8. F-test *and* distribution, analysis of variance (ANOVA)

F-test

Normally distributed data can be described by two parameters – mean and variance. We discussed testing the difference in the mean between two samples in previous chapter. However, it is also possible to test whether two samples come from population with the same variance, i.e. the null hypothesis stating:

$\sigma^2_1 = \sigma^2_2$

as usual for population parameters, we do not know the σ but they can be estimates by s^2 (sample variances). A comparison between sample variances is then done by F-test

$$F = \frac{s_1^2}{s_2^2}$$

which is a simple ratio between sample variances. The F statistic follows F distribution, shape of which is defined by two degrees of freedom – DF numerator and DF denominator. These are found as $n_1 - 1$ and $n_2 - 1$ (i.e. number of observations in corresponding sample – 1). When reporting test results in a text, both DFs must be reported (usually as subscripts). For instance, variances significantly differed between green and red apples ($F_{20,25} = 2.52$, $p = 0.015$).

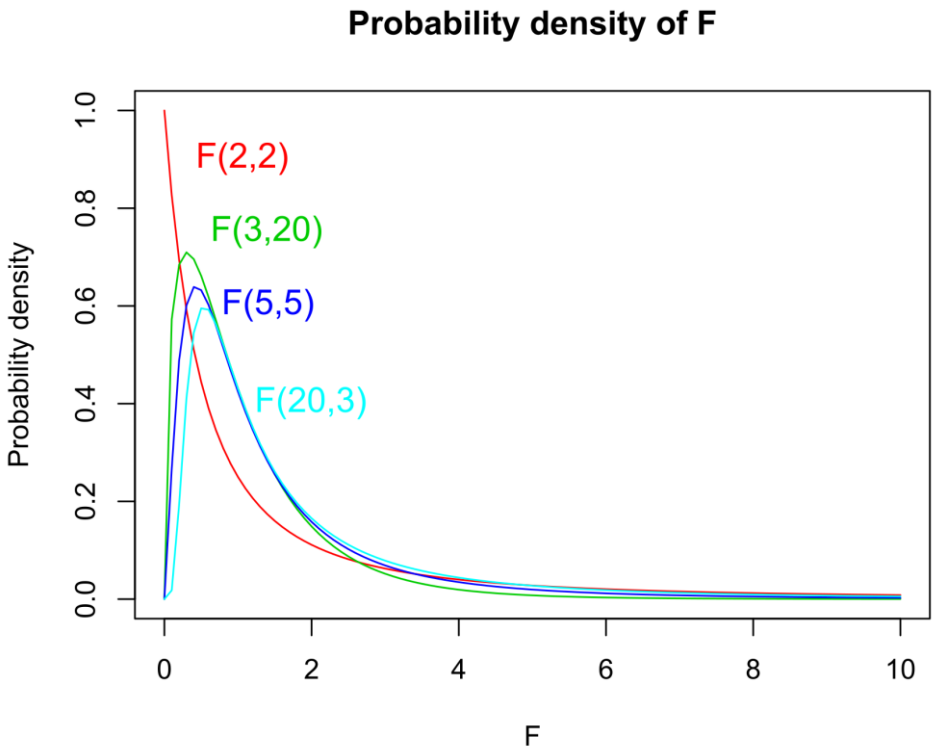

Fig. 8.1 Probability density plot of F-distributions with different DFs.

Analysis of variance (ANOVA)

F-test is rarely used to test the differences in variance between two samples because hypotheses on variance are not that common. However, F-test has its crucial application in analysis of variance.

In chapter 7, we discussed comparison between the means of two samples using t-test. A natural question however arises – what if we have more than two samples? We may try pairwise comparisons between each pair of samples. That would however lead to multiple non-independent tests and result in inflated type I error probability[1]. Therefore, we use analysis of variance (ANOVA) to solve such problems.

ANOVA tests a null hypothesis on means of multiple samples, which states that the population means are equal, i.e.

$\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$

The mechanism of ANOVA is based on decomposing the total variability into two components: 1. systematic component corresponding to differences between groups and 2. error (or residual) component corresponding to differences within groups. These differences are measured as squares. For each observation in the dataset, its total square (measuring difference between its value and the overall mean), effect square (measuring difference between corresponding group mean and the overall mean), and error square (measuring difference between the value and corresponding group mean) can be calculated (Fig 8.2).

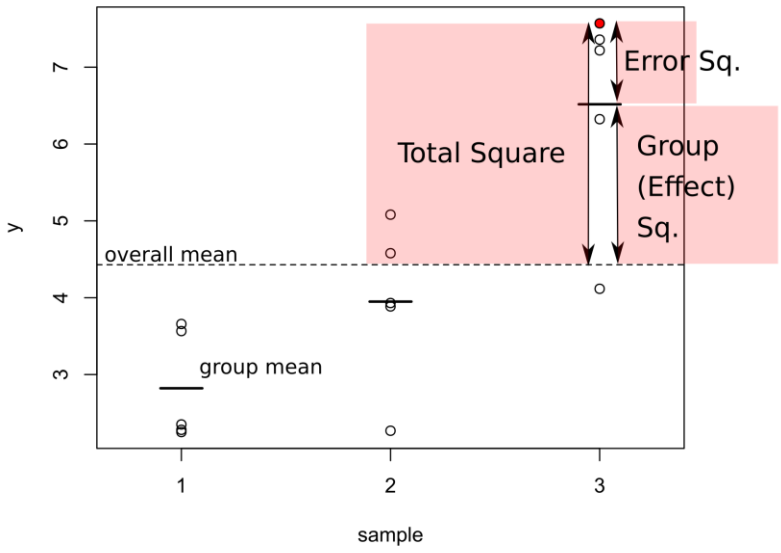

Fig. 8.2 Mechanism of ANOVA: definition of squares exemplified with the red data point.

Subsequently, we can summarize the square statistics over the whole dataset by summing to get sums of squares (SS): SS_{total}, SS_{effect}, SS_{error}. We can further calculate mean squares (MS) by dividing SS by corresponding DF, with $DF_{total} = n - 1$, $DF_{effect} = k - 1$, and

[1] This comes from the fact that if individual tests are performed at $\alpha = 0.05$, then probability of making type I error in 2 tests (i.e. making error in at least one of the test) is $p = 0.05+0.05-0.05^2 = 0.975$.

$DF_{error} = DF_{total} - DF_{effect}$, where *n* is total number of observations and *k* number of categories. Hence we get:

$MS_{effect} = SS_{effect}/DF_{effect}$

$MS_{error} = SS_{error}/DF_{error}$

and now, it comes: **the mean squares are actually variances**. As a result, we can use an F-test to test null hypothesis that MS_{effect} is higher that MS_{error} which is equivalent to the test that all means are equal:

$F_{DFeffect,DFerror} = MS_{effect}/ MS_{error}$

the corresponding *p*-value can be then found based on comparison with F distribution as in an ordinary F-test. Note, that rejecting the null hypothesis means, that *all* means are not equal, i.e. at least one of the means is significantly different from at least one other.

In addition, to the p-value, it is also possible to compute proportion of variability explained by the groups:

$r^2 = SS_{effect}/SS_{total}$

Typical report of ANOVA result in the text then reads: Means were significantly different among the groups ($r^2 = 0.70$, $F_{2,12} = 14.63$, $p = 0.0006$).

ANOVA assumptions

ANOVA application assumes that i. samples come from normally distributed populations and variances are equal among the groups. These assumptions can be checked by **analysis of residuals** as they can be restated as i. normal distribution and ii. constant variance of residuals.

There are formal tests testing for normality, such as the Shapiro-Wilk test, but their use is problematic as they test the null hypothesis that given sample comes from normal distribution. The tests are more powerful (likely to reject the null) if there are many observations, but in that case, ANOVA is rather robust to moderate violations of the assumption. By contrast, the formal tests of normality fail to identify the most problematic cases, when the assumptions are not met and also the number of observations is low.

Instead, I highly recommend visual check of the residuals. In particular, scatterplot of standardized residuals and normal quantile-quantile (QQ) plots (https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot) are informative about possible problems with ANOVA assumptions.

Post-hoc comparisons

When we get a significant result in ANOVA (and only in such case!), we may be further interested to see, which mean is different from which. Statistical theory does not provide much help here, however some pragmatic tools were developed in this respect. These are

based on the principle of pair-wise comparisons (similar to a series of pair-wise two-sample t-tests), which however control for inflation of type I error probability by adjusting the p-values upwards. An example of such test is Tukey honest significant difference test (Tukey HSD).

Results of these tests are frequently summarized in plots by letter indices with different letters indicating significant differences (Fig. 8.3)

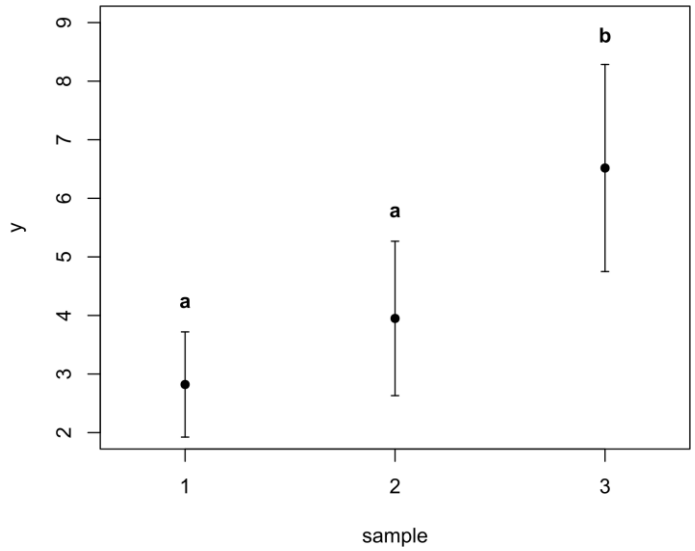

Fig. 8.3 Dotchart displaying means and 95%-confidence intervals for the means of the three samples. Means significantly different from each other at α = 0.05 are denoted by different letters (based on Tukey HSD test).

How to do in R

1. F test, F-distribution

function var.test; pf, qf for F-distribution probabilities

2. ANOVA

Function aov – accepts formula syntax. Note that the predictor must be a factor; otherwise linear regression is fitted (which is incorrect, but no warning is given).

summary (aov.object) displays the ANOVA table with SS, MS, F and p.

plot(aov.object) displays the diagnostic plots for checking ANOVA assumptions

3. Post-hoc test

tukeyHSD(aov.object)-produces just the differences between groups. Letters as in Fig. 8.3 must be produced manually.