4/15/2023

Review Exam #2
ANOVA
χ^2 Test of Independence

Hypothesis tests so far:
One-sample hypothesis tests—

- Tests of means
  - T-test – most cases!!
  - Z-test – when you have both a large sample size and the population standard deviation
- Tests of proportions

Two-sample hypothesis tests—

- Test of means
  - Dependent/paired test – one sample test for the differences between the pairs
  - Independent tests
    - Pooled – assumes population standard deviations are the same
    - Unpooled – assumes population standard deviations may be different
- Tests of proportions – differences between 2 proportions

What if there are three samples, or more?

If we are testing 3 or more means: use ANOVA (**An**alysis **o**f **Va**riance). It is equivalent to the Pooled t-test if you use it for two samples because it assumes that the standard deviations of all the datasets are equal.

Difference of proportion tests (two-samples) can be displayed in raw form in two-way tables. In such cases, there are only two columns (and two rows) of data.

The chi-squared (χ^2) test tests two-way tables for independence: with three or more columns, or three or more rows.

ANOVA – one-way ANOVA
Measurements of a continuous variable and that measured under different conditions that have three or more levels.

The null hypothesis for an ANOVA test is always the same:
All the means are the same

$$H_0: \mu_i = \mu_j, for\ all\ i \neq j$$

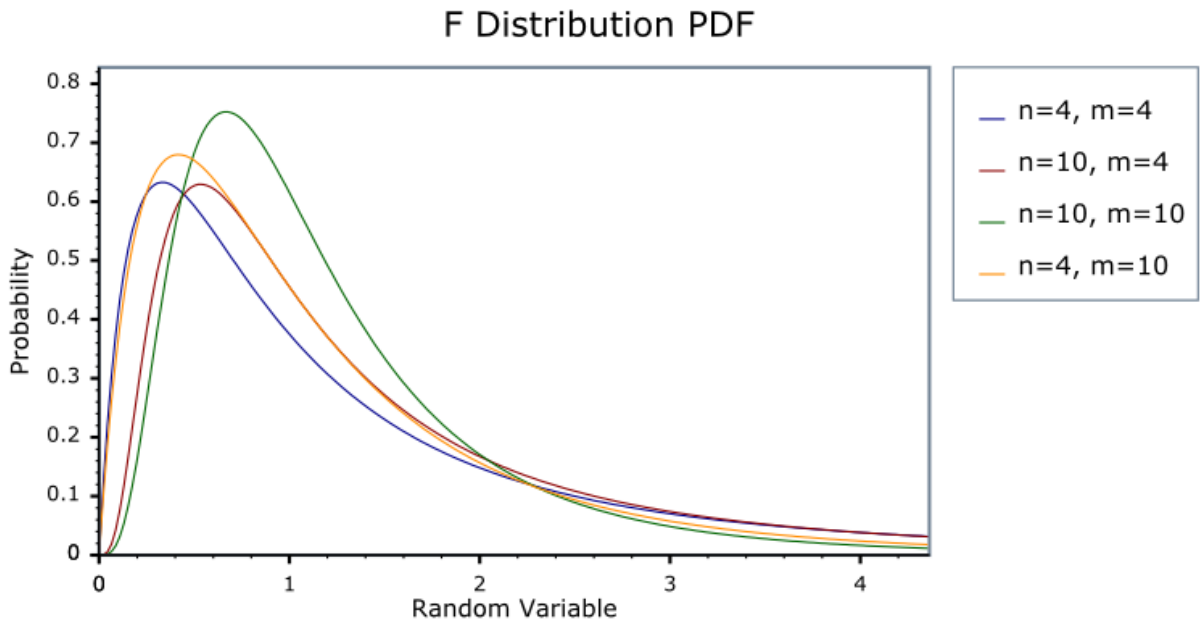
The alternative is that at least one mean is different

$$H_a: \mu_i \neq \mu_j, for\ some\ i \neq j$$

One alternative to doing an ANOVA is doing repeated two-sample t-tests. In the case of 3 levels, we could do three 2-samples t-tests. That's a lot of testing, and the number of two-sample tests increases as the number of levels increases.

If the ANOVA null is not rejected, then the means are all equal (approximately), and so we can avoid doing all the two-sample tests. If the ANOVA null is rejected, then we have good reason to wonder which groups are different from the others, and then there may be a reason to do the two-sample tests to tease out those differences.

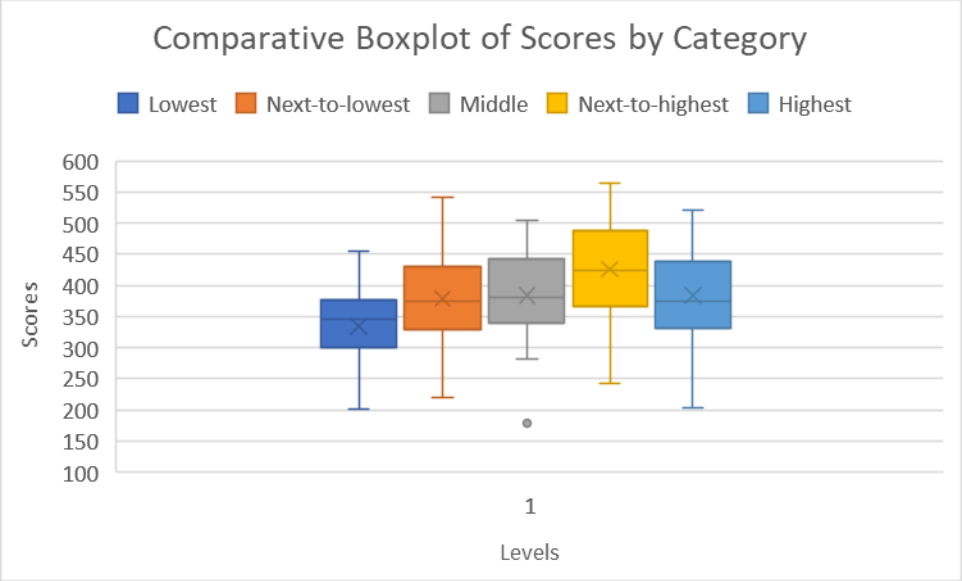
The ANOVA makes a strong claim that the sample standard deviations of each group are the same (similar enough).



ANOVA

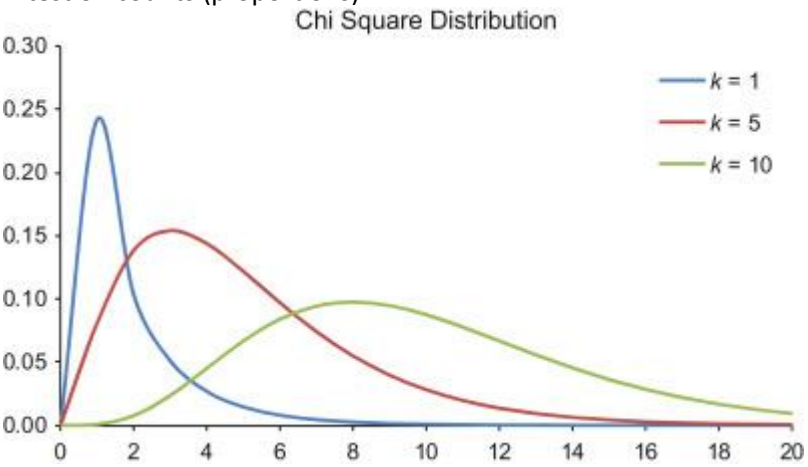
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	104807.68	4	26201.92	4.581401874	0.001772104	2.447236511
Within Groups	686303.12	120	5719.192667			
Total	791110.8	124				

The p-value is less than 0.05, so we can reject the null hypothesis. Based on the boxplot, is that at least Lowest category and the Next-to-highest category (since those are the most different) are producing this result. But the categories could be grouped in different ways, so to tease out the groupings, we would need to do additional testing.



χ^2 test

A test on counts (proportions)



Tests of independence for two-way tables.

The degrees of freedom for our test is based on the size of the two-way table: m categories in the rows and the n categories in the columns, the degrees of freedom for the test is (m-1)(n-1).

Null hypothesis for the test of independence is that the variables in the columns and rows of the table are independent.

Alternative is that they are dependent.

The test statistic is $\chi^2 = \sum \frac{(O-E)^2}{E}$

	Baseball	Basketball	Football	Total
Male	13	15	20	48
Female	23	16	13	52
Total	36	31	33	100

Men/Baseball: O is 13
Expected value: P(M)*P(B)*100= 48/100*36/100*100=48*36/100

General Rule: Row Total * Column Total ÷ Grand Total

For Excel: you need to create the two-way table, and then calculate (at a minimum) the expected table. Then the chisq.test() function will get you the p-value. Don't use any totals in the test function.

If you use the calculator, you need the values in the two-way table, and then put into a matrix.

Next: the last topics we will cover is regression.