needed:
For the R code which accomplishes these CI computations are the `chisq()` and `qf()` commands. See below for the help menu on these commands.
Example:

We will use the data set airline_stats.csv, which contains information about flight times at an airport. A picture of the partial data set is shown below.

<table>
<thead>
<tr>
<th>Year</th>
<th>Month</th>
<th>OnTime Arr</th>
<th>Late Arr</th>
<th>Cancelled</th>
<th>Diverted</th>
<th>%OnTime</th>
<th>%Late</th>
<th>%Cancelled</th>
<th>%Diverted</th>
<th>Duration</th>
<th>Flights</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>9</td>
<td>0.799</td>
<td>0.137</td>
<td>0.053</td>
<td>0.009</td>
<td>1.12</td>
<td>0.12</td>
<td>0.040</td>
<td>0.003</td>
<td>0.12</td>
<td>9.99</td>
</tr>
<tr>
<td>2003</td>
<td>1</td>
<td>0.664</td>
<td>0.189</td>
<td>0.125</td>
<td>0.050</td>
<td>1.72</td>
<td>0.20</td>
<td>0.125</td>
<td>0.050</td>
<td>0.20</td>
<td>19.17</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.439</td>
<td>0.774</td>
<td>0.092</td>
<td>0.124</td>
<td>2.34</td>
<td>0.23</td>
<td>0.092</td>
<td>0.124</td>
<td>0.23</td>
<td>19.49</td>
</tr>
<tr>
<td>2003</td>
<td>9</td>
<td>0.664</td>
<td>0.189</td>
<td>0.125</td>
<td>0.050</td>
<td>1.72</td>
<td>0.20</td>
<td>0.125</td>
<td>0.050</td>
<td>0.20</td>
<td>19.17</td>
</tr>
<tr>
<td>1999</td>
<td>6</td>
<td>0.659</td>
<td>0.171</td>
<td>0.145</td>
<td>0.015</td>
<td>1.74</td>
<td>0.23</td>
<td>0.145</td>
<td>0.015</td>
<td>0.23</td>
<td>19.08</td>
</tr>
<tr>
<td>2002</td>
<td>11</td>
<td>0.651</td>
<td>0.174</td>
<td>0.098</td>
<td>0.113</td>
<td>2.37</td>
<td>0.25</td>
<td>0.098</td>
<td>0.113</td>
<td>0.25</td>
<td>19.24</td>
</tr>
<tr>
<td>1997</td>
<td>11</td>
<td>0.650</td>
<td>0.177</td>
<td>0.133</td>
<td>0.037</td>
<td>1.75</td>
<td>0.24</td>
<td>0.133</td>
<td>0.037</td>
<td>0.24</td>
<td>19.20</td>
</tr>
<tr>
<td>2003</td>
<td>10</td>
<td>0.649</td>
<td>0.161</td>
<td>0.112</td>
<td>0.073</td>
<td>1.64</td>
<td>0.22</td>
<td>0.112</td>
<td>0.073</td>
<td>0.22</td>
<td>19.47</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.649</td>
<td>0.161</td>
<td>0.112</td>
<td>0.073</td>
<td>1.64</td>
<td>0.22</td>
<td>0.112</td>
<td>0.073</td>
<td>0.22</td>
<td>19.47</td>
</tr>
<tr>
<td>2001</td>
<td>11</td>
<td>0.647</td>
<td>0.146</td>
<td>0.100</td>
<td>0.046</td>
<td>1.70</td>
<td>0.24</td>
<td>0.100</td>
<td>0.046</td>
<td>0.24</td>
<td>17.57</td>
</tr>
<tr>
<td>2002</td>
<td>9</td>
<td>0.644</td>
<td>0.151</td>
<td>0.091</td>
<td>0.030</td>
<td>1.64</td>
<td>0.23</td>
<td>0.091</td>
<td>0.030</td>
<td>0.23</td>
<td>17.91</td>
</tr>
<tr>
<td>2004</td>
<td>9</td>
<td>0.637</td>
<td>0.152</td>
<td>0.061</td>
<td>0.020</td>
<td>1.57</td>
<td>0.20</td>
<td>0.061</td>
<td>0.020</td>
<td>0.20</td>
<td>15.94</td>
</tr>
<tr>
<td>2005</td>
<td>4</td>
<td>0.634</td>
<td>0.151</td>
<td>0.077</td>
<td>0.023</td>
<td>1.58</td>
<td>0.21</td>
<td>0.077</td>
<td>0.023</td>
<td>0.21</td>
<td>15.99</td>
</tr>
</tbody>
</table>

It includes percentages of arrivals at the airport on time, late, canceled, diverted, and the percentage of flights overtime, as well as the total number of flights during the specific month. I don't know which airport this is, except that it is a very large one.

Let's look at the “on-time arrival” statistics. We first want to construct a histogram of percentages on-time, noting where the mean of the distribution is.

```
# airport arrival times
# --------------
datal <- read.csv("airline_stats.csv")
ontimel <- datal[,3]
hist(ontimel, main="Percent On Time",
    xlab="percent")
abline(v=mean(ontimel), lty=2)
text(85,40,"dotted line is mean")
```

The output is shown below.
As we learned first semester, the first 2 lines of code are comments, the next line reads in the .csv data file from the desktop, which we made as the default directory chosen with the FILE CHANGE DIR command of R Console. The 4th line makes a variable called ontime1 from the 3rd column of data read in. The next line makes a histogram from the ontime1 variable, with the title shown and x label of percent, with default y label of frequency. The next line puts a vertical line at x = the value of the mean, which seems around about 78%. The last line is a text line centered at the location (85, 40) on the graph.

We see that the histogram seems slightly left skewed, almost normally distributed. Most of the data lie between 70% and 85%, which might be about 2 standard deviations from the mean, which is about + or - 8% to 10% from the 78%. So, we can estimate the sample standard deviation (s) to be about 4% to 5% or so.

Now let us find the 95% confidence interval (CI) of variance and standard deviation of the airport on-time distribution. We use the code shown below.

```r
var1 <- sd(ontime1)^2
n1 <- length(ontime1)
chilower <- qchisq(.025, df=n1-1)
chiupper <- qchisq(.975, df=n1-1)
lower1 <- (n1-1)*var1/chiupper
upper1 <- (n1-1)*var1/chilower
cat("CI of variance is from", lower1," to ",upper1,"\n")
cat("CI of std. dev. is from", sqrt(lower1), 
" to ",sqrt(upper1),"\n")
```

The R code input (red) and output (blue) are shown below.
We see that our 95% standard deviation CI goes from about 4.4% to 5.6% as we estimated, with variance CI going from about 19.3 to 31.3 (in percent squared).

Our code first finds the variance ($\text{var1}$) of our data $\text{ontime1}$. Note that our value of $n$ is $n1$, the length of the vector $\text{ontime1}$. We next find the lower and upper values of our 95% interval using the commands $\text{qchisq(.025, \ldots)}$ and $\text{qchisq(.75, \ldots)}$ commands, resulting in the lower and upper values $\text{lower1}$ and $\text{upper1}$, respectively. The last 2 $\text{cat()}$ commands print out these values, with text labeling them. Remember that when you use the $\text{cat()}$ command, you must end with the "\n" argument so you get a carriage return after printing each text line with value. Finally, note that we found the standard deviation CI values just by finding the square root of the respective variance values.

**HomeWork1:** Use the code shown above, and find the 90% CI for both $\sigma$ and $\sigma^2$ of the $\text{ontime}$ data of $\text{airline_stats.csv}$. Also make a $\text{qqnorm()}$ plot with the 45 degree line superimposed using the $\text{qqline()}$ command, to see how close to normal the data is to normal. Remember we did qq norm plots in one of the previous labs of STAT 457.

**95% CI of ratio of 2 sample variances:**

Let us now find the 95% CI of the variances and standard deviations of 2 distributions, the percentage on-time arrival flights compared with the on-time departure flights. Below is the code used, and the output with graphs shown. Note that we are computing the confidence interval of both variance and standard deviation ratios using the formulas given above (using $X^2$ and $F$ statistics), as well as adding a “shape curve” behind the histogram, so we can see how bell-shaped the histogram appears.

```r
# CI of ratio of variances sigma1/sigma2
# --------------------------------------
ontime.leave <- data1[,7]
var2 <- var(ontime.leave)
n2 <- length(ontime.leave)
hist(ontime.leave, main="Leave Ontime % normal density curve shadow",
  xlab="percentage", prob=TRUE)
lines(density(ontime.leave))
abline(v=mean(ontime.leave), lty=2)
text(78,.08,"dotted line is mean")
flower <- qf(.025,n1-1,n2-1)
fupper <- qf(.975,n1-1,n2-1)
cilower <- var1/var2 * flower
ciupper <- var1/var2 * fupper
cat("CI of variance ratio is ",cilower," to ",ciupper,"\n")
cat("CI of std. dev. ratio is ",
sqrt(cilower)," to ",sqrt(ciupper), "\n")
```

The output graph and output results are shown below.
Structurally, this R code is similar to that for the previously computed CI, after reading in the distribution of on-time arrival percentages (column 7 of the data file), and using the proper \texttt{qf()} commands. Also notice that we made the y axis of the histogram relative percent (called “density”) by using the argument \texttt{prob=TRUE} in the \texttt{hist()} command. When we insert the curved line superimposed on the histogram, using the \texttt{lines(density())} command, we must have the y axis of the histogram as density, not as “counts”, in order for the lines() command to work correctly.

Our results show that the ratio of standard deviations are about the same, so the distribution spread of on-time arrivals and departures at this airport are fairly consistent.

\textbf{HomeWork [2]:} Using the code above, repeat the results for a 90\% CI of comparison of variances ($\sigma_1^2/\sigma_2^2$) and standard deviations ($\sigma_1/\sigma_2$).

\textbf{Chi-squared Testing of 2 categorical variables:}

We have 3 elementary uses of comparing categorical variables with the $X^2$ statistic--[1] “Goodness of Fit”, “Test of Homogeneity”, and “Test of Independence”.

Goodness of fit uses the following hypotheses:
- \textbf{Ho:} all categories are essentially matching the model
- \textbf{Ha:} at least one category is too variant from the proposed proportion specified in the model

We will use the $X^2$ statistic to determine the probability of Ho being true. The statistic is computed: $X^2 = \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$, where “observed” = number of counts in each cell and “expected” = number of cells expected to be in each cell by our model. If the value of chi-squared is too high (therefore the difference of observed-expected is too
much), we have a low probability that Ho is true.

**Example of Goodness-of-fit:**
The data file (shown below) of baseball_birth.csv contains each month the members of the major leagues were born in an unspecified year, along with the percentage of all humans born during that time at the specified months.

```
<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>birth month</td>
<td>count</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>137</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>121</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>116</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>121</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>126</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>114</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>162</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>165</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>134</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>115</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>165</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>122</td>
</tr>
</tbody>
</table>
```

We want to test the hypotheses that major league players' birth months match that of humanity. We will do a goodness-of-fit test to determine this. See the code used below.

```r
# # goodness of fit
# # -------------------
data1 <- read.csv("baseball_birth.csv")
players <- data1[,2]
barplot(players, main="players' birth month", xlab="month", ylab="count", names=c(1:12))
chisq.test(players, p=c(.08,.07,.08,.08,.08,.09,.09,.09,.08,.09))
probvec <- c(.07,.08,.09)* sum(players)
plot(1:12,players, main="players plotted on model", xlab="month", ylab="players count", pch=16)
abline(h=probvec, lty=2)
text(4,145,labels="dashed lines = model values")
```

The output and graphs are shown below.

Possible outlier??
The R code does the following. We first read in the data file, calling the number of players listed in column 2 of the data as `players`. We make a bar plot of this number just as a preliminary graph of this study. Then we perform the `chisq.test()` on the list, comparing it to the probability model listed as a vector, `c(.08,.07,..., etc)`, which is essentially the percentages (expressed as proportions) given in the last column of data of all humans born that specific month.

The results of the chi-squared test is too high to have faith in our Ho, so we have strong evidence that the major league players don’t follow the human birth model. I made a plot of the number of players born each month, along with dotted horizontal lines representing the number representing .7, .8, and .9 of the players. We see from the graph that at least one month (the 8th month, August) seems to be an outlier, where many more players are born than should be born according to the model. This seeming outlier would seriously affect our goodness-of-fit test results. Note that I found the value where the horizontal lines occur by creating a vector `probvec`, made up of the number of players representing .7, .8, and .9 of all players counted.

HomeWork [3]: The data file `zodiak500.csv` contains birth signs of 256 of the Fortune 500 CEO's from a past year. Use a goodness-of-fit test to determine the following:

- Ho: all zodiac signs are equally represented by the CEO population (i.e., each month represents 1/12 = 0.0833 of all CEO's)
- Ha: at least one sign is inequitably represented by the CEO's

Be sure to include either a graph, a plot, or both, along with your test results, and comment on whether you trust your test results, based upon the graph, plot, etc.