The Jeffreys-Lindley's paradox

Andrew Fowlie

November 7, 2016

Table of contents

Nutshell

Jeffreys-Lindley's paradox in a nutshell

Hypothesis testing is the most controversial aspect of inference.

Frequentist methods (Neyman, Fisher, etc) and Bayesian methods *don't* always agree.

A specific example of a disagreement was given by Lindley[1], though previously noted by Jeffreys[2].

Lindley described it as a paradox. It's been somewhat controversial since.

[1] D. V. Lindley, Biometrika 44, 187–192 (1957).
[2] H. Jeffreys, (Oxford University Press, 1939).

“Paradox”

Lindley’s paradox is in fact a difficulty reconciling two paradigms — Bayesian and frequentist statistics. There is no mathematical inconsistency.

Similar in that regard to paradoxes in physics from reconciling quantum/relativistic and classical physics — think of ladder, twin, EPR parodxes etc.

Like in physics, paradoxes are useful for understanding foundations of a subject. Again, think of EPR or Maxwell’s demon.

Bayes versus frequentism

Bayes — probability is a (unique) measure of degree of belief (see e.g., Cox's theorem in Chap. 2 of Jaynes[3])

Frequentist — probability is the (asymptotic) frequency at which an outcome occurs, in a hypothetical sequence of repeated trials.

Homework: is probability a property of a coin? The coin/thrower system? Measure of degree of belief about the outcome of a coin toss? (see Chap. 10.3 of Jaynes).

Homework: is probability a property of a QM system?

[3] E. T. Jaynes, (Cambridge University Press, 2003).

Bayes versus frequentism

Bayesian probabilities can describe any hypotheses or propositions.

Figure 1: Probability that Leicester would win the Premier league? $5000/1$ betting odds. Best inference sometimes completely wrong.

Frequentist probabilities describe only repeatable events. Homework: if events are repeated identically, why is there variation in outcome?

Hypothesis testing

Bayesian I

Calculate the plausibility of a theory directly

$$p(M|D) = \frac{p(D|M) \cdot p(M)}{\sum p(D|M)p(M)} \quad (1)$$

This requires more than one model to be specified. See e.g., Gregory[4] or Bretthorst[5] or any introductory textbook.

The factor $p(D|M)$ is called the *evidence*,

$$p(D|M) = \int p(D|M,x)p(x|M)dx \quad (2)$$

You can calculate evidences with e.g., `MultiNest`[6]. In fact, we usually considered a Bayes factor, which is ratio of evidences

$$B = \frac{p(D|M_1)}{p(D|M_2)} \tag{3}$$

The Bayes factor “updates” the relative prior belief in two models with data, resulting in a posterior belief,

$$\text{Posterior odds} = \text{Bayes factor} \times \text{Prior odds} \tag{4}$$

The factors $p(M)$ and $p(x|M)$ are called priors. They reflect prior knowledge/ignorance. Priors are the most controversial ingredient. They could be selected by e.g., invariance under a symmetry or maximum entropy[7].

What if we want to make a decision? Do we announce a discovery? Do we declare a new drug safe? Decision theory: loss/utility functions are required. Evidences alone tell us "truth", not best choices.

[4] P. Gregory, (Cambridge University Press, 2005).

[5] G. L. Bretthorst, in , edited by G. R. Heidbreder, (Springer Netherlands, Dordrecht, 1996), pp. 1–42.

[6] F. Feroz, et al., Mon. Not. Roy. Astron. Soc. 398, 1601–1614 (2009), arXiv:0809.3437 [astro-ph], F. Feroz, et al., (2013), arXiv:1306.2144 [astro-ph.IM].

[7] D. A. Lavis, and P. J. Milligan, The British Journal for the Philosophy of Science 36, 193–210 (1985).

Frequentist goodness-of-fit test test I

Test with a single hypothesis (Fisher, Pearson et al).

Based around a decision — accept or reject model (cf. Fisher advocated reporting p-values). Not based around epistemology — e.g., calculate relative plausibility of two models.

Consider the type-1 error — probability of rejecting the null hypothesis, given that it was true.

Pick a “null” hypothesis that you wish to test.

Pick a “sufficient” test-statistic that measures disagreement between data and predictions. The test-statistic is a random variable and it would be convenient if it had a known distribution. Common test-statistic e.g. χ^2.

Calculate a p-value (also a random variable) — the probability of obtaining a test-statistic so extreme, were the null hypothesis true. Homework: show that p-value is uniformly distributed.

$$p\text{-value} = p(\lambda \geq \lambda_{\text{observed}} | H_0) \quad (5)$$

Frequentist goodness-of-fit test test III

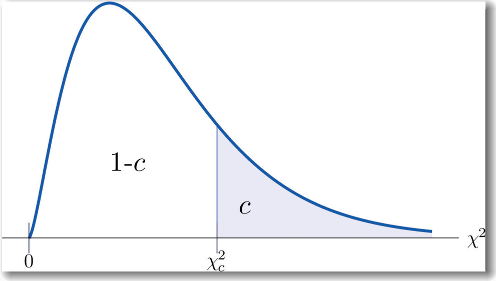

Figure 2: Tail probability.

Frequentist goodness-of-fit test test IV

Reject model if p-value less than a previously chosen threshold, e.g., 0.05. p-values are often converted into z-scores, i.e., expressed as the probability in tail of standard normal at z,

$$z = \Phi^{-1}(1 - p\text{-value}) \tag{6}$$

Homework: why did this catch on? Why report z-score rather than p-value?

This is a property of the experiment (and hypothetical pseudo-experiments): if we hypothetically repeated experiment many times, we'd reject the null hypothesis in 5% of cases, if it were true.

Frequentist hypothesis test I

Test with two hypotheses (Neyman et al).

This allows one to consider type-1 *and* type-2 errors. Type-2 error — probability of accepting null hypothesis, given that alternative was true.

Allows a notion of statistical power: for a fixed type-1 error, minimise the type-2 error (see Neyman-Pearson lemma[8] about likelihood ratios being best test-statistic).

Frequentist hypothesis test II

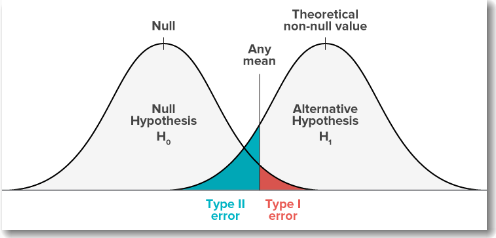

Figure 3: Type-1 and type-2 errors.

Homework: how did the goodness-of-fit test work without calculating type-2 error? What does it mean to reject a model with no alternative?

[8] J. Neyman, and E. S. Pearson, Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 231, 289–337 (1933), eprint: http://rsta.royalsocietypublishing.org/content/231/694-706/289.full.pdf.

Lindleys’ paradox

They disagree. Even with lots of data.

The frequentist and Bayesian methods needn't agree.

Folk theorem — they agree in the limit of lots of data. This is not true in model selection. In parameter inference, something like this is true (Bernstein–von Mises theorem).

Lindley provided a specific example.

The problem

Suppose we pick n samples from a normal distribution, $N(\mu, \sigma^2)$, with known variance σ^2. We want to select a model that best predicts the mean of distribution.

Frequentist p-value from goodness-of-fit

Null hypothesis, H_0: the mean $\mu = \mu_0$.

By the central limit theorem, the sample mean $\bar{x} = \sum x_i / n$, is normally distributed, $\bar{x} \sim N(\mu, \sigma^2/n)$. Let's pick a χ^2 test-statistic:

$$\chi^2 = \frac{(\bar{x} - \mu_0)^2}{\sigma^2/n} \tag{7}$$

We can calculate χ^2, and find the p-value,

$$p\text{-value} = p(\chi^2 \geq \chi^2_{\text{obs}} | H_0) \tag{8}$$

from the survival function of a χ^2-distribution.

The p-value depends on the χ^2 — for fixed χ^2, the number of samples n didn't matter.

Digression on Gaussian distribution

Why is Gaussian distribution ubiquitous?

CLT

Take n samples from a distribution of mean μ, variance σ^2. The sample mean $\bar{x} \sum x/n$ is distributed $\bar{x} \sim N(\mu, \sigma^2/n)$.

MaxEnt

If we only know the first two moments of a distribution, μ and σ^2, the distribution that maximises the Shannon entropy (i.e., uncertainty) is the Gaussian! i.e., Gaussian is most honest choice if that's all you know.[9].

[9] D. A. Lavis, and P. J. Milligan, The British Journal for the Philosophy of Science 36, 193–210 (1985).

Bayes factor I

Two models, introduced on an equal footing:

- M_1: $\mu = \mu_0$.
- M_2: μ lies inside an interval, length L, that includes μ_0 and $L \gg \sigma$. We pick a prior $p(\mu) = 1/L$

Let's calculate evidences. In M_1, it is trivial,

$$p(D|M_1) = \frac{1}{\sqrt{2\pi}\sigma/\sqrt{n}} e^{-\frac{(\bar{x}-\mu_0)^2}{2\sigma^2/n}} = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} e^{-\chi^2/2} \qquad (9)$$

Bayes factor II

In M_2, we must marginalise the μ parameter by integration,

$$p(D|M_2) = \int p(D|\mu, M_2)p(\mu|M_2)d\mu \tag{10}$$

$$= \frac{1}{L} \int \frac{1}{\sqrt{2\pi}\sigma/n} e^{-\frac{(\bar{x}-\mu)^2}{2\sigma^2/n}} d\mu \tag{11}$$

$$\approx \frac{1}{L} \tag{12}$$

Thus, we find a Bayes factor

$$B(M_1/M_2) = \frac{\sqrt{n}L}{\sqrt{2\pi}\sigma} e^{-\chi^2/2} \tag{13}$$

Bayes factor III

For fixed χ^2, as $n \to \infty$, the Bayes factor favours $\mu = \mu_0$ by a Bayes factor $B \to \infty$. This result is somewhat insensitive to choices of prior for μ in M_2.

Bartlett[10] observed the sensitivity of the Bayes factor to the width of the uniform prior L. Homework: Do we need reliable prior information about reliable interval of parameter to make inference? What does this dependence mean?

[10] 10.2307/2332888.

The "paradox"

The behaviours of the p-value and Bayes factors as functions of χ^2 and n mean that

Paradox

Taking $n \to \infty$, but fixing e.g., $\chi^2 = 25$, we would reject $\mu = \mu_0$ at 5σ. But the Bayes factor would favor $\mu = \mu_0$ by a factor $B \to \infty$.

Resolutions/discussion

Resolutions I

Lindley's paradox was invoked by advocates of Bayesian and frequentist statistics. The implications aren't agreed upon.

Trivial resolution: two methodologies answer different questions. That's no good. What if they lead to different decisions? e.g., should you announce a GW discovery?!

Should significance levels e.g., 5%, in fact be functions of sample size n, resulting in agreement between approaches?

Re-examine what $n \to \infty$ but t_n fixed means? Under alternative hypothesis, we expect $t_n \to \infty$ as $n \to \infty$.

Resolutions II

Is t_n, in this case a χ^2, well-defined under M_2 in the Bayesian analysis? There isn't a particular prediction μ_0 for the χ^2 formula.

Are point null priors inappropriate?

Frequentist p-values overstate evidence against the null hypothesis? i.e., one really cannot invert the p-value and think of probability that the null hypothesis is true.

Summary

Summary

Bayesian and frequentist methods for model selection don't always agree, even asymptotically in the limit of large statistics.

One particular disagreement noted by Lindley and described as a paradox.

Implications disputed. Both sides claimed victory.

Questions?

References

D. V. Lindley, "A statistical paradox," Biometrika 44, 187–192 (1957).

H. Jeffreys, *The Theory of Probability*, (Oxford University Press, 1939).

E. T. Jaynes, *Probability theory: the logic of science*, (Cambridge University Press, 2003).

P. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, (Cambridge University Press, 2005).

References II

G. L. Bretthorst, "An introduction to model selection using probability theory as logic," in *Maximum entropy and bayesian methods: santa barbara, california, u.s.a., 1993*, edited by G. R. Heidbreder, (Springer Netherlands, Dordrecht, 1996), pp. 1–42.

F. Feroz, M. P. Hobson, and M. Bridges, "MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics," Mon. Not. Roy. Astron. Soc. 398, 1601–1614 (2009), arXiv:0809.3437 [astro-ph].

F. Feroz, M. P. Hobson, E. Cameron, and A. N. Pettitt, "Importance Nested Sampling and the MultiNest Algorithm," (2013), arXiv:1306.2144 [astro-ph.IM].

References III

D. A. Lavis, and P. J. Milligan, "The work of e. t. jaynes on probability, statistics and statistical physics," The British Journal for the Philosophy of Science 36, 193–210 (1985).

J. Neyman, and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 231, 289–337 (1933), eprint: http://rsta.royalsocietypublishing.org/content/231/694-706/289.full.pdf.