Probability that P-value Provides Misleading Evidence Cannot be Controlled by Sample Size

Marian Grendár* and George Judge**

Abstract
A measure of statistical evidence should permit sample size determination so that the probability $M$ of obtaining (strong) misleading evidence can be held as low as desired. On this desideratum the p-value fails completely, as it leads either to an arbitrary sample size if $M \geq 0.01$ or no sample size at all, if $M < 0.01$.

Introduction
The p-value measures evidence that the data provides against a hypothesis. There are several arguments demonstrating that the p-value is not a good measure of evidence. The p-value is not coherent (cf. (Baird, 1983), (Baird, 1984), (Schervish, 1996), (Royall, 1997)), neither is it consistent (cf. (A. W. F. Edwards, 1992), (Royall, 1997), (Grendár, 2012)), and the $\alpha$-postulate (cf. (Cornfield, 1966)) upon which its use rests, is uncertain to hold true (cf. (Cornfield, 1966), (Royall, 1986)). Moreover, the p-value depends on a stopping rule; cf. (W. Edwards, Lindman, & Savage, 1963), (Royall, 1997).

A measure of evidence is consistent, if it is asymptotically impossible to obtain evidence that is strongly against $H$, if $H$ is true. Motivated by (Royall, 2000), we restate the desideratum in a form which could be more appealing to applied statisticians: A measure of statistical evidence should permit sample size determination so that the probability $M$ of obtaining misleading evidence can be held as low as desired.

---

*Bioinformatic Unit, Biomedical Center Martin, Jessenius Faculty of Medicine in Martin, Comenius University in Bratislava, Slovakia. Institute of Measurement Science, Slovak Academy of Sciences, Bratislava, Slovakia.

**Graduate School, University of California, Berkeley, the USA.
There are three main measures of statistical evidence in use: the p-value, the Bayes Factor (BF) and the Ratio of Likelihoods (RL). Both BF and RL satisfy the desideratum. The p-value fails it.

### Evidential sample size determination: Ratio of Likelihoods, Bayes Factor

The Ratio of Likelihoods (RL) measures evidence, or support, that data provide for a one point hypothesis relative to another one point hypothesis; cf. (Barnard, 1949; Hacking, 1965; A. W. F. Edwards, 1992; Royall, 1997). RL is coherent, consistent and an analogue of the α-postulate for RL does not contradict common sense. Recently, the Generalized Ratio of Likelihoods was proposed for the evidential ranking of interval hypotheses (cf. (Bickel, 2012), (Zhang, 2009)).

For a few basic models, the formulas for the evidential sample size determination in the RL framework has been developed* by (Strug, Rohde, & Corey, 2012); see also (Li, 2016).

Besides the data, bayesians consider also the prior information as the evidence and commonly measure support for a hypothesis relative to another hypothesis by the Bayes Factor (BF); cf. (Jeffreys, 1935; Kass & Raftery, 1995). BF, though consistent, is not a coherent measure of evidence; cf. (Lavine & Schervish, 1999).

For a simple models, a bayesian evidential sample size determination in the Bayes Factor framework was considered by (Katsis & Toman, 1999), (De Santis, 2004), among others.

### Evidential sample size determination: P-value

There seems to be no work addressing evidential sample size determination, in the p-value framework. Let us fill the gap.

A researcher plans an experiment in order to assess a hypothesis $H$. She intends to use the p-value to measure evidence the data provides against $H$. Before performing the experiment, the researcher wants to ensure that the number of observations will be sufficient to guarantee that the probability of obtaining strong misleading evidence is at most $k \in (0, 1)$. According to the commonly used calibration, evidence against $H$ is considered very strong if the p-value $p(X_1, \ldots, X_n)$ is smaller than 0.01 (cf. (Wasserman, 2013)).

*And contrasted with the Neyman-Pearsonian sample size determination for decision making.
Thus, the researcher wants to determine the size $n$ of the sample $X_1, \ldots, X_n$, so that the probability

$$M \triangleq \Pr(p(X_1, \ldots, X_n) \leq 0.01; H),$$

that the p-value provides strong misleading evidence against $H$ when $H$ is true, is at most $k$.

Since the p-value is uniformly distributed under $H$, the probability $M$ is 0.01, regardless of the sample size. Thus, if $k \geq 0.01$, the probability of obtaining misleading evidence is smaller than $k$ for any sample size. One observation can do it; or ten thousands, as you wish. If $k < 0.01$, then however large the sample size it would not make the probability $M$ smaller than $k$.

Consequently, the probability that the p-value provides misleading evidence cannot be controlled by sample size.

If the researcher desires to control the probability of obtaining weak evidence against $H$, i.e. $\Pr(p(X_1, \ldots, X_n) \in [0.05, 0.1]; H)$, she would end up in the same void.

**Conclusion**

The recent ASA statement (Wasserstein & Lazar, 2016) stresses that the p-value ‘does not provide a good measure of evidence regarding a model or hypothesis’. Among other flaws, the p-value does not permit setting sample size in such a way that the probability of obtaining misleading (or weak) evidence, can be as low as desired.

**Acknowledgments**

This work was supported by the project “Biomedical Center Martin” ITMS code: 26220220187, the project is co-financed from EU sources. Also supported by VEGA 2/0047/15 grant. Valuable feedback from Ján Strnádel is gratefully acknowledged.

**References**


