

STA 113 Spring 2005

I. H. Dinwoodie

Summary of One-Way Analysis of Variance (ANOVA)

Consider the data of lesion lengths on rats subject to three treatments (Placebo, Old, and New) in the file `filloon.txt`. Each treatment group corresponds to a value or “level” of a single variable (treatment type for lesions) so this is called “one-way” data at 3 levels. The data can be thought of as three groups of numbers, $x_{P,j}, x_{O,j}, x_{N,j}$, with true unknown means or expected values equal to μ_P, μ_O, μ_N for each treatment group.

One basic question is whether the true unknown means of lesion lengths in the treatment groups are all equal. This is written

$$H_0 : \mu_P = \mu_O = \mu_N.$$

The statement above of equal means is called the “null hypothesis” and is treated as the defendant in a procedure that looks for contradictory evidence.

The method of ANOVA computes first a measure of “signal” that would look for any difference in the means μ_P, μ_O, μ_N . Then ANOVA gets a measure of “noise” that gives the amount of background randomness. Finally, ANOVA computes the ratio, called the F-statistic (3.6317 on this data). The p -value (written `Prob>F` in Matlab) is a way to judge the size of this number. The smaller the p -value, the larger the signal-to-noise ratio.

The sample sizes in the three groups (the range of the index j) are 33, 32, and 34, which are not equal, so the formulas in the book on p. 415 do not apply. The measure of signal (called Mean Square Treatment, or `Groups MS` in Matlab) is

$$\text{Groups MS} = \frac{33(\bar{x}_P - \bar{x})^2 + 32(\bar{x}_O - \bar{x})^2 + 34(\bar{x}_N - \bar{x})^2}{(3 - 1)} = 42.24$$

where \bar{x}_P is the average of the lesion lengths in the Placebo group (actually 1.5456 mm), \bar{x}_O is the average of the lesion lengths in the Old treatment group, \bar{x}_N is the average of the lesion lengths in the New treatment group, and \bar{x} is the overall average across all treatment groups. This measures the variability in the treatment means across the three groups.

The “noise” measure is called Mean Square Error (or `Error MS` in Matlab), and is given by

$$\text{Error MS} = \frac{\sum_{j=1}^{33} (x_{P,j} - \bar{x}_P)^2 + \sum_{j=1}^{32} (x_{O,j} - \bar{x}_O)^2 + \sum_{j=1}^{34} (x_{N,j} - \bar{x}_N)^2}{33 - 1 + 32 - 1 + 34 - 1} = 11.63$$

This is a measure of how much variability there is within each treatment group. The quantity `s` in the output below is $\sqrt{\text{Error MS}}$.

```

anova =

    'Source'    'SS'          'df'    'MS'          'F'          'Prob>F'
    'Groups'    [ 84.4812]    [ 2]    [42.2406]    [3.6317]    [0.0302]
    'Error'     [1.1166e+03] [96]    [11.6309]    []          []
    'Total'     [1.2011e+03] [98]    []          []          []

```

```

stats =

    gnames: {3x1 cell}
           n: [33 32 34]
    source: 'anova1'
    means: [1.5456 3.7174 2.0218]
    df: 96
    s: 3.4104

```

Finally, we need a way to tell whether the signal-to-noise ratio

$$3.63 = \text{Groups MS/Error MS}$$

is large or not. We find the probability of seeing a ratio as big or bigger than 3.63 if all the means were the same. This gives us an idea of how big 3.63 is relative to all the numbers that would occur if there were no real signal, just noise or randomness. This comparison uses the F-distribution, with parameters (2,96) (called degrees of freedom) equal to the denominators in the two MS calculations. So in Matlab

```
>>1-fcdf(3.6317,2,96)
```

```
ans =
```

```
0.0302
```

Anything less than 0.05 indicates a very small probability that the signal-to-noise ratio could be so big as 3.6317, if in fact there had been no real signal. The value .03 says that you would see the signal-to-noise ratio of at least 3.63 only 3% of the time if there were no signal. So it leads to the conclusion here that there is some difference in the three means, the signal-to-noise ratio F is quite large.

Then you have to go to other methods, such as boxplots or two sample comparisons, to find out where the differences are.