Max Planck Institute for Gravitational Physics

IMPRS Lecture Series

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

Lecturer: Jonathan Gair

Winter, 2023

This course will provide a general introduction to statistics, which will be useful for researchers working in the area of gravitational wave astronomy. It will start with some of the basic ideas from classical (frequentist) and Bayesian statistics then show how some of thee ideas are or will be used in the analysis of data from current and future gravitational wave (GW) detectors. The final section of the course will introduce some advanced topics that are also relevant to GW observations. These topics will not be expounded in great depth, but some of the key ideas will be described to provide familiarity with the concepts. The aim of the course will be to establish sufficient grounding in statistics that students will be able to understand research seminars and papers, and know where to begin if carrying out research in these areas.

The lectures will be supported by a number of computer practicals. Statisticians typically use the community software package R and this is also commonly used by researchers in other disciplines. Most new statistical methods that are developed are implemented as R packages and so familiarity with R will enable the user to carry out fairly sophisticated analyses straightforwardly. However, in physics it is more common these days to use PYTHON and there are a number of libraries of statistical functions and methods available for PYTHON as well. Therefore, the practicals will use PYTHON.

Course outline

1. **(lectures 1 to 6)** Classical (frequentist) statistics and stochastic processes.

   - Random variables: definition, properties, some useful probability distributions, central limit theorem.
   - Statistics: definition, estimators, likelihood, desirable properties of estimators, Cramer-Rao bound.
   - Hypothesis testing: definition, Neyman-Pearson lemma, power and size of tests, type I and type II errors, ROC curves, confidence regions, uniformly-most-powerful tests.
   - Frequentist statistics in GW astronomy: false alarm rates, Fisher Matrix, PSD estimation.
   - Stochastic processes, optimal filtering, signal-to-noise ratio, sensitivity curves.
   - Practical: simulating random variables in PYTHON.

2. **(lectures 7 to 12)** Bayesian statistics.

   - Bayes' theorem, conjugate priors, Jeffrey's prior.
   - Bayesian hypothesis testing, hierarchical models, posterior predictive checks.
   - Sampling methods for Bayesian inference.
   - Bayesian statistics in GW astronomy: parameter estimation, population inference, model selection.
   - Practical: Introduction to using PYTHON and PYMC3 for sampling probability distributions.

3. **(lectures 13 to 16)** Introduction to machine learning

Lecture notes and problem sets

Lecture notes will be provided for the whole course. Sections in the notes that are marked with a star will not be covered in lectures, but provide additional information that can be studied in your own time. In addition notes will be provided on three advanced topics that were included in the previous incarnation of this course, but will not be covered this year. Again, this material can be studied or used as a reference, but it is optional. The advanced topics are

- Time series analysis: auto-regressive processes, moving average processes, ARMA models.
- Nonparametric regression: kernel density estimation, smoothing splines, wavelets.
- Gaussian processes, Dirichlet processes.

Two problem sets will be made available, one for each of the first two parts of the course. Questions marked with an asterix are either more difficult or similar to other questions and are not compulsory.

Part I: Frequentist Statistics

1 Random variables

In classical physics most things are deterministic. There are physical laws governing the evolution of a system which can be solved and used to predict the state of the system in the future. In reality there are many situations in which things are not (or effectively not) deterministic, and so the outcome of an experiment cannot be predicted with certainty. However, if the experiment is repeated many times some outcomes will occur more frequently than others. This notion of in-deterministicity in measurements is encoded in the concept of a *random variable*. A random variable, X, is a quantity that, when observed, can take one of a (possibly infinite) number of values. Prior to making a measurement the value of the random variable cannot be predicted, but the relative frequency of the outcomes over many experiments are described by a *probability distribution*. The value that X takes in a particular observation (or experiment), x_i, say is called a *realisation* of the random variable.

Random variables can be *discrete*, in which case the values that the variable takes are drawn from a countable set of discrete possibilities, or *continuous* in which case the random variable may take on any value within one or more ranges.

1.1 Discrete random variables

A discrete random variable X can take on any of a (possibly infinite but countable) set of possible values, $\{x_1, x_2, \ldots)\}$, which together comprise the *sample space*. The probability that X takes any particular value is represented by a *probability mass function* (pmf), which is a set of numbers $\{p_i\}$ with the properties $0 \leq p_i \leq 1$ for all i and $\sum p_i = 1$. The probability that X takes the value x_i is p_i.

1.2 Examples of discrete random variables

1.2.1 Binomial and related distributions

The Binomial distribution is the distribution of the number of success in n trials for which the probability of success in one trial is p. We write $X \sim B(n, p)$ and

$$P(X = k) = p_k = \begin{cases} \left(\begin{array}{c} n \\ k \end{array}\right) p^k (1-p)^{n-k} & \text{if } k \in \{1, \ldots, n\}, \\ 0 & \text{otherwise} \end{cases} . \tag{1}$$

When $n = 1$ this is the *Bernoulli distribution*. The binomial distribution is the distribution of the sum of n Bernoulli trials, i.e., the number of "successes" in n trials. A related distribution is the *negative binomial distribution* which has pmf

$$P(X = k) = p_k = \begin{cases} \left(\begin{array}{c} k+r-1 \\ k \end{array}\right) p^k (1-p)^{r} & \text{if } k \in \{0, 1, \ldots\}, \\ 0 & \text{otherwise} \end{cases} . \tag{2}$$

This is the distribution of the number of successes in a sequence of Bernoulli trials that will be observed before r failures have been observed. Setting $r = 1$ and $p \to (1 - p)$ this is the *geometric distribution*, which is the distribution of the number of trials required before the first success.

Another generalisation of the Binomial distribution is the *multinomial distribution.* In this case the outcome of a trial is not a binary 'success' or 'fail', but it is one of k possible outcomes. The probability of each outcome is denoted p_i with $\sum_{i=1}^{k} p_i = 1$ and the multinomial distribution describes the probability of seeing n_1 occurrences of outcome 1, n_2 occurrences of outcome 2 etc. in n trials. The pmf is

$$P(\{n1,\dots,n_k\}) = \left\{ \begin{array}{ll} \frac{n!}{n_1!n_2!\dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k} & \text{if } n_i \geq 0 \; \forall i \text{ and } \sum_{i=1}^{k} n_i = n \\ 0 & \text{otherwise} \end{array} \right. . \tag{3}$$

Applications: counting problems, e.g., distribution of events in categories or time, trials factors.

1.2.2 Poisson distribution

This is the distribution of the number of occurrences of some event in a certain time interval if that event occurs at a *rate* λ. The quantity X follows a Poisson distribution, $X \sim P(\lambda)$ if

$$P(X=k) = p_k = \left\{ \begin{array}{ll} \lambda^k \mathrm{e}^{-\lambda}/k! & \text{if } k \in \{0,1,\dots\}, \\ 0 & \text{otherwise} \end{array} \right. . \tag{4}$$

The Poisson distribution is the limiting distribution of $B(n,p)$ as $n \to \infty$, $p \to 0$ with $np = \lambda$ fixed.

Applications: distribution of number of events in a population, e.g., gravitational wave sources.

1.3 Continuous random variables

A continuous random variable can take any (usually real, but the extension to complex RVs is straightforward) value within some continuous range, or some set of ranges, which together comprise the *sample space* $\mathcal{X}$. The probability that X takes a particular value is characterised by the *probability density function* (pdf), $p(x)$. The probability that X takes a value in the range x to $x + \mathrm{d}x$ is $p(x)\mathrm{d}x$. The pdf has the properties $0 \leq p(x) \leq 1$ for all $x \in \mathcal{X}$ and

$$\int_{x \in \mathcal{X}} p(x) \mathrm{d}x = 1. \tag{5}$$

For single valued random variables with non-disjoint sample spaces continuous random variables may also be characterised by the *cumulative density function* or CDF, defined as

$$P(X \leq x) = \int_{-\infty}^{x} p(x) \mathrm{d}x. \tag{6}$$

1.3.1 Uniform distribution

X is uniform on an interval (a,b), denoted $X \sim U[a,b]$ if the pdf is constant on the interval $[a,b]$

$$p(x) = \left\{ \begin{array}{ll} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{otherwise} \end{array} \right. . \tag{7}$$

X takes values only in the range $[a,b]$.

Applications: often used as an "uninformative" prior in parameter estimation.

1.3.2 Normal distribution

X is Normal with *mean* μ and *variance* σ^2, denoted $X \sim N(\mu, \sigma^2)$ if the pdf has the form

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{8}$$

X takes all values in the range $(-\infty, \infty)$. If $\mu = 0$ and $\sigma^2 = 1$ we say that X follows a *standard Normal distribution.*

Applications: distribution of noise fluctuations in a gravitational wave detector, priors on mass distribution, most common distribution to assume in parametric statistics.

1.3.3 Chi-squared distribution

X is chi-squared with k degrees of freedom, denoted $X \sim \chi^2(k)$ or χ^2_k is the pdf has the form

$$p(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{\frac{k}{2}-1} \mathrm{e}^{-\frac{x}{2}} \tag{9}$$

Here $\Gamma(n)$ is the Gamma function, defined by

$$\Gamma(n) = \int_0^\infty x^{n-1} \mathrm{e}^{-x} \mathrm{d}x \tag{10}$$

and such that $\Gamma(n+1) = n!$. X takes non-negative real values only, $x \in [0, \infty)$. This is the distribution of the sum of the squares of n independent standard normal distributions.

There is also a *non-central chi-square distribution* which depends on two parameters — degrees of freedom, $k > 0$, as before plus a *non-centrality parameter*, $\lambda > 0$. This has the pdf

$$p(x) = \frac{1}{2} \mathrm{e}^{-\frac{(x+\lambda)}{2}} \left(\frac{x}{\lambda}\right)^{\frac{k}{4}-\frac{1}{2}} I_{\frac{k}{2}-1}(\sqrt{\lambda x}) \tag{11}$$

where $I_\nu(y)$ is the modified Bessel function of the first kind. The non-central chi-square distribution again takes non-negative values only and arises as the distribution of the sum of k independent normal distributions with equal (unit) variance, but non-zero means, denoted μ_i. The non-centrality parameter is then $\lambda = \sum_{i=1}^k \mu_i^2$.

Applications: used to test for deviations from normality, e.g., in noise fluctuations in a gravitational wave detector.

1.3.4 Student's t-distribution

X follows Student's t-distribution with $n > 0$ degrees of freedom, $X \sim t_n$, if it has pdf

$$p(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}. \tag{12}$$

The Student t-distribution arises in hypothesis testing as the distribution of the ratio of a standard Normal distribution to the square root of an independent χ^2_n distribution, normalised by the degrees of freedom. Specifically if $X \sim N(0, 1$ and $Y \sim \chi^2_n$ then $X/\sqrt{Y/n}$ follows a t_n distribution.

Applications: used for statistical test on significance of parameters in linear models, used as a "heavy-tailed" distribution for robust parameter estimation, arises naturally when marginalising over uncertainty in power-spectral density estimation.

1.3.5 F-distribution

X follows an F-distribution with degrees of freedom $n_1 > 0$ and $n_2 > 0$ if it has pdf

$$p(x) = \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} x^{\frac{n_1}{2}-1} \left(1 + \frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}} \tag{13}$$

where $B(a,b)$ is the beta function, which is given by

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\mathrm{d}x \tag{14}$$

and is related to the Gamma function through $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$. The F-distribution arises as the ratio of two independent chi-squared distributions with n_2 and n_2 degrees of freedom.

Applications: arises primarily in analysis of variance to test differences between groups.

1.3.6 Exponential distribution

X is exponential with *rate* $\lambda > 0$, $X \sim \mathcal{E}(\lambda)$ if it has pdf

$$p(x) = \begin{cases} \lambda\mathrm{e}^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

X takes positive real values only, $x \in (0, \infty)$. The exponential distribution is the distribution of the time that elapses between successive events of a Poisson process.

Applications: distribution of time lag between events, e.g., gravitational wave signals.

1.3.7 Gamma distribution

X is Gamma with parameters $n > 0$ and $\lambda > 0$, $X \sim$Gamma(n, λ), if it has pdf

$$p(x) = \begin{cases} \frac{1}{\Gamma(n)}\lambda^n x^{n-1} e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

X takes positive real values only, $x \in (0, \infty)$. The Gamma distribution is the distribution of of the sum of n exponential distributions with parameter λ.

Applications: conjugate distribution to the Poisson distribution, so useful in Bayesian analysis of rates. Useful as prior distribution whenever variable has support on $[0, \infty)$.

1.3.8 Beta distribution

X is Beta with parameters $a > 0$ and $b > 0$, $X \sim$Beta(a, b), if it has pdf

$$p(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

X takes values in the range $x \in (0, 1)$ only.

Applications: conjugate to binomial distribution. Useful as prior when variable has support on $[0, 1]$, e.g., for probabilities.

1.3.9 Dirichlet distribution

The Dirichlet distribution is a multivariate extension of the Beta distribution. A realisation of a Dirichlet random variable is a set of K values, $\{x_i\}$, satisfying the constraints $0 < x_i < 1$ for all i and $\sum_{i=1}^{K} x_i = 1$. The Dirichlet distribution is characterised by a vector of *concentration parameters* $\vec{\alpha} = (\alpha_1, \ldots, \alpha_K)$ satisfying $\alpha_i > 0$ for all i and has pdf

$$p(x) = \frac{1}{B(\vec{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}, \qquad \text{where } B(\vec{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{j=1}^{K} \alpha_j\right)}. \tag{18}$$

Applications: infinite dimensional generalisation is a Dirichlet process which is used as a distribution on probability distributions. Very important in Bayesian nonparametric analysis.

1.3.10 Cauchy distribution

X follows a Cauchy distribution (also known as a Lorentz distribution) with *location parameter* x_0 and *scale parameter* $\gamma > 0$, if it has pdf

$$p(x) = \frac{1}{\pi\gamma \left[1 + \left(\frac{x - x_0}{\gamma}\right)^2\right]}. \tag{19}$$

X takes any real value $x \in (-\infty, \infty)$. The Cauchy distribution arises as the distribution of the x intercept of a ray issuing from the point (x_0, γ) with a uniformly distributed angle. It is also the distribution of the ratio of two independent zero-mean Normal distributions.

Applications: used to model distributions with sharp features. In a gravitational wave context it is used as a model for lines in the spectral density of gravitational wave detectors, for example in BayesLine (and hence BayesWave).

1.4 Properties of random variables

The pdf (or pmf) of a random variable tells us everything about the random variable. However, it is often convenient to work with a smaller number of quantities that summarise the properties of the distribution. These characterise the 'average' value of a random variable and the spread of the random variable about the average. We summarise a few of these quantities here. They all rely on the notion of an *expectation value*, denoted $\mathbb{E}$. The expectation value of a function, $T(X)$, of a discrete random variable X is defined by

$$\mathbb{E}(T(X)) = \sum_{i=1}^{\infty} p_i t(x_i). \tag{20}$$

A similar definition holds for continuous random variables by replacing the sum with an integral

$$\mathbb{E}(T(X)) = \int_{-\infty}^{\infty} p(x) t(x) \mathrm{d}x. \tag{21}$$

1.4.1 Quantities representing the average value of a random variable

- **Mean** The mean, often denoted μ, is the expectation value of X, $\mu = \mathbb{E}(X)$.

- **Median** The median, m, is the central value of the distribution in probability, i.e., a value such that the probability of obtaining a value smaller than that or larger than that is (roughly) equal. For discrete random variables $m = x_k$, where

$$\sum_{i:x_i<x_k} p_i < 0.5 \qquad \text{and} \sum_{i:x_i\leq x_k} p_i \geq 0.5. \tag{22}$$

  For continuous random variables m is the value such that

$$\int_{-\infty}^{m} p(x)\mathrm{d}x = \int_{m}^{\infty} p(x)\mathrm{d}x = \frac{1}{2}. \tag{23}$$

- **Mode** The mode, M, is the 'most probable' value of the random variable. For discrete random variables

$$M = \mathrm{argmax}_{i\in\mathcal{X}} p_i \tag{24}$$

  and for continuous random variables

$$M = \mathrm{argmax}_{x\in\mathcal{X}} p(x). \tag{25}$$

  The mode may not be unique.

1.4.2 Quantities representing the spread of a random variable

- **Variance** The variance, often denoted σ^2, is the expectation value of the squared distance from the mean, i.e.,

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right]. \tag{26}$$

- **Standard deviation** The standard deviation is simply the square root of the variance, usually denoted σ.

- **Covariance** When considering two random variables, X and Y say, the covariance is defined as the expectation value of the product of their distance from their respective means, i.e.,

$$\mathrm{cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}(X))\,(Y - \mathbb{E}(Y))\right]. \tag{27}$$

  Here the expectation value is taken with respect to the joint distribution (see section on independence below).

- **Skewness** Given the mean, μ, and variance, σ^2, defined above, the skewness of a distribution is

$$\gamma_1 = \mathbb{E}\left[\left(\frac{x-\mu}{\sigma}\right)^3\right]. \tag{28}$$

- **Kurtosis** In a similar way, kurtosis is defined as

$$\mathrm{Kurt}(X) = \mathbb{E}\left[\left(\frac{x-\mu}{\sigma}\right)^4\right]. \tag{29}$$

  This measures the heaviness of the tails of the distribution of the random variable. The kurtosis of the Normal distribution is 3, so it is common to quote *excess kurtosis*, which is the kurtosis minus 3, i.e., the excess relative to the Normal distribution.

- **Higher moments** Higher moments can be defined in a similar way. The n'th moment about a reference value c of a probability distribution is

$$\mathbb{E}\left[(X-c)^n\right]. \tag{30}$$

  Moments are usually defined with c taken to be the mean, μ, as in the definition of skewness and kurtosis above.

1.4.3 Moment generating functions

A useful object for computing summary quantities of a probability distribution is the *moment generating function*, $M_X(t)$, which is defined as

$$M_X(t) = \mathbb{E}\left[\mathrm{e}^{tX}\right] \quad t \in \mathbb{R}. \tag{31}$$

It is clear that derivatives of this function with respect to t, evaluated at $t = 0$, give successive moments about zero of the distribution. Moment generating functions (MGFs) are defined in the same way for both discrete and continuous random variables.

In Table 1 we list these various summary quantities for the probability distributions listed earlier. Where quantities are not known in closed form they are omitted from this table.

Distribution	Mean	Median	Mode	Variance	Skewness	Excess kurtosis	MGF
Binomial(n, p)	np	$\lfloor np \rfloor$	$\lfloor (n+1)p \rfloor$	$np(1-p)$	$\frac{1-2p}{\sqrt{np(1-p)}}$	$\frac{1-6p(1-p)}{np(1-p)}$	$(1-p+p\mathrm{e}^t)^n$
Poisson(λ)	λ	$\approx \lfloor \lambda + \frac{1}{3} - \frac{0.02}{\lambda} \rfloor$	$\lceil \lambda \rceil - 1, \lfloor \lambda \rfloor$	λ	$\lambda^{-\frac{1}{2}}$	λ^{-1}	$\exp\left[\lambda(\mathrm{e}^t - 1)\right]$
Uniform$[a, b]$	$\frac{1}{2}(a+b)$	$\frac{1}{2}(a+b)$	all	$\frac{1}{12}(b-a)^2$	0	$-\frac{6}{5}$	$\frac{\mathrm{e}^{tb}-\mathrm{e}^{ta}}{t(b-a)}$
Normal(μ, σ^2)	μ	μ	μ	σ^2	0	0	$\exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right]$
χ^2_n	n	$\approx n\left(1 - \frac{2}{9n}\right)^3$	$\max(n-2, 0)$	$2n$	$\sqrt{\frac{8}{n}}$	$\frac{12}{n}$	$(1-2t)^{-k/2}$
Student's t_n	0	0	0	$\frac{n}{n-2}$	0 for $n > 3$	$\frac{6}{n-4}$ for $n > 4$	—
$\mathrm{F}(n_1, n_2)$	$\frac{n_1}{n_2-2}$	—	$\frac{n_2(n_1-2)}{n_1(n_2+2)}$	$\frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$	$\frac{(2n_1+n_2-2)\sqrt{8(n_2-4)}}{(n_2-6)\sqrt{n_1(n_1+n_2-2)}}$	see caption	—
$\mathcal{E}(\lambda)$	$\frac{1}{\lambda}$	$\frac{\ln 2}{\lambda}$	0	$\frac{1}{\lambda^2}$	2	6	$\frac{\lambda}{\lambda-t}$
Gamma(n, λ)	$\frac{n}{\lambda}$	—	$\frac{n-1}{\lambda}$	$\frac{n}{\lambda^2}$	$\frac{2}{\sqrt{n}}$	$\frac{6}{n}$	$\left(1 - \frac{t}{\lambda}\right)^{-n}$
Beta(a, b)	$\frac{a}{a+b}$	$I^{[-1]}_{\frac{1}{2}}(a, b)$	$\frac{a-1}{a+b-2}$	$\frac{ab}{(a+b)^2(a+b+1)}$	$\frac{2(b-a)\sqrt{a+b+1}}{(a+b+2)\sqrt{ab}}$	see caption	see caption
Dirichlet $(K, \vec{\alpha})$	$\frac{\alpha_i}{\sum_{j=1}^K \alpha_j}$	—	$\frac{\alpha_i - 1}{\sum_{j=1}^K \alpha_j - K}$	$\frac{\bar{\alpha}_i(1-\bar{\alpha}_i)}{\alpha_0+1}$	—	—	—
Cauchy (x_0, γ)	undefined	x_0	x_0	undefined	undefined	undefined	does not exist

Table 1: Summary of important properties of common probability distributions. The excess kurtosis of the F distribution is $12n_1(5n_2 - 22)(n_1+n_2-2) + (n_2-4)(n_2-2)^2/[n_1(n_2-6)(n_2-8)(n_1+n_2-2)]$. For the Beta$(a, b)$ distribution, the excess kurtosis is $6[(a-b)^2(a+b+1) - ab(a+b+2)]/[ab(a+b+2)(a+b+3)]$ and the MGF is $1 + \sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{a+r}{a+b+r}\right)\frac{t^k}{k!}$. For the Dirichlet distribution, the mean and variance are quoted for one component of the distribution, x_i, the parameters $\alpha_0 = \sum_{j=1}^K \alpha_j$ and $\bar{\alpha}_i = \alpha_i / \sum_{j=1}^K \alpha_j$ and the covariance $\mathrm{cov}(x_i, x_j) = -\bar{\alpha}_i\bar{\alpha}_j/(1+\alpha_0)$.

1.5 Independence

Most of the random variables described above are single valued, but a few of them, e.g., the multinomial and Dirichlet distributions, return multiple values. In other situations, several random variables might be evaluated simultaneously, or sequentially, or the same random variable might be observed multiple times. When dealing with multiple random variables, covariance as introduced above is an important concept, as is *independence*. A set of random variables $\{X_1, \ldots, X_N\}$ are said to be *independent* if

$$P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_N \leq x_N) = P(X_1 \leq x_1)P(X_1 \leq x_1) \ldots P(X_1 \leq x_1) \quad \forall\, x_1,\, x_2,\, \ldots,\, x_N. \tag{32}$$

In terms of the pdf (or pmf) the random variables are independent if their joint distribution $p(x_1, \ldots, x_N)$ can be separated

$$p(x_1, \ldots, x_N) = p_{X_1}(x_1)p_{X_2}(x_2) \ldots p_{X_N}(x_N). \tag{33}$$

Independence of two random variables implies that the covariance is 0, but the converse is not true except in certain special cases, for example for two Normal random variables.

A set of variables $\{X_i\}$ is called *independent identically distributed* or IID if they are independent and all have the same probability distribution. This situation arises often, for example when taking multiple repeated observations with an experiment.

1.6 Linear combinations of random variables

Suppose $X_1, \ldots, X_N$ are (not necessarily independent) random variables and consider a new random variable Y defined as

$$Y = \sum_{i=1}^{N} a_i X_i. \tag{34}$$

For any set of random variables

$$\mathbb{E}(Y) = \sum_{i=1}^{N} a_i \mathbb{E}(X_i), \qquad \mathrm{Var}(Y) = \sum_{i=1}^{N} a_i^2 \mathrm{Var}(X_i) + \sum_{i \neq j} a_i a_j \mathrm{cov}(X_i, X_j). \tag{35}$$

If the random variables are *independent* then the variance expression simplifies to

$$\mathrm{Var}(Y) = \sum_{i=1}^{N} a_i^2 \mathrm{Var}(X_i) \tag{36}$$

and the moment generating function of Y can be found to be

$$M_Y(t) = \prod_{i=1}^{N} M_{X_i}(a_i t). \tag{37}$$

A commonly used linear combination of random variables is the *sample mean* of a set of IID random variables, defined as

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} X_i \tag{38}$$

for which

$$\mathbb{E}(\hat{\mu}) = \mathbb{E}(X_1), \qquad \mathrm{Var}(\hat{\mu}) = \frac{1}{n}\mathrm{Var}(X_1), \qquad M_{\hat{\mu}}(t) = \left(M_{X_1}\left(\frac{t}{N}\right)\right)^N. \tag{39}$$

1.7 Laws of large numbers

Suppose that $X_1, \dots, X_n$ are a sequence of IID random variables, each having finite mean μ and variance σ^2. We denote the sum of the random variables by

$$S_n = \sum_{i=1}^{n} X_i, \qquad \text{which implies } \mathbb{E}(S_n) = n\mu, \quad \mathrm{Var}(S_n) = n\sigma^2. \tag{40}$$

Laws of large numbers tells us that the sample mean becomes increasingly concentrated around the mean of the random variable as the number of samples tends to infinity.

1.7.1 Weak law of large numbers

The *weak law of large numbers* states that, for $\epsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) \to 0, \text{ as } n \to \infty. \tag{41}$$

1.7.2 Strong law of large numbers

The *strong law of large numbers* states simply

$$P\left(\frac{S_n}{n} \to \mu\right) = 1. \tag{42}$$

1.7.3 Central limit theorem

In many applications, people assume that the data generating process is Normal. This is partially because the Normal distribution is convenient to work with and has many nice properties, but also because regardless of the distribution large samples of random variables tend to look quite Normally distributed. This fact is encoded in the *Central Limit Theorem*, which states that the standardized sample mean, S_n^*, is approximately standard Normal in the limit $n \to \infty$

$$S_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}}. \tag{43}$$

Formally the statement of the central limit theorem is

$$\mathrm{lim}_{n\to\infty} P(a \leq S_n^* \leq b) = \Phi(b) - \Phi(a) = \mathrm{lim}_{n\to\infty} P(n\mu + a\sigma\sqrt{n} \leq S_n \leq n\mu + b\sigma\sqrt{n}). \tag{44}$$

2 Frequentist statistics

In the last section we discussed the notion of a random variable. When observing phenomena in nature or performing experiments we would like to deduce the distribution of the random variable, i.e., the probability distribution from which realisations of that random variable are drawn. In **parametric inference** we assume that the distribution of the random variable takes a particular form, i.e., it belongs to a known family of probability distributions. All of the distributions that were described in the previous section are characterised by one or more parameters and so inference about the form of the distribution reduces to inference about the values of those parameters.

In *frequentist statistics* we assume that the parameters characterising the distribution are *fixed* but *unknown.* Statements about the parameters, for example *significance* and *confidence* are statements about multiple repetitions of the same observation, with the parameters fixed. Key frequentist concepts are *statistics*, *estimators* and *likelihood.*

A **statistic** is a random variable or random vector $T = t(\mathbf{X})$ which is a function of $\mathbf{X}$ but does not depend on the parameters of the distribution, θ. Its realised value is $t = t(\mathbf{x})$. In other words a statistic is a function of observed data only, not the unknown parameters.

An **estimator** is a statistic used to estimate the value of a parameter. Typically the random vector would be a set of IID random variables, $X_1, \ldots, X_n$ with pdf $p(x|\,\theta)$. A function $\widehat{\theta}(X_1, \ldots, X_n)$ of $X_1, \ldots, X_n$ used to infer the parameter values is called an **estimator** of θ; note that $\widehat{\theta}$ is a random variable with a sampling distribution in this latter context. The value of the estimator at the observed data $\widehat{\theta}(x_1, \ldots, x_n)$ is called an **estimate** of θ.

A statistic might also be used to provide an upper or lower limit for a *confidence interval* on the value of a parameter, or to evaluate the validity of a hypothesis in *hypothesis testing.*

2.1 Likelihood

Likelihood is central to the theory of frequentist parametric inference.

If an event E has probability which is a specified function of parameters $\vec{\theta}$, then the likelihood of E is $\mathbb{P}(E|\,\vec{\theta})$, regarded as a function of $\vec{\theta}$.

The likelihood, denoted $L(\vec{\theta}; \mathbf{x})$, is functionally the same as the pdf of the data generating process, the difference is that the likelihood is regarded as a function of the parameters $\vec{\theta}$ while the pdf is regarded as a function of the observed data, $\mathbf{x}$. It is often convenient to work with the **log likelihood**

$$l(\theta; \mathbf{x}) = \ln[L(\theta; \mathbf{x})] = \ln[p(\mathbf{x}|\,\theta)] \qquad (\theta \in \Theta)$$

Another useful quantity is the **score**

$$\frac{\partial l}{\partial \theta_i}$$

which is a vector that is also regarded as a function of $\vec{\theta}$ with the data fixed at the observed values.

One interpretation of likelihood is that, given data $\mathbf{x}$, the relative plausibility of or support for different values $\vec{\theta}_1$, $\vec{\theta}_2$ of $\vec{\theta}$ is expressed by

$$\frac{L(\vec{\theta}_1; \mathbf{x})}{L(\vec{\theta}_2; \mathbf{x})} \qquad \text{or} \qquad l(\vec{\theta}_1; \mathbf{x}) - l(\vec{\theta}_2; |\mathbf{x}).$$

As a result, inferences are unchanged if $L(\vec{\theta}|\mathbf{x})$ is multiplied by a positive constant (possibly depending on $\mathbf{x}$).

Typically we will be interested in cases where we observe more than one independent realisation of the random variable. For discrete random variables the combined likelihood is then the product of the likelihoods of each observed event.

Example: Poisson distribution

We observe a set $\{x_1, \ldots, x_n\}$, of n IID observations from a Poisson distribution with parameter λ. Denoting $n\bar{x} = \sum_{j=1}^{n} x_j$ the likelihood is

$$L(\theta;\mathbf{x}) = \frac{e^{-n\lambda}\lambda^{n\bar{x}}}{\prod_j x_j!} \qquad (\lambda > 0)$$

$$l(\lambda;\mathbf{x}) = \log\left(L(\lambda;\mathbf{x})\right] = -n\lambda + n\bar{x}\ln\lambda - \ln(\prod_j x_j!)$$

For continuous random variables the joint likelihood can usually be written as

$$L(\theta;\mathbf{x}) = \prod_{j=1}^{n} p(x_j|\,\theta) \qquad \Rightarrow \qquad l(\theta;\mathbf{x}) = \sum_{j=1}^{n} l(x_j|\,\theta).$$

or just $p(\mathbf{x}|\,\theta)$ for a vector $\mathbf{x}$ of random variables that are not IID. One case where this does not necessarily hold is when measurements are imperfect. Typically we cannot observe a quantity with infinite precision, but inevitably round to the nearest measurement unit. Observations of continuous random variables therefore typically involve grouping measurements into bins.

Suppose random variables $X_1, \ldots, X_n$ are IID with cumulative distribution function $P(x|\,\vec{\theta})$ and we observe that there are $n_1, \ldots, n_k$ observations in each of the k intervals $(a_0, a_1], \ldots, (a_{k-1}, a_k]$, where $-\infty \le a_0 < a_1 < \ldots < a_k \le \infty$ and $\mathbb{P}(a_0 < X_j \le a_k) = 1$. The distribution of $(N_1, \ldots N_k)$ is Multinomial with parameters $(n, p_1(\vec{\theta}), \ldots p_k(\vec{\theta}))$ with

$$p_r(\vec{\theta}) = \mathbb{P}(a_{r-1} < X_j \le a_r|\,\vec{\theta}) = P(a_r|\,\vec{\theta}) - P(a_{r-1}|\,\vec{\theta}),$$

and the likelihood is given by (3). For example, with common distribution $N(\mu, \sigma^2)$ we have

$$p_r(\mu,\sigma^2) = \Phi\left(\frac{a_r - \mu}{\sigma}\right) - \Phi\left(\frac{a_{r-1} - \mu}{\sigma}\right).$$

If observations of the IID random variables are made with a resolution (or maximum grouping error)of $\pm\frac{1}{2}h$, then we are effectively in the above situation, and a recorded value x represents a value in the range $x \pm \frac{1}{2}h$. Assuming that the grouping error is small, the likelihood is

$$\prod_{j=1}^{n}\{P(x_j + \frac{1}{2}h|\,\theta) - P(x_j - \frac{1}{2}h|\,\theta)\}. \tag{45}$$

If $p(x|\,\theta)$ does not vary too rapidly in each interval $(x_j - \frac{1}{2}h, x_j + \frac{1}{2}h)$ then (45) can be approximated by

$$\prod_{j=1}^{n}\{hp(x_j|\,\theta)\},$$

or, ignoring the constant h^n,

$$L(\theta;\mathbf{x}) \simeq \prod_{j=1}^{n} p(x_j|\,\theta).$$

which is the result we wrote down when there was no grouping error. However, this argument can fail, as illustrated in the two examples below.

Examples where this approximation fails

- Single observation from $N(\mu,\sigma^2)$

$$\begin{aligned} L(\mu,\sigma|x) &= \Phi\left\{\frac{x+\frac{1}{2}h-\mu}{\sigma}\right\} - \Phi\left\{\frac{x-\frac{1}{2}h-\mu}{\sigma}\right\} & (46) \\ &\simeq \frac{h\ \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)}{\sqrt{2\pi}\sigma} & (47) \end{aligned}$$

  if $\sigma > h$. If $\mu = x$ and $\sigma \to 0$, (46)$\to 1$ but (47)$\to \infty$.

- Uniform distribution on $[0,\theta]$, $U(0,\theta)$
  If $X_1,\ \ldots\ ,X_n$ are IID with pdf given by

$$p(x|\,\theta) = \left\{ \begin{array}{ll} \frac{1}{\theta} & (0 < x \leq \theta) \\ 0 & \text{otherwise} \end{array} \right.$$

  then

$$p(\mathbf{x}|\,\theta) = \left\{ \begin{array}{ll} \frac{1}{\theta^n} & (0 < x_{(n)} \leq \theta) \\ 0 & \text{otherwise} \end{array} \right.$$

  where $x_(i)$ denotes the i'th element in the ordered sequence of $\{x_i\}$. The likelihood is

$$L(\theta;\mathbf{x}) \simeq \left\{ \begin{array}{ll} 0 & (\theta < x_{(n)}) \\ \frac{1}{\theta^n} & (\theta \geq x_{(n)}) \end{array} \right. \qquad (48)$$

  Taking account of a grouping error of $\pm\frac{1}{2}h$, the probability assigned to $(x_j - \frac{1}{2}h, x_j + \frac{1}{2}h)$ is

$$\left\{ \begin{array}{ll} \frac{h}{\theta} & (x_j + \frac{1}{2}h < \theta) \\ \frac{\theta - x_j + \frac{1}{2}h}{\theta} & (x_j - \frac{1}{2}h \leq \theta < x_j + \frac{1}{2}h) \end{array} \right.$$

  and, if $h \leq x_{(n)} - x_{(n-1)}$,

$$L(\theta;\mathbf{x}) \propto \left\{ \begin{array}{ll} 0 & (\theta < x_{(n)} - \frac{1}{2}h) \\ \frac{\left[(\theta - x_{(n)} + \frac{1}{2}h)/h\right]^a}{\theta^n} & (x_{(n)} - \frac{1}{2}h \leq \theta < x_{(n)} + \frac{1}{2}h) \\ \frac{1}{\theta^n} & (\theta > x_{(n)} + \frac{1}{2}h) \end{array} \right. \qquad (49)$$

  where a is the number of observations equal to $x_{(n)}$. The continuous likelihood (Eq. (48)) and the likelihood accounting for grouping error (Eq. (49)) are shown in Figure 1.

  Ignoring grouping, $x_{(n)}$ is the ML estimator and has variance of order n^{-2}; with grouping the asymptotic variance is the usual $O(n^{-1})$.

To summarise: if the precision of observing the data (h) is much smaller than the variability of the data (e.g. than the standard deviation) then it is fine to use the approximation of the likelihood by the density. However, if the precision h is comparable with the variability, in order to estimate the unknown parameters reliably, one has to use the discrete version of the likelihood.

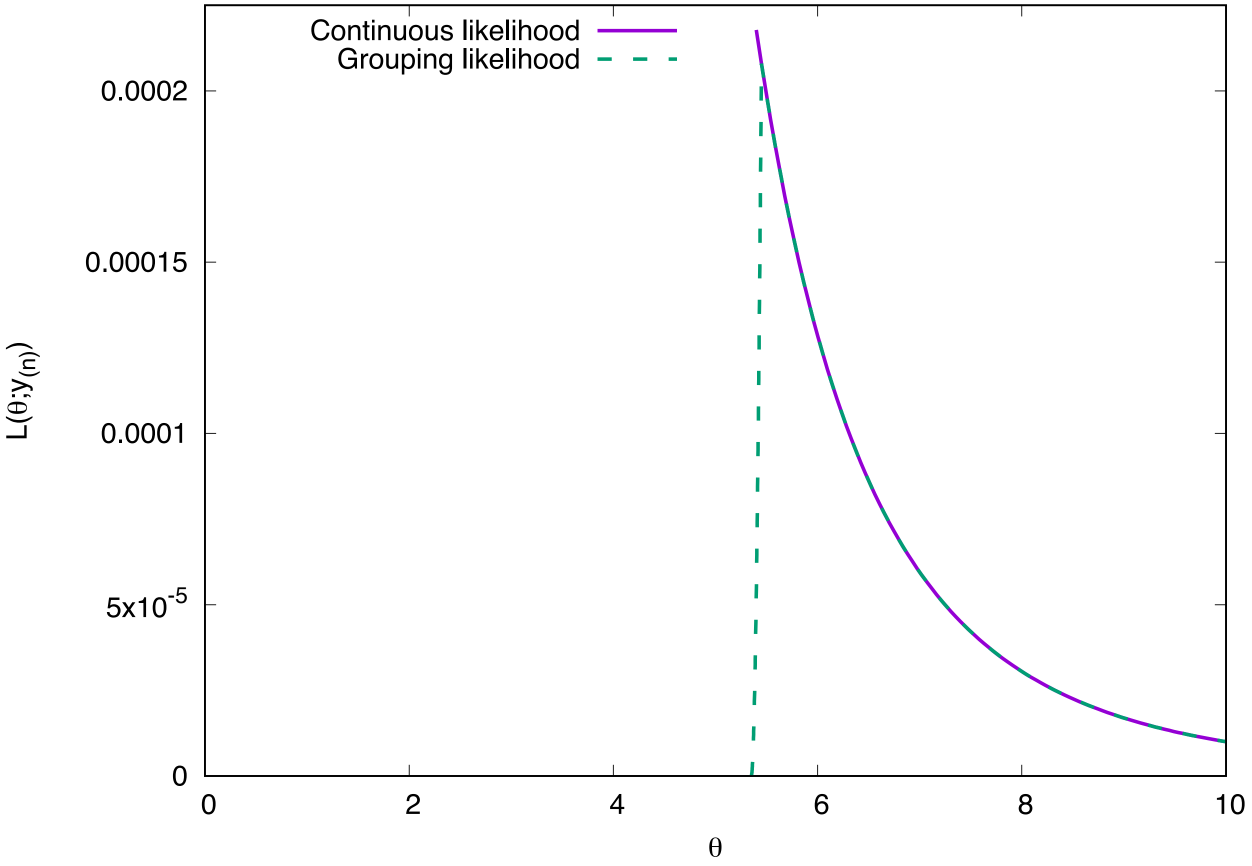


Figure 1: The continuous likelihood for the parameter, θ, of the uniform distribution, as given in Eq. (48), based on $n = 5$ observations with maximum observed value $x_{(n)} = 5.4$ (solid purple line). Also shown is the likelihood including grouping error, as given in Eq. (49), assuming that results are rounded to one decimal place, $h = 0.1$, and there are $a = 2$ observations equal to 5.4 (dashed green line).

2.2 Sufficient statistics

If a parametric form is assumed for the distribution of X, then there may exist a lower dimensional function of the vector of observations $\mathbf{x}$ that contains the same information on the value of $\vec{\theta}$ as vector $\mathbf{x}$. Such a function is called a **sufficient statistic**.

2.3 Definition

Suppose a random vector $\mathbf{X}$ has distribution function in a parametric family $\{P(\mathbf{x}|\,\theta);\theta \in \Theta\}$ and realized value $\mathbf{x}$. A statistic (recall this just means a function of observed data only) is said to be **sufficient** for $\vec{\theta}$ if the distribution of $\mathbf{X}$ given S does not depend on $\vec{\theta}$, i.e. $p_{\mathbf{X}|S}(\mathbf{X}|s,\vec{\theta})$ does not depend on $\vec{\theta}$. Note that

(i) if S is sufficient for $\vec{\theta}$, so is any one-to-one function of S.

(ii) $\mathbf{X}$ is trivially sufficient.

Examples

- Bernoulli trials : $X_1, \ \ldots \ , X_n$ take values 0 or 1 independently with probabilities $1-p$ and p; n is fixed.

$$p_{\mathbf{X}}(\mathbf{x}|\,p) = \prod_{j=1}^{n} p^{x_j}(1-p)^{1-x_j} = p^{\sum x_j}(1-p)^{n-\sum x_j} \tag{50}$$

If $S = X_1 + \ \cdots \ + X_n$, then S has the Binomial p.d.f.

$$p_S(s|\,p) = \left(\begin{array}{c} n \\ s \end{array}\right) p^s(1-p)^{n-s} \quad (s = 0, 1, \ \ldots \ , n)$$

and the p.d.f. of $\mathbf{X}$ given S is

$$\begin{aligned} p_{\mathbf{X}|s}(\mathbf{x}|s) &= \frac{\mathbb{P}(X_1 = x_1, \ \ldots \ , X_n = x_n, X_1 + \ \cdots \ + X_n = s|\,\theta)}{\mathbb{P}(X_1 + \ \cdots \ + X_n = s)} \\ &= \begin{cases} \frac{p_{\mathbf{X}}(\mathbf{x}|\,p)}{p_S(s|\,p)} & (\sum x_j = s) \\ 0 & (\sum x_j \neq s) \end{cases} \\ &= \begin{cases} \left(\begin{array}{c} n \\ s \end{array}\right)^{-1} & (\sum x_j = s) \\ 0 & (\sum x_j \neq s) \end{cases} \end{aligned}$$

This does not depend on p, so S is sufficient for p.

For example, in the case when $n = 3$ the conditional p.d.f of $\mathbf{x} = (x_1, x_2, x_3)$ given $s = \sum x_i$ is as follows:

Sample	$s=\sum x_i$			
(y_1, y_2, y_3)	0	1	2	3
(0 0 0)	1	0	0	0
(1 0 0)	0	$\frac{1}{3}$	0	0
(0 1 0)	0	$\frac{1}{3}$	0	0
(0 0 1)	0	$\frac{1}{3}$	0	0
(1 1 0)	0	0	$\frac{1}{3}$	0
(1 0 1)	0	0	$\frac{1}{3}$	0
(0 1 1)	0	0	$\frac{1}{3}$	0
(1 1 1)	0	0	0	1

- Pois(λ) , $S = X_1 + \ \cdots\ + X_n$ has distribution Pois($n\lambda$) and p.d.f.

  $$p_S(s|\,\lambda) = \frac{e^{-n\lambda}(n\lambda)^s}{s!},$$

  so the distribution of $\mathbf{X}$ given s has p.d.f.

  $$p_{X|s}(X|s) = \begin{cases} \frac{p_{\mathbf{X}}(\mathbf{x}|\,\lambda)}{p_S(s|\,\lambda)} = \frac{e^{-n\lambda}\lambda^{\sum x_j}(\prod_j x_j!)^{-1}}{\frac{e^{-n\lambda}(n\lambda)^s}{s!}} = \frac{n^{-s}s!}{\prod_j x_j!} & (\sum x_j = s) \\ 0 & (\sum x_j \neq s) \end{cases},$$

  which does not depend on λ (it is a multinomial distribution), so S is sufficient for λ.

Interpretation of sufficiency: If S is sufficient for $\vec{\theta}$, we can argue that $\mathbf{x}$ contains no information on $\vec{\theta}$ beyond what is contained in the value s of S, i.e. all the information in $\mathbf{X}$ about $\vec{\theta}$ is contained in s. This suggests that inferences about the value of $\vec{\theta}$ should be based on the value of s. The rest of the information in $\mathbf{y}$ is still relevant to testing the correctness of the assumed parametric family, e.g., by a residual analysis. Sufficiency leads to replacing $\mathbf{x}$ by s and hence to a reduction in the data, so there is an advantage in using statistical models and designs which lead to sufficient statistics of low dimensionality.

2.4 Recognizing sufficient statistics: Neyman Factorization Theorem

Theorem 2.1. *(Neyman Factorization Theorem). Let* $\mathbf{X} = (X_1, \ldots, X_n) \sim p(\mathbf{x}|\,\vec{\theta})$*. Then, statistic* $s = s(X_1, \ldots, X_n)$ *is sufficient for* θ *iff there exist functions* h *of* $\mathbf{x}$ *and* g *of* $(s, \vec{\theta})$ *such that*

$$p(\mathbf{x} \mid \vec{\theta}) = L(\vec{\theta}; \mathbf{x}) = g(s(\mathbf{x}), \vec{\theta})h(\mathbf{x}) \quad \forall \vec{\theta} \in \Theta,\ \mathbf{x} \in \mathcal{X} \tag{51}$$

Proof. Proof (discrete case only).

If s is sufficient, then the conditional p.d.f. $p_{\mathbf{X}|S}(\mathbf{x}|s)$ does not depend on $\vec{\theta}$ and we can take $h(\mathbf{x})$ to be $p_{\mathbf{X}|S}(\mathbf{x}|s)$ and $g(s;\theta)$ to be $f_S(s|\,\theta)$. Then

$$\begin{aligned} L(\vec{\theta}; \mathbf{x}) = p_{\mathbf{X}}(\mathbf{x}|\,\vec{\theta}) &= \mathbb{P}(\mathbf{X} = \mathbf{x}|\,\vec{\theta}) \\ &= \mathbb{P}(\mathbf{X} = \mathbf{x} \,\&\, S = s(\mathbf{x}) \mid \vec{\theta}) \\ &= \mathbb{P}(\mathbf{X} = \mathbf{x}|S = s(\mathbf{x}), \vec{\theta})\, \mathbb{P}(S = s(\mathbf{x})|\,\vec{\theta}) \\ &= \mathbb{P}(\mathbf{X} = \mathbf{x}|S = s(\mathbf{x}))\, \mathbb{P}(S = s(\mathbf{x})|\,\vec{\theta}) \ \text{[since } S \text{ is sufficient]} \\ &= h(\mathbf{x})g(s(\mathbf{x}), \vec{\theta}). \end{aligned}$$

Conversely, if (51) holds, then for any given s there is a subset A_s of $\mathcal{X}$ in which $s(\mathbf{x}) = s$; for $\mathbf{x}$ in A_s

$$\mathbb{P}(\mathbf{X} = \mathbf{x} | S = s, \vec{\theta}) = \frac{f_{\mathbf{X}}(\mathbf{x}| \vec{\theta})}{\sum_{\mathbf{z} \in A_s} f_{\mathbf{X}}(\mathbf{z}| \vec{\theta})} = \frac{h(\mathbf{x})}{\sum_{\mathbf{z} \in A_s} h(\mathbf{z})},$$

while for $\mathbf{x} \notin A_s$ $\mathbb{P}(\mathbf{X} = \mathbf{x} | S = s, \vec{\theta}) = 0$. Thus the conditional distribution does not depend on $\vec{\theta}$, i.e. S is sufficient for $\vec{\theta}$.

□

Note: the statistic $s(\mathbf{x})$ divides the sample space $\mathcal{X}$ into equivalence classes A_s (one for each value of s). This partitioning of $\mathcal{X}$ is unchanged if s is replaced by any one-to-one function of s.

Examples

- Bernoulli trials
$$L(p; \mathbf{x}) = p^{\sum x_j} (1-p)^{n - \sum x_j},$$
so if $s(\mathbf{x}) = \sum x_j$, we could take $h(\mathbf{x}) = 1$, $g(s, p) = p^s (1-p)^{n-s}$
[or, alternatively, we could take $h(\mathbf{x}) = \begin{pmatrix} n \\ s \end{pmatrix}^{-1}$, $g(s,p) = \begin{pmatrix} n \\ s \end{pmatrix} p^s (1-p)^{n-s}$].

- Pois(λ), with $s = \sum x_i$ we have the factorization
$$L(\lambda; \mathbf{x}) = (\prod x_j!)^{-1} \cdot e^{-n\lambda} \lambda^s$$

- The Gamma distribution $\Gamma(\alpha, \lambda)$
$$p_{\mathbf{X}}(\mathbf{x}| \alpha, \lambda) = \prod_{j=1}^{n} \left[\frac{\lambda^\alpha x_j^{\alpha-1} e^{-\lambda x_j}}{\Gamma(\alpha)} \right] = \frac{\lambda^{n\alpha} (\prod_j x_j)^{\alpha-1} e^{-\lambda \sum x_j}}{\{\Gamma(\alpha)\}^n} = 1 \cdot \frac{\lambda^{n\alpha} (s_2)^{\alpha-1} e^{-\lambda s_1}}{\{\Gamma(\alpha)\}^n}$$
Therefore, $(s_1, s_2) = (\sum x_j, \prod x_j)$ is sufficient for (α, λ).

- In a gravitational wave context, reduced order models are used to form a basis for the space of waveforms. Given a set $\{h_i(t)\}$ of basis functions that describe a waveform model, the set $\{(\mathbf{d}|\mathbf{h}_i)\}$ of overlaps of the basis functions with the data are sufficient statistics for deducing the waveform parameters.

2.5 Minimal sufficiency

(Non-trivial) sufficiency leads to a reduction in the data; sufficient statistics achieving the greatest reduction are called **minimal sufficient**, i.e. a minimal sufficient statistic is a function of all other sufficient statistics.

While such statistics are usually obvious, a general method for finding them is implied from the following lemma.

Lemma 2.1. *Consider the following partition of the sample space of* $\mathbf{X} = (X_1, \ldots, X_n) \in \mathcal{X}^n$*:* $\mathbf{x}, \mathbf{y} \in \mathcal{X}^n$ *belong to the same class of the partition if and only if* $L(\vec{\theta}; \mathbf{x}) / L(\vec{\theta}; \mathbf{y})$ *does not depend on* $\vec{\theta}$*.*

Then, any statistic defining this partition is minimal sufficient.

Example

- Weibull distribution: $\{X_1, \ \dots \ , X_n\}$ IID from Weibull with pdf

  $$p(y|\,\alpha, \lambda) = \alpha\lambda^\alpha x^{\alpha-1}\exp[-(\lambda x)^\alpha] \quad (x > 0; \alpha, \lambda > 0)$$

  Then

  $$L(\alpha, \lambda; \mathbf{x}) = \alpha^n\lambda^{n\alpha}(\prod_{j=1}^{n} x_j)^{\alpha-1}\exp(-\lambda^\alpha \sum x_j^\alpha)$$

  For $L(\alpha, \lambda; \mathbf{z})/L(\alpha, \lambda; \mathbf{x})$ not to depend on α, λ, the z_j must be some permutation of the x_j, but no other reduction in the data retains sufficiency, i.e. the order statistics $x_{(1)} \leq \ \dots \ \leq x_{(n)}$ are minimal sufficient.

2.6 Exponential families of distributions

A family of distributions indexed by a multivariate parameter $\vec{\theta} \in \Theta \subset \mathbb{R}^p$, is an **exponential family** iff for some real-valued functions $\{A_j; j = 1\dots, K\}, \{B_j; j = 1\dots, K\}, C, D$ the pdf has the form

$$p(x|\,\theta) = \exp\left\{\sum_{j=1}^{K} A_j(x)B_j(\vec{\theta}) + C(\vec{\theta}) + D(x)\right\} \ \forall x, \vec{\theta} \tag{52}$$

Given observations $\{x_1, \dots, x_n\}$, the set of K statistics $\{\sum_{j=1}^n A_i(x_j) : 1 \leq i \leq K\}$ are sufficient for $\vec{\theta}$ and they are called the natural statistics of the exponential family

In fact, for a K-dimensional parameter $\vec{\theta}$, the minimal sufficient statistic vector is also K-dimensional only for the distributions from the exponential family (under certain regularity conditions, which are the same as those that apply for the validity of the Cramer-Rao inequality described below).

Example. $N(\mu, \sigma^2)$:

$$p(x|\,\mu, \sigma) = \exp\left\{\mu\sigma^{-2}x - \frac{1}{2}\sigma^{-2}x^2 - \left(\frac{1}{2}\mu^2\sigma^{-2} + \ln\sigma + \frac{1}{2}\ln(2\pi)\right)\right\},$$

and $B_1(\mu, \sigma) = \mu\sigma^{-2}$, $B_2(\mu, \sigma) = -\frac{1}{2}\sigma^{-2}$, $A_1(x) = x$, $A_2(x) = x^2$. The vector $S = (\sum_i x_i, \sum_i x_i^2)$ based on sample $(x_1, \dots, x_n)$ is sufficient for $\vec{\theta} = (\mu, \sigma)$.

2.7 Estimators

Recall that an estimator is a statistic (i.e., a function of data only) that is used to obtain an estimate of one or more parameters of the underlying distribution. Often we consider *point estimators* which are single valued functions $\widehat{\theta}(X_1, \dots, X_n)$ of $X_1, \dots, X_n$.

Examples of point estimators:

1. if $\theta = \mathbb{E}(X)$, we can take $\widehat{\theta}$ to be mean, median, mode of the empirical distribution;

2. moment estimators, including the **sample mean**

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

   and the **sample variance**

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 .$$

3. MLE - maximum likelihood estimator, which minimizes the *score*.

Typically there will be several possible estimators of a parameter θ. To choose between estimators we will define various desirable properties: *unbiasedness*, *consistency* and *efficiency*. *Admissibility* and *sufficiency* are also desirable properties but we won't discuss these here. Sufficiency of an estimator is closely related to sufficiency of a statistic. Robustness and ease of computation are not considered in this course, but may be important in practical applications.

2.7.1 Unbiasedness

Definition 2.1. *$\widehat{\theta}$ (r.v.) is an unbiased estimator of θ iff*

$$\mathbb{E}(\widehat{\theta}) = \theta.$$

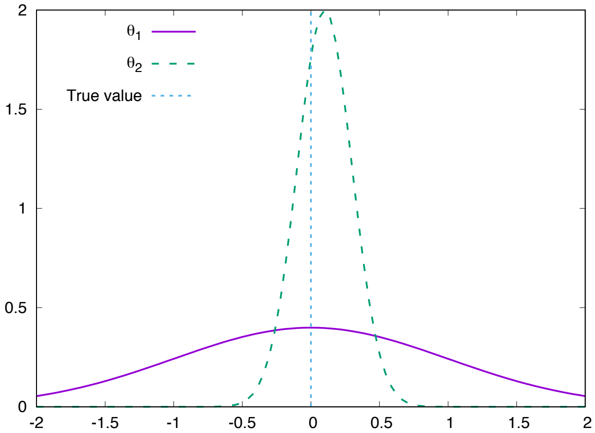
If $\mathbb{E}(\widehat{\theta}) \neq \theta$ then $\widehat{\theta}$ is a biased estimator and we define the bias function of $\widehat{\theta}$ as

$$\text{bias}(\widehat{\theta}) = \mathbb{E}(\widehat{\theta}) - \theta.$$

As an example, suppose θ is a population mean, then the sample mean $\bar{X}$ is unbiased. Also, X_1 (first observation in sample) is unbiased, and if the distribution is symmetric so is the sample median.

There are often several unbiased estimators to choose from, but which is best?

Unbiasedness is not necessarily required for all estimation problems, e.g.,



$\widehat{\theta}_1$ (with wide density) and
$\widehat{\theta}_2$ (with narrow density)
are estimators of θ;
$\widehat{\theta}_1$ is unbiased;
$\widehat{\theta}_2$ is biased;
but $\widehat{\theta}_2$ may be preferred because it is
less likely to be a long way from θ.

Biased estimators may be preferred to unbiased estimators in some circumstances. A good property is asymptotic unbiasedness.

Definition 2.2. *$\widehat{\theta}$ (r.v.) is asymptotically unbiased estimator of θ iff*

$$\mathbb{E}(\widehat{\theta}) \to \theta \quad as \quad n \to \infty.$$

2.7.2 Consistency

As sample size is increased the sampling pdf of any reasonable estimator should become more closely concentrated about θ.

Definition 2.3. *$\widehat{\theta}$ is a (weakly) consistent estimator for θ if*

$$\mathbb{P}(\mid \widehat{\theta} - \theta \mid > \epsilon) \to 0 \qquad \text{as } n \to \infty$$

for any $\epsilon > 0$.

For a particular problem, it may be difficult to verify consistency from this definition, however, a sufficient (not necessary) condition for consistency is given in the lemma below.

Lemma 2.2. *If* var $(\widehat{\theta}) \to 0$ *and* bias$(\widehat{\theta}) \to 0$ *as $n \to \infty$, then $\widehat{\theta}$ is (weakly) consistent.*

Definition 2.4. *The mean square error of an estimator $\widehat{\theta}$ is defined as*

$$\text{MSE}(\widehat{\theta}) \;\; = \;\; \mathbb{E}[(\widehat{\theta} - \theta)^2] = \text{var}(\widehat{\theta}) + [\text{bias}(\widehat{\theta})]^2.$$

Mean squared error consists of two terms: variance of $\widehat{\theta}$ and its squared bias.

The *Markov inequality* states that, for a non-negative random variable X and $a > 0$

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

which can be proved straightforwardly

$$\mathbb{E}(X) = \int_0^\infty xp(x)\text{d}x = \int_0^a xp(x)\text{d}x + \int_a^\infty xp(x)\text{d}x \geq \int_a^\infty xp(x)\text{d}x \geq a\int_a^\infty p(x)\text{d}x = a\mathbb{P}(X \geq a).$$

Setting $X = (\hat{\theta} - \theta)^2$ and $a = \epsilon^2$ we find

$$\mathbb{P}[\mid \widehat{\theta} - \theta \mid > \epsilon] \;\; \leq \;\; \frac{1}{\epsilon^2}\mathbb{E}(\widehat{\theta} - \theta)^2.$$

The term on the right had side is the mean square error. If both bias and variance tend to zero asymptotically, the mean square error tends to zero and therefore the left hand side must tend to zero. Hence we have proven Lemma 2.2.

Examples

1. Estimation of the mean of a normal distribution: using the sample mean $\bar{X}$ or median or just the value of X_1 (first observation in sample) are all unbiased estimators and have variances $\frac{\sigma^2}{n}$, $\alpha\frac{\sigma^2}{n}$ (α is a constant > 1) and σ^2. Therefore the first two are consistent. However, it is evident that X_1 is not consistent as its distribution does not change with sample size.

2. The Cauchy distribution with scale 1 and pdf $p(x|\,\theta) = \pi^{-1}[1+(x-x_0)^2]^{-1}$. In this case, the sample mean $\bar{X}$ has the same distribution as any single X_i, thus $\mathbb{P}[\mid \bar{X} - x_0 \mid > \epsilon]$ is the same for any n. This does not tend to zero as $n \to \infty$, and so $\bar{X}$ is not (weakly) consistent. (However, the sample median is a consistent estimator of x_0.)

2.8 Efficiency

Definition 2.5. *The **efficiency** of an unbiased estimator $(\widehat{\theta})$ is the ratio of the minimum possible variance to var$(\widehat{\theta})$.*

Definition 2.6. *An unbiased estimator with efficiency equal to 1 is called **efficient** or a **minimum variance unbiased estimator (MVUE)**.*

We can also define asymptotic efficiency of an (asymptotically) unbiased estimator $(\widehat{\theta})$ is the limit of the ratio of the minimum possible variance to var$(\widehat{\theta})$ as sample size $n \to \infty$.

Definition 2.7. *An estimator with asymptotic efficiency equal to 1 is called **asymptotically efficient**.*

We can compare the efficiency of two estimators in the following way.

Definition 2.8. *The **(asymptotic) relative efficiency** of two unbiased estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ is the reciprocal of the ratio of their variances, as sample size $\to \infty$: $\lim_{n\to\infty} \frac{Var(\widehat{\theta}_1)}{Var(\widehat{\theta}_2)}$.*

The definition of asymptotic relative efficiency can also be extended to asymptotically unbiased estimators. These definitions are all fine, but they rely on knowing what the smallest possible variance is. Under certain assumptions we can obtain this from the Cramér-Rao inequality.

2.8.1 Cramér-Rao lower bound (inequality)

The theorem below (Cramér-Rao inequality) provides a lower bound on the variance of an estimator. When this lower bound is attainable for unbiased estimators, it can be used in the definition of efficiency.

Regularity conditions for the Cramér-Rao inequality.

1. $\forall\theta_1,\, \theta_2 \in \Theta$ such that $\theta_1 \neq \theta_2$, $p(x \mid \theta_1) \neq p(x \mid \theta_2)$ [identifiability].
2. $\forall\theta \in \Theta$, $p(x \mid \theta)$ have common support.
3. Θ is an open set.
4. $\exists\partial p(x \mid \theta)/\partial\theta$.
5. $\mathbb{E}\left(\partial \log p(\mathbf{X}|\theta)/\partial\theta\right)^2 < \infty$.

Here $I(\theta) = \mathbb{E}\left(\frac{\partial \log f(\mathbf{X}|\theta)}{\partial\theta}\right)^2$ is the Fisher information matrix.

Theorem 2.2. *(Cramér-Rao inequality) Let $X_1, \ldots, X_n$ denote a random sample from $p(x|\,\theta)$, and suppose that $\widehat{\theta}$ is an estimator for θ. Then, subject to the above regularity conditions,*

$$\mathrm{var}(\widehat{\theta}) \geq \frac{\left(1 + \frac{\partial b}{\partial\theta}\right)^2}{I_\theta},$$

where

$$b(\theta) = \mathrm{bias}(\widehat{\theta}) \quad \text{and} \quad I_\theta = \mathbb{E}\left[\left(\frac{\partial\ell}{\partial\theta}\right)^2\right].$$

Comments

1. For unbiased $\widehat{\theta}$, the lower bound simplifies to $\mathrm{var}(\widehat{\theta}) \geq I_\theta^{-1}$.
2. I_θ is called Fisher's information about θ contained in the observations.
3. Regularity conditions are needed to change the order of differentiation and integration in the proof given below.
4. The result can be extended to estimators of functions of θ.

Proof of Theorem 2.2.*

$$\begin{aligned}\mathbb{E}[\widehat{\theta}] &= \int \ldots \int \widehat{\theta}(x_1, \ldots, x_n) \left\{\prod_{i=1}^{n} p(x_i|\,\theta)\right\} d\mathbf{x} \\ &= \int \ldots \int \widehat{\theta}(x_1, x_2, \ldots, x_n) L(\theta; \mathbf{x}) d\mathbf{x}\end{aligned}$$

$\int \ldots \int$ is a multiple integral with respect to $\mathbf{x}$=$(x_1, x_2, \ldots, x_n)$.

From the definition of bias we have

$$\theta + b = \mathbb{E}(\widehat{\theta}) = \int \ldots \int \widehat{\theta} L(\theta; \mathbf{x}) d\mathbf{x}.$$

Differentiating both sides with respect to θ gives (using regularity conditions)

$$1+\frac{\partial b}{\partial\theta}=\int\ldots\int\widehat{\theta}\frac{\partial L}{\partial\theta}d\mathbf{x}$$

since $\widehat{\theta}$ does not depend on θ. Since $l=\ln(L)$ we have

$$\frac{\partial l}{\partial\theta}=\frac{\partial\ln(L)}{\partial\theta}=\frac{1}{L}\frac{\partial L}{\partial\theta},\quad\text{and thus}\quad\frac{\partial L}{\partial\theta}=L\frac{\partial l}{\partial\theta}.$$

Thus

$$1+\frac{\partial b}{\partial\theta}=\int\ldots\int\widehat{\theta}\frac{\partial l}{\partial\theta}Ld\mathbf{x}=\mathbb{E}\left(\widehat{\theta}\frac{\partial l}{\partial\theta}\right).$$

Now use the result that for any two r.v.s $U$and V,

$$\{\mathrm{cov}(U,V)\}^2\leq\mathrm{var}(U)\mathrm{var}(V)$$

and let

$$U=\widehat{\theta},\text{ and }V=\partial l/\partial\theta.$$

Then

$$\begin{aligned}\mathbb{E}[V]&=\int\ldots\int\frac{\partial l}{\partial\theta}Ld\mathbf{x}=\int\ldots\int\frac{\partial L}{\partial\theta}d\mathbf{x}\\&=\frac{\partial}{\partial\theta}\left(\int\ldots\int L\ d\mathbf{x}\right)\quad\text{(using regularity conditions)}\\&=\frac{\partial}{\partial\theta}(1)=0.\end{aligned}$$

Hence

$$\mathrm{cov}(U,V)=\mathbb{E}(UV)=1+\frac{\partial b}{\partial\theta}.$$

Similarly

$$\mathrm{var}(V)=\mathbb{E}(V^2)=\mathbb{E}\left[\left(\frac{\partial l}{\partial\theta}\right)^2\right]=I_\theta\quad\text{(by definition of }I_\theta)$$

and since $\mathrm{var}(U)=\mathrm{var}(\widehat{\theta})$ we obtain the Cramér-Rao lower bound as

$$\mathrm{var}(\widehat{\theta})\geq\frac{\{\mathrm{cov}(U,V)\}^2}{\mathrm{var}(V)}=\frac{\left(1+\frac{\partial b}{\partial\theta}\right)^2}{I_\theta}.$$

□

The Cramér-Rao lower bound will only be useful if it is attainable or at least nearly attainable.

Lemma 2.3. *The Cramér-Rao lower bound is attainable iff there exists a function $f(x)$ of x only, and functions $a(\theta)$, $c(\theta)$ of θ only such that*

$$\frac{\partial l}{\partial \theta} = \frac{(f(x) - a(\theta))}{c(\theta)},$$

in which case $\widehat{\theta} = f(x)$ attains it. The expectation value $\mathbb{E}_\theta \hat{\theta} = a(\theta)$ and $\mathrm{d}a/\mathrm{d}\theta = c(\theta) I_\theta$.

In the derivation of the Cramér-Rao bound, it is clear that equality will be attained if and only if $\mathrm{cov}(U,V)^2$=$\mathrm{var}(U)\mathrm{var}(V)$, which holds if and only if $U = a(\theta) + c(\theta)V$. This lemma and the following corollary follow directly from this.

Corollary 2.1. *There is an unbiased estimator that attains the Cramér-Rao lower bound iff there exists a function $g(x)$ of x only such that*

$$\frac{\partial l}{\partial \theta} = I_\theta (g(x) - \theta),$$

in which case the unbiased estimator $\widehat{\theta} = g(x)$ attains it.

Lemma 2.4. *Under the same regularity conditions as for the Cramér-Rao lower bound*

$$I_\theta = -\mathbb{E}\left[\frac{\partial^2 l}{\partial \theta^2}\right]$$

This result follows from integration by parts, and dropping a boundary term by assuming that the probability density tends to zero asymptotically.

Example
$X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$, σ^2 known.
Likelihood for μ

$$L(\mu; \mathbf{x}) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\}$$

log likelihood for μ

$$l = \log L = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Thus we have

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu), \qquad \frac{\partial^2 l}{\partial \mu^2} = -\frac{n}{\sigma^2},$$

and

$$I_\theta = \mathbb{E}\left[-\frac{\partial^2 l}{\partial \mu^2}\right] = \frac{n}{\sigma^2}.$$

The lower bound for unbiased estimators is $I_\theta^{-1} = \frac{\sigma^2}{n}$. However,

$$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n},$$

so $\bar{X}$ attains its lower bound. No other unbiased estimator can have smaller variance than $\bar{X}$. Therefore $\bar{X}$ is MVUE.

Alternatively, we can use Lemma 2.3, and

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum (X_i - \mu) = \frac{n}{\sigma^2}(\bar{X} - \mu)$$

Therefore the bound is attainable.

Regularity conditions are essential to be able to use the lower bound. Consider the uniform distribution case $X_1, X_2, \ldots, X_n \sim U[0, \theta]$

$$L(\theta; \mathbf{x}) = \begin{cases} \frac{1}{\theta^n} & 0 \leq x_{(1)} \leq x_{(2)} \leq \ldots, \leq x_{(n)} \leq \theta \\ 0 & \text{elsewhere} \end{cases}$$

In the range where L is differentiable $l = -n \log \theta$

$$\frac{\partial l}{\partial \theta} = -\frac{n}{\theta} \quad \text{and} \quad \frac{\partial^2 l}{\partial \theta^2} = \frac{n}{\theta^2}.$$

Thus

$$I_\theta = \mathbb{E}\left[\left(\frac{\partial l}{\partial \theta}\right)^2\right] = \frac{n^2}{\theta^2}$$

but

$$\mathbb{E}\left[-\frac{\partial^2 l}{\partial \theta^2}\right] = \frac{-n}{\theta^2}.$$

Therefore the lower bound should be $\frac{\theta^2}{n^2}$, but

$$\text{var}\left[\frac{n+1}{n}X_{(n)}\right] = \frac{\theta^2}{n(n+2)} < I_\theta^{-1}.$$

The lower bound is violated because the regularity conditions don't hold. In particular the second condition is violated, since the support of the distribution depends on θ.

The derivation and examples above were all for a one dimensional parameter. The corresponding result for the multiple parameter case is

$$\text{cov}(t_i, t_j) \geq \frac{\partial m_i}{\partial \theta_k}\left[\mathbf{I}_\theta\right]^{-1}_{kl}\frac{\partial m_j}{\partial \theta_l}, \qquad [\mathbf{I}_\theta]_{ij} = \mathbb{E}\left[\frac{\partial l}{\partial \theta_i}\frac{\partial l}{\partial \theta_j}\right],$$

where $\mathbf{t}$ is the realised value of some multi-dimensional statistic $\mathbf{T}$ and $\mathbf{m} = \vec{\theta} + \mathbf{b} = \mathbb{E}(\mathbf{T})$.

2.9 Rao-Blackwell Theorem*

The Rao-Blackwell theorem gives a method of improving an unbiased estimator, and involves conditioning on a sufficient statistic.

Theorem 2.3. *(Rao-Blackwell theorem). Let X_1, $X_2, \ldots, X_n$ be a random sample of observations from a distribution with pdf $p(x|\,\theta)$. Suppose that S is a sufficient statistic for θ and that $\widehat{\theta}$ is any unbiased estimator for θ. Define $\widehat{\theta}_S = \mathbb{E}[\widehat{\theta} \mid S]$. Then*

(a) $\widehat{\theta}_S$ *is a function of S only;*

(b) $\mathbb{E}[\widehat{\theta}_S] = \theta$*;*

(c) $\text{var}\,\widehat{\theta}_S \leq \text{var}\,\widehat{\theta}$*.*

2.10 Maximum likelihood estimators

Definition 2.9. *The maximum likelihood estimator (MLE) is defined by* $\widehat{\theta} = \arg\max_{\theta\in\Theta} L(\theta;\mathbf{x}) = \arg\max_{\theta\in\Theta} \ell(\theta;\mathbf{x})$.

If $\exists \partial\ell/\partial\theta_j$ and Θ is open, then the MLE $\widehat{\theta}$ satisfies $\partial\ell/\partial\theta_j(\widehat{\theta}) = 0$, $j = 1,\dots,K$, $\theta \in \Theta \subset \mathbb{R}^K$.

The MLE can be biased or unbiased but it is asymptotically unbiased and efficient and it is also consistent. In fact the following lemma holds.

Lemma 2.5. *Let* $X_1,\dots,X_n \sim p(x \mid \theta)$ *IID,* $\theta \in \Theta \subset \mathbb{R}^K$. *Under the regularity conditions of Cramer-Rao inequality, the MLE asymptotically satisfies*

$$\widehat{\theta} \sim N_K(\theta, I_\theta^{-1}) \quad n \to \infty,$$

in particular, $\mathbb{E}(\widehat{\theta}) \to \theta$ *and for* $K = 1$, $Var(\widehat{\theta})/I_\theta^{-1} \to 1$ *as* $n \to \infty$.

If there exists an unbiased efficient estimator this has to be the MLE.

Lemma 2.6. *Suppose there exists an unbiased estimator* $\tilde{\theta}$ *that attains Cramer-Rao lower bound, and suppose that MLE* $\hat{\theta}$ *is the solution of* $\frac{\partial\ell}{\partial\theta} = 0$. *Then,* $\tilde{\theta} = \hat{\theta}$.

Proof. $\tilde{\theta}$ is unbiased and attains Cramer-Rao lower bound, hence, by the corollary to Lemma 2.3, $\frac{\partial\ell}{\partial\theta} = I_\theta(\tilde{\theta} - \theta)$. Then, the only solution of $\frac{\partial\ell}{\partial\theta} = 0$ is $\tilde{\theta}$, that is, $\tilde{\theta} = \hat{\theta}$. □

Thus, (under the regularity conditions of Cramer-Rao inequality) if the Cramer-Rao lower bound is attainable, the MLE attains it, thus in this case the MLE is efficient. If the bound is unattainable, then the MLE is asymptotically efficient.

2.11 Confidence intervals and regions

Point estimators provide single estimated values for parameters, but we usually also need an estimate of the uncertainty in those estimated values. These are characterised by **confidence intervals**. A confidence interval is a random variable since the ends of the interval are typically determined as a function of the observed data. The interval has the property that over many realisations of the same experiment, the intervals constructed randomly by this procedure will contain the true value of the parameter a certain fraction of the time.

Formally a set $S_\alpha(\mathbf{X})$ is a $(1-\alpha)$ **confidence region** for ψ if

$$\mathbb{P}(S_\alpha(\mathbf{X}) \ni \psi; \psi, \lambda) = 1 - \alpha \quad \forall \psi, \lambda.$$

Thus, $S_\alpha(\mathbf{X})$ is a random set of ψ-values which includes the true value with probability $1-\alpha$. If more than one value of α is considered, we usually require

$$S_{\alpha_1}(\mathbf{x}) \supset S_{\alpha_2}(\mathbf{x}) \text{ if } \alpha_1 < \alpha_2. \tag{53}$$

e.g. a 99% region contains the 95% region.

If ψ is a scalar and $S_\alpha(\mathbf{x})$ has the form $\{\psi : t^\alpha \geq \psi\}$ for some statistic t^α, then t^α is a $(1-\alpha)$ **upper confidence limit** for ψ.

If ψ is a scalar and $S_\alpha(\mathbf{x})$ has the form $\{\psi : s^\alpha \leq \psi\}$ for some statistic s^α, then s^α is a α **lower confidence limit** for ψ.

If $S_\alpha(\mathbf{x}) = \{\psi : a_\alpha(\mathbf{x}) \leq \psi \leq b_\alpha(\mathbf{x})\}$, it is a **two-sided confidence interval**.

A two-sided confidence interval is called **equitailed** if $a_\alpha(\mathbf{x})$ is the $\alpha/2$ lower confidence limit and $b_\alpha(\mathbf{x})$ is the $1 - \alpha/2$ upper confidence limit.

A **high density confidence region** is $\{\theta \in \Theta : p(\mathbf{x}|\theta) \geq K_\alpha\}$ where the constant K_α is determined by the condition $\mathbb{P}\{p(\mathbf{X}|\theta) \geq K_\alpha\} = 1 - \alpha$.

Confidence intervals/regions for estimators can be constructed by identifying **pivotal quantities**. A pivotal quantity $U = u(\mathbf{X}, \psi)$ is a scalar function of $\mathbf{X}$ and ψ with the same distribution for all ψ and λ. If u_α is the upper α point of this distribution, then

$$\mathbb{P}(u(\mathbf{X}, \psi) \leq u_\alpha) = 1 - \alpha,$$

so that the set $\{\psi : u(\mathbf{x}, \psi) \leq u_\alpha\}$ defines a $(1 - \alpha)$ confidence region for ψ.

If ψ is a scalar and $u(mathbfx, \psi)$ is monotone in ψ, this yields a one-sided interval. In this case we may also define two-sided intervals by $\{\psi : u_{\alpha_L} \leq u(\mathbf{x}, \psi) \leq u_{\alpha_U}\}$ with $\alpha_U - \alpha_L = 1 - \alpha$.

Examples of pivotal quantities

- $\mathcal{E}(\lambda)$: $2\theta \sum X_j$ which has distribution $\chi^2(2n)$;
- $N(\mu, \sigma^2)$, inference about μ with σ unknown: $\sqrt{n}(\bar{x} - \mu)/s$ which has distribution $t(n-1)$;
- Ratio of two Normal variances: $(s_1^2/\sigma_1^2)/(s_2^2/\sigma_2^2)$ which has distribution $F(n_1 - 1, n_2 - 1)$.

3 Hypothesis testing

Often when we observed data we have some ideas about the random processes that are generating the observations. Having collected data it is natural to test whether the observed data are consistent with those expectations. The idea of hypothesis testing is to say if the data provides sufficient evidence to rule out those assumptions. The emphasis is always placed in favour of the assumptions, rather than the alternative. We require strong evidence that the data are inconsistent with the assumptions before we reject them.

Formally, we suppose that we have data $\mathbf{x} = (x_1, \ldots, x_n)$ and want to examine whether they are consistent with a hypothesis H_0 (the **null hypothesis** or **hypothesis under test**) about the distribution function $F_{\mathbf{X}}$ of $\mathbf{X}$.

A hypothesis is **simple** if it defines $P_{\mathbf{X}}$ completely:

$$H_0 : P_{\mathbf{X}} = P_0$$

otherwise, it is **composite**. If $P_{\mathbf{X}}$ is parametric with more than one parameter, a composite hypothesis might specify the values of some or all of them. (e.g. one regression coefficient)

The distribution of $\mathbf{X}$ under H_0, P_0, is called null distribution.

Examples of hypotheses

- A significant trigger in a gravitational wave detector is due to instrumental fluctuations. This is a composite hypothesis as the distribution of triggers under the noise assumption is not fully specified.

- The numbers of gravitational wave events $x_1, \ldots, x_7$ observed on Monday, ..., Sunday. The null hypothesis is that all days are equally likely, i.e., the joint distribution is Multinomial$(n; \frac{1}{7}, \ldots, \frac{1}{7})$. This is a simple hypothesis.

- The right ascensions $x_1, \ldots, x_n$ angles of observed gravitational wave events. The hyypothesis that the X_j's are independently Uniform on $[0, 2\pi)$ is simple.

  Suppose we want to test that there is clustering around some angle, then we can assume that the distribution is von Mises with pdf

  $$p(x|\,\theta, \lambda) = \frac{1}{2\pi I_0(\lambda)} e^{\lambda \cos(x-\theta)}, \quad x \in \mathcal{X} = [0, 2\pi);\ \lambda \geq 0,\ 0 \leq \theta < 2\pi;$$

  for unknown λ. This is a composite hypothesis.

- The hypothesis that the number of gravitational wave events in each month $X_1, \ldots, X_n$ are independently Poisson(θ) with unknown θ is composite.

3.1 Definitions and basic concepts

1. A sample of n observations is available to make inference about parameter θ.

2. We wish to decide between two hypotheses: H_0, *the null hypothesis*, and H_1, *the alternative hypothesis.*

   H_0 is often *simple* (only one value is specified for θ)

   $$\text{i.e. } H_0 : \theta = \theta_0 \text{ (e.g. } H_0 : \mu = 100,\ H_0 : p = \tfrac{1}{2}).$$

H_1 can be simple: $H_1 : \theta = \theta_1$ but more commonly it is *composite* (more than one value is allowed for θ). The most common alternatives are

$$H_1 : \theta < \theta_0 \text{ or } H_1 : \theta > \theta_0 \text{ — one-sided/one-tailed alternative}$$
$$\text{or } H_1 : \theta \neq \theta_0 \text{ — two-sided/two-tailed alternative.}$$

3. Two possible decisions: *to reject* or *not to reject* H_0 in favour of H_1.

   The decision whether or not to reject H_0 is based on the value of a *test statistic*, which is a function of the observations.

4. Values of the test statistic for which H_0 is not rejected form the *acceptance region*, $\bar{C}$.

   Values of the test statistic for which H_0 is rejected form the *rejection region* (or critical region), C.

   The form of these regions depends on the form of H_1.

5. There are two possible types of error:

   Reject H_0 when H_0 is true — Type I error
   Fail to reject H_0 when H_0 is false — Type II error

   The probability of Type I error, denoted by α, is the **significance level (**or **size)** of the test.

   The probability of Type II error, denoted by β, is only defined uniquely if H_1 is simple. In which case

   $$\eta \; = \; 1 - \beta \text{ is the } \textbf{power} \text{ of the test.}$$

   For composite H_1, $\eta(\theta)$ is the *power function.*

Generally we consider Type-I error (false rejection) to be worse than Type-II (incorrect failure to reject) as usually in the latter case more data will be collected and the test will be re-evaluated. It is therefore usual to specify the **significance level** of the test in order to determine the threshold for rejection, or the quote a **p-value** (see next section) when quoting test results.

We can define a **test function** $\phi(x)$ such that

$$\phi(x) = \begin{cases} 1 & \text{if } t(\mathbf{x}) \in C \\ 0 & \text{if } t(\mathbf{x}) \in \bar{C} \end{cases}$$

and when we observe $\phi(\mathbf{X}) = 1$, we reject H_0. This function has the property that $\alpha = \mathbb{E}_{H_0}(\phi(\mathbf{X}))$ and $\eta = \mathbb{E}_{H_1}(\phi(\mathbf{X}))$, in which the subscript denotes the hypothesis under which the expectation value is to be calculated.

For discrete distributions, the probability that the test statistic lies on the boundary of the critical region, ∂C, may be non-zero. In that case, it is sometimes necessary to use a **randomized test**, for which the test function is

$$\phi(x) = \begin{cases} 1 & \text{if } t(\mathbf{x}) \in C \\ \gamma(\mathbf{x}) & \text{if } t(\mathbf{x}) \in \partial C \\ 0 & \text{if } t(\mathbf{x}) \in \bar{C} \end{cases}$$

for some function $\gamma(\mathbf{x})$ and we reject H_0 based on observed data $\mathbf{x}$ with probability $\phi(\mathbf{x})$.

3.2 Test statistic

Often to construct a test (i.e. the decision whether to reject H_0 or not based on observed data $\mathbf{x}$), a test statistic is used.

Definition 3.1. *A real-valued function $t(\mathbf{x})$ on $\mathcal{X}$ is a test statistic for testing H_0 iff*

(i) values of t are **ordered** *with respect to the evidence for departure from H_0*

(ii) the distribution of $T = t(\mathbf{X})$ under H_0 is known, at least approximately. For composite H_0 the distribution should be (approximately) the same for all simple hypotheses making up H_0.

For any observation $\mathbf{x}$, we measure the consistency of $\mathbf{x}$ with H_0 using the *significance probability* or the *p-value*, e.g. if larger values of t correspond to stronger evidence for departure from H_0, the p-value is defined by

$$p = \mathbb{P}(T \geq t(\mathbf{x})|\, H_0),$$

the probability (under H_0) of seeing the observed value of t or any more extreme value. The smaller the value of p the greater the evidence against H_0.

3.3 Alternative hypothesis

Can be specified or unspecified.

3.3.1 Pure significance tests

In a *pure significance test*, only the null hypothesis H_0 is explicitly specified. The p-value of the observed value under the null distribution is evaluated, and if it is sufficiently small, the null hypothesis would be rejected. Such tests are done if we want to avoid specifying a parametric family of alternative distributions.

There will often be multiple quantities that could be computed under the null hypothesis and we can choose any of them to evaluated the distribution of the test statistic. The best choice can be guided if we have a specific idea of the type of departure from H_0 we are looking for, e.g.,

- Directional data: Might look for a tendency for the observed directions to cluster about a (possibly unknown) direction. But not a specific set of alternatives such as von Mises distributions.

- Pois(θ): if the alternative is not a Poisson distribution, we might test whether variance $\neq$ expectation.

An important class of pure significance tests are *goodness of fit* tests where either the sample distribution function $\hat{P}_X(x) = \frac{1}{n}\sum_{i=1}^{n} I(x \leqslant x_i)$ or the histogram are compared to those of the null distribution.

Examples

- Event frequency on different days: $H_0: \; X_1, \ldots, X_7 \sim \text{Mult}(n; \frac{1}{7}, \ldots, \frac{1}{7})$.

  With no particular alternative we might use Pearson's χ^2 test, comparing

$$X^2 = \sum_{i=1}^{7} \frac{\left(x_i - \frac{n}{7}\right)^2}{\frac{n}{7}} \quad \text{with} \quad \chi^2_6.$$

- Right ascension of GW sources: If alternative to H_0 is clustering about the reference direction (e.g. galactic centre) we could use $\sum \cos x_j$, the projection onto the reference axis of the resultant sum vector $(\sum \cos x_j, \sum \sin x_j)$.

- $\text{Pois}(\theta)$: might use index of dispersion,

$$d = \frac{\sum (x_i - \bar{y})^2}{\bar{y}},$$

  which is approximately χ^2 with $(n-1)$ degrees of freedom under H_0 for $\theta \geq 1$.

  Note that given $\sum X_j = s$, the distribution of $X_1, \ \ldots \ , X_n$ is $\text{Mult}(s, \frac{1}{n}, \ \ldots \ , \frac{1}{n})$ and d is the χ^2 statistic for testing the fit of this distribution.

3.3.2 Specified alternative hypothesis

For a parametrised family of distributions $p(x|\,\theta)$, $\theta \in \Theta$, say $H_0 : \; \theta = \theta_0$, then

$$H_1 : \; \theta \in \Theta_1 \subset \Theta \setminus \{\theta_0\},$$

e.g. $\theta \neq \theta_0$ (two-sided), $\theta > \theta_0$ or $\theta < \theta_0$ (one-sided).

Below we consider two cases: with simple and composite alternative hypotheses (and a simple null hypothesis).

With composite alternative hypotheses, the power of the test becomes the power function defined over $\theta \in \Theta_1$:

$$\eta(\theta) = \mathbb{P}(\text{reject } \; H_0|\,\theta) = \mathbb{P}_\theta(\text{reject } \; H_0).$$

3.4 Critical regions

In § 3.2 we defined for each $\mathbf{x} \in X$ the significance probability

$$p = \mathbb{P}(T \geq t(\mathbf{x})|\,H_0)$$

associated with a test statistic t. A different, but equivalent, approach defines a test using critical regions rather than test statistics. This

(i) facilitates comparison of different tests of H_0 according to their properties under H_1;

(ii) is useful for establishing a connection between tests and confidence regions.

For any α in the interval $(0, 1)$, a subset R_α of X is a **critical region of size** α if

$$\mathbb{P}(\mathbf{X} \in R_\alpha|\,H_0) = \alpha \tag{54}$$

Interpretations of R_α:

(i) points in R_α are regarded as not consistent with H_0 at level α;

(ii) points in R_α are "significant at level α";

(iii) if $\mathbf{x} \in R_\alpha$, then H_0 is "rejected" in a test of size α.

A significance test is defined by a set of critical regions $\{R_\alpha : 0 < \alpha < 1\}$ satisfying

$$R_{\alpha_1} \subset R_{\alpha_2} \quad \text{if } \alpha_1 < \alpha_2. \tag{55}$$

Thus, for example, if data $\mathbf{x}$ are significant at the 1% level, they are also significant at the 5% level.

The **significance probability** (also called p-value) for data $\mathbf{x}$ is then defined as

$$P = \inf(\alpha; \mathbf{x} \in R_\alpha),$$

i.e. the smallest α for which $\mathbf{x}$ is significant at level α.

The definition of a test in §3.2 corresponds to critical regions of the form

$$R^t_\alpha = \{\mathbf{x} : t(\mathbf{x}) \geq t_\alpha\},$$

where t_α is the upper α point of $T = t(\mathbf{X})$ under H_0, since

$$\mathbb{P}(\mathbf{X} \in R^t_\alpha | \, H_0) = \mathbb{P}(t(X) \geq t_\alpha | \, H_0) = \alpha,$$

by the definition of t_α; also if $\alpha_1 < \alpha_2$ then $t_{\alpha_1} > t_{\alpha_2}$ and $R^t_{\alpha_1} \subset R^t_{\alpha_2}$ satisfying (55). Finally,

$$\begin{aligned} P &= \mathbb{P}(t(\mathbf{X}) \geq t(\mathbf{x}) : H_0) \\ &= \inf(\alpha; t(\mathbf{x}) \geq t_\alpha) \\ &= \inf(\alpha; \mathbf{x} \in R^t_\alpha), \end{aligned}$$

the smallest α for which $\mathbf{x}$ is significant at level α.

Example

- X_j independent $N(\mu, \sigma^2)$ (σ known and hence =1 without loss of generality) To test $H_0 : \mu = \mu_0$ vs $\mu > \mu_0$, obvious test statistics are $\bar{Y}$ or $(\bar{Y} - \mu_0)\sqrt{n}$. The significance probability is

$$P = \mathbb{P}\left((\bar{Y} - \mu_0)\sqrt{n} > (\bar{y} - \mu_0)\sqrt{n} | \, H_0\right) = 1 - \Phi((\bar{y} - \mu_0)\sqrt{n}).$$

The corresponding critical regions are $R_\alpha = \{\mathbf{x} : (\bar{y} - \mu_0)\sqrt{n} \geq \Phi^{-1}(1 - \alpha)\}$. Thus

$$\mathbb{P}(\mathbf{X} \in R_\alpha | \, H_0) = \mathbb{P}((\bar{Y} - \mu_0)\sqrt{n} \geq \Phi^{-1}(1 - \alpha)) = \alpha,$$

as required, and if $\alpha_1 < \alpha_2$, then $\Phi^{-1}(1 - \alpha_1) > \Phi^{-1}(1 - \alpha_2)$, so that $R_{\alpha_1} \subset R_{\alpha_2}$. Also

$$\begin{aligned} \inf(\alpha; \mathbf{x} \in R_\alpha) &= \inf(\alpha; (\bar{y} - \mu_0)\sqrt{n} \geq \Phi^{-1}(1 - \alpha)) \\ &= \inf(\alpha; \alpha \geq 1 - \Phi((\bar{y} - \mu_0)\sqrt{n}) \\ &= P. \end{aligned}$$

3.5 Construction of confidence intervals using critical regions

The construction of hypothesis tests leads naturally to the construction of confidence intervals and regions. For any value ψ_0 of ψ, let $R_\alpha(\psi_0)$ be a size-α critical region for testing the null hypothesis $\psi = \psi_0$ against $\psi \neq \psi_0$ (or possibly $\psi < \psi_0$ or $\psi > \psi_0$). For any $\mathbf{x}$ define

$$S_\alpha(\mathbf{x}) = \{\psi_0 : \mathbf{x} \notin R_\alpha(\psi_0)\}.$$

Then $S_\alpha(\mathbf{X})$ is a $(1-\alpha)$ confidence interval for ψ since

$$\mathbb{P}(S_\alpha(\mathbf{X}) \ni \psi_0; \psi_0, \lambda) = \mathbb{P}(\mathbf{X} \notin R_\alpha(\psi_0) : \psi_0, \lambda) = 1-\alpha \quad \forall \psi_0, \lambda$$

[$\bar{R}_\alpha(\psi_0)$ comprises $\mathbf{x}$ values judged consistent with ψ_0 (at level α), so $S_\alpha(\mathbf{x})$ comprises ψ values consistent with $\mathbf{x}$.]

If $\alpha_1 < \alpha_2$, then from (19) $\{\psi_0 : \mathbf{x} \in R_{\alpha_1}(\psi_0)\} \subset \{\psi_0 : \mathbf{x} \in R_{\alpha_2}(\psi_0)\}$, so that (53) holds.

For scalar ψ, critical regions for alternatives $\psi < \psi_0$ lead to upper confidence limits.

Example

- Exp(λ): Find the best size-α critical region for testing $\lambda = \lambda_0$ against $\lambda < \lambda_0$.

  The best size-α critical region for testing $\lambda = \lambda_0$ against $\lambda < \lambda_0$ is $R_\alpha(\lambda_0) = \{\mathbf{x} : \sum x_j > \frac{1}{2}\lambda_0^{-1}\chi^2_{2n}(\alpha)\}$. The corresponding $(1-\alpha)$ confidence region for λ is $\{\lambda_0 : \sum x_j \leq \frac{1}{2}\lambda_0^{-1}\chi^2_{2n}(\alpha)\}$ i.e. $\{\lambda_0 : \lambda_0 \leq \frac{1}{2}(\sum x_j)^{-1}\chi^2_{2n}(\alpha)\}$, so that $\frac{1}{2}(\sum x_j)^{-1}\chi^2_{2n}(\alpha)\}$ is the $(1-\alpha)$ upper confidence limit for λ.

3.6 Examples of hypothesis tests

We give three commonly encountered examples of hypothesis tests.

3.6.1 z-test

Suppose that we observe two independent samples

$$X_1, \ldots, X_n \sim N(\mu_1, \sigma^2), \qquad Y_1, \ldots, Y_m \ N(\mu_2, \sigma^2).$$

We assume additionally that σ^2 is known and we are interested in testing the hypothesis

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{versus} \quad H_1 : \mu_1 - \mu_2 \neq 0.$$

If the null hypothesis is violated we expect that the magnitude of the difference in sample means, $|\bar{X} - \bar{Y}|$, will be large. The statistic

$$Z = \left(\frac{1}{n} + \frac{1}{m}\right)^{-\frac{1}{2}} \frac{(\bar{X} - \bar{Y})}{\sigma}$$

follows a $N(0,1)$ distribution under the null hypothesis so we use a critical region of the form

$$|z| > z_{\frac{\alpha}{2}}$$

to define a test with significance α. Here $z_{\frac{\alpha}{2}}$ denotes the upper $\alpha/2$ point in the Normal distribution, i.e., the point such that

$$\mathbb{P}(X \sim N(0,1) > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}.$$

3.6.2 t-test

We now suppose that we want to test the same hypothesis as in the previous example, but assuming that σ^2 is not known. Once again, we expect the difference in sample means to be large when the null hypothesis is false, but exactly how large now depends on the unknown value of σ^2. If we use the same test statistic, but with the known variance replaced by the estimated value we have

$$T = \left(\frac{1}{n} + \frac{1}{m}\right)^{-\frac{1}{2}} \frac{(\bar{X} - \bar{Y})}{\hat{\sigma}} \qquad \text{where } \hat{\sigma}^2 = \frac{1}{m+n-2}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_i - \bar{Y})^2\right)$$

which follows a t_{m+n-2} distribution under the null hypothesis.

The critical region of a size-α test is to reject H_0 when

$$|t| > t_{\frac{\alpha}{2}},$$

where $z_{\frac{\alpha}{2}}$ denotes the upper $\alpha/2$ point in the t-distribution with $m+n-2$ degrees of freedom.

3.6.3 Analysis of variance: F-test

Suppose we have observations of random variables X_{ij} where $j = 1, \ldots, n_i$ labels different observations of one particular group, and $i = 1, \ldots, k$ labels the different groups. We denote the mean in each group by

$$\bar{X}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$$

and the overall mean by

$$\bar{X}_{\bullet\bullet} = \frac{1}{N}\sum_{ij} X_{ij}, \qquad N = \sum_{i=1}^{k} n_i.$$

We are interested in testing that the means of all the groups are equal. If this is true then we expect that the **between samples sum of squares**

$$SS_b = \sum_i n_i(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2$$

is comparable to the **within samples sum of squares**

$$SS_w = \sum_{ij}(x_{ij} - x_{i\bullet})^2.$$

If the means are different then we expect the former to be larger than the latter. Therefore, we reject the null hypothesis for large values of SS_b/SS_w. The quantity

$$F = \frac{(N-k)SS_b}{(k-1)SS_w}$$

follows an $F_{k-1,N-k}$-distribution under the null hypothesis and so our critical regions are of the form to reject H_0 when

$$F > F_{k-1,N-k}(\alpha)$$

the upper α critical point of the $F_{k-1,N-k}$ distribution.

3.7 Calculating thresholds for tests

For the examples above the test statistics followed known distributions under the null hypothesis and so the critical values can be directly calculated. This is not always possible. In other situations it might be possible to compute the mean, μ, and variance, σ^2, of the test statistic, if not its full distribution. In that case, a Normal approximation can often be used by appealing to the Central Limit Theorem.

Example: $\mathcal{E}(\lambda)$: we saw above that $X = \sum x_j$ can be used for testing $\lambda = \lambda_0$ versus $\lambda < \lambda_0$. While in this case we know the exact distribution of the test statistic, if we did not we can approximate

$$X \sim N\left(\frac{n}{\lambda_0}, \frac{n}{\lambda_0^2}\right)$$

and reject the hypothesis at significance α if

$$\frac{\lambda_0 X - n}{\sqrt{n}} > z_\alpha.$$

The power of the test can be approximated in a similar way, by writing down a Normal approximation to the distribution of the test statistic under the alternative hypothesis.

If the mean and variance cannot be easily calculated, or the form of the test statistic does not lend itself to approximation by the Central Limit Theorem, then usually the best approach is to do a **simulation study**, i.e., generate many realisations of the test statistic under H_0 and determine thresholds numerically. In principle, the power of the test can be evaluated in a similar way although this might not be practical for composite alternative hypotheses.

3.8 Multiple testing

When presented with new data, there is a temptation to keep asking different questions of the same data. When doing this you have to be careful to avoid **multiple testing** (or, in the language of the gravitational wave community **trials factors**). If you keep carrying out independent tests that have a significance of α then you would expect to reject a hypothesis every $1/\alpha$ tests purely by chance. Therefore, if you plan to carry out m independent tests and want the overall significance to be α, the significance levels applied to the individual tests must be lower.

If we carry out m independent tests, each with significance α, then the combined significance is

$$1 - (1-\alpha)^m = \alpha_c.$$

To reach a target significance of the combined tests requires using individual tests with significance $\alpha = 1 - (1-\alpha_c)^{1/m} = 1 - \exp(\log(1-\alpha_c)/m) \approx \alpha_c/m$. The first expression is the *Ŝidák correction*, while the latter correction is referred to as the *Bonferroni correction.*

It is also possible to not divide the total significance evenly between the different individual tests. The *Holm-Bonferroni method* orders the individual test p-values and then tests the i'th (starting from the smallest) at a significance level of $\alpha_c/(m-i+1)$. This approach gives better overall performance.

In practice, multiple tests on the same data will not be independent and so using the corrections based on independence will be conservative and the true significance of any rejection of the null hypothesis will be greater (i.e., the true p-value will be smaller than that estimated in this way). Understanding the dependency of multiple tests is typically highly non-trivial so it is usually best to assess the true p-value of a testing programme using simulations.

Another issue to be cautious of is changing the question based on the data. Changing the question based on what was observed can lead to results appearing significant when they are not, as the following example illustrates.

Example: LIGO/Virgo operate for 8 months from January to August and sees event counts $(1, 0, 0, 0, 0, 1, 1, 4)$. Are the 4 events in the last month unusual? A total of 7 events have been observed in 8 months, so we have a rate of $\sim 7/8$ per month. Assuming that the events are Poisson distributed with this rate, the probability that a given month would have 4 or more events in it is $\sim 1.2\%$, which would be significant at the 5% level usually used for hypothesis tests. But it is not fair to ask "Is four events in August unusual?", since we only decided to look at August in particular when we saw the data. The fair question to ask is "Is four events in one of the months unusual", which means we must multiply by 8 to account for the fact that we have 8 potentially unusual months to choose from. The resulting probability of $\sim 9.8\%$ is much less significant [1]. Note that it is perfectly fine, having made these observations, to ask "Is August unusual in the next observing run?" and specifically target the month that was an outlier in previous data in the next analysis. However, this is less sensitive than doing the test "Is any month unusual?" on all of the data from both observing runs together. Suppose in the next year we also take data from January to August and observe events $(0, 1, 0, 1, 1, 0, 0, 2)$. The probability of observing two or more events in August, given the rate of 5/8 events per month, is 13%, so this would not be considered significant. However, adding the two observing runs together we have $(1, 1, 0, 1, 1, 1, 1, 6)$ and the rate for binned observations is 4/3. The probability of seeing 6 or more events in a Poisson distribution with rate 4/3 is 0.25%, which is significant [2].

3.9 Receiver operator characteristic

As mentioned above, Type-I errors are considered to be more serious than Type-II errors and so tests are quoted by the significance level. However, there may be (infinitely) many tests with the same significance, so how do we choose between them? This is done using the power function. Clearly if one test is more powerful than another for the same significance level then it is better and should be used.

In general, one way to compare different tests is by plotting a **receiver operator characteristic** (ROC) curve. This is a plot of the power versus significance of a test, or equivalently the "detection rate" of deviations in the null hypothesis against the "false alarm rate". For a random test, i.e., we toss a coin and, regardless of the observed data, say that if it is heads we have made a detection, the ROC curve is the diagonal line. Tests that lie above the line are more powerful than random at given significance, and so the further away from the diagonal line the better the test is. ROC curves can be used to compare tests visually, or

[1] Another way to tackle this problem is to say that we expect the distribution of events across the 8 months to be Multinomial with equal probability of 0.125 in each month. The distribution of events in a specific month is Binomial with $n = 7$ and $p = 0.125$ and so the probability that a specific event will have four or more events out of the 7 is $\sim 0.6\%$, but this rises to $\sim 5.0\%$ when we compute the probability that one (unspecified) month has four or more events.

[2] In the multinomial analysis the probabilities are 12% and 0.18% respectively

by computing the area between the curve and the diagonal line. Sometimnes the curves can cross, so one test may be better at one significance level and another at another. The best test then depends on what regime you are operating in.

In the following subsections we will present a number of results that describe how to find tests that have the highest power at a given significance, under various assumptions about the hypotheses and the underlying distributions. As we shall see below, it is not always possible to find a test that is the best everywhere.

3.10 Designing the best test: simple null and alternative hypotheses

Consider null and alternative hypotheses H_0, H_1 corresponding to completely specified p.d.f.'s p_0, p_1 for $\mathbf{X}$. For these hypotheses, comparison between the critical regions of different tests is in terms of

$$\mathbb{P}(\mathbf{X} \in R_\alpha |\, H_1)$$

the **power** of a size-α critical region R_α for alternative H_1. A **best** critical region of size α is one with maximum power.

In terms of p_0, p_1, the power is

$$\begin{aligned}\int_{R_\alpha} p_1(\mathbf{x}) d\mathbf{x} &= \int_{R_\alpha} p_0(\mathbf{x}) r(\mathbf{x}) d\mathbf{x} \quad \left(\text{or } \sum_{R_\alpha} p_0(\mathbf{x}) r(\mathbf{x})\right) \\ &= \mathbb{E}\{r(\mathbf{X}) | \mathbf{X} \in R_\alpha ; H_0\}\end{aligned}$$

where

$$r(\mathbf{x}) = \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} = \frac{L(\theta; \mathrm{H}_1)}{L(\theta; \mathrm{H}_0)},$$

the **likelihood ratio** (LR) for H_1 vs H_0. We can **prove** that the power is maximized when R_α has the form $\{\mathbf{x} : r(\mathbf{x}) \geq k_\alpha\}$ or $\{\mathbf{x} : \frac{L(\theta;\mathrm{H}_1)}{L(\theta;\mathrm{H}_0)} \geq k_\alpha\}$, i.e. when R_α is a LR critical region. Thus we have the Neyman-Pearson lemma.

Theorem 3.1. *(Neyman-Pearson lemma). For any size α, the LR critical region is the best critical region for testing simple hypotheses H_0 vs H_1. (It is also better than any critical region of size $< \alpha$.)*

A LR test is a test whose critical regions are LR critical regions for all α for which such a size-α region exists (all α in the continuous case).

Examples

- Angles: If H_0, H_1 correspond to a Uniform distribution and a von Mises distribution with parameter θ_1, the LR is

$$r(\mathbf{x}) = \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} = \{2\pi I_0(\theta_1)\}^{-n} \frac{e^{\theta_1 \sum_j \cos x_j}}{(2\pi)^{-n}},$$

which is an increasing function of $t(\mathbf{x}) = \sum \cos x_j$. So the LR critical regions have the form $\{\mathbf{x} : \sum \cos x_j > t_\alpha\}$. For any α, t_α is given by $\mathbb{P}(\sum \cos X_j \geq t_\alpha |\, H_0) = \alpha$. From §3.3 $\sum \cos X_j$ is approximately $N(0, \frac{1}{2}n)$ under H_0, so t_α is approximately $\left(\frac{1}{2}n\right)^{1/2} \Phi^{-1}(1-\alpha)$. Note that the critical regions, and hence the test, do not depend on the value of θ_1.

- $\mathcal{E}(\lambda)$: $X_1, \ldots, X_n$ are i.i.d. with d.f. $1-e^{-\lambda y}$ $(y>0)$. H_0 is $\lambda=\lambda_0$; H_1 is $\lambda=\lambda_1<\lambda_0$

$$r(\mathbf{x})=\frac{p_1(\mathbf{x})}{p_0(\mathbf{x})}=\left(\frac{\lambda_1}{\lambda_0}\right)^n \exp\{(\lambda_0-\lambda_1)\sum x_j\},$$

which is increasing in $\sum x_j$. So the test is based on $\sum x_j$ or $2\lambda_0\sum X_j$, which is χ^2_{2n} under H_0, and the critical regions are $\{\mathbf{x} : \sum x_j > \frac{1}{2}\lambda_0^{-1}\chi^2_{2n}(\alpha)\}$, where $\chi^2_{2n}(\alpha)$ is the upper α point of χ^2_{2n}. The power is

$$\begin{aligned}\mathbb{P}(2\lambda_0\sum X_j > \chi^2_\alpha |\, H_1) &= \mathbb{P}\left(2\lambda_1\sum X_j > \frac{\lambda_1}{\lambda_0}\chi^2_{2n}(\alpha)|\, H_1\right)\\ &= Q_{2n}\left(\frac{\lambda_1}{\lambda_0}\chi^2_{2n}(\alpha)\right)\end{aligned}$$

where Q_{2n} is 1− distribution function for χ^2_{2n}.

For comparison, we might base a test on $x_{(1)}$, which has distribution function $1-e^{-n\lambda y}$; size α critical regions are given by $\{\mathbf{x} : x_{(1)} > -(n\lambda_0)^{-1}\ln\alpha\}$, and the power is $\alpha^{\lambda_1/\lambda_0}$, which is $< Q_{2n}\left(\frac{\lambda_1}{\lambda_0}\chi^2_\alpha\right)$ for $n>1$ and $\lambda_1<\lambda_0$, and does not depend on n.

3.11 Designing the best test: simple null and composite alternative hypotheses

Suppose now there is a parametric family $\{p(\mathbf{x}|\,\theta) : \theta\in\Theta_1\}$ of alternative p.d.f.'s for $\mathbf{X}$. The power of a size-α critical region R_α generalizes to the size-α **power function**

$$\begin{aligned} pow(\theta;\alpha) &= \mathbb{P}(\mathbf{X}\in R_\alpha|\,\theta)\\ &= \int_{R_\alpha} p(\mathbf{x}|\,\theta)dy \quad \left(\text{or} \sum_{R_\alpha} p(\mathbf{x}|\,\theta)dy\right) \quad (\theta\in\Theta_1).\end{aligned}$$

A size-α critical region R_α is then **uniformly most powerful size** α (UMP size α) if it has maximum power uniformly over Θ_1. A test is UMP if all its critical regions are UMP. More formally

Definition 3.2. *A* **uniformly most powerful** *or UMP test, $\phi_0(\mathbf{X})$, of size α is a test $t(\mathbf{x})$ for which*

(i) $\mathbb{E}_\theta\phi_0(\mathbf{X})\le\alpha \quad \forall\,\theta\in\Theta_0$;

(ii) given any other test $\phi(\cdot)$ for which $\mathbb{E}_\theta\phi(\mathbf{X})\le\alpha \quad \forall \quad \theta\in\Theta_0$, we have $\mathbb{E}_\theta\phi_0(\mathbf{X})\ge \mathbb{E}_\theta\phi(\mathbf{X}) \quad \forall \quad \theta\in\Theta_1$.

Such tests cannot be found in general, as this requires that the Neyman-Pearson test should be the same for every pair of simple hypotheses. However, for one sided testing problems, i.e., tests of the form $H_0 : \theta\le\theta_0$ against $H_1 : \theta>\theta_0$, there are a wide class of parametric families for which UMP tests exist. These are distributions that have **monotone likelihood ratio** or MLR.

Definition 3.3. *The family of densities $\{p(\mathbf{x}|\theta), \theta \in \Omega_\theta \subseteq \mathbb{R}\}$ with real scalar parameter θ is said to be of* **monotone likelihood ratio** *if there exists a function $s(\mathbf{x})$ such that the likelihood ratio*

$$\frac{p(\mathbf{x}|\theta_2)}{p(\mathbf{x}|\theta_1)}$$

is a non-decreasing function of $s(\mathbf{x})$ whenever $\theta_1 < \theta_2$.

Note that the same result applies for a non-increasing test statistic, by replacing $t(\mathbf{x})$ by $-t(\mathbf{x})$.

Theorem 3.2. *Suppose $\mathbf{X}$ has a distribution from a family that is monotone likelihood ratio with respect to some continuous test statistic $s(\mathbf{X})$ and we wish to test $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$, then a UMP test exists with critical region of the form $s \geq s_\alpha$.*

Proof. For testing $\theta = \theta_0$ against $\theta = \theta_1$ for any specific $\theta_1 \in \Theta_1$, the Neyman-Pearson lemma tells us that the most powerful critical region is given by the likelihood ratio critical region. The LR is a non-decreasing function of $s(\mathbf{y})$ for any $\theta_1 > \theta_0$, and so the critical region is of the form $s \geq s_\alpha$. s_α is determined by the size of the test and depends only on θ_0. Hence, this critical region is identical for all $\theta_1 \geq \theta_0$ and this test is UMP. □

Corollary 3.1. *If $X_1, \ldots, X_n$ are i.i.d with p.d.f. of the form*

$$p(x|\,\theta) = \exp\{a(x)b(\theta) + c(\theta) + d(x)\}$$

with θ a scalar parameter and $b(\theta)$ strictly increasing, then for testing the null hypothesis that $\theta = \theta_0$ against $\theta > \theta_0$ the LR test has critical regions corresponding to large values of $s = \sum a(x_j)$ and is UMP.

Proof For any $\theta_1 > \theta_0$, the LR is

$$\frac{p_{\mathbf{X}}(\mathbf{x}|\,\theta_1)}{p_{\mathbf{X}}(\mathbf{x}|\,\theta_0)} = \exp[\{b(\theta_1) - b(\theta_0)\}s + n\{c(\theta_1) - c(\theta_0)\}].$$

Since $b(\theta_1) > b(\theta_0)$, this is monotone likelihood ratio and so the conditions of Theorem 3.2 are satisfied. This applies to all one-parameter exponential families, e.g. Normal, Binomial, Poisson. There are similar results for $\theta < \theta_0$, when $b(\theta)$ is a decreasing function.

Example.

- Angles : take H_0 to be that angles $X_1, \ldots, X_n$ are i.i.d. and Uniform on $[0, 2\pi)$.

  A set of alternatives representing a type of symmetrical clustering about $y = 0$ has the X_j i.i.d. with von Mises p.d.f.

  $$\frac{\exp(\theta \cos x)}{2\pi I_0(\theta)} \quad (0 \leq x < 2\pi;\ \theta > 0).$$

  So we test the hypothesis $H_0 : \theta = 0$ against the alternative $\theta > 0$.

3.12 Designing the best test: composite null and alternative hypotheses

3.12.1 One-sided tests*

Previously we considered tests of hypotheses where the null hypothesis was simple. Testing composite hypotheses is more complex in general. However, the above result for monotone likelihood ratio distributions also applies to one-sided tests of the form $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$.

Theorem 3.3. *Suppose* $\mathbf{X}$ *has a distribution from a family that is monotone likelihood ratio with respect to some continuous test statistic* $s(\mathbf{X})$ *and we wish to test* $H_0 : \theta \leq \theta_0$ *against* $H_1 : \theta > \theta_0$*, then*

(a) The test

$$\phi_0(\mathbf{x}) = \begin{cases} 1 & \textit{if } s(\mathbf{x}) > s_0, \\ 0 & \textit{if } s(\mathbf{x}) \leq s_0, \end{cases} \tag{56}$$

is UMP among all tests of size $\leq \mathbb{E}_{\theta_0}\left\{\phi_0(\mathbf{X})\right\}$.

(b) Given some $0 < \alpha \leq 1$*, there exists an* s_0 *such that the tests in (a) has size exactly equal to* α.

Proof. 1. From Theorem 3.2, ϕ_0 is UMP for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$.

2. $\mathbb{E}_\theta\{\phi_0(\mathbf{x})\}$ is a non-decreasing function of θ. If we have $\theta_2 < \theta_1$ and $\mathbb{E}_{\theta_2}\{\phi_0(\mathbf{x})\} = \beta$, then the trivial test $\phi(\mathbf{x}) = \beta$ has $\mathbb{E}_{\theta_1}\{\phi(\mathbf{x})\} = \beta$. The test ϕ_0 is UMP for testing θ_2 against θ_1 and so it must be at least as good as ϕ, i.e., $\mathbb{E}_{\theta_1}\{\phi_0(\mathbf{x})\} \geq \beta$. Hence, if we construct the test with $\mathbb{E}_{\theta_0}\{\phi_0(\mathbf{x})\} = \alpha$, then $\mathbb{E}_\theta\{\phi_0(\mathbf{x})\} \leq \alpha$ for all $\theta \leq \theta_0$, so ϕ_0 is also of size α under the larger hypothesis $H_0 : \theta \leq \theta_0$.

3. For any other test ϕ that is of size α under H_0, we have $\mathbb{E}_{\theta_0}\{\phi(\mathbf{x})\} \leq \alpha$ and by the Neyman-Pearson lemma $\mathbb{E}_{\theta_1}\{\phi_(\mathbf{x})\} \leq \mathbb{E}_{\theta_1}\{\phi_0(\mathbf{x})\}$ for any $\theta_1 > \theta_0$. This shows that this test is UMP among all tests of its size.

4. If α is specified we must show that there exists a s_0 such that $\mathbb{P}_{\theta_0}\{s(\mathbf{X}) > s_0\} = \alpha$, but this follows from the assumption that $s(\mathbf{X})$ is continuous.

□

3.12.2 Two-sided tests

In more general situations we will be interested in testing hypotheses of the form $H_0 : \theta \in \Theta_0$, where Θ_0 is either an interval $[\theta_1, \theta_2]$ for $\theta_1 < \theta_2$ or a single point $\Theta_0 = \{\theta_0\}$, against the generic alternative $H_1 : \theta \in \Theta_1$, with $\Theta_1 = \mathbb{R}/\Theta_0$. For a family with monotone likelihood ratio with respect to a statistic $s(\mathbf{X})$, we might expect a good test to have a test function of the form

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{if } s(\mathbf{x}) > s_2 \text{ or } s(\mathbf{x}) < s_1, \\ \gamma(\mathbf{x}) & \text{if } s(\mathbf{x}) = s_2 \text{ or } s(\mathbf{x}) = s_1, \\ 0 & \text{if } s_1 < s(\mathbf{x}) < s_2. \end{cases}$$

Such a test is called **a two-sided test**. For such two-sided tests, we cannot usually find a UMP test. However, under certain circumstances it is possible to find a **uniformly most powerful unbiased** (UMPU) test.

Definition 3.4. *A test $\phi(\mathbf{y})$ of $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is called* **unbiased of size** *α if*

$$\sup_{\theta\in\Theta_0} \mathbb{E}_\theta \left\{\phi(\mathbf{Y})\right\} \leq \alpha$$

and

$$\mathbb{E}_\theta \left\{\phi(\mathbf{Y})\right\} \geq \alpha \textit{ for all } \theta \in \Theta_1.$$

In other words, an unbiased test is one which has higher probability of rejecting H_0 when it is false than when it is true. Note that if the power function is a continuous function of θ then an unbiased test of size α must have size equal to α on the boundary of the critical region (since the size is less than or equal to α within the critical region and greater than or equal to α outside).

Definition 3.5. *A test which is uniformly most powerful among the set of all unbiased tests is called* **uniformly most powerful unbiased**.

For a scalar exponential family of the form given in Corollary 3.1 the following theorem holds

Theorem 3.4. *If $X_1, \ \ldots \ , X_n$ are i.i.d with p.d.f. of the form*

$$p(x|\,\theta) = \exp\{a(x)b(\theta) + c(\theta) + d(x)\}$$

with θ a scalar parameter and $b(\theta)$ strictly increasing, then there exists a unique UMPU test of size α, ϕ', for testing the hypothesis $H_0 : \theta \in [\theta_1, \theta_2]$, against the generic alternative $H_1 : \theta \in \mathbb{R} - [\theta_1, \theta_2]$, of the form

$$\phi'(\mathbf{x}) = \begin{cases} 1 & \textit{if } s(\mathbf{x}) > s_2 \textit{ or } s(\mathbf{x}) < s_1, \\ \gamma_j & \textit{if } s(\mathbf{x}) = s_j, \\ 0 & \textit{if } s_1 < s(\mathbf{x}) < s_2. \end{cases} \tag{57}$$

where $S = \sum a(x_j)$, for which

$$\mathbb{E}_{\theta_j}\phi'(\mathbf{X}) = \mathbb{E}_{\theta_j}\phi(\mathbf{X}) = \alpha, \qquad j = 1, 2.$$

The boundaries of the critical region, s_1, s_2, and the rejection probabilities on the boundaries, γ_1, γ_2, are determined from the conditions $\mathbb{E}_{\theta_j}\phi'(\mathbf{X}) = \alpha$.

Example. Suppose a sample Y is drawn from an Exp(λ) distribution, so that $f(y|\lambda) = \lambda\exp(-\lambda y)$. Construct a uniformly most powerful unbiased test of size $\alpha = 0.05$ of the hypothesis $H_0 : \lambda \in [1, 2]$ against the generic alternative $\lambda \in [0, 1) \cup (2, \infty)$.

For a single sample from the exponential distribution, the sufficient statistic is the observed value, y. Using the previous result, the UMPU test is of the form (57). The probability that $s = s_i$ is zero for any single value s_i and therefore the γ_i's do not need to be determined. The boundaries of the critical region can be found from the constraints

$$\alpha = 0.05 = 1 - \exp(-s_1) + \exp(-s_2) = 1 - \exp(-2s_1) + \exp(-2s_2),$$

from which we find $s_1 = 0.02532$ and $s_2 = 3.6889$. The corresponding power function $\eta(\lambda)$ is shown in Figure 2. This shows that the test is unbiased as the probability of rejecting H_0 is less than or equal to the size α within the region defined by H_0, it is equal to α on the boundary, and greater than α everywhere outside that region.

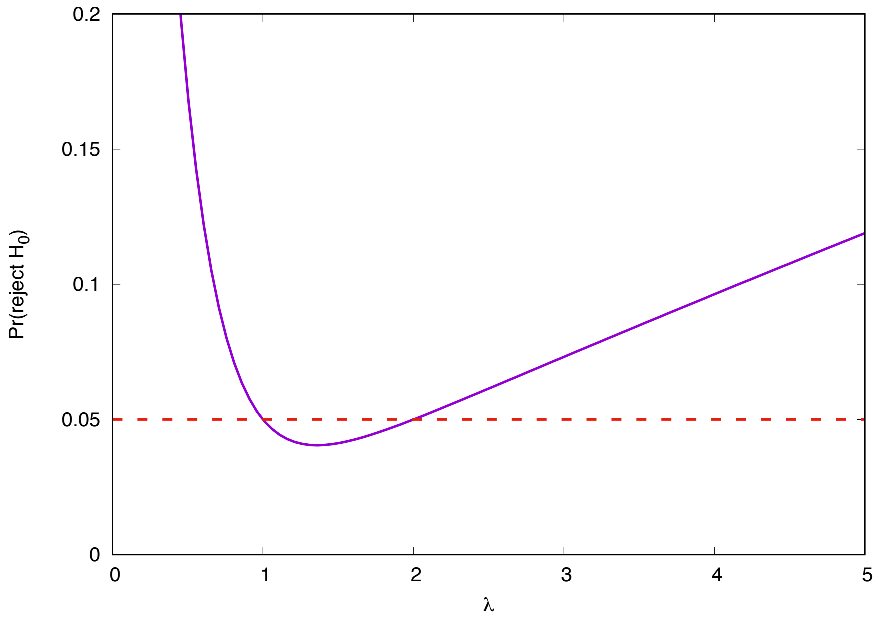


Figure 2: Power of the UMPU test of $\lambda \in [1,2]$ against a generic alternative for an exponential distribution, as a function of λ, i.e., $\mathbb{P}_\lambda(\text{reject } H_0)$. The horizontal line indicates the size of the test, $\alpha = 0.05$.

3.12.3 Testing a point null hypothesis*

A test of the null hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ can be considered as the limit of the preceding two-sided test when $\theta_2 - \theta_1 \to 0$. Therefore, as a corollary to the previous result, there must exist a unique UMPU test, ϕ', of this hypothesis of the form (57) for which

$$\mathbb{E}_{\theta_0}\{\phi'(X)\} = \alpha, \qquad \frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta\{\phi'(X)\}|_{\theta=\theta_0} = 0. \tag{58}$$

Differentiability of the power function for any test function is ensured from the assumption that the distribution is in the exponential family.

Example. Returning to the example of the preceding section of a single sample from an Exp(λ) distribution, if we instead want to test the hypothesis that $\lambda = 1$ then we proceed as before, but the constraints on the boundary of the rejection region are now

$$\begin{aligned} \alpha &= 0.05 = 1 - \exp(-t_1) + \exp(-t_2), \\ 0 &= t_1 \exp(-t_1) - t_2 \exp(-t_2), \end{aligned}$$

which can be solved numerically to give $t_1 = 0.0423633$, $t_2 = 4.76517$. The power function is shown in Figure 3. We see that it reaches a minimum of $\alpha = 0.05$ at $\theta = \theta_0$ so it is unbiased and of size α as desired.

3.13 Designing the best test: similar Tests*

So far we have focussed on tests of one-parameter distributions. However, often the distribution will depend on more than one parameter. In that case we are interested in tests

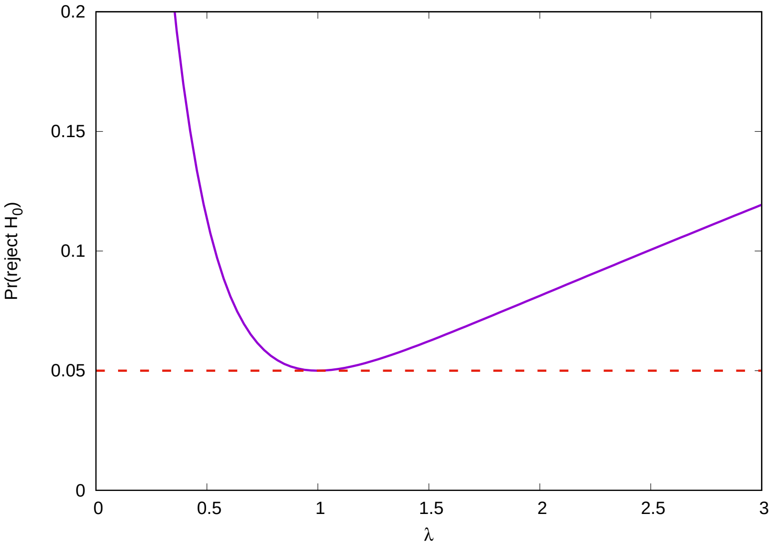


Figure 3: Power of the UMPU test of $\lambda = 1$ against a generic alternative for an exponential distribution, as a function of λ, i.e., $\mathbb{P}_\lambda(\text{reject } H_0)$. The horizontal line indicates the size of the test, $\alpha = 0.05$.

that perform as well as possible in inferring the value of one parameter of the distribution, irrespective of the value of the other parameters of the distribution. This gives rise to the notion of a **similar** test.

Definition 3.6. *Suppose $\theta = (\psi, \lambda)$ and the parameter space is of the form $\Omega_\theta = \Omega_\psi \times \Omega_\lambda$. Suppose we wish to test the null hypothesis $H_0 : \psi = \psi_0$ against the alternative $H_1 : \psi \neq \psi_0$, with λ treated as a nuisance parameter. Suppose $\phi(\mathbf{x})$, $\mathbf{x} \in \mathcal{X}$ is a test of size α for which*

$$\mathbb{E}_{\psi_0,\lambda}\left\{\phi(\mathbf{x})\right\} = \alpha \text{ for all } \lambda \in \Omega_\lambda.$$

Then ϕ is called a **similar test of size** *α.*

This definition can be extended to composite null hypotheses. If the null hypothesis is of the form $\theta \in \Theta_0$, where Θ_0 is a subset of Ω_θ, then a similar test is one for which $\mathbb{E}_\theta\left\{\phi(\mathbf{x})\right\} = \alpha$ on the boundary of Θ_0.

If a test is uniformly most powerful among all similar tests then it is called **UMP similar**. There is close connection to UMPU tests. If the power function of a test is continuous then we saw earlier that any unbiased test of size α must have size exactly equal to α on the boundary, i.e., it must be similar. In such cases, if we can find a UMP similar test and it turns out to also be unbiased, then it is necessarily UMPU.

Moreover, in many cases it is possible to demonstrate that a test which is UMP among all tests based on the conditional distribution of a statistic S given the value of an ancillary statistic A, this test is UMP among all similar tests. In particular, this applies if A is a complete sufficient statistic for the variables λ.

One common situation in which this occurs is for multi-parameter exponential families, for which the likelihood can be written

$$p(x|\theta) = \exp\left\{\sum_{i=1}^{p} A_i(x)B_i(\theta) + C(\theta) + D(x)\right\}.$$

Consider a test of the form $H_0 : B_1(\theta) \leq \theta_1^*$ against $H_1 : B_1(\theta) > \theta_1^*$. If we take $s(\mathbf{x}) = \sum_j A_1(x_j)$ and $A = (\sum_j A_2(x_j), \ldots, \sum_j A_p(x_j))$, then the conditional distribution of S given A is also of the exponential form and doesn't depend on $B_2(\theta), \ldots, B_p(\theta)$, so A is both sufficient and complete for $B_2(\theta), \ldots, B_p(\theta)$. The Conditionality Principle suggests we should make inference about $B_1(\theta)$ based on the conditional distribution of S given A. Tests constructed in this way are UMPU (Ferguson 1967). The optimal one-sided test is then of the following form. Based on observations $s_1 = \sum_j A_1(x_j)$, $s_2 = \sum_j A_2(x_j), \ldots, s_p = \sum_j A_p(x_j)$, we reject H_0 if and only if $s_1 > s_1^*$, where s_1^* is calculated from

$$\mathbb{P}_{B_1(\theta)=\theta_1^*}\left\{S_1 > s_1^* | S_2 = s_2, \ldots, S_p = s_p\right\} = \alpha.$$

It can be shown this is a UMPU test of size α.

Similarly, to construct a two-sided test of $H_0 : \theta_1^* \leq B_1(\theta) \leq \theta_1^{**}$ against $B_1(\theta) < \theta_1^*$ or $B_1(\theta) > \theta_1^{**}$, we first define the conditional power function

$$w_{\theta_1}(\phi|s_2, \ldots, s_p) = \mathbb{E}_{\theta^1}\left\{\phi(S_1)|S_2 = s_2, \ldots, S_p = s_p\right\}.$$

Then we can construct a two-sided conditional test of the form

$$\phi'(s_1) = \begin{cases} 1 & \text{if } s_s < s_1^* \text{ or } s_1 > s_1^{**}, \\ 0 & \text{if } s_1^* \leq s_1 \leq s_1^{**}, \end{cases}$$

where s_1^* and s_1^{**} are chosen such that

$$w_{\theta_1}(\phi'|s_2, \ldots, s_p) = \alpha \quad \text{when } B(\theta_1) = \theta_1^* \text{ or } B(\theta_1) = \theta_1^{**}.$$

It can be shown that these tests are also UMPU of size α. If the test is of a simple hypothesis $B(\theta_1) = \theta_1^*$ against the generic alternative $B(\theta_1) \neq \theta_1^*$ then the test is of the same form but the conditions are that the power function is equal to α and its derivative with respect to θ is equal to 0, as in Eq. (58).

3.14 Generalized likelihood ratio tests

In the previous sections we focussed on finding the "best" tests by one metric or another. However, as we have seen this is not always easy and the resulting test statistics are not always straightforward to evaluate. Under many circumstances, in the limit $n \to \infty$, the likelihood ratio follows a χ^2 distribution and so this can be used to construct a test that is valid asymptotically.

In particular, suppose we are testing $H_0 : \vec{\theta} \in \Theta_0$ versus $H_1 : \vec{\theta} \in \Theta_1$. We define the likelihood ratio

$$L_X(H_0, H_1) = \frac{\sup_{\vec{\theta}\in\Theta_1} p(x|\theta)}{\sup_{\vec{\theta}\in\Theta_0} p(x|\theta)}$$

and denote by $p = |\Theta_1 - \Theta_0|$ the difference in the numbers of degrees of freedom in the unknown parameters between the two hypotheses. Then as $n \to \infty$

$$2 \log L_X(H_0, H_1) \sim \chi_p^2$$

under H_0 and tends to be larger under H_1. Therefore critical regions of the form $2 \log L_X > \chi^2_p(\alpha)$ give tests of approximately size α.

The interpretation of p is the number of constraints that have been placed to reduce the, typically more general, alternative hypothesis, to the more restrictive null hypothesis. For example, the null hypothesis might be specified by fixing the values of p of the parameters, or by imposing p linear constraints on the parameters, or by writing the k parameters of Θ_1 as functions of an alternative $k - p$ dimensional parameter space.

4 Examples of frequentist statistics in gravitational wave astronomy

In this section we will describe some of the applications of frequentist statistical methods to gravitational wave detection. Fundamental to frequentist statistics is the likelihood. For gravitational wave detectors, we assume that the output of the detector, $s(t)$, is a linear combination of a signal, $h(t|\vec{\lambda})$, determined by a finite set of (unknown) parameters, $\vec{\lambda}$, and instrumental noise, $n(t)$. We assume in addition that the noise is Gaussian with a (usually known) power spectral density $S_h(f)$

$$s(t) = n(t) + h(t|\vec{\lambda}), \qquad \langle \tilde{n}^*(f)\tilde{n}(f')\rangle = S_h(f)\delta(f-f'). \tag{59}$$

The signal is deterministic, but the noise is a random process. The likelihood, for parameters $\vec{\lambda}$, is therefore the probability that the observed noise realisation would take the value $n(t) = s(t) - h(t|\vec{\lambda})$, which can be seen to be

$$\mathcal{L}(s|\vec{\lambda}) = p(n(t) = s(t) - h(t|\vec{\lambda})) \propto \exp\left[-\frac{1}{2}(s - h(\vec{\lambda})|s - h(\vec{\lambda}))\right] \tag{60}$$

where the noise weighted overlap is

$$(a|b) = \int_{-\infty}^{\infty} \frac{\tilde{a}^*(f)\tilde{b}(f) + \tilde{a}(f)\tilde{b}^*(f)}{S_h(f)} \mathrm{d}f.$$

These expressions will be justified more carefully in Chapter 8.

4.1 The Fisher Matrix

We introduced the Fisher Matrix in the discussion of the Cramer-Rao bound on the variance of an estimator, which, for a multivariate unbiased estimator, $\hat{\lambda}$, is given by

$$\mathrm{cov}(\hat{\lambda}_i, \hat{\lambda}_j) \geq [\mathbf{\Gamma}_\lambda]^{-1}_{ij}$$

where

$$(\mathbf{\Gamma}_\lambda)_{ij} = \mathbb{E}\left[\frac{\partial l}{\partial \lambda_i}\frac{\partial l}{\partial \lambda_j}\right].$$

In the above l denotes the log-likelihood. For the gravitational wave log-likelihood in Eq. (60), the derivative is

$$\frac{\partial l}{\partial \lambda_i} = \left(\frac{\partial h}{\partial \lambda_i}\middle| s - h(\vec{\lambda})\right) = \left(\frac{\partial h}{\partial \lambda_i}\middle| \mathbf{n}\right).$$

It therefore follows from Eq. (59) (see Chapter 8 for an explicit calculation), that

$$(\mathbf{\Gamma}_\lambda)_{ij} = \mathbb{E}\left[\frac{\partial l}{\partial \lambda_i}\frac{\partial l}{\partial \lambda_j}\right] = \left\langle \left(\frac{\partial h}{\partial \lambda_i}\middle| \mathbf{n}\right)\left(\frac{\partial h}{\partial \lambda_i}\middle| \mathbf{n}\right)\right\rangle = \left(\frac{\partial h}{\partial \lambda_i}\middle| \frac{\partial h}{\partial \lambda_j}\right).$$

The Fisher Matrix gives a lower bound on the variance of any unbiased estimator of the parameters of the signal, and hence it provides a guide to how accurately the parameters can be measured. We know that the maximum likelihood estimator is asymptotically efficient, i.e., it achieves this Fisher Matrix bound, which is why it might be expected to provide a good guide to parameter measurement precision. However, asymptotic efficiency refers to making many repeated measurements of the same parameter, which we do not typically do in gravitational wave observations. But it can be seen that the Fisher Matrix provides a good guide to measurement precision even for a single observation, as follows. We suppose that the true parameters of the signal are given by $\vec{\lambda}_0$, and expand to leading order about those parameters

$$\vec{\lambda} = \vec{\lambda}_0 + \Delta\vec{\lambda}, \qquad h(t|\vec{\lambda}) = h(t|\vec{\lambda}) + \partial_i h(t|\vec{\lambda})\Delta\lambda^i$$

where ∂_i denotes the derivative with respect to λ_i and the last term employs Einstein summation convention. This approximation is known as the **linear signal approximation**. The likelihood can then be expanded as

$$\begin{aligned}\mathcal{L}(s|\vec{\lambda}) &\propto \exp\left[-\frac{1}{2}(n - \partial_i h(t|\vec{\lambda})\Delta\lambda^i | n - \partial_j h(t|\vec{\lambda})\Delta\lambda^j)\right] \\ &= \exp\left\{-\frac{1}{2}\left[(n|n) - 2(n|\partial_i h(t|\vec{\lambda}))\Delta\lambda^i + (\partial_i h(t|\vec{\lambda})|\partial_j h(t|\vec{\lambda}))\Delta\lambda^i\Delta\lambda^j\right]\right\} \\ &= \exp\left[-\frac{1}{2}(n|n)\right]\exp\left[-\frac{1}{2}\left(\Delta\lambda^i - (\Gamma^{-1})_{ik}(n|\partial_k h(t|\vec{\lambda}))\right)\Gamma_{ij}\left(\Delta\lambda^j - (\Gamma^{-1})_{jl}(n|\partial_l h(t|\vec{\lambda}))\right)\right] \\ &\qquad \times \exp\left[-\frac{1}{2}(n|\partial_i h(t|\vec{\lambda}))(\Gamma^{-1})_{ij}(n|\partial_j h(t|\vec{\lambda}))\right]. \end{aligned} \tag{61}$$

The latter term is sub-dominant since it is $O(1)$ compare to the middle term which is of order of the signal amplitude, or SNR. The middle term is a Gaussian, centred at $\Delta\lambda^i = (\Gamma^{-1})_{ik}(n|\partial_k h(t|\vec{\lambda}))$, and with covariance matrix given by the Fisher Matrix. The latter therefore provides an estimate of the width of the likelihood distribution and hence can be used as a guide to the uncertainty. In addition, the maximum likelihood estimator

$$\widehat{\Delta\lambda}^i = (\Gamma^{-1})_{ik}(n|\partial_k h(t|\vec{\lambda}))$$

has mean and variance

$$\mathbb{E}\left(\widehat{\Delta\lambda}^i\right) = 0, \qquad \mathrm{cov}\left(\widehat{\Delta\lambda}^i, \widehat{\Delta\lambda}^j\right) = \Gamma^{-1}_{ij},$$

which again confirms the interpretation of the Fisher Matrix as the uncertainty in the parameter estimate. The fractional corrections to the Fisher Matrix estimate scale like the inverse of the signal-to-noise ratio and therefore the Fisher Matrix is a good approximation in the high signal-to-noise ratio limit.

The Fisher Matrix has been widely used in a gravitational wave context to assess the measurability of parameters using observations with present or future detectors. While the Fisher Matrix is only an approximation, it can be directly calculated by evaluating a small number of waveforms, rather than requiring samples to be obtained all over the waveform parameter space, and so it is much cheaper computationally. This makes it a good tool for Monte Carlo simulations over parameter space, to survey parameter estimation accuracies over a wide parameter range.

4.2 Matched filtering

The notion of filtering will be discussed in more detail in Chapter 8, but the basic idea is to construct a statistic based on the output of a filter, which is the convolution of the observed data with the specified filter template. When the form of the signal is known, the filter with the highest signal to noise ratio, called the optimal filter, has a frequency-domain kernel $\tilde{K}(f) \propto \tilde{h}(f)/S_h(f)$. The use of the output of this filter as a test statistic for a search can also be motivated by the frequentist concepts that we encountered in previous chapters. Suppose that we write $\mathbf{h}(\lambda) = A\hat{\mathbf{h}}(\lambda)$, where $(\hat{\mathbf{h}}(\lambda)|\hat{\mathbf{h}}(\lambda)) = 1$, to separate out the amplitude of the gravitational wave source from the other parameters. The log-likelihood can be written

$$\begin{aligned} l(\lambda) &= -\frac{1}{2}(\mathbf{s} - A\hat{\mathbf{h}}(\lambda)|\mathbf{s} - A\hat{\mathbf{h}}(\lambda)) = -\frac{1}{2}\left[(\mathbf{s}|\mathbf{s}) - 2A(\mathbf{s}|\hat{\mathbf{h}}) + A^2\right] \\ &= -\frac{1}{2}\left[(\mathbf{s}|\mathbf{s}) + (A - (\mathbf{s}|\hat{\mathbf{h}}))^2 - (\mathbf{s}|\hat{\mathbf{h}})^2\right]. \end{aligned} \tag{62}$$

For a given λ, this is maximized by the choice $A = (\mathbf{s}|\hat{\mathbf{h}})$, for which the log-likelihood $\propto (\mathbf{s}|\hat{\mathbf{h}})^2 - (\mathbf{s}|\mathbf{s})$. The maximum likelihood estimator for parameters other than the amplitude is thus given by the maximum of the optimal filter output over the parameter space. So, optimal filtering is just maximum likelihood estimation. To do this in practice, the optimal filter must be evaluated over the whole parameter space. In the analysis of gravitational wave data, from LIGO in particular, this is achieved using a **template bank**, which is a set of templates that cover the whole parameter space. The overlap of each template with the detector data is evaluated, and the maximum of those template overlaps is used as a test statistic to identify whether or not there is a signal in the data.

The question that we want to ask is "Is there a gravitational wave signal in the data?". Assuming that the parameters λ are fixed, this can formulated as a hypothesis test on the signal amplitude

$$H_0 : A = 0, \qquad \text{vs.} \qquad H_1 : A > 0.$$

From the Neyman-Pearson lemma the optimal statistic for testing the simple hypothesis $A = 0$ versus $A = A_1$ is the likelihood ratio, which is

$$\exp\left[A_1(\mathbf{s}|\hat{\mathbf{h}}(\lambda)) - \frac{1}{2}A_1^2\right].$$

This is large for large values of the optimal filter $(\mathbf{s}|\hat{\mathbf{h}}(\lambda))$ and so we deduce that the optimal filter is also the most powerful detection statistic. As the detection statistic does not depend on A_1, this test is uniformly most powerful for the composite hypothesis $A > 0$. In the more usual case that λ is unknown, although the maximum of the optimal filter statistic is still the maximum likelihood estimator, this is no longer a uniformly most powerful test, although it remains quite close to being so.

LIGO matched filtering searches typically use a large number of templates, distributed throughout the parameter space in a **template bank**. The matched filter output is evaluated for all of these templates, and the maximum filter output over the template bank is used as a detection statistic. Template banks are typically characterised by their **minimal match**, MM. This is defined as the *minimum* over all *possible signals* of the **maximum** overlap of that signal with one of the templates in the bank

$$\min_{\vec{\lambda}}\left[\max_{h_{\rm temp,i}:i=1,\ldots,N}(h(\vec{\lambda})|h_{\rm temp,i})\right] \gtrsim \mathrm{MM}$$

where $\{h_{\rm temp,i} : i = 1, \ldots, N\}$ are the N templates in the template bank. The minimal match is the worst possible detection statistic that a randomly chosen signal could have. Setting this minimal match to some value close to 1 ensures that very few signals will be missed. A typical value of the minimal match used in practice would be 0.97. For a uniform distribution of sources in a Euclidean Universe, the fraction of sources that would be missed is $1-0.97^3 = 0.087$.

Template banks can be constructed analytically using the Fisher Matrix as a metric. This follows from expanding the overlap of two normalised templates, $\hat{h}(\vec{\lambda}) = h(\vec{\lambda})/\sqrt{(h(\vec{\lambda})|h(\vec{\lambda}))}$,

$$(\hat{h}(\vec{\lambda})|\hat{h}(\vec{\lambda}+\Delta\vec{\lambda})) = (\hat{h}(\vec{\lambda})|\hat{h}(\vec{\lambda}))+\left(\hat{h}(\vec{\lambda})\left|\frac{\partial\hat{h}}{\partial\lambda_i}(\vec{\lambda})\right.\right)\Delta\lambda^i+\frac{1}{2}\left(\hat{h}(\vec{\lambda})\left|\frac{\partial^2\hat{h}}{\partial\lambda_i\partial\lambda_j}(\vec{\lambda})\right.\right)\Delta\lambda^i\Delta\lambda^j+\cdots .$$

The first term is 1 because of the normalisation. The second term vanishes since

$$(\hat{h}(\vec{\lambda})|\hat{h}(\vec{\lambda})) = 1 \quad\Rightarrow\quad \frac{\partial}{\partial\lambda_i}(\hat{h}(\vec{\lambda})|\hat{h}(\vec{\lambda})) = 0 \quad\Rightarrow\quad \left(\hat{h}(\vec{\lambda})\left|\frac{\partial\hat{h}}{\partial\lambda_i}(\vec{\lambda})\right.\right) = 0.$$

The third term can be simplified using

$$\begin{aligned}\frac{\partial}{\partial\lambda_j}\left(\hat{h}(\vec{\lambda})\left|\frac{\partial\hat{h}}{\partial\lambda_i}(\vec{\lambda})\right.\right) = 0 \quad &\Rightarrow\quad \left(\frac{\partial\hat{h}}{\partial\lambda_i}(\vec{\lambda})\left|\frac{\partial\hat{h}}{\partial\lambda_j}(\vec{\lambda})\right.\right) + \left(\hat{h}(\vec{\lambda})\left|\frac{\partial^2\hat{h}}{\partial\lambda_i\partial\lambda_j}(\vec{\lambda})\right.\right) = 0\\ &\Rightarrow\quad \left(\frac{\partial\hat{h}}{\partial\lambda_i}(\vec{\lambda})\left|\frac{\partial\hat{h}}{\partial\lambda_j}(\vec{\lambda})\right.\right) = -\left(\hat{h}(\vec{\lambda})\left|\frac{\partial^2\hat{h}}{\partial\lambda_i\partial\lambda_j}(\vec{\lambda})\right.\right). \qquad (63)\end{aligned}$$

We deduce

$$(\hat{h}(\vec{\lambda})|\hat{h}(\vec{\lambda}+\Delta\vec{\lambda})) = 1-\frac{1}{2}\Gamma_{ij}\Delta\lambda^i\Delta\lambda^j.$$

The Fisher Matrix (of normalised templates) thus provides a metric on parameter space, which can be used to place templates. This is only practical in low numbers of dimensions. In higher numbers of dimensions, it is easier to use **stochastic template banks**. A stochastic bank is constructed as follows

1. At step 1, choose the first template, $\hat{h}(\lambda_1)$, randomly from parameter space. Add it to the template bank, $\mathcal{T}$.

2. At step $i \geq 2$, set the counter to 1 and then repeat the following steps:

   (a) Draw a random set of parameter values, $\vec{\lambda}_i$, and evaluate the match, M, with the current template bank

   $$M = \left[\max_{h_{\rm temp}\in\mathcal{T}}(h(\vec{\lambda}_i)|h_{\rm temp})\right].$$

   (b) If $M < MM$, add $h(\vec{\lambda}_i)$ to the template bank and advance to step $i+1$. Otherwise, increment the counter. If the counter has reached $N_{\rm max}$, stop. Otherwise return to step (a).

4.3 LIGO searches

LIGO employs two different matched filtering algorithms to search for signals, Pycbc and Gstlal. They differ in various details, including how the template overlaps are computed. We will not discuss these in detail here, but refer the interested reader to relevant publications. For *gstlal* these are

- Cannon, K., Cariou, R., Chapman, A., et al. (2012), *Astrophys. J.* **748**, 136, doi: 10.1088/0004-637X/748/2/136.

- Privitera, S., Mohapatra, S. R. P., Ajith, P., et al. (2014), *Phys. Rev.* D **89**, 024003, doi: 10.1103/PhysRevD.89.024003

- Messick, C., Blackburn, K., Brady, P., et al. (2017), *Phys. Rev.* D **95**, 042001, doi: 10.1103/PhysRevD.95.042001

- Sachdev, S., Caudill, S., Fong, H., et al. (2019), arXiv:1901.08580

- Hanna, C., Caudill, S., Messick, C., et al. (2019), arXiv:1901.02227

For *pycbc* the relevant references are

- Nitz, A., Harry, I., Brown, D., et al. (2019), gwastro/pycbc: PyCBC Release v1.15.2, doi: 10.5281/zenodo.3596447

- Nitz, A. H., Dal Canton, T., Davis, D., & Reyes, S. (2018), *Phys. Rev.* D **98**, 024050, doi: 10.1103/PhysRevD.98.024050

- Usman, S. A., Nitz, A. H., Harry, I. W., et al. (2016), *Class. Quantum Grav.* **33**, 215004, doi: 10.1088/0264-9381/33/21/215004

Both searches adopt a traditional frequentist framework, in that the output of the pipeline is used as a detection statistic. If the detection statistic exceeds a threshold then the data is flagged as interesting, i.e., potentially containing a signal. The threshold is determined based on the behaviour of the search pipeline in the absence of any signals in the data. This background distribution is estimated using **time slides**. Both searches rely on consistency between triggers in two or more detectors. Any astrophysical gravitational wave signal must pass through both detectors within an interval of 10ms. If the data of one detector is time shifted relative to the other by more than this amount, then any coincident triggers in the two instruments must be due to instrumental noise only. By doing many different time shifts in this way, the background distribution can be estimated for much longer effective observation times.

In hypothesis testing, we discussed the notion of a significance or p-value. This makes sense if the size of the data set is fixed, but gravitational wave detectors are continuously taking data. Therefore it makes sense to quantify significance instead by a *false alarm rate* or FAR, which is the frequency at which triggers as extreme as the one observed, or more extreme, occur in the data. LIGO quotes FARs for all events that are distributed publicly.

We will now give an overview of a few techniques that are used in LIGO searches to improve their speed and efficiency.

4.3.1 Waveform consistency

The assumptions that lead to the optimal filter assume that the noise is stationary. This is approximately true for gravitational wave detectors, but they are also observed to have large glitches quite often. While the glitches do not match any of the templates well, there is often sufficient power in the glitch that they can trigger the detection statistic to exceed the threshold. To mitigate for this problem, LIGO searches use **waveform consistency** checks. These verify that after subtracting the best-fit template signal from the data, the resulting time series is consistent with being stationary Gaussian noise with the estimated PSD. If the template $\hat{h}$ coincides with the true signal, the quantity

$$\chi^2 = \sum_{k=1}^{N} \frac{|\hat{s}_k - \hat{h}_k|^2}{S_h(f_k)}$$

is the sum of squares of $N(0,1)$ distributed random variables, and hence follows a chi-squared distribution with N degrees of freedom. The mean of a χ^2_N random variable is N, so χ^2/N should be expected to be close to 1 if the template is a good match to the data, and much bigger otherwise. LIGO uses something called *effective SNR* as a detection statistic. This is defined as

$$\hat{\rho} = \frac{\rho}{(1+(\chi^2/N)^3)^{\frac{1}{6}}}.$$

For real signals, this is close to the true SNR, while for glitches it is much smaller. The effective SNR is used as the detection statistic by *pycbc*.

4.3.2 Marginalisation over phase and time

A template bank requires templates in all parameters, so it is useful to reduce the dimensionality of the parameter space whenever possible. This can be done straightforwardly for the *initial phase* and *time of coalescence*. For a monochromatic signal

$$h(t|A,f_0,t_c,\phi_0) = A\cos(2\pi f_0(t-t_c)+\phi_0) = A\cos(2\pi f_0(t-t_c))\cos\phi_0 - A\sin(2\pi f_0(t-t_c)\sin\phi_0$$

the matched filter overlap is

$$(s|h) = A\cos\phi_0 O_c - A\sin\phi_0 O_s, \qquad \text{where } O_c = (s|\cos(2\pi f_0(t-t_c))), \quad O_s = (s|\sin(2\pi f_0(t-t_c))).$$

Differentiating with respect to ϕ_0 and equating it to zero, we find that the value of ϕ_0 that maximises the overlap is

$$\tan\phi_0 = -\frac{O_s}{O_c} \qquad \Rightarrow \qquad \max_{\phi_0}(s|h)^2 = A^2(O_c^2+O_s^2).$$

If this is used instead of the standard overlap, then the template bank automatically maximises over phase and this parameter direction does not need to be covered by templates.

To maximize over the unknown coalescence time we use

$$\tilde{h}(f|A,f_0,t_c,\phi_0) = \tilde{h}(f|A,f_0,0,\phi_0)\exp(-2\pi i f t_c)$$

and observe that

$$(s|h(t|A,f_0,t_c,\phi_0)) = 2\Re\int_{-\infty}^{\infty} \frac{\tilde{s}^*(f)\tilde{h}(f|A,f_0,0,\phi_0)}{S_h(f)}\exp(-2\pi i f t_c)\mathrm{d}f.$$

This is just the inverse Fourier transform of

$$\frac{\tilde{s}^*(f)\tilde{h}(f|A, f_0, 0, \phi_0)}{S_h(f)}.$$

Inverse Fourier transforms can be computed cheaply (in $n \log n$ time) using the fast Fourier transform. Therefore, the time of coalescence can be efficiently maximized over by computing the quantity above, taking its inverse fast Fourier transform, and then finding the maximum of the components of the resulting vector.

4.3.3 The F-statistic

The F-statistic is an extension of the above ideas to more of the extrinsic parameters of the signal. It is not used so much for LIGO, but has been used extensively in LISA data analysis work (see for example Cornish & Porter (2007), *Phys. Rev.* D**75**, 021301; *Class. Quantum Grav.* **24**, 5729). The idea is to write the signal as a sum of modes, such that the coefficients depend only on a (subset of) the extrinsic parameters, and then analytically maximise over those coefficients. For SMBH binaries in LISA the decomposition takes the form

$$h(t) = \sum_{i=1}^{4} a_i(\iota, \psi, D_L, \phi_c) A^i(t|M_c, \mu, t_c, \theta, \phi)$$

where

$$\begin{aligned}
a_1 &= \Lambda[(1+\cos^2\iota)\cos 2\psi \cos\phi_c - 2\cos\iota \sin 2\psi \sin\phi_c] \\
a_2 &= -\Lambda[(1+\cos^2\iota)\sin 2\psi \cos\phi_c + 2\cos\iota \cos 2\psi \sin\phi_c] \\
a_3 &= \Lambda[(1+\cos^2\iota)\cos 2\psi \sin\phi_c + 2\cos\iota \sin 2\psi \cos\phi_c] \\
a_4 &= -\Lambda[(1+\cos^2\iota)\sin 2\psi \sin\phi_c - 2\cos\iota \cos 2\psi \cos\phi_c] \\
A_1 &= M\eta x(t) D^+ \cos(\Phi) \\
A_2 &= M\eta x(t) D^\times \cos(\Phi) \\
A_3 &= M\eta x(t) D^+ \sin(\Phi) \\
A_4 &= M\eta x(t) D^\times \sin(\Phi).
\end{aligned} \tag{64}$$

Here the waveform parameters are inclination ι, polarization angle, ψ, luminosity distance, D_L, phase at coalescence, ϕ_c, chirp mass, M_c, reduced mass ratio, μ, time of coalescence, t_c, colatitude, θ, and azimuth, ϕ. We denote the waveform phase by $\Phi(t)$ and $x = (GM\omega/c^3)^{2/3}$, where ω is the orbital frequency and $M = m_1 + m_2$ is the total mass. The quantities D^+ and D^x are the two components of LISA's time-dependent response function.

Writing $N^i = (s|A^i)$, the matched filter overlap is

$$(s|h) = a_j N^j$$

and we want to maximise this subject to the constraint that the waveform is normalised which becomes

$$a_i M^{ij} a_j = 1,' \qquad \text{where } M^{ij} = (A^i|A^j).$$

This is a standard optimisation problem with solution

$$a_i = (M^{-1})_i j N^j = M_{ij} N^j.$$

The maximized value of the log-likelihood is the F-staistic

$$\mathcal{F} = \frac{1}{2} M_{ij} N^i N^j.$$

This can be used to automatically maximise over extrinsic parameters in a search, reducing the dimensionality of the parameter space to just that of the intrinsic parameters. Note that in the above we have taken the coefficients, a_i, to be independent of one another and unconstrained, while in practice they are correlated and take a potentially limited range of values because they all depend on the same set of four extrinsic parameters. Thus, we are finding the maximum over a space that is somewhat larger than the true space, and contains some unphysical values. If there is a signal in the data, then the maximization must nonetheless still give the right extrinsic parameter values (in the absence of noise).

4.3.4 Power spectral density estimation

The likelihood contains the spectral density of noise in the detector, which is usually not known precisely. LIGO searches (and parameter estimation codes) need to use a PSD that has been estimated from the data. This is accomplished by considering a number of other sections of data, distributed either side of the section of data that is of interested because it is believed to contain a signal. The power spectrum (i.e., the norm squared of the Fourier transform) is computed for each of the empty segments, $\sigma_i^2(f)$, and then these can be combined to give an estimate of the PSD in the segment of interest. The averaging can be done by taking the mean

$$\sigma_0^2(f) = \frac{1}{2N} \sum_{k=1}^{N} (s_k^2 + s_{-k}^2)$$

but in LIGO analyses it is more usual to use the median. The median is less susceptible to outliers in the data arising from non-stationary features in the noise.

4.4 Unmodelled searches

For burst sources matched filtering cannot be used, as it is not possible to build templates of potential signals. LIGO uses a number of different searches for unmodelled sources. Again, we won't describe these in detail, but refer to papers that give full details on the algorithms:

- **Coherent Wave Burst (CWB)**:
  - S. Klimenko et al. (2016), *Phys. Rev.* D **93**, 042004, arXiv:1511.05999.
- **MBTA**:
  - Adams, T., Buskulic, D., Germain, V., et al. (2016), *Class. Quantum Grav.* **33**, 175012, doi: 10.1088/0264-9381/33/17/175012
- **SPIIR**:
  - Luan, J., Hooper, S., Wen, L., & Chen, Y. (2012), *Phys. Rev.* D **85**, 102002, doi: 10.1103/PhysRevD.85.102002
  - Hooper, S., Chung, S. K., Luan, J., et al. (2012), *Phys. Rev.* D **86**, 024012, doi: 10.1103/PhysRevD.86.024012

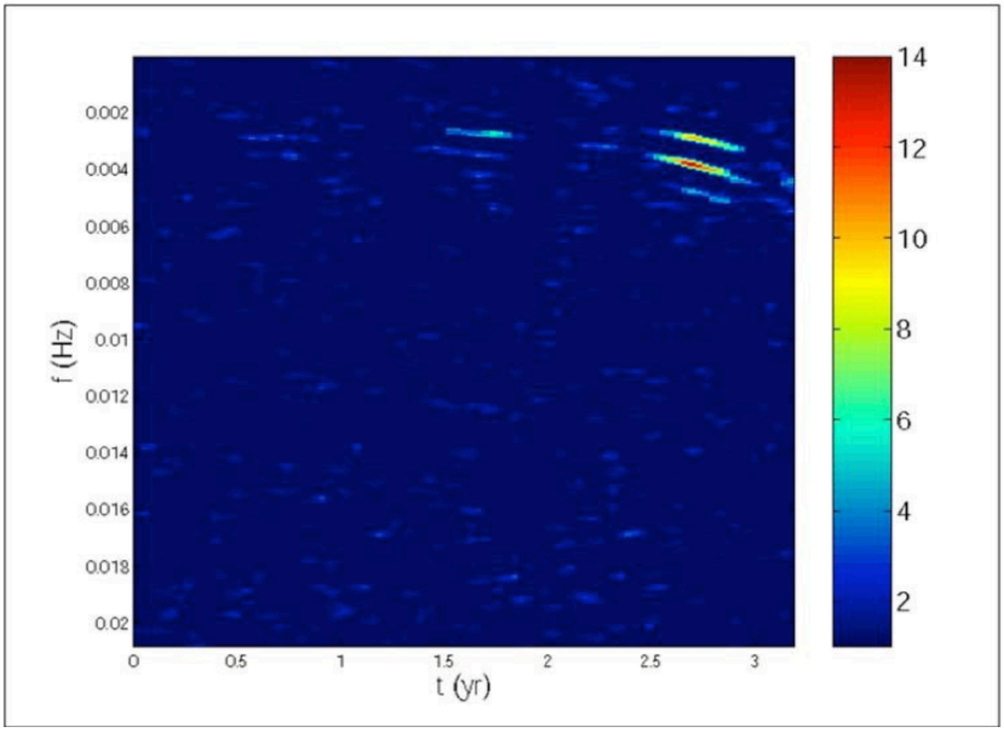


Figure 4: Example of a time-frequency spectrogram. Reproduced from Wen & Gair (2005).

  - Chu, Q. (2017), PhD thesis, University of Western Australia
  - Guo, X., Chu, Q., Chung, S. K., et al. 2018, *Co. Phys.* C **231**, 62, doi: 10.1016/j.cpc.2018.05.002

- **X-pipeline**:
  - Sutton, P. J., Jones, G., Chatterji, S., et al. (2010), *N J Phys.* **12**, 053034
  - Was, M., Sutton, P. J., Jones, G., & Leonor, I. (2012), *Phys. Rev.* D **86**, 022003

All of these algorithms search for clusters in **time-frequency spectrograms** of the data. The full data stream is divided into (usually overlapping) time segments, windowed and Fourier-transformed to obtain a frequency-domain representation of that chunk of data. The norm of these spectra is computed and they are then arranged next to one another in a grid. An example of a spectrogram is shown in Figure 4. Real astrophysical sources tend to produce coherent groups of bright pixels, or tracks, in these spectrograms. The patterns will be similar in different detectors in the network. The various time-frequency algorithms typically first evaluate bright pixels in the spectrograms, by thresholding on the power or some derived quantities. Then they cluster the pixels into groups, apply consistency criteria for the location of groups in two or more detectors in the network, and hence identify triggers of interest.

Time-frequency methods have also been applied to analysis of simulated LISA data, in the context of the LISA Mock Data Challenges (e.g., Gair, J.R. and Jones, G.J. (2007), *Class. Quantum Grav.* **24**, 1145; Gair, J.R., Mandel, I. and Wen, L. (2008), *Class. Quantum Grav.* **25**, 184031; Gair, J.R. and Wen, L. (2005), *Class. Quantum Grav.* **22**, S1359; Wen, L. and Gair, J.R. (2005), Detecting extreme mass ratio inspirals with LISA using time-frequency methods, *Class. Quantum Grav.* **22**, S445.). While these algorithms were successful in simplified situations (i.e., with many fewer sources in the data than we would expect to see in practice) they are unlikely to be very effective when applied to real LISA data, due to the very large number of expected sources that will be overlapping in both time and frequency.

4.5 Semi-coherent searches

For continuous gravitational wave signals, e.g., rotating neutron stars in LIGO data, or very long-lived inspiral signals, e.g., extreme-mass-ratio inspirals in LISA data, matched filtering is possible in the sense that templates of the signals can be generated. however, it is computationally impossible, because the number of templates required to ensure a dense coverage of parameter space is extremely large. In these cases, it is possible to use **semi-coherent** search methods. These involve dividing the data stream into shorter segments, analysing each of those segments with matched filtering, and them adding up the power in the matched filter outputs along trajectories through the segments that correspond to physical inspirals. This approach is summarised in Figure 5. The semi-coherent approach is more computationally efficient, because the number of templates required to densely cover the parameter space for shorter observation times is much smaller.

A discussion of the use of a semi-coherent technique for detection of extreme-mass-ratio inspirals may be found in Gair, J.R. et al. (2004), *Class. Quantum Grav.* **21**, S1595. In that context, the coherent phase used 2 week segments of data, out of 1 year long LISA data sets. The coherent phase also employs the $\mathcal{F}$-statistic described above to automatically maximize over some of the extrinsic parameters.The impact of using the semi-coherent method rather than fully coherent matched filtering is to increase the estimated matched-filtering signal-to-noise ratio threshold for detection from $\rho = 14$ to $\rho = 30$.

In the context of the ground-based detectors, similar methods are used to search for continuous gravitational wave signals from rotating pulsars. The most recent LIGO results from the O2 science run are described in this paper

- Abbott, B.P. et al. (2019), *All-sky search for continuous gravitational waves from isolated neutron stars using Advanced LIGO O2 data*, *Phys. Rev.* D **100**, 024004.

LIGO uses two primary search methods. The **time-domain F-statistic** uses the same technique as the EMRI search described above. In fact, the latter was based on the former. Further details can be found in

- Aasi, J. et al. (2014), *Class. Quantum Grav.* **31**, 165014
- Jaranowski, P., Królak, A. and Schutz, B.F. (1998), *Phys. Rev.* D **58**, 063001
- Astone, P., Borkowski, K.M., Jaranowski, P., Pietka M. and Królak, A. (2010), *Phys. Rev.* D **82**, 022005
- Pisarski, A. and Jaranowski, P. (2015), *Class. Quantum Grav.* **32**, 145014

LIGO also employs a second method, called **the Hough transform**. The first stage of this algorithm is the same as the stack-slide method, i.e., coherent matched filtering on shorter segments of data. The second stage is slightly different, using the Hough transform, which is a technique for edge-detection in images, to identify tracks through the coherent template overlaps that might correspond to true signals. Further details can be found in

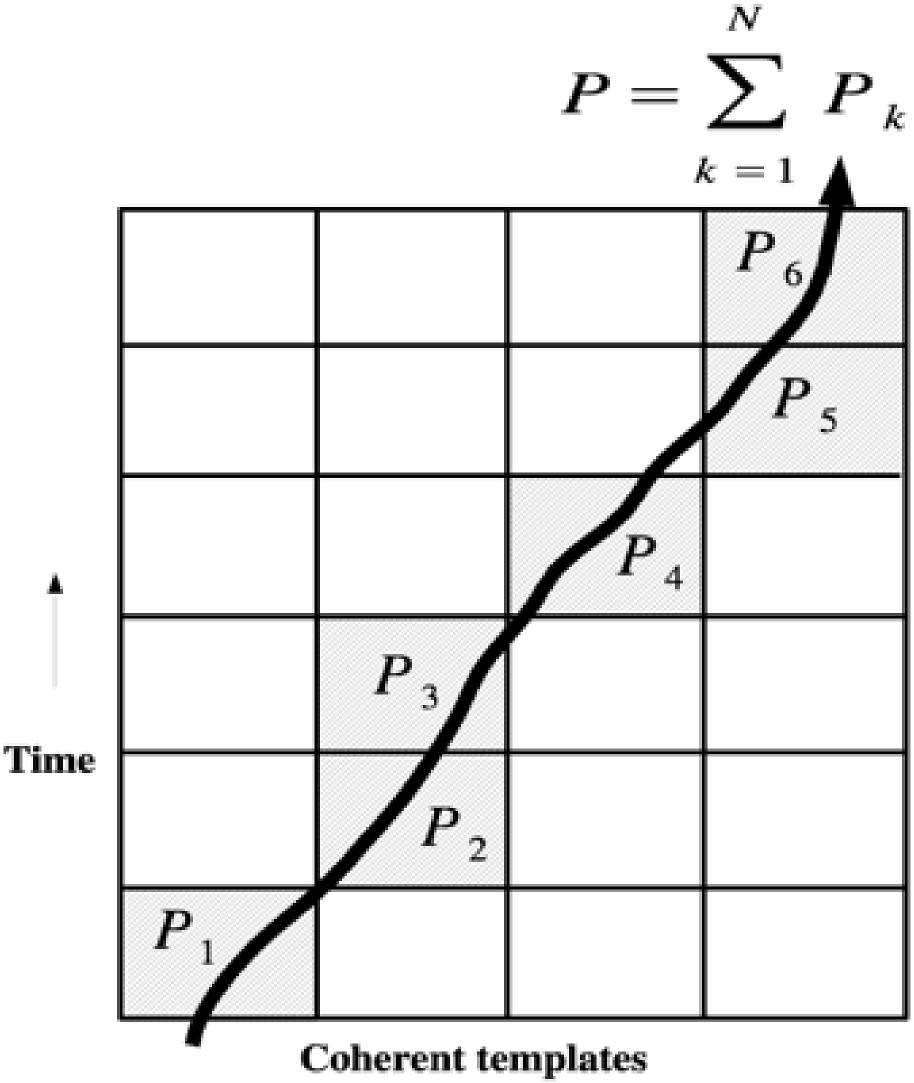


Figure 5: Illustration of the semi-coherent search method. The data is divided into shorter segments, which are searched coherently using waveform templates. The power in the templates is then summed incoherently along trajectories through the templates that correspond to EMRI inspiral trajectories. Reproduced from Gair et al. (2005).

- Astone, P., Colla, A., D?Antonio, S., Frasca, S. and Palomba, C. (2014), *Phys. Rev.* D **90**, 042002
- Antonucci, F., Astone, P., D?Antonio, S., Frasca, S. and Palomba, C. (2008), *Class. Quantum Grav.* **25**, 184015
- Krishnan, B., Sintes, A.M., Papa, M.A., Schutz, B.F., Frasca, S. and Palomba, C. (2004), *Phys. Rev.* D **70**, 082001

4.6 Searches for stochastic backgrounds

Stochastic backgrounds require different search techniques again. It is difficult to identify a background in a single detector, as it is essentially a noise source which is therefore challenging to distinguish from instrumental noise. Instead, background searches make use of multiple detectors and cross-correlate them to identify the common component of the noise. A typical detection statistic takes the form

$$\begin{aligned} Y_Q &= \int_0^T \mathrm{d}t_1 \int_0^T \mathrm{d}t_2 \; s_1(t_1) Q(t_1 - t_2) s_2(t_2) \\ &= \int_{-\infty}^{\infty} \mathrm{d}f \int_{-\infty}^{\infty} \mathrm{d}f' \, \delta_T(f - f') \tilde{s}_1^*(f) \tilde{Q}(f') \tilde{s}_2(f'). \end{aligned} \tag{65}$$

In the above, $Q(t)$ is a filter, which is analogous to the filter introduced in the single source detection case discussed earlier. The function $\delta_T(f)$ is a finite time approximation to the Dirac delta function

$$\delta_T(f) = \int_{-T/2}^{T/2} \mathrm{e}^{-2\pi i f t} \mathrm{d}t = \frac{\sin(\pi f T)}{\pi f}.$$

A generic gravitational wave background can be decomposed into a superposition of plane waves and a sum over polarisation states

$$h_{ij}(t, \vec{x}) = \int_{-\infty}^{\infty} \mathrm{d}f \int_{S^2} \mathrm{d}^{\Omega}_{\hat{k}} \mathrm{e}^{2\pi i f (t - \hat{k}\cdot\vec{x})} \mathcal{H}_A(f, \hat{k}) \mathbf{e}^A_{ij}(\hat{k}).$$

Here A labels the polarisation state, which for gravitational waves in general relativity is either plus or cross, $A = \{+, \times\}$, but in general metric theories could also include scalar and vector modes. The quantities $\mathbf{e}^A_{ij}(\hat{k})$ are the polarisation basis tensors for the individual polarisation modes

$$\mathbf{e}^+_{ij}(\hat{k}) = \hat{l}_i \hat{l}_j - \hat{m}_i \hat{m}_j, \qquad \mathbf{e}^\times_{ij}(\hat{k}) = \hat{l}_i \hat{m}_j + \hat{m}_i \hat{l}_j$$

where

$$\begin{aligned} \hat{k} &= \sin\theta \cos\phi \, \hat{x} + \sin\theta \sin\phi \, \hat{y} + \cos\theta \, \hat{z} \\ \hat{l} &= \cos\theta \cos\phi \, \hat{x} + \cos\theta \sin\phi \, \hat{y} - \sin\theta \, \hat{z} \\ \hat{m} &= -\sin\phi \, \hat{x} + \cos\phi \, \hat{y} \end{aligned} \tag{66}$$

are the standard spherical-polar coordinate basis vectors on the sky at colatitude θ and longitude ϕ. The quantities $\mathcal{H}^A(f, \hat{k})$ are the amplitudes of the various modes. For an unpolarised, stationary and statistically isotropic gravitational wave background, the expectation value of pairs of these amplitudes is given by

$$\left\langle \mathcal{H}^A(f, \hat{k}) \mathcal{H}^{A'*}(f', \hat{k}') \right\rangle = H(f) \delta(f - f') \delta^2(\hat{k}, \hat{k}') \delta_{AA'}, \tag{67}$$

where $H(f)$ is a real-valued function that depends on the energy density in the gravitational wave background and can be related to $\Omega_{\rm GW}(f)$, as introduced in the previous chapter, by

$$H(f) = \frac{3H_0^2}{32\pi^3}\frac{\Omega_{\rm GW}(f)}{|f|^3}.$$

The response of a particular gravitational wave detector, labelled by I, to a gravitational wave field can be written in the form

$$\begin{aligned} s_I(t) &= \int_{-\infty}^{\infty} {\rm d}\tau \int_{R^3} {\rm d}^3\vec{y}\, h_{ij}(t-\tau, \vec{x}-\vec{y}) R_I^{ij}(\tau, \vec{y}) \\ &= (2\pi)^3 \int_{-\infty}^{\infty} {\rm d}f \int_{R^3} {\rm d}^3\vec{k}\, \tilde{h}_{ij}(f, \vec{k}) \tilde{R}_I^{ij}(f, \vec{k}) {\rm e}^{i(2\pi f t - \vec{k}\cdot\vec{x}_I)} \end{aligned} \tag{68}$$

where $R^{ij}(t, \vec{x})$ is the impulse response of the detector, and the integral is over the spatial extent of the detector. Combining Eq. (68) with Eq. (67) we obtain

$$\langle Y_Q \rangle = \frac{T}{2}\int_{-\infty}^{\infty} \gamma_{12}(|f|)\tilde{Q}(f)H(f){\rm d}f$$

where $\gamma(|f|)$ is the **overlap reduction function**, which depends on the relative separation and orientation of the two detectors and is defined by

$$\gamma_{12}(|f|) = \int_{S^2} {\rm d}\Omega_{\hat{k}}\, \tilde{R}_1^A(f, \hat{k}) \tilde{R}_2^{A*}(f, \hat{k}) {\rm e}^{-2\pi i f \hat{k}\cdot(\vec{x}_1 - \vec{x}_2)}$$

where

$$\tilde{R}_I^A(f, \hat{k}) = (2\pi)^e \mathbf{e}_{ij}^A(\hat{k}) \tilde{R}_I^{ij}(f, 2\pi f \hat{k}).$$

The overlap reduction function for various combinations of ground-based interferometers and resonant bar detectors is shown in Figure 6. Stochastic backgrounds generated by large numbers of supermassive black hole binary inspirals are also the primary source for pulsar timing arrays. In that case, the "detector" is the measured redshift of a pulsar. The overlap reduction function for the detection of an isotropic stochastic background by cross-correlation of the measured redshifts of two different pulsars must be a function of only the angular separation between the pulsars on the sky. The resulting overlap reduction function curve is called the Hellings and Downs curve and is shown in Figure 7. Overlap reduction functions for non-isotropic backgrounds, for example anisotropic or correlated backgrounds, of backgrounds with non-GR polarisations, look different, providing a diagnostic for these physical properties of any observed stochastic background.

As in the case of the optimal filter, it is possible to maximise the signal-to-noise ratio of the filtered output. This takes a similar form to the optimal filter result

$$\tilde{Q}(f) \propto \frac{\gamma(|f|)\Omega_{\rm GW}(|f|)}{|f|^3 S_1(|f|) S_2(|f|)}$$

where $S_1(|f|)$ and $S_2(|f|)$ are the power spectral densities of the noise in the two detectors.

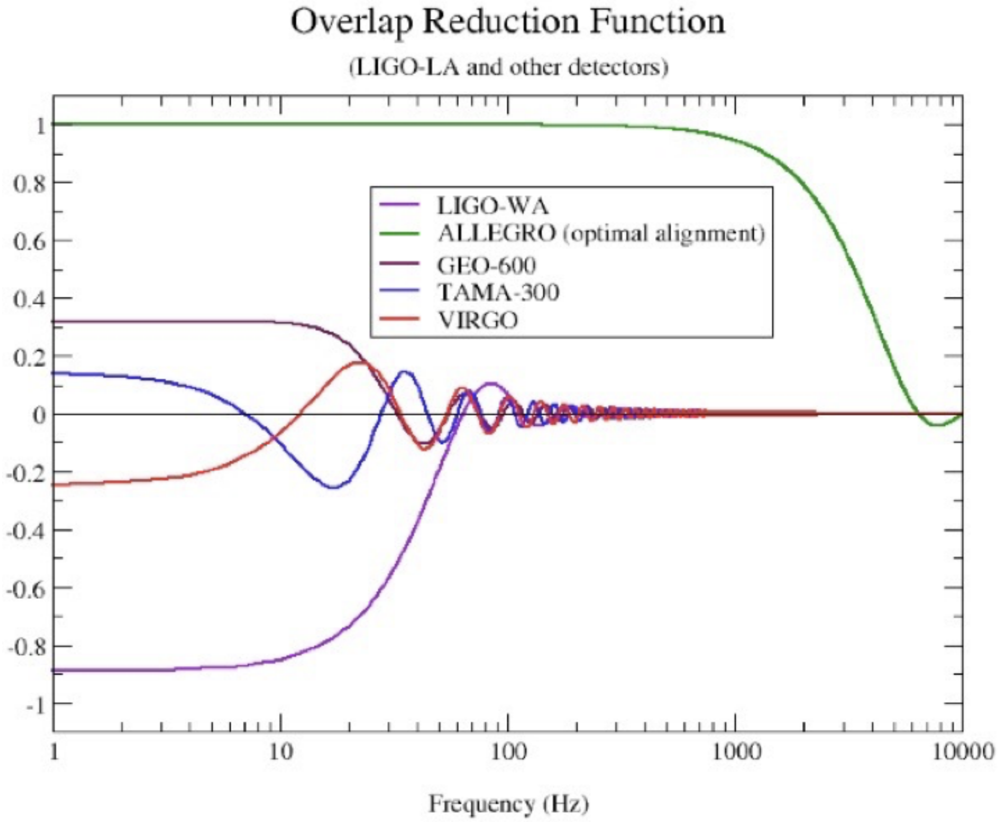


Figure 6: Overlap reduction function of the LIGO Livingston detector with LIGO Hanford (lower purple curve), Virgo (red curve), GEO (upper purple curve), TAMA (now obsolete) (blue curve) and the resonant bar detector Allegro (green curve), which was also sited in Louisiana. This was the network of detectors operating at the time of initial LIGO's science runs.

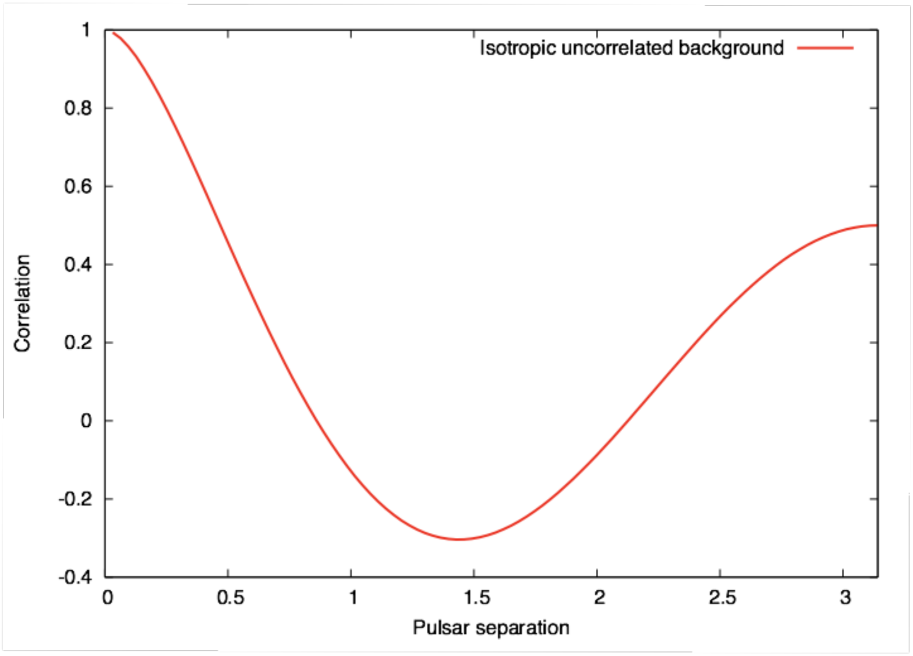


Figure 7: Overlap reduction function for the cross-correlation of the redshifts of two pulsars observed in a pulsar timing array, as a function of the angular separation of the two pulsars on the sky. This is known as the Hellings and Downs curve and the observation of a cross-correlation pattern that matches with this expectation is critical for the pulsar timing detection of gravitational waves.

5 Stochastic processes and sensitivity curves

In both frequentist and Bayesian approaches to statistical analysis, the likelihood plays a key role. This is the probability distribution from which the observed data has been drawn. In a gravitational wave context, we are typically concerned with analysing data from a noisy detector. The output from the detector, or detectors, is one or more real time series of measurements, $s_i(t)$. These measurements are a combination (usually assumed to be linear) of a signal part, $h_i(t)$, and a noise part, $n_i(t)$. The signal part is deterministic, depending only on the (unknown) parameters of the system, while the noise part is random. The likelihood is therefore a statement about the probability distribution from which the noise is drawn. The usual assumption is that the noise is generated by a **stationary**, **Gaussian random process**. In this section we will first define what this means, and discuss various approaches that are commonly used to summarise the noise properties and represent sensitivities to sources of different types.

5.1 Properties of random processes

A random process is a random sequence (often infinite in length) of values. Future values are not uniquely determined by current values, but by probability distributions that may be conditional on past values of the sequence. The observed random sequence is assumed to be drawn from *an ensemble of random processes* characterised by probability distributions

$$p_N(n_N, t_N; n_{N-1}, t_{N-1}; \ldots; n_2, t_2; n_1; t_1)\mathrm{d}n_N \mathrm{d}n_{N-1} \ldots \mathrm{d}n_2 \mathrm{d}n_1.$$

The probability distribution could be anything, but it is usual to make some simplifying assumptions, which are well motivated by observed random processes, to make computations plausible. The most commonly made assumptions are that the random process is **stationary**, **Gaussian** and **ergodic**.

A **stationary** random process is one for which the joint probability distributions for finite sets of samples depend only on time differences, not absolute time. In other words

$$p_N(n_N, t_N + \tau; \ldots; n_2, t_2 + \tau; n_1; t_1 + \tau) = p_N(n_N, t_N; \ldots; n_2, t_2; n_1; t_1) \; \forall \, \tau.$$

A random process is **Gaussian** if and only if all of its absolute probability distributions are Gaussian. In other words, for any set of N times, $\{t_1, \ldots, t_N\}$, we have

$$p_N(n_N, t_N; \ldots n_1; t_1) = A \exp\left[-\frac{1}{2} \sum_{j=1}^{N} \sum_{k=1}^{N} \alpha_{jk} (n_j - \bar{n}_j)(n_k - \bar{n}_k) \right].$$

A ensemble of random process is **ergodic** if for any process, $n(t)$, drawn from the ensemble, the new ensemble defined by $\{n(t+KT) : \; k \in \mathbb{Z}\}$ has the same probability distributions.

To understand random processes, we are interested in both their mean values and the size of random fluctuations about the mean. We assume in the following (without loss of generality) that the mean of the random process is zero. Fluctuations about the mean can be characterised by the noise power (or variance), over a certain time interval $-T/2 < t < T/2$

$$\int_{-T/2}^{T/2} |n(t)|^2 \mathrm{d}t.$$

This quantity increase with time, linearly for stationary random processes. Therefore, it is more useful to work with the average value, referred to as the **mean power** or **mean square fluctuations**

$$P_n = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} |n(t)|^2 \mathrm{d}t.$$

It is useful to analyse quantities in the Fourier domain, and so we define

$$n_T(t) = n(t)\mathbb{I}\left[|t| < T/2\right],$$

which is just the full series truncated to the interval of interest. This notation allows us to use Parseval's Theorem

$$\int_{-T/2}^{T/2} [n(t)]^2 \mathrm{d}t = \int_{-\infty}^{\infty} [n_T(t)]^2 = \int_{-\infty}^{\infty} |\tilde{n}_T(f)|^2 \mathrm{d}f = 2\int_{0}^{\infty} |\tilde{n}_T(f)|^2 \mathrm{d}f$$

and we see that the mean square fluctuations are given by

$$P_n = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} [n(t)]^2 = \lim_{T\to\infty} \frac{2}{T} \int_{0}^{\infty} |\tilde{n}_T(f)|^2 \mathrm{d}f.$$

This motivates the definition of the spectral density, $S_n^{2s}(f)$, via

$$S_n^{2s}(f) = \lim_{T\to\infty} \frac{1}{T} \left| \int_{-T/2}^{T/2} n(t) \exp(2\pi i f t) \mathrm{d}t \right|^2 .$$

This is the **two-sided spectral density**, which is defined for both positive and negative frequencies. It encapsulates the average power in the time series at frequency f. If you have a real time series, then $n^*(t) = n(t)$ and $\tilde{n}(-f) = \tilde{n}^*(f)$, so $S_n^{2s}(f) = S_n^{2s}(-f)$. The negative frequency components are required to reconstruct the time series, but they can be written down using the positive frequency components and so they are often discarded for efficiency of storage. Doing so throws away half of the power in the Fourier domain, and so for real time series we define the **one-sided spectral density**, as $S_n(f) = 2S_n^{2s}(f)$, i.e., double the two-sided spectral density defined by the previous equation. This definition ensures that when the one-sided spectral density of a real time series is plotted for positive frequencies, all of the power in the signal is represented. However, this definition also simplifies some computations, which we will discuss below.

The spectral density represents the power in the process at a particular frequency since we have

$$P_n = \int_{0}^{\infty} S_n(f) \mathrm{d}f.$$

Suppose we are interested in the properties of the process in time intervals of length Δt, with corresponding **bandwidth** $\Delta f = 1/\Delta t$. The mean square fluctuations at frequency f in intervals of length Δt, and averaged over all intervals of that length, are

$$[\Delta n(\Delta t, f)]^2 \equiv \lim_{N\to\infty} \frac{2}{N} \sum_{n=-N/2}^{N/2} \left| \frac{1}{\Delta t} \int_{n\Delta t}^{(n+1)\Delta t} n(t) \exp(2\pi i f t) \mathrm{d}t \right|^2 = \frac{S_n(f)}{\Delta t} = S_n(f)\Delta f.$$

Hence we see that the *root mean square fluctuations at frequency f and measured over a time* Δt are just $\Delta n(\Delta t, f)_{\mathrm{rms}} = \sqrt{S_n(f)\Delta f}$. The spectral density can be interpreted in this way as the size of mean square fluctuations at the specified frequency.

A property of a random process that is closely linked to the spectral density is the **auto-correlation function**. This is defined in the standard way

$$C(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} n(t)n(t+\tau)\mathrm{d}t.$$

For random processes that are ergodic (which implies they are also stationary), the averaging over time is equivalent to averaging over the ensemble

$$C(\tau) = \langle\, n(t)\, n(t+\tau)\, \rangle.$$

The auto-correlation function is the Fourier transform of the spectral density (the Wiener-Khinchin Theorem). A consequence of this result is that the expectation value of noise products can be written

$$\langle \tilde{n}^*(f)\tilde{n}(f')\rangle = \frac{1}{2}S_n(f)\delta(f-f').$$

which is a statement that fluctuations of a stationary random process at different frequencies are uncorrelated with one another. Note that the spectral density appearing here is the one-sided spectral density, and the factor of 1/2 corrects for the additional factor of 2 that appears in the definition of that quantity. This equation tells us that the spectral-density is related to the variance of the noise in the frequency domain, but the awkward factor of 1/2 means that $S_n(f)$ is twice the variance, while the two-sided spectral density would equal the variance. However, the positive and negative frequency components of the Fourier transform of a real time series are perfectly correlated. When writing down the probability distribution we should therefore only include positive frequency components, weighted by their actual variance. But, it is often more convenient to work with sums over the full frequency range. Doing this we essentially double the answer, since the negative frequency components give exactly the same result as the positive frequencies. We can compensate by dividing by twice the variance, i.e., the one-sided spectral density. So, using the one-sided spectral density with a sum over all frequencies gives the correct answer. Using the two-sided spectral density with a sum over positive frequencies only would also give the correct answer.

The spectral densities of a number of common noise processes are as follows

$$\begin{array}{ll} \textit{white noise spectrum} & S_n(f) = \mathrm{const.} \\ \textit{flicker noise spectrum} & S_n(f) \propto 1/f \\ \textit{random walk spectrum} & S_n(f) \propto 1/f^2 \end{array} .$$

We conclude this section by noting that it is also possible to define a **cross-spectral density** between two separate random processes. This is defined via

$$S_{nm}(f) = \lim_{T\to\infty} \frac{2}{T} \left[\int_{-T/2}^{T/2} n(t)\exp(-2\pi i f t)\mathrm{d}t\right]\left[\int_{-T/2}^{T/2} m(t)\exp(2\pi i f t')\mathrm{d}t'\right]$$

and is the related via Fourier transform to the cross-correlation function of the two time series

$$C_{nm}(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} n(t)m(t+\tau)\mathrm{d}t.$$

5.2 Sensitivity curves

For a Gaussian, stationary random process the spectral density conveys all the information about the statistical properties of the process. For gravitational wave detectors, it is therefore natural to plot the spectral density to characterise the detector sensitivity. But - how should sources be presented on the same diagram? There is no unique way to do this. Different types of source are best represented in different ways.

5.2.1 Burst signals

Burst signals are by definition compact in time duration, and usually also in frequency duration. It is rare that burst signals can be represented by parametric models, and so they are quite like random processes. We can characterise the burst by its frequency, f, duration, Δt, bandwidth, Δf, and its mean square amplitude, a proxy for the signal power

$$\bar{P_h} = \frac{1}{\Delta t}\int_0^{\Delta t} |h_+(t)|^2 + |h_\times(t)|^2 \mathrm{d}t = h_c^2.$$

The square root of the mean square amplitude, h_c, is called the **characteristic amplitude** of the burst. The power of the noise in the same bandwidth is $\Delta f S_n(f)$. The ratio of the power in the signal to the power in the noise is a measure of the detectability of the burst, relative to random fluctuations in the instrument. This ratio is the **signal-to-noise ratio** squared of the burst

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right)^2 = \frac{\bar{P_h}}{\Delta f S_h(f)} = \frac{h_c^2}{\Delta f S_h(f)}.$$

If the data is windowed and bandpassed in the vicinity of the burst, then we maximise the contribution of the burst to the data and the signal-to-noise ratio is the ratio of the root-mean-square (rms) signal contribution to the rms noise contribution. For a broad-band burst with $\Delta f \sim f$ we have

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right)^2 = \frac{h_c^2}{f S_h(f)}. \tag{69}$$

This motivates representing the sensitivity of a detector to bursts by plotting the quantity $f S_h(f)$ instead of the power spectral density. The detectability of a burst source with characteristic strain h_c can then be assessed by the height of h_c^2 above the curve. Note that because GWs have two polarizations, if h_c is defined from one polarization only (or by averaging, then the total SNR squared will be double as it will get a contribution from both polarization states.

5.2.2 Continuous sources

If instead of a burst we had an (optimally-oriented) monochromatic gravitational wave source

$$h_+(t) = h_0 \cos(2\pi f_0 t), \qquad h_\times(t) = \sin(2\pi f_0 t),$$

then the signal power is constant over time

$$P_h = \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} |h_+(t)|^2 + |h_\times(t)|^2 \mathrm{d}t = h_0^2.$$

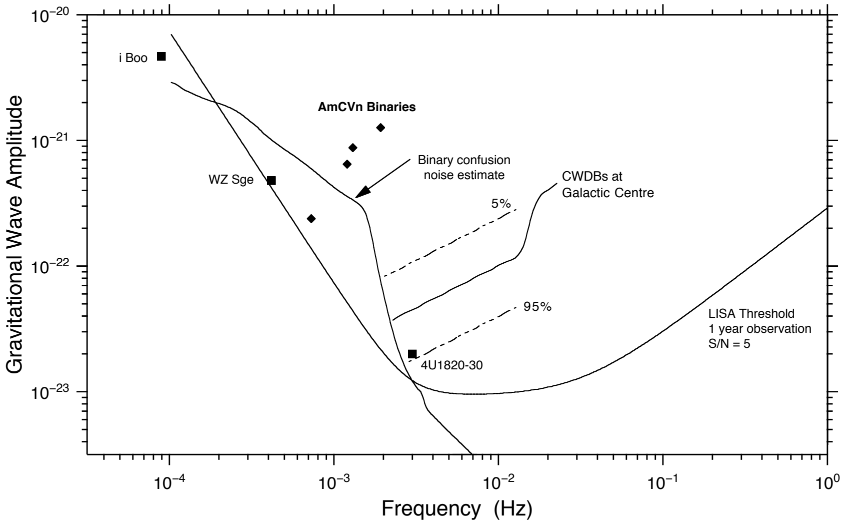


Figure 8: Strain spectral density curve for a 1 year observation with LISA and a detection threshold of $S/N = 5$. Reproduced from the LISA pre-phase A report.

This power is concentrated at f_0. When observing a finite time series of length T, we can resolve frequencies to a precision $\Delta f \sim 1/T$. The noise power in that bandwidth is $S_n(f)/T$, which motivates representing the detector sensitivity curve by plotting

$$\sqrt{S_n(f)/T} \quad \text{or} \quad \rho_{\text{thresh}}\sqrt{S_n(f)/T}$$

where ρ_{thresh} is the threshold signal-to–noise ratio needed for detection. This is called the **strain spectral density**. The advantage of rep[resenting sensitivity in this way is that the detectability of a source can be directly assessed by seeing if the source amplitude h_0 lies above or below the curve. The height above the curve is a direct estimate of the signal to noise ratio of the source. The disadvantage of this way of representing sensitivity is that it varies with the length of observation, so this must be specified. In the case of LIGO, this is not a problem, as the detectors take periodic breaks from observation. After each observing run, the length of observation is known and so the strain spectral density can be evaluated for each observing run after the fact, and used to represent the results.

An example of a strain spectral density curve is given in Fig. 8.

Finally, we note that rescaling the sensitivity according to the detection threshold is not the only type of rescaled spectral density that is encountered in the literature. The amplitude of a gravitational wave signal in a gravitational wave detector depends on the orientation of the source relative to the detector plane. The same source placed at different sky locations and orientations will have different signal-to-noise ratios. To avoid having to specify which particular choices are being made, it is useful to produce a **sky-averaged sensitivity curve**. To assess detectability of a source, its amplitude should then be assessed for optimal orientation and sky location. The height of this optimal source above the sky averaged sensitivity is the average signal-to-noise ratio squared of a source of this type. For LIGO the sky averaged sensitivity is

$$\langle S_h(f)\rangle_{\text{SA}}^{LIGO} \approx 5S_h(f)$$

while for LISA we have

$$\langle S_h(f)\rangle_{\mathrm{SA}}^{LISA} \approx \frac{20}{3} S_h(f).$$

The difference arises because of the 60° opening angle of the LISA arms ($\sin^2 60 = 3/4$).

5.2.3 Inspiralling sources

Inspiraling sources have to be treated differently to continuous sources. This is because they emit a finite amount of power in each frequency band and hence the Fourier transform at each frequency is also finite. Therefore

$$\frac{1}{\sqrt{T}}\tilde{h}(f) \Rightarrow 0 \quad \text{as} \quad T \to \infty$$

and so the strain spectral density of an inspiraling source is zero averaged over all time. Band-passing and windowing the data can recover some signal-to-noise ratio, as in the burst source case, but we can do better than that using **filtering**.

A filtered time series is defined from a kernel $K(t-t')$ via convolution

$$w(t) = \int_{-\infty}^{\infty} K(t-t')s(t')\mathrm{d}t'.$$

In the previous cases we considered signal-to-noise ratio as the ratio of the rms power in the presence of a signal to the rms power in the noise. We use an analogous definition for filtered data, but now compare the amplitude of the filter output due to the signal only, to the rms amplitude of the filtered data in the presence of noise only

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right)(t) = \frac{\int_{-\infty}^{\infty} K(t-t')h(t')\mathrm{d}t'}{\sqrt{\left\langle \left|\int_{-\infty}^{\infty} K(t-t')n(t')\mathrm{d}t'\right|^2 \right\rangle}}.$$

The rms output of the filter is the signal amplitude, S, to within a fractional error of N/S, which is the reciprocal of the signal-to-noise ratio.

The choice of the kernel is arbitrary, but it makes sense to choose the kernel that is "best" in some sense. The best kernel is the one that maximises the signal-to-noise ratio. This is most easily found by working in the Fourier domain. We use the Fourier transform definition

$$\tilde{x}(f) = \int_{-\infty}^{\infty} x(t)\exp\left[-2\pi i f t\right]\,\mathrm{d}t.$$

From the convolution theorem, the Fourier transform of the filter output is

$$\tilde{w}(f) = \tilde{K}(f)\tilde{h}(f)$$

where $\tilde{K}(f)$ and $\tilde{h}(f)$ are the Fourier transform of the kernel and waveform respectively. We have also

$$w(t) = \int_{-\infty}^{\infty} \tilde{w}(f)\exp\left[2\pi i f t\right]\,\mathrm{d}f \quad \Rightarrow \quad w(0) = \int_{-\infty}^{\infty} \tilde{w}(f)\,\mathrm{d}f.$$

Similarly

$$\begin{aligned}
N^2(0) &= \left\langle \left| \int_{-\infty}^{\infty} K(-t')n(t')\mathrm{d}t' \right|^2 \right\rangle = 4\left\langle \int_0^{\infty} \tilde{K}(f)\tilde{n}(f)\mathrm{d}f \int_0^{\infty} \tilde{K}^*(f')\tilde{n}^*(f')\mathrm{d}f' \right\rangle \\
&= 4\int_0^{\infty}\int_0^{\infty} \tilde{K}(f)\tilde{K}^*(f')\left\langle \tilde{n}^*(f')\tilde{n}(f)\right\rangle \mathrm{d}f\mathrm{d}f' \\
&= 4\int_0^{\infty}\int_0^{\infty} \tilde{K}(f)\tilde{K}^*(f')(1/2)\delta(f-f')S_h(f)\mathrm{d}f\mathrm{d}f' \\
&= \int_{-\infty}^{\infty} |\tilde{K}(f')|^2 S_h(f')\mathrm{d}f'. \qquad (70)
\end{aligned}$$

We deduce that the signal-to-noise ratio at zero lag is

$$\frac{\mathrm{S}}{\mathrm{N}} = \frac{\int \tilde{K}(f)\tilde{h}(f)\mathrm{d}f}{\sqrt{\int |\tilde{K}(f')|^2 S_h(f')\mathrm{d}f'}}$$

which can also be written as

$$\frac{\mathrm{S}}{\mathrm{N}} = \frac{(S_h K|h)}{\sqrt{(S_h K|S_h K)}}$$

by introducing the noise-weighted inner product

$$(\mathbf{h}_1|\mathbf{h}_2) = 2\int_0^{\infty} \frac{\tilde{\mathbf{h}}_1(f)\tilde{\mathbf{h}_2}^*(f) + \tilde{\mathbf{h}_1}^*(f)\tilde{\mathbf{h}}_2(f)}{S_h(f)}\,\mathrm{d}f.$$

This is of the form $\hat{\mathbf{e}}\cdot\mathbf{b}$, for a unit vector $\hat{\mathbf{e}}$ to be found. The inner product of two vectors of fixed length is maximised when they are parallel, i.e., $\hat{\mathbf{e}} \propto \mathbf{b}$. We therefore deduce that the choice which maximises the inner product is

$$S_h(f)\tilde{K}(f) \propto \tilde{h}(f) \quad \Rightarrow \quad \tilde{K}(f) \propto \frac{\tilde{h}(f)}{S_h(f)}.$$

This is the **Weiner optimal filter**. In the frequency domain the optimal filter is equal to the signal, weighted by the spectral density of the noise. A search using the optimal filter amounts to taking the inner product $(\mathbf{s}|\mathbf{h})$ of the data stream, $\mathbf{s}$, with a template of the signal, $\mathbf{h}$. This is **matched filtering**. In practice we don't know exactly what the signal is, but the parameters of the signal must be estimated from the data. In LIGO/Virgo this is done by computing the output of the optimal filter for a large number of source parameter choices which define a **template bank**. This was described in more detail in Chapter 4.

The signal-to-noise ratio of the matched filtering search that uses the optimal filter is

$$\frac{\mathrm{S}}{\mathrm{N}}[\mathbf{h}] = \frac{(\mathbf{h}|\mathbf{h})}{\sqrt{\langle(\mathbf{h}|\mathbf{n})(\mathbf{h}|\mathbf{n})\rangle}} = (\mathbf{h}|\mathbf{h})^{1/2}$$

which follows from the fact that

$$\langle(\mathbf{h}_1|\mathbf{n})(\mathbf{h}_2|\mathbf{n})\rangle = (\mathbf{h}_1|\mathbf{h}_2). \qquad (71)$$

This result is proved as follows

$$\begin{aligned}
\langle(\mathbf{h}_1|\mathbf{n})(\mathbf{h}_2|\mathbf{n})\rangle &= 4\Big\langle \int_0^\infty \frac{\tilde{\mathbf{h}}_1(f)\tilde{\mathbf{n}}^*(f)+\tilde{\mathbf{h}}_1^*(f)\tilde{\mathbf{n}}(f)}{S_h(f)}\mathrm{d}f \int_0^\infty \frac{\tilde{\mathbf{h}}_2(f')\tilde{\mathbf{n}}^*(f')+\tilde{\mathbf{h}}_2^*(f')\tilde{\mathbf{n}}(f')}{S_h(f')}\mathrm{d}f'\Big\rangle \\
&= 4\int_0^\infty\int_0^\infty \frac{\tilde{\mathbf{h}}_1(f)\tilde{\mathbf{h}}_2^*(f')\langle\tilde{\mathbf{n}}^*(f)\tilde{\mathbf{n}}(f')\rangle+\tilde{\mathbf{h}}_1^*(f)\tilde{\mathbf{h}}_2(f')\langle\tilde{\mathbf{n}}(f)\tilde{\mathbf{n}}^*(f')\rangle}{S_h(f)S_h(f')}\mathrm{d}f\mathrm{d}f' \\
&\quad + 4\int_0^\infty\int_0^\infty \frac{\tilde{\mathbf{h}}_1(f)\tilde{\mathbf{h}}_2(f')\langle\tilde{\mathbf{n}}^*(f)\tilde{\mathbf{n}}^*(f')\rangle+\tilde{\mathbf{h}}_1^*(f)\tilde{\mathbf{h}}_2^*(f')\langle\tilde{\mathbf{n}}(f)\tilde{\mathbf{n}}(f')\rangle}{S_h(f)S_h(f')}\mathrm{d}f\mathrm{d}f'.
\end{aligned} \tag{72}$$

The terms on the final line vanish because $\langle\tilde{\mathbf{n}}(f)\tilde{\mathbf{n}}(f')\rangle = 0$, i.e., the size of fluctuations in the real and imaginary components of the noise are the same. The terms on the first line are simplified using $\langle\tilde{\mathbf{n}}^*(f)\tilde{\mathbf{n}}(f')\rangle = (1/2)S_h(f)\delta(f-f')$

$$\begin{aligned}
\langle(\mathbf{h}_1|\mathbf{n})(\mathbf{h}_2|\mathbf{n})\rangle &= 4\int_0^\infty\int_0^\infty \frac{[\tilde{\mathbf{h}}_1(f)\tilde{\mathbf{h}}_2^*(f')+\tilde{\mathbf{h}}_1^*(f)\tilde{\mathbf{h}}_2(f')]\delta(f-f')}{2S_h(f')}\mathrm{d}f\mathrm{d}f' \\
&= 2\int_0^\infty \frac{[\tilde{\mathbf{h}}_1(f)\tilde{\mathbf{h}}_2^*(f)+\tilde{\mathbf{h}}_1^*(f)\tilde{\mathbf{h}}_2(f)]}{S_h(f)}\mathrm{d}f,
\end{aligned} \tag{73}$$

giving the result stated above.

The matched filtering signal-to-noise ratio simplifies to

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right)^2 = 4\int_0^\infty \frac{|\tilde{h}(f)|^2}{S_h(f)}\mathrm{d}f$$

which can also be written as

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right)^2 = 4\int_0^\infty \frac{f|\tilde{h}(f)|^2}{S_h(f)}\mathrm{d}\ln f = 4\int_0^\infty \frac{f^2|\tilde{h}(f)|^2}{fS_h(f)}\mathrm{d}\ln f. \tag{74}$$

Plotting $S_h(f)$ and $f|\tilde{h}(f)|^2$ on a logarithmic frequency plot, the integral of the ratio of the two curves "by eye" gives an estimate of the signal-to-noise ratio squared.

For a source that has amplitude h_0 at frequency f, at which point the frequency derivative is $\dot{f}$, then the stationary phase approximation gives us the scaling

$$\tilde{h}(f) \sim \frac{h_0}{2\sqrt{\dot{f}}}.$$

The analogy with the broad-band burst case described above motivates defining a characteristic strain, h_c, such that the signal-to-noise ratio squared contributed at each frequency, by each polarization of the GW, is $h_c^2/(2fS_h(f))$. The appropriate definition is

$$h_c = h_0\sqrt{\frac{2f^2}{\mathrm{d}f/\mathrm{d}t}} \sim f\tilde{h}(f).$$

Note also that since $\sqrt{\dot{E}} \sim \dddot{I}$ and $h \sim \ddot{I}/D$, we have $\sqrt{\dot{E}} \sim Dfh$ and hence

$$h_c \sim \frac{1}{D}\sqrt{\frac{\dot{E}}{\dot{f}}}$$

and this is an equality for monochromatic signals.

The characteristic strain is a measure of the signal-to-noise ratio accumulated while the frequency sweeps through a bandwidth equal to frequency. If we plot as a sensitivity curve the rms noise in a bandwidth equal to frequency, which is

$$h_n(f) \equiv \sqrt{f \langle S_h(f) \rangle_{\rm SA} \rangle}$$

then the signal-to-noise ratio squared accumulated as the inspiral proceeds from f to $2f$, summed over both polarization modes, is

$$\left(\frac{\rm S}{\rm N}\right)^2_{f \to 2f} = \left[\frac{h_c(f)}{h_n(f)}\right]^2.$$

Therefore, plotting characteristic strain on the same plot gives a quick way to see how the signal-to-noise ratio of an inspiraling source builds up over the evolution. Note that plotting the characteristic strain only makes sense if the detector sensitivity is represented by $fS_h(f)$. If the detector sensitivity is represented by $S_h(f)$ then the quantity $h_c/\sqrt{f}$ should be used to represent the signal.

In the definition of characteristic strain, $h_c = h_0\sqrt{2f^2/\dot{f}}$, the term inside the square root is proportional to the number of cycles the inspiral spends in the vicinity of the frequency f. Papers that discuss matched filtering often include the statement that the signal to noise ratio is enhanced by the number of cycles spent in the vicinity of a certain frequency. This is what they are referring to.

In Fig. 9 we give an example of a plot of the characteristic strain, reproduced from Finn and Thorne (2000). The figure shows the characteristic strain of various extreme-mass-ratio inspiral sources detectable by LISA.

As a final remark, we note that if we consider a monochromatic source and compute the SNR from Eq. (74) contributed by a single polarization state, we get $Th_0^2/S_h(f)$, while treating it as a burst source with $\Delta f = 1/T$ and using Eq. (69) we get $Th_0^2/(2S_h(f))$ from each polarization. This difference arises because in matched filtering we assume we not only know the frequency of the signal but also its phase, while the latter assume that we have only localized in frequency. In matched filtering, we can thus localize the signal to one of the two independent quadratures and hence effectively reduce the variance by an additional factor of 2.

5.2.4 Stochastic backgrounds

Stochastic backgrounds are characterised by a spectral density, so it is natural to compute the power spectral density and plot it on the same axes as the detector PSD. However, there are two caveats. Firstly, the "power" we have been talking about so far has not been a power in a physical sense since we have not specified any units for the time series (and indeed for GW strain this is dimensionless). When comparing to the noise power spectral density which is an energy density, it would be preferable to use something that represents a physical energy density if possible. Secondly, plotting two power spectral densities does not convey any information about their distinguishability. It would be preferable to represent a background in a way that conveys the detectability of the background at a glance.

The energy density carried by a gravitational wave is given by

$$\frac{{\rm d}E}{{\rm d}t{\rm d}A} \propto \dot{h}_+^2 + \dot{h}_\times^2.$$

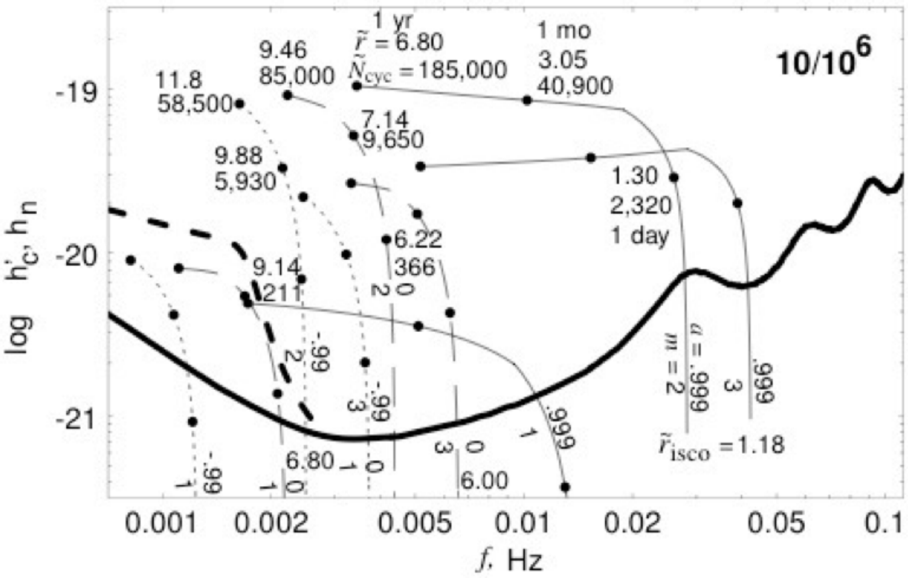


Figure 9: Characteristic strain for a number of typical extreme-mass-ratio inspiral sources observed by the classic (5 km arm length) LISA interferometer. All inspirals are circular, with $10^6 M_\odot$ central black holes and observed at a distance of 1Gpc. Curves are labelled by the spin of the central black hole, a, and the mass of the inspiraling object, m. Points on the curve correspond to 1 year, 1 month and 1 day prior to merger. The numbers above the points are the radius of the orbit (in units of M) at that time, and the number of gravitational wave cycles remaining until plunge. Reproduced from Finn and Thorne (2000).

Therefore, to obtain a physical energy density we should consider the time derivative of the strain. Differentiation with respect to time brings down a factor of frequency and so the energy spectral density is $f^2 S_h(f)$. Fluctuations of the energy spectral density in a bandwidth equal to frequency are then $f^3 S_h(f)$.

The energy density of an astrophysical or cosmological stochastic background, per logarithmic frequency interval, is often expressed as a fraction of the closure density of the Universe via

$$\Omega_{\rm GW} = \frac{8\pi G}{3H_0^2}\frac{{\rm d}E_{\rm GW}}{{\rm d}\ln f} \propto f^2 h_c^2(f).$$

The last equality defines the characteristic strain for a background, since, as argued above, a plane wave of frequency f and amplitude h_c carries an energy density $f h_c$. In the examples below we will show how to calculate the energy density for an astrophysical population of sources.

To represent backgrounds in a way that conveys their detectability directly, one can use *power-law sensitivity curves* (Thrane and Romano 2013). These are not uniquely defined, as they require some assumptions to be made about data analysis procedures and the threshold required for a detection using the defined procedure. However, given these assumptions, the procedure is as follows.

- For a given assumed power-law slope of a background, $\Omega_{\rm GW} \propto f^\beta$, compute the minimum amplitude, $A_{\rm min}(\beta)$, such that the background would be detectable by the defined procedure.

- Define the *power-law sensitivity curve*, $S_{\rm pl}(f)$, via

$$S_{\rm pl}(f) = \max\{A_{\rm min}(\beta) f^\beta : \beta \in [-\infty, \infty]\}.$$

The power-law sensitivity curve is the envelope of the minimal-detectable power-law backgrounds. It is a useful object to assess background detectability, since drawing a background of interest on the same figure gives an immediate indication of detectability. If the curve lies above the power-law sensitivity curve then it will be detectable (via the designated procedure) and otherwise it will not. An illustration of such a curve is given in Fig. 10.

5.3 Examples

We now estimate the leading order scaling of the quantities introduced above for some common astrophysical sources. Throughout we will make the usual choice of units to set $G = c = 1$.

5.3.1 Single inspiraling compact binary

We consider first the case of compact binary coalescence. We assume that we have a circular binary with component masses M_1 and M_2 and separation r. We work in the Newtonian regime where the binary components are on Keplerian orbits. We denote the total mass, M, and reduced mass, μ, by

$$M = M_1 + M_2, \qquad \mu = \frac{M_1 M_2}{M_1 + M2}.$$

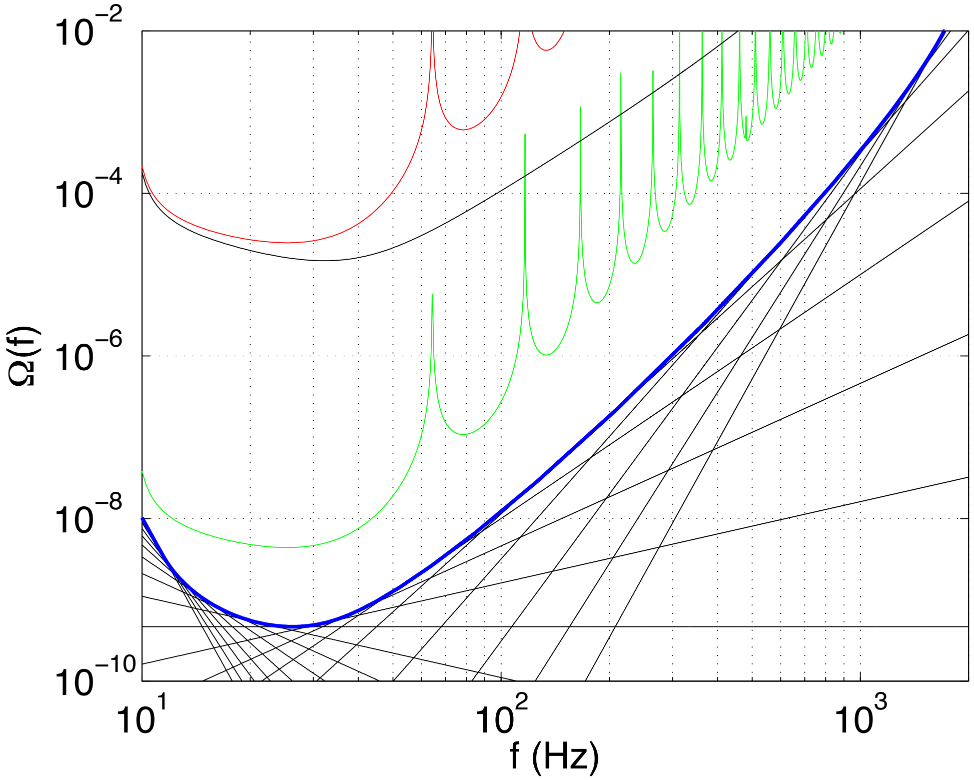


Figure 10: Power law sensitivity curve (in blue) for backgrounds detectable by ground-based interferometers, assuming the search is based on cross-correlation of the H1 and L1 detectors and the threshold for detection is a signal-to-noise ratio of 1. The solid black lines show the set of minimally-detectable power-laws that are used to generate the power-law sensitivity curve. The other curves show the detector strain spectral density instantaneously (red) and for a one year observation (green). Reproduced from Thrane and Romano (2013).

In the Newtonian two-body problem, the two objects each orbit around the centre of mass of the system. The two objects are distances r_1 and r_2 from the centre of mass respectively, where

$$r_1 M_1 = r_2 M_2 = \mu r.$$

The motion is also equivalent to that of a single body of mass μ orbiting in a fixed Newtonian potential of an object with mass M at a distance r. The orbital frequency is given by Kepler's laws

$$\omega^2 = \left(\frac{2\pi}{T}\right)^2 = (2\pi f)^2 = \frac{M}{r^3}.$$

To estimate the scaling of the gravitational wave emission we need to estimate the quadrupole moment of the binary. This can be estimated from

$$I \sim \mu r^2 \cos 2\omega t \sim \frac{M_1 M_2}{(M_1 + M_2)^{\frac{1}{3}}} \omega^{-\frac{4}{3}}.$$

At leading order, the gravitational wave strain scales like the second time derivative of the quadrupole moment divided by the distance to the source

$$h \sim \frac{\ddot{I}}{D} \sim \frac{1}{D} \frac{M_1 M_2}{(M_1 + M_2)^{\frac{1}{3}}} \omega^{\frac{2}{3}}.$$

The rate of energy loss scales like the third time derivative of I squared and so this has the scaling

$$\dot{E} \sim -\dddot{I}^2 \sim -\mu^2 M^{\frac{4}{3}} \omega^{\frac{10}{3}}.$$

Finally, we need to know how the energy relates to the orbital separation or equivalently the orbital frequency. In the Newtonian limit this follows from

$$E = -\frac{M\mu}{2r} = -\frac{\mu (M\omega)^{\frac{2}{3}}}{2}$$

from which we deduce

$$\dot{E} \sim -\mu M^{\frac{2}{3}} \omega^{-\frac{1}{3}} \dot{\omega}. \tag{75}$$

Combining this with expression (5.3.1) we obtain

$$\dot{\omega} \sim \mu M^{\frac{2}{3}} \omega^{\frac{11}{3}} = \frac{M_1 M_2}{(M_1 + M_2)^{\frac{1}{3}}} \omega^{\frac{11}{3}} = M_c^{\frac{5}{3}} \omega^{\frac{11}{3}}$$

where we have introduced the chirp mass

$$M_c = \frac{M_1^{\frac{3}{5}} M_2^{\frac{3}{5}}}{(M_1 + M_2)^{\frac{1}{5}}}.$$

We can now determine the scaling of the various quantities introduced in the previous section. From Eq. (5.2.3) and recalling $\omega = 2\pi f$, we obtain the Fourier domain amplitude

$$\tilde{h}(f) \sim \frac{h_0}{\sqrt{\dot{f}}} \sim \frac{1}{D} \frac{M_c^{\frac{5}{3}} f^{\frac{2}{3}}}{M_c^{\frac{5}{6}} f^{\frac{11}{6}}} = \frac{1}{D} M_c^{\frac{5}{6}} f^{-\frac{7}{6}}.$$

We can also deduce the characteristic strain

$$h_c(f) \sim \frac{1}{D} M_c^{\frac{5}{6}} f^{-\frac{1}{6}}.$$

5.3.2 Eccentric binaries

Eccentric binaries have gravitational wave emission at multiple harmonics of the orbital frequency (Peters and Matthews 1963). The flux of radiation at frequency nf, where n is the orbital frequency, is

$$\dot{E}_n = \frac{32}{5}\mu^2 M^{\frac{4}{3}}(2\pi f)^{\frac{10}{3}} g(n,e)$$

where $g(n,e)$ is given by

$$g(n,e) = \frac{n^4}{32}\left\{\left[J_{n-2}(ne) - 2eJ_{n-1}(ne) + \frac{2}{n}J_n(ne) + 2eJ_{n+1}(ne) - J_{n+2}(ne)\right]^2 \right.$$
$$\left. +(1-e^2)\left[J_{n-2}(ne) - 2J_n(ne) + J_{n+2}(ne)\right]^2 + \frac{4}{3n^2}\left[J_n(ne)\right]^2\right\} \qquad (76)$$

where $J_n(x)$ is the Bessel function of the first kind. The characteristic strain for an individual harmonic is therefore

$$h_{c,n}(f) = \frac{1}{\pi D}\sqrt{\frac{2\dot{E}_n(f/n)}{n\dot{f}(f/n)}} \sim M_c^{\frac{5}{6}} f^{-\frac{7}{6}} n^{\frac{2}{3}}\sqrt{g(n,e)}$$

where the argument (f/n) indicates that in order to get the contribution at frequency f from the n'th harmonic, it must be evaluated when the orbital frequency had the lower value of f/n.

It is normal to represent the contributions form individual waveform harmonics on a "waterfall plot". An example is shown in Figure 11 which is reproduced from Barack and Cutler (2004).

5.3.3 Stochastic backgrounds

The energy density in a gravitational wave background was defined in equation (5.2.4). If this background is generated by a population of individual sources, the total background can be estimated by integrating the contribution from each component in the background. The quantity of relevance is the total energy density in gravitational waves today, $\mathcal{E}_{\rm GW}$. If the sources are identical, have number density $n(z)$ and each generate a differential energy density $\mathrm{d}E/\mathrm{d}f$, then we have

$$\mathcal{E}_{\rm GW} = \int_0^\infty \rho_c c^2 \Omega_{\rm GW} \mathrm{d}\ln f = \int_0^\infty \int_0^\infty N(z) \frac{1}{(1+z)} \frac{\mathrm{d}E}{\mathrm{d}f} f \frac{\mathrm{d}f}{f} \mathrm{d}z,$$

where the factor of $(1+z)$ accounts for the fact that the energy density today is redshifted relative to the energy density at emission. We deduce

$$\rho_c c^2 \Omega_{\rm GW} = \frac{\pi}{4}\frac{c^2}{G} f^2 h_c^2(f) = \int_0^\infty \frac{N(z)}{1+z}\left(f_r \frac{\mathrm{d}E}{\mathrm{d}f_r}\right)_{|f_r = f(1+z)} \mathrm{d}z \qquad (77)$$

where the latter quantity is evaluated at the rest frame frequency, $f_r = (1+z)f$.

For a stochastic background generated by inspiraling binary sources, from Eq. (75), we have at leading order

$$f\frac{\mathrm{d}E}{\mathrm{d}f} \sim M_c^{\frac{5}{3}} f^{\frac{2}{3}}.$$

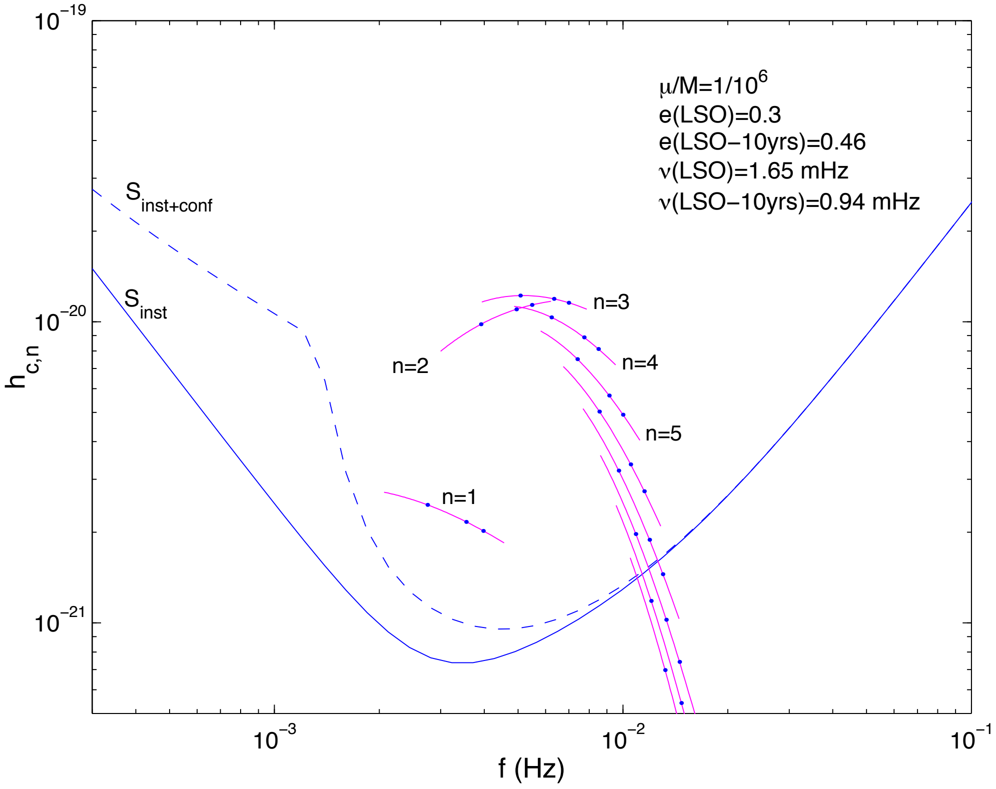


Figure 11: Characteristic strain of each harmonic in a the extreme-mass-ratio inspiral of a $1M_\odot$ black hole into a $10^6M_\odot$ black hole with eccentricity of 0.3 at plunge. Figure reproduced from Barack and Cutler (2004).

Plugging this into Eq. (77) we obtain

$$\Omega_{\mathrm{GW}}(f) \sim M_c^{\frac{5}{3}} f^{\frac{2}{3}} \int_0^{\infty} \frac{N(z)}{(1+z)^{\frac{1}{3}}} \mathrm{d}z. \tag{78}$$

We see that the energy spectral density of the background is

$$S_h(f) \sim \Omega_{\mathrm{GW}}(f)/f^3 \sim M_c^{\frac{5}{3}} f^{-\frac{7}{3}}$$

and the characteristic strain is

$$h_c(f) \sim \sqrt{\Omega_{\mathrm{GW}}(f)}/f \sim M_c^{\frac{5}{6}} f^{-\frac{2}{3}}.$$

In this case the characteristic strain scales like $f^{-2/3}$, while in the case of a single compact binary coalescence we had a scaling that was $f^{-1/6}$. This difference arises because the definition of characteristic strain relates to the signal to noise ratio that can be obtained in a search for the source of interest. For individual sources, we can perform matched filtering and enhance the signal to noise ratio coherently by the square root of the number of cycles (approximately $\sqrt{f}$, which explains the difference between $f^{-2/3}$ and $f^{-1/6}$). This is not possible for incoherent backgrounds where we can only predict the power at each frequency, not the phase.