Dec. 10 Statistics-related quotation for the day: “Math is the only subject in the curriculum where adults are proud to say they are lousy.” — Alfred S. Posamentier, former dean of School of Education, CUNY (though the original link for this quotation appears to have been removed)

Assignment: Read Chapter 25
Do exercises 1, 6, 8, 9, and 10 on pages 479-480

"Bonferroni correction” for multiple tests

Very simplistic way to (over-)correct for multiple testing problem: Take your preferred p-value cutoff, like 0.05, and divide it by the number of tests performed.

The Bonferroni correction means that we don’t declare any results statistically significant unless their p-values are smaller than this new, smaller number.

With 240 STAT 100 students in FA05, apply a Bonferroni correction to the previous situation.

"Bonferroni correction”? (Can you do the calculation?)

Remember these steps!
1. Hypotheses
2. Test statistic
3. p-value
4. Decision

How to calculate a test statistic

The test statistic in this class is always a STANDARDIZED SCORE:
\[
\frac{\text{estimate} - \text{null value}}{\text{standard deviation of estimate}}
\]

(The "estimate" is an estimate of the population parameter of interest.)

Return to chi-squared statistics

Suppose we are interested in the following research question:
Is there a difference between men and women at Penn State with respect to the proportion who have smoked marijuana?

According to the 2008 survey for this class, 42.1% of women (out of 126) versus 62.0% of men (out of 92) have smoked marijuana.

Counts and percents: Fall 2008

<table>
<thead>
<tr>
<th>Rows: Sex</th>
<th>Columns: marijuana</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Female</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>57.9%</td>
</tr>
<tr>
<td>Male</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>38.0%</td>
</tr>
</tbody>
</table>

So 42.1% of women in the sample say yes and 62.0% of men in the sample say yes.

Are they statistically significantly different?

Aside: In 2005, it was 57.6% of women and 59.7% of men.
After finding the expected counts if $H_0$ is true, compute a chi-squared statistic (refer to lecture 17 for step-by-step details).

Green: Observed counts
Red: Expected counts if skeptic is correct.

\[
\begin{array}{c|c|c}
 & (73 - 62.4)^2 & (53 - 63.6)^2 \\
62.4 & 1.80 & 1.77 \\
45.6 & 2.46 & 2.42 \\
\end{array}
\]

\[\text{Chi-Squared} = 1.80 + 1.77 + 2.46 + 2.42 = 8.45\]

Recall that we rejected if chi-squared > 3.84

\[\text{Our chi-squared value: } 8.45 \text{ (from Step 2)}\]

How about a p-value for the marijuana test?

The key is to take the square root of the chi-squared statistic and treat that as the standardized score.

Null: No difference between men & women
Alternative (2-sided): A difference exists

\[\text{Standardized score: } \sqrt{8.45} = 2.91\]

2-sided p-value: Between 2x.0013 = .0026 and 2x.005 = .01

Decision: WE DO HAVE EVIDENCE THAT A GREATER PERCENTAGE OF MEN THAN OF WOMEN HAVE SMOKED MARIJUANA.

The seven warnings of Chapter 24 paraphrased

Warning #1. The word significant has both a common and a statistical meaning. Know which is which.

Warning #2. Even small effects are often statistically significant if the sample size is large.

Warning #3. Low power (which can result from too-small samples) can lead to type 2 errors.

Warning #4. It is often helpful to consider confidence intervals together with hypothesis tests.

Warning #5. One-sided tests can miss effects in the opposite direction from the one tested.

Warning #6. Don’t cheat! The choice between one-sided and two-sided must be made before seeing the data.

Warning #7. In multiple testing situations, “rare” occurrences (low p-values) will happen occasionally even if the null is true.

Meta-analysis

- A collection of statistical techniques for combining studies.
- By combining many studies, we may sometimes be able to obtain a large “meta-study” that helps to answer difficult questions that are not clear from smaller studies.
Vote-counting method

- Simply find all studies on a particular topic and count how many had found a statistically significant result.
- This is generally a bad idea. Consider this example:

Imagine taking a single study involving 1000 participants and breaking it up into 100 studies of 10 participants each. Probably, none of the 10-participant studies would amount to anything statistically significant even if the larger study would.

Vote-counting example (p. 436, #14)

Suppose ten studies were done to assess the relationship between watching violence on television and subsequent violent behavior in children. Suppose that none of the ten studies detected a significant relationship.

Is it possible for a vote-counting procedure to detect a relationship? Is it possible for a meta-analysis to detect a relationship? Explain.

No

Yes

Using just vote-counting, we see 0 out of 10 significant results! However, a meta-analysis can combine these ten studies, giving in effect one large sample that might be enough to show a statistically significant effect.

Which studies should be included?

Different studies may differ widely in their quality of work. Often, many studies must be eliminated from a meta-analysis because it is not absolutely clear that what is being studied in them is the desired focus of the research.

A meta-analysis of the effect of behavior on blood pressure eliminated all but 26 out of 857 possible studies!

Should studies be compared or combined?

If one wishes to combine studies, make sure they're really measuring the same thing on the same population!

Consider two studies comparing surgery to relaxation for treating chronic back pain. One is conducted at a back-care specialty clinic, the other at a suburban medical center.

Where will the people with the most severe back pain go? The two studies are probably conducted on different populations. (We'll revisit this example later...)