ADVANCED STATISTICAL METHODS FOR COSMOLOGY AND ASTROPARTICLE PHYSICS

ROBERTO TROTTA, IMPERIAL COLLEGE LONDON

This is a summary handout of the lectures. While I do not intend to cover all of the material in detail, the hope is that this handout will give you the necessary information to fill in the unavoidable gaps that will be left by the lectures themselves. Additional examples and applications in cosmology are given during the lectures. An exercice sheet complements this handout and is designed to give you a hands-on experience of some of the concepts covered here.

Notice that the list of references has been kept very minimal in this handout. Please refer to [13] for a far more complete overview of the literature. Useful reference textbooks are Refs. [1, 2, 3, 4, 5, 6, 7, 8]; applications to cosmology and astrophysics can be found in e.g. [9, 10, 11, 12, 13].

Comments, suggestions and corrections are very welcome! Please contact me by e-mail: r.trotta@imperial.ac.uk

Contents

Date: Copenhagen, Nov 2011.

Guildenstern — The law of probability, as it has been oddly asserted, is something to do with the proposition that [...], if I have got this right, if six monkeys were thrown up in the air for long enough they would land on their tails about as often as they would land on their...

Rosencratz — Heads.

Tom Stoppard, *Rosencratz and Guildenstern are dead* (1966)

1. Foundational aspects

- Any physical theory must be validated against observations. The cycle theory → prediction → observation → theory forms the pillar of the scientific process.
- The comparison between theory and observation is the key phase when statistical methods are needed: *how do we learn about the world from a collection of noisy observations?*
- Statistics is at the heart of the scientific process, not merely an optional nuisance. Ernest Rutherford is reported to have said: "If you need statistics, you did the wrong experiment". This completely misses the point: if you do not need statistics, it's because you are doing the wrong kind of physics! Five reasons why you need statistics:
  - (i) The complexity of the modelling of both our theories and observations will always increase, thus requiring correspondingly more refined statistical and data analysis skills. In fact, the scientific return of the next generation of surveys will be limited by the level of sophistication and efficiency of our inference tools.
  - (ii) The discovery zone for new physics is when a potentially new effect is seen at the 3–4 σ level. This is when tantalizing suggestions for an effect start to accumulate but there is no firm evidence yet. In this potential discovery region a careful application of statistics can make the difference between claiming or missing a new discovery.
  - (iii) If you are a theoretician, you do not want to waste your time trying to explain an effect that is not there in the first place. A better appreciation of the interpretation of statistical statements might help in identifying robust claims from spurious ones.
  - (iv) Limited resources mean that we need to focus our efforts on the most promising avenues. Experiment forecast and optimization will increasingly become prominent as we need to use all of our current knowledge (*and* the associated uncertainty) to identify the observations and strategies that are likely to give the highest scientific return in a given field.
  - (v) Sometimes there will be no better data! This is the case for the many problems associated with cosmic variance limited measurements on large scales, for example in the cosmic background radiation, where the small number of independent directions on the sky makes it impossible to reduce the error below a certain level.

1.1. The meaning of probability.

- There are two different ways of understanding what probability is. The **classical (so-called "frequentist") notion of probability** is that probabilities are tied to the frequency of outcomes over a long series of trials. Repeatability of an experiment is the key concept. Most of this course will follow this notion.

  The **Bayesian outlook**[1] is that probability expresses a degree of belief in a proposition, based on the available knowledge of the experimenter. Information is the key concept. Bayesian probability theory is more general than frequentist theory, as the former can deal with unique situations that the latter cannot handle (e.g., "what is the probability that it will rain tomorrow?). Bayesian probability is briefly discussed at the end of the course.

[1] So-called after Rev. Thomas Bayes (1701(?)–1761), who was the first to introduce this idea in a paper published posthumously in 1763, "An essay towards solving a problem in the doctrine of chances".

- Let $A, B, C, \ldots$ denote propositions (e.g., that a coin toss gives tails). Let Ω describe the **sample space (or state space)** of the experiment, i.e., Ω is a list of all the possible outcomes of the experiment. If we are tossing a coin, $\Omega = \{T, H\}$, where T denotes "tails" and H denotes "head". If we are tossing a die, $\Omega = \{1, 2, 3, 4, 5, 6\}$. If we are drawing one ball from an urn containing white and black balls, $\Omega = \{W, B\}$, where W denotes a white ball and B a black ball.
- **Frequentist definition of probability:** The number of times an event occurs divided by the total number of events in the limit of an infinite series of equiprobable trials.
- The **joint probability** of A and B is the probability of A and B happening together, and is denoted by $P(A, B)$.

  The **conditional probability** of A given B is the probability of A happening given that B has happened, and is denoted by $P(A|B)$.
- The sum rule:

$$P(A) + P(\overline{A}) = 1, \tag{1}$$

where $\overline{A}$ denotes the proposition "not A".

The product rule:

$$P(A, B) = P(A|B)P(B). \tag{2}$$

By inverting the order of A and B we obtain that

$$P(B, A) = P(B|A)P(A) \tag{3}$$

and because $P(A, B) = P(B, A)$, we obtain **Bayes theorem** by equating Eqs. (46) and (47):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{4}$$

The marginalisation rule (follows from the two rules above):

$$P(A) = P(A, B_1) + P(A, B_2) + \cdots = \sum_i P(A, B_i) = \sum_i P(A|B_i)P(B_i), \tag{5}$$

where the sum is over all possible outcomes for proposition B.

- Two propositions (or events) are said to be **independent** if and only if

$$P(A, B) = P(A)P(B). \tag{6}$$

2. Random variables, parent distributions and samples

- A **random variable** (RV) is a function mapping the sample space Ω of possible outcomes of a random process to the space of real numbers.

  When tossing a coin once, the RV X can be defined as

$$X = \begin{cases} 0, & \text{if coin lands T} \\ 1, & \text{if coin lands H.} \end{cases} \tag{7}$$

When tossing a die, the RV X can be defined as

$$X = \begin{cases} 1, & \text{if a 1 is rolled} \\ 2, & \text{if a 2 is rolled} \\ 3, & \text{if a 3 is rolled} \\ 4, & \text{if a 4 is rolled} \\ 5, & \text{if a 5 is rolled} \\ 6, & \text{if a 6 is rolled.} \end{cases} \tag{8}$$

When drawing one ball from an urn containing black and white balls, the RV X can be defined as

$$X = \begin{cases} 0, & \text{if the ball drawn is white} \\ 1, & \text{if the ball drawn is black.} \end{cases} \tag{9}$$

A RV can be discrete (only a countable number of outcomes is possible, such as in coin tossing) or continuous (an uncountable number of outcomes is possible, such as in a temperature measurement). It is mathematically subtle to carry out the passage from a discrete to a continuous RV, although as physicists we won't bother too much with mathematical rigour.

- Each RV has an associated **probability distribution** to it. The probability distribution of a discrete RV is called **probability mass function** (pmf), which gives the probability of each outcome: $P(X = x_i) = P_i$ gives the probability of the RV X assuming the value x_i. In the following we shall use the shorthand notation $P(x_i)$ to mean $P(X = x_i)$.
  If X is the RV of Eq. (8), and the die being tossed is fair, then $P_i = 1/6$ for $i = 1,\dots,6$, where x_i is the outcome "a the face with i pips comes up".

  The probability distribution associated with a continuous RV is called the **probability density function** (pdf), denoted by $p(X)$. The quantity $p(x)dx$ gives the probabilty that the RV X assumes the value between x and $x + dx$.

  The choice of probability distribution to associate to a given random process is dictated by the nature of the random process one is investigating (examples follow below).
- For a discrete pmf, the **cumulative probability distribution function** (cdf) is given by

$$C(x_i) = \sum_{j=1}^{i} P(x_j). \tag{10}$$

The cdf gives the probabilty that the RV X takes on a value less than or equal to x_i, i.e. $C(x_i) = P(X \le x_i)$.

For a continuous pdf, the cdf is given by

$$P(x) = \int_{-\infty}^{x} p(y)dy, \tag{11}$$

with the same interpretation as above, i.e. it is the probability that the RV X takes a value smaller than x.

- When we make a measurement, (e.g., the temperature of an object, or we toss a coin and observe which face comes up), nature selects an outcome from the sample space with probability given by the associated pmf or pdf. The selection of the outcome is such that if the measurement was repeated an infinite number of times the relative frequency of each outcome is the same as the the probability associated with each outcome under the pmf or pdf (this is another formulation of the frequentist definition of probability given above).
- Outcomes of measurements realized by nature are called **samples**. They are a series of real numbers, $\{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_N\}$ (usually denoted by symbols with a hat in this course).

2.1. **The likelihood function.**

3. The likelihood function

- The problem of **inference** can be stated as follows: given a collection of samples, $\{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_N\}$, and a generating random process, what can be said about the properties of the underlying probability distribution?
  You toss a coin 5 times and obtain 1 head. What can be said about the fairness of the coin?
  With a photon counter you observe 10 photons in a minute. What can be said about the average photon rate from the source?
  You measure the temperature of an object twice with two different instruments, yielding the following measurements: $T = 256 \pm 10$ K and $T = 260 \pm 5$ K. What can be said about the temperature of the object?
- Schematically, we have that:

$$\begin{gathered}\text{pdf - e.g., Gaussian with a given } (\mu, \sigma) \rightarrow \text{Probability of observation} \\ \text{Underlying } (\mu, \sigma) \leftarrow \text{Observed events}\end{gathered} \tag{12}$$

The connection between the two domains is given by the **likelihood function**.

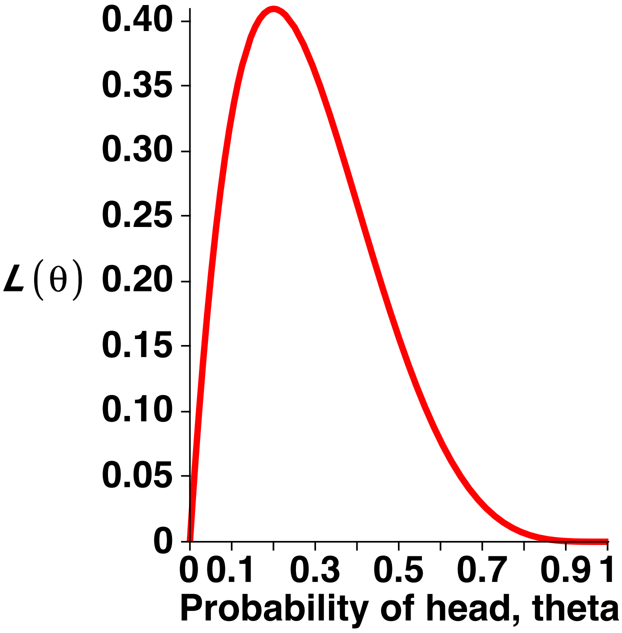

FIGURE 1. The likelihood function for the probability of heads (θ) for the coin tossing example, with $n = 5, r = 1$.

- Given a pdf or a pmf $p(X|\theta)$, where X represents a random variable and θ a collection of parameters describing the shape of the pdf[2] and the observed data $\hat{\mathbf{x}} = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_N\}$, the **likelihood function** $\mathcal{L}$ (or "likelihood" for short) is defined as

$$\mathcal{L}(\theta) = p(X = \hat{\mathbf{x}}|\theta) \tag{13}$$

i.e., the probability, **as a function of the parameters** θ, of observing the data that have been obtained. Notice that the likelihood is **not** a pdf in θ.

- In tossing a coin, let θ be the probability of obtaining heads in one throw (what we denoted previously by p. Now we use θ instead, in order to make contact with the more general formalism introduced above. Don't let yourself be thrown by the slightly different notation!). Suppose we make $n = 5$ flips and obtain the sequence $\hat{\mathbf{x}} = \{H, T, T, T, T\}$. The likelihood is obtained by taking the binomial, Eq. (84), and replacing for r the number of heads obtained ($r = 1$) in $n = 5$ trials, and looking at it **as a function of the parameter we are interested in determining, here** θ. Thus

$$\mathcal{L}(\theta) = \binom{5}{1}\theta^1(1-\theta)^4 = 5\theta(1-\theta)^4, \tag{14}$$

which is plotted as a function of θ in Fig. 1.

If instead of $r = 1$ heads we had obtained a different number of heads in our $n = 5$ trials, the likelihood function would have looked as shown in Fig. 2 for a few choices for r.

- This example leads to the formulation of the Maximum Likelihood Principle (see below): if we are trying to determine the value of θ given what we have observed (the sequence of H/T), we should choose the value that maximises the likelihood. Notice that this is **not** necessarily the same as maximising the probability of θ. Doing so requires the use of Bayes theorem, see section 3.3.

3.1. The Maximum Likelihood Principle.

[2] For example, for a Gaussian $\theta = \{\mu, \sigma\}$, for a Poisson distribution, $\theta = \lambda$ and for a binomial distribution, $\theta = p$, the probability of success in one trial.

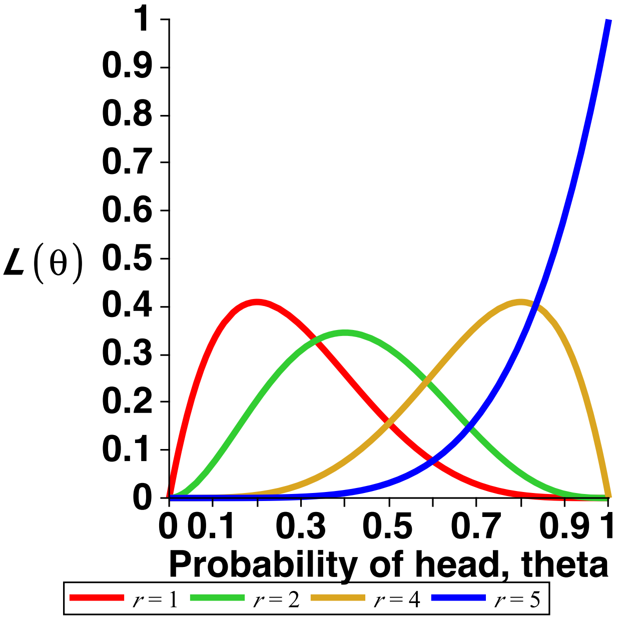

FIGURE 2. The likelihood function for the probability of heads (θ) for the coin tossing example, with $n = 5$ trials and different values of r.

- **The Maximum Likelihood Principle** (MLP): given the likelihood function $\mathcal{L}(\theta)$ and seeking to determine the parameter θ, we should choose the value of θ in such a way that the value of the likelihood is maximised. The Maximum Likelihood Estimator (MLE) for θ is thus

$$\theta_{\rm ML} \equiv \max_{\theta} \mathcal{L}(\theta) \tag{15}$$

- Properties of the MLE: it is asymptotically unbiased (i.e., $\theta_{\rm ML} \to \theta$ for $N \to \infty$, i.e., the ML estimate converges to the true value of the parameters for infinitely many data points) and it is asymptotically the minimum variance estimator, i.e. the one with the smallest errors.
- To find the MLE, we maximise the likelihood by requiring its first derivative to be zero and the second derivative to be negative:

$$\left.\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\right|_{\theta_{\rm ML}} = 0, \qquad \text{and} \qquad \left.\frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta^2}\right|_{\theta_{\rm ML}} < 0. \tag{16}$$

In practice, it is often more convenient to maximise the logarithm of the likelihood (the "log-likelihood") instead. Since log is a monotonic function, maximising the likelihood is the same as maximising the log-likelihood. So one often uses

$$\left.\frac{\partial \ln \mathcal{L}(\theta)}{\partial \theta}\right|_{\theta_{\rm ML}} = 0, \qquad \text{and} \qquad \left.\frac{\partial^2 \ln \mathcal{L}(\theta)}{\partial \theta^2}\right|_{\theta_{\rm ML}} < 0. \tag{17}$$

- **MLE of the mean of a Gaussian.** Imagine we have done N independent measurements of a Gaussian-distributed quantity, and let's denote them by $\{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_N\}$. Here the parameters we are interested in determining are μ (the mean of the distribution) and σ (the standard deviation of the distribution), hence we write $\theta = \{\mu, \sigma\}$.Then the joint likelihood function is given by

$$\mathcal{L}(\mu,\sigma) = p(\hat{\mathbf{x}}|\mu,\sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(\hat{x}_i - \mu)^2}{\sigma^2}\right), \tag{18}$$

Note: often the Gaussian above is written as

$$\mathcal{L} = L_0 \exp\left(-\chi^2/2\right) \tag{19}$$

where the so-called "chi-squared" is defined as

(20) $$\chi^2 = \sum_{i=1}^{N} \frac{(\hat{x}_i - \mu)^2}{\sigma^2}.$$

We want to estimate the (true) mean of the Gaussian. The MLE for the mean is obtained by solving

(21) $$\frac{\partial \ln \mathcal{L}}{\partial \mu} = 0 \Rightarrow \mu_{\rm ML} = \frac{1}{N}\sum_{i=1}^{N} \hat{x}_i,$$

i.e., the MLE for the mean is just the sample mean (i.e., the average of the measurements).

- **MLE of the standard deviation of a Gaussian.** If we want to estimate the standard deviation σ of the Gaussian, the MLE for σ is:

(22) $$\frac{\partial \ln \mathcal{L}}{\partial \sigma} = 0 \Rightarrow \sigma^2_{\rm ML} = \frac{1}{N}\sum_{i=1}^{N} (\hat{x}_i - \mu)^2.$$

However, the MLE above is "biased", i.e. it can be shown that

(23) $$E(\sigma^2_{\rm ML}) = (1 - \frac{1}{N})\sigma^2 \neq \sigma^2,$$

i.e., for finite N the expectation value of the ML estimator is not the same as the true value, σ^2. In order to obtain an unbiased estimator we replace the factor $1/N$ by $1/(N-1)$. Also, because the true μ is usually unknown, we replace it in Eq. (22) by the MLE estimator for the mean, $\mu_{\rm ML}$.

Therefore, **the unbiased MLE estimator for the variance** is

(24) $$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N} (\hat{x}_i - \mu_{\rm ML})^2.$$

In general, you should always use Eq. (24) as the ML estimator for the variance (and **not** Eq. (22)).
A numerical application of the above results. The surface temperature on Mars is measured by a probe 10 times, yielding the following data (units of K):

(25) $$197.2, 202.4, 201.8, 198.8, 207.6, 191.4, 201.4, 198.2, 195.7, 201.2.$$

Assuming that each measurement is independently Gaussian distributed with known variance $\sigma^2 = 5$ K^2, what is the likelihood function for the whole data set?
Answer: the measurements are independent, hence the total likelihood is the product of the likelihoods for each measurement, see Eq. (18):

(26) $$\mathcal{L}(T) = \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(\hat{T}_i - T)^2}{\sigma^2}\right)$$

What is the MLE of the mean, T_{ML}?
Answer: the MLE for the mean of a Gaussian is given by the mean of the sample, see Eq. (21), hence

(27) $$T_{\rm ML} = \frac{1}{10}\sum_{i=1}^{10} \hat{T}_i = 199.6K.$$

- **MLE for the success probability of a binomial distribution.** We go back to the coin tossing example, but this time we solve it in all generality. Let's define "success" as "the coin lands heads" (H). Having observed H heads in a number n of trials, the likelihood function of a binomial is given by (see Eq. (84)), again using the notation $\theta = p$, where the unknown parameter is p (the success probability for one trial, i.e., the probability that the coin lands H):

(28) $$\mathcal{L}(p) = P(H|p,n) = \binom{n}{H} p^H (1-p)^{n-H},$$

The Maximum Likelihood Estimator the success probability is found by maximising the log likelihood:

(29)
$$\frac{\partial \ln \mathcal{L}(p)}{\partial p} = \frac{\partial}{\partial p}\left(\ln\binom{n}{H} + H\ln p + (n-H)\ln(1-p)\right) = \frac{H}{p} - \frac{n-H}{1-p} \overset{!}{=} 0$$
$$\Leftrightarrow p_{\mathrm{ML}} = \frac{H}{n}.$$

Thus the MLE is simpy given by the observed fraction of heads, which is intuitively obvious.

- **MLE for the rate of a Poisson distribution.** The likelihood function is given by (see Eq. (87)), using the notation $\theta = \lambda$ (i.e., the parameter θ we are interested in is here the rate λ):

(30)
$$\mathcal{L}(\lambda) = P(n|\lambda) = \frac{(\lambda t)^n}{n!}\exp(-\lambda t),$$

The unknown parameter is the rate λ, while the data are the observed counts, n, in the amount of time t. The Maximum Likelihood Estimate for λ is obtained by finding the maximum of the log likelihood as a function of the parameter (here, the rate λ). Hence we need to find the value of λ such that:

(31)
$$\frac{\partial \ln P(n|\lambda)}{\partial \lambda} = 0.$$

The derivative gives

(32)
$$\frac{\partial \ln P(n|\lambda)}{\partial \lambda} = \frac{\partial}{\partial \lambda}(n\ln(\lambda t) - \ln n! - \lambda t) = n\frac{t}{\lambda t} - t = 0 \Leftrightarrow \lambda_{MLE} = \frac{n}{t}.$$

So the maximum likelihood estimator for the rate is the observed average number of counts.

- **MLE recipe:**
  (i) Write down the likelihood. This depends on the kind of random process you are considering. Identify what is the parameter that you are interested in, θ.
  (ii) Find the "best fit" value of the parameter of interest by maximising the likelihood $\mathcal{L}$ as a function of θ. This is your MLE, θ_{ML}.
  (iii) Evaluate the uncertainty on θ_{ML}, i.e. compute the confidence interval (see next section).

3.2. **Frequentist confidence intervals.** *Confidence is what you have before you understand the problem.*

Woody Allen

- Consider a general likelihood function, $\mathcal{L}(\theta)$ and let us do a Taylor expansion of the log-likelihood $\ln\mathcal{L}$ around its maximum, given by θ_{ML}:

(33)
$$\ln\mathcal{L}(\theta) = \ln\mathcal{L}(\theta_{\mathrm{ML}}) + \left.\frac{\partial \ln\mathcal{L}(\theta)}{\partial\theta}\right|_{\theta_{\mathrm{ML}}}(\theta-\theta_{\mathrm{ML}}) + \frac{1}{2}\left.\frac{\partial^2\ln\mathcal{L}(\theta)}{\partial\theta^2}\right|_{\theta_{\mathrm{ML}}}(\theta-\theta_{\mathrm{ML}})^2 + \ldots$$

The second term on the RHS vanishes (by definition of the Maximum Likelihood value), hence we can approximate the likelihood as

(34)
$$\mathcal{L}(\theta) \approx \mathcal{L}(\theta_{\mathrm{ML}})\exp\left(-\frac{1}{2}\frac{(\theta-\theta_{\mathrm{ML}})^2}{\Sigma_\theta^2}\right) + \ldots,$$

with

(35)
$$\frac{1}{{\Sigma_\theta}^2} = -\left.\frac{\partial^2\ln\mathcal{L}(\theta)}{\partial\theta^2}\right|_{\theta_{\mathrm{ML}}}.$$

So a general likelihood function can be approximated as a Gaussian around the ML value, as shown by Eq. (34). Therefore, to the extent that a probability distribution can be approximated as a Gaussian around its peak, the uncertainty around the ML value, Σ_θ, is approximately given by Eq. (35).

- Let's go back to the Gaussian problem of Eq. (18). We have seen in Eq. (21) that the sample mean is the MLE for the mean of the Gaussian. We now want to compute the uncertainty on this value. Applying Eq. (35) to the likelihood of Eq. (18) we obtain

$$\Sigma_\mu^2 = \sigma^2/N. \tag{36}$$

  This means that the the uncertainty on our ML estimate for μ (as expressed by the standard deviation Σ_μ) is proportional to $1/\sqrt{N}$, with N being the number of measurements.
- Going back to the numerical example of Eq. (25), we now wish to estimate the uncertainty on our MLE for the mean. The variance of the mean is given by $\Sigma_\mu^2 = \sigma^2/N$, where $\sigma^2 = 5$ K^2 and $N = 10$. Therefore the standard deviation of our temperature estimate $T_{\rm ML}$ is given by $\Sigma_T = 5/\sqrt{10} = 1.6$ K. The measurement can thus be summarized as $T = 199.6 \pm 1.6$ K, where the ± 1.6 K gives the range of the 1σ (or 68.3%) confidence interval (see below).
- As the likelihood function can be approximated as a Gaussian (at least around the peak), we can use the results for a Gaussian distribution to approximate the probability content of an interval around the ML estimate for the mean. The interval $[\mu_{\rm min}, \mu_{\rm max}]$ is called a 100α% **confidence interval** for the mean μ if $P(\mu_{\rm min} < \mu < \mu_{\rm max}) = \alpha$.

  So, for example, the interval $[\mu_{\rm ML} - \Sigma_\mu < \mu < \mu_{\rm ML} + \Sigma_\mu]$ is a 68.3% confidence interval for the mean (a so-called "1σ interval"), while $[\mu_{\rm ML} - 2\Sigma_\mu < \mu < \mu_{\rm ML} + 2\Sigma_\mu]$ is a 95.4% confidence interval (a "2σ interval").

  In the temperature measurement example of Eq. (25), the 68.3% confidence interval for the mean is $198.0 \text{ K} < \mu < 201.2$ K. The 95.4% confidence interval is $196.4 \text{ K} < \mu < 202.8$ K.
- Generally, the value after the "±" sign will usually give the 1σ (i.e., 68.3%) region. Sometimes you might find a notation like 50 ± 10 (95% CL), where "CL" stands for "Confidence Level". In this case, ±5 encompasses a region of 95% confidence (rather than 68.3%), which corresponds to 1.96 σ (see Table 2).
- One has to be careful with the interpretation of confidence intervals as this is often misunderstood! **Interpretation:** if we were to repeat an experiment many times, and each time report the observed 100α% confidence interval, we would be correct 100α% of the time. This means that (ideally) a 100α% confidence intervals contains the true value of the parameter 100α% of the time.
- In a frequentist sense, it does **not** make sense to talk about "the probability of θ". This is because every time the experiment is performed we get a different realization (different samples), hence a different numerical value for the confidence interval. Each time, either the true value of θ is inside the reported confidence interval (in which case, the probability of θ being inside is 1) or the true value is outside (in which case its probability of being inside is 0). **Confidence intervals do not give the probability of the parameter!** In order to do that, you need Bayes theorem.

  New Year Test sample question. The surface temperature on Mars is measured by a probe 10 times, yielding the following data (units of K):

$$191.9, 201.6, 206.1, 200.4, 203.2, 201.6, 196.5, 199.5, 194.1, 202.4 \tag{37}$$

  (i) Assume that each measurement is independently Normally distributed with known variancee $\sigma^2 = 25$ K^2. What is the likelihood function for the whole data set?

  Answer: The measurements are independent, hence the total likelihood is the product of the likelihoods for each measurement:

$$\mathcal{L}_{\rm tot}(T) = \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(\hat{T}_i - T)^2}{\sigma^2}\right) \tag{38}$$

  where $\hat{T}_i$ are the data given, T is the temperature we are trying to determine (unknown parameter) and $\sigma = 5$ K.

  (ii) Find the Maximum Likelihood Estimate (MLE) for the surface temperature, $T_{\rm ML}$, and express your result to 4 significant figures accuracy.

  Answer: The MLE for the mean of a Gaussian is given by the mean of the sample, see Eq. (21), hence

(39)
$$T_{\mathrm{ML}} = \frac{1}{10}\sum_{i=1}^{10} T_i = 199.7\mathrm{K}.$$

(iii) Determine symmetric confidence intervals at 68.3%, 95.4% and 99% around T_{ML} (4 significant figures accuracy).
Answer: The variance of the mean is given by σ^2/N, see Eq. (36). Therefore the standard deviation of our estimate T_{ML} is given by $\Sigma_T = \sigma/\sqrt{N} = 5/\sqrt{10} = 1.58$ K, which corresponds to the 68.3% interval: 199.7 ± 1.6 K, i.e. the range $[198.1, 201.3]$ K (4 s.f. accuracy). Confidence intervals at 95.4% and 99% corresponds to symmetric intervals around the mean of length 2.0 and 2.57 times the standard deviation Σ_T. Hence the required confidence intervals are $[196.5, 202.9]$ K (95.4%) and $[195.6, 203.8]$ K (99%).

(iv) How many measurements would you need to make if you wanted to have a 1σ confidence interval around the mean of length less than 1 K (on each side)?
Answer: A 1σ confidence interval lenght 1 K means that the value of Σ_T should be 1 K. Using that the standard deviation scales as $1/\sqrt{N}$, we have

(40)
$$1 = 5/\sqrt{N} \Rightarrow N = 25.$$

You would need $N = 25$ measurements to achieve the desired accuracy.

A laser beam is used to measure the deviation of the distance between the Earth and the Moon from its average value, giving the following data, in units of cm:

(41)
$$119, \quad 119, \quad 122, \quad 121, \quad 116.$$

(i) Assuming that each measurement above follows an independent Gaussian distribution of known standard deviation $\sigma = 3$ cm, write down the joint likelihood function for Δ, the deviation of the Earth-Moon distance from its average value.
Answer: The joint Gaussian likelihood function for Δ is given by

(42)
$$P(\Delta|d) \equiv \mathcal{L}(\Delta) = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(\Delta - d_i)^2}{\sigma^2}\right),$$

where $\sigma = 3$ cm and d_i are the measurements given in the question.

(ii) Compute the maximum likelihood estimate for Δ and its uncertainty, both to 3 significant figures.
Answer: The maximum likelihood estimate for Δ is found by maximising the log-likelihood function wrt Δ:

(43)
$$\frac{\partial \ln \mathcal{L}}{\partial \Delta} = -\sum_{i=1}^{5} \frac{\Delta - d_i}{\sigma^2} = 0 \rightarrow \Delta_{\mathrm{MLE}} = \frac{1}{N}\sum_{i=1}^{5} d_i$$

The numerical value is $\Delta_{\mathrm{MLE}} = 119.4 cm \approx 119$ (cm, 3 s.f.).

The uncertainty Σ on Δ is estimated from the inverse curvature of the log likelihood function at the MLE point:

(44)
$$-\frac{\partial^2 \ln \mathcal{L}}{\partial \Delta^2} = \frac{N\Delta}{\sigma^2} \rightarrow \Sigma = \left(-\frac{\partial^2 \ln \mathcal{L}}{\partial \Delta^2}\right)^{-1/2} = \frac{\sigma}{\sqrt{N}}$$

Numerically this gives $\Sigma = 3/\sqrt{5} = 1.34 \approx 1$ cm.

(iii) How would you report the measurement of Δ?
Answer: The measurement of Δ would thus be reported as $\Delta = (119 \pm 1)$ cm.

3.3. Bayes theorem.

- Bayes theorem, Eq. (48), encapsulates the notion of *probability as degree of belief.* The Bayesian outlook on probability is more general than the frequentist one, as the former can deal with unrepeatable situations that the latter cannot address. We begin with some simple definitions and consequences.

- The *joint probability* of A and B is the probability of A and B happening together, and is denoted by $P(A,B)$.

  The *conditional probability* of A given B is the probability of A happening given that B has happened, and is denoted by $P(A|B)$.
- The sum rule:

$$P(A)+P(\overline{A})=1, \tag{45}$$

  where $\overline{A}$ denotes the proposition "not A".
- The product rule:

$$P(A,B)=P(A|B)P(B). \tag{46}$$

  By inverting the order of A and B we obtain that

$$P(B,A)=P(B|A)P(A) \tag{47}$$

  and because $P(A,B)=P(B,A)$, we obtain, by equating Eqs. (46) and (47):
- *Bayes theorem*

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}. \tag{48}$$

- The marginalisation rule (follows from the two rules above):

$$P(A)=P(A,B_1)+P(A,B_2)+\cdots=\sum_i P(A,B_i)=\sum_i P(A|B_i)P(B_i), \tag{49}$$

  where the sum is over all possible outcomes for proposition B.
- Two propositions (or events) are said to be *independent* if and only if

$$P(A,B)=P(A)P(B). \tag{50}$$

- We replace in Bayes theorem, Eq. (48), $A\to\theta$ (the parameters) and $B\to d$ (the observed data, or samples), obtaining

$$P(\theta|d)=\frac{P(d|\theta)P(\theta)}{P(d)}. \tag{51}$$

  On the LHS, $P(\theta|d)$ is *the posterior probability for* θ (or "posterior" for short), and it represents our degree of belief about the value of θ after we have seen the data d.

  On the RHS, $P(d|\theta)=\mathcal{L}(\theta)$ is the likelihood we already encountered. It is the probability of the data given a certain value of the parameters. The quantity $P(\theta)$ is *the prior probability distribution* (or "prior" for short). It representes our degree of belief in the value of θ before we see the data. This is an essential ingredient of Bayesian statistics. In the denominator, $P(d)$ is a normalizing constant (often called "the evidence"), which ensures that the posterior is normalized to unity:

$$P(d)=\int d\theta P(d|\theta)P(\theta). \tag{52}$$

  The evidence is important for Bayesian model selection (see section 5).
- **Interpretation:** Bayes theorem relates the posterior probability for θ (i.e., what we know about the parameters after seeing the data) to the likelihood. It can be thought of as a general rule to update our knowledge about a quantity (here, θ) from the prior to the posterior. A result known as Cox theorem shows that Bayes theorem is the unique generalization of boolean algebra in the presence of uncertainty.
- Remember that in general $P(\theta|d)\neq P(d|\theta)$ (see ex. of pregnant woman!), i.e. the posterior and the likelihood are two different quantities with different meaning!

4. Bayesian inference

4.1. **Advantages of the Bayesian approach.** Irrespectively of the philosophical and epistemiological views about probability, as physicists we might as well take the pragmatic view that the approach that yields demonstrably superior results ought to be preferred. In many real–life cases, there are several good reasons to prefer a Bayesian viewpoint:

(i) Classic frequentist methods are often based on asymptotic properties of estimators. Only a handful of cases exist that are simple enough to be amenable to analytic treatment (in physical problems one most often encounters the Normal and the Poisson distribution). Often, methods based on such distributions are employed not because they accurately describe the problem at hand, but because of the lack of better tools. This can lead to serious mistakes. Bayesian inference is not concerned by such problems: it can be shown that *application of Bayes' Theorem recovers frequentist results (in the long run) for cases simple enough where such results exist*, while remaining applicable to questions that cannot even be asked in a frequentist context.

(ii) Bayesian inference deals effortlessly with *nuisance parameters*. Those are parameters that have an influence on the data but are of no interest for us. For example, a problem commonly encountered in astrophysics is the estimation of a signal in the presence of a background rate The particles of interest might be photons, neutrinos or cosmic rays. Measurements of the source s must account for uncertainty in the background, described by a nuisance parameter b. The Bayesian procedure is straightforward: infer the joint probability of s and b and then integrate over the uninteresting nuisance parameter b ("marginalization", see Eq. (63)). Frequentist methods offer no simple way of dealing with nuisance parameters (the very name derives from the difficulty of accounting for them in classical statistics). However neglecting nuisance parameters or fixing them to their best–fit value can result in a very serious underestimation of the uncertainty on the parameters of interest.

(iii) In many situations *prior information* is highly relevant and omitting it would result in seriously wrong inferences. The simplest case is when the parameters of interest have a physical meaning that restricts their possible values: masses, count rates, power and light intensity are examples of quantities that must be positive. Frequentist procedures based only on the likelihood can give best–fit estimates that are negative, and hence meaningless, unless special care is taken (for example, constrained likelihood methods). This often happens in the regime of small counts or low signal to noise. The use of Bayes' Theorem ensures that relevant prior information is accounted for in the final inference and that physically meaningless results are weeded out from the beginning.

(iv) Bayesian statistics only deals with the *data that were actually observed*, while frequentist methods focus on the distribution of possible data that have not been obtained. As a consequence, *frequentist results can depend on what the experimenter thinks about the probability of data that have not been observed.* (this is called the "stopping rule" problem). This state of affairs is obviously absurd. Our inferences should not depend on the probability of what could have happened but should be conditional on whatever has actually occurred. This is built into Bayesian methods from the beginning since inferences are by construction conditional on the observed data.

- The cosmology and astrophysics communities have been embracing Bayesian mehods since the turning of the Millennium, spurred by the availability of cheap computational power that has ushered in an era of high-performance computing, thus allowing for the first time to deploy the power of Bayesian statistics thanks to numerical implementations (in particular, MCMC and related techniques). The steep increase in the number of Bayesian papers in the astrophysics literature is shown in Fig. 3.

- Bayesian inference works by updating our state of knowledge about a parameter (or hypothesis) as new data flow in. The posterior from a previous cycle of observations becomes the prior for the next.

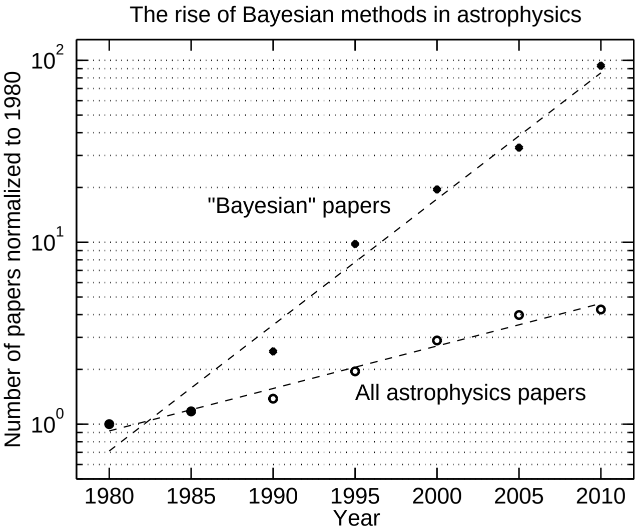

FIGURE 3. Number of articles in astronomy and cosmology with "Bayesian" in the title, as a function of publication year (upper data points) and total number of articles (lower data points) as a function of publication year. Numbers are normalized to 1980 levels for each data series. The number of Bayesian papers doubles every 4.3 years, while the total number of papers doubles "only" every 12.6 years. At the present rate, by 2060 all papers on the archive will be Bayesian. (source: NASA/ADS).

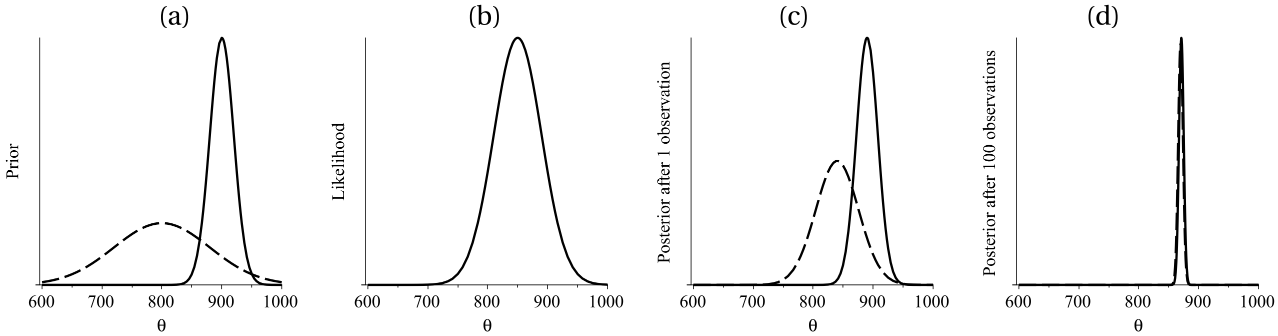

FIGURE 4. Converging views in Bayesian inference. Two scientists having different prior believes $p(\theta)$ about the value of a quantity θ (panel (a), the two curves representing two different priors) observe one datum with likelihood $\mathcal{L}(\theta)$ (panel (b)), after which their posteriors $p(\theta|d)$ (panel (c), obtained via Bayes Theorem, Eq. (48)) represent their updated states of knowledge on the parameter. This posterior then becomes the prior for the next observation. After observing 100 data points, the two posteriors have become essentially indistinguishable (d).

The price we have to pay is that we have to start somewhere by specifying an initial prior, which is not determined by the theory, but it needs to be given by the user. The prior should represent fairly the state of knowledge of the user about the quantity of interest. Eventually, the posterior will converge to a unique (objective) result even if different scientists start from different priors (provided their priors are non-zero in regions of parameter space where the likelihood is large). See Fig. 4 for an illustration.

- ▸ There is a vast literature about how to select a prior in an appropriate way. Some aspects are fairly obvious: if your parameter θ describes a quantity that has e.g. to be strictly positive (such as the number of photons in a detector, or an amplitude), then the prior will be 0 for values $\theta < 0$.

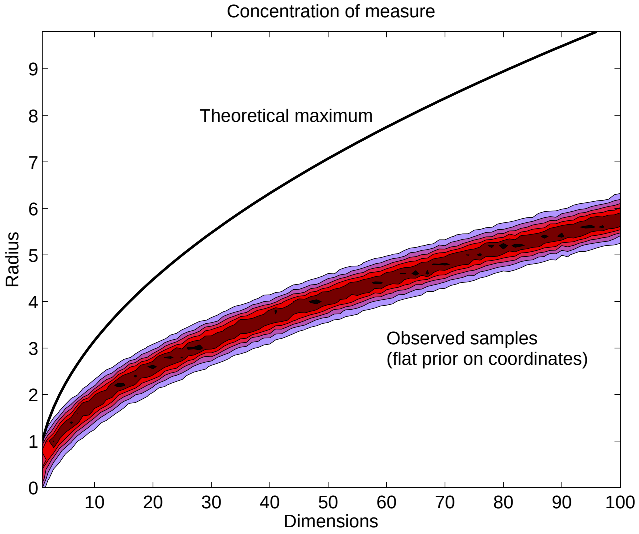

FIGURE 5. Illustration of the phenomenon of the concentration of measure in parameter spaces with a large number of dimensions. The coloured band represents the density of samples (as a function of the number of dimensions sampled) obtained with a flat prior on the axis coordinates of a D-dimensional hypercube. It can be seen that samples from the prior concentrate in a thin shell of constant variance, leaving most of the parameter space unexplored. The radius of the shell is given in the vertical axis.

A standard (but by no means harmless, see below) choice is to take a *uniform prior* (also called "flat prior") on θ, defined as:

$$P(\theta) = \begin{cases} \frac{1}{(\theta_{\max}-\theta_{\min})} & \text{for } \theta_{\min} \leq \theta \leq \theta_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{53}$$

With this choice of prior in Bayes theorem, Eq. (51), the posterior becomes functionally identical to the likelihood up to a proportionality constant:

$$P(\theta|d) \propto P(d|\theta) = \mathcal{L}(\theta). \tag{54}$$

In this case, all of our previous results about the likelihood carry over (but with a different interpretation). In particular, the probability content of an interval around the mean for the posterior should be interpreted as a statement about our degree of belief in the value of θ (differently from confidence intervals for the likelihood).

- Under a change of variable, $\Psi = \Psi(\theta)$, the prior transforms according to:

$$P(\Psi) = P(\theta)\left|\det\left(\frac{\partial\theta}{\partial\Psi}\right)\right|. \tag{55}$$

In particular, a flat prior on θ is no longer flat in Ψ if the variable transformation is non-linear.

- **Important notes about priors**:
  (i) Beware! A flat prior is far from harmless, especially in parameter spaces of high dimensionality. **Example (concentration of measure)**: sampling uniformly along each dimension $x_i \in [0,1]$ of a D-dimensional hypercube leads to the radius $r = \left(\sum_{i=1}^{D} x_i^2\right)^{1/2}$ of the samples to concentrate around the value $\langle r\rangle = (D/3)^{1/2}$ with constant variance. As a consequence, all of the samples are found on a thin shell (see Fig. 5 for an illustration). Even worse, in D dimensions the volume of the hypercube is much larger than the volume of the hypersphere, hence most of the volume is in the corners of the hypercube which are not sampled.
  (ii) A sensitivity analysis should always be performed (i.e., change the prior in a reasonable way and assess how robust your posterior is). This is seldom done in the astrophysics and cosmology literature!

(iii) There is a vast body of literature on different types of priors, when to use them and what they are good for. It is a good idea to browse the literature when faced with a new problem, as there is no point in re-inventing the wheel everytime. There are essentially two schools of thought: one maintains that priors should be chosen according to subjective degree of belief; the other, that they should be selected according to some formal rule, i.e. priors should be chosen by convention. None of the two approaches is free from difficulties, see [15]. Some examples:
  - *reference priors:* the idea is to define a prior wrt which the contribution of the data to the posterior is maximised. This is achieved by choosing a prior with maximum entropy. For example, in the case of a Gaussian likelihood this leads to the conclusion that the proper prior for the mean μ is flat on μ, while for the standard deviation σ is should be flat in $\log\sigma$ (with appropriate cutoffs of course).
  - *ignorance priors:* in 1812 Laplace set forth the principle that when nothing else is known priors should be chosen so as to give equal probability to all alternatives ("the principle of indifference"). Unfortunately this is very difficult to do in the case of continous parameters (also because indifference on a certain parameter is not invariant wrt non-linear reparameterizations). In some relatively simple cases, ignorance priors can be derived using symmetry or invariance arguments, see [3].
  - *conjugate priors:* a prior is said to be conjugate to the likelihood if the resulting posterior is of the same family as the likelihood. The covenience of having conjugate priors is that the likelihood updates the prior to a posterior which is of the same type (i.e., same distributional family). For example, Gaussian distributions are self-conjugate (i.e., a Gaussian prior with a Gaussian likelihood leads to a Gaussian posterior); the conjugate prior to both the Poisson and the exponential likelihood is the Gamma distribution; the conjugate prior to a Binomial likelihood is the Beta distribution.

4.3. **General Bayesian solution to inference problems.**

- Choose a model containing a set of hypotheses in the form of a vector of parameters, θ (e.g., the mass of an extra–solar planet or the abundance of dark matter in the Universe).
- Specify the priors for the parameters. Priors should summarize our state of knowledge about the parameters before we consider the new data, including an relevant external source of information.
- Construct the likelihood function for the measurement, which usually reflects the way the data are obtained (e.g., a measurement with Gaussian noise will be represented by a Normal distribution, while γ–ray counts on a detector will have a Poisson distribution for a likelihood). Nuisance parameters related to the measurement process might be present in the likelihood, e.g. the variance of the Gaussian might be unknown or the background rate in the absence of the source might be subject to uncertainty.
- Obtain the posterior distribution (possibly unnormalized) either by analytical or, more often, numerical methods (see below for MCMC).
- The posterior pdf for one parameter at the time is obtained by *marginalization*, i.e., by integrating over the uninteresting parameters. E.g., assume $\theta = \{\phi, \psi\}$, then the 1D posterior pdf for ϕ alone is given by

$$p(\phi|d) \propto \int \mathcal{L}(\phi,\psi) p(\phi,\psi) \mathrm{d}\psi. \tag{56}$$

  The final inference on ϕ from the posterior can then be communicated by plotting $p(\phi|d)$, with the other components marginalized over.
- An alternative statistical measure to the marginal posterior given by (56) is the *profile likelihood*, defined, say, for the parameter θ_1 as

$$\mathcal{L}(\theta_1) \equiv \max_{\theta_2,\dots,\theta_N} P(d|\theta), \tag{57}$$

where in our case $P(d|\theta)$ is the full likelihood function. Thus in the profile likelihood one maximises the value of the likelihood along the hidden dimensions, rather than integrating it out as in the marginal posterior. The profile likelihood is obtained from the samples by maximising the value of the likelihood in each bin. The profile likelihood is expected to be prior-independent, as long as the scan has gathered a sufficient number of samples in the favoured region, which is in general a difficult task for multi-dimensional parameter spaces. In fact, the profile likelihood and the Bayesian posterior ask two different statistical questions of the data: the latter evaluates which regions of parameter space are most plausible in the light of the measure implied by the prior; the former singles out regions of high quality of fit, independently of their extent in parameter space, thus disregarding the possibility of them being highly fine tuned. The information contained in both is relevant and interesting, and for non-trivial parameter spaces the two different approaches do not necessarily lead to the same conclusions (in the paradigmatic case of a Gaussian distributed quantity, both the pdf and the profile likelihood are identical and thus the question of which to choose does not arise).

- The profile likelihood can be directly interpreted as a likelihood function, except of course that it does account for the effect of the hidden parameters. Approximate confidence intervals from the profile likelihood can be obtained via the usual likelihood ratio test as follows. Starting from the best-fit value in parameter space, an $\alpha\%$ confidence interval encloses all parameter values for which minus twice the log–likelihood increases less than $\Delta\chi^2(\alpha,n)$ from the best fit value. The threshold value depends on α and on the number n of parameters one is simultaneously considering (usually $n=1$ or $n=2$), and it is obtained by solving

$$\alpha = \int_0^{\Delta\chi^2} \chi^2_n(x)dx, \tag{58}$$

where $\chi^2_n(x)$ is the chi–square distribution for n degrees of freedom.

4.4. **Markov Chain Monte Carlo in practice.**

- The purpose of the Markov chain Monte Carlo algorithm is to construct a sequence of points (or "samples") in parameter space (called "a chain"). The crucial property of the chain is that the density of samples is proportional to the posterior pdf. This allows to construct a map of the posterior distribution.
- A Markov chain is defined as a sequence of random variables $\{X^{(0)}, X^{(1)}, \dots, X^{(M-1)}\}$ such that the probability of the $(t+1)$–th element in the chain only depends on the value of the t–th element. The crucial property of Markov chains is that they can be shown to converge to a stationary state (i.e., which does not change with t) where successive elements of the chain are samples from the *target distribution*, in our case the posterior $p(\theta|d)$.
- The generation of the elements of the chain is probabilistic in nature, and is described by a *transition probability* $T(\theta^{(t)},\theta^{(t+1)})$, giving the probability of going from the point $\theta^{(t)}$ to the point $\theta^{(t+1)}$ in parameter space. A sufficient condition to obtain a Markov Chain is that the transition probability satisfy the *detailed balance condition*

$$p(\theta^{(t)}|d)T(\theta^{(t)},\theta^{(t+1)}) = p(\theta^{(t+1)}|d)T(\theta^{(t+1)},\theta^{(t)}). \tag{59}$$

This means that the ratio of the probabilities for jumping from $\theta^{(t)}$ to $\theta^{(t+1)}$ is inversely proportional to the ratio of the posterior probabilities at the two points.

The simplest (and widely used) MCMC algorithm is the **Metropolis algorithm:**

(i) Start from a random point $\theta^{(0)}$, with associated posterior probability $p_0 \equiv p(\theta^{(0)}|d)$.

(ii) Propose a candidate point $\theta^{(c)}$ by drawing from the *proposal distribution* $q(\theta^{(0)},\theta^{(c)})$. The proposal distribution might be for example a Gaussian of fixed width σ centered around the current point. The distribution q must satisfy the symmetry condition $q(x,y)=q(y,x)$.

(iii) Evaluate the posterior at the candidate point, $p_c = p(\theta^{(c)}|d)$. Accept the candidate point with probability

(60)
$$\alpha(\theta^{(0)}, \theta^{(c)}) = \min\left(\frac{p_c}{p_0}, 1\right).$$

This accept/reject step can be performed by generating a random number u from the uniform distribution $[0,1)$ and accepting the candidate sample if $u < \alpha$, and rejecting it otherwise.

(iv) If the candidate point is accepted, add it to the chain and move there. Otherwise stay at the old point (which is thus counted twice in the chain). Go back to (ii).

Notice from Eq. (60) that whenever the candidate sample has a larger posterior than the previous one (i.e., $p_c > p_0$) the candidate is always accepted. Also, in order to evaluate the acceptance function (60) only the unnormalized posterior is required, as the normalization constant drops out of the ratio. It is easy to show that the Metropolis algorithm satisfies the detailed balance condition, Eq. (59), with the transition probability given by $T(\theta^{(t)}, \theta^{(t+1)}) = q(\theta^{(t)}, \theta^{(t+1)})\alpha(\theta^{(t)}, \theta^{(t+1)})$.

- ▸ Once samples from the posterior pdf have been gathered, obtaining Monte Carlo estimates of expectations for any function of the parameters becomes a trivial task. For example, the posterior mean is given by (where $\langle\cdot\rangle$ denotes the expectation value with respect to the posterior)

(61)
$$\langle\theta\rangle \approx \int P(\theta|d)\theta d\theta = \frac{1}{M}\sum_{t=0}^{M-1}\theta^{(t)},$$

where the equality with the mean of the samples from the MCMC follows because the samples $\theta^{(t)}$ are generated from the posterior by construction.

- ▸ One can easily obtain the expectation value of any function of the parameters $f(\theta)$ as

(62)
$$\langle f(\theta)\rangle \approx \frac{1}{M}\sum_{t=0}^{M-1} f(\theta^{(t)}).$$

It is usually interesting to summarize the results of the inference by giving the 1–dimensional *marginal probability* for the j–th element of θ, θ_j, obtained by integrating out all other parameters from the posterior:

(63)
$$P(\theta_1|d) = \int P(\theta|d)d\theta_2 \ldots d\theta_n,$$

where $P(\theta_1|d)$ is the *marginal posterior* for the parameter θ_1.

- ▸ From the Markov chain it is trivial to obtain and plot the marginal posterior on the left–hand–side of Eq. (63): since the elements of the Markov chains are samples from the full posterior, $P(\theta|d)$, their density reflects the value of the full posterior pdf. It is then sufficient to divide the range of θ_1 in a series of bins and *count the number of samples falling within each bin*, simply ignoring the coordinates values $\theta_2, \ldots, \theta_n$. A 2–dimensional posterior is defined in an analogous fashion. A 1D 2–tail α% credible region is given by the interval (for the parameter of interest) within which fall α% of the samples, obtained in such a way that a fraction $(1-\alpha)/2$ of the samples lie outside the interval on either side. In the case of a 1–tail upper (lower) limit, we report the value of the quantity below (above) which α% of the sample are to be found.
- ▸ There are several important practical issues in working with MCMC methods which are worth mentioning. Especially for high–dimensional parameter spaces with multi–modal posteriors it is important *not* to use MCMC techniques as a black box, since poor exploration of the posterior can lead to serious mistakes in the final inference if it remains undetected. Some of the most relevant aspects are:

(i) Initial samples in the chain must be discarded, since the Markov process is not yet sampling from the equilibrium distribution (so–called *burn–in period*). The burn–in can roughly be assessed by looking at the evolution of the posterior density as a function of the number of steps in the chain. When the chain is started at a random point in parameter space, the posterior probability will typically be small and becomes larger at every step as the chain approaches

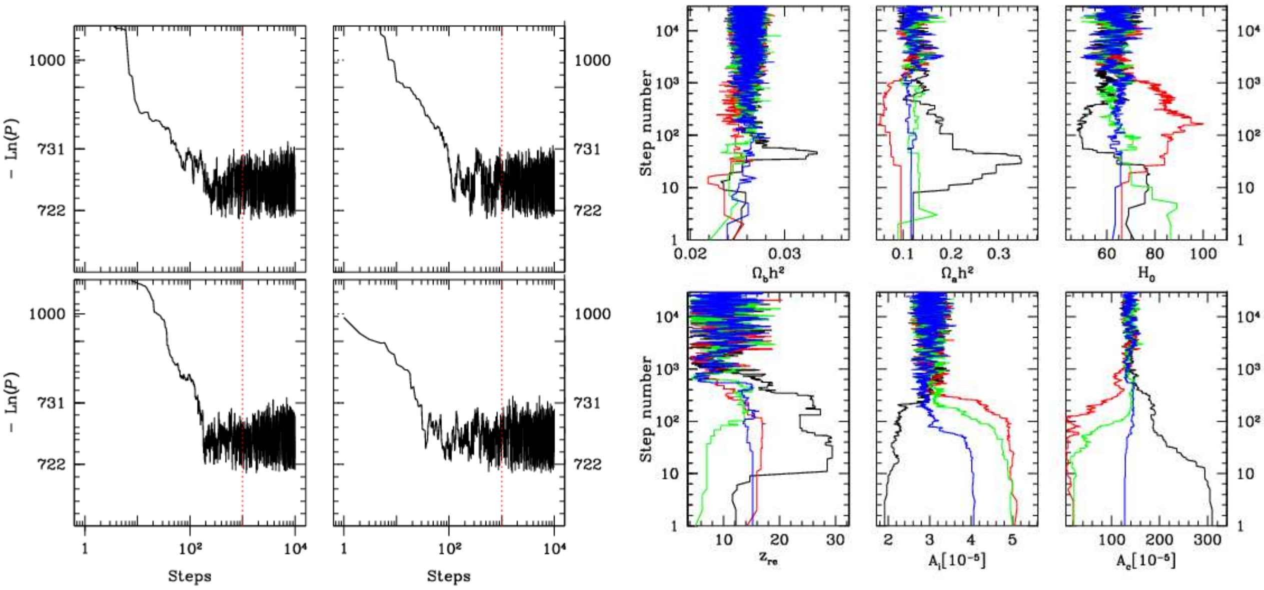

FIGURE 6. Illustration of the burn-in period. Left panel: the logarithm of the log-likelihood, $-\ln P(d|\theta)$, as a function of the step number for four Monte Carlo chains. After the burn-in period (dotted, vertical lines), the value flattens and the chains are sampling from the target distribution. Right panel: the four chains (in different colors) are started in different points of a 6-dimensional parameter space and all converge to the same region after the burn-in. The vertical axis gives the number of steps.

the region where the fit to the data is better. Only when the chain has moved in the neighborhood of the posterior peak the curve of the log posterior as a function of the step number flattens and the chain begins sampling from its equilibrium distribution. Samples obtained before reaching this point must be discarded, see Fig. 6

(ii) A difficult problem is presented by the assessment of *chain convergence*, which aims at establishing when the MCMC process has gathered enough samples so that the Monte Carlo estimate (62) is sufficiently accurate. Useful diagnostic tools include the Raftery and Lewis statistics [22] and the Gelman and Rubin criterion [23].

(iii) One has to bear in mind that MCMC is a *local algorithm*, which can be trapped around local maxima of the posterior density, thus missing regions of even higher posterior altogether. Considerable experimentation is sometimes required to find an implementation of the MCMC algorithm that is well suited to the exploration of the parameter space of interest. Experimenting with different algorithms (each of which has its own strength and weaknesses) is highly recommended.

(iv) Successive samples in a chain are in general correlated. Although this is not prejudicial for a correct statistical inference, it is often interesting to obtain *independent samples* from the posterior. This can be achieved by "thinning" the chain by an appropriate factor, i.e. by selecting only one sample every K. A discussion of samples independence and how to assess it can be found in [24], along with a convergence test based on the samples' power spectrum.

5. BAYESIAN MODEL SELECTION

5.1. **The three levels of inference.**

- It is convenient to divide Bayesian inference in three different levels:

(i) **Level 1:** We have chosen a model $\mathcal{M}_0$, assumed true, and we want to learn about its parameters, θ_0. E.g.: we assume ΛCDM to be the true model for the Universe and try to constrain its parameters. This is the usual parameter inference step.

(ii) **Level 2:** We have a series of alternative models on the table $(\mathcal{M}_1, \mathcal{M}_2, \dots)$ and we want to determine which of those is in best agreement with the data. This is a problem of model selection, or model criticism. For example, we might want to decide whether a dark energy equation of state $w = -1$ is a sufficient description of the available observations or whether we need an evolving dark energy model, $w = w(z)$.

(iii) **Level 3:** Of the N models considered in Level 2, there is no clear "best" model. We want to report inferences on parameters that account for this model uncertainty. This is the subject of Bayesian model averaging (not covered in these lectures). For example, we want to determine Ω_m independently of the assumed dark energy model.

- The Frequentist approach to model criticism is in the form of hypothesis testing (e.g., "chi-squared-per-degree-of-freedom“ type of tests). One ends up rejecting (or not) a null hypothesis H_0 based on the p-value, i.e., the probability of getting data as extreme or more extreme than what has been observed if one assumes that H_0 is true. Notice that this is *not* the probability for the hypothesis! Classical hypothesis testing assumes the hypothesis to be true and determines how unlikely are our observations given this assumption. Studying Ref. [20] is highly recommended.
- The Bayesian approach takes the view that there is no point in rejecting a model unless there are specific alternatives available: it takes therefore the form of model *comparison*. The key quantity for model comparison is the Bayesian evidence, which automatically implements a quantitative version of Occam's razor (i.e., the notion that simpler models ought to be preferred if they can explain the data sufficiently well).

5.2. **The Bayesian evidence.**

- The evaluation of a model's performance in the light of the data is based on the *Bayesian evidence*. As seen above, this is the normalization integral on the right–hand–side of Bayes' theorem, Eq. (52), which we rewrite here for a continuous parameter space $\Omega_{\mathcal{M}}$ and conditioning explicitly on the model under consideration, $\mathcal{M}$:

$$p(d|\mathcal{M}) \equiv \int_{\Omega_{\mathcal{M}}} p(d|\theta, \mathcal{M}) p(\theta|\mathcal{M}) \mathrm{d}\theta \quad \text{(Bayesian evidence)}. \tag{64}$$

- The Bayesian evidence is the average of the likelihood under the prior for a specific model choice. From the evidence, the model posterior probability given the data is obtained by using Bayes' Theorem to invert the order of conditioning:

$$p(\mathcal{M}|d) \propto p(\mathcal{M}) p(d|\mathcal{M}), \tag{65}$$

where we have dropped an irrelevant normalization constant that depends only on the data and $p(\mathcal{M})$ is the prior probability assigned to the model itself. Usually this is taken to be non–committal and equal to $1/N_m$ if one considers N_m different models.

- When comparing two models, $\mathcal{M}_0$ versus $\mathcal{M}_1$, one is interested in the ratio of the posterior probabilities, or *posterior odds*, given by

$$\frac{p(\mathcal{M}_0|d)}{p(\mathcal{M}_1|d)} = B_{01} \frac{p(\mathcal{M}_0)}{p(\mathcal{M}_1)}. \tag{66}$$

- The *Bayes factor* B_{01} is the ratio of the models' evidences:

$$B_{01} \equiv \frac{p(d|\mathcal{M}_0)}{p(d|\mathcal{M}_1)} \quad \text{(Bayes factor)}. \tag{67}$$

A value $B_{01} > (<)\ 1$ represents an increase (decrease) of the support in favour of model 0 versus model 1 given the observed data (see [16] for more details on Bayes factors).

- Bayes factors are usually interpreted against the Jeffreys' scale [2] for the strength of evidence, given in Table 1. This is an empirically calibrated scale, with thresholds at values of the odds of about 3 : 1, 12 : 1 and 150 : 1, representing weak, moderate and strong evidence, respectively.

$\vert\ln B_{01}\vert$	Odds	Probability	Strength of evidence
< 1.0	$\lesssim 3:1$	< 0.750	Inconclusive
1.0	$\sim 3:1$	0.750	Weak evidence
2.5	$\sim 12:1$	0.923	Moderate evidence
5.0	$\sim 150:1$	0.993	Strong evidence

TABLE 1. Empirical scale for evaluating the strength of evidence when comparing two models, $\mathcal{M}_0$ versus $\mathcal{M}_1$ (so–called "Jeffreys' scale"). Threshold values are empirically set, and they occur for values of the logarithm of the Bayes factor of $|\ln B_{01}| = 1.0$, 2.5 and 5.0. The right–most column gives our convention for denoting the different levels of evidence above these thresholds. The probability column refers to the posterior probability of the favoured model, assuming non–committal priors on the two competing models, i.e., $p(\mathcal{M}_0) = p(\mathcal{M}_1) = 1/2$ and that the two models exhaust the model space, $p(\mathcal{M}_0|d) + p(\mathcal{M}_1|d) = 1$.

5.3. The Occam's razor effect.

- **Example (nested models)**: Consider two competing models: $\mathcal{M}_0$ predicting that a quantity $\theta = 0$ with no free parameters, and $\mathcal{M}_1$ which assigns θ a Gaussian prior distribution with 0 mean and variance Σ^2. Assume we perform a measurement of θ described by a normal likelihood of standard deviation σ, and with the maximum likelihood value lying λ standard deviations away from 0, i.e. $|\theta_{\max}/\sigma| = \lambda$. Then the Bayes factor between the two models is given by, from Eq. (67)

$$B_{01} = \sqrt{1 + (\sigma/\Sigma)^{-2}} \exp\left(-\frac{\lambda^2}{2(1+(\sigma/\Sigma)^2)}\right). \tag{68}$$

  For $\lambda \gg 1$, corresponding to a detection of the new parameter at many sigma, the exponential term dominates and $B_{01} \ll 1$, favouring the more complex model with a non–zero extra parameter, in agreement with the usual conclusion. But if $\lambda \lesssim 1$ and $\sigma/\Sigma \ll 1$ (i.e., the likelihood is much more sharply peaked than the prior and in the vicinity of 0), then the prediction of the simpler model that $\theta = 0$ has been confirmed. This leads to the Bayes factor being dominated by the Occam's razor term, and $B_{01} \approx \Sigma/\sigma$, i.e. evidence accumulates in favour of the simpler model proportionally to the volume of "wasted" parameter space. If however $\sigma/\Sigma \gg 1$ then the likelihood is less informative than the prior and $B_{01} \to 1$, i.e. the data have not changed our relative belief in the two models.
- In the above example, if the data are informative with respect to the prior on the extra parameter (i.e., for $\sigma/\Sigma \ll 1$) the logarithm of the Bayes factor is given approximately by

$$\ln B_{01} \approx \ln(\Sigma/\sigma) - \lambda^2/2, \tag{69}$$

  where as before λ gives the number of sigma away from a null result (the "significance" of the measurement). The first term on the right–hand–side is approximately the logarithm of the ratio of the prior to posterior volume. We can interpret it as the information content of the data, as it gives the factor by which the parameter space has been reduced in going from the prior to the posterior. This term is positive for informative data, i.e. if the likelihood is more sharply peaked than the prior. The second term is always negative, and it favours the more complex model if the measurement gives a result many sigma away from the prediction of the simpler model (i.e., for $\lambda \gg 0$). We are free to measure the information content in base–10 logarithm (as this quantity is closer to our intuition, being the order of magnitude of our information increase), and we define the quantity $I_{10} \equiv \log_{10}(\Sigma/\sigma)$. Figure 7 shows contours of $|\ln B_{01}| = \text{const}$ for const $= 1.0, 2.5, 5.0$ in the (I_{10}, λ) plane, as computed from Eq. (69). The contours delimit significative levels for the strength of evidence, according to the Jeffreys' scale (Table 1). For moderately informative data ($I_{10} \approx 1-2$) the measured mean has to lie at least about 4σ away from 0 in order to robustly disfavor the simpler model (i.e., $\lambda \gtrsim 4$). Conversely, for $\lambda \lesssim 3$ highly informative data ($I_{10} \gtrsim 2$) do favor the conclusion that the extra parameter is indeed 0. In general, a large information content favors the simpler model, because Occam's razor penalizes the large volume of "wasted" parameter space of the extended model.

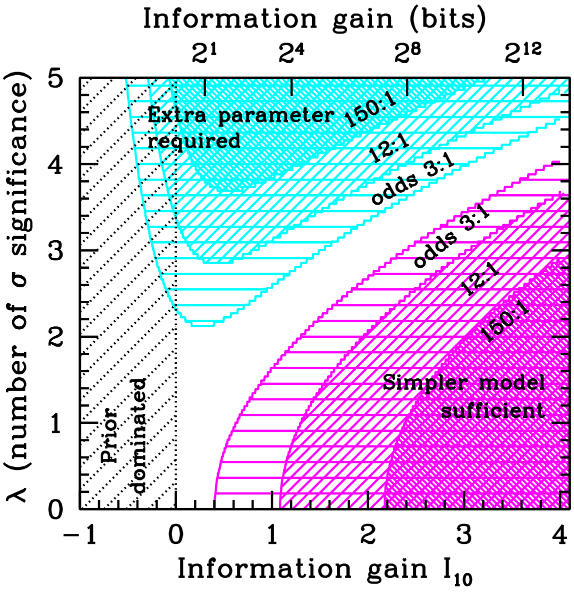

FIGURE 7. Illustration of Bayesian model comparison for two nested models, where the more complex model has one extra parameter. The outcome of the model comparison depends both on the information content of the data with respect to the *a priori* available parameter space, I_{10} (horizontal axis) and on the quality of fit (vertical axis, λ, which gives the number of sigma significance of the measurement for the extra parameter). Adapted from [14].

- An useful properties of Figure 7 is that the impact of a change of prior can be easily quantified. A different choice of prior width (i.e., Σ) amounts to a *horizontal shift* across Figure 7, at least as long as $I_{10} > 0$ (i.e., the posterior is dominated by the likelihood). Picking more restrictive priors (reflecting more predictive theoretical models) corresponds to shifting the result of the model comparison to the left of Figure 7, returning an inconclusive result (white region) or a prior–dominated outcome (hatched region). Notice that results in the 2–3 sigma range, which are fairly typical in cosmology, can only support the more complex model in a very mild way at best (odds of 3 : 1 at best), while actually being most of the time either inconclusive or in favour of the simpler hypothesis (pink shaded region in the bottom right corner).
- Notice that Bayesian model comparison is usually *conservative* when it comes to admitting a new quantity in our model, even in the case when the prior width is chosen "incorrectly" (whatever that means!). In general the result of the model comparison will eventually override the "wrong" prior choice (although it might take a long time to do so), exactly as it happens for parameter inference.
- Bayesian model selection does not penalize parameters which are unconstrained by the data. This is easily seen from Eq. (69): if a parameter is unconstrained, its posterior width σ is approximately equal to the prior width, Σ, and the Occam's razor penalty term goes to zero. In such a case, consideration of the Bayesian model complexity might help in judging model performance, see [25] for details.

5.4. Computation of the evidence.

- **Nested sampling.** A powerful and efficient alternative to classical MCMC methods has emerged in the last few years in the form of the so–called "nested sampling" algorithm, invented by John Skilling [18]. Although the original motivation for nested sampling was to compute the evidence integral of Eq. (64), the recent development of the multi–modal nested sampling technique [17] has

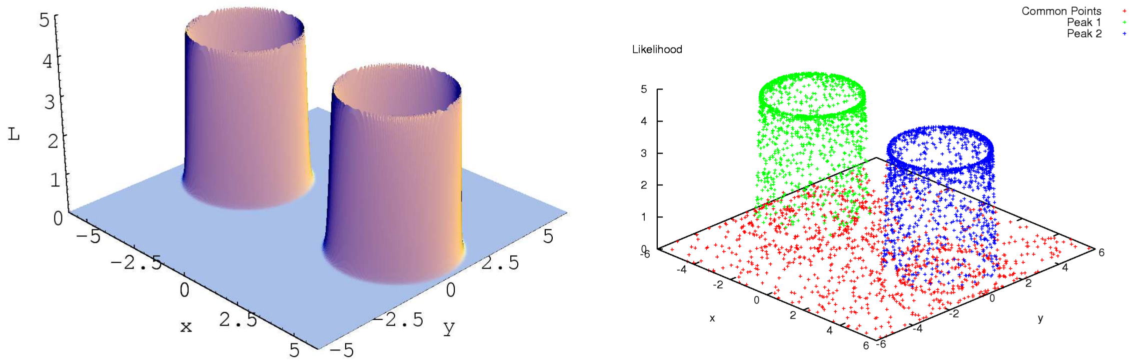

FIGURE 8. Example of posterior reconstruction using Nested Sampling. Left panel: target likelihood in a 2D parameter space (x, y). Right panel: reconstructed posterior (with flat priors) using Nested Sampling. From Ref. [17].

delivered an extremely powerful and versatile algorithm which has been demonstrated to be able to deal with extremely complex likelihood surfaces, see Fig. 8.

– The gist of nested sampling is that the multi–dimensional evidence integral is recast into a one–dimensional integral, by defining the prior volume X as $\mathrm{d}X \equiv p(\theta|\mathcal{M})\mathrm{d}\theta$ so that

$$X(\lambda) = \int_{\mathcal{L}(\theta)>\lambda} p(\theta|\mathcal{M})\mathrm{d}\theta \tag{70}$$

where $\mathcal{L}(\theta) \equiv p(d|\theta,\mathcal{M})$ is the likelihood function and the integral is over the parameter space enclosed by the iso–likelihood contour $\mathcal{L}(\theta) = \lambda$. So $X(\lambda)$ gives the volume of parameter space above a certain level λ of the likelihood.

– The Bayesian evidence, Eq. (64), can be written as

$$p(d|\mathcal{M}) = \int_0^1 \mathcal{L}(X)\mathrm{d}X, \tag{71}$$

where $\mathcal{L}(X)$ is the inverse of Eq. (70). Samples from $\mathcal{L}(X)$ can be obtained by drawing uniformly samples from the likelihood volume within the iso–contour surface defined by λ. This is the difficult part of the algorithm.

– Finally, the 1–dimensional integral of Eq. (71) can be obtained by simple quadrature, thus

$$p(d|\mathcal{M}) \approx \sum_i \mathcal{L}(X_i)W_i, \tag{72}$$

where the weights are $W_i = \frac{1}{2}(X_{i-1} - X_{i+1})$, see [18, 19] for details[3].

▸ **Thermodyamic integration.** The numerical method of choice until recently has been thermodynamic integration, which computes the evidence integral by definining

$$E(\mu) \equiv \int_{\Omega_\mathcal{M}} \mathcal{L}(\theta)^\mu p(\theta|\mathcal{M})\mathrm{d}\theta, \tag{73}$$

where μ is an annealing parameter and $\mathcal{L}(\theta) \equiv p(d|\theta,\mathcal{M})$. Obviously the desired evidence correponds to $E(1)$. One starts by performing a standard MCMC sampling with $\mu = 0$ (i.e., sampling from the prior), then gradually increases μ to 1 according to some annealing schedule. The log of

[3] Publicly available software implementing nested sampling can be found at `cosmonest.org` and `http://www.mrao.cam.ac.uk/software/cosmoclust/`. The latest release of the SUSY constraints package SuperBayeS also implements the MultiNest algorithm, see `http://superbayes.org`.

the evidence is then given by

$$\ln E(1) = \ln E(0) + \int_0^1 \frac{d\ln E}{d\mu}\mathrm{d}\mu = \int_0^1 \langle \ln \mathcal{L}\rangle_\mu \mathrm{d}\mu, \tag{74}$$

where the average log-likelihood is taken over the posterior with annealing parameter μ, i.e.

$$\langle \ln \mathcal{L}\rangle_\mu = \frac{\int_{\Omega_\mathcal{M}} (\ln \mathcal{L})\mathcal{L}(\theta)^\mu p(\theta|\mathcal{M})\mathrm{d}\theta}{\int_{\Omega_\mathcal{M}} \mathcal{L}(\theta)^\mu p(\theta|\mathcal{M})\mathrm{d}\theta}. \tag{75}$$

The drawback is that the end result might depend on the annealing schedule used and that typically this methods takes 10 times as many likelihood evaluations as parameter estimation.

- **Laplace approximation.** An approximation to the Bayesian evidence can be obtained when the likelihood function is unimodal and approximately Gaussian in the parameters. Expanding the likelihood around its peak to second order one obtains the Laplace approximation

$$p(d|\theta,\mathcal{M}) \approx \mathcal{L}_{\max} \exp\left[-\frac{1}{2}(\theta-\theta_{\max})^t L(\theta-\theta_{\max})\right], \tag{76}$$

where $\theta_{\max}$ is the maximum–likelihood point, $\mathcal{L}_{\max}$ the maximum likelihood value and L the likelihood Fisher matrix (which is the inverse of the covariance matrix for the parameters). Assuming as a prior a multinormal Gaussian distribution with zero mean and Fisher information matrix P one obtains for the evidence, Eq. (64)

$$p(d|\mathcal{M}) = \mathcal{L}_{\max}\frac{|F|^{-1/2}}{|P|^{-1/2}} \exp\left[-\frac{1}{2}(\theta_{\max}{}^t L\theta_{\max} - \overline{\theta}^t F\overline{\theta})\right], \tag{77}$$

where the posterior Fisher matrix is $F = L + P$ and the posterior mean is given by $\overline{\theta} = F^{-1}L\theta_{\max}$.

- **The Savage-Dickey density ratio.** A useful approximation to the Bayes factor, Eq. (67), is available for situations in which the models being compared are *nested* into each other, i.e. the more complex model ($\mathcal{M}_1$) reduces to the original model ($\mathcal{M}_0$) for specific values of the new parameters. This is a fairly common scenario in cosmology, where one wishes to evaluate whether the inclusion of the new parameters is supported by the data (e.g., do we need isocurvature contributions to the initial conditions for cosmological perturbations, or whether a curvature term in Einstein's equation is needed, or whether a non–scale invariant distribution of the primordial fluctuation is preferred). Writing for the extended model parameters $\theta = (\phi,\psi)$, where the simpler model $\mathcal{M}_0$ is obtained by setting $\psi = 0$, and assuming further that the prior is separable (which is usually the case in cosmology), i.e. that

$$p(\phi,\psi|\mathcal{M}_1) = p(\psi|\mathcal{M}_1)p(\phi|\mathcal{M}_0), \tag{78}$$

the Bayes factor can be written in all generality as

$$B_{01} = \left.\frac{p(\psi|d,\mathcal{M}_1)}{p(\psi|\mathcal{M}_1)}\right|_{\psi=0}. \tag{79}$$

This expression is known as the Savage–Dickey density ratio (see [14] and references therein). The numerator is simply the marginal posterior under the more complex model evaluated at the simpler model's parameter value, while the denominator is the prior density of the more complex model evaluated at the same point. This technique is particularly useful when testing for one extra parameter at the time, because then the marginal posterior $p(\psi|d,\mathcal{M}_1)$ is a 1–dimensional function and normalizing it to unity probability content only requires a 1–dimensional integral, which is simple to do using for example the trapezoidal rule.

- **Information criteria.** Sometimes it might be useful to employ methods that aim at an approximate model selection under some simplifying assumptions that give a default penalty term for more complex models, which replaces the Occam's razor term coming from the different prior volumes in the Bayesian evidence [21].

(i) **Akaike Information Criterion (AIC):** the AIC is an essentially frequentist criterion that sets the penalty term equal to twice the number of free parameters in the model, k:

$$\mathrm{AIC} \equiv -2 \ln \mathcal{L}_{\mathrm{max}} + 2k \tag{80}$$

where $\mathcal{L}_{\mathrm{max}} \equiv p(d|\theta_{\mathrm{max}}, \mathcal{M})$ is the maximum likelihood value.

(ii) **Bayesian Information Criterion (BIC):** the BIC follows from a Gaussian approximation to the Bayesian evidence in the limit of large sample size:

$$\mathrm{BIC} \equiv -2 \ln \mathcal{L}_{\mathrm{max}} + k \ln N \tag{81}$$

where k is the number of fitted parameters as before and N is the number of data points. The best model is again the one that minimizes the BIC.

(iii) **Deviance Information Criterion (DIC):** the DIC can be written as

$$\mathrm{DIC} \equiv -2 D_{\mathrm{KL}} + 2\mathcal{C}_b. \tag{82}$$

In this form, the DIC is reminiscent of the AIC, with the $\ln \mathcal{L}_{\mathrm{max}}$ term replaced by the estimated KL divergence D_{KL} and the number of free parameters by the effective number of parameters, $\mathcal{C}_b$ (see [13] for definitions).

The information criteria ought to be interpreted with care when applied to real situations. Comparison of Eq. (81) with Eq. (80) shows that for $N > 7$ the BIC penalizes models with more free parameters more harshly than the AIC. Furthermore, both criteria penalize extra parameters regardless of whether they are constrained by the data or not, unlike the Bayesian evidence. In conclusion, what makes the information criteria attractive, namely the absence of an explicit prior specification, represents also their intrinsic limitation.

References

[1] G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis* (John Wiley & Sons, Chicester, UK, 1992).

[2] H. Jeffreys, *Theory of probability*, 3rd edn , Oxford Classics series (reprinted 1998) (Oxford University Press, Oxford, UK, 1961).

[3] E. T. Jaynes, *Probability Theory. The logic of science* (Cambridge University Press, Cambridge, UK, 2003).

[4] J. M. Marin and C. P. Robert, *Bayesian Core. A Practical Approach to Computational Bayesian Statistics* (Springer, New York, 2007).

[5] D. MacKay, *Information theory, inference, and learning algorithms* (Cambridge University Press, Cambridge, UK, 2003).

[6] D. Sivia, *Data Analysis: A Bayesian tutorial* (Oxford University Press, Oxford, UK, 1996).

[7] P. Gregory, *Bayesian logical data analysis for the physical sciences* (Cambridge University Press, Cambridge, UK, 2003).

[8] G.A. Young & R.L. Smith, *Essentials of Statistical Inference*, (Cambridge University Press, Cambridge, UK, 2005).

[9] G. D'Agostini, Probability and Measurement Uncertainty in Physics - a Bayesian Primer, (hep-ph:/9512295) (1995).

[10] T. J. Loredo, From Laplace to Supernova SN 1987A: Bayesian Inference in Astrophysics, in T. Fougere (Editor) *Maximum-Entropy and Bayesian Methods*, Available from: http://bayes.wustl.edu/gregory/articles.pdf (accessed Jan 15 2008) (Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990), pp. 81–142.

[11] T. J. Loredo, The promise of Bayesian inference for astrophysics, in E. D. Feigelson and G. J. Babu (Eds) *Statistical Challenges in Modern Astronomy*, Available from: http://www.astro.cornell.edu/staff/loredo/bayes/promise.pdf (accessed Jan 15 2008) (Springer, New York, 1992), pp. 275–297.

[12] M. Hobson, A. Jaffe, A. Liddle, P. Mukherjee, *et al.* (Eds) *Bayesian Methods in Cosmology* (Cambridge University Press, Cambridge, UK, 2010).

[13] R. Trotta, Contemp. Phys. **49**, 71 (2008) [arXiv:0803.4089 [astro-ph]].

[14] R. Trotta, Mon. Not. Roy. Astron. Soc. **378**, 72 (2007) [arXiv:astro-ph/0504022].

[15] R. E. Kass & L. Wasserman, J. Am. Stat. Ass., **91**, 435, 1343–1370 (1996).

[16] R. E. Kass and A. E. Raftery, J. Am. Stat. Ass. **90**, 430, 773–795 (1995).

[17] F. Feroz and M. P. Hobson, Mon. Not. Roy. Astron. Soc., 384, 2, 449-463 (2008) arXiv:0704.3704 [astro-ph].

[18] J. Skilling, Nested sampling, in R. Fischer, R. Preuss and U. von Toussaint (Eds) *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, 735 (Amer. Inst. Phys. conf. proc. 2004), pp. 395–405.

[19] P. Mukherjee, D. Parkinson and A. R. Liddle, Astrophys. J. **638** L51–L54 (2006).

[20] T. Sellke, M. Bayarri and J. O. Berger, American Statistician **55** 62–71 (2001).

[21] A. R. Liddle, Mon. Not. Roy. Astron. Soc. **351** L49–L53 (2004).

[22] A. Raftery, Sociological Methodology **25** 111–163 (1995).

[23] A. Gelman and D.B. Rubin, Statistical Science, 7, 457-511 (1992).

[24] J. Dunkley, M. Bucher, P. G. Ferreira, K. Moodley and C. Skordis, Mon. Not. Roy. Astron. Soc. **356**, 925 (2005) [arXiv:astro-ph/0405462].

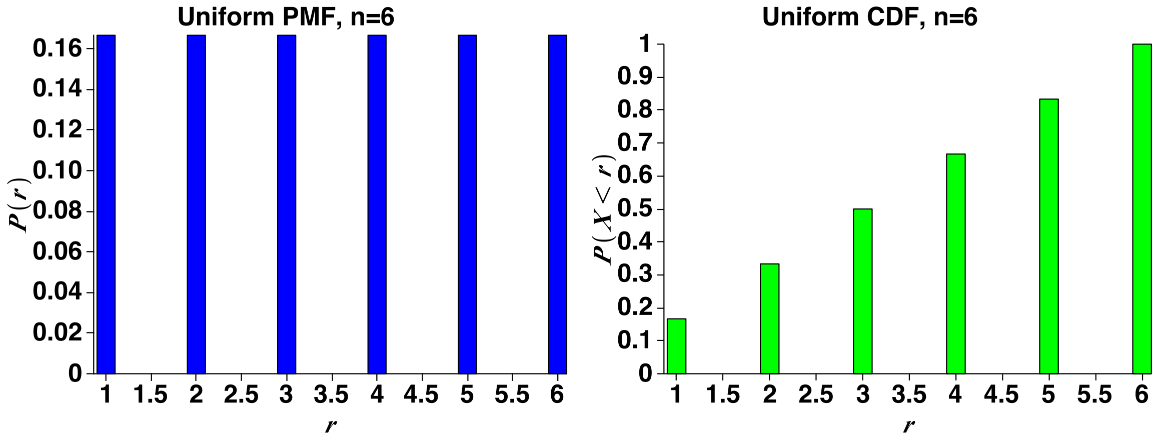

FIGURE 9. Left panel: uniform discrete distribution for $n = 6$. Right panel: the corresponding cdf.

[25] M. Kunz, R. Trotta and D. Parkinson, Phys. Rev. **D74** 023503 (2006).

APPENDIX A. SOME BACKGROUND MATERIAL

A.1. The uniform, binomial and Poisson distributions.

- **The uniform distribution:** for n equiprobable outcomes between 1 and n, the **uniform discrete distribution** is given by

$$P(r) = \begin{cases} 1/n & \text{for } 1 \le r \le n \\ 0 & \text{otherwise} \end{cases} \tag{83}$$

  It is plotted in Fig. 9 alongside with its cdf for the case of the tossing of a fair die ($n = 6$).
- **The binomial distribution:** the binomial describes the probability of obtaining r "successes" in a sequence of n trials, each of which has probability p of success. Here, "success" can be defined as one specific outcome in a binary process (e.g., H/T, blue/red, 1/0, etc). The binomial distribution $B(n,p)$ is given by:

$$P(r|n,p) \equiv B(n,p) = \binom{n}{r} p^r (1-p)^{n-r}, \tag{84}$$

  where the "choose" symbol is defined as

$$\binom{n}{r} \equiv \frac{n!}{(n-r)!r!} \tag{85}$$

  for $0 \le r \le n$ (remember, $0! = 1$). Some examples of the binomial for different choices of n, p are plotted in Fig. 10.
  A bent coin has a probability of landing heads $p = 0.7$. You toss it $n = 10$ times. What is the probability of getting 6 heads?
  Answer:

$$P(6|n = 10, p = 0.7) = \binom{10}{6} 0.7^6 0.3^4 = 0.2. \tag{86}$$

The derivation of the binomial distribution proceeds from considering the probability of obtaining r successes in n trials (p^r), while at the same time obtaining $n-r$ failures ($(1-p)^{n-r}$). The combinatorial factor in front is derived from considerations of the number of permutations that leads to the same total number of successes.

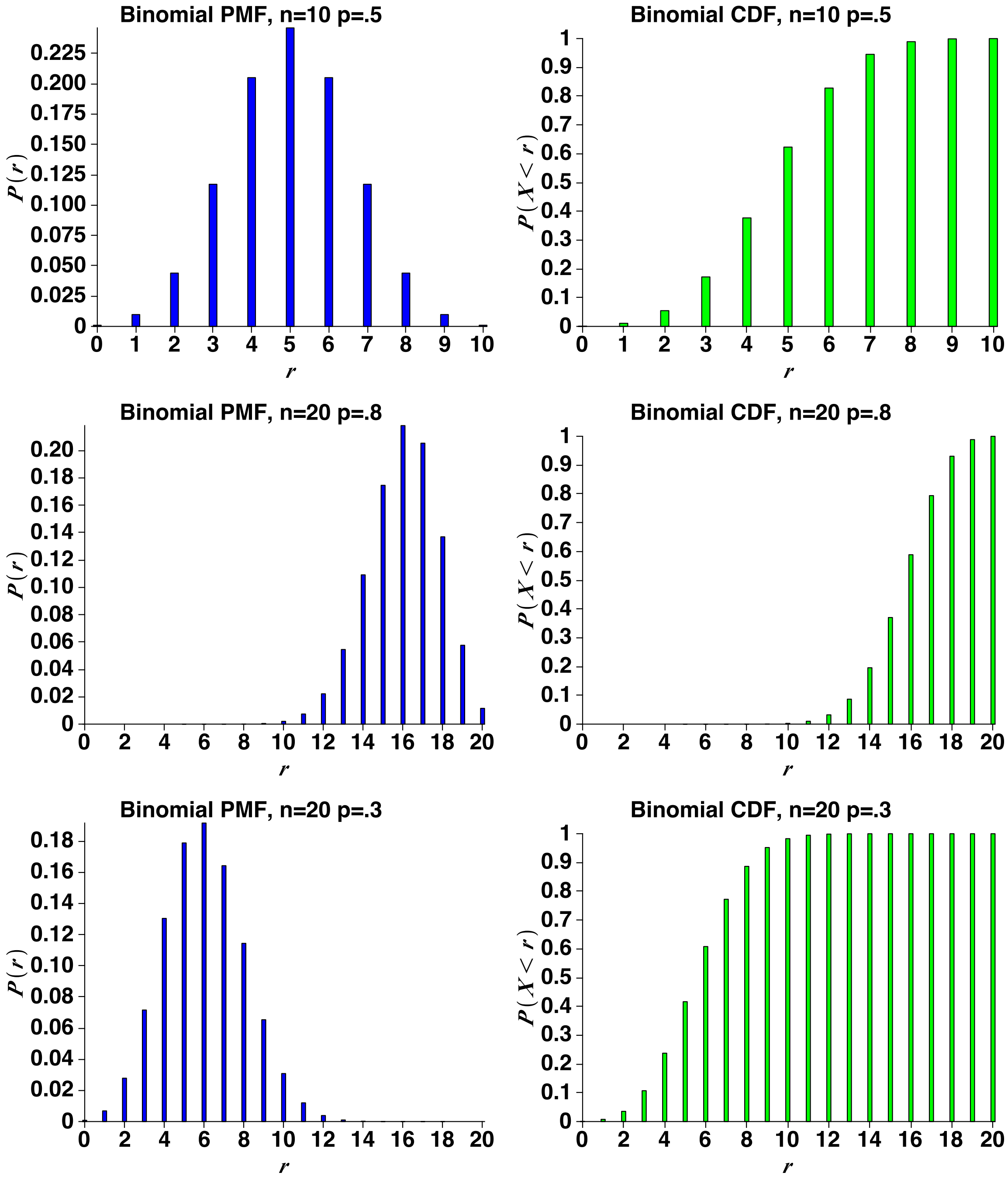

FIGURE 10. Some examples of the binomial distribution, Eq. (84), for different choices of n, p, and its corresponding cdf.

- **The Poisson distribution:** the Poisson distribution describes the probability of obtaining a certain number of events in a process where events occur with a fixed average rate and independently of each other. The process can occur in time (e.g., number of planes landing at Heathrow, number of photons arriving at a photomultiplier, number of murders in London, number of electrons at a detector, etc... in a certain time interval) or in space (e.g., number of galaxies in a patch on the sky).

  Let's assume that λ is the average number of events occuring per unit time or per unit length (depending on the problem being considered). Furthermore, λ = constant in time or space.

For example, $\lambda = 3.5$ busses/hour is the *average* number of busses passing by a particular bus stop every hour; or $\lambda = 10.3$ droplets/m^2 is the *average* number of drops of water hitting a square meter of the surface of an outdoor swimming pool in a certain day. Notice that of course at every given hour an integer number of busses actually passes by (i.e., we never observe 3 busses and one half passing by in an hour!), but that the **average** number can be non-integer (for example, you might have counted 7 busses in 2 hours, giving an average of 3.5 busses per hour). The same holds for the droplets of water.

For problems involving the time domain (e.g., busses/hour), the probability of r events happening in a time t is given by the **Poisson distribution**:

$$P(r|\lambda,t) \equiv \text{Poisson}(\lambda) = \frac{(\lambda t)^r}{r!}e^{-\lambda t}. \tag{87}$$

If the problem is about the spatial domain (e.g., droplets/m^2), the probability of r events happening in an area A is given by:

$$P(r|\lambda,A) \equiv \text{Poisson}(\lambda) = \frac{(\lambda A)^r}{r!}e^{-\lambda A}. \tag{88}$$

Notice that this is a discrete pmf in the number of events r, and **not** a continuous pdf in t or A. The probability of getting r events in a unit time interval is obtained by setting $t = 1$ in Eq. (87); similarly, the probability of getting r events in a unit area is obtained by setting $A = 1$ in Eq. (88)
A particle detector measures protons which are emitted with an average rate $\lambda = 4.5/\text{s}$. What is the probability of measuring 6 protons in 2 seconds?
Answer:

$$P(6|\lambda = 4.5\text{s}^{-1}, t = 2\text{s}) = \frac{(4.5\cdot 2)^6}{6!}e^{-4.5\cdot 2} = 0.09. \tag{89}$$

So the probability is about 9%.

The Poisson distribution of Eq. (87) is plotted in Fig. 11 as a function of r for a few choices of λ (notice that in the figure $t = 1$ has been assumed, in the appropriate units). The derivation of the Poisson distribution follows from considering the probability of 1 event taking place in a small time interval Δt, then taking the limit $\Delta t \to dt \to 0$. It can also be shown that the Poisson distribution arises from the binomial in the limit $pn \to \lambda$ for $n \to \infty$, assuming $t = 1$ in the appropriate units (see lecture).
In a post office, people arrive at the counter at an average rate of 3 customers per minute. What is the probability of 6 people arriving in a minute?
Answer: The number of people arriving follows a Poisson distribution with average $\lambda = 3$ (people/min). The probability of 6 people arriving in a minute is given by

$$P(n = 6|\lambda, t = 1\,\text{min}) = \frac{(\lambda t)^n}{n!}e^{-\lambda t} \approx 0.015 \tag{90}$$

So the probability is about 1.5%.

- The discrete distributions above depend on parameters (such as p for the binomial, λ for Poisson), which control the shape of the distribution. If we know the value of the parameters, we can compute the probability of an observation (as done it the examples above). This is the subject of **probability theory**, which concerns itself with the theoretical properties of the distributions. The inverse problem of making inferences about the parameters from the observed samples (i.e., learning about the parameters from the observations made) is the subject of statistical inference, addressed later.

A.2. **Expectation value and variance.**

- Two important properties of distributions are the **expectation value** (which controls the location of the distribution) and the **variance or dispersion** (which controls how much the distribution is spread out). Expectation value and variance are functions of a RV.

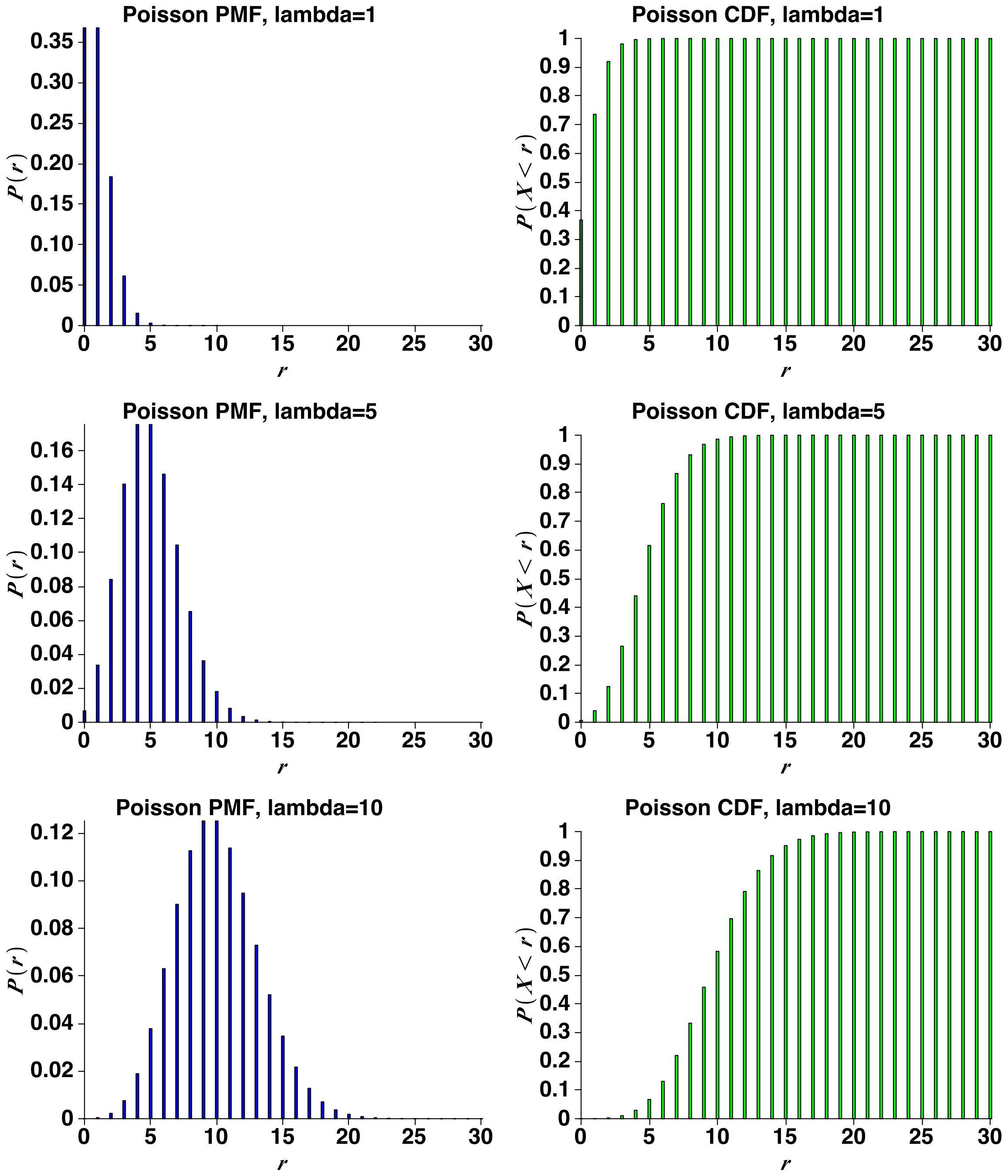

FIGURE 11. Some examples of the Poisson distribution, Eq. (87), for different choices of λ, and its corresponding cdf.

- The **expectation value** $E(X)$ (often called "mean", or "expected value"[4]) of the discrete RV X is defined as

$$E(X) = \langle X \rangle \equiv \sum_i x_i P_i. \tag{91}$$

[4] We prefer not to use the term "mean" to avoid confusion with the **sample mean**.

You toss a fair die, which follows the uniform discrete distribution, Eq. (83). What is the expectation value of the outcome?
Answer: the expectation value is given by $E(X) = \sum_i i \cdot \frac{1}{6} = 21/6$.

- The **variance or dispersion** $\mathrm{Var}(X)$ of the discrete RV X is defined as

(92)
$$\mathrm{Var}(X) \equiv E[(X - E(X))^2] = E(X^2) - E(X)^2.$$

The square root of the variance is often called "standard deviation" and is usually denoted by the symbol σ, so that $\mathrm{Var}(X) = \sigma^2$.
For the case of tossing a fair die once, the variance is given by

(93)
$$\mathrm{Var}(X) = \sum_i (x_i - \langle X\rangle)^2 P_i = \sum_i x_i^2 P_i - \left(\sum_i x_i P_i\right)^2 = \sum_i i^2 \frac{1}{6} - \left(\frac{21}{6}\right)^2 = \frac{105}{36}.$$

- For the binomial distribution of Eq. (84), the expectation value and variance are given by:

(94)
$$E(X) = np, \qquad \mathrm{Var}(X) = np(1-p).$$

(95)
$$\begin{aligned} E(x) &= \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=1}^{n} x \binom{n}{x} p^x (1-p)^{n-x} \\ &= np \sum_{x=1}^{n} x \frac{n!}{(n-x)!x!} \frac{p^x}{np} (1-p)^{n-x} = np \sum_{x=1}^{n} \frac{(n-1)!}{(n-x)!(x-1)!} p^{x-1} (1-p)^{n-x} \\ &= np \sum_{x=1}^{n} \frac{(n-1)!}{((n-1)-(x-1))!(x-1)!} p^{x-1} (1-p)^{(n-1)-(x-1)}, \end{aligned}$$

where in the first line we have made use of the fact that the $x = 0$ term in the sum is 0, hence we can sum from $x = 1$ onwards. Using the substituion $s = x - 1$ and $m = n - 1$ we obtain

(96)
$$E(x) = np \sum_{s=0}^{m} \frac{m!}{(m-s)!s!} p^s (1-p)^{m-s} = np,$$

as the sum above equals 1, being the sum of the terms of a binomial distribution which is normalized to unity total probability content.

To compute the variance, we start from by noticing that

(97)
$$\mathrm{Var}(x) = E(x^2) - E(x)^2 = E(x(x-1) + x) - E(x)^2 = E(x(x-1)) + E(x) - E(x)^2,$$

hence we only need to compute $E(x(x-1))$:

(98)
$$\begin{aligned} E(x(x-1)) &= \sum_{x=0}^{n} x(x-1) \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=2}^{n} \frac{n!}{(n-x)!(x-2)!} p^x (1-p)^{n-x} \\ &= n(n-1)p^2 \sum_{x=2}^{n} \frac{n!}{(n-x)!(x-2)!} \frac{p^x}{n(n-1)p^2} (1-p)^{n-x} \\ &= n(n-1)p^2 \sum_{x=2}^{n} \frac{(n-2)!}{((n-2)-(x-2))!(x-2)!} p^{x-2} (1-p)^{(n-2)-(x-2)}. \end{aligned}$$

Substituting $s = x - 2$ and $m = n - 2$ we obtain

(99)
$$E(x(x-1)) = n(n-1)p^2 \sum_{s=0}^{m} \frac{m!}{(m-s)!s!} p^s (1-p)^{m-s} = n(n-1)p^2.$$

Thus

(100)
$$\mathrm{Var}(x) = E(x(x-1)) + E(x) - E(x)^2 = n(n-1)p^2 + np - (np)^2 = np(1-p).$$

A fair coin is tossed N times. What is the expectation value for the number of heads, H? What is its variance? For $N = 10$, evaluate the probability of obtaining 8 or more heads.

Answer: The expectation values and variance are given by Eq. (94), with $p = 1/2$ (as the coin is fair), thus

(101) $$E(H) = Np = N/2 \quad \text{and} \quad \text{Var}(H) = Np(1-p) = N/4.$$

The probability of obtaining 8 or more heads is given by

(102) $$P(H = 8 = \sum_{H=8}^{10} P(H\text{heads}|N, p = 1/2) = \frac{1}{2^{10}} \sum_{H=8}^{10} \binom{10}{H} = \frac{56}{1024} \approx 0.055.$$

So the probability of obtaining 8 or more heads is about 5.5%.

- For the Poisson distribution of Eq. (87), the expectation value and variance are given by:

(103) $$E(X) = \lambda t, \qquad \text{Var}(X) = \lambda t,$$

while for the spatial version of the Poisson distribution, Eq. (88), they are given by:

(104) $$E(X) = \lambda A, \qquad \text{Var}(X) = \lambda A.$$

Setting $t = 1$ in the appropriate units). Since $e^x = \sum_k \frac{x^k}{k!}$ and $\frac{de^x}{dx} = e^x$, we have that

(105) $$\frac{de^x}{dx} = \sum_k k \frac{x^{k-1}}{k!} = e^x$$

and multiplying both sides by x we obtain:

(106) $$xe^x = \sum_k k \frac{x^k}{k!}.$$

The expectation value is given by

(107) $$E(n) = \sum_{n=0}^{\infty} n \cdot \text{Poisson}(\lambda) = \sum_{n=0}^{\infty} n \frac{\lambda^n}{n!} e^{-\lambda} = \lambda e^{\lambda} e^{-\lambda} = \lambda,$$

where in the penultimate step we have made use of Eq. (106).

To compute the variance, we need $\text{Var}(n) = E(n^2) - E(n)^2$. Use the same trick:

(108) $$\frac{d^2 e^x}{d^2 x} = \sum_k k(k-1) \frac{x^{k-2}}{k!} = e^x$$

hence

(109) $$x^2 e^x = \sum_k k(k-1) \frac{x^k}{k!} = \sum_k (k^2 - k) \frac{x^k}{k!}.$$

Therefore

(110) $$\sum_k k^2 \frac{x^k}{k!} = x^2 e^x + xe^x.$$

We can now compute

(111) $$E(n^2) = \sum_{n=0}^{\infty} n^2 \cdot \text{Poisson}(\lambda) = \sum_{n=0}^{\infty} n^2 \frac{\lambda^n}{n!} e^{-\lambda} = (\lambda^2 + \lambda) e^{\lambda} e^{-\lambda} = \lambda^2 + \lambda,$$

so that

(112) $$\text{Var}(n) = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

- As we did above for the discrete distribution, we now define the following properties for continuous distributions.
- The **expectation value** $E(X)$ of the continuous RV X with pdf $p(X)$ is defined as

(113) $$E(X) = \langle X \rangle \equiv \int xp(x)dx.$$

- The **variance or dispersion** $\mathrm{Var}(X)$ of the continuous RV X is defined as

$$\mathrm{Var}(X) \equiv E[(X-E(X))^2] = E(X^2) - E(X)^2 = \int x^2 p(x)dx - \left(\int xp(x)dx\right)^2. \tag{114}$$

A.3. **The exponential distribution.**

- **The exponential distribution** describes the time one has to wait between two consecutive events in a Poisson process, e.g. the waiting time between two radioactive particles decays. If the Poisson process happens in the spatial domain, then the exponential distribution describes the distance between two events (e.g., the separation of galaxies in the sky). In the following, we will look at processes that happen in time (rather than in space).
- To derive the exponential distribution, one can consider the arrival time of Poisson distributed events with average rate λ (for example, the arrival time particles in a detector). The probability that the first particle arrives at time t is obtained by considering the probability (which is Poisson distributed) that no particle arrives in the interval $[0,t]$, given by $P(0|\lambda,t) = \exp(-\lambda t)$ from Eq. (87), times the probability that one particle arrives during the interval $[t, t+\Delta t]$, given by $\lambda\Delta t$. Taking the limit $\Delta t \to 0$ it follows that the probability density (denoted by a symbol $p()$) for observing the first event happening at time t is given by

$$p(\text{1st event happens at time } t|\lambda) = \lambda e^{-\lambda t}, \tag{115}$$

where λ is the mean number of events per unit time. This is the exponential distribution.
Let's assume that busses in London arrive according to a Poisson distribution, with average rate $\lambda = 5$ busses/hour. You arrive at the bus stop and a bus has just departed. What is the probability that you will have to wait more than 15 minutes?
Answer: the probability that you'll have to wait for $t_0 = 15$ minutes or more is given by

$$\int_{t_0}^{\infty} p(\text{1st event happens at time } t|\lambda)dt = \int_{t_0}^{\infty} \lambda e^{-\lambda t}dt = e^{-\lambda t_0} = 0.29, \tag{116}$$

where we have used $\lambda = 5$ busses/hour = 1/12 busses/min.
- If we have already waited for a time s for the first event to occur (and no event has occurred), then the probability that we have to wait for another time t before the first event happens satisfies

$$p(T > t+s|T > s) = p(T > t). \tag{117}$$

This means that having waited for time s without the event occuring, the time we can expect to have to wait has the same distribution as the time we have to wait from the beginning. The exponential distribution has no "memory" of the fact that a time s has already elapsed (this is proved in Appendix **??**).
- For the exponential distribution of Eq. (115), the expectation value and variance for the time t are given by

$$E(t) = 1/\lambda, \qquad \mathrm{Var}(t) = 1/\lambda^2. \tag{118}$$

The expectation value is given by

$$E(t) = \int_0^{\infty} t\lambda e^{-\lambda t}dt = \frac{1}{\lambda}\int_0^{\infty} xe^{-x}dx = \frac{1}{\lambda}\left(-xe^{-x}\Big|_0^{\infty} - \int_0^{\infty}(-e^{-x})\,dx\right) = \frac{1}{\lambda}, \tag{119}$$

where the integral has been performed by integrating by parts. The variance can be calculated in a similar way but integrating by parts twice:

$$\mathrm{Var}(t) = E(t^2) - E(t)^2 = \int_0^{\infty} t^2\lambda e^{-\lambda t}dt - E(t)^2 = \frac{1}{\lambda}\int_0^{\infty} t^2\lambda^2 e^{-\lambda t}dt - \frac{1}{\lambda^2} \tag{120}$$

$$= \frac{1}{\lambda^2}\int_0^{\infty} x^2 e^{-x}dx - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}. \tag{121}$$

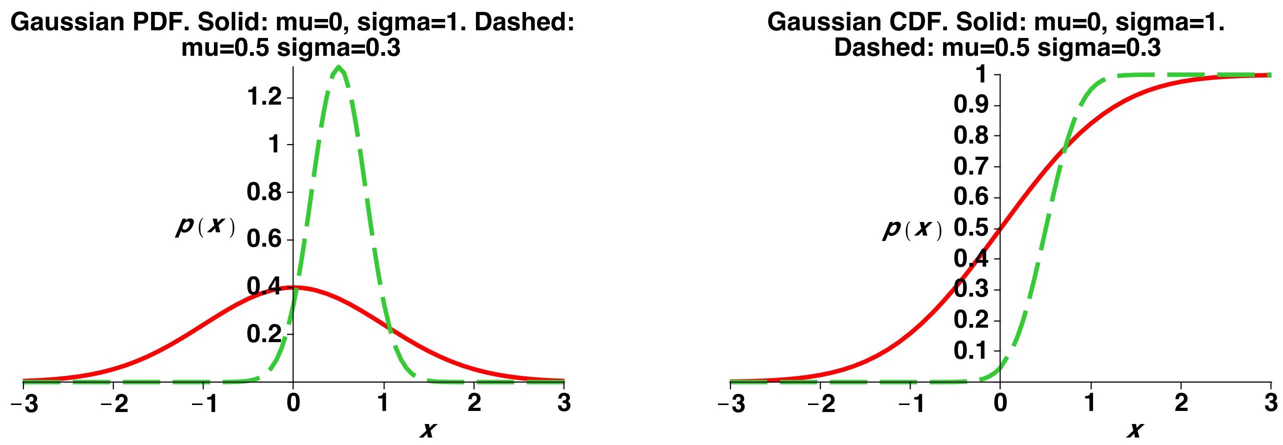

FIGURE 12. Two examples of the Gaussian distribution, Eq. (125), for different choices of μ, σ, and its corresponding cdf. The expectation value μ controls the location of the pdf (i.e., when changing μ the peak moves horizontally, without changing its shape), while the standard deviation σ controls its width (i.e., when changing σ the spread of the peak changes but not its location).

We now prove the "lack of memory" property, i.e. Eq. (117). The probability that at least one event has happened until time t is given by the cumulative distribution function

$$F_\lambda(t) = \int_0^t \lambda e^{-\lambda\tau} d\tau = \lambda \int_0^{\lambda t} \frac{1}{\lambda} e^{-x} dx = -e^{-x}\Big|_0^t = 1 - e^{-\lambda t}. \tag{122}$$

Therefore the probability that no events happen until time t is $1 - F_\lambda(t)$. Let's call this the probability that we have to wait a time $T > t$ for an event to happen. Then

$$P(T > t + s | T > s) = \frac{P(T > t + s, T > s)}{P(T > s)} = \frac{P(T > t + s)}{P(T > s)}, \tag{123}$$

where in the last passage we have used that if $T > t + s$ then it must also trivially be $T > s$ (i.e., if we have waited for a time $t + s$ we must have waited for a time s, too). Thus

$$P(T > t + s | T > s) = \frac{1 - F_\lambda(t + s)}{1 - F_\lambda(s)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = P(T > t). \tag{124}$$

A.4. **The Gaussian (or Normal) distribution.**

- The Gaussian pdf (often called "the Normal distribution") is perhaps the most important distribution. It is used as default in many situations involving continuous RV (the reason becomes clear once we have studied the Central Limit Theorem, section A.5). A heuristic derivation of how the Gaussian arises follows from the example of darts throwing (see Appendix **??**).
- The Gaussian pdf is a continuous distribution with mean μ and standard deviation σ is given by

$$p(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right), \tag{125}$$

  and it is plotted in Fig. 12 for two different choices of $\{\mu, \sigma\}$. The Gaussian is the famous bell-shaped curve.
- For the Gaussian distribution of Eq. (125), the expectation value and variance are given by:

$$E(X) = \mu, \qquad \text{Var}(X) = \sigma^2. \tag{126}$$

  The expectation value is given by

$$E(X) = \int x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx = \frac{1}{\sqrt{2\pi}} \int (\sigma t + \mu) e^{-\frac{1}{2}t^2} dt, \tag{127}$$

where we have used the variable transformation $x = \sigma t + \mu$. The first integral (containing a linear term in t) vanishes because of symmetry, hence

$$E(X) = \frac{\mu}{\sqrt{2\pi}} \int e^{-\frac{1}{2}t^2} dt = \mu \tag{128}$$

since $\int e^{-\frac{1}{2}t^2} dt = \sqrt{2\pi}$.

To compute the variance, we exploit the usual trick: $\mathrm{Var}(X) = E(X^2) - E(X)^2$, hence we need to compute

$$E(X^2) = \int x^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx \frac{1}{\sqrt{2\pi}} \int (\sigma t + \mu)^2 e^{-\frac{1}{2}t^2} dt \tag{129}$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int t^2 e^{-\frac{1}{2}t^2} dt + \frac{\mu^2}{\sqrt{2\pi}} \int e^{-\frac{1}{2}t^2} dt + \frac{\mu}{\sqrt{2\pi}} \int t e^{-\frac{1}{2}t^2} dt \tag{130}$$

$$= \sigma^2 + \mu^2 + 0, \tag{131}$$

and therefore $\mathrm{Var}(X) = \sigma^2 + \mu^2 - \mu^2 = \sigma^2$.

▸ It can be shown that the Gaussian arises from the binomial in the limit $n \to \infty$ and from the Poisson distribution in the limit $\lambda \to \infty$. As shown in Fig. 13, the Gaussian approximation to either the binomial or the Poisson distribution is very good even for fairly moderate values of n and λ. Here is an heuristic derivation of how the Gaussian arises.

Suppose we are throwing darts towards a target (located at the center of the coordinate system, at the position $x = 0, y = 0$), with the following rules:

(i) Throws are independent.
(ii) Errors in the x and y directions are independent.
(iii) Large errors are less probable than small ones.

The probability of a dart landing in an infinitesimal square located at coordinates (x, y) and of size $(\Delta x, \Delta y)$ (i.e., the dart landing in the interval $[x, x + \Delta x]$ and $[y, y + \Delta y]$) is given by:

$$p(x)\Delta x \cdot p(y)\Delta y = f(r)\Delta x \Delta y, \tag{132}$$

where $p(x)$ is the probability density of landing at position x (and similarly for $p(y)$), which is what we are trying to determine. On the l.h.s. of this equation, we can multiply the probabilities of landing in the x and y direction because of rule number (1) and (2). On the l.h.s., $f(r)$ is a function that only depends on the radial distance from the center, because of rule (2).

We now differentiate the above equation w.r.t. the polar coordinate ϕ:

$$\left(p(x)\frac{dp(x)}{d\phi} + p(y)\frac{dp(y)}{d\phi}\right)\Delta x \Delta y = 0. \tag{133}$$

(Note that the r.h.s. becomes 0 as it does not depend on ϕ). In polar coordinates, $x = r\cos\phi, y = r\sin\phi$, hence

$$\frac{dp(x)}{d\phi} = \frac{\partial p}{\partial x}\frac{\partial x}{\partial \phi} = -\frac{\partial p}{\partial x} y, \tag{134}$$

$$\frac{dp(y)}{d\phi} = \frac{\partial p}{\partial y}\frac{\partial y}{\partial \phi} = \frac{\partial p}{\partial y} x. \tag{135}$$

Eq. (133) becomes

$$\left(-p(x)\frac{\partial p}{\partial x} y + p(y)\frac{\partial p}{\partial y} x\right)\Delta x \Delta y = 0, \tag{136}$$

which implies

$$\frac{p(x)}{x}\frac{\partial p}{\partial x} = \frac{p(y)}{y}\frac{\partial p}{\partial y}. \tag{137}$$

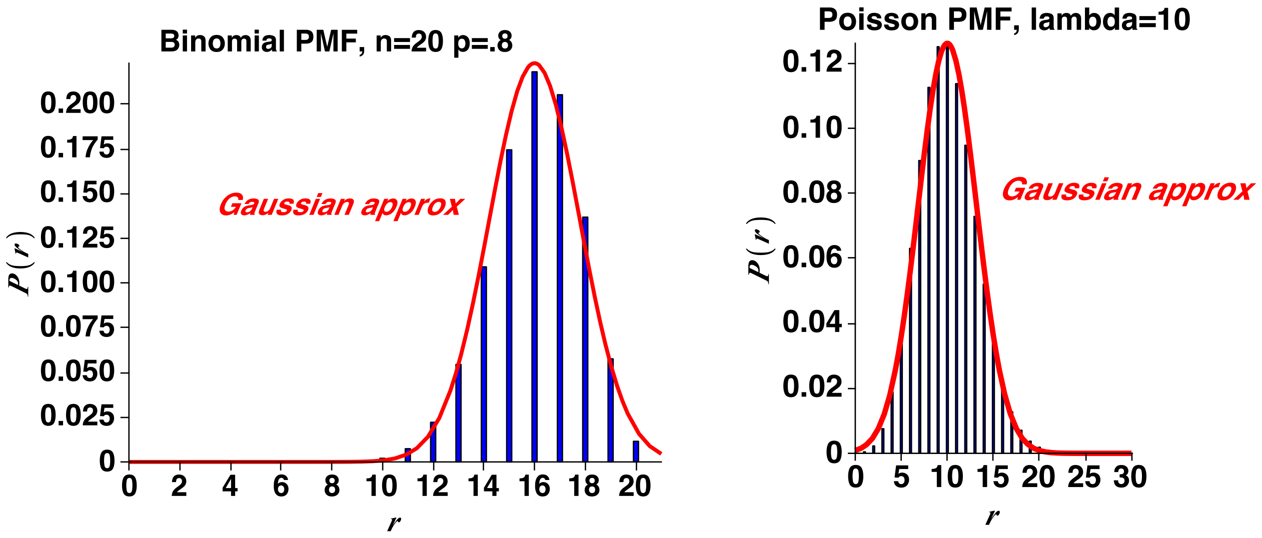

FIGURE 13. Gaussian approximation to the binomial (left panel) and the Poisson distribution (right panel). The solid curve gives in each case the Gaussian approximation to each pmf.

Since each side only depends on one of the variables, they must both equal a constant C, and we obtain the differential equation:

$$\frac{\partial p}{\partial x} = Cxp(x) \tag{138}$$

(and similarly for y). Integration gives the solution

$$p(x) = Ae^{\frac{C}{2}x^2} \tag{139}$$

and $C < 0$ because of rule (3). We thus define $C = -1/\sigma^2$. Requiring that the distribution is normalized gives $A = \frac{1}{\sqrt{2\pi}\sigma}$, and therefore $p(x)$ has the shape of a Gaussian (similarly for $p(y)$).

- The probability content of a Gaussian of standard deviation σ for a given symmetric interval around the mean of width $\kappa\sigma$ on each side is given by

$$P(\mu - \kappa\sigma < x < \mu + \kappa\sigma) = \int_{\mu-\kappa\sigma}^{\mu+\kappa\sigma} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) dx \tag{140}$$

$$= \frac{2}{\sqrt{\pi}} \int_0^{\kappa/\sqrt{2}} \exp\left(-y^2\right) dy \tag{141}$$

$$= \text{erf}(\kappa/\sqrt{2}), \tag{142}$$

where the **error function** erf is defined as

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp\left(-y^2\right) dy, \tag{143}$$

and can be found by numerical integration (also often tabulated and available as a built-in function in most mathematical software). Also recall the useful integral:

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) dx = \sqrt{2\pi}\sigma. \tag{144}$$

- Eq. (140) allows to find the probability content of the Gaussian pdf for any symmetric interval around the mean. Some commonly used values are given in Table 2.
- Measurements are often reported with the notation $T = (100 \pm 1)$ K (in this case, we assume we have measured a temperature, T). If nothing else is specified, it is usually implied that the error follows

κ "number of sigma"	$P(-\kappa < \frac{x-\mu}{\sigma} < \kappa)$ Probability content	Usually called
1	0.683	1σ
2	0.954	2σ
3	0.997	3σ
4	0.9993	4σ
5	$1-5.7\times10^{-7}$	5σ
1.64	0.90	90% probability interval
1.96	0.95	95% probability interval
2.57	0.99	99% probability interval
3.29	0.999	99.9% probability interval

TABLE 2. Relationship between the size of the interval around the mean and the probability content for a Gaussian distribution.

a Gaussian distribution. In the example above, ±1 K is the so-called "1σ interval". This means that 68.3% of the probability is contained within the range $[99,101]$ K. A "2σ interval" would have a length of 2 K on either side, so 95.4% of the probability is contained in the interval $[98,102]$ K. If one wanted a 99% interval, one would need a 2.57σ range (see Table 2). Since in this case the 1σ error is 1 K, the 2.57σ error is 2.57 K and the 99% interval is $[97.43,102.57]$ K.

A.5. **The Central Limit Theorem.**

- The Central Limit Theorem (CLT) is a very important result justifying why the Gaussian distribution is ubiquitous.
- **Simple formulation of the CLT**: Let $X_1, X_2, \dots, X_N$ be a collection of independent RV with finite expectation value μ and finite variance σ^2. Then, for $N \to \infty$, thir sum is Gaussian distributed with mean $N\mu$ and variance $N\sigma^2$.

  Note: it does not matter what the detailed shape of the underlying pdf for the individual RVs is!

  Consequence: whenever a RV arises as the sum of several independent effects (e.g., noise in a temperature measurement), we can be confident that it will be very nearly Gaussian distributed.
- **More rigorous (and more general) formulation of the CLT**: Let $X_1, X_2, \dots, X_N$ be a collection of independent RV, each with finite expectation value μ_i and finite variance σ_i^2. Then the variable

$$Y = \frac{\sum_{i=1}^N X_i - \sum_{i=1}^N \mu_i}{\sum_{i=1}^N \sigma_i^2} \qquad (145)$$

  is distributed as a Gaussian with expectation value 0 and unit variance.
- Proof: Very simple using characteristic functions.