This is a draft version of the lecture notes. We aim to keep improving it but at the current stage it is most likely far from perfect. Please contact us if you notice any typos, errors, subtle points, or if you have any questions or suggestions for improvements.

`maciej.lisicki[at]fuw.edu.pl`
`milosz.panfil[at]fuw.edu.pl`

Maciej Lisicki & Miłosz Panfil
October 17, 2022

2 Continuous probability distributions

2.1 Energy distribution of the sun's radiation

It's often convenient to work with a continuous (uncountable) sample space and continuous probability distributions. We can then use tools of mathematical analysis to extract information from them.

To understand the meaning behind the uncountable sample space let's look at the energy distribution of the sun's radiation.

Let's say we want to learn something about the Sun. The Sun emits electromagnetic waves. The electromagnetic waves are characterized by wavelengths (or frequency) and we would like to know how much energy is carried in different regions of the spectrum. So we go out to the space to get rid of the spurious effects of the atmosphere and set out the measurements.

Imagine we have a spectrometer in which we set the wavelength and then read out the energy of the radiation received per second and per area (so W/m^2). The spectrometer has a certain resolution, we can set the wavelength to multiplicities of $\Delta\lambda$. Therefore, taking a measurement we are really learning about the total power P_i carried by waves of wavelengths from $\lambda_i - \Delta\lambda/2$ to $\lambda_i + \Delta\lambda/2$.

After performing the measurements through the whole range of wavelength we know the total power P. We can then define

$$p_i = P_i/P, \tag{2.1}$$

which is the fraction of the power carried by the wavelength from $\lambda_i - \Delta\lambda/2$ to $\lambda_i + \Delta\lambda/2$. We can now reverse the thinking. If we were able to randomly observe an electromagnetic radiation emitted by the Sun, what would be a probability that its wavelength falls within $\lambda_i - \Delta\lambda/2$ to $\lambda_i + \Delta\lambda/2$. That would be exactly p_i.

Imagine now that we got a better spectrometer that allowed us to reduce the resolution to $\Delta\lambda' = \Delta\lambda/10$. Repeating the measurements and the procedure we would get numbers p'_i. And if everything is coherent we should get

$$p_i = \sum_j p'_j \tag{2.2}$$

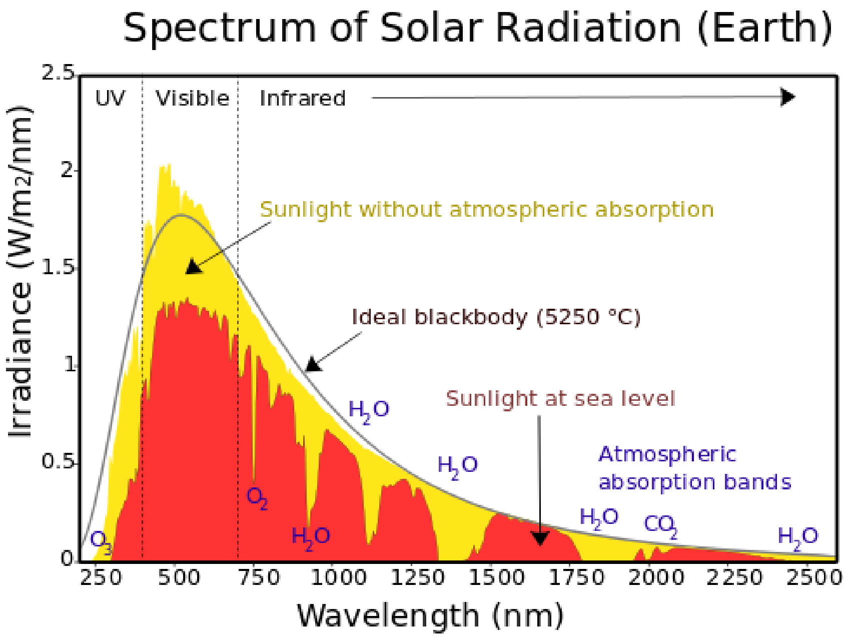

Figure 1. Solar spectrum

In the limiting procedure of the resolution $\Delta\lambda' \to 0$ we then get

$$p_i = \int_{\lambda_i - \Delta\lambda}^{\lambda_i + \Delta\lambda} p(\lambda) \mathrm{d}\lambda. \tag{2.3}$$

Essentially we can think of the probability distribution as a probability per the wavelength,

$$p(\lambda_i) \approx \frac{p_i}{\Delta\lambda}. \tag{2.4}$$

Note that, unlike discrete probabilities, probability distribution can have dimensions.

For the Sun's radiation the distribution function is very well approximated by the Planck's distribution

$$B(\lambda, T) = \frac{2hc^2}{\lambda^5} \frac{1}{e^{hc/(\lambda k_B T)} - 1}. \tag{2.5}$$

To get a probability distribution, this function needs to be normalized,

$$p(\lambda, T) = \frac{1}{N} \int_0^\infty \mathrm{d}\lambda B(\lambda, T) \tag{2.6}$$

where the normalization constant is choosen such that

$$\int \mathrm{d}\lambda \, p(\lambda, T) = 1 \tag{2.7}$$

The meaning on the formula $p(\lambda, T)$ is the following. The probability of observing a radiation of wavelengths within $\lambda, \lambda + \mathrm{d}\lambda$ given that the temperature of the Sun is T is $p(\lambda, T)\mathrm{d}\lambda$. We can use then our measurement to infer about the temperature of the Sun.

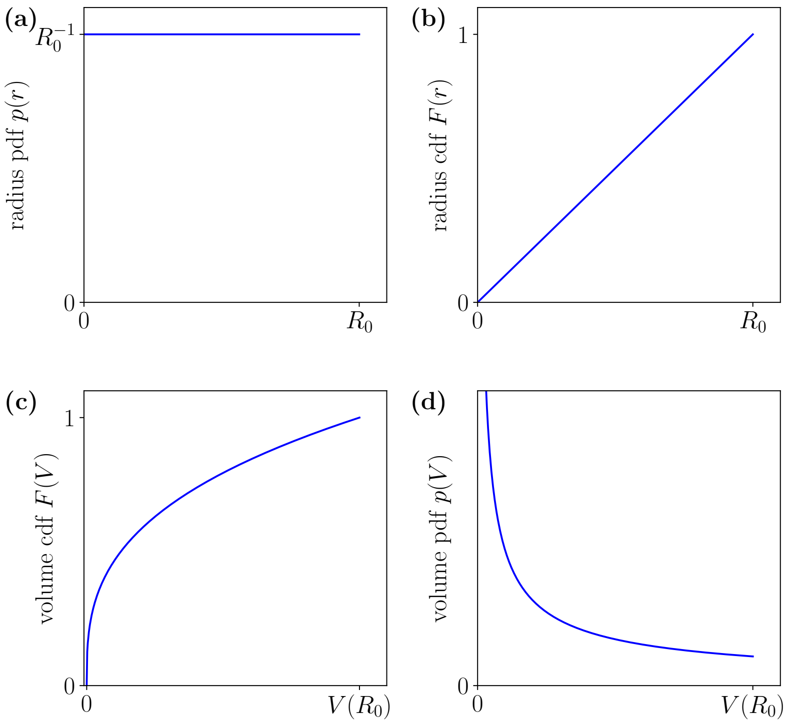

Figure 2. Transformations between probability density and cumulative distribution functions for the uniform distribution of rainfall droplets radii vs. the non-uniform corresponding distribution of volumes.

2.2 Continuous probability distributions

In the preceding section we have defined a new object – a continuous probability distribution $p(x)$, sometimes called the probability density or probability distribution function (PDF). By analogy to the discrete case, we can readily write some properties of the PDF. The completeness of probabilities means that

$$\int_\Omega p(x)\mathrm{d}x = 1. \tag{2.8}$$

The mean and variance are again defined in the standard way, as

$$\mu = \langle X \rangle = \int xp(x)\mathrm{d}x, \tag{2.9}$$

$$\sigma^2 = \langle (X - \langle X \rangle)^2 \rangle = \int (x - \langle x \rangle)^2 p(x)\mathrm{d}x. \tag{2.10}$$

2.3 Functions of random variables and the rainfall droplets distribution

The example of the laser pointer involves a transformation of a random variable, expressed by Eq. (2.22). In this case, a useful mnemonic rule is that for a continous PDF, events in corresponding intervals of the sample space have to have equal probabilities. The probability that a random variable X with a distribution $\rho_X(x)$ takes the value between x and $x + \mathrm{d}x$ is simply $\rho_X(x)\mathrm{d}x$. Now, if we are interested in the PDF of a transformed variable,

$Y = f(X)$, the value of x corresponds to y and the probability of the random variable Y taking the value in the close interval of with $\mathrm{d}y$ is $\rho_Y(y)\mathrm{d}y$. The mentioned equality of probabilities yields

$$\rho_Y(y) = \rho_X(x)\frac{\mathrm{d}x}{\mathrm{d}y}. \tag{2.11}$$

In this way, the distribution function for the new, transformed variable y can be readily written. An alternative way involves the use of the *cumulative distribution function* (CDF) F. It describes the probability that a random variable X takes the value less than x,

$$F_X(x) = P(X \leq x). \tag{2.12}$$

The probability that the random variable x takes the value in the interval $[a, b]$ is then readily calculated as

$$P(a \leq X \leq b) = F_X(b) - F_X(a), \tag{2.13}$$

from which we can deduce the relationship of the CDF with the PDF ρ_X from the fundamental theorem of calculus

$$\rho_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}. \tag{2.14}$$

Take an example of the rainfall droplets distribution. If we assume that the probability of finding a droplet of a given radius r is uniform on the interval $[0, R_0]$, what is the distribution of droplet volumes V?

The uniform distribution of droplet radii is given by

$$p(r) = \begin{cases} 1/R_0 & \text{for } r \in [0, R_0], \\ 0 & \text{for } r > R_0. \end{cases} \tag{2.15}$$

Then, the cumulative distribution function for the droplet radius is found as

$$F(r) = \int_0^r p(s)\mathrm{d}s = \frac{r}{R_0}, \tag{2.16}$$

from which the cdf for the volume $V = 4\pi r^3/3$ is readily found as

$$F(V) = \frac{1}{R_0}\left(\frac{3V}{4\pi}\right)^{1/3}, \tag{2.17}$$

and, finally, we find the PDF for the volume by differentiation $p(V) = F'(V)$, as

$$p(V) = \begin{cases} \frac{1}{3R_0}\left(\frac{3}{4\pi}\right)^{1/3} V^{-2/3} & \text{for } V \in \left[0, \frac{4}{3}\pi R_0^3\right], \\ 0 & \text{otherwise}. \end{cases} \tag{2.18}$$

We note here that the distribution of volumes is not at all uniform. Alternatively, Eq. (2.18) is obtained directly from Eq. (2.11) using the transformation $y = \frac{4}{3}\pi x^3$.

2.4 Laser pointer and the Cauchy distribution

Imagine yourself standing in front of a long straight wall, holding a laser gun (or a tiny laser pointer). You spin at random and whenever you stop, the beam points at a point x on the wall. What is the probability distribution of the position X of the illuminated point?

To answer this question, we need to relate the observable (the position x being the value of the random variable X) with the (known) distribution of the angle ϕ at which we stop spinning. If we are at a distance L away from the wall, the two quantities are related by a functional relationship

$$x = \tan\phi, \tag{2.19}$$

which implies that the underlying random variables $-\infty < X < \infty$ and $-\pi/2 < \Phi < \pi/2$ are functionally related

$$X = \tan\Phi. \tag{2.20}$$

Let us say probability distribution of Φ is $p_\Phi(\phi)$. What is the probability distribution $p_X(x)$ of X?

To answer this question we recall the relation between the probability density and actual probability. Namely, the relevant quantity is $p(x)\mathrm{d}x$ which we interpret as probability that the random variable X takes value between x and $x + \mathrm{d}x$. This quantity is "parametrisation independent". If X is functionally related to another random variable, like in (2.20), the same quantity $p(x)\mathrm{d}x$ can expressed in terms of the new variable

$$p_X(x)\mathrm{d}x = p_\Phi(\phi)\mathrm{d}\phi. \tag{2.21}$$

This can be turned into a rule for transformation of probability densities

$$p_X(x) = p_\Phi(\phi(x))\frac{\mathrm{d}\phi}{\mathrm{d}x}. \tag{2.22}$$

How does it work for the spinning laser? For $-\pi/2 < \phi < \pi/2$ the relation $x = \tan\phi$ can be inverted, $\phi = \operatorname{atan} x$, and

$$p_X(x) = \frac{p_\Phi(\operatorname{atan} x)}{1+x^2}.$$

If the probability distribution of angle ϕ is uniform, the resulting distribution is the *Cauchy distribution*

$$p_X(x) = \frac{1}{\pi}\frac{1}{1+x^2}. \tag{2.23}$$

2.5 Bertrand's paradox

It is often said that in the absence of relevant arguments, we should prescribe uniform probability to possible outcomes - *principle of indifference*. If possible outcomes form a continuous set this prescription is not straightforward because it relies on choosing *what is an independent variable.*

We first consider a discrete random variable, say a result of throwing a dice. Following the principle of indifference we assume that all possible results are equally probable,

$p_i = 1/6$ for $i = 1, \dots, 6$. Consider now square of the result. The possible outcomes are $1, 4, \dots, 36$, are again equally probable. This is the same result as if we applied the principle of indifference to the squares. For discrete random variables the principle of indifference is consistent with changing variables and therefore make sense. What is random does not depend on the label we associate to it. Things change for continuous variables.

A simple illustration of the difficulty with specifying the word random for continuous variables is the Bertrand's paradox https://en.wikipedia.org/wiki/Bertrand_paradox_(probability)

2.6 Multivariate probability distributions

Suppose we are measuring the outcome of an experiment by recording two separate quantities that describe it, say X_1 and X_2. They might be similar in nature, e.g. fluctuating currents in an electric system, or completely different, like the position of a tracer particle and its temperature. However, they are both random variables, i.e. their values are not deterministic. A priori we do not know whether they are related in any. The relevant statistical description of the system involves then a two-dimensional probability distribution function $P_2(x, y)$, which has the natural interpretation that

$$P_2(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2$$

is the probability that the random variables X and Y have values in the range $[x_1, x_1+\mathrm{d}x_1]$ and $[x_2, x_2 + \mathrm{d}x_2]$, respectively. The normalisation of this distribution reads

$$\iint P_2(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 = 1. \tag{2.24}$$

We can recover the familiar one-dimensional probability distribution by integrating out the second degree of freedom,

$$P_1(x_1) = \int P_2(x_1, x_2)\mathrm{d}x_2, \qquad P_1(x_2) = \int P_2(x_1, x_2)\mathrm{d}x_1, \tag{2.25}$$

called the *marginal distributions*. Importantly, we consider two random variables to be statistically independent if their joint PDF factorises

$$P_2(x_1, x_2) = P_1(x_1)P_1(x_2) \tag{2.26}$$

To characterise a two-dimensional distribution, we can calculate the average values

$$\langle X_1 \rangle = \iint x_1 P_2(x_1, x_2)\mathrm{d}x_2\mathrm{d}x_1 = \int x_1 P_1(x_1)\mathrm{d}x_1, \tag{2.27}$$

and similarly for $\langle X_2 \rangle$. Instead of variance for a single variable, we now define the covariance

$$\mathrm{Cov}(X_1, X_2) = \langle X_1 X_2 \rangle - \langle X_1 \rangle \langle X_2 \rangle. \tag{2.28}$$

In a similar way, higher-order probability distributions, P_N are constructed, for N-dimensional random variables forming a random vector $\boldsymbol{X} = (X_1, X_2, \cdots, X_N)$. The average becomes then an N-dimensional vector, and the rank of the covariance matrix is $N \times N$.

2.7 Independent random variables and the Maxwell distribution

Two random variables X_1 and X_2 are called independent if their joint probability distribution factorises into one-dimensional distributions

$$P_2(x_1, x_2) = P_1(x_1)P_2(x_2). \tag{2.29}$$

Note that in this case the covariance vanishes, $\mathrm{Cov}(X_1, X_2) = 0$. Random variables with a vanishing covariance are called *uncorrelated*, but independence is a stronger property—it implies the lack of correlation but not the other way around. To illustrate the usefulness of independence, we will derive the Maxwell distribution of the velocities of particles in a gas of non-interacting particles at temperature T.

James Clerk Maxwell solved this problem in 1860 using very minimal assumptions on the symmetries of the joint distribution of the three Cartesian velocity components. He assumed that the distributions for each component should be independent and identical. Thus we can factorise the joint probability distribution $P_3(v_x, v_y, v_z)$ as

$$P_3(v_x, v_y, v_z) = f(v_x)f(v_y)f(v_z). \tag{2.30}$$

He also noted that a rotation of the reference frame cannot change the distribution, so it can only depend on the magnitude of the velocity vector

$$P_3(v_x, v_y, v_z) = \Psi(v_x^2 + v_y^2 + v_z^2). \tag{2.31}$$

Given that the system should be isotropic, for simplicity we now consider a point $v_y = v_z = 0$. Assuming $f(0) = c$, we have

$$c^2 f(v_x) = \Psi(v_x^2), \tag{2.32}$$

and repeating the same reasoning we get

$$\Psi(v_x^2)\Psi(v_y^2)\Psi(v_z^2) = c^6\Psi(v_x^2 + v_y^2 + v_z^2). \tag{2.33}$$

Changing variables to $\alpha = v_x^2$, $\beta = v_y^2$, $\gamma = v_z^2$, we rewrite this equation as

$$\Psi(\alpha)\Psi(\beta)\Psi(\gamma) = c^6\Psi(\alpha + \beta + \gamma). \tag{2.34}$$

We now take a derivative with respect to α, finding

$$\Psi'(\alpha)\Psi(\beta)\Psi(\gamma) = c^6\Psi'(\alpha + \beta + \gamma). \tag{2.35}$$

Dividing by Eq. (2.34), we arrive at

$$\frac{\Psi'(\alpha)}{\Psi(\alpha)} = \frac{\Psi'(\alpha + \beta + \gamma)}{\Psi(\alpha + \beta + \gamma)}. \tag{2.36}$$

Now, taking $\alpha = \gamma = 0$, we get an equation

$$\frac{\Psi'(\beta)}{\Psi(\beta)} = \frac{\Psi'(0)}{\Psi(0)} = \text{const} = -\lambda, \tag{2.37}$$

which can be readily integrated to yield

$$\Psi(\beta) = A\exp(-\lambda\beta), \tag{2.38}$$

so for the general case we have

$$\Psi(\alpha+\beta+\gamma) = B\exp\left[-\lambda(\alpha+\beta+\gamma)\right], \tag{2.39}$$

where the constant $B = (\lambda/\pi)^{3/2}$ is found from the normalisation condition using

$$\int e^{-\lambda v_x^2}\mathrm{d}v_x = \sqrt{\frac{\pi}{\lambda}}. \tag{2.40}$$

Finally, the constant λ can be interpreted as the mean kinetic energy of the particles by $\frac{1}{2}m\langle v^2\rangle = \frac{3m}{4\lambda}$. It can be related to the temperature by equipartition principle, $\frac{1}{2}m\langle v^2\rangle = \frac{3}{2}k_BT$, so that $\lambda = m/2k_BT$.

We conclude that the resulting probability distribution for each Cartesian component of the velocity is Gaussian,

$$f(v_x) = \sqrt{\frac{m}{2\pi k_BT}}\exp\left(-\frac{mv^2}{k_BT}\right), \tag{2.41}$$

a distribution that emerges in many physical contexts naturally, as we shall see in the next section.

2.8 Functions of multivariate random variables

Suppose we have two random variables, X_1 and X_1, with known distributions. What is the distribution of the variable $Y = X_1 + X_2$? The methodology for transforming PDFs, presented in Sec. 2.3 works well only is the dimensionality of the initial and final variables are the same. If this is not the case, we need a more general way.

To illustrate this, let us define multivariate random vectors $\boldsymbol{X} = (X_1,\ldots,X_r)$ and $\boldsymbol{Y} = (Y_1,\ldots,Y_s)$ with the PDFs $\rho_{\boldsymbol{X}}(x_1,\ldots,x_r)$ and $\rho_{\boldsymbol{Y}}(y_1,\ldots,y_s)$, respectively. Assume that they are functionally related, so that

$$\boldsymbol{Y} = \boldsymbol{f}(\boldsymbol{X}) = (f_1(X_1,\ldots,X_r),\ldots,f_s(X_1,\ldots,X_r)). \tag{2.42}$$

In the equidimensional case of $r = s$, it is straightforward to generalise the results of Sec. 2.3 to find

$$\rho_{\boldsymbol{X}}(x_1,\ldots,x_r)\mathrm{d}x_1\ldots\mathrm{d}x_r = \rho_{\boldsymbol{Y}}(y_1,\ldots,y_r)\mathrm{d}y_1\ldots\mathrm{d}y_r. \tag{2.43}$$

When $r \neq s$, however, the solution can be formally written as

$$\rho_{\boldsymbol{Y}}(\boldsymbol{Y}) = \langle\delta(\boldsymbol{f}(\boldsymbol{X}) - \boldsymbol{Y})\rangle_{\boldsymbol{X}}\,, \tag{2.44}$$

which can be regarded as an average with respect to the random variable $\boldsymbol{X}$ of a Dirac delta function centred at such values, that correspond to $\boldsymbol{f}(\boldsymbol{X})$. In other words, we only select such realisations of the random variable $\boldsymbol{X}$ that the values of $\boldsymbol{f}(\boldsymbol{X})$ correspond to $\boldsymbol{Y}$ and average out. We can write the resulting PDF for $\boldsymbol{Y}$ more explicitly as

$$\rho_{\boldsymbol{Y}}(\boldsymbol{Y}) = \int \mathrm{d}\boldsymbol{X}\delta(\boldsymbol{f}(\boldsymbol{X}) - \boldsymbol{Y}) \tag{2.45}$$

$$= \int \delta(f_1(x_1, \ldots, x_r) - y_1) \cdots \delta(f_s(x_1, \ldots, x_r) - y_s)\mathrm{d}x_1 \ldots \mathrm{d}x_r. \tag{2.46}$$

This formulation is general, but rather abstract. We therefore will restrict our attention to an an illustrative example mentioned in the introductory paragraph. The initial set of random variables $\boldsymbol{X} = (X_1, X_2)$ is now two-dimensional, and the result is a one-dimensional random variable Y. For the recipe mentioned above, our function now reads $f(X_1, X_2) = X_1 + X_2$. The sought for distribution of the sum, by virtue of Eq. (2.44), reads

$$\rho_Y(y) = \langle \delta(x_1 + x_2 - y) \rangle_{\boldsymbol{X}} = \iint \mathrm{d}x_1 \mathrm{d}x_2 \delta(x_1 + x_2 - y) \rho_{\boldsymbol{X}(x_1, x_2) = \int \mathrm{d}x_1 \rho_{\boldsymbol{X}(x_1, y - x_1)},} \tag{2.47}$$

where in the last equality we performed the integral over x_2 with the Dirac delta function. To transform this equation further, we need additional assumptions. If the random variables X_1 and X_2 are independent, then we have $\rho(x_1, x_2) = \rho_1(x_1)\rho_2(x_2)$. Then, Eq. (2.47) becomes

$$\rho_Y(y) = \int \rho_1(x_1)\rho_2(y - x_1)\mathrm{d}x_1, \tag{2.48}$$

so the PDF for the sum of two independent random variables becomes a convolution.

One can show that the average of the sum is the sum of averages

$$\langle Y \rangle = \langle X_1 \rangle + \langle X_2 \rangle \,, \tag{2.49}$$

regardless of whether X_1 and X_2 are independent or not. The second rule is that if X_1 and X_2 are uncorrelated, then the variance follows

$$\sigma_Y^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2. \tag{2.50}$$

2.9 Central limit theorem and the Gaussian distribution

Say we throw a dice 2 times and sum the result. What's the probability of different outcomes? Assuming that each result of a single throw is equally probable we can make the following table We see that from uniform probability we got probability peaked at the

result	2	3	4	5	6	7	8	9	10	11	12
how many ways	1	2	3	4	5	6	5	4	3	2	1
probability	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	...				

Table 2. Possible combinations of genes in an offspring

central result. Summing more numbers we find the following distributions

The central limit theorem states that the sum of many random variables approaches a universal function irrespective of the details of distribution of a single variable, as long as

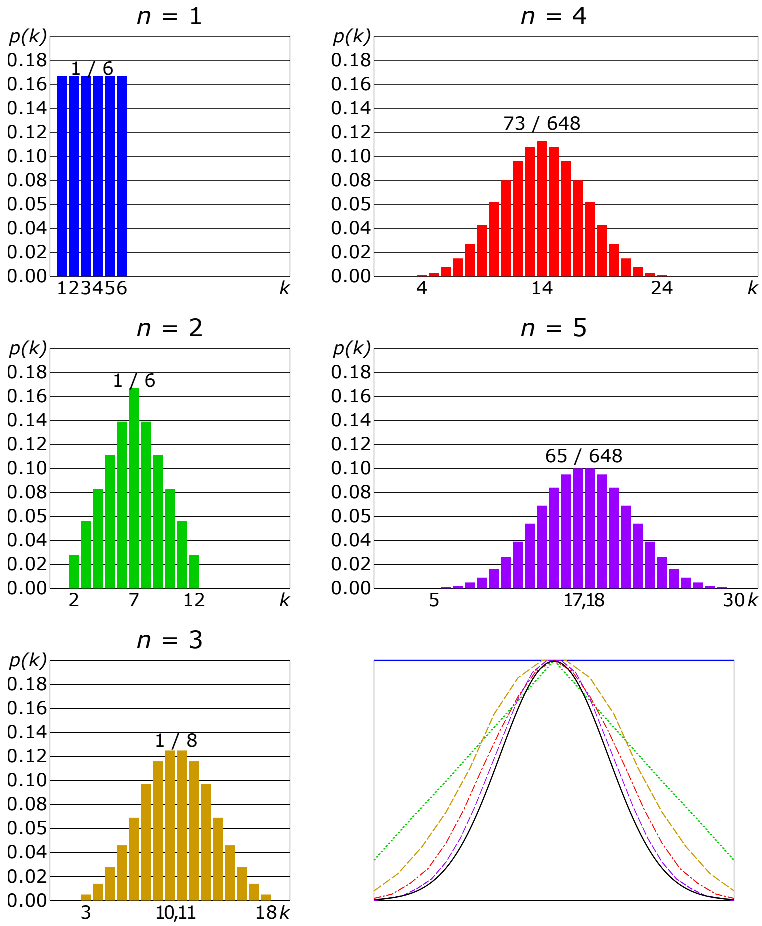

Figure 3. Illustration of the central limit theorem. By Cmglee - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=18918612.

variables are independent. More formally, let us say that variables X_i follow a distribution with a mean mu and variance σ^2. Then random variable $Y_n = \sqrt{n}(\bar{X}_n - \mu)$, where

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{2.51}$$

in the limit of $n \to \infty$ has a Gaussian distribution centered around 0 and with variance σ^2, namely

$$\lim_{n\to\infty} p(Y_n) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}Y_n^2/\sigma^2} \tag{2.52}$$

2.10 Benford's first-digit law

Frank Benford 1938 [1]

https://en.wikipedia.org/wiki/Benford's_law, https://pdodds.w3.uvm.edu/files/papers/others/1881/newcomb1881a.pdf Simon Newcomb [2]

2.11 Buffon's needle experiment

George-Louis Leclerc, Comte de Buffon (1707-1788), a French naturalist and mathematician, posed the following problem in 1733 [3]

> Suppose that we have a floor made of parallel strips of wood, each of the same width, and we drop a needle on the floor. What is the probability that the needle will lie across a line between two strips?

Rephrasing mathematically, we need to find the probability that a needle of length ℓ dropped on a plane ruled d units apart will land across a line.

Solution by geometric probability The easiest way to see the solution for a short needle, with $\ell < d$, uses geometric probability and was proposed by Joseph-Émile Barbier in 1860 [4]. Suppose that the needle makes an acute angle θ with the horizontal axis upon landing on the xy plane. The distance from the centre of the needle to the nearest horizontal line is x. The needle lies across a line only if $x < \ell \cos\theta/2$. We assume now that the values of x and θ are randomly selected when the needle lands. Since $0 < x < d/2$ and $0 < \theta < \pi/2$, the sample space for (x, θ) is a rectangle with dimensions $\frac{d}{2} \times \frac{\pi}{2}$.

The probability that the needle crosses the nearest line is the fraction of the sample space that intersects with $x \leq \ell \cos\theta/2$ for all possible angles. The area of the entire sample space is

$$A_{\text{sample}} = \frac{\pi D}{4}, \tag{2.53}$$

while the area covered by all the intersecting configurations is found by direct integration

$$A_{\text{event}} = \int_0^{\pi/2} \frac{\ell}{2} \cos\theta \mathrm{d}\theta = \frac{\ell}{2}. \tag{2.54}$$

The probability reads thus

$$P = \frac{A_{\text{event}}}{A_{\text{sample}}} = \frac{2\ell}{\pi d}. \tag{2.55}$$

There are many ways to find this result. In the following, we present a more formal alternative.

Solution by a 2D PDF We first identify our random variables in the problem: it is a pair (X, Θ), the distance to the closest line and the angle the needle makes with the line. In each experiment, a pair of results (x, θ) is randomly selected when the needle falls onto the floor. Such pairs create the joint sample space, in which we have a joint probability distribution $P_{X,\Theta}(x, \theta)$. We now assume that these random variables are independent, so we can write

$$P_{X,\Theta}(x, \theta) = P_X(x) P_\Theta(\theta). \tag{2.56}$$

The next assumption is that both variables are uniformly distributed, i.e. the probability of getting any value of x and θ is the same throughout their ranges, so

$$P_X(x) = \frac{2}{d}, \qquad P_\Theta(\theta) = \frac{2}{\pi}. \tag{2.57}$$

We want to find the probability that the needle crosses a line when it falls. This means that for a given θ, we need to be sufficiently close to the nearest line, so that $x < \ell \cos\theta/2$, as before. We just need to count in all the cases that satisfy this condition. Hence

$$P = \int_0^{\pi/2} \mathrm{d}\theta \int_0^{\ell\cos\theta/2} \mathrm{d}x \frac{4}{\pi d} = \frac{2\ell}{\pi d} \int_0^{\pi/2} \cos\theta \mathrm{d}\theta = \frac{2\ell}{\pi d}. \tag{2.58}$$

If the needle is longer than the distance between the lines, the calculation is more subtle and the resulting probability reads

$$P = \frac{2\ell}{\pi d} - \frac{2}{\pi d}\left[\sqrt{\ell^2 - d^2} + d \arcsin\frac{d}{\ell}\right] + 1. \tag{2.59}$$

Let us now focus on short needles. Eq. (2.55) can be rearranged to

$$\pi = \frac{2\ell}{Pd}. \tag{2.60}$$

In 1812 Laplace suggested that if we run Buffon's experiment multiple times to find P, we will be able to estimate the value of π. Suppose that we now drop N needles and M out of them land crossing a line. Then we can estimate $P \approx M/N$ and

$$\pi \approx \frac{2\ell N}{Md}. \tag{2.61}$$

An Italian mathematician, Mario Lazzarini, set out to this task in 1901 [5]. In his study, he took $\ell = 2.5$ cm, $d = 3$ cm, and found $M = 1808$ for $N = 3408$. Thus, our estimation of π is

$$\hat{\pi} = \frac{2 \times 2.5 \times 3408}{3 \times 1808} = 3.1415926... \tag{2.62}$$

An astonishing accuracy of six digits! However, later readers felt uneasy about this result and examined it critically [6].

First, why did Lazzarini measure 3408 tosses and not a nicer number? If we increased or decreased this number by one, to 3407 or 3409, the accuracy of the estimation is much worse, 3.1433... and 3.1398..., respectively; was Lazzarini extremely lucky? Or did he stop when he reached his goal?

Second, we may use the tools of simple statistical analysis to examine the repeated trials experiment, as discussed by Badger [6]. If we want to be 95% confident that our estimation of π has an accuracy of six digits, which means that $|\pi - \hat{\pi} < 0.5 \times 10^{-6}$, how many throws should we count? In other words, we are looking for N such that

$$P\left(\left|\pi - \frac{5N}{3M}\right| < \epsilon\right) \geq 0.95, \tag{2.63}$$

where $\epsilon = 0.5 \times 10^{-6}$. The condition $|\pi - 5N/3M| < \epsilon$ can be approximated by $|M - 5N/3\pi| < 5N\epsilon/(3\pi^2)$. Now, M is binomially distributed with parameters N and $p = 5/3\pi$, so its expectation is Np and its variance is $Np(1-p)$. Using the normal approximation, to have

$$P\left(|Np - M| < \frac{5N\epsilon}{3\pi^2}\right) = 0.95, \tag{2.64}$$

we would need

$$\frac{5N\epsilon}{3\pi^2} \approx 1.95Np(1-p), \tag{2.65}$$

which translates to $N \approx 134 \times 10^{12}$, so Lazzarini was a bit short of this number!

Third, by carefully looking at the measured lengths of the needle and the grid, we must conclude that they are known up to a certain accuracy. If the resolution was ± 0.0005 cm, which would be the state-of-the-art precision in 1901, we could perhaps find four significant digits of the estimation.

These three facts perhaps convey a lesson in 'experimental design.' In fact, Lazzarini's fraction simplifies to

$$\hat{\pi} = \frac{355}{113}, \tag{2.66}$$

a rational approximation for π known at least from the fifth century! This suggests that the number of tosses might have been planned as a multiple of this fraction to yield the expected result. It might also be an example of confirmation bias, when Lazzarini continued his work until a satisfactory approximation emerged. There is more scepticism of the statistical nature as regards Lazzarini's work [6], but there are also works which defend him, claiming the publication to be planned as a joke [7]. Either way, let us conclude with the uplifting reassurance that proper statistical analysis can debunk and pinpoint inconsistent, misprinted, erroneous, or falsified measurements.