Chi-square is the most generally useful of inferential statistics because it is the most versatile. It can be applied with qualitative data or quantitative data. It makes no statistical presumptions (i.e. it is "nonparametric"). It can be used to test a distribution's "goodness of fit" as well as to test the significance of the association between two variables in a contingency table with any number of rows and any number of columns. Its statistical basis is easy to understand, and Microsoft Excel supplies everything needed to use it, including "the table."

The only rule to remember is that each expected frequency should be 5 or greater.

Here is what you are to do. First, read the primer on chi-square on pages 4 and 5 below. Then, complete the following four exercises to submit hardcopy at our next session.

1) Hypothesis testing: With respect to the data in Table 6.6 on page 132 of Rudestam and Newton (reproduced below), apply chi-square to compute the probability reportable in the statement, "The relationship between parenthood and the belief that abortion is wrong is significant at p = _____ " (this statement is not in Rudestam and Newton).

<table>
<thead>
<tr>
<th>Ever Had Children?</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>abortion always wrong</td>
<td>330</td>
<td>180</td>
</tr>
<tr>
<td>abortion almost always wrong</td>
<td>180</td>
<td>120</td>
</tr>
<tr>
<td>abortion never wrong</td>
<td>90</td>
<td>100</td>
</tr>
</tbody>
</table>

Perform the computations by hand to compute the value of chi-square, then perform them through a worksheet in Excel using CHIINV. Use symbolic expressions to compute the expected frequencies, and use the CHITEST statistical function to compute the probability. Compute the table's degrees of freedom by inspection.

Hand-in two copies of the worksheet, one showing the results and one showing the formulas (CTRL + ` toggles between results and the formulas, where ` is the grave accent, which is the leftmost key on the top row).
2) Testing goodness of fit: Table 6.5 on page 128 of Rudestam and Newton shows the age
distribution in a random sample of the U.S. population from a survey by the National Opinion
Research Center (the NORC). The representativeness of the NORC sample will be evidenced if the
observed age distribution matches the data from the U.S. Census Bureau. The NORC data is
reproduced below:

<table>
<thead>
<tr>
<th>age group</th>
<th>18-25</th>
<th>26-35</th>
<th>36-45</th>
<th>46-55</th>
<th>56-65</th>
<th>66-75</th>
<th>76-89</th>
</tr>
</thead>
<tbody>
<tr>
<td>percentage</td>
<td>12.2</td>
<td>22.2</td>
<td>23.1</td>
<td>17.7</td>
<td>10.5</td>
<td>8.4</td>
<td>6.0</td>
</tr>
</tbody>
</table>

If the U.S. Census Bureau reports the following percentages, what is the probability that chance
alone accounts for the difference between the sample's age distribution and the population's?

<table>
<thead>
<tr>
<th>age group</th>
<th>18-25</th>
<th>26-35</th>
<th>36-45</th>
<th>46-55</th>
<th>56-65</th>
<th>66-75</th>
<th>76-89</th>
</tr>
</thead>
<tbody>
<tr>
<td>percentage</td>
<td>13.6</td>
<td>18.0</td>
<td>23.4</td>
<td>18.5</td>
<td>10.9</td>
<td>8.2</td>
<td>7.5</td>
</tr>
</tbody>
</table>

Hand in two copies of the worksheet, one showing the results and one showing the formulas.

3) CHITEST shows the probability that the observed frequencies could have resulted from chance.
This is the value used for rejecting or accepting the null hypothesis and is the finding you'll report.

CHIINV shows the value of chi-square itself. Its first argument is the probability (e.g. the cell
holding the CHITEST function); its second argument is the table's degrees of freedom.

CHIDIST shows the same probability that CHITEST shows, except that instead of computing it
from the observed and expected frequencies, it computes it from the value of chi-square and the
table's degrees of freedom. If you were to calculate chi-square by hand, you could use CHIDIST to
get the probability.

3.a.i On the worksheet for exercise 1, use CHIINV to view the value of chi-square.
3.a.ii Use CHIDIST to get the probability from the value of chi-square given by CHIINV.

3.b.i On the worksheet for exercise 2, use CHIINV to view the value of chi-square.
3.b.ii Use CHIDIST to get the probability from the value of chi-square given by CHIINV.

Hand in two copies of the worksheet, one showing the results and one showing the formulas.
4) The statistical significance of the relationship between two variables is different from the strength of their association (the ability to predict the value of one based on the other). A weak association, with a large N, may be highly significant. Conversely, an apparently strong association based on a small N may be fortuitous (not significant).

Two different measures of contingency association are shown below, Yule's Q and epsilon. Like the Pearson coefficient of correlation, both of these descriptive statistics range between -1 and 1. The difference between them is that epsilon maximizes only with an "if-and-only-if" the relationship between the independent variable and the dependent variable (e.g. working out and muscular physique). Yule's Q maximizes with an "if-then" relationship; in other words when the independent variable tends to be necessary but not sufficient (e.g. not smoking and longevity).

\[
\text{Yules' Q} = \frac{ad - bc}{ad + bc}
\]

\[
\text{epsilon} = \frac{ad - bc}{(a+c)(b+d)}
\]

Suppose a sample of 30 yields the following data. Q's value of 0.391 is indicative of a worthy association, but chi-square's probability of 0.269 suggests that the vicissitudes of sampling could yield data like this over 25% of the time when in fact there is no association at all.

\[
\begin{array}{c|cc}
\text{Independent variable} & \text{high} & \text{low} \\
\text{Dependent variable} & \text{high} & a & b \\
& \text{low} & c & d \\
\end{array}
\]

\[
\begin{array}{c|cc}
\text{Independent variable} & \text{high} & \text{low} \\
\text{Dependent variable} & \text{high} & 8 & 5 \\
& \text{low} & 7 & 10 \\
\end{array}
\]

\[
\begin{array}{c|cc}
\text{Independent variable} & \text{high} & \text{low} \\
\text{Dependent variable} & \text{high} & a & b \\
& \text{low} & c & d \\
\end{array}
\]

Assuming the proportionality of observations holds constant (so Q remains fixed), re-compute \textsc{CHITEST} with sample sizes of 60, 90, 120, 150, and 180 to show how significance increases.

Hand in two copies of the worksheet, one showing the results and one showing the formulas.
A Primer on Chi-Square (X²)

As all inferential statistics, chi-square is a tool for deciding whether chance variation in sampling or some causative phenomenon is responsible for the observed results.

Always working with tables, chi-square operates by adding-up, cell-by-cell, the difference between the observed count and the count that would have been expected on the basis of chance. The larger this total, the more compelling the evidence that something other than chance is responsible. Exactly how large a sum is needed to rule out chance as the responsible agent depends on the number of cells and the certainty of the decision. The number-of-cells factor is called the "degrees of freedom." The certainty factor is called the "significance level." The number associated with a level of significance for a number of degrees of freedom is called the "critical value." It is this value that needs to be exceeded for the null hypothesis (the hypothesis of "no difference" or "no effect"), \( H_0 \), to be rejected in favor of the alternative hypothesis, \( H_a \).

Excel gives critical values with the \texttt{CHIINV} function. For example, the value that must be exceeded in a chi-square test of significance at the .01 level with 8 degrees of freedom is \( \text{CHIINV}(0.01,8) \).

The \texttt{CHITEST} function reports the probability that chance accounts for the disparity between observed and expected frequencies. The first argument is the table of observed frequencies. The second argument is a corresponding table of statistically expected frequencies. You need to provide both. For a contingency table, expected frequencies are computed from the row sums and the column sums of the data's tabulation (i.e. the table of observed frequencies). For instance:

\[
\begin{array}{c|c|c|c}
\text{the data} & \text{calculated data} \\
\text{(observed frequencies)} & \text{(expected frequencies)} \\
\hline
a & b & c & (a+b+c) \\
\hline
d & e & f & (d+e+f) \\
\hline
a+d & b+e & c+f & N = a+b+c+d+e+f \\
\hline
\end{array}
\]

When chi-square is applied to contingency tables, the expected frequency for the cell in row \( r \), column \( c \) is the (sum_of_row\_r) times the (sum_of_column\_c) divided by \( N \), the size of the data set. The number of degrees of freedom is the table's (number_of_rows - 1) times (number_of_columns - 1). For the 3x2 table shown, the degrees of freedom is (3-1)*(2-1) = 2.

When chi-square is applied as a goodness of fit test, it is as if the data's table has a single row. The degrees of freedom is the number of cells minus 1. The expected frequencies come from the theoretical expectations.

If you were to compute chi-square by hand, \texttt{CHIDIST} would give you the probably that such a value could have resulted by chance. The first argument is the computed value of chi-square; the second argument is number of degrees of freedom.
The manual computation of chi-square is based on the differences between the observed frequency and the expected frequency of each cell in the table:

<table>
<thead>
<tr>
<th>cell</th>
<th>observed frequency</th>
<th>expected frequency</th>
<th>(O-E)</th>
<th>(O-E)^2</th>
<th>(O-E)^2 / E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>O</td>
<td>E</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The sum of this column is chi-square

When chi-square has only one degree of freedom, the computation is slightly different. It is known as Yates' correction for discontinuity:

<table>
<thead>
<tr>
<th>cell</th>
<th>observed frequency</th>
<th>expected frequency</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>O</td>
<td>E</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The sum of this column is chi-square