university of groningen

DLG

Lausanne 2009

NWO Netherlands Organisation for Scientific Research

The persistent experimenter

Katie Steele

London School of Economics and Political Science

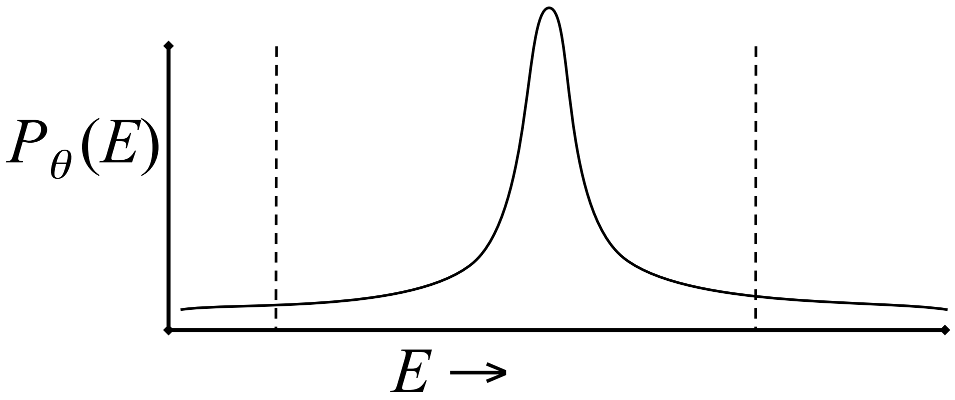
as told by

Jan-Willem Romeijn

University of Groningen

Stopping rules

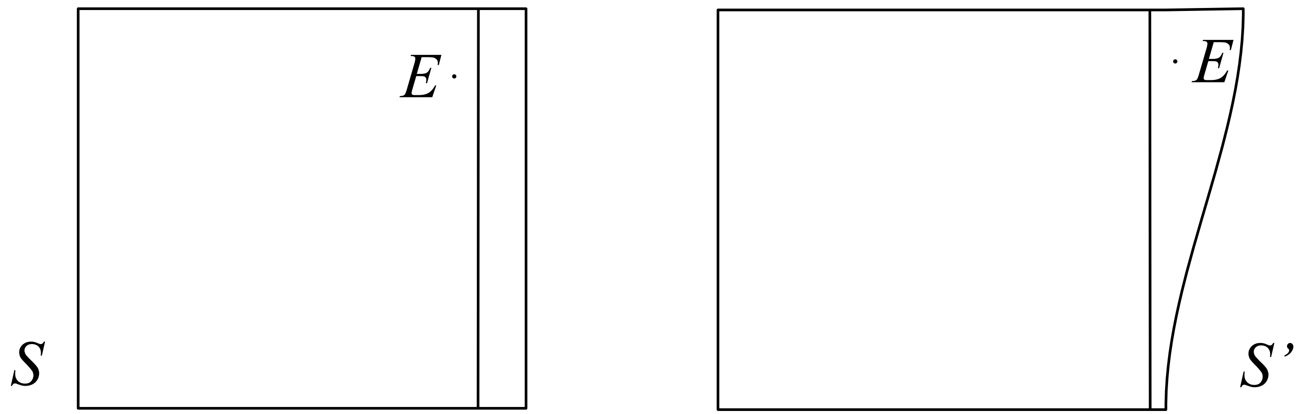
Classical statistical testing is based on a function that selects a region in sample space for which the data are improbable.



We may reject the hypothesis θ if the data that we obtain falls within this region.

Stopping rules

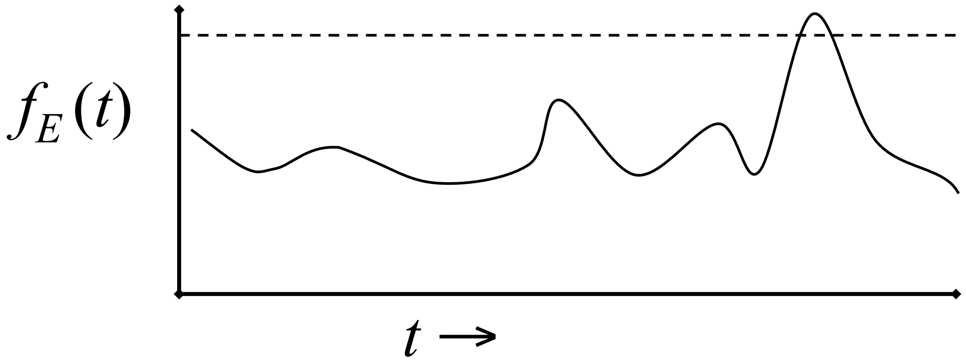
The determination of this region depends on the sample space in its entirety.



Hence the region of data for which some hypothesis can be rejected depends on the sampling plan.

Persistent experimenting

We might think that the sampling plan should not matter to the result of the statistical analysis.



But it turns out that if we allow an experimenter to gather data indefinitely, she can almost always reject the hypothesis with arbitrarily low error.

Persistent experimenting

If in fact the null hypothesis is true, the probability that this freak rejection occurs within a given finite number of trials decreases with the total number of trials.

$$\lim_{T \to \infty} \int_0^T P(\exists t : f_E(t) > r)\, dt = 1$$

$$\lim_{t \to \infty} P(f_E(t) > r) = 0$$

Nevertheless the measure of the set of sequences of trials in which the rejection is never warranted is zero: a bullet the Bayesian will have to bite.

Bringing in decisions

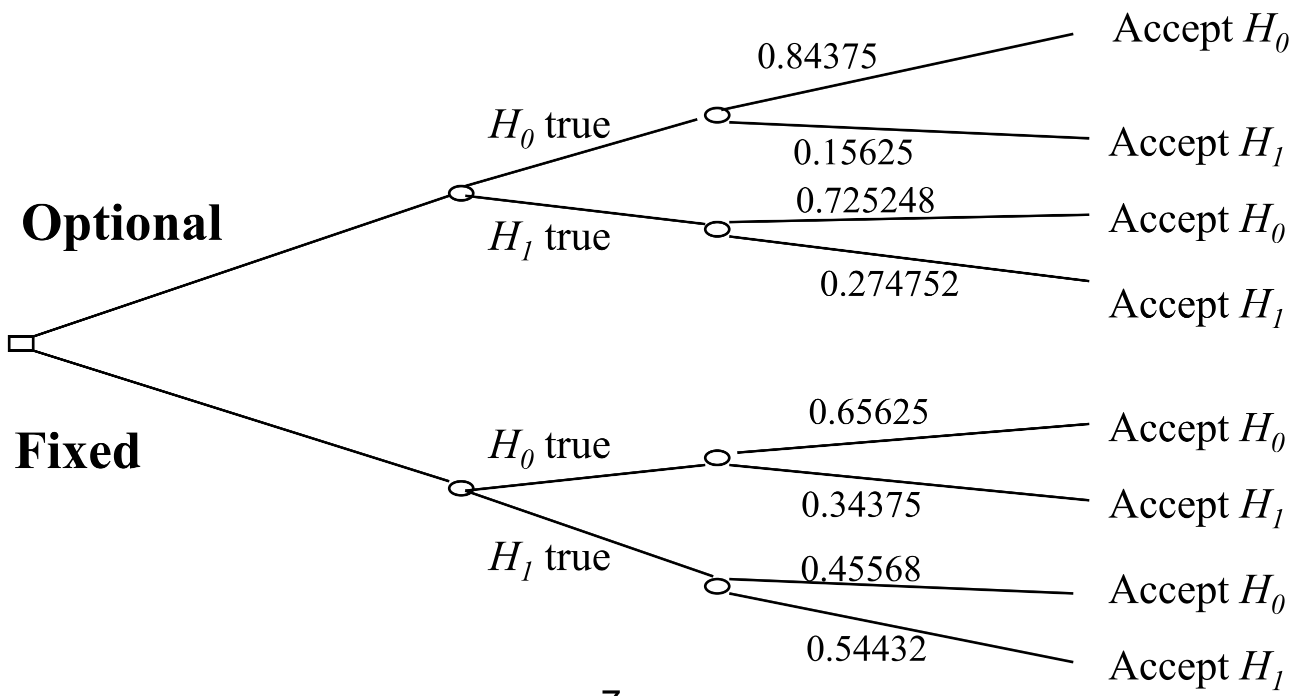
Steele puts the conflict on optional stopping in a decision-theoretic perspective.

	H_0 **true**	H_0 **false**
accept H_0	*U1*	*U2*
reject H_0	*U3*	*U4*

She first notes that, given the data, there will always be some representation of a classical statistician in terms of priors and utilities, justifying the decision.

Bringing in decisions

Steele employs an example with 6 trials to illustrate how to decide between different test procedures, either with or without optional stopping.



Bringing in decisions

Under certain conditions, optional stopping has lower utility than fixing the number of observations beforehand.

Utility truth	Utility false	EU optional stopping	EU fixed sample size
2	1	1.5595	1.600285
4	1	2.6785	2.800855
10	1	6.0355	6.402565

But we cannot apply this idea across data and maintain the invariance of priors and utility over these data sets. Each procedure-data-combination has its own Bayesian counterpart.

Some problems

Steele notes some potential weaknesses in this approach to the issue, and I have a few more.

- We might want to factor into the model that obtaining data is costly. This makes optional stopping more favourable.

- If we are less easy on rejecting hypotheses, we can diminish the false rejection rate considerably. It is unclear whether this saves optional stopping.

- Classical statistics already is oriented on decision, so a Bayesian decision-theoretic analysis of it may not be convincing to all sides of the debate.

Mayo vs. Berger and Wolpert

The expected utility of optional stopping depends also on the priors of the hypotheses under consideration.

> *Mayo (2001, p. 397–8) reports that Bayesians say confusing things about the optional stopping test, things that imply that they actually agree with the classical approach in these cases (that stopping rules/error probabilities matter). Apparently, Berger and Wolpert (1988) claim that in order to avoid the foregone conclusion described at the top of this section,* ***the Bayesian might assign some positive probability to H_0: $\mu = 0$, 'perhaps to reflect a suspicion that the agent is using stopping rule T-2 because he thinks the null hypothesis is true'****.*

We might try to use this fact in an explanation of why the stopping rule matters to a Bayesian after all.

Mayo vs. Berger and Wolpert

This requires setting up the problem with an experimenter informing the decision maker, who might take the potential use of optional stopping as a signal of the opinion of an epistemic peer.

- If the experimenter only cares about going home to swim, the optional stopping is not an informative signal at all.

- The effect of finding out that the experimenter was optionally stopping is working in the wrong direction.

In any case, Steele is succesful in making the stopping rule matter to the Bayesian, countering the negative reading of Berger and Wolpert by Mayo.

Pre- and Post-analysis

Steele remarks that when deciding between test procedures, we must clearly separate the choices before and after we have obtained data.

$$E[U] = \int_{\Theta} \int_{\substack{\text{Sample}\\ \text{space}}} P_\theta(E)\, U(\text{decision over } \theta \text{ by } E)\, dE$$

$$f_t(E) \quad \sim \quad \text{Sample space}$$

There is an interesting parallel between a Bayesian pre-analysis and a classical post-analysis: both depend on what one might observe.

Further arguments?

Perhaps we can give further reasons for choosing the Bayesian or the classical outlook on trials.

- The likelihood principle may be independently motivated or rejected.

- Decisions are part and parcel of both approaches, but there are differences in how these decisions are separated from inference.

- We might want to allow for optional stopping on pragmatic grounds; both approaches would deal with that differently.

Thanks!

For questions email:

k.steele@lse.ac.uk

But I am curious to hear and happy to forward your queries:

j.w.romeijn@rug.nl