# Making sense of data: introduction to statistics for gravitational wave astronomy

## Problem Sheet 2: Bayesian Statistics

---

Questions marked with a * are a selection that will give experience of all aspects of the course. For IMPRS students taking this course, these should be completed and handed in to be marked.

1. ***** A motorist travels regularly from Berlin to Golm. On each occasion he chooses a route at random from four possible routes. From experience, the probabilities of completing the journey in under 1 hour via these routes labelled 1 to 4 are 0.2, 0.5, 0.8 and 0.9 respectively. Given that on a certain occasion they complete the journey in under 1.5 hours, calculate the probability that they travelled on each of the possible routes.

2. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables such that $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$, where $\sigma^2$ is known.

   (a) Show that Jeffreys' prior for $\mu$ is of the form $p(\mu) \propto 1$ for $-\infty < \mu < \infty$.

   (b) Hence show that the posterior distribution for $\mu$ is also normal, with mean and variance to be specified.

   (c) Suppose that we observe data $x_1, \ldots, x_{10}$ such that $\bar{x} = 10.1$. Assuming that $\sigma^2 = 1$, show that a 95% highest posterior density interval for $\mu$ is $(9.480, 10.720)$.

3. A chemist is interested in the maximum possible yield produced by a certain process. Due to the large variability in the data, they assume that, given a scalar $\theta$, each yield $x_i$, $i = 1, \ldots, n$, is independent of the other yields and follows a uniform distribution $U[0, \theta]$, so that

$$p(x_i|\theta) = \frac{1}{\theta}, \qquad \text{for } 0 < x_i < \theta.$$

   Before the chemist sees any data, they assume a Pareto prior distribution for $\theta$, so that

$$p(\theta) = \begin{cases} \frac{a x_o^a}{\theta^{a+1}} & \text{for } \theta \geq x_0; \\ 0 & \text{otherwise,} \end{cases}$$

   where $a > 0$ and $x_o > 0$ are known parameters for the prior Pareto distribution, specified by the chemist. Note that the mean of a Pareto distribution is given by $ax_0/(a-1)$, for $a > 1$, whilst the median if $x_0 2^{1/a}$.

   (a) Calculate the posterior distribution of $\theta$.

   (b) Suppose that the chemist specifies a Pareto prior distribution with $a = 2$, $x_0 = 0.1$. Consider observed data $\mathbf{x} = \{x_1, x_2, x_3\} = \{3, 10, 17\}$. Obtain the posterior distribution and indicate how the expert's beliefs have changed after observing the data, using point summary statistics.

(c) Suppose instead that the chemist specified the alternative prior $\theta \sim U(0, 15)$. What are the implications for the given observed data?

**4. ***

(a) Consider the general hypothesis testing problem

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 \theta \in \Theta_1,$$

such that the union of $\Theta_0$ and $\Theta_1$ is the whole of the parameter space $\Theta$. Letting $p_0$ and $p_1$ denote the prior probabilities for the null hypothesis and the alternative hypothesis respectively, show that the posterior probability of $H_0$ is given by

$$\mathbb{P}(H_0|\mathbf{x}) = \mathbb{P}(\theta \in \Theta_0|\mathbf{x}) = \frac{p_0}{p_0 + p_1/B_{01}},$$

where $B_{01}$ denotes the Bayes factor of $H_0$ to $H_1$.

(b) Now suppose that we observe data $\mathbf{x} = \{x_1, \ldots, x_n\}$, such that

$$X_i \sim^{iid} N(\mu, \sigma^2),$$

where $\sigma^2$ is known. We wish to test

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu = \mu_1.$$

Show that the Bayes factor is given by

$$B_{01} = \exp\left(-\frac{n(\mu_0 - \mu_1)(\mu_0 + \mu_1 - 2\bar{x})}{2\sigma^2}\right).$$

Calculate the Bayes factor for $H_0$ against $H_1$ when $\mu_0 = 0$, $\mu_1 = 1$, $\sigma^2 = 1$, $n = 9$ and $\bar{x} = 0.645$. What is your conclusion? What happens as we increase $n$, with all other values fixed?

(c) Finally, suppose that we observe data $\mathbf{x} = \{x_1, \ldots, x_n\}$, such that

$$X_i \sim^{iid} N(\mu, \sigma^2),$$

as before. Now we want to test the hypotheses:

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0.$$

We specify $p(\mu|H_1) \sim N(0, \tau^2)$. Calculate the Bayes factor for $H_0$ against $H_1$. Comment on the limiting case were we make the prior on $\mu$ under $H_1$ increasingly vague, i.e., $\tau^2 \to \infty$.

**5. ***We observe data $\mathbf{x} = \{x_1, \ldots, x_m\}$ from a multinomial distribution, $\mathbf{X} \sim \text{MN}(N, \mathbf{p})$, and wish to make inference on the parameeters $\mathbf{p} = \{p_1, \ldots, p_m\}$ (note that $\sum_i p_i = 1$). We set a prior on the unkonw parameters $\mathbf{p}$ of the form

$$\mathbf{p} \sim \text{Dir}(\alpha_1, \ldots, \alpha_m).$$

(a) Determine the corresponding posterior distribution for the parameters $\mathbf{p}$.

(b) Calculate the Bayes estimate for the parameters, assuming a quadratic loss function.

(c) We throw a 6-sided die 60 times and record the number of times that we observed the number $i = 1, \ldots, 6$, which we denote by $x_i$. LEt $p_i$ denote the associated probability of throwing the number $i = 1, \ldots, 6$ and set $\mathbf{p} = \{p_1, \ldots, p_6\}$. We observe the data $\mathbf{x} = \{10, 12, 12, 8, 7, 11\}$ and specify a Uniform prior on $\mathbf{p}$, which is $\text{Dir}(\alpha_1, \ldots, \alpha_6)$ for $\alpha_i = 1$ for $i = 1, \ldots, 6$. Determine the posterior mean for each $p_i$.

6. Suppose that $X_1, \ldots, X_n \sim^{iid} N(\mu, \sigma^2)$ where both $\mu$ and $\sigma^2$ are unknown. We specify the priors

$$\mu \sim N(0, s^2), \qquad \text{and } \sigma \sim U[0, T],$$

where $T$ is "large".

(a) Using a transformation of variables, calculate the corresponding prior on $\sigma^2$.

(b) Calculate the posterior conditional distribution of $\mu$ and $\sigma^2$ (i.e., the posterior distribution for $\mu$, treating $\sigma^2$ as fixed, and the posterior distribution of $\sigma^2$, treating $\mu$ as fixed).

**7.** *Suppose that we wish to simulate observations from the Pareto distribution with pdf

$$p(\theta) = \begin{cases} \frac{a x_o^a}{\theta^{a+1}} & \text{for } \theta \geq x_0; \\ 0 & \text{otherwise,} \end{cases} .$$

Derive the cumulative distribution function for $\theta$ and hence describe an algorithmic procedure for sampling random variables from $p(\theta)$ using the method of inversion.

**8.** * A biologist is interested in estimating the annual survival probability of a give species of deer, denoted by $\phi$. data are collected via a radio-tagging experiment which initially places radio-tags on a total of $N$ animals in year 0. Let $p_t$ denote the probability an individual dies within the interval $(t-1, t]$ for $t = 1, \ldots, T$ and $p_{T+1}$ the probability that they survive until the end of the study. Assuming survivability in each year is independent, and animals are independent of each other, we have

$$p_t = \begin{cases} (1-\phi) & t = 1 \\ (1-\phi)\phi^{t-1} & t = 2, \ldots, T \\ \phi^T & t = T+1 \end{cases} .$$

Let $X_t$ denote the number of individuals that die in the interval $(t-1, t]$ for $t = 1, \ldots, T$ and $X_{T+1}$ the number of individuals that survive until the end of the study. The corresponding likelihood is

$$p(\mathbf{x}|\theta) = \frac{N!}{\prod_{t=1}^{T+1} x_t!} \prod_{t=1}^{T+1} p_t^{x_t}.$$

Finally we specify the prior $\phi \sim$Beta$(\alpha, \beta)$.

(a) Show that the posterior distribution for the survival probability is of the form

$$p(\phi|\mathbf{x}) \sim \text{Beta}\left(\alpha + \sum_{t=1}^{T+1}(t-1)x_t, \beta + \sum_{t=1}^{T} x_t\right).$$

(b) To obtain a set of sampled realisations from the Beta distribution of interest the biologist intends to implement a rejection sampling algorithm. However, due to their limited computational skills they are only able to simulate random deviates from a $U[a, b]$ distribution. Describe a rejection sampling algorithm that the biologist can implement using their limited computational skills.

(c) We specify the prior $\phi \sim$Beta$(1, 1)$ and observe data such that the posterior $p(\phi|\mathbf{x} \sim$Beta$(91, 9)$. We obtain the following posterior summary statistics for $\phi$: posterior mean $\mathbb{E}(\phi) = 0.910$, posterior standard deviation 0.028, posterior median 0.913, 95% symmetric credible interval of $(0.847, 0.958)$ and a 95% highest posterior density interval of $(0.832, 0.949)$. Without conducting an analysis state why at least one of these summary statistics must be incorrect.

9. Radio-tagging data involves placing a radio-tag on a number of individuals and (assuming no radio failures) recording the number of deaths that occur at a series of successive "capture" times. We assume that only a single radio-tagging event occurs where a total of $n$ lambs are "tagged". We let $x_t$ denote the number of sheep that are subsequently recorded as having died within the interval $(t?1, t]$ (assuming tagging occurs at time 0), for $t = 1, \ldots, T$. We let $x_{T+1}$ denote the number of individuals that survive until time $T$ (i.e. the end of the study). The corresponding likelihood function is a function of the survival probabilities of the sheep. We assume two distinct survival probabilities: $\phi_1$ corresponding to first-year survival probability and $\phi_a$ the "adult" survival probability (i.e. older than first-years). The likelihood is given by

$$p(\mathbf{x}|\phi_1, \phi_a) \propto \prod_{i=1}^{T} p_i^{x_i}$$

where

$$p_i = \begin{cases} 1 - \phi_1 & i = 1 \\ \phi_1(1 - \phi_a) & i = 2 \\ \phi_1 \phi_a^{i-2}(1 - \phi_a) & i = 3, \ldots, T \\ \phi_1 \phi_a^{T-1} & i = T + 1 \end{cases}.$$

Without any prior information on $\phi_1$ or $\phi_a$ we set priors $\phi_1 \sim U[0, 1]$ and $\phi_a \sim U[0, 1]$ independently. Describe a Gibbs sampling algorithm for obtaining sample from the posterior distribution. Comment on the result.

**10.** ***Suppose that we wish to use the Metropolis-Hastings algorithm to generate a sample from $N(0, \sigma^2)$, and that we use the proposal $q(x, y) = N(ax, \tau^2)$ for $?1 < a < 1$.

(a) What is the corresponding acceptance probability $\alpha(x, y)$?

(b) For what value of $\tau^2$ would this particular sampler never reject the candidate value?

(c) What happens if $a = 0$?

11. Show that the Metropolis-Hastings algorithm for target distribution $\pi(x)$ generates a reversible Markov chain, such that for $x \neq y$

$$p(x)\mathcal{K}_H(x, y) = \pi(y)\mathcal{K}_H(y, x),$$

where $\mathcal{K}_H(x, y) = q(y|x)\alpha(x, y)$ and $\alpha(x, y)$ is the acceptance probability.

Hence show that

$$\int \pi(x)\mathcal{K}_H(x,y)\,\mathrm{d}x = \pi(y).$$

12. **Analysis of binomial data: drug**. Consider the example from lecture 4 where a new drug is being considered for relief of chronic pain, with the success rate $\theta$ being the proportion of patients experiencing pain relief. In the past, drugs of this type have shown variable pain relief rates, with a mean of 40% and a standard deviation of 10%. We have seen that these could be translated into a Beta$(9.2, 13.8)$ distribution. This drug had 15 successes out of 20 patients.

    (a) Calculate the posterior distribution of the success rate $\theta$.

    (b) What is the posterior mean and 95% highest posterior density (HPD) interval for the response rate?

    (c) Compute a symmetric 95% credible interval. Compare this to the 95% HPD interval.

    (d) What is the probability that the true success rate is greater than 0.6?

    (e) How is this value affected if a uniform prior is adopted? And how is it affected in the case that Jeffreys' prior is adopted?

    (f) Using the original Beta$(9.2, 13.8)$ prior, suppose 40 more patients were entered into the study. What is the chance that at least 25 of them experience pain relief? *Hint*: You might want to use the `beta` and `gamma` functions implemented in `R`.

    (g) We might ask whether the observed data is 'compatible' with the expressed prior distribution. One method is to calculate the predictive probability of observing such an extreme number of successes under this prior: this is a standard $p$-value but where the null hypothesis is a distribution. Use the predictive distribution for 20 future patients to find the probability of getting at least 15 successes (i.e., at least 15 patients experiencing pain relief). Do you think this suggests the data are incompatible with the prior?

    (h) Suppose that most drugs (95%) are assumed to come from the stated Beta$(9.2, 13.8)$ prior, but there is a small chance that the drug might be a 'winner'. 'Winners' are assumed to have a prior distribution with mean 0.8 and standard deviation 0.1.

        (i) What Beta distribution might represent the 'winners' prior? Remember that a Beta$(a, b)$ distribution has mean $a/(a+b)$ and variance $ab/\{(a+b)^2(a+b+1)\}$.

        (ii) Plot the mixture prior.

        (iii) What is now the chance that the response rate is greater than 0.6?
        *Hint*: You might start by showing that if

        $$\theta \sim \pi\text{Beta}(a_1, b_1) + (1-\pi)\text{Beta}(a_2, b_2),$$

        then

        $$\theta \mid y \sim \omega_1\text{Beta}(a_1 + y, b_1 + n - y) + (1-\omega_1)\text{Beta}(a_2 + y, b_2 + n - y),$$

        where

        $$\omega_1 = \pi\frac{B(a_1+y, b_1+n-y)}{B(a_1,b_1)}\left(\pi\frac{B(a_1+y, b_1+n-y)}{B(a_1,b_1)} + (1-\pi)\frac{B(a_2+y, b_2+n-y)}{B(a_2,b_2)}\right)^{-1}.$$

Here $y$ denotes the number of successes.

(iv) For this mixture prior, repeat the prior/data compatibility test performed previously. Are the data more compatible with this mixture prior?

(i) Repeat the above analysis numerically using `pystan`.

(i) Compute the posterior mean, standard deviation and a 95% credible interval. Compare with the exact results.

(ii) What is the probability that the true success rate is greater than 0.6. Compare with the exact result.

(iii) Suppose 40 more patients were entered into the study. What is the chance that at least 25 of them experience pain relief? Compare with the exact result.

(iv) Conduct the 'prior/data compatibility check', i.e., calculate the predictive probability of observing at least 15 successes under this prior. Compare with the exact result.

(v) For the mixture prior, what is now the chance that the response rate is greater than 0.6? Compare with the exact result.

(vi) Under this mixture prior, what is the posterior predictive probability that at least 25 out of 40 new patients experience pain relief?

(vii) For this mixture prior, repeat the prior/data compatibility test performed previously. Are the data more compatible with this mixture prior? Compare with the exact result.

**13.** ***We have data on the winning men's long jump distances (m) from 1900 through 2008, as follows

$$\begin{aligned}\text{Year} = \{&1900, 1904, 1906, 1908, 1912, 1920, 1924, 1928, 1932, 1936, 1948, 1952, 1956,\\ &1960, 1964, 1968, 1972, 1976, 1980, 1984, 1988, 1992, 1996, 2000, 2004, 2008\},\\ \text{Jump} = \{&7.185, 7.340, 7.200, 7.480, 7.600, 7.150, 7.445, 7.730, 7.640, 8.060, 7.825, 7.570,\\ &7.830, 8.120, 8.070, 8.900, 8.240, 8.350, 8.540, 8.540, 8.720, 8.670, 8.500, 8.550,\\ &8.590, 8.370\}.\end{aligned}$$

Fit a linear regression of the distances as a function of Olympic year:

$$Jump = \beta_0 + \beta_1 \text{Year} + \epsilon$$

assuming $\epsilon \sim N(0, \sigma^2 = 1/\tau^2)$. Use the following priors for the three parameters:

$$\begin{aligned}\beta_0, \beta_1 &\sim \text{Normal}\,(\mu_0 = 0, \tau_0 = 0.001)\\ \tau &\sim \text{Gamma}\,(a = 0.1, b = 0.1)\end{aligned}$$

(a) Fit the model using `pystan`. Generate trace plots, posterior distributions and summary statistics. Compute also the autocorrelation function, the effective sample size and the Gelman-Rubin statistic.

(b) Try centring the "Year" covariate, i.e., subtract its average value and use the centred covariate as the explanatory variable for sampling.

(c) Now repeat the analysis using "robust regression" by replacing the Normal distribution with a Student-$t$ distribution. Try fixing the degrees of freedom to $\nu = 3$ and allowing this to be a model parameter with a suitable Gamma prior. How do your results change?

**14.** ***A** study is conducted to measure the log-concentration of a particular chemical in soil. It can be assumed that the log-concentration follows a Normal distribution with mean $\mu$ and variance $\sigma^2$. A set, $\{y_i\}$, of $n = 10$ measurements are made of the log-concentration in soils in a certain region of the UK with the following results

$$-0.566, 3.74, 5.55, -1.90, -3.54, 5.16, -1.76, 4.08, 4.62, 0.732.$$

The primary parameter of interest is $\mu$, the unknown mean of the distribution. We assume initially that $\sigma^2 = 30$ is known.

(a) We consult a panel of 10 UK experts and they believe that $\mu \approx 1$, but could take values in the range $[0, 2]$. Construct a suitable conjugate prior based on this information and obtain the posterior distribution from the experimental data. Include a brief justification of your choice of prior.

(b) Suppose now that we had also consulted a panel of 5 US experts, and their opinion was that $\mu \approx 5$ but could be between 3 and 7. Derive a suitable mixture prior that combines the opinions of both sets of experts and obtain the posterior distribution for this new prior.

(c) Now suppose that $\sigma^2$ is also unknown. Using a suitable prior on the precision, $\tau = 1/\sigma^2$, and the same mixture prior on $\mu$, obtain samples from the posterior distribution numerically. You may wish to use the various convergence diagnostics that were discussed in lectures to verify the accuracy of your posterior distribution, but you do not need to report the results of such checks in your solution. Obtain estimates of the posterior mean, median, and the lower and upper quartiles from your samples.

(d) Based on your results, what is the posterior probability that $\mu < 1$? If we take 5 additional measurements, what is the probability that at least one of them will return a negative log-concentration?

(e) In building the mixture prior in part (b) you would have used some weighting between the two groups of experts, $p(\mu) = wp_1(\mu) + (1 - w)p_2(\mu)$. We will now include the weight $w$ as an additional model parameter. In the case that $\sigma^2 = 30$ is known, obtain the combined posterior distribution on $(\mu, w)$, using a flat prior for $w \in [0, 1]$. Obtain also the marginal distributions on $\mu$ and $w$ and comment on the result.