Math 266 - Probability & Statistics - Spring 2013 - Prof. Janeba
Analysis of Variance, or ANOVA:

Suppose we have samples from several groups and wish to determine if the groups' averages are the same or not. For example, suppose we draw a simple random sample of 6 students from each of three dorms, and find the following G.P.A.s:

Dorm A	Dorm B	Dorm C
2.75	2.39	3.04
2.92	2.92	3.17
2.94	3.15	3.27
3.18	3.19	3.36
3.23	3.33	3.62
3.43	3.49	3.98

The question is, are the three dorms' averages for GPAs the same? The samples' averages of GPAs are NOT the same, but perhaps we just picked unrepresentative samples from some or all of the dorms, by the luck of the draw. This will be our null hypothesis: The averages of each dorms' students' GPAs are all equal (and, p.s., it's because each sample is just a separate sample from a common population - so not only should the means be the same, but so should the variances.)

We will evaluate our null hypothesis by analyzing the variance of individual student GPAs in two ways. Remember that, like the standard deviation, variance is a measure of how widely spread out the data is; if the variance is small, it means the list of numbers is tightly clustered.

Since our hypothesis says that all the dorms' averages (and variances) of GPA are the same, then each sample of 6 students is like 6 values of the same random variable X. We will try to estimate that X's variance two ways:

1. By making a list of sample (dorm) averages (one average per sample) - note these are three values of the random variable $\overline{X}$. We will use the sample SD of those values of $\overline{X}$ (i.e. $S_{\overline{X}}$) to estimate $\sigma_{\overline{X}}$. We use that estimate for $\sigma_{\overline{X}}$ in the reverse of the usual calculation, to calculate/estimate σ_X.
2. By finding the variances of the individual samples (not of their averages, as in #1), and averaging them.

If these two estimates agree closely, then our null hypothesis, that the averages of each dorms' students' GPAs is the same, is plausible. More broadly, it is plausible that each sample is taken from the same overarching population. In this case, the samples vary between themselves about as much as one would expect from samples that vary internally as much as our samples do.

If these two estimates of σ_X differ greatly: Our estimate of $\sigma_{\overline{X}}$ will tend to give a bigger prediction for σ_X than the average of the variances of the samples. This means that the samples are spread out (from each other) more than samples from a single random variable should be - therefore, we are led to believe that we weren't sampling from the same random variable after all, and that the three dorms' GPAs are actually different.

Here's the details:

3. State the null hypothesis: All dorms have the same average for their students' GPAs (and the same variance).
4. Describe the random variable: The GPA of a randomly chosen student from *any* of these dorms. We have three samples of this variable, each drawn from a specific dorm. Since the sample averages have the same expected value, pooling them won't change that expected value (or *average*). We attempt to estimate the pooled samples' variance in two ways:
5. Treat the three columns as three samples. Their averages should vary by about $\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$; in this case $n = 6$.

	Dorm A	Dorm B	Dorm C
	2.75	2.39	3.04
	2.92	2.92	3.17
	2.94	3.15	3.27
	3.18	3.19	3.36
	3.23	3.33	3.62
	3.43	3.49	3.98
average	3.075	3.0783	3.4067

Find the averages of the columns/samples as shown, and then compute the sample SD (or $S_{\overline{x}}$) of those averages (use a calculator). We use sample S.D. since we have only three samples from which we are estimating the S.D. of the collection of all possible samples, i.e. $\sigma_{\overline{X}}$. The sample S.D. of 3.075, 3.0783, and 3.4066 is 0.1905.

But remember, $\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$, so $0.1906 \approx \frac{\sigma_X}{\sqrt{6}}$, and thus $\sigma_X = 0.1906\sqrt{6} \approx .4669$. So our first estimate for the σ_X^2 is $\sigma_X^2 = (.4669)^2 \approx 0.2180$.

6. The next step is simpler: We estimate the variance of the box by finding the variance predicted by each individual sample, and average them. To find the variance of the first column, find its sample SD and square: $(0.2486)^2 = 0.06180$. Repeating for each column, we get:

	Dorm A	Dorm B	Dorm C
	2.75	2.39	3.04
	2.92	2.92	3.17
	2.94	3.15	3.27
	3.18	3.19	3.36
	3.23	3.33	3.62
	3.43	3.49	3.98
σ_X^2 est	.06180	0.1498	0.1171

Now average these variance estimates: $\frac{.06180+0.1498+0.1171}{3} \approx 0.1096$.

7. Now we compute the ratio
$\frac{\text{variance estimate from variance of column averages}}{\text{variance estimate from average of variances of columns}} = \frac{0.2180}{0.1096} \approx 1.989$
Note: We can remember this as "the variance of the averages over the average of the variances," or as "variance between sample (averages) over variance within samples." Briefly, the numerator and denominator of our fraction have (almost) Chi-squared distributions (if you multiplied them by their degrees of freedom, see below), and as they are, the fraction is a random variable with a well-studied distribution, $F-$ distribution. We can find out more about the $F-$ distribution in chapter 8.6.

8. The degrees of freedom for the numerator are (# of columns -1)=(# samples -1), 2 in this case. The degrees of freedom for the denominator are (total # of table entries - # of columns), 15 in this case.

9. We can use a calculator/computer to see that: $P(F_{2,15} \geq 1.989) = 0.1713$. Since we computed F based on H_o, this means that: If H_o is true, and the separate dorms are really just three samples from the same overall population, the chance that the between-sample variance is this much larger than the within-sample variance is about 17%. While this is low, it's not low enough to convince most statisticians to reject H_o.

10. Alternately, if we decide in advance to test at the $\alpha =$, say, .05, we can look in a table for the F-statistic and see if $F = 1.989$ is higher than the 5% significance level (the level that makes $P = .05$): 3.68 is the F-table's entry, so since our F is lower, there is *more* than a 5% chance of this much variation between samples, and we *accept* the null hypothesis.