Early in Ch 8 we did 2-sample test of \( (2) \) proportions on 2 populations, each with 2 categories:

\[ \pi_{1G} - \pi_{2G} = ? \Rightarrow \text{test } \pi_1 - \pi_2 (\pi_1 + \pi_2 \neq 1) \]

\[ z = \frac{(\hat{p}_1 - \hat{p}_2) - \Delta}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \Rightarrow p\text{-value} = P(z \geq z_{obs}) \]

Data: \( \begin{array}{c} 13/15 \end{array} \begin{array}{c} 20/40 \end{array} \)

\[ \hat{p}_{1\text{obs}} = \frac{13}{13 + 15} \quad \hat{p}_{2\text{obs}} = \frac{20}{20 + 40} \]

These were the values used for the test.

Last time we did chi-squared test of \( k \) proportions on 1 population with \( k \) categories:

\[ \pi_1 = ?, \pi_2 = ?, \ldots, \pi_k = ? \quad (\sum_i \pi_i = 1) \]

\[ \chi^2 \text{-sqd test: } \chi^2 = \sum_i \left( \frac{\text{obs}_i - \text{exp}_i}{\sqrt{\text{exp}_i}} \right)^2 \Rightarrow p\text{-value} = P(\chi^2 > \chi_{obs}^2) \]

Data: \( \begin{array}{c} 14/28/44 \end{array} \)

These were the counts used for the test.

Now, \( r \) populations, each with \( k \) categories:

\[ \chi^2 = \sum_i \left( \frac{\text{obs}_i - \text{exp}_i}{\sqrt{\text{exp}_i}} \right)^2 \quad \text{df} = (r-1)(k-1) \]

Again: chi-sqrd!
Example: Suppose we have 2 Computer Makers: Apple, Dell
Suppose there are 3 CPU makers: Intel, IBM, AMD

Are Apple & Dell different in terms of the CPUs they use?
Are the CPU makers different in terms of the Computer makers they sell to?

H₀: Comp. and CPU are independent
H₁: They are not independent.

Here are the data
(i.e. The Obs. counts) =>

<table>
<thead>
<tr>
<th></th>
<th>Intel</th>
<th>IBM</th>
<th>AMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple</td>
<td>435</td>
<td>58</td>
<td>89</td>
</tr>
<tr>
<td>Dell</td>
<td>375</td>
<td>50</td>
<td>84</td>
</tr>
<tr>
<td>Total</td>
<td>810</td>
<td>108</td>
<td>173</td>
</tr>
</tbody>
</table>

It can be shown (FYI, below):

 Expected: \[
\begin{pmatrix}
\frac{(582)(810)}{1091} & \ldots & \frac{(582)}{509} \\
\frac{(810)}{509} & \ldots & \frac{(108)}{509} \\
\frac{(173)}{509} & \ldots & \frac{(173)}{509}
\end{pmatrix}
\]

Counts:

\[
\begin{pmatrix}
432.1 & 57.6 & 92.3 \\
377.9 & 50.4 & 80.7
\end{pmatrix}
\]

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

\[= .019 + .0028 + .0118 + .0022 + .0032 + .0135\]

\[= 0.3\]

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

Cannot reject H₀ in favor of H₁ at α = .01 (.05)

There is no evidence that Comp. and CPU are not independent.
(I.e. They could be independent.)
Test of independence

Here is the general procedure: This test is called the chi-squared test of independence of two categorical r.v.'s, one with r levels, the other with k levels.

\[ H_0: \text{2 categ. vars. are indep.} \]
\[ H_i: \text{not indep.} \]

In such problems, the data are shown as a **Contingency Table**:

\[
\begin{array}{c|ccc}
X = A, Y = 1 & 2 & 3 \\
\hline
A & a & b & c \\
B & d & e & f \\
\end{array}
\]

The test of \( H_0 \) (i.e., independence) turns out to be a chi-squared test but with df = \((r-1)(k-1)\).

The only question is what are the expected counts (under \( H_0 \)). It can be shown (see FRI page, below):

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

Remember this result a "row x col. marginals".

Recall: counts, not props. Table VII.

I.e., compute \( X^2_{\text{obs}} = \sum_{\text{all cells}} \frac{(\text{obs.} - \text{exp.})^2}{\text{exp.}} \), and \( p\text{-value} = p(X^2 > X^2_{\text{obs}}) \).
Derivation of “row x col marginal” method for computing
The C-table of expected counts. For simplicity, let \( r = 2, k = 2. \)

Start with the observed counts:

<table>
<thead>
<tr>
<th>Category</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>population A</td>
<td>( a )</td>
<td>( b )</td>
</tr>
<tr>
<td>population B</td>
<td>( c )</td>
<td>( d )</td>
</tr>
<tr>
<td>a+c</td>
<td>b+d</td>
<td>( n )</td>
</tr>
</tbody>
</table>

Even though we talk about population A, B, ..., the counts \( a, b, c, d, \) are the observed sample/data.

Switch to proportions:

\[
\begin{align*}
\text{pop. A} & \left( \frac{\pi_{A1}}{\pi_{A2}} \right) \\
\text{pop. B} & \left( \frac{\pi_{B1}}{\pi_{B2}} \right)
\end{align*}
\]

\( \pi_{A1} \) = true prop. of category 1 in pop. A

Etc.

If \( H_0 = T \), i.e., if \( X \) and \( Y \) are independent, then \( \pi_{A1} = \pi_{B1} \) and \( \pi_{A2} = \pi_{B2} \).

But if \( \pi_{A1} = \pi_{B1} \), then it must be that \( \pi_{A1} = \pi_{B1} = \pi_1 \),

where \( \pi_1 \) = prop. of category 1, period!

So, if \( H_0 = \text{True} \), how many do we expect to get in pop. A?

Answer: \( \frac{(a+b) \pi_1}{\pi_1} \)

# in pop. A

We can estimate \( \pi_1 \) with the sample prop. of 1's: \( \frac{a+c}{n} \).

\( \therefore \) Answer: \( \frac{(a+b)(a+c)}{n} \).

And that is the 1st element of the expected C-Table. Etc.
Interpretation / Diagnosis

So, based on this data, we cannot say that there is a difference between the computer makers in terms of the CPU they use. Mathematically, the reason is that $x^2_{obs}$ was too small.

But suppose, $x^2_{obs}$ had turned out to be huge. Then we could conclude that there is a difference between comp. makers in terms of the CPU they use. Then we can look at the relative size of the various terms in $x^2_{obs}$ to see which ones make the $x^2_{obs}$ big.

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

In short, if we had gotten a “statistically significant” result, i.e., if the p-value were less than 0.05, then we would say that there is evidence that the comp. makers are different in terms of the CPU they use. But we could go further and say the biggest difference between the comp. makers is in their use of the AMD chip.

The “signs” can be interpreted, too, to indicate the “direction” of the association.
We learned chi-squared test of k specific proportions in 1 pop.

\[ H_0: \pi_1 = \pi_0, \pi_2 = \pi_0, \ldots, \pi_k = \pi_0 \]

\[ H_1: \text{At least 1 of these is wrong.} \]

chi-squared dist. with df = k-1

\[ \text{How does proportion of } \sqrt{X=0} \text{ something (e.g. tornadoes, } X=1) \text{ vary across } k \text{ categories (or } k \text{ levels of a categorical var. } Y)? \]

\[ 1 \text{ 2-level } X, 1 \text{ k-level } Y. \]

And the chi-squared test of independence:

\[ H_0: X \text{ and } Y \text{ are indep.} \]

\[ H_1: X \text{ and } Y \text{ are not indep.} \]

chi-squared dist. with df = (k-1)(r-1)

\[ X,Y = 2 \text{ catg./discr. r.v.s.} \]

This test is equivalent to a "test of homogeneity of population across categories", which we have skipped this quarter.

In regression we studied how does 1 (or more) continuous var. \( X \) affect another continuous var. \( Y \)

\[ X,Y = 2 \text{ continuous vars.} \]

How about how does 1 catg./discr. var. \( X \) affect 1 continuous r.v. \( Y \)?

This question requires comparing \( k \) means, i.e.

\[ M_1 = \text{mean of } Y \text{ for } X=1, \ M_2 = \text{mean of } Y \text{ for } X=2, \ldots, \ M_k = \text{mean of } Y \text{ for } X=k \]

And the question of whether \( X \) affects \( Y \) becomes

\[ H_0: M_1 = M_2 = \ldots = M_k \]

\[ H_1: \text{At least 2 } \mu \text{'s are different.} \]

Once again, this method compares the mean of a continuous r.v. at different levels of a catg./discr. r.v.

Example 1: Does knowledge of religion depend on religion?
**U.S. Religious Knowledge Survey**

**Executive Summary**

Atheists and agnostics, Jews and Mormons are among the highest-scoring groups on a new survey of religious knowledge, outperforming evangelicals and Protestants, mainline Protestants and Catholics on questions about the core teachings, history and leading figures of major world religions.

On average, Americans correctly answer 16 of the 32 religious knowledge questions on this survey (by the Pew Research Center’s Forum on Religion & Public Life). Atheists and agnostics average 26.9 correct answers; Jews and Mormons do about as well, averaging 25.5 and 20.3 correct answers, respectively. Protestants as a whole average 14 correct answers. Catholics as a whole. 14.7. Atheists and agnostics, Jews and Mormons perform better than other groups on this survey even after controlling for differing levels of education.

**Average number of questions answered correctly (dots)**

- Atheist/agnostic: 26.9
- Jewish: 25.5
- Mormon: 20.3
- White evangelical Protestant: 17.6
- White mainline Protestant: 15.6
- White Catholic: 16.0
- Nothing in particular: 15.2
- Black Protestant: 13.4
- Hispanic Catholic: 11.6

Looks like Atheists know most!?
Example 2: Does fullness of note sheet have an effect on test scores?

<table>
<thead>
<tr>
<th>note-sheet fullness</th>
<th>mean test score</th>
</tr>
</thead>
<tbody>
<tr>
<td>not-so-full 1</td>
<td>0.6437</td>
</tr>
<tr>
<td></td>
<td>0.7205</td>
</tr>
<tr>
<td></td>
<td>0.7179</td>
</tr>
<tr>
<td></td>
<td>0.7201</td>
</tr>
<tr>
<td>very full 5</td>
<td>0.7142</td>
</tr>
</tbody>
</table>

Again, looking at means is not enough. Must also look at variance.

We are going to learn this stuff today.

Note that this test is just a generalization of the 2-sample/pop test (for comparing $\mu_1, \mu_2$) to the case of $k$ populations.
Example 9.1 (p. 422–423)

Does data provide evidence that:
- mean vibration varies across k types of bearings?
- mean computer speed varies among computers?
- mean detection error varies among detection algorithms?
- Are k means different?

Data:

<table>
<thead>
<tr>
<th>Brand</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( \bar{y}_1 = 13.68 \quad \bar{y}_2 = 15.97 \quad 13.67 \quad 14.73 \quad 13.58 \)

\( s_1 = 1.194 \quad s_2 = 1.167 \)

We are dealing with 5 pop. means.

\( H_0: \mu_1 = \mu_2 = \ldots = \mu_5 \)

\( H_1: \) At least 2 \( \mu \)'s are diff.

The Way ANOVA answers that question is by finding out how much of the total variability in \( y \) is within each category/pop., and how much is between the categories/populations.

Important & powerful idea
Recall the decomposition of SST from regression. Similarly, the total variability in the $y_{ij}$ is:

\[ \text{SST} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \]

- $\sum_{i=1}^{k} n_i$ is the total sample size across all populations.
- $\bar{y}$ is the grand mean.

\[ \bar{y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} = \frac{1}{n} \sum_{i=1}^{k} n_i \bar{y}_i \]

- $\bar{y}_i$ is the sample mean in the $i^{th}$ pop.
- $\bar{y}$ is the sample mean across all populations.

SS between group

SS explained

Sample mean $\bar{y}_i$

Sample means

SS unexplained

Variation between groups

Variation within groups

\[ \text{SS} : \text{Total} = \text{between} + \text{within} \]

\[ \text{df} : n-1 = (k-1) + (n-k) \]

- $k$ = # of levels in 1 factor (predictor)
- $n$ = # of pops.

Note: in linear regression:

\[ \text{df} : n-1 = k + [n-(k+1)] \]

- $k$ = # of $y$'s
Now, we can compare $SS_{\text{between}}$ and $SS_{\text{within}}$:

**Theorem:**

If $H_0 = \text{True}$, $F = \frac{SS_{\text{between}}/(k-1)}{SS_{\text{within}}/(n-k)} \equiv \frac{MS_{\text{between}}}{MS_{\text{within}}}$

has an $F$-distribution with params $df = (k-1, n-k)$

All we need is Table VIII

$p$-value $= p(F > F_{\text{obs}})$

The assumption of this theorem is that the $y$'s in each of the $k$ populations are normal, and that they all have the same variance, i.e. $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$. (Called homoscedasticity!)

Use qq-plots to test this. See optional hiv for understanding this.
Consider 5 brands of computers. A code has been run on each of the brands 6 times, and the completion times have been recorded. Here are the data:

<table>
<thead>
<tr>
<th>Brand</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\bar{y}_1 = 13.68, \quad \bar{y}_2 = 15.97, \quad 13.67, \quad 14.73, \quad 13.58
\]

\[
\begin{align*}
\text{Var. between} & \quad \bar{y}_i = 13.68 & \bar{y}_2 = 15.97 & 13.67 & 14.73 & 13.58 \\
\text{Var. within} & \quad s_1 = 1.194 & s_2 = 1.167 & \cdots & \cdots & \cdots \\
\end{align*}
\]

\[
\bar{y} = \frac{\sum_{i=1}^{5} \left( \frac{n_i}{n} \right) \bar{y}_i}{n} = \frac{6}{30} (13.68) + \cdots = 14.22
\]

\[
\begin{align*}
SS_{\text{between}} &= \sum_{i=1}^{5} n_i (\bar{y}_i - \bar{y})^2 = 6(13.68 - 14.22)^2 + \cdots = 30.88 \\
SS_{\text{within}} &= \sum_{i=1}^{5} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \frac{5}{30} (n_i - 1) s_i^2 = (6-1)(1.194)^2 + \cdots = 22.83
\end{align*}
\]

\[
F = \frac{30.88/(5-1)}{22.83/(30-5)} = 8.45
\]

\[
\text{df} = (5-1, 30-5)
\]

\[
p\text{-value} = P(F > F_{0.05}) = P(F > 8.45) < 0.001 \quad \text{Table VIII.}
\]

\[
\alpha = 0.01
\]

Conclusion: Reject \( H_0 (\mu_1 = \mu_2 = \cdots) \) in favor of \( H_1 \) (at least 2 \( \mu \)'s are diff).

In English: Brand has an effect on speed.
Most software produce an ANOVA Table (See pre-lab)

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Group</td>
<td>k-1</td>
<td>SS_between</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Group</td>
<td>n-k</td>
<td>SS_within</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>n-1</td>
<td>SSTotal</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For the above example (from R):

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor</td>
<td>5-1</td>
<td>30.85</td>
<td>7.71</td>
<td>8.44</td>
<td>.00018</td>
</tr>
<tr>
<td>Error</td>
<td>30-5</td>
<td>22.84</td>
<td>0.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>30-1</td>
<td>53.7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Summary:

2, t
H₀: \( \mu = \mu_0 \)  \( H_1: \mu \neq \mu_0 \)

\( \sigma^2 \) known/unknown

2
H₀: \( \pi = \pi_0 \)  \( H_1: \pi \neq \pi_0 \)

2 (No t). "large sample"

2,t
H₀: \( \mu_1 - \mu_2 = \Delta_0 \)  \( H_1: \mu_1 - \mu_2 \neq \Delta_0 \)  \( \text{indep. or paired} \)

Chi-sq
H₀: \( \pi_1 = \pi_0, \pi_2 = \pi_2, \ldots, \pi_k = \pi_k \)  \( H_1: \text{At least 1 is wrong} \)

Chi-sq
H₀: \{ 2 cats./discrete variables are independent \( H_1: \text{not.} \)

2 pops are homogeneous w.r.t. k categories skipped

F
H₀: \( \mu_1 = \mu_2 = \ldots = \mu_k \)  \( H_1: \text{at least 2 } \mu \text{s are diff.} \)

Note that the ANOVA F-test is a generalization of the 2-sample t-test to more than 2 populations.
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: 157-162) 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. Suppose we want to apply the chi-squared test to this problem. I hope it's clear that only the second test (i.e. test of independence) is possible.

a) What are the categorical variables whose independence can be tested? When you identify them, make sure you also state the number of levels for each.
b) State the hypotheses.
c) Write the data in the form of a contingency table.
d) Compute the expected counts.
e) Compute the p-value (or specify a range for it).
f) State the conclusion "in English."

The following data refer to the melting temperature, y (in some unit), of a certain material at four different pressures, x (in some unit).

<table>
<thead>
<tr>
<th>Pressure</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6</td>
<td>59.5, 53.3, 56.8, 63.1, 58.7</td>
</tr>
<tr>
<td>3.8</td>
<td>55.2, 59.1, 52.8, 54.4</td>
</tr>
<tr>
<td>6.0</td>
<td>51.7, 48.4, 53.9, 49.0</td>
</tr>
<tr>
<td>10.2</td>
<td>44.6, 48.5, 41.0, 47.3, 46.1</td>
</tr>
</tbody>
</table>

a) Make a comparative boxplot of y for the four pressure levels. By R.
b) Based on the above boxplot, would you say there is a difference in the mean melting temperature for at least 2 of the pressure levels?
c) At alpha = 0.05, is there evidence that the mean melting temperature at the at least 2 of the four pressure levels are different? Report the p-value, and state the conclusion clearly. By R; see prelab to see how to do 1-way ANOVA in R.
d) Write code to compute the above p-value "by hand," i.e. without using aov() or lm(), but using the basic formulas for SS_between, SS_within, etc.
e) After (or before) a 1-way ANOVA test, one should check the two assumptions that the y's are normally distributed within each group, and with the same variance. To that end, make a plot that shows four qplots (one for each pressure level) superimposed onto a single figure; make sure that the four qplots have different colors. Are the 4 qplots reasonably straight, and do they have approximately equal slopes? Hint: in the first call to plot(), use xlim=c(-2,2) and ylim=range(y). Use the "By hand" method for making qplots.

Consider the data you collected. Take one of the continuous variables (call it y) and the categorical (or discrete) variable with 3 or more levels (call it x). Since x is discrete/categorical, we can consider each level as a different population. Eg, if your x has 3 levels (say H, M, L), separate the corresponding y's into 3 populations.
a) Do 1-way ANOVA to test if any of the k populations have different means. Report the p-value, and the conclusion in English.
b) Make q-plots of y for each of the levels of X. It's important to have all q-plots superimposed onto a single figure. See "By hand" q-plots in prelab. For example, if your x has 3 levels, then you need to have 3 q-plots superimposed on one plot; e.g.

Recall that equal-slopes translate to equal variances, and so this will be a way of visually checking the homoscedasticity assumption mentioned above (in the lecture note).