5.01 Analysis of Variance: One-way ANOVA

ANalysis Of VAriance allows us to compare means of more than two groups. It's a method of analysis we use in research designs with a quantitative response variable and one or more categorical independent variables.
The simplest type of ANOVA is one-way ANOVA, with just one independent variable that distinguishes three or more groups. In this video I'll explain the basics of one-way ANOVA and the logic behind using *variances* to decide something about *means*.

In ANOVA independent variables are often referred to as factors. The factors are categorical variables that represent **groups** - often experimental or quasi-experimental **conditions**, also referred to as the **levels** of a factor. People can also refer to groups or levels as the **cells** in a factorial design.

In one-way ANOVA there's just one factor, with three or more levels. Suppose we want to compare healthiness of three groups of cats that consume different diets: Raw meat, canned food and dry food. A veterinarian rates their health on a scale from zero to ten.

How do we determine whether one or more groups differ in healthiness from the other groups? The first thing you might think of is to create an extended version of the t-test, by adding the third mean to the formula for the t-test. This will not work however, because the shape of the resulting test-statistic distribution does not correspond to a stable, known probability distribution, so we can't calculate p-values.

The next thing you might think to do is to perform three pairwise t-tests, comparing two group means at a time. This is not a good idea because the **Family-Wise Error Rate (FWER)** will be inflated. The error rate is the probability of falsely rejecting the null hypothesis - and falsely concluding there is a significant effect, when the null is in fact true. If we perform one test the error rate is equivalent to the significance level, which we determine ourselves and normally set at five percent.

When we perform more than one test, the *Family-Wise* Error Rate refers to the probability of ***at least*** *one of these tests* resulting in a false rejection of the null hypothesis. This probability is approximately equal to the number of tests times the significance level. If we want to keep the Family-Wise Error Rate at the desired significance level of $\alpha = 0.05$ for example, we could correct for this inflation by dividing the significance level we use for the individual tests by the number of tests.

So if we perform three pairwise tests, the individual significance levels would become 0.050 / 3 = 0.017. The p-values for the individual tests are much less likely to be smaller than 0.017 compared to 0.050. In other words, if we apply this correction we have much less power to detect a difference between the groups, when there is a true difference in the population.

If we use ANOVA we don't have to worry about an inflated error rate. ANOVA allows us to decide whether the groups are samples from the same population distribution, with one and the same mean, or whether they are from different population distributions with different means. It does so by comparing the variance in the response variable *between* the groups with the variance *within* the groups.

So how can variances tell us something about means? Well it requires an assumption and a trick: First we *assume* the variance is the same in populations. If there is any difference between the populations this should be a difference in the means only. The *trick* is to estimate the population variance in two different ways.
The first method will always result in a fairly accurate and precise estimate of the population variance, whether the population means are different or the same. The second method will produce a fairly accurate and precise population variance estimate if the means are the same, but it will *overestimate* the population variance if the group means differ. So we can detect a difference in means by observing a discrepancy between the two methods of estimating the variance.

The first method that is robust whether the population means differ or not, estimates the population variance with the **within-group variance**: the variance within each group, averaged over the groups. If the group sizes are equal the formula for estimating this variance is very easy, the variances are added and divided by the number of groups: $MS_{within} = \frac{\sum s_j^2}{g}$. If the group sizes differ we employ this formula with the raw sums of squares in each group: $MS_{within} = \frac{\sum\sum(y_{ij} - \bar{y}_j)^2}{n-g} = \frac{SS_{within}}{n-g}$

The second method, that is *not* robust when the population means differ, estimates the population variance with the **between-group variance**: the variance of the means. If the population means are the same we still expect to find some differences in the *sample* means. More population variation will generally result in more variation in sample means, so although it is not a very efficient way of estimating the variance in the population, the variation in

sample means can be used to estimate the population variance.
However, if the population means differ this will result in additional variation in the sample means, resulting in an overestimation of the population variance.
To calculate the between-group variance we need to know the grand mean, the mean of the means, which we calculate by multiplying each group mean with the number of observations in that group, adding these together and then dividing by the total number of groups: $\bar{y} = \frac{\sum n_j \cdot \bar{y}_j}{g}$.
We calculate the between group variance by taking each group mean, subtracting the grand mean and squaring the difference, multiplying by group size, summing these squared differences and dividing by the number of groups minus one: $MS_{between} = \frac{\sum n_j (\bar{y}_j - \bar{y})^2}{g-1}$.

Finally, we compare the two estimates of the population variance by considering their ratio, which is associated with the F probability distribution. We divide the between-group variance by the within-group variance. If the population means are the same, we expect both methods to result in the roughly the same, fair estimate. So under the null hypothesis we expect a ratio close to one. It will never be smaller than zero, since there is always some variation in the samples and variances are always positive.

If the population means differ, we expect the between-group variance in the sample to overestimate the population variance and to be larger than the within-group variance. In this case we expect a ratio larger than one. There is no maximum value.

If the ratio is so large that the associated p-value is smaller than the significance level we reject the null hypothesis that the population means are equal and accept the alternative hypothesis that at least one of the groups differs from the rest. We don't know which group or groups differ and in what direction. The ratio of the between and within group variances provides an overall test, equivalent to the overall test in multiple regression.