STATISTICAL INFERENCE

INTRODUCTORY NOTES

DR ROBERTO TROTTA

This is a collection of introductory notes for the International elite PhD Course given at the Niels Bohr Institute, Copenhagen, 6-10th Oct 2014. Students are advised to study this material prior to the course, in order to maximise their learning outcome during the course itself. The material will be covered during the first day of the course.

This symbol ⚠ denotes material that is particularly important.

Every effort has been made to correct any typos, but invariably some will remain. I would be grateful if you could point them out to me by e-mailing your corrections to: `r.trotta@imperial.ac.uk`.

Contents

1. Introduction

The purpose of physics is to learn about natural phenomena in the world. In order to do that, theoretical models (e.g., Newton's theory of gravitation) have to be validated through observations of the phenomena they aim to describe (e.g., measurement of the time it takes for an apple to fall). Thus an essential part of physics is the quantitative comparison of its theories (i.e., models, equations) with observations (i.e., data, measurements). This leads to validate theories or to refute them.

Measurements always come with uncertainties associated with them, for example because of noise in the measurement instrument. Statistics is the tool by which we can extract information about physical quantities from noisy data. The purpose of this short course is to give you an appreciation of the fundamental principles underpinning statistical inference (i.e., the process by which we reconstruct quantities of interest from noisy data) and to give you the tools to approach and solve some of the most common inference problems in the physical sciences.

Statistics addresses several relevant questions for physicists:

(i) How can we learn about regularities in the physical world given that any measurement is subject to a degree of randomness?
(ii) How do we quantify our uncertainty about observed properties in the world?
(iii) How can we make predictions about the future from past experience and theoretical models? (not covered in this course).

2. Probabilities

- 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.
- Let $A, B, C, \dots$ 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.

➽ **Example 1** 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".

[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".

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. (2) and (3):

⚠

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

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

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

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

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

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

3. 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.

➽ **Example 2** 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 $P(X = x_i)$.

➽ **Example 3** 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".

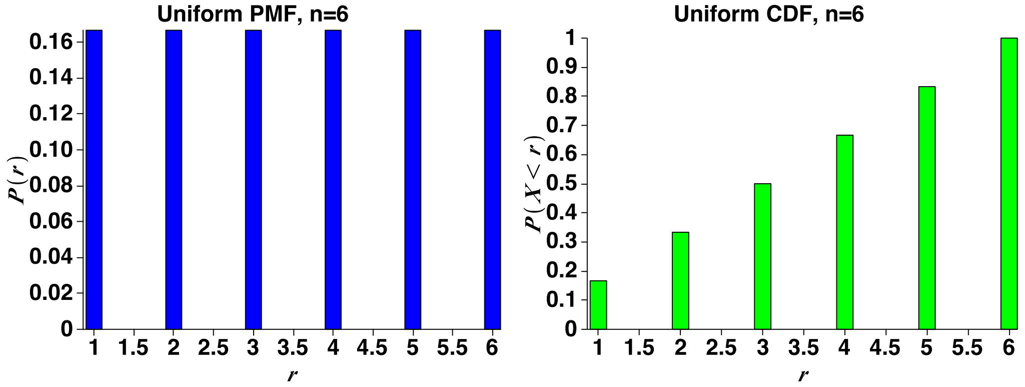

FIGURE 1. Left panel: uniform discrete distribution for $n = 6$. Right panel: the corresponding cdf.

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).

4. THE UNIFORM, BINOMIAL AND POISSON DISTRIBUTIONS

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

⚠

$$P(r) = \begin{cases} 1/n & \text{for } 1 \le r \le n \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

It is plotted in Fig. 1 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:

$$P(r|n,p) \equiv B(n,p) = \binom{n}{r} p^r (1-p)^{n-r}, \tag{13}$$

where the “choose” symbol is defined as

$$\binom{n}{r} \equiv \frac{n!}{(n-r)!r!} \tag{14}$$

for $0 \leq r \leq n$ (remember, $0! = 1$). Some examples of the binomial for different choices of n, p are plotted in Fig. 2.

➽ **Example 4**

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:

$$P(6|n=10, p=0.7) = \binom{10}{6} 0.7^6 0.3^4 = 0.2. \tag{15}$$

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.

- **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.

➽ **Example 5**

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**:

$$P(r|\lambda, t) \equiv \text{Poisson}(\lambda) = \frac{(\lambda t)^r}{r!} e^{-\lambda t}. \tag{16}$$

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:

$$P(r|\lambda, A) \equiv \text{Poisson}(\lambda) = \frac{(\lambda A)^r}{r!} e^{-\lambda A}. \tag{17}$$

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

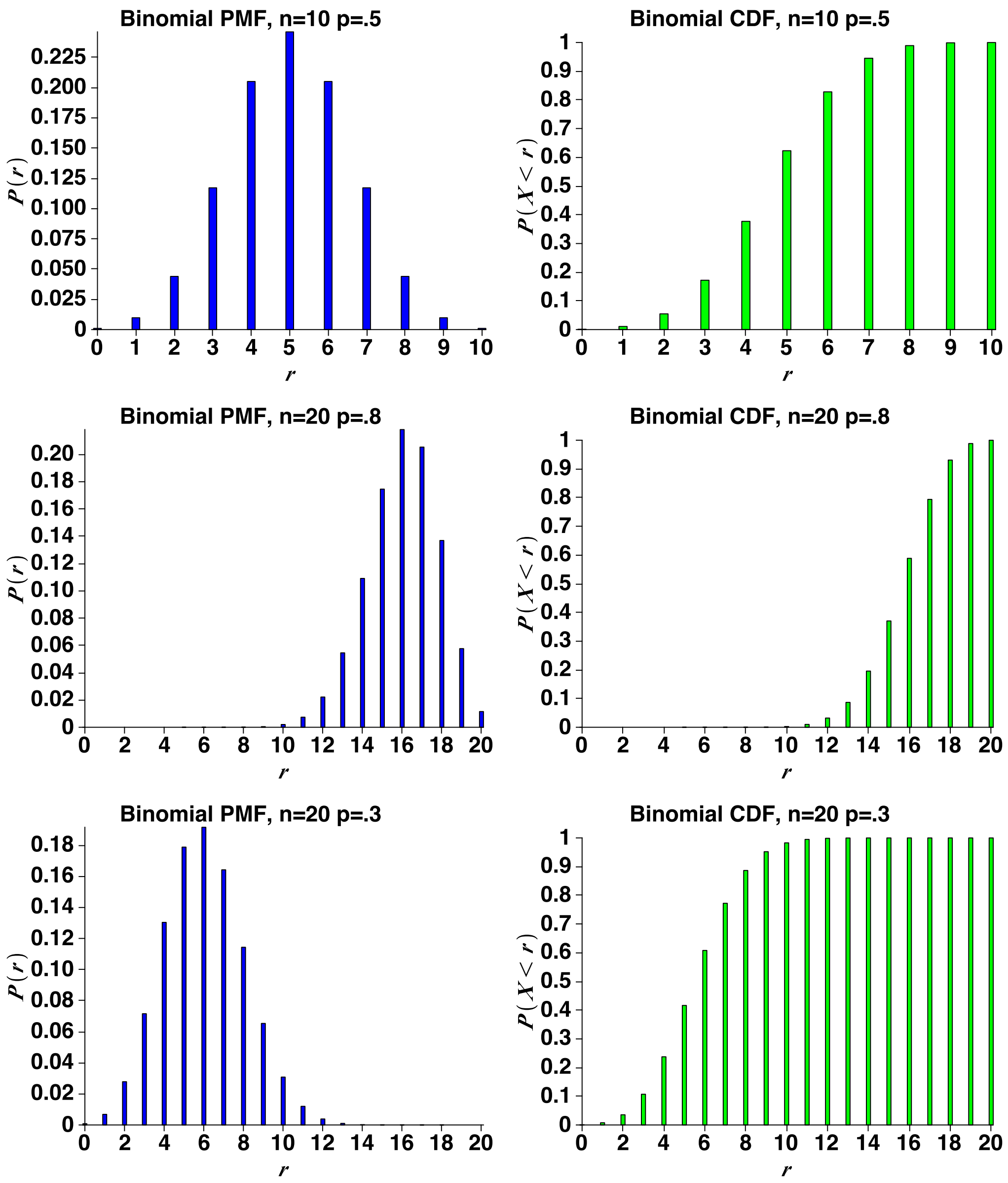

FIGURE 2. Some examples of the binomial distribution, Eq. (13), for different choices of n, p, and its corresponding cdf.

$t = 1$ in Eq. (16); similarly, the probability of getting r events in a unit area is obtained by setting $A = 1$ in Eq. (17)

Example 6 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:

$$P(6|\lambda = 4.5\text{s}^{-1}, t = 2\text{s}) = \frac{(4.5 \cdot 2)^6}{6!} e^{-4.5 \cdot 2} = 0.09. \tag{18}$$

So the probability is about 9%.

The Poisson distribution of Eq. (16) is plotted in Fig. 3 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

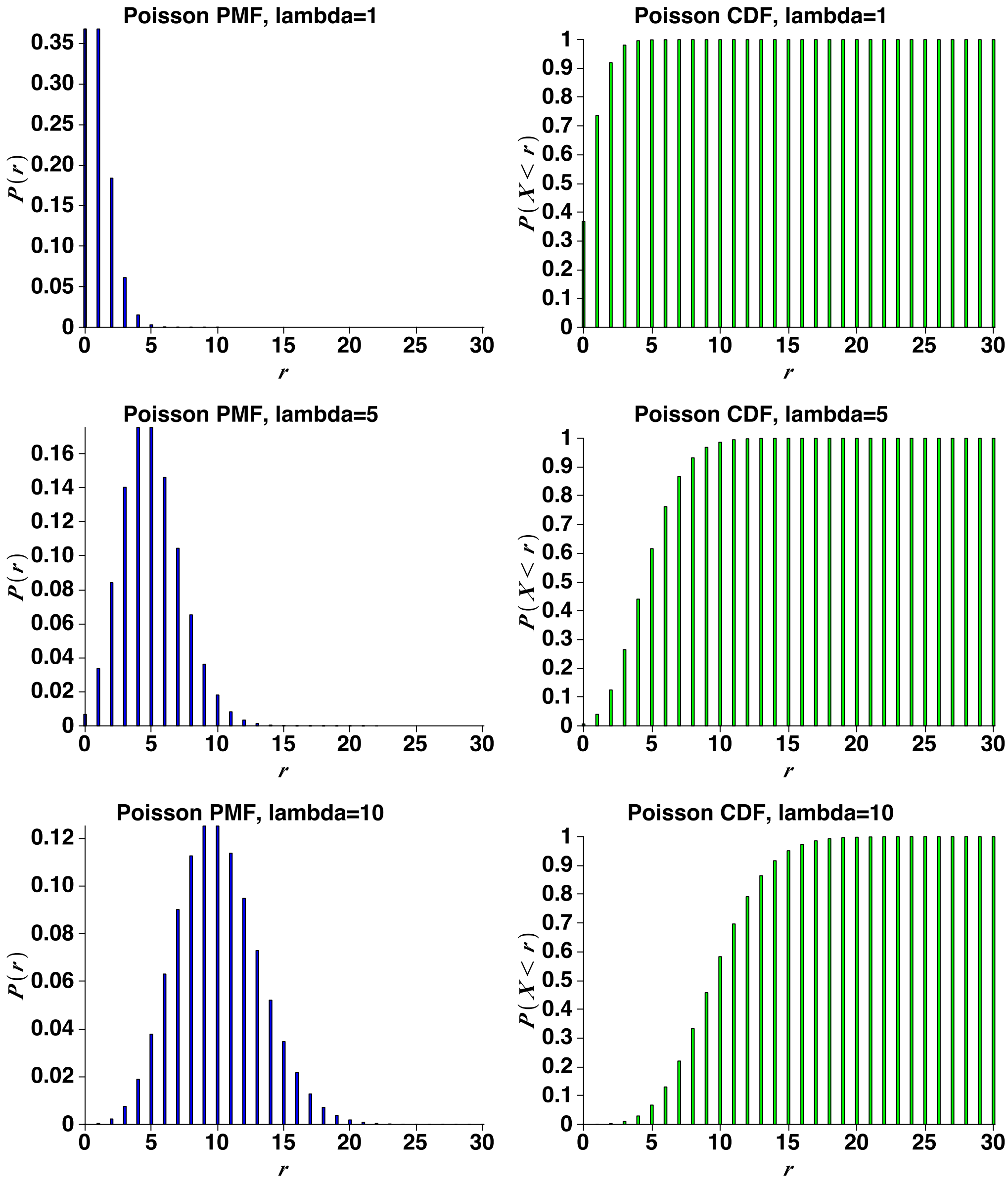

FIGURE 3. Some examples of the Poisson distribution, Eq. (16), for different choices of λ, and its corresponding cdf.

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).

➽ Example 7 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

$$P(n = 6|\lambda, t = 1\,\text{min}) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \approx 0.015 \tag{19}$$

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.

5. 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.
- The **expectation value** $E(X)$ (often called "mean", or "expected value"[2]) of the discrete RV X is defined as

$$E(X) = \langle X \rangle \equiv \sum_i x_i P_i. \tag{20}$$

➠ **Example 8** You toss a fair die, which follows the uniform discrete distribution, Eq. (12). 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

$$\mathrm{Var}(X) \equiv E[(X - E(X))^2] = E(X^2) - E(X)^2. \tag{21}$$

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

➠ **Example 9** For the case of tossing a fair die once, the variance is given by

$$\mathrm{Var}(X) = \sum_i (x_i - \langle X \rangle)^2 P_i = \sum_i x_i^2 P_i - \left(\sum_i x_i P_i\right)^2 = \sum_i i^2 \frac{1}{6} - \left(\frac{21}{6}\right)^2 = \frac{105}{36}. \tag{22}$$

- For the binomial distribution of Eq. (13), the expectation value and variance are given by:

$$E(X) = np, \qquad \mathrm{Var}(X) = np(1-p). \tag{23}$$

➠ **Example 10** 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. (23), with $p = 1/2$ (as the coin is fair), thus

$$E(H) = Np = N/2 \quad \text{and} \quad \mathrm{Var}(H) = Np(1-p) = N/4. \tag{24}$$

The probability of obtaining 8 or more heads is given by

$$P(H = 8 = \sum_{H=8}^{10} P(H\,\text{heads}|N, p = 1/2) = \frac{1}{2^{10}} \sum_{H=8}^{10} \binom{10}{H} = \frac{56}{1024} \approx 0.055. \tag{25}$$

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

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

- For the Poisson distribution of Eq. (16), the expectation value and variance are given by:

$$E(X) = \lambda t, \qquad \mathrm{Var}(X) = \lambda t, \tag{26}$$

while for the spatial version of the Poisson distribution, Eq. (17), they are given by:

$$E(X) = \lambda A, \qquad \mathrm{Var}(X) = \lambda A. \tag{27}$$

A proof of Eqs. (23) and (27) is given in Appendix A

- 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

$$E(X) = \langle X \rangle \equiv \int xp(x)dx. \tag{28}$$

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

$$\mathrm{Var}(X) \equiv E[(X - E(X))^2] = E(X^2) - E(X)^2 = \int x^2 p(x)dx - \left(\int xp(x)dx\right)^2. \tag{29}$$

6. 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. (16), 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

$$p(\text{1st event happens at time } t|\lambda) = \lambda e^{-\lambda t}, \tag{30}$$

where λ is the mean number of events per unit time. This is the exponential distribution.

Example 11 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

$$\int_{t_0}^{\infty} p(\text{1st event happens at time } t|\lambda)dt = \int_{t_0}^{\infty} \lambda e^{-\lambda t} dt = e^{-\lambda t_0} = 0.29, \tag{31}$$

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

$$p(T > t+s|T > s) = p(T > t). \tag{32}$$

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 A.3).

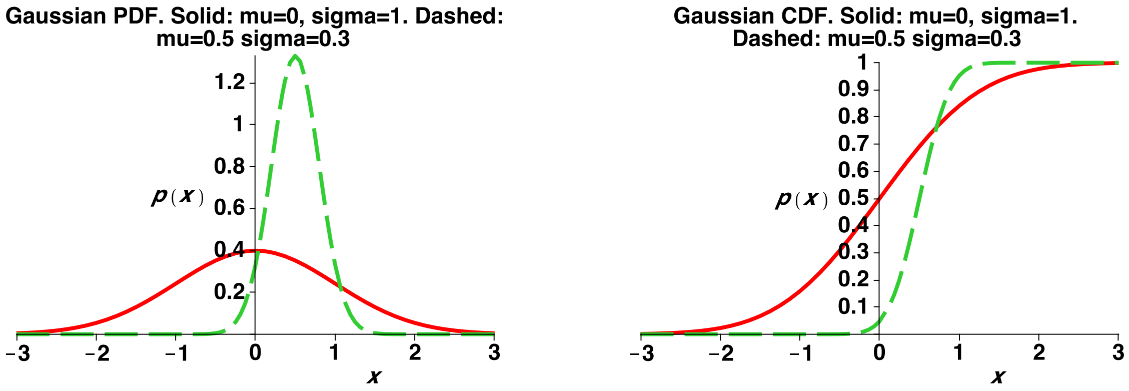

FIGURE 4. Two examples of the Gaussian distribution, Eq. (34), for different choices of μ, σ, and its corresponding cdf. The expectation value μ controls the location of the pdf (i.e., when changing μ the peak moves horizontally, without changing its shape), while the standard deviation σ controls its width (i.e., when changing σ the spread of the peak changes but not its location).

- For the exponential distribution of Eq. (30), the expectation value and variance for the time t are given by

$$E(t) = 1/\lambda, \qquad \mathrm{Var}(t) = 1/\lambda^2. \tag{33}$$

This is proved in the Appendix.

7. 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 8). A heuristic derivation of how the Gaussian arises follows from the example of darts throwing (see Appendix A.5).
- The Gaussian pdf is a continuous distribution with mean μ and standard deviation σ is given by

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right), \tag{34}$$

and it is plotted in Fig. 4 for two different choices of $\{\mu, \sigma\}$. The Gaussian is the famous bell-shaped curve.
- For the Gaussian distribution of Eq. (34), the expectation value and variance are given by:

$$E(X) = \mu, \qquad \mathrm{Var}(X) = \sigma^2. \tag{35}$$

This is proven in Appendix A.4.
- 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. 5, the Gaussian approximation to either the binomial or the Poisson distribution is very good even for fairly moderate values of n and λ.

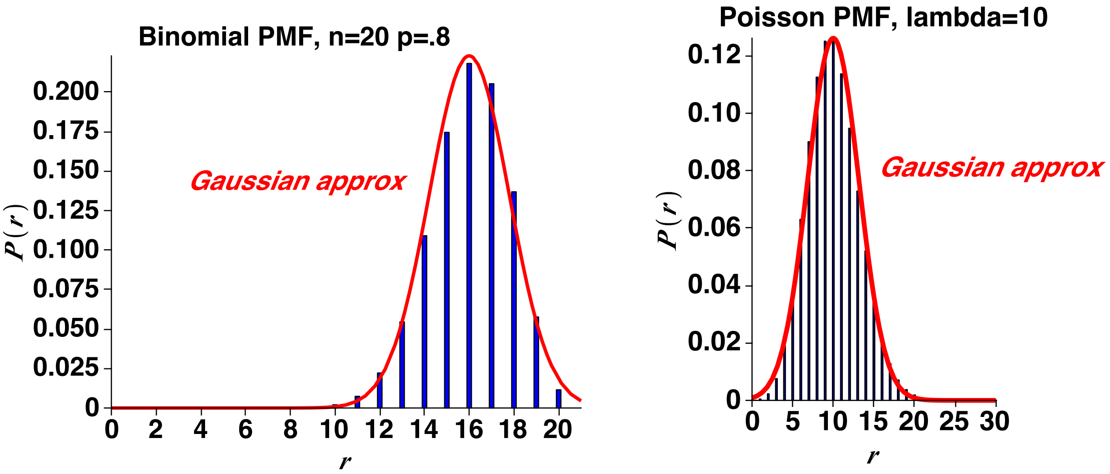

FIGURE 5. Gaussian approximation to the binomial (left panel) and the Poisson distribution (right panel). The solid curve gives in each case the Gaussian approximation to each pmf.

- 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

$$P(\mu-\kappa\sigma < x < \mu+\kappa\sigma) = \int_{\mu-\kappa\sigma}^{\mu+\kappa\sigma} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) dx \tag{36}$$

$$= \frac{2}{\sqrt{\pi}} \int_0^{\kappa/\sqrt{2}} \exp\left(-y^2\right) dy \tag{37}$$

$$= \mathrm{erf}(\kappa/\sqrt{2}), \tag{38}$$

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

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp\left(-y^2\right) dy, \tag{39}$$

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:

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) dx = \sqrt{2\pi}\sigma. \tag{40}$$

- Eq. (36) 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 1.

➽ **Example 12**

- 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 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 1). 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.

8. THE CENTRAL LIMIT THEOREM

- The Central Limit Theorem (CLT) is a very important result justifying why the Gaussian distribution is ubiquitous.

κ	$P(-\kappa < \frac{x-\mu}{\sigma} < \kappa)$	Usually called
"number of sigma"	Probability content	
1	0.683	1σ
2	0.954	2σ
3	0.997	3σ
4	0.9993	4σ
5	$1-5.7\times 10^{-7}$	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 1. Relationship between the size of the interval around the mean and the probability content for a Gaussian distribution.

⚠

- **Simple formulation of the CLT**: Let $X_1, X_2, \ldots, X_N$ 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 $X_1, X_2, \ldots, X_N$ be a collection of independent RV, each with finite expectation value μ_i and finite variance σ_i^2. Then the variable

(41)
$$Y = \frac{\sum_{i=1}^N X_i - \sum_{i=1}^N \mu_i}{\sum_{i=1}^N \sigma_i^2}$$

  is distributed as a Gaussian with expectation value 0 and unit variance.
- Proof: not required. Very simple using characteristic functions.

9. THE LIKELIHOOD FUNCTION

- The problem of **inference** can be stated as follows: given a collection of samples, $\{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N\}$, and a generating random process, what can be said about the properties of the underlying probability distribution?

➽ **Example 13** You toss a coin 5 times and obtain 1 head. What can be said about the fairness of the coin?

➽ **Example 14** With a photon counter you observe 10 photons in a minute. What can be said about the average photon rate from the source?

➽ **Example 15** 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:

(42)
$$\begin{gathered}\text{pdf - e.g., Gaussian with a given } (\mu,\sigma) \rightarrow \text{Probability of observation}\\ \text{Underlying } (\mu,\sigma) \leftarrow \text{Observed events}\end{gathered}$$

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

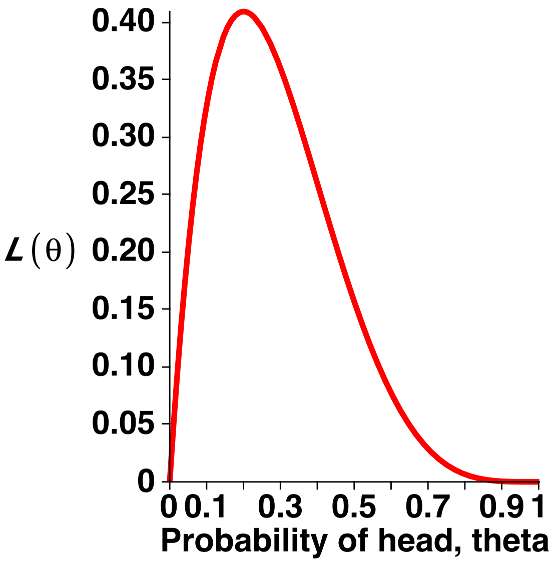

FIGURE 6. The likelihood function for the probability of heads (θ) for the coin tossing example, with $n = 5, r = 1$.

- 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[3] 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

$$\mathcal{L}(\theta) = p(X = \hat{\mathbf{x}}|\theta) \tag{43}$$

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.

➽ Example 16

- 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. (13), 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

$$\mathcal{L}(\theta) = \binom{5}{1}\theta^1(1-\theta)^4 = 5\theta(1-\theta)^4, \tag{44}$$

which is plotted as a function of θ in Fig. 6.

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. 7 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 13.

[3] 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.

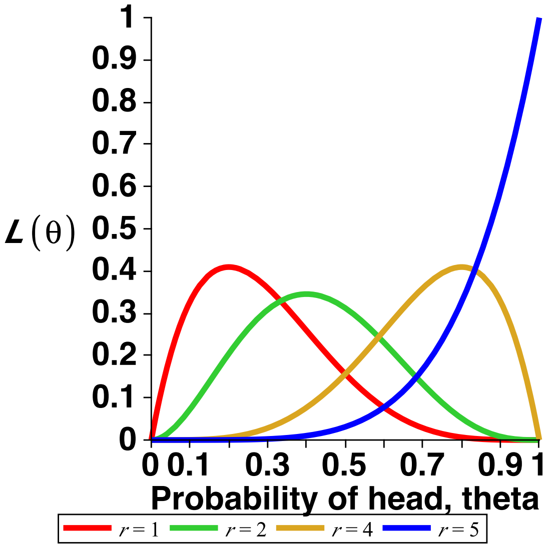

FIGURE 7. The likelihood function for the probability of heads (θ) for the coin tossing example, with $n = 5$ trials and different values of r.

10. THE MAXIMUM LIKELIHOOD PRINCIPLE

- **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

(45)
$$\theta_{\rm ML} \equiv \max_\theta \mathcal{L}(\theta)$$

- 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:

(46)
$$\left.\frac{\partial \mathcal{L}(\theta)}{\partial\theta}\right|_{\theta_{\rm ML}} = 0, \qquad \text{and} \qquad \left.\frac{\partial^2 \mathcal{L}(\theta)}{\partial\theta^2}\right|_{\theta_{\rm ML}} < 0.$$

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

(47)
$$\left.\frac{\partial \ln \mathcal{L}(\theta)}{\partial\theta}\right|_{\theta_{\rm ML}} = 0, \qquad \text{and} \qquad \left.\frac{\partial^2 \ln \mathcal{L}(\theta)}{\partial\theta^2}\right|_{\theta_{\rm ML}} < 0.$$

➽ Example 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 $\{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N\}$. 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

(48)
$$\mathcal{L}(\mu,\sigma) = p(\hat{\mathbf{x}}|\mu,\sigma) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2}\frac{(\hat{x}_i-\mu)^2}{\sigma^2}\right),$$

Note: often the Gaussian above is written as

$$\mathcal{L} = L_0 \exp\left(-\chi^2/2\right) \tag{49}$$

where the so-called “chi-squared” is defined as

$$\chi^2 = \sum_{i=1}^{N} \frac{(\hat{x}_i - \mu)^2}{\sigma^2}. \tag{50}$$

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

$$\frac{\partial \ln \mathcal{L}}{\partial \mu} = 0 \Rightarrow \mu_{\rm ML} = \frac{1}{N}\sum_{i=1}^{N} \hat{x}_i, \tag{51}$$

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

Example 18 ▸ **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:

$$\frac{\partial \ln \mathcal{L}}{\partial \sigma} = 0 \Rightarrow \sigma^2_{\rm ML} = \frac{1}{N}\sum_{i=1}^{N} (\hat{x}_i - \mu)^2. \tag{52}$$

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

$$E(\sigma^2_{\rm ML}) = (1 - \frac{1}{N})\sigma^2 \neq \sigma^2, \tag{53}$$

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. (52) by the MLE estimator for the mean, $\mu_{\rm ML}$.

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

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N} (\hat{x}_i - \mu_{\rm ML})^2. \tag{54}$$

In general, you should always use Eq. (54) as the ML estimator for the variance (and **not** Eq. (52)).

Example 19 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):

$$197.2, 202.4, 201.8, 198.8, 207.6, 191.4, 201.4, 198.2, 195.7, 201.2. \tag{55}$$

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. (48):

$$\mathcal{L}(T) = \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(\hat{T}_i - T)^2}{\sigma^2}\right) \tag{56}$$

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. (51), hence

$$T_{\rm ML} = \frac{1}{10}\sum_{i=1}^{10} \hat{T}_i = 199.6K. \tag{57}$$

➽ **Example 20**

- **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. (13)), 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):

(58)
$$\mathcal{L}(p) = P(H|p,n) = \binom{n}{H} p^H (1-p)^{n-H},$$

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

(59)
$$\frac{\partial \ln \mathcal{L}(p)}{\partial p} = \frac{\partial}{\partial p}\left(\ln\binom{n}{H} + H \ln p + (n-H)\ln(1-p)\right) = \frac{H}{p} - \frac{n-H}{1-p} \overset{!}{=} 0$$
$$\Leftrightarrow p_{\mathrm{ML}} = \frac{H}{n}.$$

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

➽ **Example 21**

- **MLE for the rate of a Poisson distribution.** The likelihood function is given by (see Eq. (16)), using the notation $\theta = \lambda$ (i.e., the parameter θ we are interested in is here the rate λ):

(60)
$$\mathcal{L}(\lambda) = P(n|\lambda) = \frac{(\lambda t)^n}{n!}\exp(-\lambda t),$$

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:

(61)
$$\frac{\partial \ln P(n|\lambda)}{\partial \lambda} = 0.$$

The derivative gives

(62)
$$\frac{\partial \ln P(n|\lambda)}{\partial \lambda} = \frac{\partial}{\partial \lambda}(n \ln(\lambda t) - \ln n! - \lambda t) = n\frac{t}{\lambda t} - t = 0 \Leftrightarrow \lambda_{MLE} = \frac{n}{t}.$$

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).

11. Confidence intervals

- 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}:

(63)
$$\ln \mathcal{L}(\theta) = \ln \mathcal{L}(\theta_{\mathrm{ML}}) + \frac{\partial \ln \mathcal{L}(\theta)}{\partial \theta}\Big|_{\theta_{\mathrm{ML}}}(\theta - \theta_{\mathrm{ML}}) + \frac{1}{2}\frac{\partial^2 \ln \mathcal{L}(\theta)}{\partial \theta^2}\Big|_{\theta_{\mathrm{ML}}}(\theta - \theta_{\mathrm{ML}})^2 + \ldots$$

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

$$\mathcal{L}(\theta) \approx \mathcal{L}(\theta_{\rm ML}) \exp\left(-\frac{1}{2}\frac{(\theta-\theta_{\rm ML})^2}{\Sigma_\theta^2}\right) + \dots, \tag{64}$$

with

$$\frac{1}{{\Sigma_\theta}^2} = -\frac{\partial^2 \ln \mathcal{L}(\theta)}{\partial \theta^2}\Big|_{\theta_{\rm ML}}. \tag{65}$$

So a general likelihood function can be approximated as a Gaussian around the ML value, as shown by Eq. (64). 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. (65).

➧ **Example 22**

- Let's go back to the Gaussian problem of Eq. (48). We have seen in Eq. (51) 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. (65) to the likelihood of Eq. (48) we obtain

$$\Sigma_\mu^2 = \sigma^2/N. \tag{66}$$

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.

➧ **Example 23**

- Going back to the numerical example of Eq. (55), 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 $P(\mu_{\rm min} < \mu < \mu_{\rm max}) = \alpha$.

➧ **Example 24**

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").

➧ **Example 25**

In the temperature measurement example of Eq. (55), the 68.3% confidence interval for the mean is 198.0 K $< \mu <$ 201.2 K. The 95.4% confidence interval is 196.4 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 1).
- 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.

➽ **Example 26** New Year Test sample question. The surface temperature on Mars is measured by a probe 10 times, yielding the following data (units of K):

$$191.9, 201.6, 206.1, 200.4, 203.2, 201.6, 196.5, 199.5, 194.1, 202.4 \tag{67}$$

(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:

$$\mathcal{L}_{\rm tot}(T) = \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(\hat{T}_i - T)^2}{\sigma^2}\right) \tag{68}$$

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. (51), hence

$$T_{\rm ML} = \frac{1}{10}\sum_{i=1}^{10} T_i = 199.7\text{K}. \tag{69}$$

(iii) Determine symmetric confidence intervals at 68.3%, 95.4% and 99% around $T_{\rm ML}$ (4 significant figures accuracy).
Answer: The variance of the mean is given by σ^2/N, see Eq. (66). Therefore the standard deviation of our estimate $T_{\rm ML}$ 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

$$1 = 5/\sqrt{N} \Rightarrow N = 25. \tag{70}$$

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

➽ **Example 27** 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:

$$119, \quad 119, \quad 122, \quad 121, \quad 116. \tag{71}$$

(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

$$P(\Delta|d) \equiv \mathcal{L}(\Delta) = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(\Delta - d_i)^2}{\sigma^2}\right), \tag{72}$$

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:

(73)
$$\frac{\partial \ln \mathcal{L}}{\partial \Delta} = -\sum_{i=1}^{5} \frac{\Delta - d_i}{\sigma^2} = 0 \rightarrow \Delta_{\rm MLE} = \frac{1}{N}\sum_{i=1}^{5} d_i$$

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

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

(74)
$$-\frac{\partial^2 \ln \mathcal{L}}{\partial \Delta^2} = \frac{N\Delta}{\sigma^2} \rightarrow \Sigma = \left(-\frac{\partial^2 \ln \mathcal{L}}{\partial \Delta^2}\right)^{-1/2} = \frac{\sigma}{\sqrt{N}}$$

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.

12. Propagation of errors

- Suppose we have measured a quantity x obtaining a measurement $\bar{x} \pm \sigma_x$. How do we propagate the measurement onto another variable $y = y(x)$?
- Taylor expanding $y(x)$ around $\bar{x}$ we obtain:

(75)
$$y(x) \approx y(\bar{x}) + (x - \bar{x})\frac{\partial y}{\partial x}\Big|_{x=\bar{x}} + \dots$$

Truncating the expansion at linear order, the expectation value of y is given by:

(76)
$$E(y) \approx E(y(\bar{x})) + \frac{\partial y}{\partial x}\Big|_{x=\bar{x}} E(x - \bar{x}) = y(\bar{x})$$

because $E(x - \bar{x}) = 0$.

The variance of y is given by:

(77)
$$V(y) = E([y(x) - E(y(x))]^2) = E([y(x) - y(\bar{x})]^2) = \left(\frac{\partial y}{\partial x}\Big|_{x=\bar{x}}\right)^2 \sigma_x^2.$$

So the variance on y is related to the variance on x by

(78)
$$\sigma_y^2 = \left(\frac{\partial y}{\partial x}\Big|_{x=\bar{x}}\right)^2 \sigma_x^2.$$

- Generalization to functions of several variables: if $y = y(x_1, \dots, x_N)$ then

(79)
$$\sigma_y^2 = \sum_{i=1}^{N}\left(\frac{\partial y}{\partial x_i}\Big|_{\mathbf{x}=\bar{\mathbf{x}}}\right)^2 \sigma_{x_i}^2.$$

Example 28 A couple of common cases are the following:

(i) Linear relationship: $y = ax$. Then $\sigma_y = a\sigma_x$.

(ii) Product or ratio: e.g. $y(x_1, x_2) = x_1 \cdot x_2$ or $y(x_1, x_2) = x_1/x_2$. Then

(80)
$$\frac{\sigma_y^2}{y^2} = \frac{\sigma_{x_1}^2}{x_1^2} + \frac{\sigma_{x_2}^2}{x_2^2}.$$

- **Systematic vs random errors:** errors are often divided in this two categories. Any measurement is subject to statistical fluctuations, whice means that if we repeat the same measurement we will obtain every time a slightly different outcome. This is a **statistical (or random) error**. Random errors manifest themeselves as noise in the measurement, which leads to variability in the data each time a measurement is made.

  On the other hand, **systematic errors** do not lead to variability in the measurement, but are the cause for data to be systematically "off" all the time (e.g., measuring a current in A while the apparatus really gives mA would lead to a factor of 1000 systematic error all the time). Systematic errors are usually more difficult to track down. They might arise by experimental mistakes, or because of unmodelled (or unrecognized) effects in the system you are measuring.

13. Bayesian statistics

- Bayes theorem, Eq. (4), 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 replace in Bayes theorem, Eq. (4), $A \to \theta$ (the parameters) and $B \to d$ (the observed data, or samples), obtaining

$$P(\theta|d) = \frac{P(d|\theta)P(\theta)}{P(d)}. \tag{81}$$

  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 (hence the name). This is an essential ingredient of Bayesian statistics.

  In the denominator, $P(d)$ is a normalizing constant (often called "the evidence"), than ensures that the posterior is normalized to unity:

$$P(d) = \int d\theta P(d|\theta)P(\theta). \tag{82}$$

  The evidence is important for Bayesian model selection (not covered in this course).
- **Interpretation:** Bayes theorem relates the posterior probability for θ (i.e., what we know about the parameter after seeing the data) to the likelihood and the prior (i.e., what we knew about the parameter before we saw the data). 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 $P(\theta|d) \neq P(d|\theta)$, i.e. the posterior $P(\theta|d)$ and the likelihood $P(d|\theta)$ are two different quantities with different meaning!

➽ **Example 29** We want to determine if a randomly-chosen person is male (M) or female (F). We make one measurement, giving us information on whether the person is pregnant (Y) or not (N). Let's assume we have observed that the person is pregnant, so $d = Y$.

The likelihood is $P(d = Y|\theta = F) = 0.03$ (i.e., there is a 3% probability that a randomly selected female is pregnant), but the posterior probability $P(\theta = F|d = Y) = 1.0$, i.e., if we have observed that the person is pregnant, we are sure she is a woman. This shows that the likelihood and the posterior probability are in general different!

This is because they mean two different things: the likelihood is the probability of making the observation if we know what the parameter is (in this example, if we know

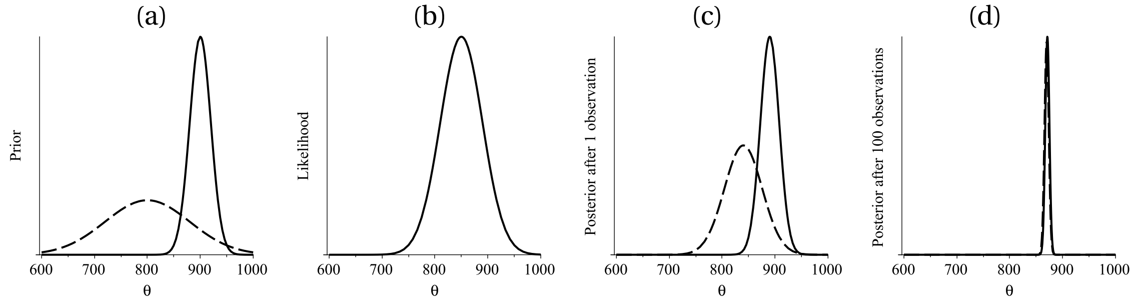

FIGURE 8. Converging views in Bayesian inference. Two scientists having different prior believes $P(\theta)$ about the value of a quantity θ (panel (a), the two curves representing two different priors) observe one datum with likelihood $\mathcal{L}(\theta)$ (panel (b)), after which their posteriors $P(\theta|d)$ (panel (c), obtained via Bayes Theorem, Eq. (4)) 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).

that the person is female); the posterior is the probability of the parameter given that we have made a certain observation (in this case, the probability of a person being female if we know she is pregnant). The two quantities are related by Bayes theorem (prove this in the example given here).

- 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.
- 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. 8 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 a mass), then the prior will be 0 for values $\theta < 0$.

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

(83)
$$P(\theta) = \begin{cases} \frac{1}{(\theta_{\max}-\theta_{\min})} & \text{for } \theta_{\min} \le \theta \le \theta_{\max} \\ 0 & \text{otherwise} \end{cases}$$

With this choice of prior in Bayes theorem, Eq. (81), the posterior becomes functionally identical to the likelihood up to a proportionality constant:

(84)
$$P(\theta|d) \propto P(d|\theta) = \mathcal{L}(\theta).$$

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).

Example 30

- Let's look once more to the temperature estimation problem of Eq. (55). The Bayesian estimation of the temperature proceeds as follows. We first need to specify the likelihood function – this is the same as before, and it is given by Eq. (55). If we want to estimate the temperature, we need to compute the posterior probability for T, given by (up to a normalization constant)

$$P(T|d) \propto \mathcal{L}(T)P(T) \tag{85}$$

where the likelihood $\mathcal{L}(T)$ is given by Eq. (55). We also need to specify the prior, $P(T)$. For this particular case, we know that $T > 0$ (the temperature in K of an object needs to be positive) and let's assume we know that the temperature cannot exceed 300 K. Therefore we can pick a flat prior of the form

$$P(T) = \begin{cases} \frac{1}{300} & \text{for } 0\text{ K} \leq T \leq 300\text{ K} \\ 0 & \text{otherwise.} \end{cases} \tag{86}$$

The posterior distribution for T then becomes

$$P(T|d) \propto \begin{cases} \frac{\mathcal{L}(T)}{300} & \text{for } 0\text{ K} \leq T \leq 300\text{ K} \\ 0 & \text{otherwise.} \end{cases} \tag{87}$$

So the posterior is identical to the likelihood (up to a proportionality constant), at least within the range of the flat prior. Hence we can conclude that the posterior is going to be a Gaussian (just like the likelihood) and we can immediately write the 68.3% posterior range of T as $198.0\text{ K} < \mu < 201.2\text{ K}$. This is **numerically** identical to our results obtained via the MLE. However, in this case the **interpretation** of this interval is that "after seeing the data, and given our prior as specified in Eq. (86), there is 68.3% probability that the true value of the temperature lies within the range $198.0\text{ K} < \mu < 201.2\text{ K}$".

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

$$P(\Psi) = P(\theta)\left|\frac{d\theta}{d\Psi}\right|. \tag{88}$$

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

Appendix A. Supplementary material

A.1. **The binomial distribution.** Here we prove Eq. (23).

$$\begin{aligned} E(x) &= \sum_{x=0}^{n} x\binom{n}{x} p^x(1-p)^{n-x} = \sum_{x=1}^{n} x\binom{n}{x} p^x(1-p)^{n-x} \\ &= np\sum_{x=1}^{n} x\frac{n!}{(n-x)!x!}\frac{p^x}{np}(1-p)^{n-x} = np\sum_{x=1}^{n}\frac{(n-1)!}{(n-x)!(x-1)!}p^{x-1}(1-p)^{n-x} \\ &= np\sum_{x=1}^{n}\frac{(n-1)!}{((n-1)-(x-1))!(x-1)!}p^{x-1}(1-p)^{(n-1)-(x-1)}, \end{aligned} \tag{89}$$

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

$$E(x) = np\sum_{s=0}^{m}\frac{m!}{(m-s)!s!}p^s(1-p)^{m-s} = np, \tag{90}$$

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

$$\text{Var}(x) = E(x^2) - E(x)^2 = E(x(x-1)+x) - E(x)^2 = E(x(x-1)) + E(x) - E(x)^2, \tag{91}$$

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

$$\begin{aligned}E(x(x-1)) &= \sum_{x=0}^{n} x(x-1)\binom{n}{x} p^x(1-p)^{n-x} = \sum_{x=2}^{n} \frac{n!}{(n-x)!(x-2)!} p^x(1-p)^{n-x} \\ &= n(n-1)p^2 \sum_{x=2}^{n} \frac{n!}{(n-x)!(x-2)!} \frac{p^x}{n(n-1)p^2} (1-p)^{n-x} \\ &= n(n-1)p^2 \sum_{x=2}^{n} \frac{(n-2)!}{((n-2)-(x-2))!(x-2)!} p^{x-2}(1-p)^{(n-2)-(x-2)}.\end{aligned} \tag{92}$$

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

$$E(x(x-1)) = n(n-1)p^2 \sum_{s=0}^{m} \frac{m!}{(m-s)!s!} p^s(1-p)^{m-s} = n(n-1)p^2. \tag{93}$$

Thus

$$\mathrm{Var}(x) = E(x(x-1)) + E(x) - E(x)^2 = n(n-1)p^2 + np - (np)^2 = np(1-p). \tag{94}$$

A.2. **The Poisson distribution.** Here we prove Eq. (27) (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

$$\frac{de^x}{dx} = \sum_k k \frac{x^{k-1}}{k!} = e^x \tag{95}$$

and multiplying both sides by x we obtain:

$$xe^x = \sum_k k \frac{x^k}{k!}. \tag{96}$$

The expectation value is given by

$$E(n) = \sum_{n=0}^{\infty} n \cdot \mathrm{Poisson}(\lambda) = \sum_{n=0}^{\infty} n \frac{\lambda^n}{n!} e^{-\lambda} = \lambda e^{\lambda} e^{-\lambda} = \lambda, \tag{97}$$

where in the penultimate step we have made use of Eq. (96).

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

$$\frac{d^2e^x}{d^2x} = \sum_k k(k-1)\frac{x^{k-2}}{k!} = e^x \tag{98}$$

hence

$$x^2e^x = \sum_k k(k-1)\frac{x^k}{k!} = \sum_k (k^2-k)\frac{x^k}{k!}. \tag{99}$$

Therefore

$$\sum_k k^2 \frac{x^k}{k!} = x^2e^x + xe^x. \tag{100}$$

We can now compute

$$E(n^2) = \sum_{n=0}^{\infty} n^2 \cdot \mathrm{Poisson}(\lambda) = \sum_{n=0}^{\infty} n^2 \frac{\lambda^n}{n!} e^{-\lambda} = (\lambda^2+\lambda)e^{\lambda}e^{-\lambda} = \lambda^2 + \lambda, \tag{101}$$

so that

$$\mathrm{Var}(n) = \lambda^2 + \lambda - \lambda^2 = \lambda. \tag{102}$$

A.3. **The exponential distribution.** Here we prove Eq. (33). The expectation value is given by

$$E(t) = \int_0^\infty t\lambda e^{-\lambda t} dt = \frac{1}{\lambda}\int_0^\infty x e^{-x} dx = \frac{1}{\lambda}\left(-xe^{-x}\Big|_0^\infty - \int_0^\infty (-e^{-x})\,dx\right) = \frac{1}{\lambda}, \tag{103}$$

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

$$\mathrm{Var}(t) = E(t^2) - E(t)^2 = \int_0^\infty t^2\lambda e^{-\lambda t} dt - E(t)^2 = \frac{1}{\lambda}\int_0^\infty t^2\lambda^2 e^{-\lambda t} dt - \frac{1}{\lambda^2} \tag{104}$$

$$= \frac{1}{\lambda^2}\int_0^\infty x^2 e^{-x} dx - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}. \tag{105}$$

We now prove the "lack of memory" property, i.e. Eq. (32). The probability that at least one event has happened until time t is given by the cumulative distribution function

$$F_\lambda(t) = \int_0^t \lambda e^{-\lambda\tau} d\tau = \lambda\int_0^{\lambda t} \frac{1}{\lambda} e^{-x} dx = -e^{-x}\Big|_0^t = 1 - e^{-\lambda t}. \tag{106}$$

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

$$P(T > t+s | T > s) = \frac{P(T > t+s, T > s)}{P(T > s)} = \frac{P(T > t+s)}{P(T > s)}, \tag{107}$$

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

$$P(T > t+s | T > s) = \frac{1 - F_\lambda(t+s)}{1 - F_\lambda(s)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = P(T > t). \tag{108}$$

A.4. **The Gaussian distribution.** Here we prove Eq. (35). The expectation value is given by

$$E(X) = \int x\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx = \frac{1}{\sqrt{2\pi}}\int (\sigma t + \mu) e^{-\frac{1}{2}t^2} dt, \tag{109}$$

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

$$E(X) = \frac{\mu}{\sqrt{2\pi}}\int e^{-\frac{1}{2}t^2} dt = \mu \tag{110}$$

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

$$E(X^2) = \int x^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx \frac{1}{\sqrt{2\pi}}\int (\sigma t + \mu)^2 e^{-\frac{1}{2}t^2} dt \tag{111}$$

$$= \frac{\sigma^2}{\sqrt{2\pi}}\int t^2 e^{-\frac{1}{2}t^2} dt + \frac{\mu^2}{\sqrt{2\pi}}\int e^{-\frac{1}{2}t^2} dt + \frac{\mu}{\sqrt{2\pi}}\int t e^{-\frac{1}{2}t^2} dt \tag{112}$$

$$= \sigma^2 + \mu^2 + 0, \tag{113}$$

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

A.5. **Heuristic derivation of the Gaussian distribution.** 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:

$$p(x)\Delta x\cdot p(y)\Delta y = f(r)\Delta x\Delta y, \tag{114}$$

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:

$$\left(p(x)\frac{dp(x)}{d\phi}+p(y)\frac{dp(y)}{d\phi}\right)\Delta x\Delta y = 0. \tag{115}$$

(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

$$\frac{dp(x)}{d\phi} = \frac{\partial p}{\partial x}\frac{\partial x}{\partial \phi} = -\frac{\partial p}{\partial x}y, \tag{116}$$

$$\frac{dp(y)}{d\phi} = \frac{\partial p}{\partial y}\frac{\partial y}{\partial \phi} = \frac{\partial p}{\partial y}x. \tag{117}$$

Eq. (115) becomes

$$\left(-p(x)\frac{\partial p}{\partial x}y + p(y)\frac{\partial p}{\partial y}x\right)\Delta x\Delta y = 0, \tag{118}$$

which implies

$$\frac{p(x)}{x}\frac{\partial p}{\partial x} = \frac{p(y)}{y}\frac{\partial p}{\partial y}. \tag{119}$$

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

$$\frac{\partial p}{\partial x} = Cxp(x) \tag{120}$$

(and similarly for y). Integration gives the solution

$$p(x) = Ae^{\frac{C}{2}x^2} \tag{121}$$

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 $p(y)$).