Fisherin näkemyksistä testaamisesta

Bernard W. Lindgren kirjoittaa (Statistical Theory, 1976, 3. laitos, s:t 289–290):

The discussion of hypothesis testing so far has been in tradition of J. Neyman and E.S. Pearson, who developed these ideas in the early 1930s. Prior to that time, hypotheses were subjected to "significance testing", a process vigorously championed by R. A. Fisher, who, though admitting that decision making is proper in industry, argued that it has no place in scientific inquiry. Belittling the Neyman–Pearson theory, he continued to be a "significance tester" along with many followers, who, even now, will reject but not accept an hypothesis.

In significance testing the investigator begins with a null "hypothesis" — an hypothesis that he would like to disprove on the basis of the data – – . – – The distribution of [the test statistic] T is calculated, and an observed value of **T** is judged according to this null distribution as being reasonable (not unexpected), or so far out in the tail of the distribution (so "improbable") as to present a strong argument for the falcity of H_0. Such an extreme value is said to be "significant" and the null hypothesis is rejected when it it observed. – –

If T turns out to have a value that is not "significant", then no conclusion is drawn. Says R.A. Fisher [10, p.45])[1]: "A test of hypothesis contains no criterion for "accepting" a hypothesis. According to circumstances it may or may not influence its acceptability." Perhaps one reason for this is that there may be many other hypotheses or models that could equally well account for a value of T that is not significant. Indeed, when one asserts something as mathematically precise as $\mu = 17$ – – he is almost sure to be wrong – – .

A significance tester – – may make a mistake in his inference, owing to quirks in the data — but only one type of mistake, that of rejecting H_0 when it is true. The α, computed as the probablity of rejecting H_0 when it is true, is not to be taken as the probability of an erronous decision: "The calculation is based solely on a hypothesis, which, in the light of the evidence, is often not believed to be true at all, so that the actual probability of errojnous decision, supposing a phrase to have any meaning, may be, for this reason only, much less than the frequence specifying the level of significance." (Fisher, [10], p. 45.) And again: ([10], p. 47): "In general, tests of significance are based on hypothetical probabilities calculated from their null hypotheses. They do not generally lead to any probability statements about the real world, but a rational and well-defined measure of reluctance to the acceptance of the hypotheses they test."

In any case, a significance level is arbitrary, a point well put by Fisher ([10], p. 45) – – . – – Fisher seems to imply – – that improbability alone is evidence of incorrectness of the model in which it is calculated.

A. W. Edwards kirjoittaa (Likelihood, 1992, laajennettu laitos, s. 177):

In what may be called the Karl Pearson — "Student" — Fisher approach, by contrast, hypotheses are tested one at a time without specified alternatives. A null hypothesis is set up and "tested" against data: "It is merely something set up like a coconut to stand until it is hit" (Jeffreys).

[1] Fisher, R. A. (1973): Statistical Methods and Scientific Inference, 3. laitos. Hafner Press. New York.