3 Statistical inference

> If your experiment needs statistics, you ought to have done a better experiment.
> *Ernest Rutherford*

In this section, we review statistical inference from both the Bayesian and frequentist perspectives. Our discussion of frequentist and Bayesian upper limits, and the example given in Sect. 3.5 comparing Bayesian and frequentist analyses is modelled in part after Röver et al. (2011). Readers interested in more details about Bayesian statistical inference should see, e.g., Howson and Urbach (1991), Howson and Urbach (2006), Jaynes (2003), Gregory (2005) and Sivia and Skilling (2006). For a description of frequentist statistics, we recommend Helstrom (1968), Wainstein and Zubakov (1971) and Feldman and Cousins (1998).

3.1 Introduction to Bayesian and frequentist inference

Statistical inference can be used to answer questions such as "Is a gravitational-wave signal present in the data?" and, if so, "What are the physical characteristics of the source?" These questions are addressed using the techniques of classical (also known as *frequentist*) inference and *Bayesian* inference. Many of the early theoretical studies and observational papers in gravitational-wave astronomy followed the frequentist approach, but the use of Bayesian inference is growing in popularity. Moreover, many contemporary analyses cannot be classified as purely frequentist or Bayesian.

The textbook definition states that the difference between the two approaches comes down to their different interpretations of probability: for frequentists, probabilities are fundamentally related to frequencies of events, while for Bayesians, probabilities are fundamentally related to our own knowledge about an event. For example, when inferring the mass of a star, the frequentist interpretation is that the star has a true, fixed (albeit unknown) mass, so it is meaningless to talk about a probability distribution for it. Rather, the uncertainty is in the data, and the relevant probability is that of observing the data d, given that the star has mass m. This probability distribution is the *likelihood*, denoted $p(d|m)$. In contrast, in the Bayesian interpretation the data are known (after all, it is what is measured!), and the mass of the star is what we are uncertain about,[5] so the relevant probability is that the mass has a certain value, given the data. This probability distribution is the *posterior*, $p(m|d)$. The likelihood and posterior are related via Bayes' theorem:

$$p(m|d) = \frac{p(d|m)\,p(m)}{p(d)}, \tag{3.1}$$

[5] In some treatments, the Bayesian interpretation is equated to philosophical schools such as Berkeley's empiricist idealism, or subjectivism, which holds that things only exist to the extent that they are perceived, while the frequentist interpretation is equated to Platonic realism, or metaphysical objectivism, holding that things exist objectively and independently of observation. These equivalences are false. A physical object can have a definite, Platonic existence, and Bayesians can still assign probabilities to its attributes since our ability to measure is limited by imperfect equipment.

where $p(m)$ is the prior probability distribution for m, and the normalization constant,

$$p(d) = \int p(d|m)\,p(m)\,dm, \tag{3.2}$$

is the *marginalized likelihood*, or *evidence*. For uniform (flat) priors the frequentist confidence intervals for the parameters will coincide with the Bayesian credible intervals, but the interpretation remains quiet distinct.

The choice of prior probability distributions is a source of much consternation and debate, and is often cited as a weakness of the Bayesian approach. But the choice of probability distribution for the likelihood (which is also important for the frequentist approach) is often no less fraught. The prior quantifies what we know about the range and distribution of the parameters in our model, while the likelihood quantifies what we know about our measurement apparatus, and, in particular, the nature of the measurement noise. The choice of prior is especially problematic in a new field where there is little to guide the choice. For example, electromagnetic observations and population synthesis models give some guidance about black hole masses, but the mass range and distribution is currently not well constrained. The choice of likelihood can also be challenging when the measurement noise deviates from the stationary, Gaussian ideal. More details related to the choice of likelihood and choice of prior will be given in Sect. 3.6.

In addition to parameter estimation, statistical inference is used to select between competing models, or hypotheses, such as, "is there a gravitational-wave signal in the data or not?" Thanks to GW150914 and GW151226, we know that gravitational-wave signals *are* already present in existing data sets, but most are at levels where we are unable to distinguish them from noise processes. For detection we demand that a model for the data that includes a gravitational-wave signal be favored over a model having no gravitational-wave signal. In Bayesian inference a detection might be announced when the odds ratio between models with and without gravitational-wave signals gets sufficiently large, while in frequentist inference a detection might be announced when the p-value for some test statistic is less than some prescribed threshold. These different approaches to deciding whether or not to claim a detection (e.g., Bayesian model selection or frequentist hypothesis testing), as well as differences in regard to parameter estimation, are described in the following subsections. Table 2 provides an overview of the key similarities and differences between frequentist and Bayesian inference, to be described in detail below.

3.2 Frequentist statistics

As mentioned above, classical or *frequentist* statistics is a branch of statistical inference that interprets probability as the "long-run relative occurrence of an event in a set of identical experiments." Thus, for a frequentist, probabilities can only be assigned to propositions about outcomes of (in principle) repeated experiments (i.e., *random variables*) and not to hypotheses or parameters describing the state of nature, which have fixed but unknown values. In this interpretation, the measured data are drawn

Table 2 Comparison of frequentist and Bayesian approaches to statistical inference

Frequentist	Bayesian
Probabilities assigned only to propositions about outcomes of repeatable experiments (i.e., random variables), not to hypotheses or parameters which have fixed but unknown values	Probabilities can be assigned to hypotheses and parameters since probability is degree of belief (or confidence, plausibility) in any proposition
Assumes measured data are drawn from an underlying probability distribution, which assumes the truth of a particular hypothesis or model (likelihood function)	Same
Constructs a statistic to estimate a parameter or to decide whether or not to claim a detection	Needs to specify prior degree of belief in a particular hypothesis or parameter
Calculates the probability distribution of the statistic (sampling distribution)	Uses Bayes' theorem to update the prior degree of belief in light of new data (i.e., likelihood "plus" prior yields posterior)
Constructs confidence intervals and p-values for parameter estimation and hypothesis testing	Constructs posteriors and odds ratios for parameter estimation and hypothesis testing/model comparison

See Sects. 3.2 and 3.3 for details

from an underlying probability distribution, which assumes the truth of a particular hypothesis or model. The probability distribution for the data is just the likelihood function, which we can write as $p(d|H)$, where d denotes the data and H denotes an hypothesis.

Statistics play an important role in the frequentist framework. These are random variables constructed from the data, which typically estimate a signal parameter or indicate how well the data fit a particular hypothesis. Although it is common to construct statistics from the likelihood function (e.g., the maximum-likelihood statistic for a particular parameter, or the maximum-likelihood ratio to compare a signal-plus-noise model to a noise-only model), there is no a priori restriction on the form of a statistic other than it be *some* function of the data. Ultimately, it is the goal of the analysis and the cleverness of the data analyst that dictate which statistic (or statistics) to use.

To make statistical inferences in the frequentist framework requires knowledge of the probability distribution (also called the *sampling distribution*) of the statistic. The sampling distribution can either be calculated analytically (if the statistic is sufficiently simple) or via Monte Carlo simulations, which effectively construct a histogram of the values of the statistic by simulating many independent realizations of the data. Given a statistic and its sampling distribution, one can then calculate either *confidence intervals* for parameter estimation or p-values for hypothesis testing. (These will be discussed in more detail below). Note that a potential problem with frequentist statistical inference is that the sampling distribution depends on data values that were *not* actually observed, which is related to how the experiment was carried out *or might have been* carried

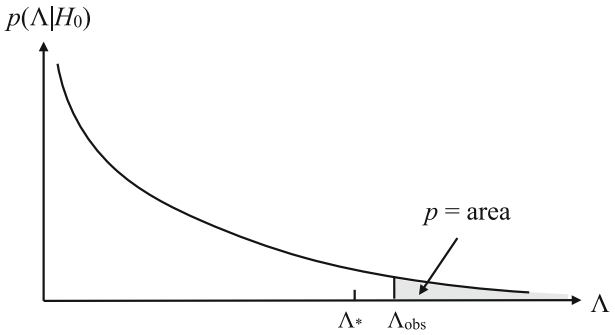

Fig. 3 Definition of the p-value (or significance) for frequentist hypothesis testing. The value of p equals the area under the probability distribution $p(\Lambda|H_0)$ for $\Lambda \geq \Lambda_{\rm obs}$

out. The so-called *stopping problem* of frequentist statistics is an example of such a problem (Howson and Urbach 2006).

3.2.1 Frequentist hypothesis testing

Suppose, as a frequentist, you want to test the hypothesis H_1 that a gravitational-wave signal, having some fixed but unknown amplitude $a > 0$, is present in the data. Since you cannot assign probabilities to hypotheses or to parameters like a as a frequentist, you need to introduce instead an alternative (or *null*) hypothesis H_0, which, for this example, is the hypothesis that there is no gravitational-wave signal in the data (i.e., that $a = 0$). You then argue for H_1 by arguing *against* H_0, similar to proof by contradiction in mathematics. Note that H_1 is a *composite* hypothesis since it depends on a range of values of the unknown parameter a. It can be written as the union, $H_1 = \cup_{a>0} H_a$, of a set of simple hypotheses H_a each corresponding to a single fixed value of the parameter a.

To rule either in favor or against H_0, you construct a statistic Λ, called a *test* or *detection statistic*, on which the statistical test will be based. As mentioned above, you will need to calculate analytically or via Monte Carlo simulations the sampling distribution for Λ under the assumption that the null hypothesis is true, $p(\Lambda|H_0)$. If the observed value of Λ lies far out in the tails of the distribution, then the data are most likely not consistent with the assumption of the null hypothesis, so you reject H_0 (and thus accept H_1) at the $p * 100\%$ level, where

$$p \equiv \text{Prob}(\Lambda > \Lambda_{\rm obs}|H_0) \equiv \int_{\Lambda_{\rm obs}}^{\infty} p(\Lambda|H_0)\, d\Lambda. \tag{3.3}$$

This is the so-called p-value (or *significance*) of the test; it is illustrated graphically in Fig. 3. The p-value required to reject the null hypothesis determines a *threshold* Λ_*, above which you reject H_0 and accept H_1 (e.g., claim a detection). It is related to the *false alarm probability* for the test as we explain below.

The above statistical test is subject to two types of errors: (i) type I or *false alarm* errors, which arise if the data are such that you reject the null hypothesis (i.e., $\Lambda_{\rm obs} >$

Λ_*) when it is actually true, and (ii) type II or *false dismissal* errors, which arise if the data are such that you accept the null hypothesis (i.e., $\Lambda_{\rm obs} < \Lambda_*$) when it is actually false. The false alarm probability α and false dismissal probability $\beta(a)$ are given explicitly by

$$\alpha \equiv \text{Prob}(\Lambda > \Lambda_* | H_0), \tag{3.4}$$

$$\beta(a) \equiv \text{Prob}(\Lambda < \Lambda_* | H_a), \tag{3.5}$$

where a is the amplitude of the gravitational-wave signal, assumed to be present under the assumption that H_1 is true. To calculate the false dismissal probability $\beta(a)$, one needs the sampling distribution of the test statistic assuming the presence of a signal with amplitude a.

Different test statistics are judged according to their false alarm and false dismissal probabilities. Ideally, you would like your statistical test to have false alarm and false dismissal probabilities that are both as small as possible. But these two properties compete with one another as setting a larger threshold value to minimize the false alarm probability will increase the false dismissal probability. Conversely, setting a smaller threshold value to minimize the false dismissal probability will increase the false alarm probability.

In the context of gravitational-wave data analysis, the gravitational-wave community is (at least initially) reluctant to falsely claim detections. Hence the false alarm probability is set to some very low value. The best statistic then is the one that minimizes the false dismissal probability (i.e., maximizes detection probability) for fixed false alarm. This is the *Neyman–Pearson criterion*. For medical diagnosis, on the other hand, a doctor is very reluctant to falsely dismiss an illness. Hence the false dismissal probability will be set to some very low value. The best statistic then is the one which minimizes the false alarm probability for fixed false dismissal.

3.2.2 Frequentist detection probability

The value $1 - \beta(a)$ is called the *detection probability* or *power* of the test. It is the fraction of times that the test statistic Λ correctly identifies the presence of a signal of amplitude a in the data, for a fixed false alarm probability α (which sets the threshold Λ_*). A plot of detection probability versus signal strength is often used to show how strong a signal has to be in order to detect it with a certain probability. Since detection probability does not depend on the observed data—it depends only on the sampling distribution of the test statistic and a choice for the false alarm probability—detection probability curves are often used as a *figure-of-merit* for proposed search methods for a signal. Figure 4 shows a detection probability curve, with the value of a needed to be detectable with 90% frequentist probability indicated by the dashed vertical line. We will denote this value of a by $a^{90\%,\text{DP}}$. Note that as the signal amplitude goes to zero, the detection probability reduces to the false alarm probability α, which for this example was chosen to be 0.10.

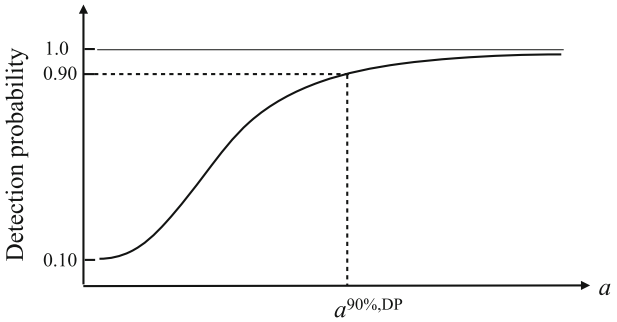

Fig. 4 Detection probability as a function of the signal amplitude for a false alarm probability equal to 10%. The value of a needed for 90% detection probability is indicated by the *dashed vertical line* and is denoted by $a^{90\%,DP}$

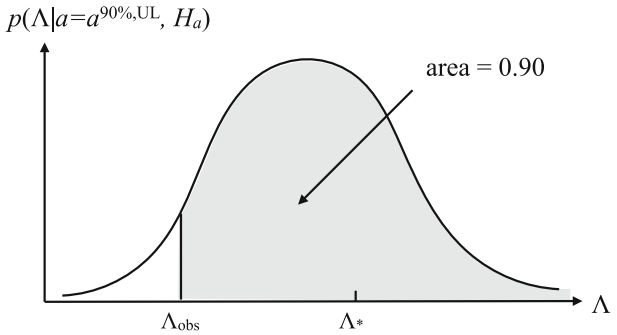

Fig. 5 Graphical representation of a frequentist 90% confidence level upper limit. When $a = a^{90\%,UL}$, the probability of obtaining a value of the detection statistic $\Lambda \geq \Lambda_{obs}$ is equal to 0.90

3.2.3 Frequentist upper limits

In the absence of a detection (i.e., if the observed value of the test statistic is less than the detection threshold Λ_*), one can still set a bound (called an *upper limit*) on the strength of the signal that one was trying to detect. The upper limit depends on the observed value of the test statistic, Λ_{obs}, and a choice of confidence level, CL, interpreted in the frequentist framework as the long-run relative occurrence for a set of repeated identical experiments. For example, one defines the 90% confidence-level upper limit $a^{90\%,UL}$ as the minimum value of a for which $\Lambda \geq \Lambda_{obs}$ at least 90% of the time:

$$\text{Prob}(\Lambda \geq \Lambda_{obs} | a \geq a^{90\%,UL}, H_a) \geq 0.90. \tag{3.6}$$

In other words, if the signal has an amplitude $a^{90\%,UL}$ or higher, we would have detected it in at least 90% of repeated observations. A graphical representation of a frequentist upper limit is given in Fig. 5.

3.2.4 Frequentist parameter estimation

The frequentist prescription for estimating the value of a particular parameter a, like the amplitude of a gravitational-wave signal, is slightly different than the method used to claim a detection. You need to first construct a statistic (called an *estimator*) $\hat{a}$ of the parameter a you are interested in. (This might be a maximum-likelihood estimator of a, but other estimators can also be used). You then calculate its sampling distribution $p(\hat{a}|a, H_a)$. Note that statements like

$$\text{Prob}(a - \Delta < \hat{a} < a + \Delta) = 0.95, \tag{3.7}$$

which one constructs from $p(\hat{a}|a, H_a)$ make sense in the frequentist framework, since $\hat{a}$ is a random variable. Although the above inequality can be rearranged to yield

$$\text{Prob}(\hat{a} - \Delta < a < \hat{a} + \Delta) = 0.95, \tag{3.8}$$

this should *not* be interpreted as a statement about the probability of a lying within a particular interval $[\hat{a} - \Delta, \hat{a} + \Delta]$, since a is not a random variable. Rather, it should be interpreted as a probabilistic statement about the *set of intervals* $\{[\hat{a} - \Delta, \hat{a} + \Delta]\}$ for all possible values of $\hat{a}$. Namely, in a set of many repeated experiments, 0.95 is the fraction of the intervals that will contain the true value of the parameter a. Such an interval is called a 95% *frequentist confidence interval*. This is illustrated graphically in Fig. 6.

It is important to point out that an estimator can sometimes take on a value of the parameter that is *not physically allowed*. For example, if the parameter a denotes the amplitude of a gravitational-wave signal (so physically $a \geq 0$), it is possible for $\hat{a} < 0$ for a particular realization of the data. Note that there is nothing mathematically wrong with this result. Indeed, the sampling distribution for $\hat{a}$ specifies the probability of

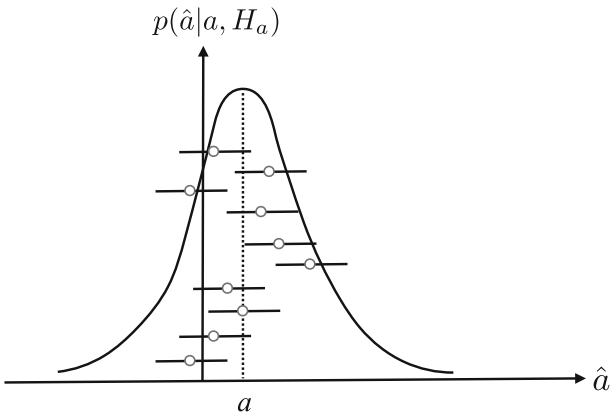

Fig. 6 Definition of the frequentist confidence interval for parameter estimation. *Each circle* and *line* represents a measured interval $[\hat{a} - \Delta, \hat{a} + \Delta]$. The set of all such intervals will contain the true value of the parameter a (indicated here by the *dotted vertical line*) CL $*$ 100% of the time, where CL is the confidence level

obtaining such values of $\hat{a}$. It is even possible to have a confidence interval $[\hat{a}-\Delta, \hat{a}+\Delta]$ all of whose values are unphysical, especially if one is trying to detect a weak signal in noise. Again, this is mathematically allowed, but it is a little awkward to report a frequentist confidence interval that is completely unphysical. We shall see that within the Bayesian framework unphysical intervals and unphysical posteriors never arise, as a simple consequence of including a prior distribution on the parameter that requires $a > 0$.

3.2.5 Unified approach for frequentist upper limits and confidence intervals

Frequentists also have a way of avoiding unphysical or empty confidence intervals, which at the same time *unifies* the treatment of upper limits for null results and two-sided intervals for non-null results. This procedure, developed by Feldman and Cousins (1998), also solves the problem that the choice of an upper limit or two-sided confidence interval leads to intervals that do not have the proper coverage (i.e., the probability that an interval contains the true value of a parameter does not match the stated confidence level) if the choice of reporting an upper limit or two-sided confidence interval is *based on the data* and not decided upon before performing the experiment.

The basic idea underlying this unified approach to frequentist intervals is a new specification (or *ordering*) of the values of the random variable to include in the acceptance intervals for an unknown parameter. If we let a denote the parameter whose value we are trying to determine, and $\hat{a}$ be an estimator of a with sampling distribution $p(\hat{a}|a, H_a)$, then the choice of acceptance intervals becomes, for each value of a, how do we choose $[\hat{a}_1, \hat{a}_2]$ such that

$$\text{Prob}(\hat{a}_1 < \hat{a} < \hat{a}_2) \equiv \int_{\hat{a}_1}^{\hat{a}_2} p(\hat{a}|a, H_a)\, d\hat{a} = \text{CL}, \tag{3.9}$$

where CL is the confidence level, e.g., CL $= 0.95$. The ordering principle proposed by Feldman and Cousins (1998) is based on the ranking function

$$R(\hat{a}|a) \equiv \frac{p(\hat{a}|a, H_a)}{p(\hat{a}|a, H_a)\big|_{a=a_{\text{best}}}}, \tag{3.10}$$

where a_{best} is the value of the parameter a that maximizes the sampling distribution $p(\hat{a}|a, H_a)$ for a given value of $\hat{a}$. The prescription then for constructing the acceptance intervals is to find, for each allowed value of a, values of $\hat{a}_1$ and $\hat{a}_2$ such that $R(\hat{a}_1|a) = R(\hat{a}_2|a)$ and for which (3.9) is satisfied. The set of all such acceptance intervals for different values of a forms a *confidence belt* in the $\hat{a}a$-plane, which is then used to construct an upper limit or a two-sided confidence interval for a particular observed value of the estimator $\hat{a}$, as explained below and illustrated in Fig. 7.

As a specific example, let us suppose that $\hat{a}$ is Gaussian-distributed about a with variance σ^2:

$$p(\hat{a}|a, H_a) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(\hat{a}-a)^2}{\sigma^2}}, \tag{3.11}$$

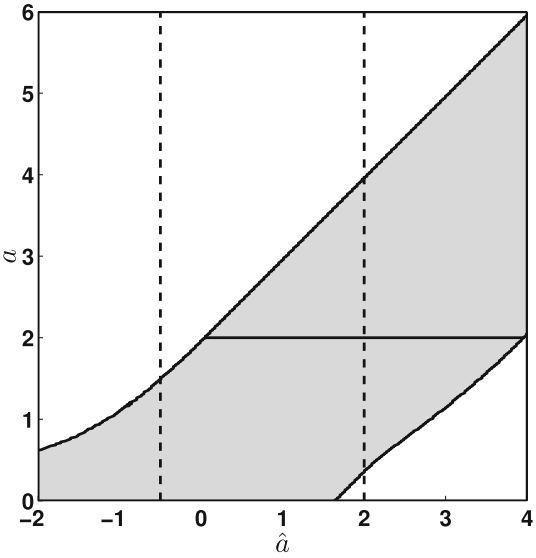

Fig. 7 Confidence belt for 95% confidence-level intervals for a Gaussian distribution with mean $a > 0$. (The values for a and $\hat{a}$ are given here in units of σ). The *solid horizontal line* shows the acceptance interval for $a = 2.0$. The *two dashed vertical lines* correspond to two different observed values for the estimator $\hat{a}$: $\hat{a} = -0.5$, which has a 95% confidence-level upper limit $a \leq 1.5$; and $\hat{a} = 2$, which has a 95% confidence-level two-sided interval $a \in [0.35, 3.95]$

and that the unknown parameter a represents the amplitude of a signal, so that $a > 0$. (Recall that it is possible, however, for the estimator $\hat{a}$ to take on negative values). Then $a_{\text{best}} = \hat{a}$ if $\hat{a} > 0$, while $a_{\text{best}} = 0$ if $\hat{a} \leq 0$, for which

$$p(\hat{a}|a, H_a)\Big|_{a=a_{\text{best}}} = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma}, & \hat{a} > 0 \\ \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\frac{\hat{a}^2}{\sigma^2}\right], & \hat{a} \leq 0 \end{cases} \tag{3.12}$$

and

$$R(\hat{a}|a) = \begin{cases} \exp\left[-\frac{1}{2}\frac{(\hat{a}-a)^2}{\sigma^2}\right], & \hat{a} > 0 \\ \exp\left[-\frac{1}{2}\frac{(-2\hat{a}a+a^2)}{\sigma^2}\right], & \hat{a} \leq 0 \end{cases}. \tag{3.13}$$

The confidence belt constructed from this ranking function is shown in Fig. 7. The solid horizontal line at $a = 2$ shows the corresponding 95% confidence-level acceptance interval for this ranking function. The two dashed vertical lines correspond to two different observed values for the estimator $\hat{a}$, leading to a 95% confidence-level upper limit and two-sided interval, respectively.

3.3 Bayesian inference

In the following subsections, we again describe parameter estimation and hypothesis testing, but this time from the perspective of Bayesian inference.

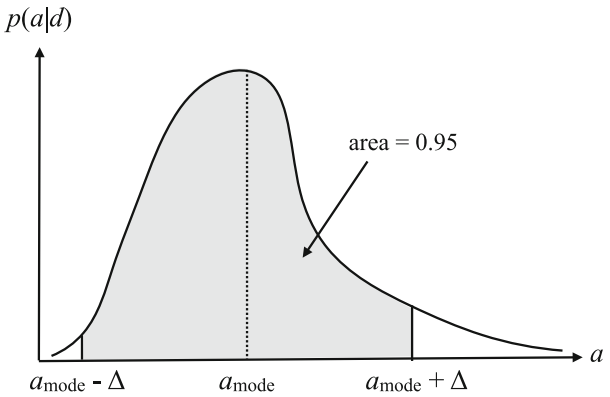

Fig. 8 Definition of a Bayesian credible interval for parameter estimation. Here we construct a symmetric 95% credible interval centered on the mode of the distribution

3.3.1 Bayesian parameter estimation

In Bayesian inference, a parameter, e.g., a, is estimated in terms of its posterior distribution, $p(a|d)$, in light of the observed data d. As discussed in the introduction to this section, the posterior $p(a|d)$ can be calculated from the likelihood $p(d|a)$ and the prior probability distribution $p(a)$ using Bayes' theorem

$$p(a|d) = \frac{p(d|a)\,p(a)}{p(d)}. \tag{3.14}$$

The posterior distribution tells you everything you need to know about the parameter, although you might sometimes want to reduce it to a few numbers—e.g., its mode, mean, standard deviation, etc.

Given a posterior distribution $p(a|d)$, a Bayesian confidence interval (often called a *credible interval* given the Bayesian interpretation of probability as degree of belief, or state of knowledge, about an event) is simply defined in terms of the area under the posterior between one parameter value and another. This is illustrated graphically in Fig. 8, for the case of a 95% symmetric credible interval, centered on the mode of the distribution $a_{\rm mode}$. If the posterior distribution depends on two parameters a and b, but you really only care about a, then you can obtain the posterior distribution for a by marginalizing the joint distribution $p(a, b|d)$ over b:

$$p(a|d) = \int db\; p(a, b|d) = \int db\; p(a|b, d)p(b), \tag{3.15}$$

where the second equality follows from the relationship between joint probabilities and conditional probabilities, e.g., $p(a|b, d)p(b) = p(a, b|d)$. Variables that you don't particularly care about (e.g., the variance of the detector noise as opposed to the strength of a gravitational-wave signal) are called *nuisance parameters*. Although nuisance parameters can be handled in a straight-forward manner using Bayesian

inference, they are problematic to deal with (i.e., they are a nuisance!) in the context of frequentist statistics. The problem is that marginalization doesn't make sense to a frequentist, for whom parameters cannot be assigned probability distributions.

The interpretation of Bayes' theorem (3.14) is that our prior knowledge is updated by what we learn from the data, as measured by the likelihood, to give our posterior state of knowledge. The amount learned from the data is measured by the information gain

$$I = \int da\; p(a|d) \log\left(\frac{p(a|d)}{p(a)}\right). \tag{3.16}$$

Using a natural logarithm gives the information in *nats*, while using a base 2 logarithm gives the information in *bits*. If the data tells us nothing about the parameter, then $p(d|a) = \text{constant}$, which implies $p(a|d) = p(a)$ and thus $I = 0$.

3.3.2 Bayesian upper limits

A Bayesian upper limit is simply a Bayesian credible interval for a parameter with the lower end point of the interval set to the smallest value that the parameter can take. For example, the Bayesian 90% upper limit on a parameter $a > 0$ is defined by:

$$\text{Prob}(0 < a < a^{90\%,\text{UL}}|d) = 0.90, \tag{3.17}$$

where probability is interpreted as degree of belief, or state of knowledge, that the parameter a has a value in the indicated range. One usually sets an upper limit on a parameter when the mode of the distribution for the parameter being estimated is not sufficiently displaced from zero, as shown in Fig. 9.

3.3.3 Bayesian model selection

Bayesian inference can easily be applied to multiple models or hypotheses, each with a different set of parameters. In what follows, we will denote the different models

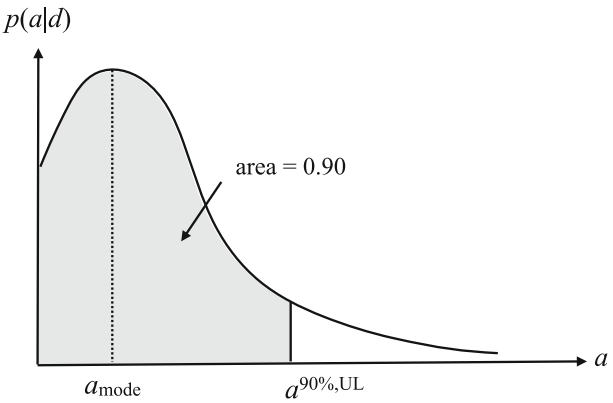

Fig. 9 Bayesian 90% credible upper limit for the parameter a

by $\mathcal{M}_\alpha$, where the index α runs over the different models, and the associated set of parameters by the vector $\boldsymbol{\theta}_\alpha$. The joint posterior distribution for the parameters $\boldsymbol{\theta}_\alpha$ is given by

$$p(\boldsymbol{\theta}_\alpha|d, \mathcal{M}_\alpha) = \frac{p(d|\boldsymbol{\theta}_\alpha, \mathcal{M}_\alpha)p(\boldsymbol{\theta}_\alpha|\mathcal{M}_\alpha)}{p(d|\mathcal{M}_\alpha)}, \tag{3.18}$$

and the model evidence is given by

$$p(d|\mathcal{M}_\alpha) = \int p(d|\boldsymbol{\theta}_\alpha, \mathcal{M}_\alpha)p(\boldsymbol{\theta}_\alpha|\mathcal{M}_\alpha)\, d\boldsymbol{\theta}_\alpha, \tag{3.19}$$

where we marginalize over the parameter values associated with that model. The posterior probability for model $\mathcal{M}_\alpha$ is given by Bayes' theorem as

$$p(\mathcal{M}_\alpha|d) = \frac{p(d|\mathcal{M}_\alpha)p(\mathcal{M}_\alpha)}{p(d)}, \tag{3.20}$$

where the normalization constant $p(d)$ involves a sum over all possible models:

$$p(d) = \sum_\alpha p(d|\mathcal{M}_\alpha)p(\mathcal{M}_\alpha). \tag{3.21}$$

Since the space of all possible models is generally unknown, the sum is usually taken over the subset of models being considered. The normalization can be avoided by considering the posterior odds ratio between two models:

$$\mathcal{O}_{\alpha\beta}(d) = \frac{p(\mathcal{M}_\alpha|d)}{p(\mathcal{M}_\beta|d)} = \frac{p(\mathcal{M}_\alpha)}{p(\mathcal{M}_\beta)}\,\frac{p(d|\mathcal{M}_\alpha)}{p(d|\mathcal{M}_\beta)}. \tag{3.22}$$

The first ratio on the right-hand side of the above equation is the *prior* odds ratio for models α, β, while the second term is the evidence ratio, or *Bayes factor*,

$$\mathcal{B}_{\alpha\beta}(d) \equiv \frac{p(d|\mathcal{M}_\alpha)}{p(d|\mathcal{M}_\beta)}. \tag{3.23}$$

The prior odds ratio is often taken to equal unity, but this is not always justified. For example, the prior odds that a signal is described by general relativity versus some alternative theory of gravity should be much larger than unity given the firm theoretical and observational footing of Einstein's theory.

While the foundations of Bayesian inference were laid out by Laplace in the 1700s, it did not see widespread use until the late twentieth century with the advent of practical implementation schemes and the development of fast electronic computers. Today, Monte Carlo sampling techniques, such as Markov Chain Monte Carlo (MCMC) and Nested Sampling, are used to sample the posterior and estimate the evidence (Skilling 2006; Gair et al. 2010). Successfully applying these techniques is something of an art, but in principle, once the likelihood and prior have been written down, the implementation of Bayesian inference is purely mechanical. Calculating the likelihood and choosing a prior will be discussed in some detail in Sect. 3.6.

3.4 Relating Bayesian and frequentist detection statements

It is interesting to compare the Bayesian model selection calculation discussed above to frequentist hypothesis testing based on the *maximum-likelihood ratio*. For concreteness, let us assume that we have two models $\mathcal{M}_0$ (noise-only) and $\mathcal{M}_1$ (noise plus gravitational-wave signal), with parameters $\boldsymbol{\theta}_n$ and $\{\boldsymbol{\theta}_n, \boldsymbol{\theta}_h\}$, respectively. The frequentist detection statistic will be defined in terms of the ratio of the maxima of the likelihood functions for the two models:

$$\Lambda_{\mathrm{ML}}(d) \equiv \frac{\max_{\boldsymbol{\theta}_n} \max_{\boldsymbol{\theta}_h} p(d|\boldsymbol{\theta}_n, \boldsymbol{\theta}_h, \mathcal{M}_1)}{\max_{\boldsymbol{\theta}'_n} p(d|\boldsymbol{\theta}'_n, \mathcal{M}_0)}. \tag{3.24}$$

As described above, the Bayes factor calculation also involves a ratio of two quantities, the model evidences $p(d|\mathcal{M}_1)$ and $p(d|\mathcal{M}_0)$, but instead of maximizing over the parameters, we marginalize over the parameters:

$$\mathcal{B}_{10}(d) = \frac{\int d\boldsymbol{\theta}_n \int d\boldsymbol{\theta}_h \; p(d|\boldsymbol{\theta}_n, \boldsymbol{\theta}_h, \mathcal{M}_1) p(\boldsymbol{\theta}_n, \boldsymbol{\theta}_h|\mathcal{M}_1)}{\int d\boldsymbol{\theta}'_n \; p(d|\boldsymbol{\theta}'_n, \mathcal{M}_0) p(\boldsymbol{\theta}'_n|\mathcal{M}_0)}. \tag{3.25}$$

These two expressions can be related using Laplace's approximation to individually approximate the model evidences $p(d|\mathcal{M}_1)$ and $p(d|\mathcal{M}_0)$. This approximation is valid when the data are *informative*—i.e., when the likelihood functions are peaked relative to the joint prior probability distributions of the parameters. For an arbitrary model $\mathcal{M}$ with parameters $\boldsymbol{\theta}$, the Laplace approximation yields:

$$\int d\boldsymbol{\theta} \; p(d|\boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M}) \simeq p(d|\boldsymbol{\theta}_{\mathrm{ML}}, \mathcal{M}) \frac{\Delta V_{\mathcal{M}}}{V_{\mathcal{M}}}, \tag{3.26}$$

where $\boldsymbol{\theta}_{\mathrm{ML}} \equiv \boldsymbol{\theta}_{\mathrm{ML}}(d)$ maximizes the likelihood with respect to variations of $\boldsymbol{\theta}$ given the data d; $\Delta V_{\mathcal{M}}$ is the characteristic spread of the likelihood function around its maximum (the volume of the uncertainty ellipsoid for the parameters); and $V_{\mathcal{M}}$ is the total parameter space volume of the model parameters. Applying this approximation to models $\mathcal{M}_0$ and $\mathcal{M}_1$ in (3.25), we obtain

$$\mathcal{B}_{10}(d) \simeq \Lambda_{\mathrm{ML}}(d) \frac{\Delta V_1 / V_1}{\Delta V_0 / V_0}, \tag{3.27}$$

or, equivalently,

$$2 \ln \mathcal{B}_{10}(d) \simeq 2 \ln \left(\Lambda_{\mathrm{ML}}(d)\right) + 2 \ln \left(\frac{\Delta V_1 / V_1}{\Delta V_0 / V_0} \right). \tag{3.28}$$

The second term on the right-hand side of the above equation is negative and penalizes models that require a larger parameter space volume than necessary to fit the data. This is basically an *Occam penalty factor*, which prefers the simpler of two models that fit the data equally well. The first term has the interpretation of being the squared

Table 3 Bayes factors and their interpretation in terms of the strength of the evidence in favor of one model relative to the other

$\mathcal{B}_{\alpha\beta}(d)$	$2\ln\mathcal{B}_{\alpha\beta}(d)$	Evidence for model $\mathcal{M}_\alpha$ relative to $\mathcal{M}_\beta$
<1	<0	Negative (supports model $\mathcal{M}_\beta$)
1–3	0–2	Not worth more than a bare mention
3–20	2–6	Positive
20–150	6–10	Strong
>150	>10	Very strong

Adapted from Kass and Raftery (1995)

signal-to-noise ratio of the data, assuming an additive signal in Gaussian-stationary noise, and it can be used as an alternative frequentist detection statistic in place of $\Lambda_{\rm ML}$.

Table 3 from Kass and Raftery (1995) gives a range of Bayes factors and their interpretation in terms of the strength of the evidence in favor of one model relative to another. The precise levels at which one considers the evidence to be "strong" or "very strong" is rather subjective. But recent studies (Cornish and Sampson 2016; Taylor et al. 2016a) in the context of pulsar timing have been trying to make this correspondence a bit firmer, using *sky* and *phase scrambles* to effectively destroy signal-induced spatial correlations between pulsars while retaining the statistical properties of each individual dataset. This is similar to doing time-slides for LIGO analyses, which are used to assess the significance of a detection.

Taylor et al. (2016a) even go so far as to perform a *hybrid* frequentist-Bayesian analysis, doing Monte Carlo simulations: (i) over different noise-only realizations, and (ii) over different sky and phase scrambles, which null the correlated signal. These simulations produce different null *distributions* for the Bayes factor, similar to a null-hypothesis distribution for a frequentist detection statistic (in this case, the log of the Bayes factor). The significance of the measured Bayes factor is then its corresponding p-value with respect to one of these null distributions. The utility of such a hybrid analysis is its ability to better assess the significance of a detection claim, especially when there might be questions about the suitability of one of the models (e.g., the noise model) used in the construction of a likelihood function.

3.5 Simple example comparing Bayesian and frequentist analyses

To further illustrate the relationship between Bayesian and frequentist analyses, we consider in this section a very simple example—a constant signal with amplitude $a > 0$ in white, Gaussian noise (zero mean, variance σ):

$$d_i = a + n_i, \quad i = 1, 2, \ldots, N, \tag{3.29}$$

where the index i labels the individual samples of the data. The likelihood functions for the noise-only and signal-plus-noise models $\mathcal{M}_0$ and $\mathcal{M}_1$ are thus simple Gaussians:

$$p(d|\mathcal{M}_0) = \frac{1}{(2\pi)^{N/2}\sigma^N} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^N d_i^2}, \tag{3.30}$$

$$p(d|a,\mathcal{M}_1) = \frac{1}{(2\pi)^{N/2}\sigma^N} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^N (d_i-a)^2}. \tag{3.31}$$

We will assume that the value of σ is known a priori. Thus, the noise model has no free parameters, while the signal model has just one parameter, which is the amplitude of the signal that we are trying to detect. We will choose our prior on a to be flat over the interval $(0, a_{\max}]$, so $p(a) = 1/a_{\max}$.

It is straight-forward exercise to check that the maximum-likelihood estimator of the amplitude a is given by the sample mean of the data:

$$\hat{a} \equiv a_{\rm ML}(d) = \frac{1}{N}\sum_{i=1}^N d_i \equiv \bar{d}. \tag{3.32}$$

This is an unbiased estimator of a and has variance $\sigma_{\hat{a}}^2 = \sigma^2/N$ (the familiar variance of the sample mean). Thus, the sampling distribution of $\hat{a}$ is simply

$$p(\hat{a}|a,\mathcal{M}_1) = \frac{1}{\sqrt{2\pi}\sigma_{\hat{a}}} e^{-\frac{1}{2\sigma_{\hat{a}}^2}(\hat{a}-a)^2}, \tag{3.33}$$

where $\hat{a}$ can take on either positive or negative values (even though $a > 0$).

To compute the posterior distribution $p(a|d,\mathcal{M}_1)$ for the Bayesian analysis, we first note that

$$\sum_{i=1}^N (d_i - a)^2 = N(\mathrm{Var}[d] + (a-\hat{a})^2). \tag{3.34}$$

The model evidence $p(d|\mathcal{M}_1)$ is then given by

$$p(d|\mathcal{M}_1) = \frac{e^{-\frac{\mathrm{Var}[d]}{2\sigma_{\hat{a}}^2}}\left[\mathrm{erf}\left(\frac{a_{\max}-\hat{a}}{\sqrt{2}\sigma_{\hat{a}}}\right) + \mathrm{erf}\left(\frac{\hat{a}}{\sqrt{2}\sigma_{\hat{a}}}\right)\right]}{2a_{\max}\sqrt{N}(2\pi)^{(N-1)/2}\sigma^{(N-1)}}, \tag{3.35}$$

and the posterior distribution is given by

$$p(a|d,\mathcal{M}_1) = \frac{1}{\sqrt{2\pi}\sigma_{\hat{a}}} e^{-\frac{(a-\hat{a})^2}{2\sigma_{\hat{a}}^2}} 2\left[\mathrm{erf}\left(\frac{a_{\max}-\hat{a}}{\sqrt{2}\sigma_{\hat{a}}}\right) + \mathrm{erf}\left(\frac{\hat{a}}{\sqrt{2}\sigma_{\hat{a}}}\right)\right]^{-1}. \tag{3.36}$$

Note that this is simply a *truncated* Gaussian on the interval $a \in (0, a_{\max}]$, with mean $\hat{a}$ and variance $\sigma_{\hat{a}}^2$.

The above calculation shows that $\hat{a}$ is a *sufficient statistic* for a. This means that the posterior distribution for a can be written simply in terms of $\hat{a}$, in lieu of the individual samples $d \equiv \{d_1, d_2, \ldots, d_N\}$. The Bayes factor

$$\mathcal{B}_{10}(d) = \frac{p(d|\mathcal{M}_1)}{p(d|\mathcal{M}_0)}, \tag{3.37}$$

is given by

$$\mathcal{B}_{10}(d) = e^{\frac{\hat{a}^2}{2\sigma_{\hat{a}}^2}} \left(\frac{\sqrt{2\pi}\sigma_{\hat{a}}}{a_{\max}}\right) \frac{1}{2}\left[\mathrm{erf}\left(\frac{a_{\max}-\hat{a}}{\sqrt{2}\sigma_{\hat{a}}}\right) + \mathrm{erf}\left(\frac{\hat{a}}{\sqrt{2}\sigma_{\hat{a}}}\right)\right]. \tag{3.38}$$

In the limit where $\hat{a}$ is tightly peaked away from 0 and $a_{\max}$, the Bayes factor simplifies to

$$\mathcal{B}_{10}(d) \simeq e^{\frac{\hat{a}^2}{2\sigma_{\hat{a}}^2}} \left(\frac{\sqrt{2\pi}\sigma_{\hat{a}}}{a_{\max}}\right). \tag{3.39}$$

If we take the frequentist detection statistic to be twice the log of the maximum-likelihood ratio, $\Lambda(d) \equiv 2\ln\Lambda_{\mathrm{ML}}(d)$, then

$$\Lambda(d) = \frac{\hat{a}^2}{\sigma_{\hat{a}}^2} = \frac{\bar{d}^2}{\sigma^2/N} \equiv \rho^2, \tag{3.40}$$

which is just the squared signal-to-noise ratio of the data. Furthermore, taking twice the log of the approximate Bayes factor in (3.39) gives

$$2\ln\mathcal{B}_{10}(d) \simeq \Lambda(d) + 2\ln\left(\frac{\sqrt{2\pi}\sigma_{\hat{a}}}{a_{\max}}\right), \tag{3.41}$$

where the first term is just the frequentist detection statistic and second term expresses the Occam penalty. This last result is consistent with the general relation (3.28) discussed in the previous subsection.

The statistical distribution of the frequentist detection statistic can be found in closed form for this simple example. Since a linear combination of Gaussian random variables is also Gaussian-distributed, Λ is the *square* of a (single) Gaussian random variable $\rho = \bar{d}\sqrt{N}/\sigma$. Moreover, since ρ has mean $\mu \equiv a\sqrt{N}/\sigma$ and unit variance, the sampling distribution for Λ in the presence of a signal is a *noncentral chi-squared* distribution with one degree of freedom and non-centrality parameter $\lambda \equiv \mu^2 = a^2N/\sigma^2$:

$$p(\Lambda|a,\mathcal{M}_1) = \frac{1}{2}e^{-(\Lambda+\lambda)/2}\left(\frac{\Lambda}{\lambda}\right)^{-1/4} I_{-1/2}(\sqrt{\lambda\Lambda}), \tag{3.42}$$

where $I_{-1/2}$ is a modified Bessel function of the first kind of order $-1/2$. In the absence of a signal (i.e., when a and hence λ are equal to zero), Λ is given by an (ordinary) chi-squared distribution with one degree of freedom:

$$p(\Lambda|\mathcal{M}_0) = \frac{1}{\sqrt{2}\Gamma(1/2)}\Lambda^{-1/2}e^{-\Lambda/2}, \tag{3.43}$$

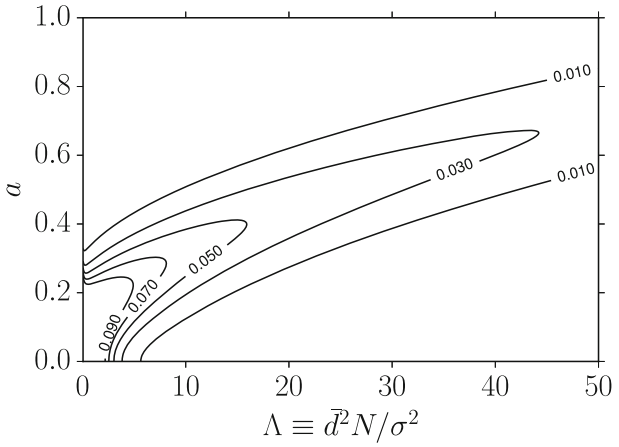

Fig. 10 Equal-probability contour plot for the frequentist detection statistic $\Lambda \equiv \bar{d}^2 N/\sigma^2$ for a signal with amplitude $a > 0$. The *contours* correspond to the values $p(\Lambda|a, \mathcal{M}_1) = 0.01, 0.03, 0.05, 0.07$, and 0.09

where Γ is the gamma function. Substituting explicit expressions for $I_{-1/2}(\sqrt{\lambda\Lambda})$ and $\Gamma(1/2)$, we find:

$$p(\Lambda|\mathcal{M}_0) = \frac{1}{\sqrt{2\pi\Lambda}} e^{-\Lambda/2}, \tag{3.44}$$

$$p(\Lambda|a, \mathcal{M}_1) = \frac{1}{\sqrt{2\pi\Lambda}} \frac{1}{2} \left[e^{-\frac{1}{2}(\sqrt{\Lambda}-\sqrt{\lambda})^2} + e^{-\frac{1}{2}(\sqrt{\Lambda}+\sqrt{\lambda})^2} \right]. \tag{3.45}$$

An equal-probability contour plot of the sampling distribution of the detection statistic is shown in Fig. 10. The fact that we are able to write down *analytic* expressions for the sampling distributions for the detection statistic Λ is due to the simplicity of the signal and noise models. For more complicated real-world problems, these distributions would need to be generated *numerically* using fake signal injections and time-shifts to produce many different realizations of the data (signal plus noise) from which one can build up the distributions.

It is also important to point out that Λ is *not* a sufficient statistic for a, due to the fact that Λ involves the *square* of the maximum-likelihood estimate $\hat{a}$—i.e., $\Lambda = \hat{a}^2 N/\sigma^2$. Thus, we cannot take $p(\Lambda|a, \mathcal{M}_1)$ conditioned on Λ (assuming a flat prior on a from $[0, a_{\max}]$) to get the posterior distribution for a given d, since we would be missing out on data samples that give negative values for $\hat{a}$. Another way to see this is to start with $p(\Lambda|a, \mathcal{M}_1)$ given by (3.45), and then make a change of variables from Λ to $\hat{a}$ using the general transformation relation

$$p_Y(y)\, dy = p_X(x)\, dx \quad \Rightarrow \quad p_X(x) = \left[p_Y(y)\, |f'(x)| \right]_{y=f(x)}. \tag{3.46}$$

This leads to

$$\tilde{p}(\hat{a}|a, \mathcal{M}_1) = \frac{1}{\sqrt{2\pi}\sigma_{\hat{a}}} \left[e^{-\frac{1}{2\sigma_{\hat{a}}^2}(\hat{a}-a)^2} + e^{-\frac{1}{2\sigma_{\hat{a}}^2}(\hat{a}+a)^2} \right], \tag{3.47}$$

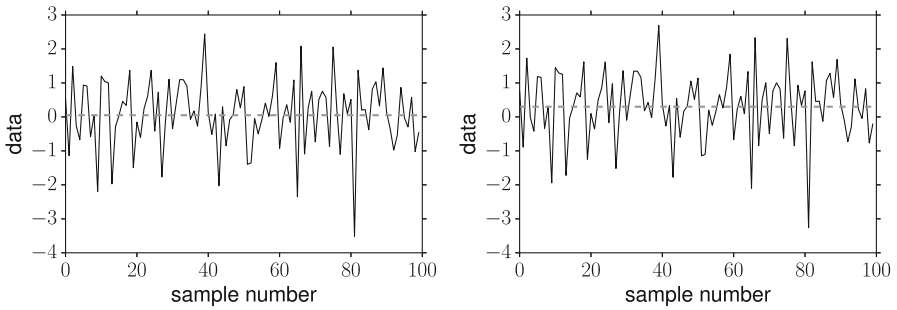

Fig. 11 Examples of simulated data for weak (*left panel*) and strong (*right panel*) signals injected into the data—$a_0 = 0.05$ and 0.3, respectively

which is properly normalized for $\hat{a} > 0$, but differs from (3.33) due to the second term involving $\hat{a} + a$. Thus, we need to construct $p(a|d)$ from (3.33)—and *not* from (3.47)—if we want the posterior to have the proper dependence on a.

3.5.1 Simulated data

For our example, we will take $N = 100$ samples, $\sigma = 1$, and $a_{\max} = 1.0$. We also simulate data with injected signals having amplitudes $a_0 = 0.05$ and 0.3, respectively. Since the expected signal-to-noise ratio, $a\sqrt{N}/\sigma$, is given by 0.5 and 3.0, these injections correspond to *weak* and (moderately) *strong* signals. Single realizations of the data for the two different injections are shown in Fig. 11. The noise realization is the same for the two injections.

3.5.2 Frequentist analysis

Given the values for N, σ, and the probability distributions (3.44) and (3.45) for the frequentist detection statistic Λ, we can calculate the detection threshold for fixed false alarm probability α (which we will take to equal 10%), and the corresponding detection probability as a function of the amplitude a. The detection threshold turns out to equal $\Lambda_* = 2.9$ (so 10% of the area under the probability distribution $p(\Lambda|\mathcal{M}_0)$ is for $\Lambda \geq \Lambda_*$). The value of the amplitude a needed for 90% confidence detection probability with 10% false alarm probability is given by $a^{90\%,\mathrm{DP}} = 0.30$. (These results for the detection threshold and detection probability do *not* depend on the particular realizations of the simulated data). The corresponding curves are shown in Fig. 12.

The sample mean of the data for the two simulations are given by $\bar{d} = 0.085$ and 0.335, respectively. Since $\hat{a} = \bar{d}$, these are also the values of the maximum-likelihood estimator of the amplitude a. The corresponding values of the detection statistic are $\Lambda_{\mathrm{obs}} = 0.72$ and 11.2 for the two injections, and have p-values equal to 0.45 and 9.0×10^{-4}, as shown in Fig. 13. The 95% frequentist confidence interval is given simply by $[\hat{a} - 2\sigma_{\hat{a}}, \hat{a} + 2\sigma_{\hat{a}}]$, since $\hat{a}$ is Gaussian-distributed, and has values

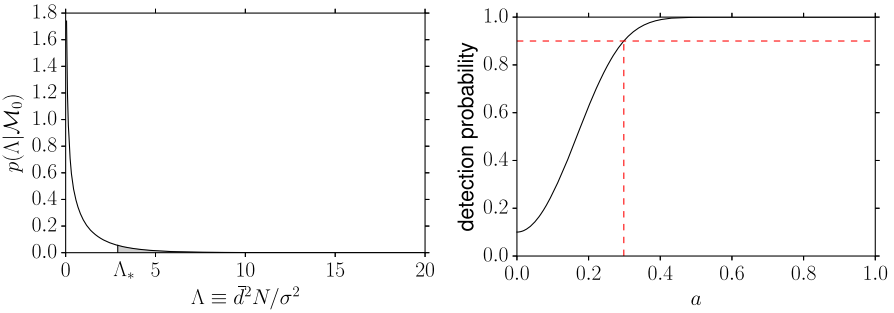

Fig. 12 *Left panel* Probability distribution for the frequentist detection statistic Λ for the noise-only model. The threshold value of the statistic for 10% false alarm probability is $\Lambda_* = 2.9$. *Right panel* Detection probability as a function of the amplitude a. The value of the amplitude needed for 90% confidence detection probability with 10% false alarm probability is $a^{90\%,\mathrm{DP}} = 0.30$

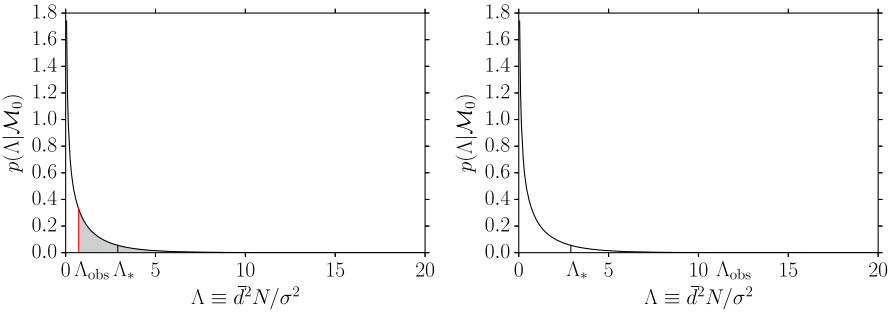

Fig. 13 Graphical representation of the p-value calculation for the weak (*left panel*) and strong (*right panel*) injections. For the weak injection, $\Lambda_{\mathrm{obs}} = 0.72$ is marked by the *red vertical line*, with corresponding p-value 0.45. For the strong injection, $\Lambda_{\mathrm{obs}} = 11.2$ is sufficiently large that the corresponding *red vertical line* is not visible on this graph. The p-value for the strong injection is 9.0×10^{-4}

$[-0.11, 0.29]$ and $[0.14, 0.54]$, respectively. These intervals contain the true value of the amplitudes for the two injections, $a_0 = 0.05$ and 0.3.

The 90% confidence-level frequentist upper limits are $a^{90\%,\mathrm{UL}} = 0.20$ and 0.46, respectively. Figure 14 shows the probability distributions for the detection statistic Λ conditioned on these upper limit values for which the probability of obtaining $\Lambda \geq \Lambda_{\mathrm{obs}}$ is equal to 0.90.

3.5.3 Bayesian analysis

The results of the Bayesian analysis for the two different injections are summarized in Fig. 15. The plots show the posterior distribution for the amplitude a given the value of the maximum-likelihood estimator $\hat{a}$, which (as we discussed earlier) is a sufficient statistic for the data d. Recall that the posterior for a for this example is simply a truncated Gaussian from 0 to a_{max} centered on $\hat{a}$, which could be negative, see (3.36). The left two panels show the graphical construction of the Bayesian 90% upper limit and 95% credible interval for the amplitude a for the weak injection,

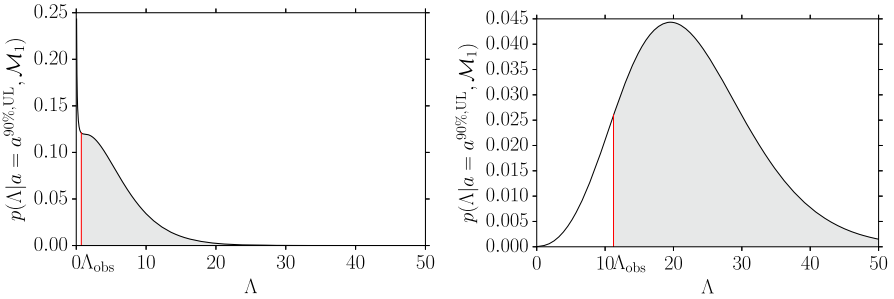

Fig. 14 Probability distributions for the frequentist detection statistic Λ, conditioned on the value of the amplitude a for which the probability of obtaining $\Lambda \geq \Lambda_{\mathrm{obs}}$ is equal to 0.90. These define the 90% confidence-level frequentist upper limits $a^{90\%,\mathrm{UL}} = 0.20$ and 0.46, respectively. The *red vertical lines* mark the value of Λ_{obs} for the weak (*left panel*, $\Lambda_{\mathrm{obs}} = 0.72$) and strong (*right panel*, $\Lambda_{\mathrm{obs}} = 11.2$) injections

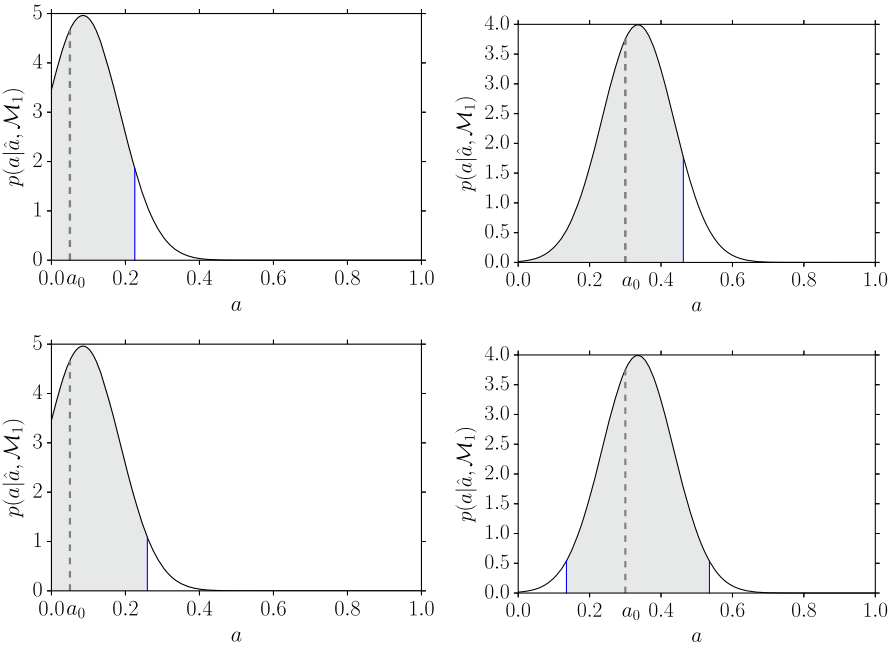

Fig. 15 Posterior distributions for the amplitude a given the value of the maximum-likelihood estimator $\hat{a}$. The *left two panels* are for the weak injection; the *right two panels* are for the strong injection. The *top two plots* illustrate the graphical construction of Bayesian 90% upper limits for the two injections; the *bottom two plots* illustrate the graphical construction of the Bayesian 95% credible intervals. The *dashed vertical lines* indicate the values of the injected signal amplitude a_0, which equal 0.05 and 0.3, respectively

$a^{90\%,\mathrm{UL}} = 0.23$ and [0, 0.26]. The right two panels show similar plots for the strong injection, $a^{90\%,\mathrm{UL}} = 0.46$ and [0.14, 0.54].

Finally, the Bayes factor for the signal-plus-noise model $\mathcal{M}_1$ relative to the noise-only model $\mathcal{M}_0$ can be calculated by taking the ratio of the marginalized likelihood $p(d|\mathcal{M}_1)$ given by (3.35) to $p(d|\mathcal{M}_0)$ given by (3.30). Doing this, we find 2 ln

Table 4 Tabular summary of the frequentist and Bayesian analysis results for the simulated data (both weak and strong injections)

	(Weak injection, $a_0 = -0.05$)		(Strong injection, $a_0 = 0.3$)	
	Frequentist	Bayesian	Frequentist	Bayesian
Detection threshold (Λ_*)	2.9	–	2.9	–
Detection statistic ($\Lambda_{\rm obs}$)	0.72	–	11.2	–
p-value	0.45	–	9.0×10^{-4}	–
90% upper limit	0.20	0.23	0.46	0.46
95% interval	$[-0.11, 0.29]$	$[0, 0.26]$	$[0.14, 0.54]$	$[0.14, 0.54]$
ML estimator ($\hat{a}$)	0.085	0.085	0.335	0.335
Bayes factor ($2 \ln \mathcal{B}_{10}$)	–	-2.2	–	9.2
Laplace approximation	–	-2.0	–	8.5

A dash indicates that a particular quantity is not relevant for either the frequentist or Bayesian analysis

$B_{10} = -2.2$ and 9.2 for the weak and strong signal injections, respectively. The Laplace approximation to this quantity is given by (3.41), with values -2.0 and 8.5, respectively.

3.5.4 Comparison summary

Table 4 summarizes the numerical results for the frequentist and Bayesian analyses. We see that the frequentist and Bayesian 90% upper limits and 95% intervals numerically agree for the strong injection, but differ slightly for the weak injection. The interpretation of these results is different, of course, for a frequentist and a Bayesian, given their different definitions of probability. But for a moderately strong signal in noisy data, we expect both approaches to yield a confident detection as they have for this simple example.

3.6 Likelihoods and priors for gravitational-wave searches

To conclude this section on statistical inference, we discuss some issues related to calculating the likelihood and choosing a prior in the context of searches for gravitational-wave signals using a network of gravitational-wave detectors.

3.6.1 Calculating the likelihood

Defining the likelihood function (for either a frequentist or Bayesian analysis) involves understanding the instrument response and the instrument noise. The data collected by gravitational-wave detectors comes in a variety of forms. For ground-based interferometers such as LIGO and Virgo, the data comes from the error signal in the differential arm-length control system, which is non-linearly related to the laser phase difference, which in turn is linearly related to the gravitational-wave strain. For pulsar timing arrays, the data comes from the arrival times of radio pulses (derived from

the folded pulse profiles), which must be corrected using a complicated timing model that takes into account the relative motion of the telescopes and the pulsars, along with the spin-down of the pulsars, in addition to a variety of propagation effects. The timing residuals formed by subtracting the timing model from the raw arrival times contain perturbations due to gravitational waves integrated along the line of sight to the pulsar. For future space-based gravitational-wave detectors such as LISA, the data will be directly read out from phase meters that perform a heterodyne measurement of the laser phase. Synthetic combinations of these phase read outs (chosen to cancel laser phase noise) are then linearly proportional to the gravitational-wave strain.

Since gravitational waves can be treated as small perturbations to the background geometry, the time delays or laser phase/frequency shifts caused by a gravitational wave can easily be computed. These idealized calculations have then to be related to the actual observations, either by propagating the effects through an instrument response model, or, alternatively, inverting the response model to convert the measured data to something proportional to the gravitational-wave strain. (For example, most LIGO analyses work with the calibrated strain, rather than the raw differential error signal). If we assume that the gravitational-wave signal and the instrument noise are linearly independent, then the data taken at time t can be written as

$$d(t) = h(t) + n(t), \tag{3.48}$$

where $h(t)$ is shorthand for the gravitational-wave metric perturbations $h_{ab}(t, \vec{x})$ convolved with the instrument response function and converted into the appropriate quantity—phase shift, time delay, differential arm length error, etc. (A detailed calculation of $h(t)$ and the associated detector response functions will be given in Sect. 5.2). As mentioned above, the data $d(t)$ may be the quantity that is measured directly, or, more commonly, some quantity that is derived from the measurements such as timing residuals or calibrated strain. In any analysis, it is important to marginalize over the model parameters used to make the conversion from the raw data.

The likelihood of observing $d(t)$ is found by demanding that the residual

$$r(t) \equiv d(t) - \bar{h}(t), \tag{3.49}$$

be consistent with a draw from the noise distribution $p_n(x)$:

$$p(d(t)|\bar{h}(t)) = p_n(r(t)) = p_n(d(t) - \bar{h}(t)). \tag{3.50}$$

Here $\bar{h}(t)$ is our model[6] for the gravitational-wave signal. The likelihood of observing a collection of discretely-sampled data $d \equiv \{d_1, d_2, \ldots, d_N\}$, where $d_i \equiv d(t_i)$, is then given by $p(d|\bar{h}) = p_n(r)$, where $r \equiv \{r_1, r_2, \ldots, r_N\}$ with $r_i \equiv r(t_i)$. Since instrument noise is due to a large number of small disturbances combined with counting noise in the large-number limit, the central limit theorem suggests that the noise distribution can be approximated by a multi-variate normal (Gaussian) distribution:

[6] Since the model $\bar{h}(t)$ will differ from the actual $h(t)$, we use an overbar for the model to distinguish the two.

$$p(d|\bar{h}) = \frac{1}{\sqrt{\det(2\pi C_n)}}\, e^{-\frac{1}{2}\sum_{i,j} r_i \left(C_n^{-1}\right)_{ij} r_j}, \tag{3.51}$$

where C_n is the noise correlation matrix, with components

$$(C_n)_{ij} = \langle n_i n_j \rangle - \langle n_i \rangle \langle n_j \rangle. \tag{3.52}$$

If the noise is stationary, then the correlation matrix only depends on the lag $|t_i - t_j|$, and the matrix C_n can be (approximately) diagonalized by transforming to the Fourier domain, where r_i should then be interpreted as $\tilde{r}(f_i)$ (see Appendix D.6 for a more careful treatment of discrete probability distributions in the time and frequency domain). In practice, the noise observed in most gravitational-wave experiments is neither stationary nor Gaussian (Sect. 9 and Appendix C), but (3.51) still serves as a good starting point for more sophisticated treatments. The Gaussian likelihood (3.51) immediately generalizes for a network of detectors:

$$p(d|\bar{h}) = \frac{1}{\sqrt{\det(2\pi C_n)}} e^{-\frac{1}{2}\sum_{Ii,Jj} r_{Ii} \left(C_n^{-1}\right)_{Ii,Jj} r_{Jj}}, \tag{3.53}$$

where I, J labels the detector, and i, j labels the discrete time or frequency sample for the corresponding detector. Note here that the parameters $\boldsymbol{\theta}$ appearing in (3.18) are the individual time or frequency samples $\bar{h}_i$.

3.6.2 Choosing a prior

For Bayesian inference, it is also necessary to define a model $\mathcal{M}$ for the gravitational-wave signal, which is done by placing a prior $p(\bar{h}|\mathcal{M})$ on the samples $\bar{h}_i$. In some cases, a great deal is known about the signal model, such as when approximate solutions to Einstein's equations provide waveform templates. In that case the prior can be written as

$$p(\bar{h}|\mathcal{M}) = \delta(\bar{h} - \bar{h}(\boldsymbol{\theta}, \mathcal{M}))\, p(\boldsymbol{\theta}|\mathcal{M}). \tag{3.54}$$

Marginalizing over $\bar{h}$ converts the posterior $p(\bar{h}|d)$ to a posterior distribution for the signal parameters $p(\boldsymbol{\theta}|d, \mathcal{M})$. In other cases, such as for short-duration bursts associated with certain violent astrophysical events, much less is known about the possible signals and weaker priors have to be used. Models using wavelets, which have finite time-frequency support, and priors that favor connected concentrations of power in the time-frequency plane are commonly used for these "unmodeled burst" searches. At the other end of the spectrum from deterministic point sources are the statistically-isotropic stochastic backgrounds that are thought to be generated by various processes in the early Universe, or through the superposition of a vast number of weak astrophysical sources. In the case of Gaussian stochastic signals, the prior for a signal $\bar{h} = (\bar{h}_+(\hat{n}), \bar{h}_\times(\hat{n}))$ coming from direction $\hat{n}$ direction has the form

$$p(\bar{h}|\mathcal{M}) = \frac{1}{2\pi S_h} e^{-(\bar{h}_+^2(\hat{n}) + \bar{h}_\times^2(\hat{n}))/2S_h}, \tag{3.55}$$

where S_h is the power spectrum of the background. As we shall show in Sect. 4, marginalizing over $\bar{h}$ converts the posterior $p(\bar{h}|d)$ to a posterior $p(S_h|d, \mathcal{M})$ for S_h.

4 Correlations

Correlation is not cause, it is just a 'music of chance'. *Siri Hustvedt*

Stochastic gravitational waves are indistinguishable from unidentified instrumental noise in a single detector, but are correlated between pairs of detectors in ways that differ, in general, from instrumental noise. Cross-correlation methods basically use the random output of one detector as a template for the other, taking into account the physical separation and relative orientation of the two detectors. In this section, we introduce cross-correlation methods in the context of both frequentist and Bayesian inference, analyzing in detail a simple toy problem (the data are "white" and we ignore complications that come from the separation and relative orientation of the detectors—this we discuss in detail in Sect. 5). We also briefly discuss possible alternatives to cross-correlation methods, e.g., using a null channel as a noise calibrator.

The basic idea of using cross-correlation to search for stochastic gravitational-waves can be found in several early papers (Grishchuk 1976; Hellings and Downs 1983; Michelson 1987; Christensen 1990, 1992; Flanagan 1993). The derivation of the likelihood function in Sect. 4.2 follows that of Cornish and Romano (2013); parts of Sect. 4.4 are also discussed in Allen et al. (2003) and Drasco and Flanagan (2003).

4.1 Basic idea

The key property that allows one to distinguish a stochastic gravitational-wave background from instrumental noise is that the gravitational-wave signal is correlated across multiple detectors while instrumental noise typically is not. To see this, consider the simplest possible example, i.e., a single sample of data from two colocated and coaligned detectors:

$$\begin{aligned} d_1 &= h + n_1, \\ d_2 &= h + n_2. \end{aligned} \tag{4.1}$$

Here h denotes the common gravitational-wave signal and n_1, n_2 the noise in the two detectors. To cross correlate the data, we simply form the product of the two samples, $\hat{C}_{12} \equiv d_1 d_2$. The expected value of the correlation is then

$$\langle \hat{C}_{12} \rangle = \langle d_1 d_2 \rangle = \langle h^2 \rangle + \langle n_1 n_2 \rangle + \cancelto{0}{\langle h n_2 \rangle} + \cancelto{0}{\langle n_1 h \rangle} = \langle h^2 \rangle + \langle n_1 n_2 \rangle, \tag{4.2}$$

since the gravitational-wave signal and the instrumental noise are uncorrelated. If the instrumental noise in the two detectors are also uncorrelated, then

$$\langle n_1 n_2 \rangle = 0, \tag{4.3}$$