Analysis of Variance R^2 and F-test

Elad Gilboa

December 23, 2011

the total sample variance $s_p = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. now assume that our n samples are a collections from two different groups of samples, i.e., $n_1 + n_2 = n$. we can write the sample deviation (variance) s_p as:

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n_1+n_2-1}\sum_{i=1}^{n_1+n_2}(x_i - \bar{x})^2 \quad (1)$$

$$= \frac{1}{n_1+n_2-1}\left(\sum_{i=1}^{n_1}(x_{1i} - \bar{x})^2 + \sum_{i=1}^{n_2}(x_{2i} - \bar{x})^2\right) \quad (2)$$

where $\bar{x}$ is the mean for the entire joined sample.

$$\bar{x} = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1}x_i + \sum_{i=1}^{n_2}x_i\right). \quad (3)$$

In the sums from the right side, we can add and subtract a term $\bar{x}_1$, $\bar{x}_2$, respectively, which are the sample means of each group.

$$\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1 + \bar{x}_1 - \bar{x})^2 = \sum_{i=1}^{n_1}\left((x_{1i} - \bar{x}_1) + (\bar{x}_1 - \bar{x})\right)^2 \quad (4)$$

$$= \sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + 2\underbrace{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)}_{0}(\bar{x}_1 - \bar{x}) + \sum_{i=1}^{n_1}(\bar{x}_1 - \bar{x})^2 \quad (5)$$

$$= \sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + n_1(\bar{x}_1 - \bar{x})^2 \quad (6)$$

$$(7)$$

Putting it all together, we get that the total deviation equals the explained deviation, which is what we would have predicted from having two samples with different means, and the unexplained deviation, which is the result of random fluctuations due to noise.

$$\underbrace{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}_{\text{Total deviation}} = \underbrace{\frac{1}{n-1}\sum_{r=1}^{2}\sum_{i=1}^{n_r}(x_{ri} - \bar{x}_r)^2}_{\text{Unexplained deviation}} + \underbrace{\frac{1}{n-1}\sum_{r=1}^{2}n_r(\bar{x}_r - \bar{x})^2}_{\text{Explained deviation}} \quad (8)$$

Here is when the two tests diverge. The R^2 test checks for the $\frac{\text{explained variance}}{\text{total variance}}$, where the F-test checks for $\frac{\text{explained variance}}{\text{unexplained variance}}$.

1 R^2 Coefficient of determination test

In statistics, the coefficient of determination R^2 is used in the context of statistical models whose main purpose is the prediction of future outcomes on the basis of other related information. It is the proportion

of variability in a data set that is accounted for by the statistical model. It provides a measure of how well future outcomes are likely to be predicted by the model. As I understand it, in our case, our model is having two groups at location $\bar{x}_1$, and $\bar{x}_2$.

To get $R^2 = \frac{\text{explained variance}}{\text{total variance}}$, we divide Eq (8) by the left side term:

$$\begin{aligned} 1 &= \frac{\frac{1}{n-1}\sum_{r=1}^{2}\sum_{i=1}^{n_r}(x_{ri}-\bar{x}_r)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} + \frac{\frac{1}{n-1}\sum_{r=1}^{2} n_r(\bar{x}_r-\bar{x})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} && (9) \\ &= C^2 + R^2 && (10) \end{aligned}$$

where C is some constant and R^2 is what we are looking for. As both C^2 and R^2 are positive and their sum is 1, this means $0 \leq R^2 \leq 1$. Solving for R^2 and using Eq. (3), we get

$$\begin{aligned} R^2 &= \frac{\sum_{r=1}^{2} n_r(\bar{x}_r-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} && (11) \\ &= \frac{\frac{\left(\sum_{i=1}^{n_1} x_{1i}\right)^2}{n_1} + \frac{\left(\sum_{i=1}^{n_2} x_{2i}\right)^2}{n_2} - G}{\sum_{i=1}^{n_1} x_{1i}^2 + \sum_{i=1}^{n_2} x_{2i}^2 - G} && (12) \end{aligned}$$

where $G = \frac{\left(\sum_{i=1}^{n_1} x_{1i} + \sum_{i=1}^{n_2} x_{2i}\right)^2}{n_1+n_2}$

2 F-test

An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis. It is most often used when comparing statistical models that have been fit to a data set, in order to identify the model that best fits the population from which the data were sampled. In this case we will be comparing two models: the samples came from two distinct distribution that is stimulus dependent, or the samples came from a single distribution.

In the F-test we are looking for $F = \frac{\text{explained variance}}{\text{unexplained variance}}$, which can also be written as $F = \frac{\text{between-group variability}}{\text{within-group variability}}$.

$$F = \frac{\frac{\left(\sum_{i=1}^{n_1} x_{1i}\right)^2}{n_1} + \frac{\left(\sum_{i=1}^{n_2} x_{2i}\right)^2}{n_2} - G}{\sum_{i=1}^{n_1} x_{1i}^2 + \sum_{i=1}^{n_2} x_{2i}^2 - \left(\frac{\left(\sum_{i=1}^{n_1} x_{1i}\right)^2}{n_1} + \frac{\left(\sum_{i=1}^{n_2} x_{2i}\right)^2}{n_2}\right)}(n_1+n_2-2) \quad (13)$$