1 CHI-SQUARED TEST

Motivation:
Is there an association between two categorical variables or are they independent? We have count data in a two-way table like the following, where Factor X takes one of \( r \) values (or levels) and Factor Y takes one of \( c \) values. Here, \( r = c = 2 \).

<table>
<thead>
<tr>
<th>Factor X</th>
<th>Level A</th>
<th>Level B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>( n_{1A} )</td>
<td>( n_{1B} )</td>
<td>( n_1 = n_{1A} + n_{1B} )</td>
</tr>
<tr>
<td>Level 2</td>
<td>( n_{2A} )</td>
<td>( n_{2B} )</td>
<td>( n_2 = n_{2A} + n_{2B} )</td>
</tr>
<tr>
<td>Total</td>
<td>( n_A = n_{1A} + n_{2A} )</td>
<td>( n_B = n_{1B} + n_{2B} )</td>
<td>( n = n_1 + n_2 = n_A + n_B )</td>
</tr>
</tbody>
</table>

We can use the table to calculate various probabilities. We can compute the marginal distributions from the margins of the table (i.e. the row and column totals). The marginal distributions summarize each factor independently. As an example consider the marginal distribution of Level A,

\[
P(\text{Level A}) = \frac{n_A}{n}
\]

or the marginal distribution of Level 2,

\[
P(\text{Level 2}) = \frac{n_2}{n}
\]

We can also find the conditional distributions using the cells of the 2-way table. The cells represent the intersection of a given level of one factor with a given level of the other factor. For example, the conditional distribution of Level B given Level 1 is

\[
P(\text{Level B}| \text{Level 1}) = \frac{n_{1B}}{n_1}
\]

or the conditional distribution of Level 2 given Level A

\[
P(\text{Level 2}| \text{Level A}) = \frac{n_{2A}}{n_A}
\]

Hypothesis formulation:

\( H_0 \): There is no association between row and column variables (the two factors are independent).

\( H_A \): \( H_0 \) is not true.

The Chi-Square Statistic, \( \chi^2 \):

When \( H_0 \) is true, the expected count in each cell is computed based on the marginal distributions as follows:

\[
\text{expected count} = \frac{\text{row total} \times \text{column total}}{\text{table total}}
\]

We can then compute the test statistic:

\[
\chi^2 = \sum_{i,j} \frac{(\text{observed}_{ij} - \text{expected}_{ij})^2}{\text{expected}_{ij}}
\]

P-value:

When \( H_0 \) is true, the \( \chi^2 \) test statistic has the \( \chi^2 \) distribution with \((r - 1)(c - 1)\) degrees of freedom (denoted \( \chi^2_{(r-1)(c-1)} \)). Note that this assumes Factor X has \( r \) levels and Factor Y has \( c \) levels, so in the case above \((r - 1)(c - 1) = 1 \).
Assumptions:
1. The data represent one or more random samples from an observational study or a randomized experiment.
2. All expected counts have numerical values $\geq 1$.
3. Most expected counts have numerical values $\geq 5$. There is no more than one smallish value for every 5 expected counts.

Calculator:
If you are given $\chi^2$ and can calculate the degrees of freedom, use 2ND>DISTR>X2cdf($X^2$, 1E99, (r-1)(c-1)) to find the P-value.
Otherwise, enter the count data into a matrix the same way as it is presented in the two-way table cells and use the $X^2$-Test function. Namely,
1. Enter data into matrix $[A]$ by 2ND>MATRIX>EDIT
2. Calculate the $\chi^2$ statistic and P-value with STAT>TESTS>X2-Test>Observed=$[A]$. This function will calculate the expected values and store them in a matrix that you specify (default is $[B]$).

Example Problem:
In the 2008 General Social Survey conducted by National Opinion Research Center at the University of Chicago, participants were asked if they favor or oppose the death penalty for persons convicted of murder and if they think the use of marijuana should be made legal or not. A two-way table of counts for the responses to these two questions is as follows:

<table>
<thead>
<tr>
<th>Death Penalty?</th>
<th>Legal</th>
<th>Not Legal</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favor</td>
<td>313</td>
<td>461</td>
<td>774</td>
</tr>
<tr>
<td>Oppose</td>
<td>165</td>
<td>247</td>
<td>412</td>
</tr>
<tr>
<td>Total</td>
<td>478</td>
<td>708</td>
<td>1186</td>
</tr>
</tbody>
</table>

a. Compute the proportion of respondents who favor the death penalty in each marijuana opinion group.
b. The proportions you computed are part of
   A. the marginal distribution of marijuana legalization opinion.
   B. the conditional distribution of marijuana legalization opinion, given death penalty opinion.
   C. the marginal distribution of death penalty opinion.
   D. the conditional distribution of death penalty opinion, given marijuana legalization opinion.
c. Is there evidence that the opinions regarding the death penalty and marijuana are significantly associated? Run the appropriate inference test and give the numerical values of the test statistic and the test P-value. Show what you typed in your calculator.
d. Assuming the conditions for the $\chi^2$ test are met, conclude in context using a significance level of 0.05.

Solutions:
a. Legal: 0.655, Not legal: 0.651
b. D
c. Enter the data into a 2 x 2 matrix, then use X2-Test $\Rightarrow \chi^2 = 0.017, P = 0.896$
d. We fail to reject the null hypothesis ($P > 0.05$). We are not able to conclude that there is a relationship between opinion about the death penalty and opinion about marijuana legalization.
The Pew Research Center conducted a survey of a random sample of 2250 American adults. Each participant was asked about their place of residence (urban, suburban, or rural area) and whether they ever go online to email or use the internet. Here are the summary findings:

<table>
<thead>
<tr>
<th></th>
<th>Not online</th>
<th>Online</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban</td>
<td>107</td>
<td>656</td>
<td>763</td>
</tr>
<tr>
<td>Suburban</td>
<td>145</td>
<td>892</td>
<td>1037</td>
</tr>
<tr>
<td>Rural</td>
<td>90</td>
<td>360</td>
<td>450</td>
</tr>
</tbody>
</table>

a. Compute the proportion of respondents who do not go online in each residential group.

b. The proportions you computed are part of
   A. the marginal distribution of online presence.
   B. the conditional distribution of online presence, given residential group.
   C. the marginal distribution of residential group.
   D. the conditional distribution of residential group, given online presence.

c. Is there evidence that online presence is significantly associated with residential type? Run the appropriate inference test and give the numerical values of the test statistic and the test P-value. Show what you typed in your calculator.

d. What are the assumptions for this inference procedure and are they met? Explain briefly.

e. Conclude in context using a significance level of 0.05.

Solutions:

a. Urban: 0.1402, Suburban: 0.1398, Rural: 0.2000
b. B
c. Enter the data into a $3 \times 2$ matrix, then use $\chi^2$-Test $\Rightarrow \chi^2 = 10.06, P = 0.0066$
d. Randomness: yes, data are a random sample

*Large enough expected counts*; yes, smallest expected count is $68 > 5$
e. There is significant evidence ($P < 0.05$) of an association between place of residence and online presence.
For each following question, choose the appropriate inference procedure:

(A) One sample or matched pairs $t$

(B) Two sample $t$

(C) ANOVA

(D) One sample $z$

(E) Chi-square for two-way tables

1. A recent study examined whether light-emitting eReaders at bedtime may impair sleep quality. A random sample of 12 healthy adults slept in the lab on two different days. In random order, participants read for 30 minutes a print book one night and a light-emitting eReader the other night. Scalp electrodes were used to measure how long (in minutes) it took participants to reach a deep sleep stage.

2. A study enrolled patients who had regained weight after gastric bypass surgery and randomly assigned them to either a secondary stomach reduction or to a sham procedure. The objective was to see if the secondary stomach reduction leads, on average, to a greater percentage weight loss from baseline than the sham procedure.

3. A study is interested in opinions about the gender pay gap in the American workforce. A random sample of adult U.S. residents is selected, and they are asked whether they think that there is a lot or some discrimination against women in businesses today. The study then compares the responses of men and women in the sample.

4. A study examined the efficacy of acetaminophen for treating osteoarthritis of the knee in 779 patients who had knee pain during physical activity. The participants were randomly assigned to a 6-week treatment with 4 grams per day of acetaminophen or to a placebo. The study found that 213 of the 405 subjects in the acetaminophen group achieved noticeable pain relief, compared with 194 of the 374 subjects in the placebo group.

5. A study recruits 50 patients with chronic lower back pain and assigns them randomly to either biweekly yoga sessions or biweekly physical therapy sessions. The subjects complete a McGill Pain Questionnaire to evaluate their pain level (score of 0 to 10, 10 being the worst) after 5 weeks on the program. The objective of the study is to determine whether the physical therapy is significantly better than yoga at managing pain level.

6. Some types of cheese are aged to allow flavor to develop. A dairy company selects a random sample of 20 cheddar cheese wheels (commercially produced cheeses are often as large as a cart wheel, hence the name). For each wheel, a little piece of cheese is taken when the wheel is one month old, and another piece is taken when the wheel is 4 months old. A quality-control engineer records lactic acid concentration each time, in order to assess whether the aging process significantly changes lactic acid concentration in cheddar cheese.

7. The National Assessment of Educational Progress (NAEP) was used to compare the impact of testing conditions on test scores of native Spanish speakers. The study compared the test scores of three independent random samples of native Spanish-speaking students who were tested with either a dual-language test, an English test with dual-language instructions, or an English-only test.

Solutions:
1. A (matched pairs) 2. B (two sample $t$ test) 3. E (Chi squared) 4. E (Chi squared) 5. B (two sample $t$ test) 6. A (matched pairs) 7. C (ANOVA)