Data Modeling for the Sciences

Applications, Basics, Computations

Steve Pressé and Ioannis Sgouralis

(This draft was last modified on January 8, 2022)

Preface

Data analysis courses that go beyond teaching elementary topics, such as fitting residuals, are rarely offered to students in the Natural Sciences. As a result, data analysis, much like programming, is something often learned and improvised "on the job". Yet, with an explosion of experimental methods generating large quantities of diverse data, we believe that students and researchers alike would benefit from a clear presentation of methods of analysis many of which have only become feasible due to the practical needs and computational advances of the last decade or two.

The framework for data analysis that we provide here is inspired by exciting new developments in Data Science, Machine Learning and Statistics in a language accessible to the broader community of Natural scientists. As such, this text is ambitiously aimed at making topics such as statistical inference, computational modeling and simulation approachable to the Natural Sciences.

It is also a goal of ours, if nothing else, to help develop an appreciation for data-driven modeling and what data analysis choices are available to us alongside what approximations are inherent to the choices explicitly or implicitly made. We do so because theory in the Physical Sciences has traditionally provided limited emphasis on data-driven approaches. Indeed, the prevailing philosophy is that models are first proposed and then verified or otherwise disproven by experiments. But this approach is not data-centric. Nor is it rigorous except for the cleanest of experimental data sets as one's perceived choice in how to compare models and experiments may have dramatic consequences in whether the model is ultimately falsified. As we move toward monitoring events on smaller or faster timescales or complex events otherwise sparsely sampled, examples of clean data are already few and far between.

We designed the text as a self-contained single semester course in data analysis, statistical modeling and inference. We have now use it in teaching at Arizona State University since 2017 to first year Chemistry and Physics graduate students as well as upper-level undergraduates. While the text is appropriate for upper-level undergraduates in the Physical Sciences, its intended audience is at the master's level. The concepts presented herein are self-contained though a basic course in computer programming and prior knowledge of undergraduate level calculus is assumed.

Our text places equal emphasis on explaining the foundations of existing methods and their implementation. It correspondingly places little emphasis on formal proofs and research active topics at the forefront of data analysis yet to be settled. Along core sections, we have interspersed sections and topics designated by an asterisk. These contain more advanced materials that may be included at the instructor's discretion and are otherwise not necessary upon a first reading.

The text begins with a survey of modeling concepts to motivate the problem of parameter estimation from given data. This leads to a discussion of frequentist and Bayesian inference tools. Along the way, we introduce computational techniques including Monte Carlo methods that are necessary for a comprehensive exposition of the most recent advances. The second half of the text is devoted to specific models starting from basic mixture models followed by Gaussian processes, the hidden Markov model, its adaptations as well as its generalization to state-space models and continuous time representations.

Finally, we made clear choices on what topics to include in the book. These were sometimes based on personal interest though, most often, these choices were based on what we believe is most relevant. To keep our presentation streamlined, however, we have excluded many topics. Some of these are topics that we perceive as easier for students to understand after reading this book, such as special cases or otherwise specialized generalizations of the topics covered herein.

Tempe, AZ
Knoxville, TN
January 8, 2022

Thanks

Many thanks to Weiqing Xu for generating figures in the earlier chapters as well as Sina Jazani and Zeliha Kilic for helping write portions of the text. Special thanks to Julian Lee and Corey Weistuch for reading over the text. We also thank the many other members of the Pressé lab and students at Arizona State University taking CHM/PHY 598 ("Unraveling the noise") who have suggested many revisions and identified typos across earlier drafts of the text. Any remaining typos and omissions are ours alone.

Short contents

Contents

*This is an advanced topic and could be skipped on a first reading.

†This is an advanced topic and could be skipped on a first reading.

‡This is an advanced topic and could be skipped on a first reading.

*This is an advanced topic and could be skipped on a first reading.
†This is an advanced topic and could be skipped on a first reading.

Part I

Concepts from modeling, inference, and computing

Chapter 1

Probabilistic modeling and inference

By the end of this chapter, we will have presented
- *Data oriented modeling*
- *Random variables and their properties*
- *An overview of inverse problem solving*

1.1 Modeling with data

If experimental observations or, put differently, binaries on a screen were all we ever cared about, then no experiment would require modeling or interpretation and the remainder of this book would be unnecessary. But binaries on a screen do not constitute knowledge. They constitute *data*. Put differently, Quantum Mechanics like any scientific knowledge is not self-evident from the pixelated outcome on a camera chip of a modern incarnation of a Young's two-slit interference experiment.

In the Natural Sciences, *models* of physical systems provide mathematical frameworks in which we unify disparate pieces of information. These include conceptual notions such as symmetries, fundamental constituents and other postulates as well as scientific *measurements* and, even more generally, empirical observations of any form. If we think of direct observations as data in particular, at least for now, we can think of mathematical models as a way of compressing or summarizing all these data.

Data summaries may be used to make predictions about physical conditions we may encounter in the future, such as in new experiments, or to interpret and describe an underlying physical system already probed in past experiments. For example, with time-ordered data we may be interested in learning equations of motion or kinetic schemes. Or, already knowing a kinetic scheme sufficiently well from past experiments or fundamental postulates, we may only be interested in learning the noise characteristics of a new piece of equipment on which future experiments will be run. Thus, models may be aimed at discovering new Science as well as at devising careful controls to get a better handle on error bars and, more broadly, even at designing new experiments altogether.

1.1.1 Why do we obtain models from raw data?

Experimental data rarely provide direct insight on the physical conditions and systems of interest. At the very least, measurements are *corrupted* by unavoidable noise and, as a result, models obtained from experimental data are unavoidably probabilistic. So, we ask: *how should we, the scientific community, go about obtaining models from imperfect data?*

Note 1.1: Obtaining models from data

Data can be time and labor intensive to acquire. Perhaps more importantly, every datum in a dataset encodes information. In light of this, we re-pitch our question and ask: *how should we go about obtaining models* efficiently *and* without compromising *the information encoded in the data?*

The key is to start from the data acquired in the experiments and arrive at models with a minimal amount of pre-processing, if at all. This is because obtaining a model from quantities derived from the data, as opposed to directly from the data, is necessarily *equal to* or *worse than* obtaining the model from the data directly since derived quantities contain as much as or less information than the data themselves. For instance, fitting histogrammed data is an information inefficient and unreliable approach to obtaining models as it demands downsampling via binning and a more or less arbitrary choice of bin sizes.

Besides information efficiency, obtaining models from unprocessed data also has another critical advantage that gets to the heart of scientific practice. While error bars around individual data points may be imperfectly known, they are, by construction, *better characterized* than error bars around derived quantities. Thus error bars around models determined from derived quantities are necessarily only as good as, but often less reliable, than error bars around models determined from the data. Unfortunately, as error bars around derived quantities can become too difficult to compute in practice, they are often ignored altogether. Nevertheless, error bars are a cornerstone of modern scientific research. They not only help quantify reproducibility but they also directly inform error bars around the models obtained and, as such, inspire the formulation of new competing models.

Putting it all together, it becomes clear that a model is *best informed*, and has the *most reliable error bars*, when learned from the data available in as raw a form as accessible from the experiments. This is true so long as it is computationally feasible to obtain models from such raw data and, as we will see in subsequent chapters, we are far from reaching computational bottlenecks in most problems of interest across the Natural Sciences.

1.1.2 Why do we formulate models with random variables?

If there is no uncertainty involved, a physical system is adequately described using deterministic variables. For example, Newtonian mechanics are expressed in terms of momenta, positions, and forces. However, when a system involves any degree of uncertainty, either due to noise, poor characterization of some or all of its constituents, features as of yet unresolved or otherwise fundamentally stochastic, then it is better described by *random variables*. This is true of the probabilistic nature of Quantum Mechanics as well as Statistical Physics and, as we illustrate herewith, also of Data Analysis.

In this book we focus on the latter case; namely *stochastic systems*. Stochasticity in our systems arise due to inherent randomness in the physical phenomena of interest or due to measurement noise or both. We represent observations generated by stochastic systems as *random variables*. This is because, as we will see, random variables are mathematical notions that can reproduce naturally stochastic relationships between uncertain effects and observations; while, their deterministic counterparts cannot.

Note 1.2: Measurement noise

It is sometimes thought that models with probabilistic formulations are only required when the quantities of interest are inherently probabilistic. Nevertheless, measurement noise corrupts experimental observations irrespective of the quantities themselves being probabilistic or not. Consequently, probabilistic models are *always required* whenever models are informed by experimental output.

Random variables are abstract notions that most often represent numbers or collections of numbers. However, random variables are generic notions and they may also be non-numeric such as: labels for grouping data, *e.g.* group A, group B; logical indicators, *e.g.* true, false; functions, *e.g.* trajectories or energy potentials. In all cases, numeric or not, random variables may be *discrete*, *e.g.* dice rolls, coin flips, photon counts, bound energy states or *continuous*, such as temperatures, pressures or distances. Further, random variables may be finite collections of individual quantities, *e.g.* measurements acquired during an experiment or even infinite ones, *e.g.* successive positions on a *Brownian particle's* trajectory. At any rate, random variables have unique properties, which we will shortly explore, that allow us to use them in the construction and evaluation of meaningful probabilistic models.

Commonly, we imagine a random variable, which we denote with W, as being instantiated or assigned a specific value realized at w as a result of performing a measurement which amounts to a *stochastic event*. That is, we think of a measurement output w as a *stochastic realization* of W. Our stochastic events entail randomness inherited through W and influencing the assigned values w.

Stochastic events may encompass *physical* events, like the occurrence of chemical reactions or events in a cell's life cycle. Stochastic events may also encompass *conceptual* events, like an idealized version of a real-life system expressed in terms of fair coin tosses or, even, like instantaneously learning the spin orientation of a faraway particle given a local measurement of another spin to which the first is entangled.

Example 1.1: The photo-electric effect

When a photon falls into certain materials, a photo-electron is sometimes emitted and sometimes not. Such a phenomenon provides the basis for a stochastic event.

In the photo-electric setting, it is often convenient to formulate a random variable W that counts the number of photo-electrons emitted. This random variable may take values $w = 0, 1, 2, \cdots$.

Throughout this introductory chapter, we will distinguish between a random variable W and its realizations, w, *i.e.*, the particular values that W attains or may attain. Due to clarity, we will be strict in our notation but, as we move forward, in the subsequent chapters we will gradually relax our convention since what is meant by w, serving as a proxy for either the value w or the variable W, will be clear from the context of the discussion.

To develop a model, we imagine a *prototype experiment* as a sequence of stochastic events that produce N numeric measurements or, more generally, observations of any kind. We typically use w_n to denote the n^{th} observation and use $n = 1, \cdots, N$ to index them. Individual observations in our experiment may be scalar values, for example $w_n = 20.1°\text{C}$ or $w_n = 0.74\ \mu\text{m}^3$ for typical measurements of room temperature or an *E. coli*'s volume, respectively. Individual observations may also be non-numeric, such as $w_n =$ p.R83SfsX15 for descriptions of gene variations. In general, we do not require that each observation in our experiment be of the same type; that is, w_1 may be a temperature while w_2 may be a volume.

As we will often do, we gather every observation conveniently together in a list

$$w_{1:N} = \{w_1, w_2, \cdots, w_N\}$$

and use subscripts $1:N$ to indicate that the list $w_{1:N}$ gathers every single w_n with an index n ranging between 1 and N. Unless explicitly needed to help draw attention to the subscript, for clarity, we may sometimes suppress this subscript and write simply w for the entire list.

As we have already mentioned, the observations $w_{1:N}$ are better understood as realizations of appropriate random variables $W_{1:N} = \{W_1, W_2, \cdots, W_N\}$ that we use to formulate our model. In principle, each observation w_n may be obtained under different conditions, and so our $W_{1:N}$ may gather random variables W_n with different properties.

Example 1.2: Photo-electric assessments

Consider a simple experiment where a photon source successively sends N bursts of photons that impinge upon a photo-electric material. Suppose further that, each time a photon burst is sent out from the source, we assess how many photo-electrons are produced. In this case, our assessments can be modeled by $W_{1:N} = \{W_1, W_2, \ldots, W_N\}$, where each variable W_n may take values $w_n = 0, 1, 2, \cdots$.

Provided that the intensity of the photon bursts remains constant, it is reasonable to consider all variables in $W_{1:N}$ as maintaining the same statistical properties; however, if the intensity of each burst changes over time, then our model needs to account for different statistics for each W_n.

1.1.3 Why do our models have parameters?

Models are mathematical formulations to which we associate parameters. Both models and their associated parameters are specialized to particular systems, experiments or experimental setups. Assuming a model structure encoded in $W_{1:N}$ and provided observed values $w_{1:N}$, our main objective in Data Analysis becomes the estimation of the model's associated parameters.

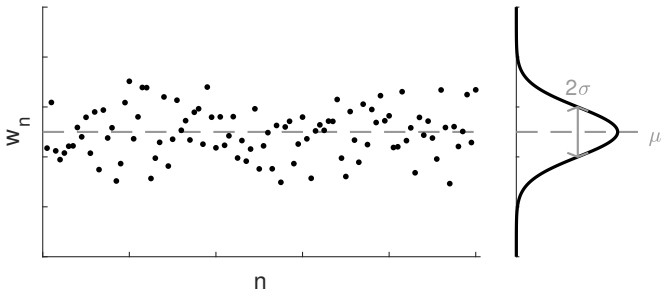

Figure 1.1: On the left hand side we show the output of an experiment after successive trials which we index with n. On the right hand side we find a histogram of the data with very fine bin sizes that assumes the shape of a Gaussian distribution. We denote the mean of this distribution by μ and standard deviation by σ.

Example 1.3: Normal random variables

The mean of a sequence of identical random variables W_n is only probabilistically related to each measured value w_n. For the simple example of a normally distributed sequence W_n, what we call the "model" is the Normaldistribution (often also termed the Gaussian distribution) and its associated parameters; the mean μ and variance σ^2 which indicate the center and spread of the values $w_{1:N}$, respectively. These are collectively described by the list of model parameters $\theta = \{\mu, \sigma^2\}$. As illustrated in fig. 1.1, and as we will see in detail in later chapters, θ can be estimated from $w_{1:N}$.

In the previous example, the Gaussian forms a simple model that contains two parameters, namely the mean μ and the variance σ^2, that we gather in θ. More generally, our models may contain K individual parameters that we may also gather in a list $\theta_{1:K} = \{\theta_1, \theta_2, \ldots, \theta_K\}$.

Typically, the parameters $\theta_{1:K}$ represent quantities we care to *estimate*, for example μ and σ^2. A model is deemed *specified* when *numerical values* are assigned to $\theta_{1:K}$. Thus, specifying a model is understood as being equivalent to assigning values to $\theta_{1:K}$. Similarly, deriving error bars around the assigned values of $\theta_{1:K}$ is equivalent to deriving error bars around the model.

As we invariably always face some degree of measurement noise, we formulate an experiment's results $w_{1:N}$ as probabilistically related to the parameters $\theta_{1:K}$. In the context of our *prototype experiment*, we incorporate such relations through the random variables $W_{1:N}$ and in the next section we lay down some necessary concepts.

Note 1.3: Modeling terminology

Strictly speaking, by model in this chapter we mean the mathematical formulation itself alongside numerical values for its associated parameters. When we speak of measurements, observations, assessments, or data points we refer to the random variables $W_{1:N}$ and their realizations $w_{1:N}$. Similarly, by calibrating a model we imply selecting the correct values for its associated parameters (and sometime also characterizing their uncertainty). Selecting both model parameters and their uncertainty is collectively referred to as model estimation or model training.

1.2 Working with random variables

Before we embark on specific modeling and estimation strategies, we begin by exploring some important notions that we need in order to work with random variables and the distributions from which they are sampled.

That is, just as we can easily deduce derivatives and integrals of complex functions by remembering a few simple rules of Calculus, we can similarly deduce probability distributions of complex models by remembering a few simple rules of probability that we put forth in this section.

As we will soon start using random variables not only to represent measurements W, but also other relevant quantities of our model, we begin using R to label generic random variables.

1.2.1 How to assign probability distributions

In any model, a random variable R is *drawn* or *sampled from* some *probability distribution*. We label such a distribution with $\mathbb{P}$ and we write

$$R \sim \mathbb{P}.$$

In the language of Statistics this reads "the random variable R is sampled from the probability distribution $\mathbb{P}$" or "R follows the statistics of $\mathbb{P}$".

In a statistical formulation like $R \sim \mathbb{P}$, we use $\mathbb{P}$ as a notational shorthand that summarizes the most important characteristics of the variable R. These include a description of the values r that R may take and a recipe to compute probabilities associated with these r's. As we will see in many cases, most often we work with probability density functions that are specified by the coinciding probability distributions. In such cases, it is more convenient to think of $R \sim \mathbb{P}$ as a compact way of communicating: (i) what the possible values r of R are; and (ii) that these values obey the probability density $p(r)$ associated with $\mathbb{P}$.

Example 1.4: The normal distribution

We previously encountered the normal distribution, Normal(μ, σ^2). A shorthand like

$$R \sim \text{Normal}\left(\mu, \sigma^2\right)$$

captures the following pieces of information:

- The particular values r that R attains are real numbers ranging from $-\infty$ to $+\infty$.
- The probability density $p(r)$ of R depends on two parameters, μ and σ^2, and has the form

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

Furthermore, the two parameters μ and σ^2 can be interpreted as the mean and the variance of R, respectively, since integration of the density leads to

$$\begin{aligned}(\text{Mean of } R) &= \int_{-\infty}^{+\infty} dr\, r p(r) = \mu, \\ (\text{Variance of } R) &= \int_{-\infty}^{+\infty} dr\, (r-\mu)^2 p(r) = \sigma^2.\end{aligned}$$

Using the density $p(r)$, we can also compute the probability of measuring any value r between some specified $r_{\min}$ and $r_{\max}$. In particular, this is

$$\int_{r_{\min}}^{r_{\max}} dr\, p(r) = \frac{1}{2}\left[\text{erf}\left(\frac{r_{\max}-\mu}{\sigma\sqrt{2}}\right) - \text{erf}\left(\frac{r_{\min}-\mu}{\sigma\sqrt{2}}\right)\right] \tag{1.1}$$

where erf is the error function and it is defined by an integral

$$\text{erf}\,(r) = \frac{2}{\sqrt{\pi}} \int_0^r dr'\, e^{-\frac{(r')^2}{2}}.$$

Example 1.5: The exponential distribution

The exponential distribution arises in many applications. A shorthand like

$$R \sim \text{Exponential}\,(\lambda)$$

captures the following pieces of information:

- The particular values r that R attains are real numbers ranging from 0 to ∞.

- The probability density $p(r)$ of R depends on one parameter, λ, and has the form

$$p(r) = \lambda e^{-\lambda r}.$$

The parameter λ can be interpreted as the reciprocal of the mean of R, since integration of the density leads to

$$(\text{Mean of } R) = \int_0^\infty dr\, r p(r) = \frac{1}{\lambda}.$$

Through the density $p(r)$, we can also compute the probability of measuring any value r between some specified $r_{\min}$ and $r_{\max}$. In particular, this is

$$\int_{r_{\min}}^{r_{\max}} dr\, p(r) = e^{-\lambda r_{\min}} - e^{-\lambda r_{\max}}. \tag{1.2}$$

Throughout this book, we extensively use several common distributions. In examples 1.4 and 1.5 we introduced two of them though more are to come. As these will appear frequently, to refer back to them, we adopt a convention that we summarize in appendix B. Briefly, we use $R \sim \text{Normal}(\mu, \sigma^2)$ and $\text{Normal}(\mu, \sigma^2)$ to denote a normal random variable and the normal distribution, respectively. Furthermore, we use $\text{Normal}(r; \mu, \sigma^2)$ to distinguish the associated density. According to our convention, the values r of the random variable R do *not* appear in the distribution $\text{Normal}(\mu, \sigma^2)$; while, they *do* appear in the density $\text{Normal}(r; \mu, \sigma^2)$. In the latter, we separate with ";" the variable values r from the parameters μ and σ^2. We apply the same convention to the other distributions and densities as well.

As we distinguish between a random variable R and its values r, for clarity, in this chapter we also distinguish between a probability distribution $\mathbb{P}$ and its density $p(r)$. However, in subsequent chapters, we relax this convention whenever there is no ambiguity.

Distributions on random variables with probability density functions

For a random variable R whose distribution has a probability density, we can compute the probability of attaining any of the values gathered in η, where η is a collection of r values, by the integral

$$P_\eta = \int_\eta dr\, p(r). \tag{1.3}$$

In this integral, $p(r)$ is the *probability density function* of $R \sim \mathbb{P}$ and its precise form is characteristic of the distribution $\mathbb{P}$. For instance, as we have seen on examples 1.4 and 1.5, a normal distribution $\text{Normal}(\mu, \sigma^2)$ has a normal density $p(r) = \exp\left(-\frac{(r-\mu)^2}{2\sigma^2}\right)/\sqrt{2\pi\sigma^2}$ and an exponential distribution $\text{Exponential}(\lambda)$ has an exponential density $p(r) = \lambda e^{-\lambda r}$. For these two, eq. (1.3) reduces to eqs. (1.1) and (1.2), respectively.

Note 1.4: Probability density over inadmissible values

By convention, the density $p(r)$ is assumed equal to zero over inadmissible values r. For example, if the random variable models a distance, $p(r) = 0$ for $r < 0$; or, if the random variable models a temperature reported in absolute units, $p(r) = 0$ for $r < 0$.

By definition, the area, or more generally the volume, underneath an entire probability density $p(r)$ must be equal to 1. This is called the normalization condition and it means that an η that includes *every* admissible value r has probability 1. For instance, from eqs. (1.1) and (1.2) we can see that the probabilities of sampling any real scalar value is equal to 1 for either normal of exponential random variables.

As can be seen from eq. (1.3), a density $p(r)$ is *unitful* and its units are determined by normalization. Since $\int dr\, p(r) = 1$, where the region of integration includes every admissible value, the density $p(r)$ has the *reciprocal units* of r. So, if r is a length (in cm), the density $p(r)$ has units of reciprocal length (1/cm); or, if r is a time (in s), the density $p(r)$ has units of frequency (Hz).

Note 1.5: Re-sampling the same value of a random variable

Equation (1.3) already signals that the probability of sampling a continuous scalar random variable between some values $r_{\min}$ and $r_{\max}$ is

$$P_{r_{\min},r_{\max}} = \int_{r_{\min}}^{r_{\max}} dr\, p(r). \tag{1.4}$$

This brings up an interesting point: there is a vanishingly small probability for the same value of a continuous scalar random variable to be sampled twice with finite samplings. In fact, the probability of sampling *any particular value* is 0 as we can see by having coinciding $r_{\min}$ and $r_{\max}$ in the integral eq. (1.4). This indicates that, when thinking about continuous variables, we need to consider *intervals* of values rather than isolated values.

This feature generalizes to any continuous random variable that need not necessarily be scalar; but, as we will see shortly, it does not carry over the values of discrete random variables which can re-occur even in finite samplings. For example, a roll of 4 will almost certainly re-occur multiple times in a total of 1000 rolls of a fair dice.

For a random variable R, it is also possible, and often useful, to *transform* its density $p(r)$ into a density $q(v)$ over another random variable V with values that are related by a given function $v = f(r)$. For example, such a transformation occurs when we want to apply a change to our coordinate system or simply otherwise re-parametrize our model. Because we require the transformation $R \mapsto V$ to leave unaffected the probabilities we compute either using the initial or transformed variables, the two densities must satisfy

$$\int_{\eta} dr\, p(r) = \int_{f(\eta)} dv\, q(v).$$

In this equality, $f(\eta)$ is a collection that contains the transformed values $v = f(r)$ of all r in the initial collection η. In the most general setting, it is hard to relate mathematically the densities $p(r)$ and $q(v)$ any further. However, provided $f(r)$ is a *differentiable* function that can be *inverted uniquely*, as is often the case, we may apply a change of variables on the right-hand-side integral to reach $\int_{\eta} dr\, p(r) = \int_{\eta} dr\, |J_{r\mapsto v}| q(v)$, where $|J_{r\mapsto v}|$ is the determinant of the transformation's Jacobian. In turn, since such an equality holds for any η, we may drop the integrals to reach a simpler form

$$q(v) = \frac{1}{|J_{r\mapsto v}|} p(r). \tag{1.5}$$

Example 1.6: Rescaling of random variables

Any physical quantity measured in real-life experiments almost always carries units. For practical reasons, often we need to convert between quantities reported in one system of units to another. Unit conversion itself is an example of variable transformation.

For concreteness, we consider a random variable R reported in some units and suppose $v = \xi r$ where v is expressed in different units from r. Here, ξ is the conversion factor, for example ξ could be 100 cm/m for v expressed in terms of centimeters and r in terms of meters. In this example, both random variables R and V are scalar, and so the Jacobian reduces to a simple derivative. More specifically, $|J_{r\mapsto v}| = |f'(r)| = \xi$ and so, the densities are

$$q(v) = \frac{p(r)}{\xi}.$$

Example 1.7: Coordinate transformation of spatial random variables

Measurements of position are reported with respect to certain frames of reference. Changing the frame of reference is another example of a variable transformation.

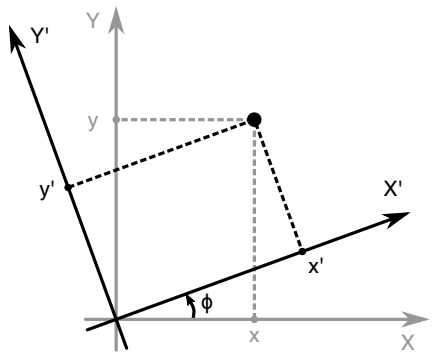

Figure 1.2: A random Cartesian position in the initial (X, Y) and the transformed (X', Y') frames of reference.

For concreteness, we consider a bivariate random variable (X, Y) that models a location in the Cartesian plane, and suppose that (X', Y') is the same location in another Cartesian frame of reference rotated through an angle ϕ about the origin, see fig. 1.2. In this case, the original and transformed positions are related through

$$x' = x\cos\phi + y\sin\phi, \qquad y' = -x\sin\phi + y\cos\phi,$$

and the Jacobian of the transformation has the form

$$J_{(x,y)\mapsto(x',y')} = \begin{bmatrix} \frac{\partial x'}{\partial x} & \frac{\partial x'}{\partial y} \\ \frac{\partial y'}{\partial x} & \frac{\partial y'}{\partial y} \end{bmatrix} = \begin{bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{bmatrix}.$$

Since $|J_{(x,y)\mapsto(x',y')}| = \cos^2\phi + \sin^2\phi = 1$, the densities in the two coordinate systems are

$$q(x', y') = p(x, y).$$

Distributions on random variables with discrete values

If $\rho_{1:M} = \{\rho_1, \rho_2, \ldots, \rho_M\}$ gathers every admissible value of a *discrete* random variable R, then its probability density has the generic form

$$p(r) = \pi_{\rho_1}\delta_{\rho_1}(r) + \cdots + \pi_{\rho_M}\delta_{\rho_M}(r) = \sum_{m=1}^{M} \pi_{\rho_m}\delta_{\rho_m}(r), \tag{1.6}$$

where π_{ρ_m} are the probabilities of the individual values ρ_m contained in $\rho_{1:M}$. The *Dirac* terms $\delta_\rho(r)$ are specified by the properties

$$\delta_\rho(r) = 0, \qquad r \neq \rho$$
$$\int dr\, \delta_\rho(r) = 1$$

where the integral is taken over every meaningful value of r; see appendix C.

Note 1.6: What is a discrete random variable?

One way to gain some intuition about discrete random variables is to consider *limiting* cases of continuous ones. For instance, we may consider acquiring measurements where ewe wish to distinguish between M distinct scalar values $\rho_{1:M}$. In a real-life experiment, our acquisitions are contaminated with noise and for this reason our measurements

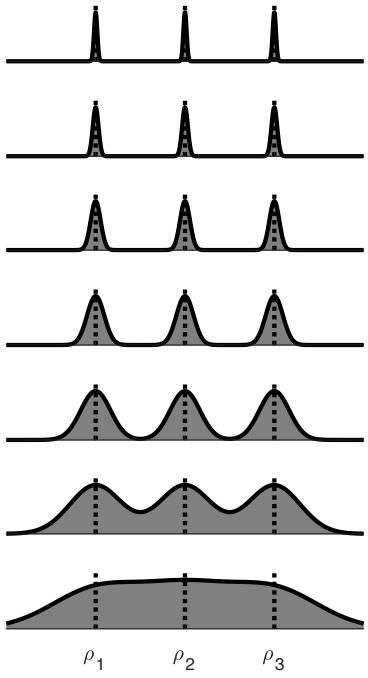

Figure 1.3: A discrete random variable arises as an idealization of increasingly less noisy measurements. Here, the bottom panel shows the probability distribution of noisy measurements. Upper panels show the probability distribution of successively clearer scenarios.

are generally scattered *around* $\rho_{1:M}$. So, we may model the measurements with a random variable $R \sim \mathbb{P}$ which, due to the noise, attains continuous values; fig. 1.3.

In a noisy scenario, the scattering of R around $\rho_{1:M}$ is wide, and our measurements are found generally anywhere around and between $\rho_{1:M}$. In this case, a fine separation between the outcomes $\rho_{1:M}$ might be impossible. However, in increasingly less noisy scenarios, our measurement distribution $\mathbb{P}$ concentrates around the outcomes giving rise to clearly isolated peaks; fig. 1.3.

At the extreme limit of an idealized noiseless scenario, the distribution $\mathbb{P}$ places all of its probability thinly around $\rho_{1:M}$ and so its density $p(r)$ becomes a train of Dirac terms as in eq. (1.6).

Normalization, in the case of a discrete random variable, reads $1 = \sum_{m=1}^{M} \pi_{\rho_m} \int dr\, \delta_{\rho_m}(r)$, where the integral over r spans any admissible and inadmissible value. Since probabilities π_{ρ_m} are dimensionless, this implies that each $\delta_{\rho_m}(r)$ on the right hand side of eq. (1.6) has dimensions of reciprocal r, similar to the density $p(r)$. As such, normalization of a discrete random variable's density can also take the equivalent form $1 = \sum_{m=1}^{M} \pi_{\rho_m}$.

One way to represent the distribution of a random variable with a density as in eq. (1.6) is

$$R \sim \text{Categorical}_{\rho_{1:M}}(\pi_{\rho_{1:M}})$$

where $\text{Categorical}_{\rho_{1:M}}(\pi_{\rho_{1:M}})$ denotes the Categorical *distribution* with outcomes $\rho_{1:M}$ and associated probabilities $\pi_{\rho_{1:M}}$. A random variable drawn from this distribution samples an outcome, ρ_m, in proportion to that outcome's probability, π_{ρ_m}; see fig. 1.4.

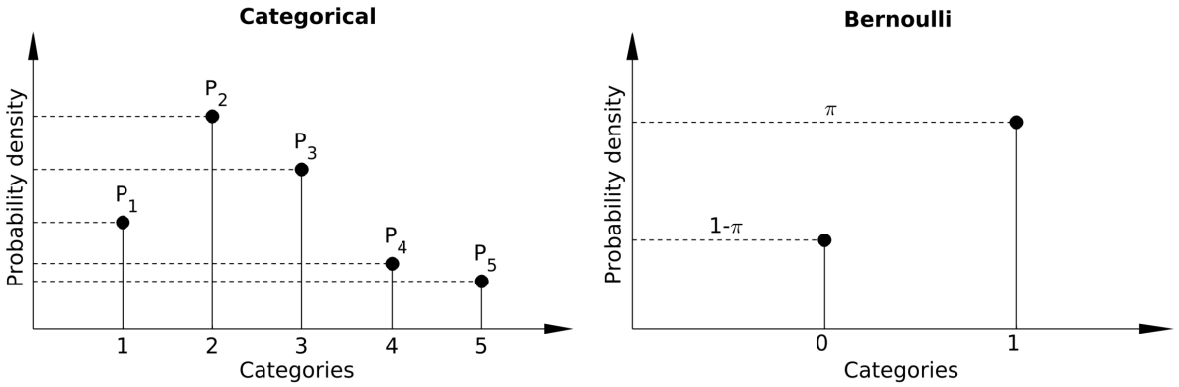

Figure 1.4: On the left hand side we plot the associated probabilities π and $1-\pi$ of the Bernoulli distribution located at its two possible outcome locations (0 and 1). On the right hand side we plot the associated probabilities $\pi_{\rho_{1:5}}$ with 5 outcomes, $\rho_{1:5}$, of the Categorical distribution.

Example 1.8: Dice rolls modeled as Categorical random variables

Rolling a common dice leads to one out of six outcomes that we idealize as the faces marked with the numbers "1" through "6". Provided we identify these outcomes with the categories ρ_m, for $m = 1, \ldots, 6$, we can model a dice roll as a Categoricalrandom variable

$$R \sim \text{Categorical}_{\rho_{1:6}}(\pi_{\rho_{1:6}})$$

where the probability a face marked with "m", or category ρ_m, is π_{ρ_m}. As we know, fair dice have equiprobable faces, $\pi_{\rho_1} = \cdots = \pi_{\rho_6} = 1/6$; but loaded dice do not follow these probabilities.

The simplest example of a Categorical distribution is the Bernoulli distribution which is the special case having just two outcomes that conventionally we identify with the numbers 1 and 0, and respective probabilities $\pi_1 = \pi$ and $\pi_2 = 1 - \pi$; see fig. 1.4. We often write $\text{Bernoulli}(\pi)$ instead of the more elaborate $\text{Categorical}_{1,0}(\pi, 1-\pi)$.

Example 1.9: Coin flips modeled as Bernoullirandom variables

An ideal coin flip has only two outcomes: "heads" or "tails". Provided we identify these with the numbers 1 and 0, respectively, we can model a coin flip as a Bernoulli random variable

$$R \sim \text{Bernoulli}(\pi) \tag{1.7}$$

where π is the probability of "heads". Here, specifying the probability of tails, $1-\pi$, is redundant since, by normalization, it is uniquely determined by π.

If, instead, we want to avoid identifying "heads" and "tails" with 1 and 0, we can also model a coin flip as a Categorical random variable

$$R \sim \text{Categorical}_{\text{heads,tails}}(\pi, 1-\pi). \tag{1.8}$$

Essentially, the only difference between eq. (1.7) and eq. (1.8) is in the meaning we assign to the values r, with the latter representation here having an *interpretational advantage* over the former.

Distributions on random variables *without* probability density functions†

Since in later chapters we formulate models with random variables to which we *cannot* assign a probability density function, for example random variables that are functions or random variables that are probability distributions

†This is an advanced topic and could be skipped on a first reading.

themselves, we also need to account for appropriate distributions on those. In such cases recipes to compute probabilities are case specific and, in general, the description of the associated distributions is considerably more complicated. In examples 1.10 and 1.11 we provide only a sneak preview.

Example 1.10: The standard Brownian motion

We will examine Brownian motion in more detail in chapter 2. As we will see, standard Brownian motions in 1D are random variables that represent functions from a time interval spanning 0 to some positive T to the real line. To denote them we write

$$X \sim \text{BMotion}_T^{\text{1D}}\left(D\right)$$

where the parameter D in the Brownian motion is a positive real scalar and, as we will see, can be interpreted as a diffusion coefficient of a particle diffusing in 1D.

A shorthand like this captures the following pieces of information:

- The realizations of X are functions $x(\cdot)$ that, to any time t between 0 and T, assign $x(t)$ which is a position on the real line.
- Any realization of X, is initialized at the origin, *i.e.* $x(0) = 0$.
- For any choice of times t and t' between 0 and T, the difference $x(t) - x(t')$ between the values $x(t)$ and $x(t')$ of any realization $x(\cdot)$ is a random variable itself.
- The random variable $x(t) - x(t')$ has a probability density given by

$$p\left(x(t) - x(t')\right) = \frac{1}{\sqrt{4\pi D|t-t'|}} \exp\left(-\frac{\left(x(t) - x(t')\right)^2}{4D|t-t'|}\right).$$

Example 1.11: The Gaussian process

We will examine *Gaussian processes* in more detail in **??**. As we will see, Gaussian processes are random variables that represent functions from a space S to the real numbers. To denote them we will write

$$F \sim \text{GaussianP}_S\left(h(\cdot), c(\cdot,\cdot)\right).$$

A shorthand like this captures the following pieces of information:

- The realizations of F are functions $f(\cdot)$ that, to any point x in S, assign $f(x)$ which is a real number.
- The parameter $h(\cdot)$ is a function that, to every point x in S, assigns $h(x)$ which is also a real number.
- The parameter $c(\cdot,\cdot)$ is a function that, to every points x and x' in S, assigns $c(x,x')$ which is a non-negative real number.
- For any choice $x_1, \ldots, x_M$ of any finite number M of points in S, the values $[f(x_1), \cdots, f(x_M)]$ form a random array.
- The random array $[f(x_1), \ldots, f(x_M)]$ has a probability density

$$p\left([f(x_1), \cdots, f(x_M)]\right) = \text{Normal}_M\left(\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_M) \end{bmatrix}; \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_M) \end{bmatrix}, \begin{bmatrix} c(x_1,x_1) & \cdots & c(x_1,x_M) \\ \vdots & \ddots & \vdots \\ c(x_M,x_1) & \cdots & c(x_M,x_M) \end{bmatrix}\right).$$

1.2.2 How to simulate probability distributions

So far we have discussed random variables and probability distributions from which random variables are drawn. What we discuss next is how to run *simulations*. That is, how to generate values or sample random variables in a computer using their probability distributions. Random simulations are useful when when we seek to re-create *in silico* repetitions of our prototype experiment. In subsequent chapters, we will see that we can use random sampling not only to re-create an experiment's results but also to draw inferences *from* an experiment's results.

Continuous random variables

For a random variable $R \sim \mathbb{P}$ that takes scalar real values r, its *probability cumulative function* is a function $C(r)$ given by

$$C(r) = \int_{-\infty}^{r} dr'\, p(r') \tag{1.9}$$

where $p(r)$ is the probability density associated with $\mathbb{P}$. From this definition, we see that a cumulative function is dimensionless and increases monotonically between 0 to 1. This is a characteristic that we can use to develop a method from which to generate random values r of R on a computer.

For instance, starting with the density $p(r)$, we first calculate its cumulative function $C(r)$. We then generate a random value, call it u, uniformly between 0 and 1, and ask: *for what value r is the cumulative function equal to u?* In other words, we find $r = C^{-1}(u)$, where $C^{-1}(u)$ is the inverse function of $C(r)$. This method, often termed the *fundamental theorem of simulation*, is summarized in algorithm 7.3 and is visually illustrated in fig. 1.5.

Algorithm 1.1: Fundamental theorem of simulation for continuous variables

To simulate a continuous random variable $R \sim \mathbb{P}$

- First, find the cumulative function $C(r)$ and its inverse $C^{-1}(u)$.
- Then, repeat the following steps
  - Generate $u \sim \text{Uniform}_{[0,1]}$.
  - Set $r = C^{-1}(u)$.

Upon completion, this algorithm generates values r according to $\mathbb{P}$.

Why the algorithm works? Here, we have a uniform random variable u whose density is $g(u) = 1$ for all values between 0 and 1. When we set $r = C^{-1}(u)$, effectively we perform a transformation of random variables. As we saw earlier, the density $h(r)$ of the transformed variable is given by eq. (1.5) which, in this setting, takes the form

$$h(r) = \frac{1}{|J_{u \mapsto r}|} g(u).$$

Here, because both our variables are scalar, the Jacobian of the transformation is given by the derivative

$$J_{u \mapsto r} = \frac{d}{du} C^{-1}(u) = \frac{1}{\frac{d}{dr} C(r)} = \frac{1}{p(r)}.$$

Considered together, the last two equalities lead to $h(r) = p(r)$. In other words, the values r generated in algorithm 7.3, indeed, follow the desired density $p(r)$.

Example 1.12: Simulating from an exponential distribution

We consider a random variable $R \sim \text{Exponential}(\lambda)$. As we saw in example 1.5, this random variable takes real scalar values and its density is

$$p(r) = \begin{cases} \lambda e^{-\lambda r}, & r \geq 0 \\ 0, & r < 0 \end{cases}.$$

To apply the fundamental theorem of simulation, we first compute the cumulative function and its inverse. These are

$$C(r) = 1 - e^{-\lambda r}, \qquad C^{-1}(u) = -\frac{1}{\lambda} \log(1 - u).$$

By means of algorithm 7.3, we then sample R as follows:

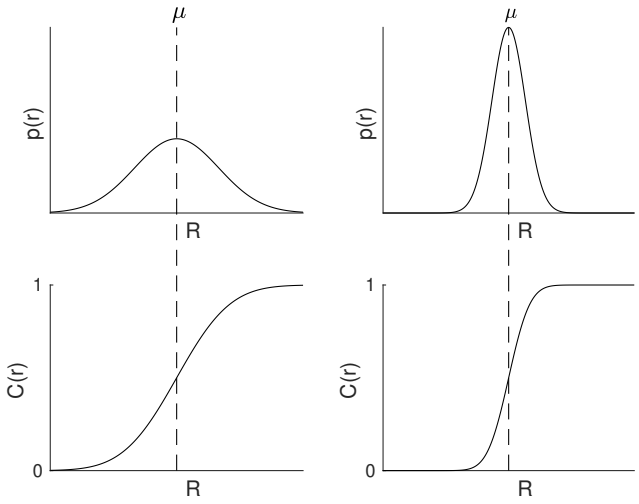

Figure 1.5: On the top rows we have PDFs broadly and tightly centered around their mean μ. In the bottom row, we have the coinciding CDF. The CDF has a sharp slope near $r = \mu$ for the PDF more tightly centered on its mean. Thus, for this special case in applying algorithm 7.3, most values of u would coincide with values of r near μ. By contrast, if the PDF were broader near μ, the slope of the CDF would be smaller and values of u would coincide with a broader range of values of r.

- First, we generate u from a uniform distribution between 0 and 1

$$u \sim \text{Uniform}_{[0,1]}.$$

- Then, we compute r from

$$r = -\frac{1}{\lambda}\log(1-u). \tag{1.10}$$

Since $v = 1 - u$ is also uniformly distributed between 0 and 1, for computational efficiency, when sampling exponential random variables, we generate $v \sim \text{Uniform}_{[0,1]}$ in the first place and then use $r = -\frac{1}{\lambda}\log v$ instead of eq. (1.10). In this way, we speed up the execution of the algorithm by avoiding the computation of the difference $1 - u$.

Note 1.7: Distribution functions

So far, we have encountered three important functions $p(r), C(r), C^{-1}(u)$ associated with a random variable $R \sim \mathbb{P}$. These are very common in the literature and below we summarize some terms that are used to designate them.

- $p(r)$ is occasionally termed *probability density function* or *PDF*.
- $C(r)$ is occasionally termed *probability cumulative function, cumulative distribution function* or *CDF* .
- $C^{-1}(u)$ is occasionally termed *probability quantile function, inverse cumulative distribution function* or *ICDF* .

Because all these functions are tightly associated with the distribution $\mathbb{P}$, often we refer to them simply as *distribution functions*.

Discrete random variables

We can use a similar procedure to sample discrete random variables too. In particular, for $R \sim \text{Categorical}_{\rho_{1:M}}(\pi_{\rho_{1:M}})$, the cumulative function is

$$C(\rho_m) = \sum_{m'=1}^{m} \pi_{\rho_{m'}}$$

or, more concretely, it has the form

$$C(\rho_1) = \pi_{\rho_1}, \qquad C(\rho_2) = \pi_{\rho_1} + \pi_{\rho_2}, \qquad \cdots \qquad C(\rho_M) = \pi_{\rho_1} + \pi_{\rho_1} + \cdots + \pi_{\rho_M}.$$

To sample an outcome r, as with continuous random variables, we also need to generate $u \sim \text{Uniform}_{[0,1]}$. However, now a problem concerning the inversion of $C(r)$ arises. Namely, there may be no r such that $C(r) = u$. For this reason, instead of searching for outcomes such that $C(r) = u$, we search for the *lowest* value r such that $u \leq C(r)$. This version of the fundamental theorem of simulation is summarized in algorithm 1.3.

Algorithm 1.2: Fundamental theorem of simulation for discrete variables

To simulate a random variable $R \sim \text{Categorical}_{\rho_{1:M}}(\pi_{\rho_{1:M}})$

- Generate $u \sim \text{Uniform}_{[0,1]}$.
- Find the lowest m such that $u \leq \pi_{\rho_1} + \pi_{\rho_2} + \cdots + \pi_{\rho_m}$.
- Set $r = \rho_m$.

At first, it might appear that this algorithm depends on the particular labeling of $\rho_{1:M}$ and that it would lead to different realizations r if the labels m had been assigned differently over the categories ρ_m. However, since relabeling of $\rho_{1:M}$ involves also a similar relabeling of $\pi_{\rho_{1:M}}$, this is *not* the case. In other words, this algorithm realizes each outcome $r = \rho_m$ with probability π_{ρ_m}, even when the labels m are reassigned over ρ_m.

Why does this algorithm work? We consider the limiting scenario of note 1.6. Namely, we have a continuous random variable $R \sim \mathbb{P}$ in an experiment that aims to distinguish between the outcomes $\rho_{1:M}$. Generally, in a noisy experiment, $\mathbb{P}$ has a wide PDF and a CDF that increases smoothly from 0 to 1. In a less noisy experiment, however, the PDF has peaks and this means that the CDF retains smooth albeit clearly visible steps. As the noise level reduces, the PDF's peaks become more prominent resulting in the CDF's steps becoming sharper. In the extreme case of a noiseless scenario, the CDF forms perfectly sharp steps mathematically represented by discontinuities; fig. 1.6. Seen as a limiting case of sampling continuous random variables, algorithm 1.3 is the direct analog of algorithm 7.3.

Example 1.13: Simulation of Bernoulli random variables

Consider $R \sim \text{Bernoulli}(\pi)$. In this case, the cumulative function has a very simple form

$$C(1) = \pi, \qquad C(0) = 1.$$

To sample r, according to the fundamental theorem of simulation

- first, we generate $u \sim \text{Uniform}_{[0,1]}$.
- then, if $u \leq \pi$ we set $r = 1$, else we set $r = 0$.

The two steps are illustrated in fig. 1.7.

Steve to Ioannis: you alternate between "cumulative function" and CDF. Can we stick to one term or is there a subtle reason I am missing?

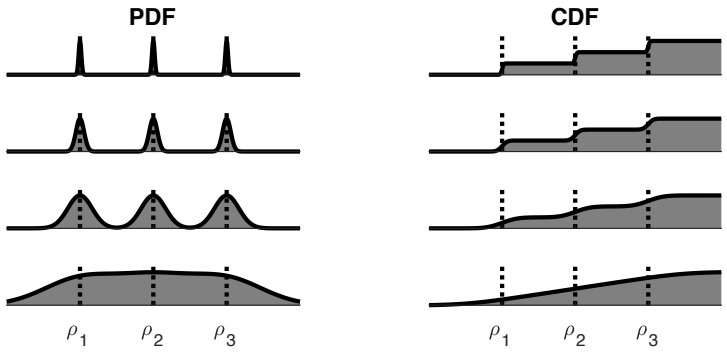

Figure 1.6: The CDF of a discrete random variable arises as an idealization of increasingly less noisy measurements. Here, the left panels show the probability distributions of noisy measurements and the right panels illustrate the corresponding CDFs.

Note 1.8: How the fundamental theorems of simulation work

There exists an intuitive explanation for the fundamental theorem of simulation that we illustrate in **??**. Steve to Ioannis: can you reproduce fig. fig. 1.6 but with the appropriate horizontal and vertical lines to illustrate where u falls and what R it coincides with?

That is, we imagine a stick of unit length with a break point just as shown in fig. 1.7. The portion of the stick before the break point has length π. The remainder of the stick has length $1 - \pi$. We now sample a uniform random variable, u. If u falls before the break point, outcome 0 is realized. Otherwise outcome 1 is realized.

A similar logic holds for the Categorical distribution. **??** shows the discrete steps in the cumulative function of a discrete distribution. A draw from the uniform distribution can be visualized as the dotted horizontal line of **??**. The value of the abscissa that coincides with the location where the dotted line intersects with the CDF dictates the value realized by the discrete random variable.

The next logical leap we need to take is to think of a continuous distribution as the limit of a Categorical with very many closely packed realizations of the random variable. The reasoning underlying how we go about sampling from a continuous distribution then follows from the argument put forward for sampling from a Categorical. This is shown in **??**.

1.2.3 How to combine probability distributions

In a model of a prototype experiment, we most often have random variables with different properties. When handling such a complex model, we need to work simultaneously with more than one distribution. Here, we present bookkeeping rules that help us combine and manipulate multiple distributions.

Joint and marginal distributions

Provided the random variables in our model are independent from each other, for example because they may encode physical processes or observations that exert no influence upon each other, we may write

$$R_1 \sim \mathbb{P}_1, \qquad R_2 \sim \mathbb{P}_2, \qquad \cdots \qquad R_N \sim \mathbb{P}_N.$$

Each distribution $\mathbb{P}_1, \mathbb{P}_2, \ldots, \mathbb{P}_N$ is, in turn, associated with its own density $p_1(r_1), p_2(r_2), \ldots, p_N(r_N)$. As there is little chance of confusion, commonly we simply write $p(r_n)$ instead of $p_n(r_n)$.

Occasionally, we also encounter models with multiple random variables

$$R_1 \sim \mathbb{P}, \qquad R_2 \sim \mathbb{P}, \qquad \cdots \qquad R_N \sim \mathbb{P}, \tag{1.11}$$

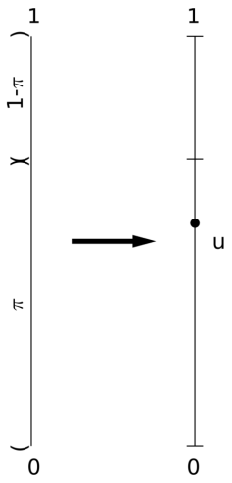

Figure 1.7: Here we consider a probability, ranging from 0 to 1, segmented into two portions of weight π and $1-\pi$ separated by a break point. We imagine these to be the probability of sampling outcome 0 or outcome 1 in a Bernoulli trial. To determine which outcome we select, we draw a uniform random number u. In this figure, the u sampled fell below the break point. As such, we select outcome 0 for this Bernoulli trial.

that are independent and which also follow identical distributions $\mathbb{P}$, for example random variables that may model independent observations obtained from a time invariant system. On such occasions, we might abbreviate eq. (1.11) into $R_1, R_2, \ldots, R_N \overset{iid}{\sim} \mathbb{P}$ and speak of *independent and identically distributed*, or simply *iid*, random variables. Essentially, we mean that all densities $p(r_1), p(r_2), \ldots, p(r_N)$ happen to have the same form. In the iid setting, and only when there is no chance of confusion, we might refer to each one of the densities simply as $p(r)$. However, as we will see shortly, even with iid variables, most of the times we run into complicated settings and what we mean by $p(r)$ needs explicit clarification.

Following our convention, we denote the density of a single variable R_n as $p(r_n)$ and call it a *marginal density*. When multiple random variables $R_{1:N}$ arise in the same setting and a density gathers all of them, we write $p(r_{1:N}) = p(r_1, r_2, \ldots, r_N)$. We refer to $p(r_{1:N})$ as a *joint density*.

A marginal density $p(r_n)$ is related to a joint density $p(r_{1:N})$ through an integration over the entire range spanned by $r_{1:n-1}$ and $r_{n+1:N}$. That is,

$$p(r_n) = \underbrace{\int dr_1 \cdots \int dr_{n-1} \int dr_{n+1} \cdots \int dr_N}_{\text{everything but } r_n} \; p(r_{1:N}). \tag{1.12}$$

Colloquially, we refer to the integration over variables, *i.e.* from-right-to-left in eq. (1.12), as a "marginalization". We refer to the reverse process, *i.e.* from-left-to-right in eq. (1.12), as a "de-marginalization" or a "completion".

Example 1.14: Marginalization

Similarly, we may obtain distributions over any subset of the variables in $R_{1:N}$. For concreteness, we consider a total of $N = 5$ variables and suppose that we wish to obtain a distribution over R_2 and R_4 only. In this case, marginal and joint densities are linked by

$$p(r_2, r_4) = \underbrace{\int dr_1 \int dr_3 \int dr_5}_{\text{everything but } r_2 \text{ and } r_4} \; p(r_{1:5}).$$

Note 1.9: Box-Muller simulation of normal random variables

We are now ready to discuss the simulation of a random variable $X \sim \text{Normal}(\mu, \sigma^2)$. As the cumulative function of X does not have a closed form and we cannot use the fundamental theorem of simulation, we follow a different approach that relies on joint distributions.

We start by considering two iid random variables, X, Y. In particular,

$$X, Y \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2).$$

The associated joint density reads

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

On X and Y, as we illustrate on fig. 1.8, we perform three successive transformations:

- a linear transformation from x and y to

$$x' = \frac{x-\mu}{\sigma}, \qquad y' = \frac{y-\mu}{\sigma}$$

- a non-linear transformation from Cartesian (x', y') to polar coordinates (ρ, ϕ) with

$$x' = \rho \cos\phi, \qquad y' = \rho \sin\phi$$

- a non-linear transformation from ρ to λ with

$$\lambda = \rho^2.$$

The advantage of applying these transformations is that the resulting density over λ and ϕ is separable

$$p(\lambda, \phi) = \frac{1}{2} \exp\left(-\frac{\lambda}{2}\right) \frac{1}{2\pi} = \text{Exponential}\left(\lambda; \frac{1}{2}\right) \text{Uniform}_{[0,2\pi]}(\phi).$$

The cumulative function over λ and ϕ can now be computed analytically. So, by generating two uniform random samples $u_1, u_2 \overset{iid}{\sim} \text{Uniform}_{[0,1]}$, we can readily obtain random samples from the radial and polar angle distribution

$$\rho = \sqrt{\lambda} = \sqrt{-2 \log u_1}, \qquad \phi = 2\pi u_2.$$

Transforming back to our original variables, we arrive at

$$x = \mu + \sigma x' = \mu + \sigma\rho\cos\phi = \mu + \sigma\sqrt{-2\log u_1}\cos(2\pi u_2).$$

This algorithm for sampling normal random variables is termed the *Box-Muller algorithm*. As can be seen, with little additional computational cost, this method also provides another normal sample

$$y = \mu + \sigma\sqrt{-2\log u_1}\sin(2\pi u_2)$$

which is independent of x.

Conditional distributions

The order in which random variables arise in a model may be irrelevant, for example random variables modeling an experiment's observations that exert no influence upon each other, such as individual test scores, biometric measurements collected from a group of unrelated participants, or observations generated by a system at equilibrium. On the other hand, the order in which random variables arise *may be important*, for example random variables modeling observations of time-dependent phenomena, such as successive measurements of the number of cells in a growing cell culture or the number of molecules available to react in a chain of chemical reactions.

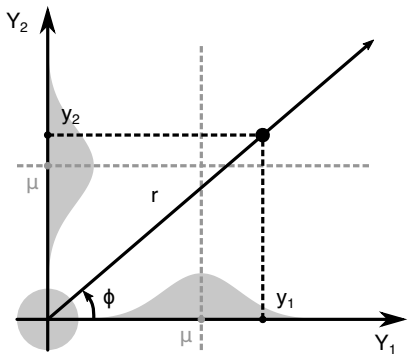

Figure 1.8: Sampling independent random variables $X, Y \sim$ Normal(μ, σ^2) according to the Box-Muller method of note 1.9 by, equivalently, sampling an r and ϕ in polar coordinates.

To express *dependencies* among two random variables, R_1 and R_2, we write

$$R_2|r_1 \sim \mathbb{P}(r_1). \tag{1.13}$$

This reads "the random variable R_2, given the realization r_1 of the random variable R_1, is sampled from the probability distribution $\mathbb{P}(r_1)$" and means that the values of r_2 are associated with a density $p(r_2|r_1)$ that depends upon r_1. We designate a distribution that depends upon the value of another random variable like $\mathbb{P}(r_1)$ and the associated density $p(r_2|r_1)$ as *conditionals*.

Note 1.10: How to avoid inaccuracies in specifying variable dependencies

In the setting of eq. (1.13), the random variable R_1 is sampled from its own (marginal) distribution that needs to be specified *separately*. In a complete model, both random variables $R_1 \sim \mathbb{P}_1$ and $R_2|r_1 \sim \mathbb{P}_2(r_1)$ need to be specified adequately.

In a properly formulated model, the distribution of R_1 *must not* depend upon r_2 and, for this reason, the description of R_1 should precede that of $R_2|r_1$. If this is not possible, then we need to describe the random variables together through a single (joint) distribution $(R_1, R_2) \sim \mathbb{P}$.

Ideally, proper descriptions of models involving multiple random variables, that depend upon each other, should be given in a nested fashion. For example

$$\begin{aligned} R_1 &\sim \mathbb{P}_1 \\ R_2|r_1 &\sim \mathbb{P}_2(r_1) \\ R_3|r_2, r_1 &\sim \mathbb{P}_3(r_2, r_1) \\ &\text{etc...} \end{aligned}$$

A necessary condition (although not always sufficient) for a reliable description of a probabilistic model, no matter how convincing the involved arguments may be and no matter how intuitive the involved distributions may appear, is that *every single distribution* $\mathbb{P}_1, \mathbb{P}_2(r_1), \mathbb{P}_3(r_2, r_1), \ldots$ be specified *clearly and explicitly*.

Whenever a supposedly flawless model cannot be put in a nested form as above, even when random variables are grouped and joint distributions are applied, the model most likely contains flaws such as tautologies or contradictions. Consequently, such a model is inappropriate for quantitative or even qualitative use.

In note 1.10, we consider nested variable dependencies. In particular, R_3 depends on the realizations r_2 and r_1, in turn, R_2 depends on the realization r_1, and finally R_1 depends on no other realization. In the most general case, the probability distribution over our last random variable, say R_N, may depend on the realization of all

previous random variables, $r_{1:N-1}$, and the same happens for all other variables up to the very first one r_1. Because of this hierarchy, which resembles successive generations of variables, to simulate a nested model we may use a sampling algorithm termed *ancestral sampling* as detailed below.

Algorithm 1.3: Ancestral sampling

To draw values for a group of random variables $R_{1:N}$, we proceed as follows:

- Find the density $p(r_1)$ associated with $R_1 \sim \mathbb{P}_1$.
- Sample r_1 using $p(r_1)$.
- for n from 2 up to N, repeat:
  - Find the density $p(r_n|r_{1:n-1})$ associated with $R_n|r_{1:n-1} \sim \mathbb{P}_n(r_{1:n-1})$.
  - Sample r_n using $p(r_n|r_{1:n-1})$.

Because we use ancestral sampling and hierarchal models extensively in the following chapter, we describe here methods to obtain the necessary conditional densities. Our starting point is the full joint density $p(r_{1:N})$.

Conditional and joint densities are related to each other through the *chain rule* which, in the most general setting, reads

$$p(r_{N:1}) = p(r_N|r_{N-1:1}) \cdots p(r_2|r_1)p(r_1).$$

In the simplest case, consisting of only two random variables, the chain rule reads

$$p(r_2, r_1) = p(r_2|r_1)p(r_1).$$

From this we immediately see that a conditional density over r_2 is normalized irrespective of r_1, *i.e.*

$$\int dr_2\, p(r_2|r_1) = \int dr_2\, \frac{p(r_2, r_1)}{p(r_1)} = \frac{\int dr_2\, p(r_2, r_1)}{p(r_1)} = \frac{p(r_1)}{p(r_1)} = 1.$$

Additionally, from the chain rule, we obtain two equalities, $p(r_2, r_1) = p(r_2|r_1)p(r_1)$ and $p(r_1, r_2) = p(r_1|r_2)p(r_2)$, that we can combine to obtain another important rule, namely *Bayes' rule*, which most often is written in the form

$$p(r_2|r_1) = \frac{p(r_1|r_2)p(r_2)}{p(r_1)}, \qquad p(r_1) \neq 0. \tag{1.14}$$

As we will see in subsequent chapters, eq. (1.14) is an indispensable tool in Data Analysis.

Example 1.15: Modeling dynamical systems

Ioannis: do you want these to be W's or, more intuitively, R's? Steve: This is more delicate than it seems. Let's keep W, otherwise we need to say something about state variables. I think W entails less overhead definitions. Dependency among variables is especially important when the physical system of interest evolves over time. In this dynamical setting, which we explore in the next chapter, our prototype experiment is temporally structured: causality indicates that the last measurement W_N may be influenced by all preceding measured values, $w_{1:N-1}$; the penultimate measurement, W_{N-1}, may be influenced by all of its preceding ones $w_{1:N-2}$; and so forth.

With the rules of joint and conditional distributions, we can work out the densities of such models in the most general setting. For instance,

$$p(w_{1:N}) = p(w_N|w_{1:N-1})p(w_{N-1}|w_{1:N-2}) \cdots p(w_2|w_1)p(w_1).$$

It follows that if we need to sample realizations of $W_{1:N}$, we need 1 marginal and $N-1$ *different* conditional distributions. As a result, this sampling may become infeasible unless we make some assumptions.

- One drastic assumption, often too crude for realistic dynamical systems, is to assume that all variables are independent, which in this particular case is equivalent to assuming $p(w_n|w_{1:n-1}) = p(w_n)$. Under this

assumption, the joint density factorizes into a product of densities

$$p(w_{1:N}) = p(w_1)p(w_2)\cdots p(w_N) = \prod_{n=1}^{N} p(w_n). \tag{1.15}$$

- Another, less drastic, and often realistic, assumption is to consider $p(w_n|w_{1:n-1}) = p(w_n|w_{n-1})$. Under this assumption, the joint density also factorizes into a product of densities

$$p(w_{1:N}) = p(w_1)p(w_2|w_1)\cdots p(w_N|w_{N-1}) = p(w_1)\prod_{n=2}^{N} p(w_n|w_{n-1}). \tag{1.16}$$

Under these two assumptions, the total number of different probability distributions, that are needed to sample $W_{1:N}$, reduces from N to 1 and 2, respectively.

Somewhat pedantically, in deriving eq. (1.15), we invoked a so-called *0th order Markov assumption*

$$p(w_n|w_{1:n-1}) = p(w_n),$$

while, in deriving eq. (1.16) we invoked a so-called *1st order Markov assumption*

$$p(w_n|w_{1:n-1}) = p(w_n|w_{n-1}).$$

In principle, we can also invoke higher order assumptions where a measurement w_n is influenced by more than 1 past measurement; however, as we will see in the subsequent chapters, such assumptions are rarely used in practice, either because a 1st order assumption is already sufficient or because they lead to models with prohibitive computational cost.

1.3 Data-driven modeling and inference

Our emphasis from now on is not as much on mathematical rigor as it is focused on problem-formulation and problem-solving. But, *of what problem exactly?* With our basic notions laid down, we are now ready to define our problem concretely.

In the data-centric context that is most appropriate for the Physical and Natural Sciences, we envision being provided information on a physical system such as:

- *how this system behaves* under relevant, well or poorly characterized, conditions;
- *how observations are acquired* on this system;
- *specific values* of acquired observations.

These are the *data* and they serve as our input or starting point. Our primary task is to analyze the data and we tackle Data Analysis with the framework introduced in section 1.1.2. More specifically, within the framework set by the prototype experiment, which we need to specialize to particular real-life scenarios, our goal is to use the acquired values of the observations to infer a model. However, before we can infer a model, we first need to go through a *synthesis stage* in order to develop the necessary mathematical formulation that is meaningful for the system at hand.

During the synthesis stage, we utilize the available information on our system to formulate the probability distribution $p(w_{1:N}|\theta_{1:K})$ that best describes our experiment. For example, in this stage we consider physical laws, intrinsic dynamics, and noise properties, which, although non-numeric, in a very concrete sense are part of our given data. At this stage, we also decide on parameters $\theta_{1:K}$ and commonly assign physical meaning to all or some of them.

The synthesis stage concludes with the establishment of a *generative model*; that is, a quantitive description of *how our experiment's measurements are generated*, see fig. 1.9. Our generative model links mathematically our unknowns $\theta_{1:K}$ with our knowns $w_{1:N}$ and, in principle, could be simulated in a computer.

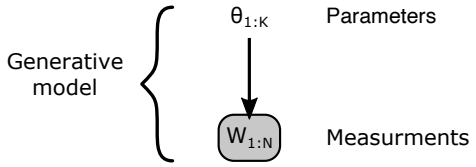

Figure 1.9: Ioannis: typo in your figure: measurement A generative model describes how measurements are generated. Implicitly, it encodes any influence the parameters $\theta_{1:K}$ exert upon the measurements $W_{1:N}$.

Note 1.11: A model's likelihood

The probability distribution $p(w_{1:N}|\theta_{1:K})$, which is mathematically established in a generative model, is a key quantity. This distribution is termed the *likelihood*. The term follows from the notion that $p(w_{1:N}|\theta_{1:K})$ quantifies the likelihood of observing (sampling) the sequence of observations $w_{1:N}$ in our prototype experiment which is influenced by the parameters $\theta_{1:K}$.

During the analysis stage, once we have formulated $p(w_{1:N}|\theta_{1:K})$ adequately, we apply the measured values of $w_{1:N}$ to compute estimates of the parameters $\theta_{1:K}$. Traditionally, we call these values *estimators* and denote them with $\hat{\theta}_{1:K}$. We will see in chapter 3 that a likelihood provides us with a *universal* strategy to obtain $\hat{\theta}_{1:K}$ needed to specify uniquely a model we wish to learn. The challenge, however, is that we are also often interested in error bars around $\hat{\theta}_{1:K}$ or, put differently, whole probability distributions over $\theta_{1:K}$. For this reason, in chapter 4, we will consider an extended strategy that uses more than an experiment's likelihood.

The first stage in our workflow takes more of a *modeling perspective;* while, the second stage takes more of a *computational perspective.* Nonetheless, as we discuss in example 1.16, both stages in the solution of our problem are important and both stages pose unique challenges. As we will see in the subsequent chapters, often we have to devise comprehensive approaches that deal with the challenges arising in both stages simultaneously.

Example 1.16: Likelihood based modeling and inference

As a concrete example, we imagine an experiment idealized as having one of two (discrete) measurement outcomes, for example the emission or not of a photo-electron as described in example 1.2. For simplicity, we may denote these outcomes with $\rho_1 = 1$ and $\rho_2 = 0$, respectively.

If we idealize individual assessments as iid, meaning that each measurement is independent of the others as in eq. (1.15), then the mathematical form of the likelihood is readily derived. In particular, the model responsible for generating the data takes the form

$$W_n|\pi \sim \text{Bernoulli}(\pi), \qquad n = 1, \ldots, N,$$

and, as of yet, has one unspecified parameter, namely π, which is the probability that a single assessment measures a photo-electron. Our goal is therefore to estimate π.

Now we ask: *Given this generative model, what is the likelihood of our measurements?* This likelihood is the probability of observing the sequence $w_{1:N}$ and we may compute it as following

$$p(w_{1:N}|\pi) = \prod_{n=1}^{N} p(w_n|\pi) = \prod_{n=1}^{N} \text{Bernoulli}(w_n;\pi) = \prod_{n=1}^{N} \pi^{w_n}(1-\pi)^{1-w_n} = \pi^M(1-\pi)^{N-M}$$

where we assumed that, within $w_{1:N}$, the first outcome, ρ_1, has been observed in total M times and the second outcome, ρ_2, has been observed the remainder of the times, $N - M$.

We can estimate a value for the parameter π by asking: *Which value of π makes our observations most likely?* This is equivalent to asking which value of π makes $p(w_{1:N}|\pi)$ highest. Essentially, we need to seek for the maximizer of $p(w_{1:N}|\pi)$. For instance, solving $\frac{d}{d\pi}p(w_{1:N}|\pi) = 0$, we find $\hat{\pi} = M/N$, as expected intuitively.

For this example, we assumed that both outcomes, ρ_1 and ρ_2, are observed at least once and so $0 < \hat{\pi} < 1$, because $0 < M < N$. Yet had $M = 0$ or $M = N$, we may had erroneously concluded, due to limited data, that $\hat{\pi} = 0$ or $\hat{\pi} = 1$. In turn, if we had used this estimate to predict the outcome of future experiments, we would had erroneously concluded that this would *conclusively* be either ρ_2 or ρ_1, respectively. Thus, even this toy example forebodes our need to go beyond approaches that rely exclusively on likelihoods.

In the Physical Sciences, data-driven approaches are sometimes termed *inverse methods*, *inverse problems*, or *inverse modeling*. Yet, as example 1.16 illustrates, there is nothing backward about obtaining models starting from the data and these, somewhat unfortunate terms, arose only because traditional approaches–that is, obtaining models from the ground-up with a combination of first principles and data-fitting–came historically first and are now termed forward (or direct).

Note 1.12: Inverse modeling

Data-driven model inference essentially is an inverse problem. Solving an inverse problem is the opposite of solving a direct problem. Briefly, in a *direct problem* we seek to determine an effect knowing its cause; while, in an *inverse problem* we seek to recover the cause knowing only the effect.

Inverse problems arise mainly when we need to interpret indirect physical measurements of unknown or partially known origin. For instance, when we are interested in elucidating the dynamics of complex biomolecules observed indirectly through fluorescence microscopy. In an experiment we acquire images (measurements) with all sorts of artifacts that subsequently need to be removed in order to reveal the positions or dynamics of the biomolecules we are interested in. By contrast, simulating possible measurements invoking a physical model, established or tentative, that *predicts* the positions or dynamics sought after and subsequently checking whether they are in agreement or disagreement with the observed measurements is forward modeling and essentially involves solving only direct problems; see fig. 1.10.

A problem, whether direct or inverse, is *well-posed* when it meets the following conditions

- the problem has a solution;
- the solution is unique;
- the solution does not differ substantially unless the supplied data differ substantially too.

These conditions are mathematically known as *existence, uniqueness,* and *stability*, respectively. If a problem fails to satisfy one or more of them, it is *ill-posed*.

Direct problems are well-posed when the effects (data) we are after are well-defined, single-valued, and depend continuously on their causes. Often this is the case when we seek to reproduce observations mathematically or computationally. On the other hand, solutions to inverse problems do *not always exist*, or when they exist they *may not be unique* or they may *change dramatically* even when the supplied data (effects) differ only insignificantly. As a result, inverse problems are commonly ill-posed and solving them can be very challenging.

Throughout the subsequent chapters, with the use of the appropriate random variables and probability distributions, we will see how inverse problems can be formulated as statistical problems and how solutions to these problems can be computed robustly and efficiently.

Forward modeling has had its role to play and is heavily showcased throughout Physics where disparate observations were unified into predictive frameworks inspired by logic, symmetries and fundamental postulates. Undoubtedly, the forward approach has been tremendously successful. To wit, among others, it predicted the magnetic moment of the electron to a spectacular number of significant digits. But there are limitations to this historically successful approach.

While forward modeling historically came first, inverse methods, spurred in equal parts by advances in probability theory and motivating data-centric questions in the Natural Sciences, also arose. Today, large swathes of complicated enough physical and chemical systems, in addition to Life and Social Sciences, are not naturally

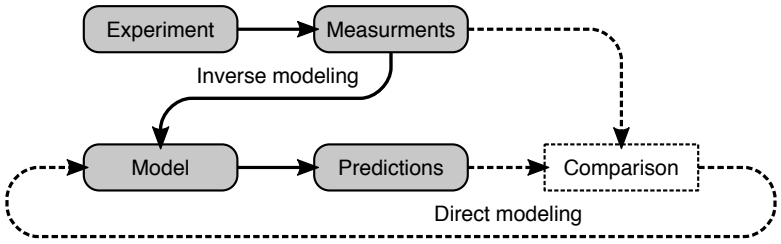

Figure 1.10: Ioannis: typo in fig. Measurement Illustration of the direct and inverse modeling paradigms. In the direct paradigm, a model is adjusted until its predictions agree with an experiment's measurements. By contrast, in the inverse paradigm, a model is inferred from the experiment's measurements without adjustments.

modeled from the ground-up, *i.e.* starting from first principles. Instead, in these cases, observations often only suggest loose couplings between variables of interest and probabilistic relations between various quantities that must be aided by data.

The forward approach is different from the philosophy we adopt here. Instead, we use the first principles only to motivate forms for our generative models. Beyond this, we are motivated by the practice of Statistics that instead attempts, from the onset, to be as agnostic about the model parameters (or the model itself) as possible and learn parameters and models self-consistently from the available data as efficiently as computationally possible.

1.4 Exercise problems

Exercise 1.1: Math warm-up

Evaluate the following by hand.

1. Gaussian integral: $\int_{-\infty}^{+\infty} dx\, e^{-(x-\mu)^2/(2\sigma^2)}$, assume μ and σ^2 are real scalars and $\sigma^2 > 0$.
2. Gaussian moments: $\int_{-\infty}^{+\infty} dx\, x^2 e^{-(x-\mu)^2/(2\sigma^2)}$, assume μ and σ^2 are real scalars and $\sigma^2 > 0$
3. Gaussian product: Show that the product of N Gaussians remains a Gaussian.
4. Gamma-function integral: Assume n is a positive integer and show that $\int_0^\infty dx\, x^n e^{-x}$ is equal to $n!$.
5. Gamma-function integral: Assume n is positive integer, a is a positive real scalar and show that $\int_0^\infty dx\, x^n e^{-x/a}$ is equal to $a^{n+1} n!$.
6. Geometric series: $\sum_{n=0}^{N} x^n$, assume N is a positive integer.
7. Poisson variance: $\sum_{n=0}^{\infty} n^2 \lambda^n \exp(-\lambda)/n!$, assume λ is real and positive.

Exercise 1.2: Permutations and combinations

Consider integers $N = 1, 2, \ldots$ and $M = 0, 1, \ldots, N$.

1. Show that the total number of distinct arrangements (permutations) of M objects selected out of N distinct objects is $\frac{N!}{(N-M)!}$.
2. Show that if we ignore the arrangement of the objects (combinations) the total number drops to $\frac{N!}{M!(N-M)!}$.
3. Show that the total number of different combinations of N distinct objects is 2^N.

Exercise 1.3: Cumulative probability function

Explain why the cumulative probability function in eq. (1.9) takes only values between 0 and 1.

Exercise 1.4: Cumulative and quantile functions of exponential random variables

Verify the formulas of $C(r)$ and $C^{-1}(r)$ in example 1.12.

Exercise 1.5: Sum of random variables

Consider two independent random variables R_1 and R_2 with densities $p_1(r_1)$ and $p_2(r_2)$, respectively. Use a transformation to show that the density $p_3(r_3)$ of a random variable R_3, with values $r_3 = r_1 + r_2$, is equal to the convolution $p_3(r_3) = (p_1 * p_2)(r_3)$.

Exercise 1.6: Minimum of exponential random variables

Consider two exponential random variables $R_1 \sim$ Exponential(λ_1) and $R_2 \sim$ Exponential(λ_2). Show that the random variable R_3, with values $r_3 = \min(r_1, r_2)$, follows an Exponential$(\lambda_1 + \lambda_2)$ distribution.

Exercise 1.7: A sanity check on random variable rescaling

Verify that the density $q(v)$ of the rescaled random variable V in example 1.6 has the correct units and that it is properly normalized.

Exercise 1.8: Linear transformations

In examples 1.6 and 1.7 we have seen how to obtain the probably density of random variables under rescaling and rotation. However, these two separate operations can be combined in a single more general one. For instance, consider a bivariate random variable (X, Y) and suppose that (X', Y') is the random variable under a linear transformation

$$x' = Ax + By + C, \qquad y' = Dx + Ey + F,$$

where A, B, C, D, E, and F are constants. Find the probability density of (X', Y') in terms of $p(x, y)$ and to avoid degeneracies consider only the case with $AE - BD \neq 0$.

Exercise 1.9: Division of random variables

Consider a random variable V with values $v = 1/x$, where X is a scalar random variable with density $p(x)$. Compute the density $q(v)$ in terms of $p(x)$.

Exercise 1.10: Spherical coordinate transformations

Consider a tri-variate random variable (X, Y, Z) that models a position in the Cartesian space. Use a transformation of random variables to relate the probability density $p(x, y, z)$ with the probability density $q(r, \phi, \theta)$ of the same position in spherical coordinates (R, Φ, Θ).

Exercise 1.11: Gamma random variables and derivatives

Suppose $R_1 \sim$ Gamma(α_1, β) and $R_2 \sim$ Gamma(α_2, β) are independent random variables. Find the probability densities of the random variables V_1, V_2, V_3 with values

$$v_1 = r_1 + r_2, \qquad v_2 = \frac{r_1}{r_1 + r_2}, \qquad v_3 = \frac{r_1}{r_2}.$$

Exercise 1.12: Manipulating transformed densities

Consider iid random variables R_1, R_2, R_3 with a common density $p(r)$. Further, assume R_1, R_2, R_3 are random variables that can only be realized as real and positive scalars. Find, in terms of $p(r)$, the probability that the polynomial $r_1 x^2 + r_2 x + r_3$ has real roots.

Exercise 1.13: The Weibull distribution

Consider a continuous random variable $X \sim \mathbb{P}$ that takes real scalar values and whose probability density is

$$p(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}}, \qquad x > 0$$

for appropriate values α and β.

1. Describe an algorithm that uses the fundamental theorem of simulation to simulate the random variable X.
2. Implement your algorithm.
3. Use simulations to verify that your implementation produce variables with the correct statistics.

Exercise 1.14: A fair dice

Use the fundamental theorem of simulation to simulate a roll of a fair dice. Generate several rolls and verify that indeed the dice simulated is fair.

Exercise 1.15: Poisson convolution

Show that the convolution of multiple Poisson distributions remains a Poisson distribution.

Exercise 1.16: Label invariance of the fundamental theorem of simulation

1. Apply the fundamental theorem of simulation to simulate draws from $\text{Categorical}_{\rho_1,\rho_2,\rho_3}(\pi_{\rho_1}, \pi_{\rho_2}, \pi_{\rho_3})$.
2. Verify that ρ_1, ρ_2, ρ_3 are realized with probabilities $\pi_{\rho_1}, \pi_{\rho_2}, \pi_{\rho_3}$, respectively.
3. Apply a relabeling of ρ_1, ρ_2, ρ_3 and verify that the fundamental theorem of simulation keeps yielding realizations with the correct probabilities.

Exercise 1.17: Normal random variables

1. Implement the Box-Muller algorithm of note 1.9 and generate a large number of $\text{Normal}(\mu, \sigma^2)$ random values.
2. Use your generated values to construct histograms of the $\text{Normal}(\mu, \sigma^2)$ probability density and verify that your implementation yields the correct statistics.

Exercise 1.18: Exponential rate

1. Sample 100 exponential random variables (the rate, λ must be specified by hand). Use your generated data to construct histograms of the CDF and PDF (with reasonable bin sizes selected at will).
2. Use the mean of your data to estimate the rate, λ. Then estimate the mean by fitting the CDF as well as the PDF (using whichever preferred criterion, like minimizing a mean square difference between the fit and the data, to find the best λ). Which of the three methods for estimating λ seems to yield more accurate results? Why?

Exercise 1.19: A loaded dice

A dice is rolled 120 times yielding the results:

face	"1"	"2"	"3"	"4"	"5"	"6"
number of appearances	15	34	18	19	19	15

Reason, based on a likelihood approach, that the dice is loaded.

Exercise 1.20: The point spread function in fluorescence microscopy

In fluorescence microscopy, light is detected at positions that are only probabilistically related to the position from which it is emitted. Under ideal imaging conditions, each photon emitted from a light emitter at position $(x_\star, y_\star)$ is detected independently of the other photons at a position (X, Y) that is randomly distributed according to the *Airy probability density*

$$p(x,y) = \frac{4\pi n_\alpha^2}{\lambda^2} \left(\frac{J_1\left(\frac{2\pi n_\alpha}{\lambda}\sqrt{(x-x_\star)^2 + (y-y_\star)^2}\right)}{\frac{2\pi n_\alpha}{\lambda}\sqrt{(x-x_\star)^2 + (y-y_\star)^2}} \right)^2$$

where λ is the photon's wavelength, n_α is the microscope's numerical aperture and $J_1(\cdot)$ is the 1st Bessel function of the first kind. Typical values for the parameters are $\lambda = 510$ nm and $n_\alpha = 1.40$.

1. Verify that the probability density $p(x, y)$ is properly normalized and determine the units of all quantities involved.
2. Apply a transformation from Cartesian to polar coordinates and chance the photon detection position from (X, Y) to radius and azimuth (R, Φ) relative to $(x_\star, y_\star)$.
3. Use the Airy density of (X, Y) to derive the density of (R, Φ).
4. Apply a fixed grid of radii $r_{1:M}$ and tabulate the corresponding probability density at these distances.
5. Use your tabulation and numerical integration to approximate the cumulative function $C(r_m)$ at the grid's radii $r_{1:M}$.
6. Use interpolation to approximate the cumulative function $C(r)$ at radii between $r_{1:M}$.
7. Use the fundamental theorem of simulation and your interpolated $C(r)$ to simulate the detection position of a large number of photons.
8. Summarize your simulated positions in a histogram and verify that indeed your implementation produces photon detections from the correct distribution.