Max Planck Institute for Gravitational Physics
IMPRS Lecture Series

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

Problem Sheet 1: Frequentist Statistics and Stochastic Processes

---

IMPRS students taking this course should complete the questions in the first part of this sheet and hand them in to be marked. The questions in the second part of the sheet, labelled "Additional questions', are for personal study and do not need to be handed in.

Questions on Frequentist Statistics

1. For the Beta(a, b) distribution, find the mean, mode, variance, skewness and excess kurtosis.

2. Suppose $X \sim N(0, 1)$ with pdf

$$p(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}$$

   and $Y \sim \chi^2_n$ with pdf

$$p(y) = \frac{1}{2^{\frac{n}{2}}\Gamma(n/2)} y^{\frac{n}{2}-1} \mathrm{e}^{-\frac{y}{2}},$$

   and assume that X and Y are independent. Show that the distribution of $T = X/\sqrt{Y/n}$ has pdf

$$p(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}.$$

   This is the Student t-distribution with n degrees of freedom.

3. **Gravitational Wave Birthday Problem(s):**

   (a) How many gravitational wave sources would we have to observe before it is more likely than not we will have two events on the same date (i.e., day and month)?

   (b) Suppose we have observed n GW events in a particular category, say binary black hole mergers and then observe an event in a new category. What is the probability that the new event is on the same date as one of the previously observed events (consider both the case that we know all the events in the category are on different days, and the case where this is not specified)?

   (c) Given a rate of gravitational wave events of one per week, how many events would we have to observe before having a greater than 50% chance that two events were observed within 24 hours? [Hint: consider the distribution of the minimum difference between successive events and compute the probability that this is less than 1 day.]

   (d) (OPTIONAL) Given a rate of gravitational wave events of one per week, how long would we expect to wait before having a greater than 50% chance of observing two events in 24 hours? This latter question is considerably more difficult to answer than the previous one, but gives a very similar answer.

4. Independent Bernoulli r.v.s. $X_1, X_2, \ldots, X_n$ are such that the probability of X_i taking the value 1 depends on an explanatory variable z, which takes corresponding values $z_1, z_2, \ldots, z_n$.

   Show that for the model

$$\rho_j \quad = \quad \log \left\{ \frac{\Pr(X_j = 1)}{\Pr(X_j = 0)} \right\} \quad = \quad \alpha + \beta z_j,$$

   the minimal sufficient statistic for (α, β) is $\left(\sum_{j=1}^{n} X_j,\ \sum_{j=1}^{n} z_j X_j\right)$; the quantity ρ_j is called the logistic transformation.

5. Let $X_1, X_2, \ldots, X_n$ be a random sample from $U[0, \theta]$.

   (a) Find the p.d.f. of $X_{(n)}$, the largest of the X_is.

   (b) Show that $2\bar{X}$ (where $\bar{X}$ is the sample mean) and $(n+1)X_{(n)}/n$ are both unbiased consistent estimators of θ, and compare their variances.

6. Suppose that $x_1, \ldots, x_n$ form a random sample from a distribution with probability density function

$$f(x|\sigma^2) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right) \qquad (x > 0).$$

   Obtain the Cramér-Rao lower bound when the parameter of interest is $\theta = \sigma^2$.

   Determine whether the bound is attainable, and if it is attainable give the estimator which attains the bound.

7. Let $x_1, \ldots, x_n$ denote a random sample from a distribution with probability density function

$$p(x \mid \theta) = \frac{x}{\theta} \exp\left(-\frac{x^2}{2\theta}\right) \qquad (x > 0)$$

   where θ is a positive constant.

   (a) Obtain a minimal sufficient statistic for θ based on $x_1, \ldots, x_n$, and explain why it is minimal sufficient.

   (b) Show that the most powerful test of size α of

$$\begin{aligned} H_0 &: \theta = \theta_0, \\ \text{against } H_1 &: \theta = \theta_1 \qquad (\theta_1 > \theta_0), \end{aligned}$$

   involves a minimal sufficient statistic.

   Deduce the form of the uniformly most powerful test of $H_0 : \theta = \theta_0$ against the composite alternative hypothesis $H_1' : \theta > \theta_0$.

   (c) Let $Y_i = X_i^2/\theta$, $i = 1, \ldots, n$, where X_i is defined as above. Show that Y_i has an exponential distribution with mean 2, i.e. a χ_2^2 distribution.

   Deduce the critical value of the uniformly most powerful test of $H_0 : \theta = 1$ against $H_1' : \theta > 1$ in (b) when there are five observations and the size of the test is 5%. Find the power of the test as a function of θ.

8. Consider a simple post-Newtonian, frequency-domain model of a gravitational waveform

$$\tilde{h}(f) = \mathcal{A} f^{-\frac{7}{6}} e^{i\psi(f)},$$
$$\mathcal{A} = \frac{1}{\sqrt{30}\pi^{\frac{2}{3}}} \frac{\mathcal{M}^{\frac{5}{6}}}{D_L},$$
$$\psi(f) = 2\pi f t_c - \phi_c + \frac{3}{128}(\pi \mathcal{M} f)^{-5/3} \left\{ 1 + \left(\frac{3715}{756} + \frac{55}{9}\eta \right) \eta^{-\frac{2}{5}} (\pi \mathcal{M} f)^{\frac{2}{3}} \right\}. \quad (1)$$

Here $\mathcal{M}$ is the chirp mass, $\eta = m_1 m_2/(m_1+m_2)^2$ is the reduced mass ratio, ϕ_c and t_c are the phase and time of coalescence and D_L is the luminosity distance. Assuming that the source is observed in a frequency interval $[f_{\rm min}, f_{\rm max}]$ and the detector PSD is constant in that interval and equal to Σ^2, compute a Fisher Matrix to estimate the precision of parameter measurement uncertainties for the set $\{\mathcal{M}, \eta, \phi_c, t_c, D_L\}$.

Questions on Stochastic Processes

9. This question is focussed on calculating properties of gravitational wave source populations. For two different classes of source you will estimate properties such as the characteristic strains, signal-to-noise ratios and background spectrum generated by a population of such sources. There are two sub-parts to this question. At the end of each there is an optional part which is more open ended and involves some numerical investigation.

   (a) This question is concerned with properties of a background generated by a population of massive black holes. For the first few parts of this question you can assume that these binaries are identical, i.e., they all have the same values for the two masses, m_1 and m_2, and hence the derived quantities of total mass, $M = m_1 + m_2$, reduced mass, $\mu = m_1 m_2/M$, and chirp mass,

$$\mathcal{M}_c = \frac{m_1^{\frac{3}{5}} m_2^{\frac{3}{5}}}{M^{\frac{1}{5}}}.$$

   In the final part of this question you'll be asked to think about and play around with different assumptions about the mass distribution in the binaries.

   We will use geometric units throughout, i.e., we set $c = G = 1$ so we don't need to worry about keeping track of these factors.

      i. Assuming that the binary is Newtonian and circular, derive the following characteristic properties of the emitted gravitational waves [1].

         A. The GW amplitude scales like

$$h \sim \frac{1}{D} \mathcal{M}_c^{\frac{5}{3}} f^{\frac{2}{3}}.$$

         B. The GW energy loss scales like

$$\dot{E}_{\rm GW} \sim \mathcal{M}_c^{\frac{10}{3}} f^{\frac{10}{3}}.$$

[1] You may find it useful to recall that the energy of a Newtonian binary of semi-major axis a is $E = -M\mu/(2a)$ and the frequency is related to the semi-major axis via $2\pi f = \sqrt{M/a^3}$.

C. The rate of change of frequency scales like

$$\dot{f} \sim \mathcal{M}_c^{\frac{5}{3}} f^{\frac{11}{3}}.$$

D. The Fourier transform of $h(t)$ scales like

$$\tilde{h} \sim \frac{1}{D} \mathcal{M}_c^{\frac{5}{6}} f^{-\frac{7}{6}}.$$

E. The characteristic strain scales like

$$h_c \sim \frac{1}{D} \mathcal{M}_c^{\frac{5}{6}} f^{-\frac{1}{6}}.$$

F. The energy density of a GW background generated by a population of these sources scales like

$$\Omega_{\rm GW}(f) \sim \mathcal{M}_c^{\frac{5}{3}} f^{\frac{2}{3}}$$

In the above f denotes the orbital frequency of the binary and D the distance to the source.

ii. We now suppose that there is an additional process driving the evolution of the binaries, stellar hardening. This leads to an evolution of the semi-major axis, a, of the binary of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{a}\right) = k \frac{\rho_*}{\sigma^3} \frac{m_2}{a}$$

where k is a numerical constant, ρ_* is the stellar density and σ is the velocity dispersion of the stars. Show that this equation implies the energy loss from stellar hardening scales like

$$\dot{E}_{\rm hard} \propto \frac{\rho_* m_2 \mu}{\sigma^3} M^{\frac{2}{3}} f^{\frac{2}{3}}. \tag{2}$$

iii. Derive an expression for the corresponding GW background energy density, $\Omega_{\rm GW}(f)$.

iv. Assuming that the sources are at a common redshift (or the population is dominated by sources at a particular redshift), show that the spectrum is a broken power-law and find the asymptotic slopes at low and high frequency.

v. If a broken power law background were detected, what would it tell you about the relevant physical processes influencing the population? How is that information encoded in the background?

vi. (OPTIONAL) Try varying the assumptions about the population, e.g., the masses of the binary components or the properties of the stellar system or the redshift distribution of the sources. Compute the background from such populations numerically and explore how these assumptions change the results qualitatively and quantitatively.

(b) We now consider a population of burst sources. We will represent these as sine-Gaussians with a four parameter waveform family

$$h(t) = \frac{A}{D} \cos(2\pi f_0 t) \mathrm{e}^{-\frac{1}{2}Qt^2}.$$

i. Find the average waveform power

$$\langle h^2 \rangle = \frac{1}{2T} \int_{-T}^{T} h^2(t) \mathrm{d}t$$

and show that, for a reasonable choice of T, e.g., $T = 2/\sqrt{Q}$, it scales like A^2/D^2 as expected. Why is this a 'reasonable choice' for T?

ii. Compute the Fourier transform of $h(t)$. Hence deduce the bandwidth of the source is $\Delta f \sim \sqrt{Q}/\pi$.

iii. Estimate the SNR that might be possible in a burst search using

$$\left(\frac{S}{N}\right)^2 \approx \frac{\langle h^2 \rangle}{\Delta f S_n(f)}.$$

You may assume white noise, $S_n(f) = \sigma^2$, for simplicity.

iv. Compute the SNR that would be obtained if the source was found by matched filtering and compare it to the burst search SNR. Comment on your answer.

v. Compute the energy distribution for one of these bursts, $\mathrm{d}E/\mathrm{d}f$.

vi. Find an expression for the GW background energy density produced by a population of these bursts, assuming that A and Q are constant and the distribution of f_0 is a power law $\mathrm{d}n/\mathrm{d}f_0 \propto f_0^\alpha$ with $\alpha > -1$.

vii. Assuming the sources are at a common redshift, show that the asymptotic slope of the background is f^3 at low frequency and $f^{3+\alpha}$ at high frequency.

viii. (OPTIONAL) Use simulations on a computer to explore how things change under modifications of these various assumptions.

Additional questions on Frequentist statistics

10. Show that the t_n distribution with pdf

$$p(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}.$$

is properly normalised, i.e., the integral of the pdf is 1.

11. Derive the moment generating function for the exponential distribution, $\mathcal{E}(\lambda)$ and the Gamma(n, λ) distribution. Hence deduce that the distribution of the sum of n IID $\mathcal{E}(\lambda)$ random variables is Gamma(n, λ).

12. **Gravitational wave physicist birthday cake problem:** It is traditional at the Alfred Embleton Institute for gravitational wave physics that when one member of the institute has a birthday, they bring cake to share with the other members of the group. One student, Andrew Antony, is very fond of cake and would like to eat it at least once every two weeks.

    (a) Given that the institute has n members, compute the probability distribution of the maximum separation between birthdays. How large must n be such that the probability that the maximum separation is less than two weeks is greater than 50%?

    (b) The director of the institute, Alice Bunton, is concerned that the cakes are bad for the health of the researchers in her institute, and therefore wants to make sure these celebrations do not occur too often. Find the distribution of the minimum separation between birthdays. What is the maximum n should be to ensure the probability that the minimum separation is greater than 2 weeks is at least 50%?

    [*Note: all similarities to real institutes and researchers are purely coincidental.*]

13. A life test is conducted by installing n items of equipment at time 0 and recording at times $h, 2h, \ldots, mh$ the numbers n_r of items failing in the intervals $(r-1)h$ to rh $(r = 1, 2, \ldots, m)$, m and h being a fixed integer and a fixed time interval respectively. The time to failure is modelled as an Exponential (λ) distribution, and the lifetimes of different items are assumed independent.

    (a) Find the likelihood function for λ.

    (b) Hence determine sufficient statistics for λ.

14. Let $X_1, X_2, \ldots, X_n$ be independent r.v.s where X_i has p.d.f. $\theta_i e^{-\theta_i x}$, $x > 0$ where $\theta_i = (\alpha + i\beta)$ and α, β are unknown parameters.

    Find sufficient statistics for (α, β).

15. Let $X_1, X_2, \ldots, X_n$ be a random sample from the exponential distribution with p.d.f. $p(x|\,\lambda) = \lambda e^{-\lambda x} \quad x > 0, \quad \lambda > 0$.

    Find the maximum likelihood estimator, its mean and variance and the Cramer-Rao bound on the variance of unbiased estimators of λ. Hence show that the maximum likelihood estimator for λ is biased, consistent and asymptotically efficient.

16. Let $X_1, X_2, \ldots, X_n$ denote n independent, identically distributed random variables with a Bernoulli density $p(x|p) = p^x(1-p)^{1-x}$ for $x = 0, 1$. Show that X_1 is an unbiased estimator for p and compute its variance. Show that $S = \sum X_i$ is a sufficient statistic. Use the Rao-Blackwell theorem to obtain an estimator of lower variance and compute its variance.

17. **Linear modelling:** Consider observations

$$y_i = \beta^T \mathbf{x}_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$, β is a vector of parameters and $\mathbf{x}_i$ is a vector of k covariates for each observation y_i.

   (a) Show that the maximum likelihood estimate for β is

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

   where $\mathbf{X}$ is the *design matrix*, defined by $X_{ij} = (\mathbf{x}_i)_j$.

   (b) Show that the distribution of $\hat{\beta}$ is

$$N\left(\beta, \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right).$$

   (c) Show that the quantity

$$\hat{\sigma}^2 = \frac{\mathbf{y}^T\mathbf{y} - \hat{\beta}^T\mathbf{X}^T\mathbf{y}}{n-k}$$

   is an unbiased estimator of the variance σ^2. In fact it is relatively straightforward to show that this quantity is independent of $\hat{\beta}$ and follows a chi-squared distribution.

   (d) For a fixed constant vector $\mathbf{c}$, show that

$$\frac{\mathbf{c}^T\hat{\beta} - \mathbf{c}^T\beta}{\hat{\sigma}\sqrt{\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}}$$

   follows a t-distribution and hence deduce a 95% confidence interval for $\mathbf{c}^T\beta$.

18. Let $x_1, \ldots, x_n$ be observations of independent random variables $X_1, \ldots, X_n$ from the distribution with the probability density function

$$p(x_i \mid \theta) = \frac{(z_i\theta)^a}{\Gamma(a)} x_i^{a-1} e^{-\theta z_i x_i}, \quad x_i > 0,$$

with known covariates $z_i > 0$ and known $a > 0$, that is, $X_i \sim \Gamma(a, z_i\theta)$.

   (a) Derive the form of the most powerful test of size α, of the simple null hypothesis $H_0\colon \theta = 1$ against the simple alternative hypothesis $H_1\colon \theta = \theta_1$ $(\theta_1 > 1)$.

   (b) Deduce the form of the uniformly most powerful (UMP) test of the simple hypothesis $H_0\colon \theta = 1$ against the alternative hypothesis $H_1\colon \theta > 1$.

   (c) Does there exist a UMP test of the simple hypothesis $H_0 : \theta = 1$ against the alternative hypothesis $H_1\colon \theta \neq 1$?

(d) For observed data with $a = 2$, $n = 311$ and $\sum_i z_i y_i = 571$, test the hypothesis that $\theta = 1$ against $\theta > 1$.

[*Hint: use the Central Limit Theorem to find an approximate distribution of the test statistic.*]

(e) Find the power of the test $H_0 : \theta = 1$ against the alternative hypothesis $H_1 : \theta = 3$ as a function of n for $a = 2$ and $\alpha = 0.05$. Find the smallest n such that the power of the test is greater than 0.9.

[*Hint: use the Central Limit Theorem to find an approximate distribution of the test statistic.*]

(f) Determine the best critical regions of size α, of the simple null hypothesis $H_0 : \theta = \theta_0$ against the simple alternative hypothesis $H_1 : \theta = \theta_1$ $(\theta_1 > \theta_0)$. Use these critical regions to construct a one-sided 90% confidence interval for θ for the data given in (d).

19. Let $X_1, X_2, \ldots, X_n$ denote n independent, identically distributed random variables having a Poisson distribution with mean λ.

(a) Derive the form of the most powerful test, of size α, of the simple null hypothesis $H_0 \colon \lambda = \lambda_0$ against the simple alternative hypothesis $H_1 \colon \lambda = \lambda_1$ $(\lambda_1 > \lambda_0)$.

Deduce the form of the uniformly most powerful (UMP) test of the simple hypothesis $H_0 : \lambda = \lambda_0$ against the composite alternative hypothesis $H_1 : \lambda > \lambda_0$.

(b) Determine the moment generating function of X_i, and hence show that $\sum X_i$ has a Poisson distribution with parameter $n\lambda$.

Explain how the distribution of $\sum X_i$ may be used to determine a critical region for the test in (a), and obtain the critical value for a test with a nominal level of 5% when $n = 10$ and $\lambda_0 = 1$. Compare this critical value with an approximate critical value obtained by using a normal approximation to the distribution of $\sum X_i$.

(c) Calculate the power of the test in (b) when $\lambda = 2$.

(d) Suppose now that we require a test of $H_0 \colon \lambda = \lambda_0$ against the alternative $H_1 \colon \lambda \neq \lambda_0$. Determine whether a uniformly most powerful test exists. Calculate (approximate) critical values of a two-sided 5% level test obtained by using a normal approximation to the distribution of $\sum X_i$ when $n = 10$ and $\lambda_0 = 1$.