Review: ANOVA and Measures of Association

XDASI Fall 2021

11/18/2021

Contents

ANOVA

ANOVA and related non-parametric methods allow us to compare differences in a ***quantitative*** variable between ***categorical*** groups. Using the sums-of-squares method, we saw that we can partition the total variation among and between groups, as shown in W&S Fig, 15.1-2:

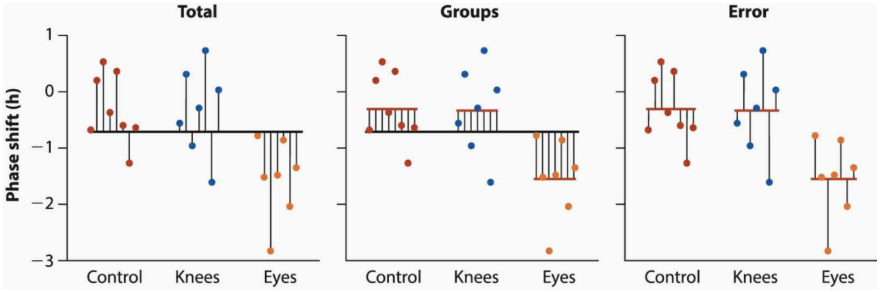

Figure 1: W&S Fig. 15.1-2: Partitioning the total variation

For any individual data point, we can separate its distance from the overall mean into two parts: its distance to its own group mean, and the distance from its group mean to the overall mean.

$$(Y_{ij} - \bar{Y}) = (Y_{ij} - \bar{Y}_j) + (\bar{Y}_j - \bar{Y})$$

where

- $\bar{Y}$ is the overall mean,
- $\bar{Y}_j$ is the mean of group j, for $j \in \{1..m\}$ and m is the number of groups
- Y_{ij} is the value of data point j in group i, for $i \in \{1..n_j\}$ and n_j is the number of data points in group j.

It's easy to see that the group terms on the right ($\bar{Y}_j$) cancel each other out.

Sums of Squares

To get the total variation in the data, we just sum up all the squared differences from every data point to the grand mean. Similarly, we can sum up the the squared differences from each data point to its individual group mean, and the squared differences between the group means and the total mean:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}-\bar{Y})^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}-\bar{Y}_j)^2 + \sum_{j=1}^{m} n_j(\bar{Y}_j-\bar{Y})^2$$
$$SS_{total} \quad = \quad SS_{error} \quad + \quad SS_{group}$$

Mean Squared Error

Recall that the ***variance*** of a random sample is defined as $s^2 = \frac{\sum(y_i-\bar{Y})^2}{n-1}$, where the degrees of freedom are $df = n-1$.

So, the variance is essentially the ***mean sum of squares*** of a random variable. More generally, we can write $MSS = \frac{SS}{df}$. For ANOVA with m groups, we can write:

$$MS_{group} = \frac{SS_{group}}{df_{group}} = \frac{\sum_j n_j(\bar{Y}_j-\bar{Y})^2}{m-1}$$

$$MS_{error} = \frac{SS_{error}}{df_{error}} = \frac{\sum_i\sum_j(y_{ij}-\bar{Y}_j)^2}{N-m}$$

where N is the total number of points, $N = \sum_j n_j$, and m is the number of groups.

F-statistic

The F-ratio is the ratio of the between-group variance to the within-group variance:

$$F = \frac{MS_{group}}{MS_{error}} = \frac{SS_{group}/df_{group}}{SS_{error}/df_{error}}$$

If there is no difference between the groups, then $F = 1$; otherwise, $F > 1$, and the upper-tail probability determines how significant the differences between groups are.

The critical value for the F-statistic is determined by the allowable Type I error rate, α (usually $\alpha = 0.05$), and the degrees of freedom for SS_{group} and SS_{error}:

$$F_{crit} = F_{(1-\alpha),(m-1),(N-m)}$$

Interestingly, when there are only two groups, and so $df = 1$, the F-statistic is the same as the square of the 2-sample t-statistic: $F = t^2$. This may seem a little weird, but just take my word for it. Recently someone posted a proof of this on their blog, which you can read if you want! [1]

R-squared

R^2 is the ***amount of variation in the data that is explained by the groups***, i.e. the proportion of the total variation that is due to variation between groups:

$$R^2 = \frac{SS_{group}}{SS_{total}}$$

R^2 ranges from $0 \leq R^2 \leq 1$.

- If the groups are all drawn from the same population, then almost none of the total variation will be due to differences between groups, so R^2 will be close to zero.
- On the other hand, if the groups are very different, then R^2 will be close to 1, since the differences between groups will account for almost all of the variation in the dataset.

Nonparametric alternatives to ANOVA

ANOVA assumes that the data are normally distributed and that they have approximately equal variances. When these are not met, other options are available:

- ***Kruskal-Wallace*** test
- ***pairwise Wilcoxon Rank Sum test*** using ***adjusted p-values*** for multiple hypothesis testing; most commonly these are
  - ***Bonferroni*** (which controls for the family-wise error, FWR) and is the most conservative
  - ***Benjamini-Hochberg*** (which controls for the false discovery rate, FDR)

Measures of Association

To measure the strength of association between ***two quantitative variables***, we have seen that we can use the ***covariance***, or the more useful ***linear correlation coefficient***.

Covariance

The covariance can easily tell us if there is a positive, negative, or no association between two variables. It is defined as:

$$Cov(X, Y) = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{n - 1}$$

Again, notice that this equation has exactly the same form as that for the variance of a single variable: $s^2 = \frac{\sum(x_i - \bar{X})(x_i - \bar{X})}{n-1}$! **Cool.** So now we've seen that the ***mean squared error*** and the ***covariance*** both have the same form as the ***variance***! Keep this in mind for later.

The ***drawback of the covariance*** is that it is not so easy to interpret:

[1] https://canovasjm.netlify.app/2018/10/29/when-does-the-f-test-reduce-to-t-test/

- it is not bounded, i.e. it ranges from $-\infty$ to $+\infty$
- it varies with sample size
- it changes with scale, even when the underlying relationship is the same (e.g. X ranging from 0-20 vs. 0-40)

Correlation

To solve these problems, we can simply ***normalize*** the covariance by the variance of the individual variables. This is called the ***linear correlation coefficient, r***:

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{Cov(X,Y)}{s_X s_Y}$$

The correlation coefficient r is easier to interpret because has the following advantages:

- its range is bounded: $-1 \leq r \leq 1$
- it is unaffected by the scale of the data

r quantifies the ***strength*** of a linear relationship between two variables. A single variable is perfectly correlated with itself, so $Cor(X,X) = 1$. When two variables are completely uncorrelated, $r = 0$ *(warning: the reverse is not true, e.g. $y = x^2$!)*

One caveat to be aware of is that r may differ depending on the ***range*** of the data analyzed. This is illustrated in W&S Fig. 16.4-1, which shows that computing r over just a small portion of the available data does not reveal the same correlation.

Our ***confidence*** in the significance of any non-zero correlation depends on the ***value of r***, coupled with the ***amount of data***. For example, any two points will have a correlation of 1, but that's not very significant! On the other hand, the chance of being able to draw a straight line through 3 random points is very low.

The ***p-value*** for r quantifies the probability that some number of random data points will show a certain correlation, and it depends ***only*** on the value of r and the number of data points. It is calculated in the usual way following a t-distribution, using the test statistic

$$t = \frac{r}{SE_r} \quad , \; where \; SE_r = \sqrt{\frac{1-r^2}{n-2}}$$

The higher the r, and the more data we have, the better confidence we can have in our ability to make inferences about our data. This means that for the same r, the dataset with a lot more data points will have a lower p-value. On the other hand, a low r could have a significant p-value and still have very low predictive value due to the large amount of variation in the data.

To get a feel for r, check out ***this fun online demo*** from the W&S online companion site.

R-squared

One problem with r is that it's not so easy to compare two different r values. Is $r = 0.8$ twice as good as $r = 0.4$? Well, it's not really that clear.

However, something interesting happens if we take the square of r:

$$r^2 = \frac{Cov(X,Y)^2}{Var(X)Var(Y)} = R^2$$

Now we have a new variable, R^2, that ranges from $0 \leq R^2 \leq 1$. So, it does not contain any information about the ***direction*** of the relationship.

However, R^2 is great because it measures the ***amount of variation*** that's due to the ***association*** between two variables, relative to their ***individual variation***. It has a formal name, which is the ***coefficient of determination***. To help you remember the proper name, just ask the question, *"How much of the variation is **determined** by the relationship between variables?"*

Unlike r, R^2 ***can*** be used to directly compare two different sets of data. For example, an $R^2 = 0.5$ means that the association between variables explains 50% of the total variation in the data, whereas $R^2 = 0.25$ explains 25% of the variation in the data. So, the amount of variation explained by the association between X and Y in the first dataset is two times that for the second dataset.

Although it can be formulated in different ways, R^2 represents the same idea whether we are looking at variation between groups in ANOVA, or association between two (or more) quantitative variables: it is the ***fraction of the total variance that is explained by the association between variables, relative to the total variance***.

We will revisit R^2 again from a slightly different perspective as related to least squares linear regression.

Note: *r^2 only quantifies the association betwen two variables, whereas R^2 is a more general measure that can be applied to any number of variables. They are only the same in the special case of two variables.*

Spearman's Rank Correlation

When the data do not follow a bivariate normal distribution (such as when the variation in Y varies with X, or there are outliers, or there is a nonlinear relationship), then other approaches must be used to test for correlation between two quantitative variables.

The same options exist as for univariate data: try transforming the data to make it look more normal, or use a rank-based test. Spearman's Rank Correlation computes a correlation using the ranks of the data, called r_S, and signficance is calculated using a t-statistic in the usual manner.