Math 244 Lecture Notes

CHAPTER 26 DAY ONE: ANALYSIS OF VARIATION (ANOVA)

Overview: Last class, we compared multiple counts/proportions. This time, we will look at comparing multiple averages.

Example 1. Mr. Wherry likes to make bread. Recently, he has begun to experiment with different brands of flour to see if the average amount of time needed to knead dough varies by company. His initial data for the needed time in minutes is listed below:

Arthur's Flour	Bob's Red Mill	Cost Smart
20	23	35
30	26	36
25	27	37
28	33	38
26	30	36
26	31	30
21	31	37
22	32	38
26	29	37
24	30	36
	29	37
	32	37

(a) Calculate the average amount of time needed to knead for ALL brands.

(b) Calculate the average and standard deviation for each brand. As a reminder,

$$s^2 = \Sigma \frac{(x_i - \bar{x})^2}{n - 1}.$$

(c) Does there appear to be a difference in brands?

It feels like we should be doing a hypothesis test. But it's unclear what the model should be and how it should work. Let's look at some examples to see what's going on:

Example 2. Which of the following sets of side-by-side box-plots appear to have a difference in average?

Observations:

We have to study two types of variation: (1) Variation between groups and (2) Variation within groups. We will then compare the two types of variation.

Let's bring that variance equation up again:

$$s^2 = \Sigma \frac{(x_i - \bar{x})^2}{n - 1}.$$

In the most general sense, we have

$$Variance = \Sigma \frac{(value - center)^2}{df}.$$

Variation Between Groups: We want to compare each of the individual averages. What is the "center"? How will we do this?

Variation Between Groups can be found using the formula:

It has a *df* of ____

The variation between groups for the flour brands was _____

Variation Within Groups: We need something to compare with the variation between groups (comparing centers). Luckily, we have the individual spreads of those boxes that inform us of the individual spread that is typical. So, let's **POOL** together each of those variances to get a collective variance.

Variation Within Groups (Based off the SE's) can be found using the formula:

$$s^2_{pool} =$$

It has a *df* of ____

The variation within groups for the flour brands was _____

Episode IV: A New Distribution Okay. We have two variances, but how do we compare them? Informally, we consider the Variation Between Groups to be the amount of spread between the centers and the Variation Within Groups to be the amount of spread we are allowed in typical situations. As such...

$F = \frac{Var\ Between}{Var\ Within}$ is the basis for the F-Distribution. You can think of F as ROUGHLY being the % of allowed variation taken by the differences.


Which way do we shade?

Assume that the averages are very similar...

(a) Should the variation between groups be large or small?

(b) Should our F ratio be large or small?

(c) Do we want to reject or fail to reject the null-hypothesis?

(d) Do we want a large or small P-value?
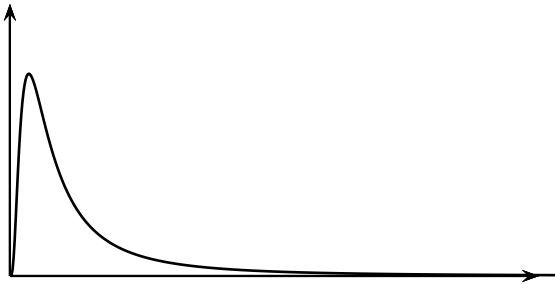
(e) SHADE IN THE CURVE:

Assume that the averages are very different...

(a) Should the variation between groups be large or small?

(b) Should our F ratio be large or small?

(c) Do we want to reject or fail to reject the null-hypothesis?

(d) Do we want a large or small P-value?

(e) SHADE IN THE CURVE:
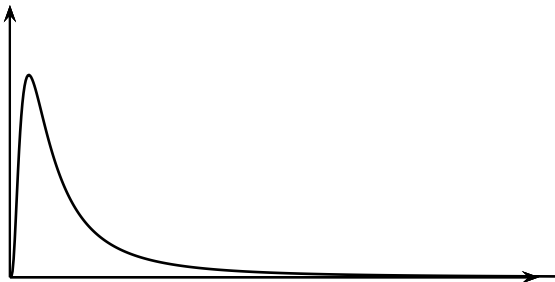
Our F-Score:

OBSERVATION:

Much like χ^2, ANOVA, which uses the F-distribution, is always a right-tailed test. We can find the P-value by doing the following:

TI-83 and TI-84: [2nd]$\rightarrow$[Vars]$\rightarrow$[Fcdf].
TI-89: [Apps]$\rightarrow$[Stat/List]$\rightarrow$[F5:Dist]$\rightarrow$[Fcdf].

The general setup is the same for all calculators. We need to type in "Fcdf(F-score, ∞, $df_{Between}$,df_{Within})". Our P-value=_____

Example 3. Assuming that the null-hypothesis is that all brands of flour work equally well on average and the alternative hypothesis is that at least one brand is different, what conclusion can we draw based off our P-value?

How can we figure out which group is different?

Here is the general process from start to finish:

STEP I: State the hypotheses. $H_0 : \mu_1 = ... = \mu_k$ and H_A : At least one average is different.

STEP II: When comparing multiple averages, we use the F-Distribution. Where...

$$\text{Variation Between Groups} = \Sigma \frac{n_i(\bar{x}_i - \bar{X})^2}{k - 1}$$

with k the number of groups and $\bar{X}$ the overall average AND

$$\text{Variation Within Groups} = s^2_{pool} = \Sigma \frac{df_i(s_i)^2}{N - k}$$

with N the total number of values.

$$F = \frac{\text{Variation Between Groups}}{\text{Variation Within Groups}}$$

has df's of $k - 1$ and $N - k$ in that order.

The assumptions are

    (a) Random

    (b) Independence/10% Rule

    (c) Large n:

    (d) Indepedent Groups

    (e) Equal Spread:

STEP III: ANOVA is always a right-tailed test.

We find the P-value using "Fcdf(F-score, ∞, $df_{Between}$,df_{Within})"

STEP IV: Conclusion

Example 4. Mr. Wherry wants to know if there is an average difference in GPA (out of 4.0) for daytime versus evening students. Daytime: $\bar{x} = 2.86$, $s = 1.4$, and $n = 20$. Evening: $\bar{x} = 3.16$, $s = 1.6$, and $n = 30$. Use ANOVA to compare.