Foundations of the Formal Sciences VI, Amsterdam

5.5. 2006

Surprise and Evidence in Statistical Model Checking

Jan Sprenger, University of Bonn

Department of Philosophy

I. P-values

What do they measure?

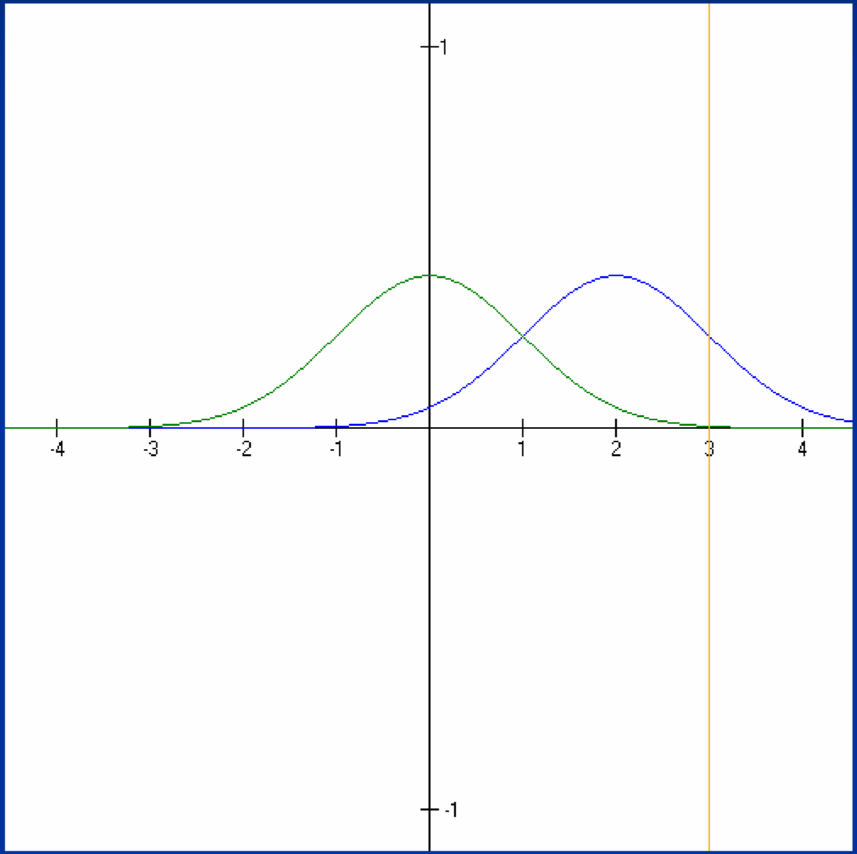
P-values

- Take a null (default) model H_0 and an alternative model H_1 and test them against each other
- Choose a statistic (function of the data) T that measure the distance to the null model H_0
- Then the P-value is $P(T(X) > T(x_0) \,|\, H_0)$

  x_0 = observed value
- Typical choices:

  T = probability density, identity, …

An example



- For the observed value x_0=3 and T=id, the P-value of the null model [green line] is $P(X > x_0 \mid H_0) < 0.01$
- P-values near 0 indicate significant deviation from the null!

P-values in applied statistics

- Low P-values (e.g. < 0.05) = „statistical significance“
- P-Values indicate whether an observed result is „significant evidence against the null model“

But that is relative to the choice of an alternative model!

P-values in applied statistics

- in standard applications, a null model H_0 is compared to an alternative model H_1

Siginificant result
=> the best of all models?

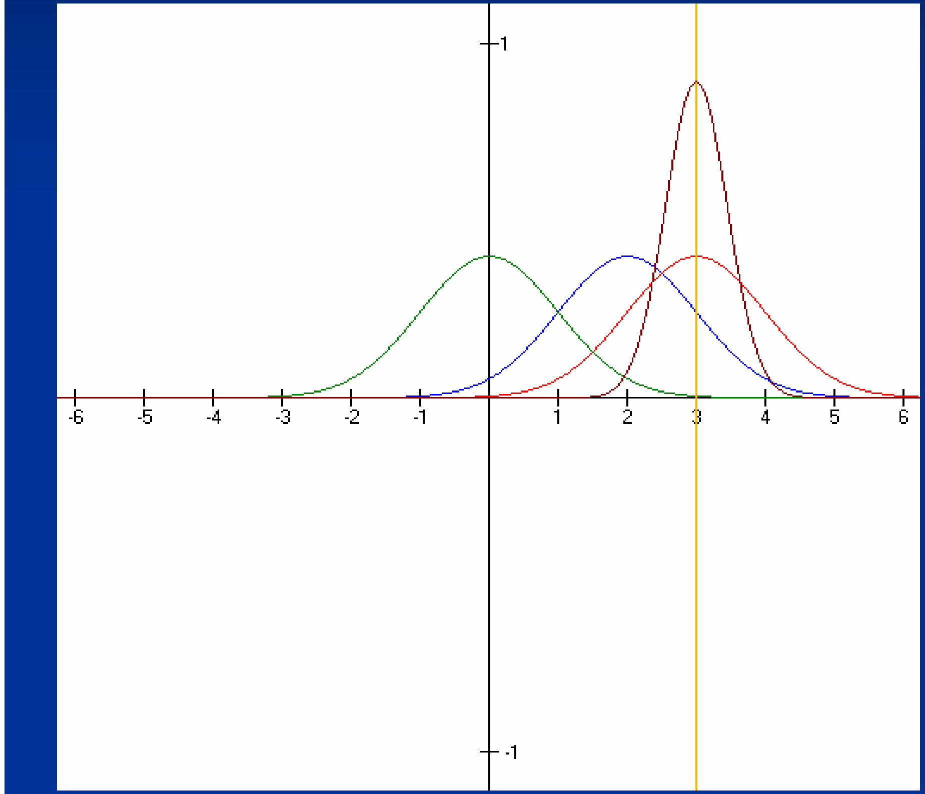
=> „Evidence for a model (simpliciter)“ is an improper way of speaking!

P-Values in Applied Statistics

„Evidence for a model“ is relative to an (implicit) alternative!

- a model will never achieve perfect fit with the data
- thus, evidence for a model is taken to mean that other models fit/predict the data worse

Significance in practice



- We test N(0,1) [green] against N(2,1) [blue]
- The actual result x=3 [yellow] is „significant evidence“ against N(0,1) and for N(2,1)
- However, it is still better evidence for N(3,1) [red] (or N(3,0.2)) [violet]!

P-values = measures of evidence?

What do we expect from a relevant and fruitful concept of evidence?

Comparative character

No dependence on counterfactual outcomes

Continuity in the likelihoods

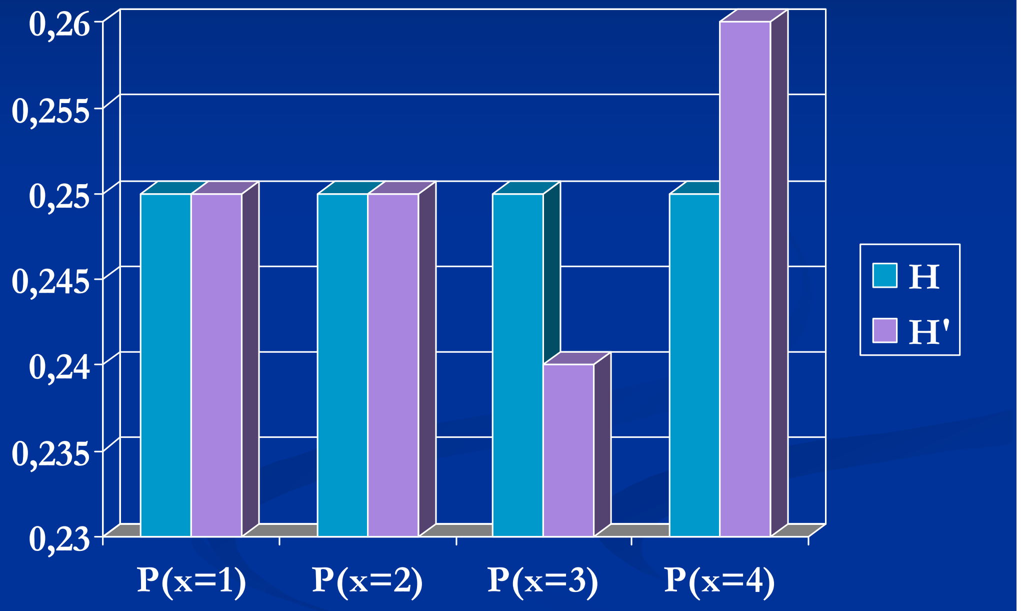
P-values = measures of evidence?

Unfortunately, P-values do not fulfil those conditions!

- They depend on the likelihood of outcomes other than the observed outcome x_0
- They are not comparative measures, but depend on only one distribution
- Finally, they are not continuous functions of the probability density

Discontinuity of P-values

- To recap: the p-value summarizes the probabilities of results that are less likely than the actual result
- If we observe x=1, the P-values for H and H´ should not be too different
- But in fact, P_H=0 and $P_{H‘}$=0,25!



P-values as measures of evidence

Conclusions:

- Evidence is an essentially comparative concept
- P-values are inadequate measures of evidence

But what are they good for?

II. Measures of surprise

A new rationale for P-values?

The point of surprise measures

- guiding the development of models at preliminary stags of model analysis.

- valuable when models are only tentatively proposed and accepted

- surprise measures are supposed to indicate the need for modification of the model

The point of surprise measures (II)

- Measures of surprise describe the *relative expectedness* of the actual result (relative to other possible results)
- a measure of surprise has to depend on the *probability of counterfactual outcomes*.
- They are *non-comparative*

=> measures of surprise are fundamentally different from measures of evidence.

Are P-values good measures of surprise?

- They depend on counterfactual outcomes, are non-comparative…
- But the discontinuity in the probability density is still a major problem!
- However, there are suitable modifications of P-values that are reasonable measures of surprise (cf. Howard 2007)

Surprise and Evidence

Surprise and Evidence play different epistemological roles! (exploratory model analysis versus model selection)

P-values have often been regarded as measures of evidence, however, if they have any value at all, then as measures of surprise!

Lessons for statisticians

- If P-values are taken as measures of evidence, then because the „distance statistic“ is a monotonous function of a measure of evidence (e.g the likelihood ratio)!
- Statisticians should be more aware of the *surprise-measuring role of P-values*, especially in two-tailed testing problems!

Lessons for statisticians (II)

- P-values should not be used for „significance testing“
- it is important to clearly separate the epistemic roles of exploratory model analysis and model selection!

Thanks a lot
for your attention!!!

© by Jan Sprenger, Cologne,
May 2007