The Analysis of Variance
(Devore Chapter Ten)

MATH-252-01: Probability and Statistics II[*]

Spring 2019

Contents

Tuesday 19 March 2019

Guest Lecturer: Dr. Robert Parody; independent notes provided for reference

1 The ANOVA F-Test

We now consider a generalization of the two-sample problem, where we now have multiple samples, each possibly from a different population. Following the notation of Devore, we will call the number of samples I and label the individual samples with $i = 1, \ldots, I$. The size of sample i will be written J_i, and we'll label the individual observations in the ith sample with $j = 1, \ldots, J_i$, and write x_{ij} as the j observed value in the ith sample, and X_{ij} as the corresponding random variable. Sometimes (e.g., in Section 10.1 of Devore) we'll assume all the samples are of the same size, which we'll call J, so that the total number of observations is IJ; in general it is $n = \sum_{i=1}^{I} J_i$. The different populations from which the samples are drawn are sometimes called *treatments*, based on an experimental design in which we have I treatments (e.g., medicines), each of which is applied to J test subjects and the responses $\{x_{ij}\}$ recorded.

We'll consider the most basic test, in which the samples are all assumed to come from normal populations with the same variance σ^2, i.e., $X_{ij} \sim N(\mu_i, \sigma^2)$. The null hypothesis H_0 will be that all of the populations are identical, $\mu_1 = \mu_2 = \ldots = \mu_I$, while the alternative hypothesis H_a is that at least one mean μ_i is different from at least one of the others. An obvious starting point for a statistic comparing the means of the I populations is to construct the sample mean of each of the I samples, which we write as

$$\overline{X}_{i\bullet} = \frac{1}{J_i} \sum_{j=1}^{J_i} X_{ij} \qquad (1.1)$$

As with all sample means, we know $E(\overline{X}_{i\bullet}) = \mu_i$ and $V(\overline{X}_{i\bullet}) = \sigma^2/J_i$. Now, if we knew the common variance σ^2 and had a

[*] Copyright 2019, John T. Whelan, and all that

hypothesized value μ for the common mean, we could use the fact that

$$\frac{\overline{X}_{i\bullet} - \mu}{\sigma/\sqrt{J_i}} \tag{1.2}$$

is standard normal if H_0 is true to construct a test statistic

$$\sum_{i=1}^{I} \frac{(\overline{X}_{i\bullet} - \mu)^2}{\sigma^2/J_i} = \frac{\sum_{i=1}^{I} J_i(\overline{X}_{i\bullet} - \mu)^2}{\sigma^2} \tag{1.3}$$

which, if H_0 is true, should be a chi-squared random variable with I degrees of freedom. If H_a is true, the statistic should have larger than expected values because of the mismatch between each μ_i and μ.

Of course, we don't know μ or σ, so we have to estimate them from the data, using the "grand mean"

$$\overline{X}_{\bullet\bullet} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J_i} X_{ij} = \frac{\sum_{i=1}^{I} J_i \overline{X}_{i\bullet}}{\sum_{i=1}^{I} J_i} \tag{1.4}$$

and corresponding variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{I}\sum_{j=1}^{J_i}(X_{ij} - \overline{X}_{\bullet\bullet})^2 \tag{1.5}$$

which would point towards a statistic something like

$$\frac{\sum_{i=1}^{I} J_i(\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})^2}{S^2} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J_i}(\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})^2}{\frac{1}{n-1}\sum_{i=1}^{I}\sum_{j=1}^{J_i}(X_{ij} - \overline{X}_{\bullet\bullet})^2} \tag{1.6}$$

The numerator and denominator look sort of similar, and have interesting interpretations as two of three obvious measures of variability of the data:

- The "total sum of squares" (SST)

$$\text{SST} = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(X_{ij} - \overline{X}_{\bullet\bullet})^2 \tag{1.7}$$

  measures the overall variability of the data set as a whole.
- The "treatment sum of squares" (SSTr)

$$\text{SSTr} = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})^2 \tag{1.8}$$

  measures the variability between samples (treatments).
- The "error sum of squares" (SSE)

$$\text{SSE} = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(X_{ij} - \overline{X}_{i\bullet})^2 \tag{1.9}$$

  measures the variability within samples (treatments).

An important identity is

$$\begin{aligned}\text{SST} &= \sum_{i=1}^{I}\sum_{j=1}^{J_i}[(X_{ij} - \overline{X}_{i\bullet}) - (\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})]^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J_i}(X_{ij} - \overline{X}_{i\bullet})^2 - 2\sum_{i=1}^{I}\sum_{j=1}^{J_i}(X_{ij} - \overline{X}_{i\bullet})(\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet}) \\ &\quad + \sum_{i=1}^{I}\sum_{j=1}^{J_i}(\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})^2 = \text{SSE} + \text{SSTr}\end{aligned} \tag{1.10}$$

where the cross term is zero because

$$\sum_{j=1}^{J_i}(X_{ij} - \overline{X}_{i\bullet}) = \sum_{j=1}^{J_i} X_{ij} - J_i\overline{X}_{i\bullet} = 0 \tag{1.11}$$

The test statistic we've been considering is thus

$$\frac{\text{SSTr}}{(n-1)\text{SST}} = \frac{\text{SSTr}}{(n-1)(\text{SSE}+\text{SSTr})} = \frac{1}{(n-1)(1+\text{SSE}/\text{SSTr})} \tag{1.12}$$

This means we can get an equivalent test by talking about SSTr/SSE. If the means of the different populations are different, SSTr will be be bigger than expected, because the different treatments (populations) will have more variability in their estimated means. Both SSTr and SSE are, up to a scaling, chi-squared random variables when H_0 is true:

$$\frac{\text{SSTr}}{\sigma^2} = \sum_{i=1}^{I} \frac{(\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})^2}{\sigma^2/J_i} \sim \chi^2(I-1) \tag{1.13}$$

and

$$\frac{\text{SSE}}{\sigma^2} = \sum_{i=1}^{I}\sum_{j=1}^{J_i} \frac{(X_{ij} - \overline{X}_{i\bullet})^2}{\sigma^2} \sim \chi^2(N-I) \tag{1.14}$$

The number of degrees of freedom in SSTr is $I-1$ because we've used one degree of freedom among the I means to construct the grand mean $\overline{X}_{\bullet\bullet} = \frac{\sum_{i=1}^{I} J_i \overline{X}_{i\bullet}}{\sum_{i=1}^{I} J_i}$. The numer of degrees of freedom in in SSE is $n-I$ because we've used I degree of freedom among the n observations to construct the I individual means $\{\overline{X}_{i\bullet}|i=1,\ldots,I\}$. It's less obvious but still straightforward to show that these two chi-squared random variables are independent and therefore that

$$F = \frac{\text{SSTr}/[(I-1)\cancel{\sigma^2}]}{\text{SSE}/[(N-I)\cancel{\sigma^2}]} = \frac{\text{MSTr}}{\text{MSE}} \tag{1.15}$$

follows the F-distribution introduced in the last lecture, with parameters $\nu_1 = I-1$ and $\nu_2 = n-I$. (If all the samples have the same size $J_i = J$, then $n = IJ$, so $\nu_2 = IJ - I = I(J-1)$.)

The testing procedure is then this: given the data, construct the "mean square for treatments"

$$\text{MSTr} = \frac{1}{I-1}\sum_{i=1}^{I} J_i(\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})^2 \tag{1.16}$$

which is an estimate of the common variance σ^2 if the means $\{\mu_i\}$ are all the same and the "mean square for error"

$$\text{MSE} = \frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{J_i}(X_{ij} - \overline{X}_{i\bullet})^2 \tag{1.17}$$

which is an estimate of σ^2 using all the samples with no assumption about the means. The ratio of these

$$f = \frac{\text{MSTr}}{\text{MSE}} \tag{1.18}$$

is a test statistic which is $F(I-1, n-I)$ distributed if H_0 is true. We perform an upper-tailed test on this F statistic.

Because this test involves comparing two different estimates of the variance, it is part of the field known as Analysis of Variance, or ANOVA. Because there is only one set of populations in the model, this is known as one-way ANOVA. Statistical software typically outputs the various quantities into a table of the form

Source of		**Sum of**	**Mean**	
Variation	**dof**	**Squares**	**Square**	f
Treatments	$I-1$	SSTr	MSTr	MSTr/MSE
Error	$n-I$	SSE	MSE	
Total	$n-1$	SST		

Practice Problems

10.1, 10.3, 10.5, 10.9, 10.27

Thursday 21 March 2019

Guest Lecturer: Dr. Joseph Voelkel; independent notes provided for reference

Dr. Voelkel's notes available at `http://ccrg.rit.edu/~whelan/courses/2019_1sp_MATH_252/TukeyMultipleComparisons1up.pdf`

2 Tukey's Multiple-Comparisons Test

If the ANOVA F-test rejects the null hypothesis, we can say the if the data $\{x_{ij}\}$ represent I random samples from normal distributions with the same variance, $X_{ij} \sim N(\mu_i, \sigma^2)$, the data are inconsistent with all of the $\{\mu_i\}$ being equal. However, the test results don't say anything about which $\{\mu_i\}$ are different. To extract that information, an additional test is necessary. For simplicity, we'll limit attention to the case where all of the samples are of the same size J. The obvious check for the $\{\mu_i\}$ being different is to compare the treatment means $\{\overline{x}_{i\bullet} | i = 1, \ldots, I\}$ to each other, but how different do they have to be to be significant? Each treatment mean, as a random variable, is $\overline{X}_{i\bullet} \sim N(\mu_i, \sigma^2/J)$, so the scale is going be be set by $\sigma/\sqrt{J}$, and our best estimate of σ^2 is

$$\text{MSE} = \frac{1}{IJ - I} \sum_{i=1}^{I} \sum_{j=1}^{J} (X_{ij} - \overline{X}_{i\bullet})^2 \tag{2.1}$$

We have to be a bit careful, though, since the more treatments there are, the more likely we are to have two of the sample means differ just by chance. One approach is to set a *simultaneous* confidence interval on all of the differences between the $\{\mu_i\}$. This is done by noting that

$$\frac{\overline{X}_{i\bullet} - \mu_i}{\sigma/\sqrt{J}} \sim N(0, 1) \tag{2.2}$$

and

$$(IJ - I)\frac{\text{MSE}}{\sigma^2} \sim \chi^2(IJ - I) \tag{2.3}$$

This means that the combination[1]

$$\frac{(\overline{X}_{i\bullet} - \mu_i) - (\overline{X}_{j\bullet} - \mu_j)}{\sqrt{\text{MSE}/J}} = \frac{(\overline{X}_{i\bullet} - \overline{X}_{j\bullet}) - (\mu_i - \mu_j)}{\sqrt{\text{MSE}/J}} \tag{2.4}$$

is the difference between two standard normal random variables divided by the square root of a chi-squared random variable which has been divided by its number of degrees of freedom. It turns out, given the way these variables were constructed, they are all independent random variables. There is a less well-known probability distribution for this situation known as the *Studentized range distribution*. If you have a sample of size m from a standard normal distribution (i.e., with $\{\frac{\overline{X}_{i\bullet} - \mu_i}{\sigma/\sqrt{J}}\}$ with $m = I$) and an independent chi-squared random variable with ν degrees of freedom ($(IJ - I)\frac{\text{MSE}}{\sigma^2}$ with $\nu = IJ - I$), the largest difference between all the possible pairs of standard normals, scaled by the square root of the chi-squared per degree of freedom, is a random variable which follows a studentized range distribution with parameters m and ν. The probability density function of this distribution isn't even written down in closed form, but the percentiles and tail probabilities are tabulated basically because of Tukey's test.[2] Devore collects some of them in Table A.10 and refers to the 100 × $(1 - \alpha)$th percentile as $Q_{\alpha,m,\nu}$.

[1] Note that we're now using j to label a different treatment, not an observation within treatment i.

[2] They're not in Python, but in R the cdf is `ptukey()` and the quantiles are `qtukey()`. (R does not bother to define `dtukey()` for the distribution function or `rtukey()` to generate random variables from the distribution.)

So the probabilistic statement refers to the maximum difference between offsets between the sample and population mean for each treatment, and can be written

$$1-\alpha = P\left(\max_{i,j=1,\dots,I}\frac{(\overline{X}_{i\bullet}-\overline{X}_{j\bullet})-(\mu_i-\mu_j)}{\sqrt{\text{MSE}/J}} \le Q_{\alpha,I,IJ-I}\right) \tag{2.5}$$

If this is true of the maximum, it must be true for each of the possible pairs of treatments, so:

$$\begin{aligned}1-\alpha = P\Big(&\forall_{i,j=1,\dots,I}\overline{X}_{i\bullet}-\overline{X}_{j\bullet}-Q_{\alpha,I,IJ-I}\sqrt{\text{MSE}/J}\\ &\le \mu_i-\mu_j \le \overline{X}_{i\bullet}-\overline{X}_{j\bullet}+Q_{\alpha,I,IJ-I}\sqrt{\text{MSE}/J}\Big)\end{aligned} \tag{2.6}$$

I.e., for each pair of treatments i and j, we can construct a confidence interval of

$$\overline{x}_{i\bullet}-\overline{x}_{j\bullet}\pm Q_{\alpha,I,IJ-I}\sqrt{\text{MSE}/J} \tag{2.7}$$

and there is a probability $1-\alpha$ that this all of the intervals will include the corresponding difference of means. So, for each difference of means whose confidence interval does not include zero, we declare that difference to be significant.

Practice Problems

10.xx, 10.xx, 10.xx