Is the p-value a good measure of evidence? An asymptotic consistency criterion

marian.grendar@savba.sk[1]

Seminar at Comenius University, Bratislava

Nov 14, 2011

[1]KM FPV UMB BB

Intro

- Frequentist decision making: Neyman Pearson Wald
- Bayesian belief revision
- Evidence: what is the extent of support in data for a hypothesis H_1 relative to H_2

Setup, measure of evidence

r.v. $X \in \mathrm{R}^K$ with pmf/pdf $f_X(x \mid \theta)$, where $\theta \in \Theta \subseteq \mathrm{R}^L$.
partition of Θ: Θ_1, Θ_2
associate Θ_j with the hypothesis H_j, $j = 1, 2$.
$X_1^n \triangleq X_1, \ldots, X_n \sim f_X(x \mid \theta)$ random sample from $f_X(x \mid \theta)$.

Measure of evidence $\epsilon(H_1, H_2, X_1^n)$,

- in data X_1^n,
- for the hypothesis $H_1 : X_1^n \sim f_X(x \mid \theta)$ where $\theta \in \Theta_1$,
- relative to $H_2 : X_1^n \sim f_X(x \mid \theta)$ where $\theta \in \Theta_2$, i

is a mapping $\epsilon(H_1, H_2, X_1^n) : \Theta_1 \times \Theta_2 \times (\mathrm{R}^K)^n \to \mathrm{R}$.

Measure of evidence *against* H_1, relative to H_2 denoted $\epsilon(\neg H_1, H_2, X_1^n)$.
Calibration. The partition of R that corr. to the extreme values of evidence that corr. to the strongest evidence is denoted S.

Measures of evidence

- Fisherian: p-value
- Likelihood-wallahs': ratio of likelihoods, extended ratio of likelihoods, ...
- Bayesian: Bayes factor, posterior odds, ratio of posterior modes, ...

Measures of evidence: Fisherian

The p-value $\pi(\neg H_1, \cdot, X_1^n) = \inf\{\alpha : T(X_1^n) \in R_\alpha\}$, where $T(\cdot)$ is a test statistic, α is the size of the test with the rejection region R_α for H_1.

Measures evidence in a data X_1^n, against a hypothesis H_1.
The smaller the p-value, the stronger the evidence against H_1 in the data.
The p-value smaller than 0.01 suggests a very strong evidence against H_1; i.e., $S = (0, 0.01)$

Measures of evidence: likelihood-wallahs'

The ratio of likelihoods $r_{12}(H_1, H_2, X_1^n) = f(X_1^n \mid H_1)/f(X_1^n \mid H_2)$.

Measures evidence in a data for a simple hypothesis H_1, rel. to a simple hypothesis H_2.
$r_{12} >$ app.30 suggests a very strong evidence for H_1 rel. to H_2; i.e., $S = [30, \infty)$.

For general Θ_1, Θ_2, the extended ratio of likelihoods is $r_{12}^e(H_1, H_2, X_1^n) \triangleq \sup_{\Theta_1} f(X_1^n \mid \theta)/\sup_{\Theta_2} f(X_1^n \mid \theta)$.

Measures of evidence: Bayesian

Bayesian?

The Bayes Factor.
$b_{12} = \int_{H_1} f(X_1^n \mid \theta) q(\theta)\, d\theta / \int_{H_2} f(X_1^n \mid \theta) q(\theta)\, d\theta$, where $q(\cdot)$ is the prior.
$b_{12} > 150$, strong evidence for H_1 rel. to H_2.

The posterior odds. $p_{12} = b_{12}\, q(\Theta_1)/q(\Theta_2)$.

Ratio of posterior modes.

Criteria for a measure of evidence

Coherence. Gabriel, '69; $\Rightarrow$ Schervish, '96; Lavine & Schervish, '99; Bickel, '08.

If $H : \theta \in \Theta$ implies $H' : \theta \in \Theta'$ (i.e., $\Theta \subset \Theta'$), then the evidence for H' should be at least as large as for H.

Incoherent measures of evidence: p-value (Schervish, '96), Bayes factor (Lavine & Schervish, '99), ...

Coherent measures of evidence: ratio of likelihoods, extended ratio of likelihoods, posterior odds, ...

Motivation for consistency requirement

Sellke, Bayarri, Berger, '01

Ex.: yield (per hectare) of corn of sort D_l, $l = 1, 2, \ldots$.
Θ_1 corr. to 'mean yield is uninteresting',
Θ_2 corr. to 'mean yield is interesting'.
Experiment with corn D_l gives a random sample X_1^n.

Some sorts of corn give interesting mean yield, some give the uninteresting one.

I.e., some experimental data X_1^n come from H_1, other data sets are from H_2.

Consistency criterion

Data-sampling scheme:

1. First, θ is drawn from a pdf (or pmf) $p(\theta)$.
2. Given θ, a random sample X_1^n is drawn from $f_X(x \mid \theta)$.

We say that a measure of evidence $\epsilon(\neg H_1, H_2, X_1^n)$ against H_1, relative to H_2, is *consistent*, if

$$\lim_{n \to \infty} Pr(H_1 \mid \epsilon(\neg H_1, H_2, X_1^n) \in S) = 0.$$

The probability that θ is in Θ_1, given that the measure of evidence $\epsilon(\neg H_1, H_2, X_1^n)$ strongly testifies against H_1, relative to H_2, should go to zero, as the sample size n goes beyond any limit.

Is the p-value consistent?

It is assumed that X is a continuous random variable and the test statistic T is such that it rejects H_1 when observed value t of T is large. Then the p-value is

$$\pi(\neg H_1, \cdot, X_1^n) = \sup_{\Theta_1} Pr(T > t \,|\, \theta).$$

Also, $S = (0, \alpha_S)$, typically $\alpha_S = 0.01$.

Ex.: let X be a gaussian r.v., $\sigma^2 = 1$,
let $\Theta_1 = \{\theta_1\}$, $\Theta_2 = \{\theta_1 + \delta\}$, $\delta > 0$.
Let $w = p(\Theta_1)$, $w \in (0, 1)$.
And, let $T(X_1^n) = \sqrt{n}(\overline{x} - \theta_1)$ be the test statistic, and
$R_\alpha = \{X_1^n : T(X_1^n) > z_{1-\alpha}\}$ be the rejection region.

Is the p-value consistent? (cont'd)

Under H_1, the p-value is a uniform random variable, so $Pr(\pi(\neg H_1, \cdot, X_1^n) \in S \,|\, \Theta_1) = \alpha_s$.

Under H_2, the power of the test is $Pr(\pi(\neg H_1, \cdot, X_1^n) \in S \,|\, \Theta_2) = 1 - \Phi(z_{1-\alpha_s} - \sqrt{n}\delta)$.
It converges to 1, as $n \to \infty$.

Thus,

$$\lim_{n\to\infty} Pr(H_1 \,|\, \epsilon(\neg H_1, \cdot, X_1^n) \in S) = \frac{\alpha_S w}{1 - w(1 - \alpha_S)}.$$

The p-value is not consistent, in this example.

Is the p-value consistent? (cont'd)

Proposition 1
Let $p(\theta)$ be such that $w \triangleq \int_{\Theta_1} p(\theta)$ is $w \in (0,1)$. And, let T, R_α, be such that $Pr(\pi(\neg H_1, \cdot, X_1^n) \in S \,|\, \Theta_2) \to 1$, as $n \to \infty$ (i.e., for $\theta \in \Theta_2$, the power of the test T converges to 1). Then,

$$\lim_{n\to\infty} Pr(H_1 \,|\, \pi(\neg H_1, \cdot, X_1^n) \in S) = \frac{\alpha_S w}{1 - w(1-\alpha_S)}.$$

Is the p-value consistent? (cont'd)

Let $\alpha_S = 0.01$.

w	$\alpha_S\, w/(1 - w(1 - \alpha_S))$
0.5	0.0099
0.9	0.0826
0.999	0.9090

The greater the relative presence of data sets from H_1, the higher the asymptotic probability that the data come from H_1, when the p-value strongly testifies against H_1.

Is the ratio of likelihoods consistent?

Ex. (cont'd) $r_{21}(\neg H_1, H_2, X_1^n)$, as a measure of evidence against H_1, very strongly supports H_2 over H_1, if $r_{21} > k_S > 1$, so $S = [k_S, \infty)$.

Clearly,
$Pr(r_{21}(\neg H_1, H_2, X_1^n) \in S \,|\, \Theta_1) = 1 - \Phi(\log k_S/\delta\sqrt{n} + \sqrt{n}\delta/2)$,
which, under the assumption $\delta > 0$, converges to 0, as $n \to \infty$.

And,
$Pr(r_{21}(\neg H_1, H_2, X_1^n) \in S \,|\, \Theta_2) = 1 - \Phi(\log k_S/\delta\sqrt{n} - \sqrt{n}\delta/2)$,
which, under the assumption $\delta > 0$, converges to 1, as $n \to \infty$.

Thus, $\lim_{n\to\infty} Pr(H_1 \,|\, r_{21}(\neg H_1, H_2, X_1^n) \in S) = 0$.

Is the ratio of likelihoods consistent? (cont'd)

Proposition 2
For point sets Θ_1, Θ_2, and $p(\theta)$ such that $\int_{\Theta_1} p(\theta) \in (0,1)$, the ratio of likelihoods r_{21} is a consistent measure of evidence, i.e.,

$$\lim_{n\to\infty} Pr(H_1 \mid r_{21}(\neg H_1, H_2, X_1^n) \in S) = 0.$$

Is the extended ratio of likelihoods consistent?

Proposition 3
Let $f_X(x \mid \theta)$ and Θ_1, Θ_2 be such that the maximum likelihood estimators $\hat{\theta}_j(\Theta_j)$, restricted to Θ_j, are consistent estimators of θ, $j = 1, 2$. And let the maximum likelihood estimators $\hat{\theta}_j(\Theta_i)$, restricted to Θ_i, converge in probability to some finite $\bar{\theta}_j$, $i, j \in \{1, 2\}, i \neq j$. Let $p(\theta)$ be such that $\int_{\Theta_1} p(\theta) \in (0, 1)$. Then the extended ratio of likelihoods r^e_{21} is a consistent measure of evidence against H_1, relative to H_2.

Summary

- Consistency is a simple, self-evident criterion on a measure of evidence.
- The p-value is not consistent. Its inconsistency is a direct consequence of the fact that the p-value is uniformly distributed.
- Likelihood-based measures of evidence (ratio of likelihoods, extended ratio of likelihoods) and the Bayes factor, posterior odds, ratio of posterior modes, are consistent.
- Due to the inconsistency of the p-value, SBB's calibration of p-val to the consistent Bayes factor, becomes asymptotically meaningless.

The paper is available at arXiv.

References

- Bickel, D. R. (2008). The strength of statistical evidence for composite hypotheses with an application to multiple comparisons, COBRA preprint series, paper 49. To appear in *Statist. Sinica*.
- Lavine, M., and Schervish, M. J. (1999). Bayes factors: what they are and what they are not, *Amer. Statist.*, 53(2):119-122.
- Schervish, M. (1996). P values: what they are and what they are not, *Amer. Statist.*, 50:203-206.
- Sellke, T., Bayarri, M. J., and Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *Amer. Statist.*, 55(1):62-71.