Statistical Astronomy 1 in a nutshell

1 The meaning of 'probability'

The idea of probability has a number of interpretations and the word 'probability' means something slightly different in each. Much confusion can come from not appreciating this, so to summarise:

frequentist probability Here the probability of something is the limiting relative frequency of its occurrence in identical repeated trials. Only random variables (RVs) can have frequentist probabilities because only RVs return different values under the same conditions. Measurements made in an experiment are generally RVs, but the parameters of the system under test are not.

combinatorial probability Here we think of probability as number of favourable outcomes in which an event can occur divided by the total number of equivalent possible outcomes. This only works when you can enumerate the total number of outcomes, so is useful for the outcome of throwing a die, but not for (say) taking a length measurement.

Bayesian probability Here we think of probability as a number between 0 and 1 that quantifies our degree of belief in a proposition. A Bayesian probability may be numerically equal to the equivalent frequentist or combinatorial value under particular conditions, but it can also be computed in circumstances when the other two are meaningless. You can, for example, talk about the Bayesian probability of a parameter of a system, for example that Mars has three moons (although observational data indicates $P(\text{3-moons}) \ll 1$!). However, you cannot talk about the frequentist probability of this statement as Mars either has three moons or it doesn't, so the statement is not a random variable.

2 Manipulating probabilities

As well as the probability of a statement A, written Prob(A) or $P(A)$, we can talk about the **probability density function** (PDF) of a continuous variable x, written $p(x)$. This is defined so that the probability that x has a value between x and $x + \delta x$ is $p(x)\,\delta x$, as δx approaches zero. By definition, if x is well-defined then

$$\int p(x)\,\mathrm{d}x = 1,$$

where the integral is over all possible x values.

Frequentist, combinatorial and Bayesian probabilities and probability densities all follow the same basic sum and product rules. The table below show important relationships seen in Bayesian analysis:

probability:	$P(A)$
conditional probability:	$P(A \mid B)$
NOT *A*:	$\bar{A}$
prob of *A* OR *B*:	$P(A + B)$
prob of *A* AND *B*:	$P(A, B)$, alternatively $P(AB)$
sum rule:	$P(A) + P(\bar{A}) = 1$
extended sum rule:	$P(A + B) = P(A) + P(B) - P(A, B)$
product rule:	$P(A, B) = P(A)\,P(B \mid A) = P(B)\,P(A \mid B)$

We write the probability of A given that C is true as $P(A \mid C)$. This is the **conditional probability of *A* on *C***.

A and B are **independent** if $P(A \mid B) = P(A)$, and $P(B \mid A) = P(B)$. The product rule for independent quantities $A, B, C, D, \ldots$ reduces to

$$P(A, B, C, D, \ldots) = P(A)\,P(B)\,P(C)\,P(D)\ldots .$$

Joint probabilities are always commutative in the sense that $P(A, B) = P(B, A)$.

A and B are **mutually exclusive** if $P(A, B) = 0$. The extended sum rule for mutually exclusive quantities $A, B, C, D, \ldots$ simplifies to

$$P(A + B + C + D + \ldots) = P(A) + P(B) + P(C) + P(D) + \ldots$$

Cumulative probabilities represent the probability that a parameter is less than some value. If the parameter has a PDF $p(x)$ the cumulative probability to x_0 is

$$C(x_0) \equiv P(x < x_0) = \int_{-\infty}^{x_0} p(x)\,\mathrm{d}x.$$

Similarly, for a discrete probability distribution we sum the probabilities for $x \leq x_0$ to get the cumulative.

Marginal probabilities can be formed by integrating over one or more of the associated joint parameters of a PDF, so

$$p(x) = \int_{-\infty}^{\infty} p(x, y)\,\mathrm{d}y.$$

3 Moments, measures, and changes of variable

For a PDF $p(x)$ we can define the following useful quantities. Note that these are mathematical definitions and their statistical interpretation depends on the meaning of p:

mean: $\langle x \rangle = \int_{-\infty}^{\infty} x\,p(x)\,\mathrm{d}x$

mean square: $\langle x^2 \rangle = \int_{-\infty}^{\infty} x^2 p(x)\,\mathrm{d}x$

variance: $\mathrm{var}(x) = \langle x^2 \rangle - (\langle x \rangle)^2$

standard deviation: $\mathrm{SD}(x) = [\mathrm{var}(x)]^{1/2}$

mode: $p(x)$ has a maximum value at x_{mode}

median: $\int_{-\infty}^{x_{\mathrm{median}}} p(x)\,\mathrm{d}x = 0.5.$

If $z(x)$ is a unique functional mapping of x, then we can uniquely determine the PDF of z by change of variable:

$$p(z) = p(x)\left|\frac{\mathrm{d}x}{\mathrm{d}z}\right|.$$

More generally,

$$p(y_1, y_2, y_3, \ldots, y_n) = p(x_1, x_2, x_3, \ldots, x_n)\,|J|$$

where $|J|$ is the determinant of the Jacobian matrix

$$J_{ij} = \frac{\partial x_i}{\partial y_j}.$$

4 Probability distributions

We consider three important distributions:

4.1 Poisson distribution

The Poisson probability distribution,

$$P(N \mid \lambda) = \frac{\lambda^N \exp(-\lambda)}{N!},$$

appears quite often in astronomy. We can interpret it as the probability of seeing N events in any particular experiment if the expected number of events is λ. So, if the expected event rate is μ, then over a time τ we have $\lambda = \mu\tau$ and

$$P(N \mid \tau, \mu) = \frac{(\mu\tau)^N \exp(-\mu\tau)}{N!}.$$

Events are Poissonian if they (i) do not depend on each other in any way, (ii) are stationary (i.e., the average rate is a constant) and (iii) cannot occur simultaneously. The Poisson distribution is the limit of the binomial distribution as the number of binomial trials approaches infinity. The Poisson distribution is a 'discrete' distribution, that is it gives probabilities for discrete values of the integer N. As N gets large it converges on the (continuous) central distribution (see below).

4.2 Uniform distribution

The uniform distribution is the simplest continuous probability density function. A quantity x is uniformly distributed between a and b ($b > a$) if

$$p(x) = \begin{cases} 1/(b-a) & a < x < b \\ 0 & \text{otherwise.} \end{cases}$$

A uniform distribution for a quantity is sometimes an appropriate description of prior ignorance of its value. Formally we must define a and b for this PDF to be normalised. Sometimes it is acceptable to use the 'improper' (un-normalised) form $p(x) =$ constant, but take care!

4.3 Central distribution

Also known as the '**Gaussian**' or '**Normal**' distribution. For a continuous parameter x with mean μ and variance σ^2 it has the form

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

When used in a frequentist analysis, μ and σ are the mean and standard deviation of the random variable x. In a Bayesian context μ and σ represent the location and spread of our uncertainty, and the central distribution presents these two parameters with minimal further information content.

The central distribution can be extended to many dimensional parameters $\boldsymbol{x} = (x_1, x_2, x_3, \ldots, x_n)$ as

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\mathbb{C}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\mathbb{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

where $\boldsymbol{\mu}$ is the vector of means ($\mu_i = \langle x_i \rangle$) and $\mathbb{C}$ is the covariance matrix:

$$\mathbb{C}_{ij} = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle.$$

If the x_i are independent, and each has the same variance σ^2, this reduces to

$$p(\boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}|\boldsymbol{x}-\boldsymbol{\mu}|^2\right].$$

This form of the central distribution is seen quite often in data analysis. If $\boldsymbol{x}$ is a set of independent data measurements and $\boldsymbol{m}(\boldsymbol{a})$ a set of model predictions for these data, the likelihood (see below) of the model parameters $\boldsymbol{a}$ is

$$p(\boldsymbol{x} \mid \boldsymbol{a}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}|\boldsymbol{x}-\boldsymbol{m}(\boldsymbol{a})|^2\right].$$

The data $\boldsymbol{x}$ may have thousands, or even millions, of dimensions whereas there are commonly only 10 or fewer dimensions to the model parameters $\boldsymbol{a}$ in astronomical problems.

5 Bayes' Theorem

The product rule

$$P(A, B \mid C) = P(A \mid C)\, P(B \mid A, C)$$
$$P(B, A \mid C) = P(B \mid C)\, P(A \mid B, C)$$

together with commutation gives us *Bayes' Theorem*:

$$P(A \mid B, C) = \frac{P(A \mid C)\, P(B \mid A, C)}{P(B \mid C)}.$$

The importance of this relationship in Bayesian data analysis becomes clear when you consider A to be a statement about a system you are observing (e.g., "This rock has a mass of 2 g"), and B to be an experimental measurement of the system (e.g., "the mass of the rock as measured by my balance is 2.1 g"). $p(A \mid B)$ is the probability that the mass is 2 g when the balance says 2.1 g. It depends on $P(A)$ (your degree of belief in the mass being 2 g before the measurement), $P(B \mid A)$ (the probability that the balance would read 2.1 g if the true mass was 2 g) and $p(B)$ (the probability of measuring 2 g on the balance irrespective of A). C is assumed true throughout, and represents the background assumptions made in the analysis and not tested. We can usefully express Bayes' Theorem in terms of probability densities and give names to the three terms on the right-hand side:

$$p(\text{mass} \mid \text{measurement}) = \frac{p(\text{mass})\, p(\text{measurement} \mid \text{mass})}{p(\text{measurement})} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}.$$

The **prior probability** of the rock's mass represents how your degree of belief in its possible values is spread over those values before ('prior' to) making the measurement. The **likelihood** of the mass is the probability of getting the data you saw, given any particular true value of the rock's mass. This likelihood term encapsulates the uncertainty in your measurement method and/or instrumentation. Note the careful use of language here: $p(x \mid y)$ is the *likelihood* of y and the *probability density* of x; the same expression interpreted from two perspectives.

The **evidence** in the denominator does not depend on any particular value of the rock's mass. It is simply the integral of the likelihoods for each, weighted by the prior,

$$p(B \mid C) = \int p(A \mid C)\, p(B \mid A, C)\, \mathrm{d}A,$$

and can be thought of as the likelihood of your background assumptions after the experiment has been performed, or the probability of the data given those assumptions. The evidence can be largely ignored in simple parameter estimation problems because it is not a function of the parameter, and therefore plays the role of a normalising constant.

The expression we seek, $p(\text{mass} \mid \text{measurement})$ is the PDF of the mass parameter after ('a posteriori') the measurement. It is referred to as the **posterior probability** for the mass and is our updated version of the prior probability.

6 Practical parameter estimation

Parameter estimation is the process of inferring the probable values that parameters of a system can have when constrained by new data and prior information. You can think of the parameters

as the 'knobs' you can turn to adjust the model you are comparing to the data. The process of parameter estimation can be broken down into steps:

Identify the source of uncertainty In any inference problem there is a source of uncertainty. This can be through ignorance of some detail of the observation or experiment, or noise in a measurement.

Write down the likelihood of the parameters This is usually the stage that requires the most thought. The likelihood of the parameters is the probability (or PDF) of the data assuming values for the parameters. In the case of a noisy measurement, it is simply the probability/PDF that the difference between the model prediction and the data is entirely due to the noise, so you will need to assume an expression for the noise probability/ PDF. This makes up the bulk of the 'background assumptions' in the parameter estimation process, and it commonly takes the form of a Gaussian or Poisson distribution. Remember that noise in two measurements may not be independent, and this can add significant complexity to the likelihood. However, correlated noise can be handled quite straightforwardly if it is Gaussian, through the covariance matrix.

Identify the prior What do you already know about the parameters? You need to encapsulate this knowledge/ignorance in a prior probability/ PDF. Don't be shy: write down something with a justification. The worst that can happen is this prior becomes the new posterior, indicating you have learned nothing from the data.

Compute the posterior Algebraically, this is the easy step: the posterior is simply the normalised product of the prior and the likelihood you computed above. The posterior is a joint probability/PDF for all the parameters given the new data, and has the dimensions of the number of parameters. Although easy to write down algebraically, it can be hard to explore numerically.

Interpret the posterior Posteriors with more than two or three dimensions need to be summarised in some way, as we cannot fully represent them in a plot. Commonly we could generate a marginal PDF for each parameter and compute the mode/mean/median of each distribution as a point estimate of the parameter, together with the shortest credible interval. The extent of this interval defines the 'error bars' on the point estimate. Alternatively one could generate a 'corner plot' which shows the two-dimensional marginals of parameters taken in pairs. At the very least one could apply the 'Laplace approximation', expanding the log posterior, L, as a truncated Taylor series about the maximum, to give a Gaussian PDF, and compute the width of this Gaussian:

$$\sigma \simeq \left(- \left. \frac{\mathrm{d}^2 L}{\mathrm{d}x^2} \right|_{L_{\max}} \right)^{-1/2} .$$