# Stats4Astro Astrostatistics School

Roberto Trotta, Imperial College London

Autrans, October 2017

## 1 Exercises

(i) A coin is tossed $N = 250$ times and it returns $H = 140$ heads.

(a) Evaluate the evidence that the coin is biased using Bayesian model comparison. Assume a uniform prior between 0 and 1 for the probability of heads in a single toss.

(b) How many heads would you have to observed in $N = 250$ flips to obtain "strong" evidence for bias (on the Jeffreys' scale)?

(c) Contrast your findings with the usual (frequentist) hypothesis testing procedure (i.e. testing the null hypothesis that $p_H = 0.5$). Can you rule out the null hypothesis at the 95% confidence level?

(d) Discuss the dependency of the Bayesian result on the choice of prior for the probability of heads, by giving the "biased coin" model the maximum possible advantage under the prior.

(ii) In 1919 two expeditions sailed from Britain to measure the light deflection from stars behind the Sun's rim during the solar eclipse of May 29th. Einstein's General Relativity predicts a deflection angle

$$\alpha = \frac{4GM}{c^2 R},$$

where $G$ is Newton's constant, $c$ is the speed of light, $M$ is the mass of the gravitational lens and $R$ is the impact parameter. It is well known that this result it exaclty twice the value obtained using Newtonian gravity. For $M = M_\odot$ and $R = R_\odot$ one gets from Einstein's theory that $\alpha = 1.74$ arc seconds.

The team led by Eddington reported $1.61 \pm 0.40$ arc seconds (based on the position of 5 stars), while the team headed by Crommelin reported $1.98 \pm 0.16$ arc seconds (based on 7 stars).

What is the Bayes factor between Einstein and Newton gravity from those data? Comment on the strength of evidence.

(iii) Assume that the combined constraints from CMB, BAO and SNIa on the density parameter for the cosmological constant can be expressed as a Gaussian posterior distribution on $\Omega_\Lambda$ with mean 0.7 and standard deviation 0.05. Use the Savage-Dickey density ratio to estimate the Bayes factor between a model with $\Omega_\Lambda = 0$ (i.e., no cosmological constant) and the $\Lambda$CDM model, with a flat prior on $\Omega_\Lambda$ in the range $0 \leq \Omega_\Lambda \leq 2$. Comment on the strength of evidence in favour of $\Lambda$CDM.

(iv) If the cosmological constant is a manifestation of quantum fluctuations of the vacuum, QFT arguments lead to the result that the vacuum energy density $\rho_\Lambda$ scales as

$$\rho_\Lambda \sim \frac{c\hbar}{16\pi} k_{\rm max}^4 \tag{1}$$

where $k_{\rm max}$ is a cutoff scale for the maximum wavenumber contributing to the energy density (see e.g. [**?**]). Adopting the Planck mass as a plausible cutoff scale (i.e., $k_{\rm max} = c/\hbar M_{\rm Pl}$) leads to "the cosmological constant problem", i.e., the fact that the predicted energy density

$$\rho_\Lambda \sim 10^{76} {\rm GeV}^4 \tag{2}$$

is about 120 orders of magnitude larger than the observed value, $\rho_{\rm obs} \sim 10^{-48} {\rm GeV}^4$.

(a) Repeat the above estimation of the evidence in favour of a non-zero cosmological constant, adopting this time a flat prior in the range $0 \leq \Omega_\Lambda/\Omega_\Lambda^{\rm obs} < 10^{120}$. What is the meaning of this result? What is the required observational accuracy (as measured by the posterior standard deviation) required to override the Occam's razor penalty in this case?

(b) It seems that it would be very difficult to create structure in a universe with $\Omega_\Lambda \gg 100$, and so life (at least life like our own) would be unlikely to evolve. How can you translate this "anthropic" argument into a quantitative statement, and how would it affect our estimate of $\Omega_\Lambda$ and the model selection problem?

## 2 Solutions

(i) (a) In this model comparison problem, we are comparing model $\mathcal{M}_0$ that the coin is fair (i.e., $p_H = 0.5$) with a model $\mathcal{M}_1$ where the probability of heads is $\neq 0.5$. We begin by assigning under model 1 a flat prior to $p_H$ between 0 and 1.

The Bayes factor (or ratio of the two models' evidences) is given by

$$B = \frac{P(H=140|\mathcal{M}_1)}{P(H=140|\mathcal{M}_0)} = \frac{\frac{H!(N-H)!}{(H+1)!}}{(1/2)^N}\Big|_{N=250,H=140} = \frac{\frac{140!110!}{251!}}{(1/2)^{250}} \approx 0.48 \sim 2:1 \tag{3}$$

(notice that we have cancelled the "choose" terms in the numerator and denominators above). So there is not even weak evidence in favour of the model that the coin is biased.

(b) The log of the Bayes factor is plotted as a function of $H$ in Fig. 1. By inspection it is apparent that values $107 \leq H \leq 143$ favour the fair coin model ($\ln B < 0$). In order to obtain "strong evidence" in favour of the biased coin model ($\ln B > 5$), it is necessary that either $H < 94$ or $H > 31$.

(c) The Frequentist hypothesis testing procedure is to compute the tail probability of obtaining data as extreme or more extreme than have been observed under the null hypothesis, i.e., that the coin is fair. This gives the p-value:

$$\text{p-value} = \left(\frac{1}{2}\right)^N \sum_{H=H_{\text{obs}}}^{N} \binom{N}{H} \approx 0.033 \tag{4}$$

So for a Frequentist, the data would exclude the null hypothesis that the coin is fair at more than the 95% CL.

(d) How does the Bayesian result depend on the choice of prior for the alternative hypothesis? Above we have given to $p_H$ a flat prior between 0 and 1. If we wanted to give the maximum possible advantage to a model where the coin is not fair, we could put all of its prior probability in a delta-function concentrated at the value of $p_H$ that maximizes the probability of what has been observed. So under this maximally advantageous model for the unfairness hypothesis (let's call this $\mathcal{M}_2$), we would select a "prior" (in quotation marks, for this prior is actually selected after the data have been gathered, so we are effectively using the data twice here!) of the form $P(p_H) = \delta(p_H - H/N)$. In this case the odds in favour of this new model are

$$B = \frac{P(H=140|\mathcal{M}_2)}{P(H=140|\mathcal{M}_0)} = \frac{(H/N)^H(1-H/N)^{N-H}}{(1/2)^{250}}\Big|_{H=140,H=250} \approx 6.1. \tag{5}$$

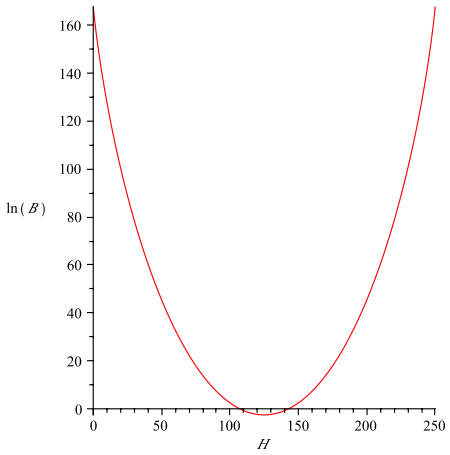


Figure 1: Natural log of the Bayes factor between the model "the coin is biased" (with flat prior) and the model "the coin is fair", as a function of the number of heads ($H$) in 250 flips, see Eq. (3). Values $\ln B > 0$ favour the biased coin model. The Jeffreys' threshold for "strong evidence" is at $\ln B = 0$

Even in this most favourable setup for the hypothesis that the coin is biased, we find only weak evidence (odds of 6 to 1) against the model of a fair coin. Therefore we can safely conclude that the data do not warrant to conclude that the coin is unfair.

(ii) We are comparing here two models which both make exact predictions for the deflection angle, with no free parameters. If you prefer, you might consider the prior on $\alpha$ under each theory to be a delta-function centered at the predicted value. This of course neglects the uncertainty associated with $M_\odot$ and $R_\odot$.

In this case, the evidence is thus simply the likelihood function for the observed data under each theory (you can convince yourself that this is correct by explicitly computing the evidence for each model assuming the delta-function prior above). This gives for the Bayes factor in favour of Einstein gravity vs Newton (assuming Gaussian likelihoods)

$$B = \frac{\mathcal{L}_0 \exp\left(-\frac{1}{2}\frac{(\hat{\alpha}-\alpha_E)^2}{\sigma^2}\right)}{\mathcal{L}_0 \exp\left(-\frac{1}{2}\frac{(\hat{\alpha}-\alpha_N)^2}{\sigma^2}\right)} \tag{6}$$

where $\alpha_E = 1.74''$, $\alpha_N = 0.87''$, $\hat{\alpha}$ is maximum likelihood value of the experiment and $\sigma$ is the standard deviation.

Using the supplied data from Eddington, one obtains $B \sim 5$, so "weak evidence" in favour of Einstein theory according to the Jeffreys' scale for the strength of evidence. The Crommelin data instead give $B \sim 10^{10}$,

so very strong evidence for Einstein. Notice that this comes about because the measurement from Crommelin is on the high side (i.e., higher than Einstein prediction, even), and therefore the assumed Gaussian tail becomes tiny for $\alpha = \alpha_N$. It is worth noticing that, although the above calculation is formally correct, it is likely to overestimate the evidence against Newton, because the Gaussian approximation made here is certain to break down that far into the tails (i.e, $\alpha_N$ is $\sim 11\sigma$ away from the value measured by Crommelin. No distribution is exactly valid that far into the tails!).

(iii) Here we are comparing two nested model, $\mathcal{M}_0$ with $\Omega_\Lambda = 0$ and a more complicated model, $\mathcal{M}_1$,where $\Omega_\Lambda \leq 0$ and a flat prior $P(\Omega_\Lambda|\mathcal{M}_1) = 1/2$ for $0 \leq \Omega_\Lambda \leq 2$ and 0 elsewhere (notice that the prior needs to be normalized, hence the factor 1/2). We can therefore use the Savage-Dickey density ratio to compute the Bayes factor between $\mathcal{M}_0$ and $\mathcal{M}_1$:

$$B_{01} = \frac{P(\Omega_\Lambda = 0|\text{CMB+BAO+SN}, \mathcal{M}_1)}{P(\Omega_\Lambda = 0|\mathcal{M}_1)} = \frac{\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(0-\hat{\Omega}_\Lambda)^2}{\sigma^2}\right)}{1/2}, \quad (7)$$

where we have assumed thart the posterior under $\mathcal{M}_1$ can be approximated as a Gaussian of mean $\hat{\Omega}_\Lambda = 0.7$ and standard deviation $\sigma = 0.05$. Numerical evaluation gives $B_{01} \sim 10^{-42}$, so with this prior the model that $\Omega_\Lambda = 0$ can be ruled out with very strong evidence. Another way of looking at this result is the following: if, after having seen the data, you remain unconvinced that indeed $\Omega_\Lambda > 0$, this means that the ratio in your relative degree of prior belief in the two models should exceed $P(\mathcal{M}_0)/P(\mathcal{M}_1) > 10^{42}$.

(iv) (a) The calculation of the Bayes factor proceeds as above, but this time with a much larger prior range for the alternative model, $\Omega_\Lambda > 0$. This means that the prior height, $P(\Omega_\Lambda = 0|\mathcal{M}_1)$ , appearing in the denominator of Eq. (7) is very small, i.e. $P(\Omega_\Lambda = 0|\mathcal{M}_1) = 10^{-120}$, as the prior needs to be normalized. Repeating the above calculation, we get for the Bayes factor in favour of $\mathcal{M}_0$ (i.e., that $\Omega_\Lambda = 0$)

$$B_{01} \sim \frac{10^{-42}}{10^{-120}} \sim 10^{88}. \quad (8)$$

Now the Bayes factor is positive (and huge), a reflection of the enormous amount of prior range wasted by $\mathcal{M}_1$. Therefore under this new prior, the Bayesian model comparison favour the hypothesis that there is no cosmological constant despite the fact that the likelihood peaks about $0.7/0.05 \sim 14\sigma$ away from $\Omega_\Lambda = 0$. This is an extreme example of Occam's penalty.

In order for the Occam's factor to be overruled by the likelihood, we require that $B_{01} = 1$ (i.e., equal odds for the two models). This translates in the approximate condition for the number of sigma detection,

$\lambda$:

$$\exp\left(-\frac{1}{2}\lambda^2\right) \sim 10^{-120}, \tag{9}$$

where we have dropped the term $1/\sigma$ in front of the likelihood for simplicity (as the likelihood is going to be dominated by the exponential anyhow). Solving for $\lambda$ gives

$$\lambda \sim \sqrt{240 \ln 10} \approx 23. \tag{10}$$

So we would need a $\sim 23\sigma$ detection of $\Omega_\Lambda > 0$ to override completely the Occam's razor penalty.

(b) The outcome of the model comparison changes dramatically if one is willing to impose a much more stringent upper cutoff to the prior range of $\Omega_\Lambda$, based e.g. on anthropic arguments. The observations that structures cannot form if $\Omega_\Lambda \gg 100$ (and therefore there would be no observers to measure dark energy, see e.g. the original argument by Weinberg [**?**]) can be approximately translated in a prior range extending perhaps to $\Omega_\Lambda \sim 10^3$. With this choice of range, the Bayes factor becomes

$$B_{01} \sim \frac{10^{-42}}{10^{-3}} \sim 10^{-39}, \tag{11}$$

thus swinging back to support $\mathcal{M}_1$ with enormous odds. This illustrate that Bayesian model comparison can be difficult (and strongly dependent on the theoretical prior range adopted) in cases where there is no compelling (or unique) argument to define the prior.