

## LECTURE 27: ANALYSIS OF VARIANCE

### I. Analysis of Variance

- a. We've previously used the t-test to compare the means of two populations. But sometimes we want to compare the variance (is one sample more consistent than another). Other times we want to compare multiple samples at the same time.
  - i. Why don't we just run the t-test several times, once for each pair of samples? Because (A) that gets really time consuming and (B) the Central Limit Theorem. Remember CLT states that sample means will approximate a normal distribution. Any two samples might be different based on chance (say, two on opposite sides of the distribution). You'll commit Type I Error.
  - ii. Example: You're curious if a particular type of occupation likes a new product more than others. You might compare just teachers and factory workers. Then you add writers, and then police officers, and then musicians, etc. The more points of comparison you have, the more likely a difference will be statistically significant by chance. Type I Error.
  - iii. Example: You're curious if a new soap is better than a current brand. So you ask the treatment and control a battery to assess based on several metrics: smell; color; shape; feeling of cleanliness right after use; cleanliness an hour after use; six hours after use; quickness to lather; etc. The more metrics you add, the more likely you are to conclude its different based solely on chance. Type I Error.
- b. We can correct this problem with *Analysis of Variance (ANOVA)*—a method for testing if the means of several groups are equal; it is a generalized t-test.
- c. ANOVA is determined by the F-ratio, or

$$F = \frac{\text{estimation of } \sigma^2 \text{ between different samples}}{\text{estimation of } \sigma^2 \text{ within different samples}} = \frac{ns_{\bar{x}}^2}{\sum s^2/k}$$

- i. The variance in the numerator assumes each sample average comes from the same standard deviation (in other words, the null hypothesis is assumed to be true and there is no difference

between samples). It's calculated by treating each mean as an observation. That's multiplied by the number of observations in each sample.

1. For example, if the averages are 62, 66, and 52, then the average of those averages is 60.
  2. The variance is thus  $(2^2+6^2+8^2)/(3-1) = 52$ .
  3. If each sample mean comes from a sample of 5, the numerator is 260.
- ii. In the denominator, we add the variances from each sample and then divide by the number of samples; it's our average variance. The variable  $k$  is the number of samples.
    1. Suppose the variances for each sample are: 27.5, 26.5, and 31.0; the average variance is thus 28.33.
  - iii. Note this equation assumes the sample size in each sample,  $n$ , is the same.
- d. The bigger the F-ratio, the higher the chance that you'll reject the null hypothesis. If you have a lot of observations in each sample ( $n$  is large), that makes  $F$  large.
- i. The F-ratio in our example is thus  $260/28.33 = 9.18$ .
- e. ANOVA relies on the F-distribution, which is really a family of distributions depending on the degrees of freedom in the numerator and the degrees of freedom in the denominator.
- i. The distribution is positively skewed and ranges from zero to infinity.
  - ii. For the numerator,  $df_1 = k - 1$ ; in our example, it's 2.
  - iii. For the denominator,  $df_2 = nk - k^1$ ; in our example, it's 12.
  - iv. Both of these degrees of freedom would output significance levels: 0.05 would have a critical score of 3.89; 0.01 would have a critical score of 6.93.
  - v. In other words, it's significant at 99% confidence.

## II. Using ANOVA:

- a. *Two-sample for variance*—check to see if two samples have the same variance.
- b. *Single factor*—check if samples are different. For example, “Do these different routes from A to B take different amounts of time?”
- c. *Two factor*—check if samples are different across samples and within samples. This calculation is different from the single factor even for

---

<sup>1</sup> Again, this assumes the number of observations in each sample are the same. If it's not equal across samples, you can change this to simply all observations summed across all samples. 15 in this case.

the differences across samples (which both return). For example, “Do these different routes from A to B take different amounts of time *and*—since several different drivers drove each route—does the average time between drivers differ?”

- i. In other words, two factor ANOVA checks if the average is different between one type of sample (here, the driver)...

	<b>Travel Time from Starbrick to Warren</b>			
<b>Driver</b>	<i>U.S. 6</i>	<i>West End</i>	<i>Hickory</i>	<i>Rte. 59</i>
<i>Adams</i>	18	17	21	22
<i>Brockman</i>	16	23	23	22
<i>Callahan</i>	21	21	26	22
<i>Davies</i>	23	22	29	25
<i>Edgeworth</i>	25	24	28	28

- ii. ...and then uses that information to check if the average is different between another type of sample (here, the route).

	<b>Travel Time from Starbrick to Warren</b>			
<b>Driver</b>	<i>U.S. 6</i>	<i>West End</i>	<i>Hickory</i>	<i>Rte. 59</i>
<i>Adams</i>	18	17	21	22
<i>Brockman</i>	16	23	23	22
<i>Callahan</i>	21	21	26	22
<i>Davies</i>	23	22	29	25
<i>Edgeworth</i>	25	24	28	28

- d. It’s crucial to understand that the two factor ANOVA assessment will get a different result than the one factor ANOVA, even for the same treatment. If we did one factor ANOVA for routes, we’d find the routes take the same time. A two factor ANOVA would find that the route does matter. Different drivers are faster at different routes.

- i. Which one to use depends on if your secondary treatment factor (here, the driver) is important to consider. Which one to use is often a judgment call.