In Ch. 7, we learned how to build CIs for either 1 prop, \( \pi \), or the difference between 2 props, \( \pi_1 - \pi_2 \), where \( \pi = \) prop of something (e.g. boys) in population 1, and \( \pi_2 = \) same thing \( z \).

We also learned how to do hyp. tests on \( \pi \) or \( \pi_1 - \pi_2 \). [Note \( \pi_1 + \pi_2 \neq 1 \), because \( \pi_1, \pi_2 \) are 2 different populations]

But in all of these situations, the 2 pops have 2 categories (boy/girl) and \( \pi_i \) is the prop. of 1 of them.

The tornado/climate e.g. in last lecture deals with the situation where a population has 3 categories. On the last page of last lecture, we generalized to a population with \( k \) categories.

We learned that the relevant dist. is chi-squared with df=\( k-1 \). And the quantity that follows that dist. is

\[
\chi^2 = \sum_{i=1}^{k} \left( \frac{\text{obs}_i - \text{exp}_i}{\sqrt{\text{exp}_i}} \right)^2
\]

where \( \text{obs}_i \) and \( \text{exp}_i \) are observed and expected counts in the \( i \)th category (still of 1 population). The latter are computed assuming \( H_0 \) is true, where

\[
H_0 : \pi_1 = \pi_0, \pi_2 = \pi_0, \ldots, \pi_k = \pi_0 \quad \text{[This time \( \pi_1 + \pi_2 + \ldots = 1 \]} \]

\[
H_1 : \text{At least one of these is wrong}
\]
Just FYI: The chi-squared density function is

\[ f(x) = \frac{x^{(df/2)-1} e^{-x/2}}{\Gamma(df/2)} \frac{1}{2^{df/2}} \]

But all we need is Table VII.

Note that the above \( H_0, H_1 \) is just a generalization of
\( H_0: \pi = \pi_0 \) (\( z \)-test).
\( H_1: \pi \neq \pi_0 \)
to more than 2 categories in the population.

However, there are \( \square \) 1-sided/2-sided varieties of chi-sq.

When \( \chi^2_{obs} \) is small (say no), then the observed counts are consistent with the expected counts if \( H_0 = i \)
(i.e. \( \pi_1 = \pi_{o1}, \pi_2 = \pi_{o2}, ..., \pi_k = \pi_{ok} \)). So, if \( \chi^2_{obs} \) is large, then at least one of these specifications is wrong.

In other words, the appropriate hypotheses are
\( H_0: \pi_1 = \pi_{o1}, \pi_2 = \pi_{o2}, ..., \pi_k = \pi_{ok} \)
\( H_1: \) At least one of these specifications is wrong.

And it is the "At least" which gives us

\[ p\text{-value} = \text{prob}(\chi^2 \geq \chi^2_{obs}) \]

(Table VII)

I.e. We are always interested in the upper tail area only.

Said differently, for the chi-squared test of the above \( H_0/H_1 \), the \( p\)-value is only the right area, because violation of each part of \( H_0 \), increases \( \chi^2 \).
Further Diagnosis:

The magnitude of the $k$ terms in $X^2_{obs}$ is useful for diagnosing which of the $k$ proportions differ most from the null values.

Example: In the tornado example,

Suppose we had found $p$-value < $\alpha$, i.e. there is evidence that climate does affect tornadic activity. In that case $X^2_{obs}$ would have been very large (so that $p$-value would be small).

The 3 terms contributing to $X^2_{obs}$ are

- $1.27$ El Niño (large)
- $0.009$ La Niña (small)
- $0.49$ Normal

Then we could conclude that it is the El Niño years which are most different (in terms of tornadic activity) from what would be expected by chance (i.e., if climate had no effect on tornadic activity).

In other words, we could conclude that the effect of climate on tornados is most in the El Niño years. Tornadic activity in La Niña years, in fact, seems to be pretty close to what one would expect by chance.
The chi-sqrd dist. shows up in 2 other situations.

2) \text{[r pops across k populations]:}

- \text{Ho: r pops are homogeneous w.r.t. k categories}
- \text{H}_i: \text{not a big concept!}

\text{Homogeneous means}

- \text{pop 1: } \frac{\sum_{i=1}^{k} \pi_i^1}{k} = 1
- \text{pop 2: } \frac{\sum_{i=1}^{k} \pi_i^2}{k} = 1
- \text{pop r: } \frac{\sum_{i=1}^{k} \pi_i^r}{k} = 1

\text{Not homogeneous means that at least 1 of these is wrong.}

\text{E.g., for } k = 2 \text{ the above } \text{Ho/H}_i

\text{Translate to:}

- \text{pop 1: } \pi_1^1 \rightarrow \pi_1^2 \rightarrow \text{homogeneous}
- \text{pop 2: } \pi_2^1 \rightarrow \pi_2^2 \rightarrow \text{homogeneous}

\text{Note: This is diff. from } \pi_1 = \pi_2, \pi_3 = \pi_2, \ldots

3) \text{Independence of 2 categorical variables, one with k levels, the other with r levels.}

\text{[r x k contingency table]:}

- \text{Ho: 2 cat. vars. are independent}
- \text{H}_i: \text{not independent}
In such problems, the data look like this:

\[
\begin{array}{c|ccc}
\text{category} & 1 & 2 & 3 \\
\hline
\text{categ 1 AND in pop 1} & a & b & c \\
\end{array}
\]

\[
\begin{array}{c|ccc}
\text{pop } A & a+b+c \\
\hline
\text{pop } B & d+e+f \\
\end{array}
\]

\[
\frac{a+d}{n} \quad \frac{b+e}{n} \quad \frac{c+f}{n}
\]

The observed counts are from the sample/data.

The good news is that one can test homogeneity with a chi-squared test, but with \( df = (k-1)(r-1) \).

\[
\text{counts, not props}
\]

I.e., compute \( X^2_{\text{obs}} = \sum_{\text{all cells}} \frac{(\text{exp} - \text{obs})^2}{\text{exp}} \), and \( p\text{-value} = P(X^2 > X^2_{\text{obs}}) \).

Table VII.

The only question is what are the expected counts?

Expected counts:

Assuming \( H_0 = T \).

\[
\begin{pmatrix}
\frac{(a+b+c)(a+b)}{n} & \frac{(a+b+c)(b+c)}{n} & \ldots \\
\frac{d+e+f}{n} & \ldots & \ldots
\end{pmatrix}
\]

Remember this result a "row \times \text{col. marginals}".
Eg. Are Boys & Girls homogenous wrt. to their belief in afterlife?

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>Un.</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boys</td>
<td>435</td>
<td>58</td>
<td>89</td>
</tr>
<tr>
<td>Girls</td>
<td>375</td>
<td>50</td>
<td>84</td>
</tr>
</tbody>
</table>

\[ \begin{array}{ccc}
582 & & \\
810 & 108 & 173 \\
1109 & & \\
\end{array} \]

\[ \frac{(582)(810)}{1109} = 432.1 \quad 57.6 \quad 92.3 \]
\[ \frac{(58)(50)}{81} = 37.9 \quad 50.4 \quad 80.7 \]

\[ \chi^2 = \frac{(435 - 432.1)^2}{432.1} + \frac{(58 - 57.6)^2}{57.6} + \ldots \]

\[ = .019 + .0028 + .0118 + .022 + .0032 + .0135 = 0.3 \]

\[ df = (2-1)(3-1) = 2 \implies p-value > 0.1 \text{ (huge)} \]

Cannot reject Ho in favor of H1, at \( \alpha = .01 \text{ or } .05 \)

I.e. there is no evidence to think that Boys and Girls are not homogeneous wrt. their belief in afterlife.

```
Also
```

There is no evidence to think that Gender and Belief in afterlife are not independent.

(i.e. They "are" independent.)
Further Diagnosis:

So, based on this data, we cannot say that there is a difference between boys & girls w/r/t their belief in afterlife. Mathematically, the reason is that \( \chi^2_{obs} \) was too small.

But suppose, \( \chi^2_{obs} \) had turned out to be huge. Then we could conclude that there is a difference. Then, just as we did with the chi-sq test of prev. lecture, we can look at the relative size of the various terms in \( \chi^2_{obs} \).

In this example, the big terms are 0.118, 0.135, which correspond to the "Xo" category.

In short, if the result had turned out to be statistically significant (i.e. \( \chi^2_{obs} = \text{huge}, p\text{-value}<\alpha \) ), then we could go further and say the difference between boys & girls w/r/t their belief in afterlife is mostly in the non-believer category.
Summary: Chi-sq is shows up in 3 situations:

I) 1 pop. (1 variable) with k categories.

\[ H_0: \pi_1 = \pi_{o1}, \pi_2 = \pi_{o2}, \ldots, \pi_k = \pi_{ok} \]
\[ H_1: \text{At least one of } \pi_i \text{ is wrong.} \]

Because there is only 1 pop, must have \( \sum_{i=1}^{k} \pi_{oi} = 1 \).

II) r pops. each with k categories:

\[ \pi_{11} \quad \pi_{12} \quad \pi_{13} \ldots \pi_{1k} \]
\[ \pi_{21} \quad \pi_{22} \quad \pi_{23} \ldots \pi_{2k} \]
\[ \vdots \]
\[ \pi_{r1} \quad \pi_{r2} \quad \pi_{r3} \ldots \pi_{rk} \]

\[ \sum_{i=1}^{k} \pi_{ii} = 1 \]

\[ \underbrace{\sum_{i=1}^{k} \pi_{ii}}_{\text{Each pop}} = 1 \]

\[ \underbrace{\sum_{i=1}^{k} \pi_{ii}}_{\text{All pops}} = 1 \]

\( H_0: \text{The above } \pi_i \text{'s are equal as specified.} \)
(ie. The r pops are homogeneous w.r.t. The k categories)

\( H_1: \text{For at least 1 of the k categories, the proportions are not equal for all pops.} \)
(ie. The r pops are not homogeneous ---)

III) Are 2 categorical variables independent?

\[ H_0: \text{They are.} \quad H_1: \text{They are not.} \]
The accompanying data resulted from an experiment in which seeds of five different types were planted and the number that germinated within 5 weeks of planting was observed for each seed type ("Nondestructive Optical Methods of Food Quality Evaluation," Food Science and Nutr., 1984: 232-279). Carry out a chi-squared test at level .01 to see whether the proportion of seeds that germinate in the specified period varies according to type of seed.

<table>
<thead>
<tr>
<th>Seed type</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germinated</td>
<td>31</td>
<td>57</td>
<td>87</td>
<td>52</td>
<td>10</td>
</tr>
<tr>
<td>Failed to germinate</td>
<td>7</td>
<td>33</td>
<td>60</td>
<td>44</td>
<td>19</td>
</tr>
</tbody>
</table>

Specifically,

a) Does the statement of the problem require a test of homogeneity of 2 populations with respect to 5 categories, or vice versa?
b) Compute the p-value corresponding to your answer in part a.
c) State your conclusion "in English."
d) Diagnose the various terms in X_obs^2.

For parts b and d, use the following:

\[
\chi^2_{\text{obs}} = 3.20 + 0.25 + 0.0001 + 0.42 + 3.00 \quad \text{Germinated}
\]
\[
+ 4.65 + 0.37 + 0.0002 + 0.61 + 4.37 \quad \text{Failed}
\]

Have you ever wondered whether soccer players suffer adverse effects from hitting "headers"? The authors of the article "No Evidence of Impaired Neurocognitive Performance in Collegiate Soccer Players" (The Amer. J. of Sports Medicine, 2002: 157Â162) investigated this issue. The paper reported that 45 of the 91 soccer players in their sample had suffered concussion, 28 of 96 nonsoccer athletes had suffered concussion, and only 8 of 53 student controls had suffered concussion. Denote

\[ p_{i1} = \text{pop. proportion of concussions among soccer players}, \]
\[ p_{i2} = \text{pop. proportion of concussions among non-soccer players}, \]
\[ p_{i3} = \text{pop. proportion of concussions among control group}. \]

Set up this problem as a test of homogeneity of three populations with respect to 2 categories. Specifically,

a) State the hypotheses in terms of \( p_{i1}, p_{i2}, p_{i3}. \)
b) Write the data in the form of a contingency table.
c) Compute the expected counts.
d) Compute the p-value (or specify a range for it).
e) State the conclusion "in English."
f) Diagnose the various terms appearing in \( X_{\text{obs}}^2. \)