VU Data Science in Astrophysics (Summer 2024)
Chapter 1: Statistics and Spatial Statistics

Oliver Hahn*

March 21, 2024

*oliver.hahn@univie.ac.at

1 Random Variables and Distributions

1.1 Definitions: Samples, Random Variables, Probabilities

We will understand in this lecture all **data** as the result of a **random process**. This does not mean that it is 'random', in the sense of 'arbitrary', but that the specific data we are given (e.g. a measurement, a simulation outcome) must always be seen in relation to some underlying large space. The specific data we are given is just one element of this space.

Sample Space Ω. The sample space is the set of all possible data that we could have obtained. This set can be finite or infinite: e.g. if the data is the mass of a star, then the **sample space is continuous**, specifically a real number (the mass measurement divided by some unit, e.g. solar masses)

$$\text{sample space of the mass of a star in } M_\odot: \quad \Omega = \mathbb{R}$$

If the space is one-dimensional, we say that it is **univariate**, otherwise it is called **multivariate**.

If the data is e.g. the spectral class of a star, then the **sample space is discrete** ('discrete' meaning countable, not necessarily finite), specifically

$$\text{sample space of stellar spectral class:} \quad \Omega = \{O, B, A, F, G, K, M\}$$

A note on **ordered sample spaces**: a priori we have no order of the elements of a sample space, i.e. we can not simply compare two elements $x_i, x_j \in \Omega$ as $x_i < x_j$. This is e.g. true for our discrete sample space of stellar spectral classes (is 'O' < 'K'?) but it is also true if our sample space is e.g. the position of a star on the sky. For many machine learning applications, it is however common to associate discrete sample spaces with integer numbers, e.g. 'O'= 1, 'B'= 2, ...

A note on **digitization**: digital information is by definition always discrete. An image e.g. might consist of $N \times M$ pixels with e.g. a 8 bit brightness value. Assuming the brightness is expressed as an integer $\mathbb{B}_8 \equiv \{0, 1, \ldots, 2^8 - 1\} = \{0, 1, \ldots, 255\}$, the sample space of such an image is thus $\Omega = \mathbb{B}_{24}^{N \times M}$, i.e. there are about 10^{157286} different images. Only a tiny fraction of these corresponds to images that we would consider 'meaningful' (e.g. an image of a galaxy, or a cat).

Again, we have a priori no ordering on this space (i.e. it does not make sense to compare two images via '<' without first defining what we mean by '<'). Note that once we introduce a distance measure, we can define an ordering on the space, but this is not a priori given. As an example, we can define a distance measure on the space of images as the sum of the absolute differences of the pixel values, i.e. the L_1-norm. Once we have that, we can define an ordering given by the distance e.g. from a given reference image, say the distance of a given image from a reference image of a cat.

Random Variable. A random variable X is an element from the sample space, i.e. $X \in \Omega$ (i.e. Ω encompasses all possible outcomes for X), which is drawn from some random process $\mathcal{P}_X$, for which we write $X \sim \mathcal{P}_X$. The 'random process' should not be thought of as something arbitrary, it can be e.g. a measurement of an atomic spectrum in the lab, or an image of a galaxy obtained from a telescope, or data from a simulation. The random variable is the result of this process, and it is random in the sense that we do not know what the outcome will be before we have observed it.

Probability. In the discrete case, the probability $p_i \geq 0$ to draw the value $X_i = x_i$, where $x_i \in \Omega$ can be any element of the sample space, is written as

$$\mathbb{P}[X_i = x_i] = p_i \geq 0. \tag{1}$$

The sum of all probabilities p_i is always 1, i.e.

$$\sum_{x_i \in \Omega} p_i = 1, \tag{2}$$

which reflects that the sample space Ω must be complete, i.e. encompass all possible outcomes. Note that it can include impossible outcomes, which have $p_i = 0$. The joint probability to have $X_1 = x_1$ and $X_2 = x_2$ and $\ldots, X_n = x_n$ is written as $\mathbb{P}[X_1 = x_1, \ldots, X_n = x_n]$ or $\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}]$ where assemble the variables into a vector $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ and $\boldsymbol{x} = \{x_1, \ldots, x_n\}$.

Finally, we define the **conditional probability** that $X = x$ given that $Y = y$ as

$$\mathbb{P}\left[X = x \mid Y = y\right] := \frac{\mathbb{P}[X = x, Y = y]}{\mathbb{P}[Y = y]} \quad \text{iff} \quad \mathbb{P}[Y = y] > 0. \tag{3}$$

If the probability that $X = x$ and $Y = y$ are **statistically independent**, then

$$\mathbb{P}\left[X = x \mid Y = y\right] = \mathbb{P}\left[X = x\right] \quad \text{and} \quad \mathbb{P}\left[Y = y \mid X = x\right] = \mathbb{P}\left[Y = y\right] \tag{4}$$

and as a consequence

$$\mathbb{P}[X = x, Y = y] = \mathbb{P}[X = x]\, \mathbb{P}[Y = y], \tag{5}$$

i.e. the probabilities are simply multiplied.

Product rule of probabilities. The joint probability of many variables can be written as conditional probabilities of single variables, i.e.

$$\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}] = \mathbb{P}[X_1 = x_1] \prod_{i=2}^{n} \mathbb{P}\left[X_i = x_i \mid X_1 = x_1, \ldots X_{i-1} = x_{i-1}\right] \tag{6}$$

and therefore

$$\begin{aligned} \mathbb{P}[X = x, Y = y, Z = z] = \mathbb{P}\left[X = x \mid Y = y, Z = z\right] \\ \times \mathbb{P}\left[Y = y \mid Z = z\right] \\ \times \mathbb{P}\left[Z = z\right]. \end{aligned} \tag{7}$$

Theorem 1. *(Bayes 1763) Bayes' famous theorem allows us to reverse conditional probabilities. From eq. (3) follows immediately that*

$$\mathbb{P}\left[X = x \mid Y = y\right] = \frac{\mathbb{P}\left[Y = y \mid X = x\right]\ \mathbb{P}\left[X = x\right]}{\mathbb{P}\left[Y = y\right]}. \tag{8}$$

It is mostly used in statistical inference in the sense of (evidence for model given data)$\propto$(likelihood for data, given model)$\times$(prior evidence for the model parameters). We will come back to this in later lectures.

1.2 The Probability Distribution Function (PDF)

In the continuous case, the possible outcomes and their frequency of occurence (or degree of belief that they will occur) can be fully described by the PDF $p_X : \Omega \to \mathbb{R}_{0+}$ which assigns to every element of the sample space Ω a number ≥ 0, the 'probability density' – where $p_X \equiv 0$ for impossible outcomes, and otherwise $p_X > 0$, as before. We require that the PDF is normalised, i.e. that

$$p_X(x) \geq 0 \qquad \text{and} \qquad \int_\Omega \mathrm{d}x\, p_X(x) = 1. \tag{9}$$

In the **continuous univariate case**, the PDF is defined so that the probability to find X between two values $a, b \in \Omega$ is given by[1]

$$\mathbb{P}[a \leq X \leq b] = \int_a^b \mathrm{d}x\, p_X(x). \tag{10}$$

Note that this definition of probablity implies that the probability to find a specific value is always zero in the continuous case, i.e. $\mathbb{P}[X = x_i] = 0$ (mathematically we say 'a set of measure zero').

Theses concepts are readily generalised to the multivariate case, i.e. the case when Ω is an n-dimensional space. In this case, let $\boldsymbol{x} \in \Omega$ be an n-dimensional vector, then we have to generalise the integral to higher dimensional spaces, i.e.

$$p_{\boldsymbol{X}}(\boldsymbol{x}) \geq 0 \qquad \text{and} \qquad \int_\Omega \mathrm{d}^n x\, p_{\boldsymbol{X}}(\boldsymbol{x}) = 1. \tag{11}$$

And we have to generalise the probability property (10) to subsets $S \subseteq \Omega$ so that

$$\mathbb{P}[\boldsymbol{X} \in S] = \int_S \mathrm{d}^n x\, p_{\boldsymbol{X}}(\boldsymbol{x}), \tag{12}$$

[1]Note that we have made a notational short cut here: Mathematically correctly, the integral should read $\int_a^b \mathrm{d}\mu(x)\, p_X(x)$, since the interval $[a, b]$ might not overlap with Ω everywhere and we should employ Lebesgue integration, we will glance over this generously however and implicitly assume that the correct measure is adopted.

1.3 Marginals and Conditional Probabilities

Let us consider an N-dimensional sample space Ω with PDF $p_{\boldsymbol{X}}$ where $\boldsymbol{X} = \{X_1, \ldots, X_N\}$. We define the **marginal** with respect to the i-th variable for discrete distributions as the sum

$$p_{\boldsymbol{Y}} = \sum_{x_i} p_{\boldsymbol{X}}(\boldsymbol{x}), \tag{13}$$

and for continuous distributions as the integral

$$p_{\boldsymbol{Y}} = \int \mathrm{d}x_i \, p_{\boldsymbol{X}}(\boldsymbol{x}), \tag{14}$$

where $\boldsymbol{Y} = \boldsymbol{X}_{\setminus i} = \{Y_1, \ldots, Y_{N-1}\}$ no longer includes X_i (we say we 'marginalised over X_i'). To make this more explicit, in the two-variate case, we obtain by marginalisation a univariate PDF $p_X(x) = \int \mathrm{d}y \; p_{X,Y}(x, y)$ which is independent of the second variable Y.

Bivariate distributions along with their marginals are readily graphically shown using the SEABORN PYTHON library. Assuming we have set of sampled data points $\{X_1, X_2\}$, the joint distribution along with the marginals is readily displayed (see Fig. 1) using the following code:

```
import seaborn as sns

h = sns.jointplot(v1,v2 )
h.set_axis_labels('variable 1', 'variable 2')
```

Note that the histograms just count the events from the empirical distribution $\hat{p}$ falling into a finite interval, we will discuss histograms in more detail in the next lecture.

While in the code snippet above, we have simply passed NUMPY arrays as arguments, SEABORN works even better if we use PANDAS data frames. Let us assume, we have a new data set $\{X_1, X_2, I\}$ where $I \in \{0, 1\}$ is a binary label grouping the two first random variables (which come from a continuous distribution) into two classes. In this case we would can use PANDAS to visualize all this data together as follows

```
import pandas as pd

# create a pandas data frame from numpy arrays v1, v2, and i
df = pd.DataFrame({
  'variable 1': v1,
  'variable 2': v2,
  'class': i})
df.head(n=3)

sns.jointplot(data=df, x="variable 1", y="variable 2", hue="class")
```

Note that in this case PANDAS automatically chooses to represent the marginals $\hat{p}_{X_1}$ and $\hat{p}_{X_2}$ in the form of kernel density estimates. Again, we will discuss what this is in detail in the next lecture, when we discuss density estimation.

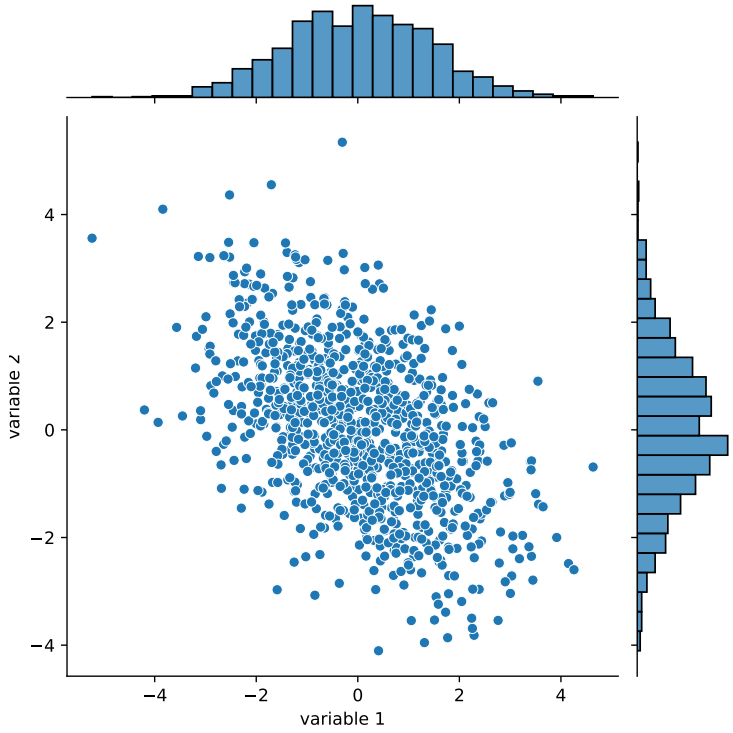

Figure 1: Scatter plot of 1000 data points $\{X_1, X_2\}$ from a bivariate random process (here we have two correlated normal variates) along with the marginals shown in the form of histograms. The histogram at the top shows the distribution marginalised over the second variable, i.e. a histogram of $\hat{p}(x_1)$, while the histogram on the right shows the distribution marginalised over the first variable, i.e. a histogram of $\hat{p}(x_2)$

1.4 Common probability distributions

A few common probability distributions that we will encounter repeatedly are

The Dirac δ distribution. The Dirac distribution is the 'single outcome' distribution where X is always μ. Its PDF is defined so that it vanishes everywhere except at μ yet is still normalised and we write

$$p_\delta(x;\, \mu) = \delta_D(x - \mu). \tag{15}$$

Every discrete distribution can be mapped to a sum of δ-distributions weighted by their respective probabilities p_i. As a consequence, we can also represent any empirical data $\mathcal{S} = \{\boldsymbol{X}_i \in \Omega\}_{i=1\ldots n}$ through its **empirical distribution**

$$\hat{p}_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \delta_D(\boldsymbol{x} - \boldsymbol{X}_i). \tag{16}$$

Note that we indicate the fact that it is empricial with the hat '$\hat{}$'.

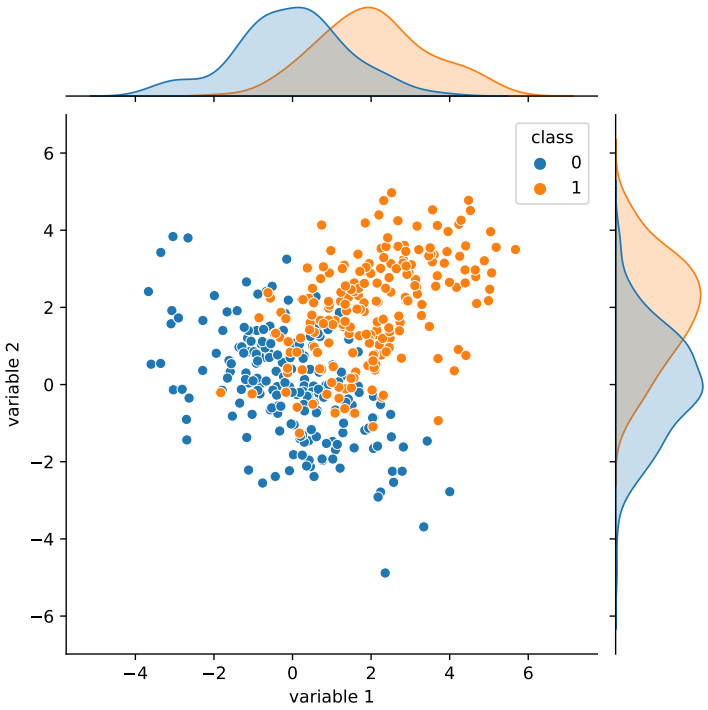

Figure 2: Scatter plot of 400 data points $\{X_1, X_2, I\}$ from a tri-variate random process (two correlated normal variates with an additional binary class label) along with the marginals shown in the form of kernel density estimates (KDEs).

1.4.1 Important Continuous Distributions

Uniform distribution. The uniform distribution $U(a, b)$ has a constant probabilty to find $x \sim U(a, b)$ in the interval $[a, b]$ and zero outside, i.e.

$$p_U(x;\, a, b) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} . \tag{17}$$

Most commonly we consider the unit uniform distribution $U(0, 1)$.

Normal distribution. The univariate normal distribution $\mathcal{N}(\mu, \sigma^2)$ is characterised by only two numbers, its mean μ and its variance σ^2, the PDF has the form

$$p_{\mathcal{N}}(x;\, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] . \tag{18}$$

We abbreviate the normal distribution of zero mean and unit variance as $\mathcal{N} = \mathcal{N}(0, 1)$.

In the multi-variate (specifically N-variate) case, these become the mean N-vector μ

and the $N \times N$ covariance matrix $\mathbf{\Sigma}$, and one has

$$p_{\mathcal{N}}(x;\, \boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^N \det \mathbf{\Sigma}}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right]. \tag{19}$$

The χ and χ^2 distributions. Assume we draw k uncorrelated numbers from a standard normal distribution, i.e. $X_i \sim \mathcal{N}$, $i = 1, \ldots, k$. Then we can form two new random variables

$$A = \sum_{i=1}^{k} (X_k)^2 \qquad \text{and} \qquad B = \sqrt{\sum_{i=1}^{k} (X_k)^2}. \tag{20}$$

The variable A will follow a so-called χ^2-distribution with k degrees of freedom, while the variable B will follow a so-called χ-distribution with k degrees of freedom. They have the PDFs

$$p_{\chi}(x;\, k) = \begin{cases} \frac{x^{k-1} \exp[-x^2/2]}{2^{(k-2)/2}\Gamma\left(\frac{k}{2}\right)} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}, \tag{21}$$

and

$$p_{\chi^2}(x;\, k) = \begin{cases} \frac{x^{(k-2)/2} \exp[-x/2]}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

where $\Gamma(x)$ is the gamma function, which is $\Gamma(n) = (n-1)!$ and $\Gamma(\frac{1}{2} + n) = \sqrt{\pi}\frac{(2n-1)!!}{2^n}$ for $n \in \mathbb{N}$, where the double factorial $n!! = n\,(n-2)\,(n-4)\,\cdots$.

1.4.2 Important Discrete Distributions

The Bernoulli distribution. The Bernoulli distribution is a distribution over a binary variable $X \in \{0, 1\}$. The probability for one state over the other is given by a parameter $\phi \in [0, 1]$ so that

$$\mathbb{P}[X = 1] = \phi \qquad \mathbb{P}[X = 0] = 1 - \phi. \tag{23}$$

It is interesting to note that we can assign a PDF to this distribution, which gives the right statistics but is continuous

$$p_X(x) = \phi^x (1-\phi)^{1-x}. \tag{24}$$

The categorial distribution. The generalisation of the Bernoulli distribution from two states to k states, each with a probability $\boldsymbol{p} = \{p_i\}_{i=1\ldots k} \in [0, 1]^{k-1}$. Note that the kth probability is of course given by $p_k = 1 - \sum_{i=1}^{k-1} p_i$. This distribution is often used for distributions over categories of objects, e.g. our spectral classes of stars from the beginning. Sometimes this distribution is also called the 'Multinoulli' distribution.

1.5 Cumulative Distributions

The **Cumulative Distribution Function (CDF)** $C : \Omega \to [0,1]$ is defined as the integral over the PDF up to a given value $X = x$

$$C_X(x) := \int_{-\infty}^{x} \mathrm{d}x'\, p_X(x') \qquad \text{so that} \qquad \mathbb{P}[a \leq X \leq b] = C_X(b) - C_X(a) \tag{25}$$

and also the probability that the random variable is below a given value is expressed directly by the value of the CDF at that value $\mathbb{P}[X \leq a] = C(a)$.

The **empirical CDF (eCDF)** can be easily obtained for univariate data which can be sorted. Let us assume that we have N measurements taken from a random process. Then the eCDF for a sample $\mathcal{S} = \{X_i \in \Omega\}_{i=1\ldots n}$ is defined as

$$\hat{C}_X(x) := \frac{\text{number of elements in } \mathcal{S} \text{ with } X_i \leq x}{\text{number of elements in } \mathcal{S}} = \frac{1}{n}\sum_{i=1}^{n} \Theta_H(x - X_i), \tag{26}$$

where Θ_H is the Heaviside function[2]. It can also be obtained by integrating over the empirical distribution $\hat{p}_X$. The eCDF can be numerically easily computed by rank ordering all elements of $\mathcal{S}$, i.e. by determining the ordered set $\mathcal{S}_O = (X_{j_1}, X_{j_2}, \ldots, X_{j_n})$ where $X_{j_a} \leq X_{j_{a+1}}$, and plotting the values X_j against the rank. In code this would be achived as follows

```
15  import numpy as np
16  import matplotlib.pyplot as plt
17  import scipy.special as sp
18
19  # get 100 normally distributed numbers as example
20  S = np.random.randn( 200 )
21  # sort S in ascending order
22  S_o = np.sort( S )
23
24  # plot
25  fig = plt.figure()
26  plt.step( S_o, np.arange(len(S_o))/len(S_o), label='empirical CDF' )
27  x = np.linspace(-3,3,100)
28  plt.plot( x, 1/2+1/2*sp.erf(x/np.sqrt(2)), 'k--', alpha = 0.5, label='CDF for ...' )
29  plt.eventplot(S, lineoffsets=-0.1, linelengths = 0.05, lw=0.5, colors='k')
```

Unfortunately, the concept of the eCDF is not readily generalised to the multivariate case. Note that the eCDF is readily obtained from the raw data, quite in contrast to estimates of the PDF. In fact, estimating the PDF is much more complicated and we will dedicate next week's lab course only to the topic of density estimation.

[2]The Heaviside step function is defined as $\Theta_H(x) = \begin{cases} 0 & \text{if} \quad x \leq 0 \\ 1 & \text{if} \quad x > 0 \end{cases}$. The derivative of the Heaviside function is the Dirac distribution $\frac{\mathrm{d}\Theta_H}{\mathrm{d}x} = \delta_D(x)$.

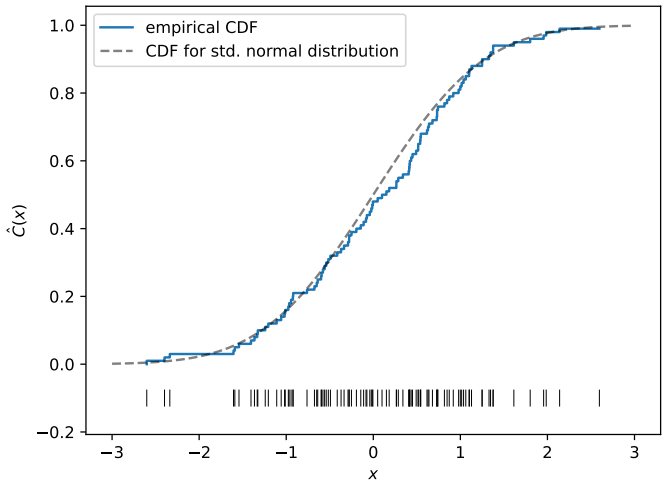

Figure 3: Empirical cumulative distribution function obtained from 100 normally distributed random numbers. The blue line indicates the eCDF, while the dashed gray line indicates the expected CDF for the standard normal distribution $C(x) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}(x/\sqrt{2})$. The black vertical dashes indicate the actual samples $\{X_i\}$, they can also be thought of as representing the empirical distribution $\hat{p}_X = \frac{1}{n}\sum_i \delta_D(x - X_i)$.

1.6 Summary statistics

Very often we do not know the underlying distribution that gave rise to the data we obtained. In this case, **summary statistics** provides an agnostic way to characterise quantitatively properties of a distribution.

1.6.1 Mean and variance

The **expectation value** of a function f of a random variable X is defined as

$$\mathbb{E}[f(X)] = \int_\Omega \mathrm{d}x\ f(x)\ p_X(x) \tag{27}$$

We find the expectation value to obey

$$\text{linearity} \qquad \mathbb{E}[\alpha f(X) + \beta g(X)] = \alpha\,\mathbb{E}[f(X)] + \beta\,\mathbb{E}[g(X)] \tag{28}$$

and define

$$\text{mean} \qquad \mu_X := \mathbb{E}[X] \qquad = \int_\Omega \mathrm{d}x\ x\ p_X(x) \tag{29}$$

$$\text{variance} \qquad \sigma_X^2 := \mathbb{E}[(X - \mathbb{E}[X])^2] \qquad = \int_\Omega \mathrm{d}x\ (x - \mu_X)^2\ p_X(x) \tag{30}$$

Multivariate/Joint Distributions. In the presence of multiple random variables $\boldsymbol{X} = (X_1, \ldots, X_N)$ with $X_i \in D_i \subseteq \Omega$ with $i = 1, \ldots, N$ the *multivariate* joint probability distribution is defined on the N-dimensional reals as $p_{\boldsymbol{X}} : \mathbb{R}^N \to \mathbb{R}_{0+}$. The normalisation is now

$$\mathbb{E}[1] = \int_\Omega \mathrm{d}^N x \, p_{\boldsymbol{X}}(\boldsymbol{x}) \stackrel{!}{=} 1. \tag{31}$$

For multivariate distributions, the n-th moment becomes a rank n tensor. In particular, the mean is now a vector of length N and the variance becomes the **covariance matrix**[3]

$$\begin{aligned} \text{mean vector} \qquad \boldsymbol{\mu}_{\boldsymbol{X}} &:= \mathbb{E}[\boldsymbol{X}] && (32) \\ &= \int_\Omega \mathrm{d}^N x \; \boldsymbol{x} \; p_{\boldsymbol{X}}(\boldsymbol{x}) \\ \text{covariance matrix} \qquad \boldsymbol{\Sigma}_{\boldsymbol{X}} &:= \mathbb{E}[(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}}) \otimes (\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})] && (33) \\ &= \int_\Omega \mathrm{d}^N x \; (\boldsymbol{x} - \boldsymbol{\mu}_X) \otimes (\boldsymbol{x} - \boldsymbol{\mu}_X) \; p_{\boldsymbol{X}}(\boldsymbol{x}) \end{aligned}$$

If the random variables are **mutually statistically independent**, then

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{N} p_{X_i}(x_i) \qquad \text{and} \qquad \boldsymbol{\Sigma}_{\boldsymbol{X}} = \mathrm{diag}(\sigma^2_{X_1}, \ldots, \sigma^2_{X_N}). \tag{34}$$

1.6.2 Moment and cumulant expansions

Moments of a distribution. Generally, we can expand the probability distribution function in terms of monomials, that together will fully describe the PDF. The expectation values of the monomials we call **(raw) moments of the distribution**. In the univariate case, we have

$$k\text{-th moment} \qquad m_X^{(k)} := \mathbb{E}[X^k] \qquad = \int_\Omega \mathrm{d}x \; x^k \; p_X(x) \tag{35}$$

In the multivariate case, the moments become tensors: the k-th moment is a rank k tensor and we have

$$k\text{-th moment} \qquad m_{\boldsymbol{X}}^{(k)} := \underbrace{\mathbb{E}[\boldsymbol{X}^{\otimes k}]}_{=\mathbb{E}[\boldsymbol{X} \otimes \cdots \otimes \boldsymbol{X}]} \qquad = \int_\Omega \mathrm{d}x \; \boldsymbol{x}^{\otimes k} \; p_{\boldsymbol{X}}(\boldsymbol{x}) \tag{36}$$

[3] Note that the tensor product between two vectors is $\boldsymbol{a} \otimes \boldsymbol{b} = (a_i b_j)$.

Cumulants. Note that the sequence of moments is not independent, the k-th moment contains contributions from all lower order moments. It is therefore desirable to isolate the 'intrinsic' part of a moment from the low order 'baggage'.

We find the 'connected' components, or **cumulants** $c_{\boldsymbol{X}}^{(k)}$ as the intrinsic contribution of each moment, one finds

$$c_{\boldsymbol{X}}^{(0)} = m_{\boldsymbol{X}}^{(0)} = 1, \qquad c_{\boldsymbol{X}}^{(1)} = m_{\boldsymbol{X}}^{(1)} = \boldsymbol{\mu}_{\boldsymbol{X}}, \qquad c_{\boldsymbol{X}}^{(2)} = m_{\boldsymbol{X}}^{(2)} - m_{\boldsymbol{X}}^{(1)} \otimes m_{\boldsymbol{X}}^{(1)} = \boldsymbol{\Sigma}_{\mathbf{X}}, \qquad \dots \tag{37}$$

The cumulants are called 'connected' because they have an intimate connection to the graphical representation shown in Figure 4, where we give them in index notation.

Theorem 2. *(Marcinkiewicz 1939) The Dirac δ-distribution and the normal distribution $\mathcal{N}$ are the only distributions with a finite number of non-zero cumulants $c_X^{(k)}$. In the case of the Dirac distribution, all cumulants $k > 1$ vanish, for the normal distribution all cumulants $k > 2$ vanish.*

Theorem 3. *(Isserlis 1918 / Wick 1955) Let $(X_1, \dots, X_n)$ be a random variable drawn from a $n-$variate normal distribution with zero mean. Then the following relation holds:*

$$\mathbb{E}[X_1 \dots X_n] = \begin{cases} \sum_{p \in P_n^2} \prod_{\{i,j\} \in p} \mathbb{E}[X_i X_j] & \text{if } n \text{ is even} \\ 0 & \text{if } n \text{ is odd} \end{cases} \tag{38}$$

where the sum runs over all distinct pairings of $\{1, \dots, n\}$, i.e. all distinct ways of partitioning $\{1, \dots, n\}$ into pairs $p = \{i, j\}$ and the product runs over all pairs in p.

For example, for $n = 4$ then we have e.g.

$$\mathbb{E}[X_1 X_2 X_3 X_4] = \mathbb{E}[X_1 X_2]\, \mathbb{E}[X_3 X_4] + \mathbb{E}[X_1 X_3]\, \mathbb{E}[X_2 X_4] + \mathbb{E}[X_1 X_4]\, \mathbb{E}[X_2 X_3]\ ,$$

or e.g.

$$\mathbb{E}[X_1^4] = 3\mathbb{E}[X_1^2]\, \mathbb{E}[X_1^2]\ .$$

A consequence of Marcinkiewicz' theorem is of course that moment/cumulant expansions are not particularly useful to describe distributions that are far from Gaussian (or Dirac), aka 'non-Gaussian'. However, in that case cumulants $k > 2$ are useful to quantify the deviation from Gaussianity. Very commonly used are the next two cumulants, in dimensionless (aka 'normalized') form, given here for an univariate distribution:

$$k = 3: \qquad c_X^{(3)} = m_X^{(3)} - 3c_X^{(2)} c_X^{(1)} - \left(c_X^{(1)}\right)^3$$

often expressed as the dimensionless skewness $S = c_X^{(3)} / \left(c_X^{(2)}\right)^{3/2}$, and

$$k = 4: \qquad c_X^{(4)} = m_X^{(4)} - 6c_X^{(2)} \left(c_X^{(1)}\right)^2 - 3\left(c_X^{(2)}\right)^2 - 4c_X^{(3)} c_X^{(1)} - \left(c_X^{(1)}\right)^4$$

with dimensionless kurtosis $K = c_X^{(4)} / \left(c_X^{(2)}\right)^2$. The coefficients and the combinations of cumulants contributing are conveniently read off from the diagrammatic representation in Figure 4.

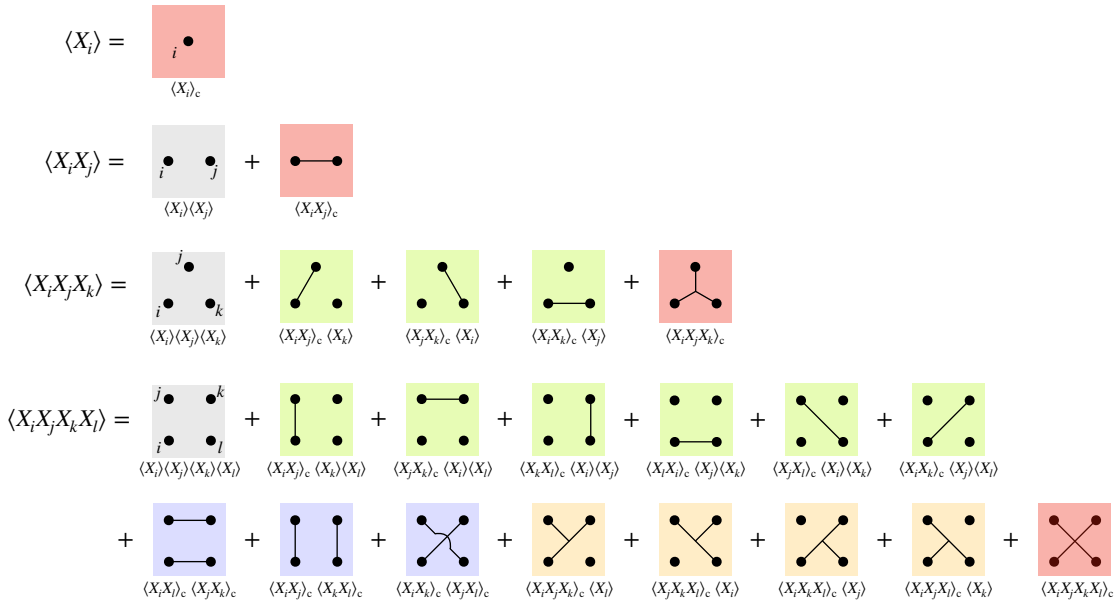

Figure 4: Expansion of moments into cumulants from first to fourth order and their representation as undirected graphs. Note that Red shaded components are the (fully) connected components of each order, all other diagrams of a given order can be composed out of products of lower order connected components. Note that if the variables have vanishing mean, i.e., $\mathbb{E}[X_i] = 0$ etc., then the number of non-zero diagrams is dramatically reduced, i.e., all grey, green and orange contributions vanish. If furthermore X_i is Gaussian then all connected cumulants vanish except those that are powers of the second cumulant (e.g., the only non-zero terms in the 4th-order moment stems from the blue contributions). Note that for notational simplicity, in the diagram, we have replaced expectation values with angle brackets.

Standardization of random variables. As we have seen, simple summary statistics such as the moments depends sensitively on the magnitude of low order moments. In order to reduce such relatively trivial effects, a common trick is to **standardize variables**. Note that the linear transformation of a normally distributed random variable yields a normally distributed random variable, i.e.

$$\text{if} \quad X \sim \mathcal{N}(\mu, \sigma^2), \quad \text{then} \quad (Y = a\,X + c) \sim \mathcal{N}(\mu + c,\, a^2\sigma^2). \tag{40}$$

or in the multi-variate case, given an $m \times m$ matrix $\mathbf{A}$ and an m-vector $\mathbf{c}$,

$$\text{if} \quad X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \text{then} \quad (\boldsymbol{Y} = \mathbf{A}\boldsymbol{X} + \boldsymbol{c}) \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \boldsymbol{c}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top). \tag{41}$$

We might therefore as well standardize the random variables by rescaling $\{X_j\}$ to

$$\tilde{X}_j := \frac{X_j - \mu_{X_j}}{\sqrt{\sigma^2_{X_j}}} \quad \text{so that} \quad \mathbb{E}[\tilde{X}_j] = 0 \quad \text{and} \quad \mathbb{E}[\tilde{X}_j^2] = 1. \tag{42}$$

Note that standardization in the multivariate case is done by treating each dimension separately rather than taking the covariance matrix into account, i.e. normalization is performed using $\sigma^2_{X_j} = \Sigma_{jj}$.

In the same spirit, the **Pearson correlation coefficient** is defined as the standardized covariance between any pair of random variables (X_1, X_2) as

$$\rho_{X_1,X_2} = \mathbb{E}[\tilde{X}_1 \tilde{X}_2] = \frac{\mathbb{E}[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})]}{\sqrt{\sigma^2_{X_1}\, \sigma^2_{X_2}}} \,. \tag{43}$$

a positive number indicates a positive correlation, a negative number a negative correlation aka anti-correlation.

Non-linearity and non-Gaussianity. A corollary that follows from the linear transformation property of normal variates above is that a non-linear transformation of a normal variate always yields a 'non-Gaussian' distribution. The consequence is that a non-linear process that starts from normally distributed data will typically lead to a non-Gaussian distribution.

Rank order statistics. Due to the sensitivity of statistics to non-linear maps, another common technique is to replace the values of a random variable X with its rank, i.e. given the data $\{X_j\}_{j=1\ldots n}$ sort in ascending fashion so that $X_{i_1} \leq X_{i_2} \leq \cdots \leq X_{i_n}$. The rank r is then the position in this sorted list, i.e. i_r. We denote this rank transformation through the map $R(X)$, which maps from $\Omega \to \mathbb{N}$. The **Spearman rank correlation coefficient** is then obtained by computing the Spearman correlation coefficient for the ranks instead of for the values, i.e.

$$\rho^{\text{Spearman}}_{X,Y} = \rho_{R(X),R(Y)} \tag{44}$$

Since the rank order is invariant under any monotonic (nonlinear) function, the Spearman correlation coefficient is also invariant under monotonic transformations (such as taking a logarithm).

1.7 Empirical estimators of moments and cumulants

Typically, we do not know the underlying distribution that gave rise to the data we obtained. In this case, **summary statistics** provides an agnostic way to characterise quantitatively properties of a distribution. We need a way to estimate the moments and cumulants from the data.

1.7.1 Frequentist estimators of moments.

While the expressions for the moments we have given above can be computed when the true underlying distribution $p_{\boldsymbol{X}}$ is known, it is straightforward to obtain empirical estimators by simply replacing $p_{\boldsymbol{X}}$ with the empirical distribution $\hat{p}_{\boldsymbol{X}}$ obtained from a data sample $\{\boldsymbol{X}_j\}_{j=1,\ldots,n}$. One finds the following frequentist estimators, e.g.

$$\hat{m}_{\boldsymbol{X}}^{(1)} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{X}_j =: \langle \boldsymbol{X} \rangle \tag{45a}$$

$$\hat{m}_{\boldsymbol{X}}^{(2)} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{X}_j \otimes \boldsymbol{X}_j =: \langle \boldsymbol{X} \otimes \boldsymbol{X} \rangle \tag{45b}$$

$$\vdots$$

Analogously to the empirically estimated moments, we denote empirically estimated cumulants as

$$\hat{c}_{\boldsymbol{X}}^{(1)} =: \langle \boldsymbol{X} \rangle_c \,, \qquad \hat{c}_{\boldsymbol{X}}^{(2)} = \langle \boldsymbol{X} \otimes \boldsymbol{X} \rangle_c \,, \qquad \ldots \tag{46}$$

We therefore have the following estimators for the mean and variance along with the variance of the mean

$$\hat{\mu}_X = \frac{1}{n}\sum_{j=1}^{n} X_j, \qquad\qquad \hat{\sigma}_X^2 = \frac{1}{n}\sum_{j=1}^{n} (X_j - \hat{\mu}_X)^2 \,. \tag{47}$$

We can prove that these estimators are unbiased, i.e. that the expectation value of the estimator is equal to the true value of the parameter, i.e. $\mathbb{E}[\hat{\mu}_X] = \mu_X$ and $\mathbb{E}[\hat{\sigma}_X^2] = \sigma_X^2$. The proof is straightforward, we have

$$\mathbb{E}[\hat{\mu}_X] = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} X_j\right] = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}[X_j] = \mu_X, \tag{48}$$

and similarly for the variance, we have

$$\mathbb{E}[\hat{\sigma}_X^2] = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} (X_j - \hat{\mu}_X)^2\right] = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}[(X_j - \hat{\mu}_X)^2] = \sigma_X^2. \tag{49}$$

We can also show that they are obtainable from a maximum likelihood estimator, i.e. that they are the most likely estimator given the data. The likelihood function for the parameters μ and σ^2 given the data assuming a normal distribution is:

$$L(\mu, \sigma^2 | X_1, ..., X_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}} \tag{50}$$

Taking the natural logarithm of the likelihood function gives the log-likelihood function:

$$l(\mu, \sigma^2 | X_1, ..., X_n) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

To find the maximum likelihood estimates, we take the derivative of the log-likelihood function with respect to μ and σ^2, and set them equal to zero. The derivative with respect to μ is:

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)$$

The derivative with respect to σ^2 is:

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

Setting both equal to zero and solving for μ and σ^2 gives the maximum likelihood estimate of mean and variance, which turn out to be identical to 47.

1.7.2 Bayesian estimators of moments.

Let us assume that there is a joint distribution $p_{\boldsymbol{X}}(\boldsymbol{X}; \boldsymbol{\theta})$ of the data and its summary statistics, which implies that the data has a common mean and variance. In the Bayesian framework, we can obtain estimators for the moments by assuming a prior distribution $p(\boldsymbol{\theta})$ over the parameters $\boldsymbol{\theta} = (\mu, \sigma^2)$ of the distribution The posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{X})$ is then obtained by applying Bayes' theorem, i.e.

$$p(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\boldsymbol{X})} \propto p(\boldsymbol{X}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}), \tag{51}$$

where $p(\boldsymbol{X}|\boldsymbol{\theta})$ is the likelihood function $L(\mu, \sigma^2|\boldsymbol{X})$ from eq. (50) above. If we have no prior information, then the prior should not be informative, in which case the Maximum Likelihood Estimator (MLE) is the same as the Maximum A Posteriori (MAP) estimator (since the likelihood and the posterior are equivalent in this case). If one tries to be more accurate, the prior should include information such as the fact that the prior distribution must have a rescaling invariance, e.g. $p(\mu, \sigma^2) = b\, p(\mu + c, b^2\sigma^2)$ for any constant $b, c \in \mathbb{R}$, $b > 0$. Taking such considerations into account, slightly modified estimators can be obtained, see e.g. this article by Travis Oliphant. Note that in the limit of large sample sizes, all the estimators converge to the same value.

1.8 Estimating errors on summary statistics

Typically we want to give also confidence intervals for the summary statistics. There are multiple choices, of increasing sophistication. If we want to report a precise measurement, it is critically important to report an accurate estimate of the error of the summary statistic.

1.8.1 Error of the mean

Since the sample mean is an unbiased estimator of the population mean, its variance is the variance of the population divided by the sample size

$$\sigma^2_{\hat{\mu}_X} = \frac{\sigma^2_X}{n} \approx \frac{\hat{\sigma}^2_X}{n}. \tag{52}$$

The proof is straightforward, we have for the variance of the mean

$$\begin{aligned}
\sigma^2_{\hat{\mu}_X} &= \mathbb{E}[(\hat{\mu}_X - \mu_X)^2] = \mathbb{E}\left[\left(\frac{1}{n}\sum_{j=1}^{n} X_j - \mu_X\right)^2\right] = \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{j=1}^{n}(X_j - \mu_X)\right)^2\right] \\
&= \frac{1}{n^2}\sum_{j=1}^{n}\mathbb{E}\left[(X_j - \mu_X)^2\right] + \frac{2}{n^2}\sum_{i<j}\mathbb{E}\left[(X_i - \mu_X)(X_j - \mu_X)\right] \\
&= \frac{1}{n^2}\sum_{j=1}^{n}\mathbb{E}\left[(X_j - \mu_X)^2\right] = \frac{\sigma^2_X}{n}\ ,
\end{aligned} \tag{53}$$

where we had to assume that the random variables are independent, i.e. $\mathbb{E}[(X_i - \mu_X)(X_j - \mu_X)] = 0$ for $i \neq j$.

1.8.2 Confidence intervals

More generally, we would like to predict confidence intervals, i.e. given a value α we would like to find the interval $[\mu_1, \mu_2]$ such that the probability that the true mean μ lies in the interval $[\mu_1, \mu_2]$ is α. The confidence interval is then given by

$$\alpha = \mathbb{P}[\mu_1 \leq \hat{\mu} \leq \mu_2] = \int_{\mu_1}^{\mu_2} p(\mu)\mathrm{d}\mu, \tag{54}$$

An estimator for confidence interval is then given by

$$\hat{\mu}_1 = \hat{\mu}_X - z_\alpha\sqrt{\frac{\hat{\sigma}^2_X}{n}}, \qquad \hat{\mu_2} = \hat{\mu}_X + z_\alpha\sqrt{\frac{\hat{\sigma}^2_X}{n}}, \tag{55}$$

where z_α is the $(1+\alpha)/2$ quantile of the standard normal distribution. For example, the 95% confidence interval is given by $z_{0.95} \approx 1.96$. The desired value can be obtained using the `scipy.stats.norm.ppf` function, which returns the quantile of the standard normal distribution for a given probability. For example, the following code snippet computes the 95% confidence interval for the mean of the data:

```
>>> from scipy.stats import norm

>>> def calculate_z_alpha(alpha):
      # Calculate the z_{alpha} factor
      return norm.ppf((1 + alpha)/2)
```

Using the Bayesian estimator of means and variances, confidence intervals can be even more conveniently computed using the `scipy.stats.bayes_mvs( data, alpha )` function, which returns the mean, variance and standard deviation of the data along with the confidence intervals. For example, the following code snippet computes the mean, variance and standard deviation of the data along with the 95% confidence intervals (which is the default):

```
>>> from scipy import stats
>>> data = [6, 9, 12, 7, 8, 8, 13]
>>> mean, var, std = stats.bayes_mvs(data, alpha=0.95)
>>> mean
Mean(statistic=9.0, minmax=(7.103650222612533, 10.896349777387467))
>>> var
Variance(statistic=10.0, minmax=(3.176724206..., 24.45910382...))
>>> std
Std_dev(statistic=2.9724954732045084,
      minmax=(1.7823367265645143, 4.945614605014631))
```

1.8.3 Estimating errors using resampling

Typically, we do not have access to an ensemble of realisations of a random process and therefore do not have access to the true distribution of the summary statistic so that we have to resort to either MLE or Bayesian estimators, or we can also employ resampling methods, which have the advantage that they do not require any assumptions about the underlying distribution.

In the absence of a known distribution, we can estimate the error of a summary statistic by resampling the data. The **bootstrap** method is a Monte Carlo method that estimates the error of a summary statistic by resampling the data with replacement. The **jackknife** method is a similar method that estimates the error of a summary statistic by resampling the data without replacement, which we will not discuss here.

Bootstrap resampling. The bootstrap method is a Monte Carlo method that estimates the error of a summary statistic by resampling the data with replacement. Specifically, the bootstrap method estimates the error of a summary statistic by generating B bootstrap samples $\mathcal{S}_b$, $b = 1, \ldots, B$ from the data $\mathcal{S}$, where each bootstrap sample $\mathcal{S}_b = \{X_{i_1}, \ldots, X_{i_m}\}$ is generated by randomly sampling m data points from the full dataset $\mathcal{S} = \{X_1, \ldots, X_n\}$ with replacement. The value of m of the bootstrap samples can be the same as the size of the original dataset, i.e. $m = n$, but can also be chosen to be smaller, e.g. $m = n/2$.

The bootstrap method then estimates the error of a summary statistic by computing first the summary statistic for each bootstrap sample $\mathcal{S}_b$ separately and then considering the sampled distribution of the summary statistic. Specifically, let $\hat{\theta}_b$ be the statistic obtained from the b-th subsample, and $\hat{\theta}$ be either the statistic obtained from the whole sample, or the mean of the $\hat{\theta}_b$. Then the bootstrap method estimates the error of the

summary statistic as

$$\hat{\sigma}_\theta^2 = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b - \hat{\theta})^2 \,. \tag{56}$$

To obtain reliable estimates, the number of bootstrap samples B should be large, typically $B \approx 1000$. The bootstrap method is particularly useful when the underlying distribution is unknown, or when the distribution is known but the summary statistic is not analytically tractable. It is generally advisible to check the actual distribution of the bootstrap samples, and to check for convergence of the bootstrap error estimate with increasing number of bootstrap samples. Bootstrap resampling is implemented in the `scipy.stats.bootstrap( data, statistics, n_resample, confidence_level )` function which performs $B =$ `n_resample` bootstrap resamples of `data` and evaluates the given `statistic` for them. The method returns the $\alpha =$`confidence_level` confidence interval, the bootstrap distribution, and the standard error, see this link for details.

Jackknife resampling. In some cases (such as the spatial statistics we consider below), it is not desirable to resample the data with replacement, as duplicate data is inadmissible. The jackknife is similar to the bootstrap method, but estimates the error of a summary statistic by subsampling the data without replacement. Specifically, the jackknife method estimates the error of a summary statistic by generating n jackknife samples $\mathcal{S}_{(i)}$, $i = 1, \ldots, n$ from the data $\mathcal{S}$, where each jackknife sample $\mathcal{S}_{(i)}$ is generated by removing the i-th data point from the full dataset $\mathcal{S} = \{X_1, \ldots, X_n\}$. Errors on the desired statistics are then computed as in the bootstrap case from the jackknife samples.

An interesting variant is the **delete-d** jackknife. Instead of 1 now d data points are deleted from the dataset and the summary statistic is computed for the reduced dataset. There are now $\binom{n}{d}$ possible ways to delete d data points from the dataset. Again, one can compute the summary statistic for each of these reduced datasets and then compute the variance of the summary statistic. If the computation of the summary statistic is very expensive, one can also compute the summary statistic for a subset of the reduced datasets.

This approach is easily implemented using the `numpy.random.Generator.choice` function, which directly allows to sample without replacement. The following code snippet demonstrates how to use the `numpy.random.Generator.choice` function to generate a jackknife sample of size `size` from the data, more details can be found here:

```
>>> import numpy as np
>>> data = np.array([1, 2, 3, 4, 5])
>>> rng = np.random.default_rng()
>>> jackknife_sample = rng.choice(data, size=3, replace=False)
>>> jackknife_sample
array([3, 1, 5])
```

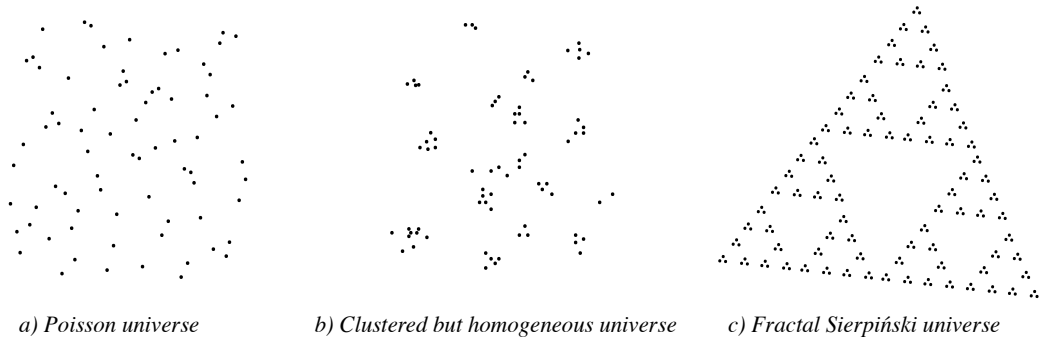

Figure 5: Point sets with different spatial correlations: a) a Poisson universe that is generated by a Poisson process with no correlation structure, b) a clustered Universe where it is more likely to find a point close to another point, but which appears as homogeneous (i.e. uniform) when averaged on scales larger than the clusters and holes, and c) a fractal Sierpiński universe which is also clustered but has no homogeneity scale since it is self-similar. Our own Universe belongs to class b).

1.9 Spatial statistics

Spatial statistics is of particular importance in astrophysics and cosmology as it of the foundation of a quantitative statistical analysis of phenomena that occur in space. Often gravity, due to its long-range interactions, leads to large-scale spatial correlations in systems that are of great interest to understand astrophysical phenomena from star formation in turbulent molecular clouds, to the turbulent intracluster medium of galaxy clusters, to the large-scale matter distribution in the Universe.

Let us assume we have a distribution of point-like objects in space (of any dimension n) – we can think of stars, quasars, or galaxies, e.g. if we like. How can we quantify their distribution? Point-like objects can show different clustering behavior at fixed mean spatial density (see Fig. 5) that can be distinguished e.g. by their N-point functions, as we will see.

1.9.1 Number density fields

We can think of these objects simply as multi-variate random variates $\boldsymbol{X}$ drawn from a random process. The PDF of this process $p_{\boldsymbol{X}}(\boldsymbol{x})$ directly is directly proportional to the number density function

$$n(\boldsymbol{x}) := N\, p_{\boldsymbol{X}}(\boldsymbol{x}), \tag{57}$$

where N is a normalisation constant that yields the expected number of objects in a fixed finite volume V over many realisation of the process. The mean number density of

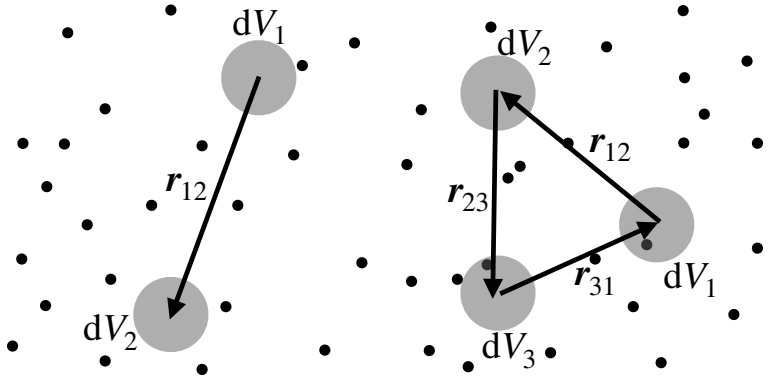

Figure 6: The two-point correlation function (left) quantifies the excess probability (with respect to a Poisson process), to find galaxies in the volume elements $\mathrm{d}V_1$ and $\mathrm{d}V_2$ separated by the vector $\boldsymbol{r}_{12}$. The three-point correlation function (right) generalises this concept to a triangle of volume elements.

objects is then given by

$$\bar{n} := \frac{N}{V}. \tag{58}$$

It is common to split the field $n(\boldsymbol{x})$ into its mean and dimensionless fluctuating part, so that

$$n(\boldsymbol{x}) =: \bar{n}(1 + \delta(\boldsymbol{x})). \tag{59}$$

This allows us to write the probability to find an object in a region R as

$$\mathbb{P}[\boldsymbol{X} \in R] = \int_R \mathrm{d}^n x \; n(\boldsymbol{x}) = \bar{n} \int_R \mathrm{d}^n x \; (1 + \delta(\boldsymbol{x})) \tag{60}$$

Poisson processes. A Poisson process is a random process which has a uniform PDF $p_{\boldsymbol{X}}(\boldsymbol{x}) = \text{const.}$ (for $\boldsymbol{x} \in \Omega$).

1.9.2 Two-point correlation functions

The simplest concept (and yet a state of the art tool in large-scale structure cosmology) of spatial statistics is the two-point correlation function. Let us consider the joint probability to find an object in volume V_1 at the same time as another object in volume V_2. Clearly this is in general a joint probability that we can write as

$$P_{12} := \mathbb{P}[\boldsymbol{X}_1 \in V_1, \boldsymbol{X}_2 \in V_2] = \int_{V_1} \mathrm{d}^n x_1 \int_{V_2} \mathrm{d}^n x_2 \; n^{(2)}(\boldsymbol{x}_1, \boldsymbol{x}_2) \tag{61}$$

Stationary/Homogeneous Processes. We say that the random process that generated the random variates $\boldsymbol{X}$ is **stationary** or **homogeneous** if P_{12} is independent of the explicit position of the volume V_1. This means that there are no regions that are special and the statistics P_{12} is translation-invariant. In this case, the joint density function must also be translation invariant. This implies that

$$n^{(2)}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \lambda(\boldsymbol{x}_2 - \boldsymbol{x}_1) =: \lambda(\boldsymbol{r}_{12}), \tag{62}$$

i.e. it can only depend on the modulus of the relative vector $\boldsymbol{r}_{12}$.

Isotropic Processes. If the random process that generates $\boldsymbol{X}$ is statistically isotropic about point $\boldsymbol{x}_1$, i.e. it has no preferred directions, then the summary statistic must be rotation invariant around $\boldsymbol{x}_1$, i.e. it can only depend on the distance of $\boldsymbol{x}_2$ from $\boldsymbol{x}_1$ not the direction. In this case, we have that

$$n^{(2)}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \kappa(\boldsymbol{x}_1, \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|) =: \kappa(\boldsymbol{x}_1, r_{12}), \tag{63}$$

where we have defined the distance $r_{12} := \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|$.

Stationary and isotropic processes. If the random process that generates $\boldsymbol{X}$ is both stationary and isotropic, then the joint density function must be a function of the distance only, i.e. we have

$$n^{(2)}(\boldsymbol{x}_1, \boldsymbol{x}_2) = C(r_{12}). \tag{64}$$

On large scales, in the Universe, we are often interested in stationary and isotropic correlators, since by the cosmological principle, on the largest scales, the Universe is statistically homogeneous (another word for stationary) and isotropic.

Poisson processes. Poisson processes are trivially homogeneous and isotropic since we have simply

$$n^{(2)}(\boldsymbol{x}_1, \boldsymbol{x}_2) = n(\boldsymbol{x}_1)\, n(\boldsymbol{x}_2) = \bar{n}^2 \tag{65}$$

if self-pairs are neglected.

The two-point correlation function. The two-point correlation function is commonly defined so that it reflects the excess clustering of objects over a Poisson process. We therefore write

$$\text{in the homogeneous anisotropic case:} \quad n^{(2)}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \bar{n}^2\,[1 + \xi_{\text{aniso}}(\boldsymbol{r}_{12})] \tag{66a}$$

$$\text{in the homogeneous isotropic case:} \quad n^{(2)}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \bar{n}^2\,[1 + \xi_{\text{iso}}(r_{12})] \tag{66b}$$

In the theory of random fields, it is not hard to show that the two-point function is essentially the infinite dimensional covariance matrix, i.e. we can think of it as the correspondence

$$\Sigma_{ij} \to n^{(2)}(\boldsymbol{x}_1, \boldsymbol{x}_2) \tag{67}$$

where each index i and j now becomes a spatial point $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Naturally, this implies that a Gaussian random field (i.e. a field whose $n(\boldsymbol{x})$ at every point follows a normal distribution) is fully described by the value of its mean and the values of its two-point function.

1.9.3 N-point correlation functions

Just like non-Gaussian distributions require higher order cumulants, the concept of the correlation function can be extended to include more vertices. The structure of the higher order correlation functions is quite analogous to the moment/cumulant constructions we encountered before (see also Figs. 4 and 6).

For the **three-point correlation function** one finds

$$n^{(3)}(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) = \bar{n}^3 \left(1 + \xi(r_{12}) + \xi(r_{23}) + \xi(r_{13}) + \zeta(r_{12}, r_{23}, r_{13})\right) \tag{68}$$

where $\zeta(a, b, c)$ is called the connected three-point correlation function. Note that the three vectors must form a triangle, otherwise it does not define a three-point structure (and its expectation value vanishes). Higher orders can be easily constructed from the diagrams in Fig. 4.

1.9.4 Estimators for the two-point correlation function

The two-point correlation function can be easily computed in some cases using the discrete Fourier transform (if the sample space Ω is an n-torus). In more general cases, other estimators have to be used. Monte-Carlo integration is most commonly used in these cases when Ω has a non-trivial structure. For a galaxy survey like the SDSS-BOSS survey (see Fig.7), the sampling space not only has a non-trivial shape on the sky, but it also has varying completeness over the survey area (meaning that the probability that a galaxy is included in the survey changes over the sky).

In this case, it is much easier to employ Monte-Carlo integration. Let us assume $D = \{\boldsymbol{X}_j\}_{j=1,\dots,N}$ are the galaxy positions, and we have also generated a Poisson-like sample of random points $R = \{\boldsymbol{Y}_j\}_{j=1,\dots,M}$ that however reflects the exact extent of the sample space Ω with its possibly non-uniform selection function. Define also

$$DD(r) = \#\text{ pairs } (\boldsymbol{x} \in D, \boldsymbol{y} \in D) \text{ with } r - \Delta r \leq \|\boldsymbol{x} - \boldsymbol{y}\| < r + \Delta r \tag{69a}$$

$$RR(r) = \#\text{ pairs } (\boldsymbol{x} \in R, \boldsymbol{y} \in R) \text{ with } r - \Delta r \leq \|\boldsymbol{x} - \boldsymbol{y}\| < r + \Delta r \tag{69b}$$

$$DR(r) = \#\text{ pairs } (\boldsymbol{x} \in D, \boldsymbol{y} \in R) \text{ with } r - \Delta r \leq \|\boldsymbol{x} - \boldsymbol{y}\| < r + \Delta r \tag{69c}$$

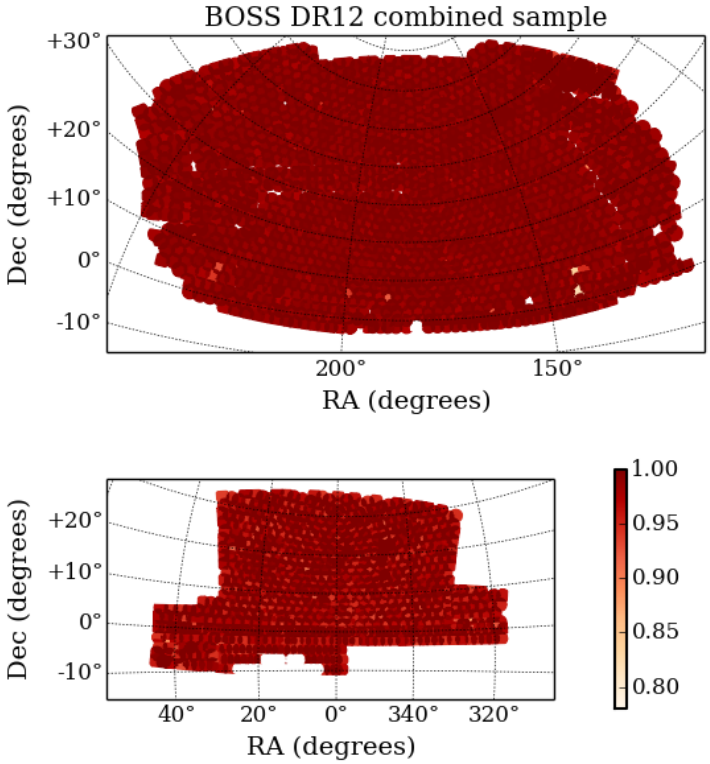

Figure 7: Angular mask (right ascension RA vs declination DEC), ranging from zero to one, showing the completeness on the sky of the SDSS-III BOSS DR12 survey (northern part; from Alam et al. 2016).

A simple estimator for the two-point correlation function (cf. Peebles & Hauser 1974) can be obtained as

$$\hat{\xi}_{PH}(r) = \left(\frac{M}{N}\right)^2 \frac{DD(r)}{RR(r)} - 1, \tag{70}$$

i.e. the number of pairs in the data, relative to the number of pairs in the random test sample. This estimator has however long been shown to be inaccurate for complex survey geometries. Instead, a very robust estimator is the Landy & Szalay (1993) estimator

$$\hat{\xi}_{LS}(r) := 1 + \left(\frac{M}{N}\right)^2 \frac{DD(r)}{RR(r)} - 2\frac{M}{N}\frac{DR(r)}{RR(r)}. \tag{71}$$

Typically the random data requires a larger number of points than the data set to get converged results ($M \sim 10N$).

A simple optimization of this method is to not double-count the pairs. When only correlating unique pairs, the estimator becomes

$$\hat{\xi}_{LS}(r) := 1 + \frac{M(M-1)}{N(N-1)} \frac{\widehat{DD}(r)}{RR(r)} - \frac{M(M-1)}{NM} \frac{\widehat{DR}(r)}{\widehat{RR}(r)} \tag{72}$$

where now $\widehat{DD}, \widehat{DR}, \widehat{RR}$ refelect the number of unique pairs.

Note that the LS estimator can effectively be derived as $\left(\frac{D}{R} - 1\right)^2 = \frac{DD-2DR+RR}{RR}$.

The variance of this estimator can be crudely estimated as

$$\hat{\sigma}^2_{\hat{\xi}_{LS}}(r) \approx \frac{1+\xi(r)}{DD(r)} \approx \frac{1}{DD(r)} \, . \tag{73}$$

A more in-depth discussion of the two-point correlation function and its estimators can be found e.g. in this master thesis by Marian Biermann. A proposal for more optimal estimators can be found in this paper by Smith & Marian 2015.

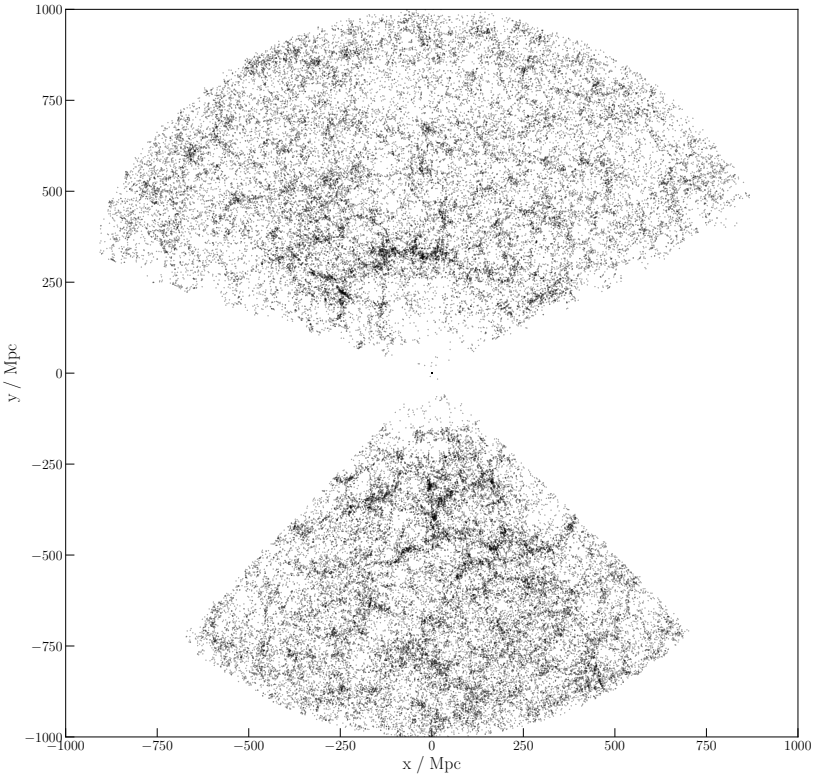

Figure 8: The distribution of galaxies in the SDSS survey in physical space.