Lecture 22 (CH 9)

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-sqd dist. with df = k-1

\{ \text{How does proportion of } \chi^2 = 0 \text{ (something e.g. tornadoes, } X = 1) \}
\text{vary across k categories (or k levels of a categorical var. Y)?}
\text{1 2-level X, 1 k-level Y.}

And the chi-squared test of independence:

\[ H_0: X \text{ and } Y \text{ are independent} \]
\[ H_1: X \text{ and } Y \text{ are not independent} \]

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

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

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)?

\[ X, Y = 2 \text{ Continuous var.s.} \]

How about how does 1 Categ./Discr. var. (X) affect 1 Continuous r.v. (Y)?

This question requires comparing k means, i.e.

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

And the question of whether X affects Y becomes

\[ H_0: M_1 = M_2 = \cdots = M_k \quad (\text{Not } M_i = M_{01}, M_2 = M_{02}, \ldots) \]
\[ H_1: \text{At least 2 } \mu_s \text{ are different.} \]

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

Example: Does knowledge of religion depend on religion?
The Science of "Disestimation": The Shortcomings of Opinion Polls

Why we shouldn't put our faith in opinion polls

By Charles Seife  |  December 14, 2010  |  19

Average number of questions answered correctly (dots)

- Atheist/agnostic
- Jewish
- Mormon
- White evangelical Protestant
- White Catholic
- White mainline Protestant
- Nothing in particular
- Black Protestant
- Hispanic Catholic

Margin of error

\[
\text{by model 1 (e.g., regression)} \quad \text{by model 2 (e.g., neural net)}
\]

Different?

Moral: even though we are testing whether several means are equal, we must pay attention to variance! Hence the name ANOVA.
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</td>
<td>0.6437</td>
</tr>
<tr>
<td>2</td>
<td>0.7205</td>
</tr>
<tr>
<td>3</td>
<td>0.7179</td>
</tr>
<tr>
<td>4</td>
<td>0.7201</td>
</tr>
<tr>
<td>very full</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 5 types of bearings?
- mean computer speed varies across computers?
- mean detection error varies across detection algorithms?
- Are k means different?

Data: Brand 1 | 2 | 3 | 4 | 5
--- | --- | --- | --- | ---
18.1 | | | | |
15.0 | | | | |
14.0 | | | | |
10.6 | | | | |

\[ \bar{y}_1 = 13.6 \quad \bar{y}_2 = 15.7 \quad \bar{y}_3 = 13.7 \quad \bar{y}_4 = 14.7 \quad \bar{y}_5 = 13.8 \]

We are dealing with 5 pop. means.

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

H1: At least 2 \( \mu \)'s are different.

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/psps.

Important & powerful idea
Recall the decomposition of SST from regression. Similarly,

\[ S_{yy} = \frac{k}{n} \sum_{i=1}^{k} \left( \bar{y}_{ij} - \bar{y} \right)^2 \]

Grand mean

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

\[ j^{th} \text{ response in the } i^{th} \text{ pop/cate} \]

\[ S_{yy} = \frac{k}{n} \sum_{i=1}^{k} n_i \left( \bar{y}_i - \bar{y} \right)^2 + \frac{k}{n} \sum_{i=1}^{k} \left( \frac{n_i}{n} \right) \sum_{j=1}^{n_i} \left( y_{ij} - \bar{y}_i \right)^2 \]

\[ \text{SST} = S_{\text{between group}} + S_{\text{within group}} \]

\[ S_{\text{Treat}} \leftarrow \text{Treatment} \]

\[ S_{\text{SSE}} \]

\[ S_{\text{explained}} \leftarrow \text{Variation between groups} \]

\[ S_{\text{unexplained}} \leftarrow \text{Variation within groups} \]

\[ \sim \text{Sample var. of sample means} \]

\[ \sim \text{Sample mean of sample variances} \]

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

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

\[ k = \# \text{ of levels in 1 factor (predictor)} \]

\[ n = \# \text{ of pops.} \]

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

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

**Theorem:**

If $H_0 = \text{True}$, $F = \frac{SS_{between}/(k-1)}{SS_{within}/(n-k)} = \frac{MS_{between}}{MS_{within}}$ has an $F$-distribution with $df = (k-1, n-k)$.

All we need is Table VIII.

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

One 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., $\sigma^2_y = \sigma^2_z = \ldots = \sigma^2_k$. (Called homoscedasticity.)

Use qq plots to test this. See HW 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></td>
<td>13.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>15.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>14.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\overline{y}_1 = 13.68, \quad \overline{y}_2 = 15.97, \quad 13.67, \quad 14.73, \quad 13.08
\]

\[
\text{Var. between} \quad \sigma^2 = \frac{1}{5} (11.94 + \cdots) = 14.22
\]

\[
\text{Var. within} \quad \sigma^2 = \frac{1}{5} (1.167 + \cdots) = 22.83
\]

\[
\overline{y} = \frac{\sum_i \left( \frac{n_i}{n} \right) \overline{y}_i}{6} = \frac{5}{6} (13.68) + \cdots = 14.22
\]

\[
\text{SS battlefield} = \sum_i n_i (\overline{y}_i - \overline{y})^2 = 6 (13.68 - 14.22)^2 + \cdots = 30.88
\]

\[
\text{SS within} = \sum_i \sum_j (y_{ij} - \overline{y}_i)^2 = \frac{5}{6} (n_i - 1) 5.2^2 = (6-1) (1.194)^2 + \cdots = 22.83
\]

\[
F = \frac{30.88/(5-1)}{22.83/(30-5)} = 8.45\left(\text{for df } (5-1, 30-5)\right)
\]

\[
p\text{-value} = p(F > F_{0.01}) = p(F > 8.45) < 0.001 \quad \text{Table VIII.}
\]

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

In English: Brand has an effect on speed.

Which 2? Section 9.3

Skipped.
Most software produce an ANOVA Table (see prelab)

<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 (factor)</td>
<td>k-1</td>
<td>8S\text{between} from formula</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Group (error)</td>
<td>n-k</td>
<td>8S\text{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>0.00018</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:

1. \( Z \):
   - \( H_0: \mu = \mu_0 \)
   - \( H_1: \mu \neq \mu_0 \)
   - \( \sigma \) known/unknown
   - \( Z \) (No t), "large sample"

2. \( t \):
   - \( H_0: \theta = \theta_0 \)
   - \( H_1: \theta \neq \theta_0 \)
   - \( t \) (No Z), "large sample"

3. \( t \):
   - \( H_0: \mu_1 - \mu_2 = \Delta_0 \)
   - \( H_1: \mu_1 - \mu_2 \neq \Delta_0 \)
   - independent or paired

Chi-squared:

- \( H_0: \chi_1^2 = \chi_2^2 = \chi_3^2 = \cdots = \chi_k^2 = \chi_0^2 \)
- \( H_1: \) At least 1 is wrong

Chi-squared:

- \( H_0: \) 2 categorical variables are independent
- \( H_1: \) not
- \( 2 \) pops are homogeneous w.r.t. \( k \) categories

F:

- \( H_0: \mu_1 = \mu_2 = \cdots = \mu_k \)
- \( H_1: \) at least 2 \( \bar{\mu} \)'s are different

Note that the ANOVA F-test is a generalization of the 2-sample t-test to more than 2 populations.
I have said that the F-test of k means is a generalization of the 2-sample t-test of 2 means. Here, we will see some of the math. Looking at the formulas for the F-test, specialized to k = 2, also assume \( n_1 = n_2 = n/2 \).

a) Show that \( \bar{Y} = \frac{1}{2} (\bar{Y}_1 + \bar{Y}_2) \)

b) Show that \( \text{SS}_{\text{between}} = \frac{1}{n} (\bar{Y}_1 - \bar{Y}_2)^2 \)

c) Show that \( \text{SS}_{\text{within}} = \frac{1}{2} (n-2) \left( S_1^2 + S_2^2 \right) \)

d) Show that \( F = \frac{1}{2} n \left( \frac{\bar{Y}_1 - \bar{Y}_2}{S_1^2 + S_2^2} \right) \)

(Note that in a 2-sample test of \( H_0 : \mu_1 - \mu_2 = 0 \), \( H_1 : \mu_1 - \mu_2 \neq 0 \), we have \( t = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \sqrt{n} \frac{\bar{Y}_1 - \bar{Y}_2}{S_1^2 + S_2^2} = \sqrt{n} \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{S_1^2 + S_2^2}} \).

and so, the F in part d) is nothing but \( t^2 \).)

**hw-lect22-2**

By hand or by R (see prelab)

Return to the data you collected. Take one of the continuous variables (call it \( y \)) and the categorical/discrete variable with 3 or more levels (call it \( x \)). Since \( x \) is discrete/categorical we can consider each level as a different population. E.g., if your \( x \) has 3 levels (say, \( H, M, L \)) separate the corresponding \( y \)'s into 3 categories.

a) Do 1-way ANOVA to test if any of the \( k \) populations have different means. Report the p-value, and the conclusion.

b) Note: in a previous hw, you made qq plots of the \( y \)'s in each population. Return to that work, and comment on whether their slopes are comparable (i.e., approximately equal). Recall that equal slopes means equal variances, and so this will be a way of visually checking the homoscedasticity assumption mentioned above (in the lecture notes).