The ultimate note on statistical tests

(Ultimate because it is most likely the last one before the exam, written 2007-11-17)

In the course, we have seen the following statistical tests:

- t-test (3 variants in Chapter 3 and t-tests on individual coefficients in linear models in chapters 5 and 6)
- F-test (not really the details) in anova tables for linear models for testing the whole model
- Shapiro test for normality (no details)

The fundamental approach in statistical tests is always the same:

We have a random sample from a (often infinite) population. The population is either explicit or hypothetical.

We have a precise hypothesis called the null hypothesis (H0) to be tested. Loosely speaking, "being tested" means that H0 is accused of being wrong. The hypothesis is on some property of the population from which the sample was drawn.

Often, H0 states that the effect of some independent variable on another dependent variable is zero (thus null-hypothesis). The independent variable can by either a grouping (gender of a living being, type of machine, ..., called qualitative variables, classifications, logical variables if there are only 2 groups) or a numeric variable (height of a being, weight or power of a machine, ..., called quantitative variable or sometimes a measure). Often, H0 can then be the conventional wisdom or a harmless situation (all individuals are equal), nothing needs to be done.

There always is an alternative hypothesis. Often the alternative hypothesis is simply that H0 is not correct, it is then not necessarily stated in an explicit manner.

We compute a test statistic (t in case of t-test, F for the F-test, W for the Shapiro test. Many other exist). Since we are working with a sample, the test statistic for the sample need not have exactly the same value as it would for the whole population.

The value of **the test statistic is interpreted through a p-value**, rather than directly.

As long as we have good statistical software, we need not always bother about the detailed method for computing the test statistic. Studying the computation might nevertheless help us understand when the test can be applied as well as explaining why we use that statistic.

A fundamental quality of the test statistic is that it is constructed in such a way that it has a known probability distribution if H0 is correct.

This probability distribution of the test statistic is used to compute the probability of getting a value that far or further from the population value (or the expected value) for the statistic. This is

the p-value.
Inference is made from the p-value in the same way for all tests: A small p-value means that the outcome (such a sample) was unlikely if H0 was correct. A very small p-value means that H0 is unbelievable, we reject it and adopt the alternative hypothesis. Normally, we define a threshold for when we have a "very small p-value". Often, it is taken to be 0.05 (5%), but other values may be used. The threshold defines the rejection region for the test (below 5% or other threshold value fixed beforehand).

If H0 is rejected, we may need a new value to believe for the parameter in question. Normally that is computed from the sample, together with a relevant confidence interval. For example, if we infer from the sample that the mean of a variable is not 0, we can adopt the sample mean as our best estimate of the true value.
A 5% rejection threshold for rejection goes hand in hand with 95% confidence interval for the parameter.

Statistical tests do not prove anything. They are like trials before the court: H0, claiming to be true, can be proven very unlikely, in which case the verdict is guilty, H0 is not correct. Even though H0 is not proven unlikely, we can not directly conclude it is correct and inncocent, it may simply be released due to lack of evidence. The burden of proof is on the claim that H0 is guilty. Really proving H0 is the task of the Laws of the disipline where the data came from.
The only rule of thumb we have for directly interpreting test statistics if for the t-value: If the absolute value of t is smaller than 1.5, we never reject H0 (p-value much higher than 0.05), if it is larger than 4 we always do (p-value very small), if it is in between we need to choose a threshold for rejection and look at the p-value.