Lectures on Statistical Mechanics

Werner Krauth

September 30, 2019

Lecture 1

Probability theory

In statistical mechanics, probabilities occupy, quite naturally, a central place. They are also present in daily life. In the present chapter, we present key concepts of the theory of probabilities. This provides us with a language and an understanding of the nature of probabilities.

in order to avoid misconceptions that arise from carrying over daily-life concepts of probabilities to where they do not apply. In Lecture 2, we will do the same for statistics, before plunging into statistical mechanics proper.

1.1 Probabilities

Probability theory and statistics are both concerned with the relation between data-generating processes and observed data (see Fig. 1.1, adapted from [1]):

- Probability theory describes properties of observations given a data-generating process. Such a data-generating process may describe the repeated throwing of two dice or the Boltzmann probability distribution of statistical physics, itself related to the time evolution of mechanical systems. It may also correspond to a process that modifies beliefs about next week's weather in Paris, or about the mass of protons given 17 experimental findings and theoretical constraints. Probabilities appear in everyday-life thinking and decision-making. Humans' "... judgment and decision-making under uncertainty" [2] is a powerful factor in social psychology and economics. This insight led to the Nobel prize in Economics in 2002.[1]

- Statistics analyzes the observed data from an unknown data-generating process, and possibly from additional information. In statistical inference, one then tries to analyze the data-generating process. Statistics makes "hard", mathematically proven statements. Nevertheless, each application of statistics to concrete data is outside of mathematics, as concrete data are not mathematical objects.

[1] see https://www.nobelprize.org/prizes/economic-sciences/2002/kahneman/biographical

In "Statistics", one often uses the singular word "a statistic", which denotes "a function of the data" (see [1, p 137]). In the literature, we have to watch out for the difference between "statistic" and "statistics". Perfect pitch is required to even notice the difference between "statistics" (as "functions of data"), and statistics, the science of data.

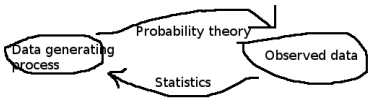

Figure 1.1: Probability and statistics: a first attempt at a description.

The above definitions imply that if, in statistical physics, we theoretically compute correlation functions of a given hamiltonian (as we will do later, in Section **??**), we are in fact working within probability theory (see Fig. 1.1). Equilibrium statistical physics, that we will be concerned with in these lectures, is applied probability theory, and the data-generating process is the Boltzmann distribution. Nevertheless, the very concept of statistical physics, namely the description of mechanical systems through probabilities (in other words, randomness) is a proper application of statistics. Let us conclude the introduction of technical terms with that of "stochastics" (that dates from the 1930s), which relates to random variables indexed by a discrete or real parameter, that can be taken as time.

On the other hand, reading off the temperature from a thermometer is an application of statistics. When, in computational physics, we sample a given probability distribution, we are in probability theory, but when we analyze the date, even for a known data-generating process, we have turned into statisticians. In our world of computer simulation and experiment, data, and of big data, the "statistics" aspects of "statistical" physics is of great importance.

1.1.1 Axioms of probability theory

Probability theory used to be an ill-defined discipline, and not really a part of core mathematics. This changed in the early 20th century, when Kolmogorov established the axioms of probability theory (there are no corresponding axioms of statistics). The basic concept is that of a sample space, denoted Ω. It is the set of all possible outcomes of an experiment $\omega \in \Omega$. ω is called a sample outcome or a realization. Sample spaces are somewhat in the background in all of probability theory [1], as one usually considers concentrates on random variables (see Section 1.2). Nevertheless, they are aways there.

Example of sample spaces are: Two coin tosses, where the sample space is $\{HH, HT, TH, TT\}$ and one sample is given by the result of the two coin tosses. This choice of sample space is not unique, though. For the same problem of the coin toss, one can also use a much larger set, partitioned into two disjoint subsets, the "Heads" event and the "Tails" event. The sample space could be the unit interval square with Lebesgue measure, and any of the quadrants are identified with the $\{HH, HT, TH, TT\}$. **FIXME** {this comes from Terence Tao post}.

Other example: Monte Carlo landing path. A sample is a pebble position (x, y), and an event is a subset of the square [2]).

Events are subsets of the sample space Ω (rather than single elements of the sample space). A probability is a function $\mathbb{P}$ that assigns a real number $\mathbb{P}(A)$ to each event A (that is, to a subset of the sample space, not to an individual sample). Probabilities must satisfy three axioms:

1. Probabilities are non-negative
$$\mathbb{P}(A) \geq 0 \quad \forall A. \tag{1.1}$$
2. Probabilities are normalized
$$\mathbb{P}(\Omega) = 1. \tag{1.2}$$
3. Probabilities add up: if $A_1, A_2, A_3, \ldots$ are disjoint, then
$$\mathbb{P}(\cup A_i) = \sum \mathbb{P}(A_i). \tag{1.3}$$

The axioms of probability theory assign a probability to an "event", a subset of sample space. Only in the case of a discrete state space can this subset be shrunk to a single point. In statistical physics, this translates to the fact that what will later be called the Boltzmann weight $\exp(-\beta E)$ is (for a continuous state space Ω that may correspond to the configuration space of the physical model) each infinitesimal event (subset) $[x, x + \mathrm{d}x\,]$ has Boltzmann probability $\mathbb{P} \propto \exp[-\beta E(x)]\,\mathrm{d}x$.

1.1.2 Interpretation of probabilities

Kolmogorov's axioms of probability theory provide a theoretical framework for speaking about probabilities, but they do not fix their interpretation. There are two main interpretations of probabilities. One is called the "frequentist" interpretation, and probabilities is the number of times an event happens, whereas the "Bayesian" approach interpretes probabilities as degrees of belief that a certain outcome takes place. All the theorems that we will discuss (in particular the inequalities and limit theorems) rely solely on the axioms, but not on the interpretation we provide for them:

1. In the above die example, with $\Omega = \{1, 2, 3, 4, 5, 6\}$ and $\mathbb{P}(1) = \cdots = \mathbb{P}(6) = 1/6$, we may suppose that the probability $1/6$ corresponds to the limiting frequency of observing a given outcome. In (the field of applied probability theory called) statistical physics, the frequentist interpretation plays a major role, and all quantities that are computed are explicitly meant to correspond to time averages.
2. The probability that it will rain tomorrow over Paris clearly treats $\{R, S\}$, where R stands for the outcome that there is rain tomorrow, and S that there is no rain. When we state $\mathbb{P}(R) = 0.65$, this might express our belief about tomorrow's weather. In the "belief" interpretation, the statement "The probability that Einstein drank tea on 1st Nov 1912 is 0.3" makes sense.

[2] The subset must be measurable (see [1, p. 13]

In common language, the words "likelihood" and "probabilities" are almost the same, and Merriam–Webster online dictionary defines the word "likely" as something quothaving a high probability of occurring or being true : very probable rain is likely today. In statistics, likelyhoods (as introduced by R. Fisher in 1923) are no probabilities (see Section **??**, they do not satisfy Kolmogorov's axioms). Our beliefs about tomorrow's weather in Paris are probabilities, not likelihoods.

1.1.3 Independence of events

Two events A and B are independent if

$$\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B) \tag{1.4}$$

A set J of events is independent if for any finite subset:

$$\mathbb{P}\left(\bigcap_{i\in J} A_i\right) = \prod_{i\in J} \mathbb{P}(A_i) \tag{1.5}$$

Independence has two origins:

1. We may suppose independence, that is, add it by construction. In the previous example of two coin tosses, with the sample space $\Omega = \{HH, HT, TH, TT\}$, we may consider two events. Let E_1 correspond to the event that the first toss realizes a head ($E_1 = \{HH, HT\}$) and let E_2 be the event that the second toss realizes a tail $E_2 = \{HT, TT\}$. We have that $\mathbb{P}(E_1) = 1/2$ and $\mathbb{P}(E_2) = 1/2$ and $\mathbb{P}(E_1 \cap E_2) = 1/4 = \mathbb{P}(E_1)\mathbb{P}(E_2)$, so that E_1 and E_2 are independent. We note that events with non-zero probabilities cannot be independent if they are disjoint. Indeed, the independent events E_1 and E_2 are not disjoint ($E_1 \cap E_2 = \{HT\}$).

2. Independence may just come out, by accident. Take a single throw of a die, and consider the events $A = \{2,4,6\}, B = \{1,2,3,4\}, AB = \{2,4\}$. If we suppose that each individual throw has probability $1/6$, then $\pi(A) = 3/6, \pi(B) = 4/6, \pi(AB) = 2/6$, which by accident satisfy $\mathbb{P}(A)\mathbb{P}B = (3/6)(4/6) = (2/6) = \mathbb{P}(AB)$.

1.2 Random variables

A random variable $\xi : \Omega \to R$ is a function (rigorously: a measurable map) that assigns a real number $\xi(\omega)$ to each outcome ω.

Random variable can be discrete or continuous. For the sample space of Fig. **??**, we may define a random variable that assigns the value $x = 1$ for all ω inside the unit circle, and $x = 0$ for all ω inside the square but outside the circle.

For discrete random variables, we can define the probability of a value x as the probability with which it takes on the value x:

$$f_\xi(x) = \mathbb{P}\left[\omega \in \Omega : \xi(\omega) = x\right] \tag{1.6}$$

and this can be written in short hand (without referring to the sample space), as:

$$f_\xi(x) = \mathbb{P}(\xi = x) \tag{1.7}$$

In the case of the standard problem, we have $f_\xi(x=0) = \mathbb{P}(\omega : \xi(\omega = 0) = 1 - \pi/4$ Likewise, $f_\xi(x=1) = \mathbb{P}(\omega : \xi(\omega = 1) = \pi/4$.

Next, we define the cumulative distribution function of a random variable:

$$F_\xi(x) = \mathbb{P}(\xi \leq x) \tag{1.8}$$

For a discrete random variable, the

$$F_\xi(x) = \mathbb{P}(\xi \leq x) = \sum_{x_i \leq x} f_\xi(x_i). \tag{1.9}$$

For a discrete random variable, the probability function or probability mass function is defined as

Probability density of a continuous random variable:

$$\mathbb{P}(a < \xi < b) = \int_a^b f_\xi(x)dx \tag{1.10}$$

we then have that $F(x) = \int_{-\infty}^{x} f_\xi(t)dt$ and also that $f_\xi(x) = F'_\xi(x)$ for all points where F is differentiable.

When considering random variables, the sample space discussed in Section **??** often appears to not really be there. But it is truly present (see [1, p. 19]), simply because any statement on probabilities of random variables (such as eq. (1.7)) may be expanded to a statement on a sample space on which this random variable is defined (such as eq. (1.6)).

We also note that a discrete random variable can coexist with a continuous sample space

Examples of random variables

Exponential random variable:

Bernoulli random variable:

Uniform random variable:

1.2.1 Independent random variables

Two random variables ξ, η are independent (written as $\xi \perp\!\!\!\perp \eta$) if $\forall A, B$, subsets of $\mathbb{R}$,

$$\mathbb{P}(\xi \in A, \eta \in B) = \mathbb{P}(\xi \in A)\mathbb{P}(\eta \in B) \tag{1.11}$$

There is a theorem stating: If ξ and η have joint probability distribution $f_{\xi,\eta}$ then:

$$\xi \perp\!\!\!\perp \eta \Leftrightarrow f_{\xi,\eta}(x,y) = f_\xi(x)f_\eta(y) \tag{1.12}$$

(This is Wasserman th 3.30). We sometimes write $\mathbb{P}(x,y) = \mathbb{P}(x)\mathbb{P}(y)$.

Let's note the sum of independent random variables, through a convolution:

$$f_{\xi+\eta}(x) = \int_{-\infty}^{\infty} \mathrm{d}y \; f_\xi f_\eta(x-y) \tag{1.13}$$

FIXME {Explain connection between independence of events and independence of random numbers. How does the disjointness come into play?}

1.2.2 Expectation of a random variable

The "Expectation", "expected value", "mean", "mean value" or or "first moment" of a random variable is defined as

$$\mathbb{E}(\xi) = \int x \mathrm{d}F_\xi(x) = \begin{cases} \sum_x x f(x) & \text{if } \xi \text{ is discrete} \\ \int \mathrm{d}x f(x) x & \text{if } \xi \text{ is continuous.} \end{cases} \tag{1.14}$$

Common symbols in use for the expectation are $\mathbb{E}(\xi)$ (or $\mathbb{E}\xi$), $\langle\xi\rangle$ or μ or μ_ξ. None of the denominations or notations is more noble, mathematical, correct, or stately than the other[1, p. 47]. We use here the word expectation and the symbol $\mathbb{E}$ (rather than $\langle\ldots\rangle$) in order to avoid the confusion with sample means.

Not all random variables have an expectation, but two random variables ξ and η with finite expectation satisfy

$$\mathbb{E}(\xi+\eta) = \mathbb{E}\xi + \mathbb{E}\eta \quad \text{or}: \langle\xi+\eta\rangle = \langle\xi\rangle + \langle\eta\rangle \tag{1.15}$$

and this is true whether or not the random variables are independent.

In addition that the mean of a function of the data is the function of the mean of the data (aka "rule of the lazy statistician") is in fact a tiny theorem.

$$\eta = r(\xi) \tag{1.16}$$

$$\mathbb{E}\eta = \mathbb{E}r(\xi) = \sum r(x) f(x) = \int dx r(x) f(x) \tag{1.17}$$

The rule is easy to rationalize for discrete random variables with a uniform distributions. Take for example a die with six equal faces, but paint the numbers $1^2, 2^2, \ldots, 6^2$ onto them. It is clear that the expectation is $(1^2+\cdots+6^2)/6 = 91/6$. Just like the rule of eq. (**??**), the eq. (1.17) is one of the tools of the trade, and simulators and experimentalists hand it down from one generation to the next: The same samples (for instance of the Boltzmann distribution) can be used to compute any observable expectations that we can compute from the samples. Many other rules and dear habits have however no mathematical foundation.

Consider a random variable $\xi \sim \text{Exponential}(\beta)$. What is its mean value $\mathbb{E}(\xi)$,

We first note that the distribution is normalized, because of

$$\frac{1}{\beta}\int_0^\infty \mathrm{d}x \ \exp(-x/\beta) = 1$$

The mean value of this distibution is $\mu_X = \beta$, because of

$$\frac{1}{\beta}\int_0^\infty \mathrm{d}x \ x \exp(-x/\beta) = \beta$$

What is the tail probability $P(|X-\mu_X| \geq k\sigma_X)$ (with $k \geq 1$) for $X \sim \text{Exponential}(\beta)$? Compare this tail probability to the bound you obtain from the Chebychev inequality.

The tail probability is $\exp(-k)$. Chebychev gives $1/k^2$. The two functins are the same for $k = 2W(1/2) = 0.703467..$, where W is the Product Log (the Lambert W function) (but our derivation does not apply to this case). For all $k \geq 1$, the Chebychev bound is not tight.

1.2.3 Variance of random variable

The variance of a random variable is defined as

$$\mathrm{Var}\,(\xi) = \left\{ \begin{array}{c} \text{Average squared distance} \\ \text{from mean} \end{array} \right\} = \mathbb{E}\left[(\xi - \mathbb{E}\,\xi)^2\right]. \tag{1.18}$$

Other notations for $\mathrm{Var}\,(\xi)$ are σ^2 or σ_ξ^2 or $V(\xi)$ or $V\xi$ (σ is the standard deviation). Again, using one or the other notation does not make us into a better person, but it helps to remember that the variance does not need to exist, that it's related to the square, and that it is the square of the standard deviation:

$$\sqrt{\mathrm{Var}} = \text{standard deviation} \tag{1.19}$$

Also it is useful to remember the two properties (Wasserman Th 4.1):

$$\mathbb{E}\,(a\xi + b) = a\mathbb{E}\,(\xi) + b \tag{1.20}$$

$$\mathrm{Var}\,(a\xi + b) = a^2\,\mathrm{Var}\,\xi \tag{1.21}$$

The variance of an exponential random variable ξ is $\mathrm{Var}_= \beta^2$ because $\mathrm{Var}_\xi = \mathbb{E}\,(\xi^2) - \mathbb{E}\,(\xi)^2$ and

$$\frac{1}{\beta}\int_0^\infty \mathrm{d}x\; x^2 \exp\left(-x/\beta\right) = 2\beta^2$$

The standard distribution of an exponential random variable ξ is $\sigma_\xi = \sqrt{\mathrm{Var}\,(\xi)} = \beta$.

Variance of sum of independent random numbers

For independent random variables $\xi_i, i = 1, \ldots, N$, the following is true:

$$\mathrm{Var}(\xi_1 + \cdots + \xi_N) = \begin{cases} \sum_{i=1}^N \mathrm{Var}\,(\xi_i) \\ N\,\mathrm{Var}\,(\xi_i) \quad \text{iid random variables} \end{cases} \tag{1.22}$$

To show this, we may restrict ourselves to random variables ξ_i with zero expectation, for which we have:

$$\begin{aligned} \mathrm{Var}\,(\xi_1 + \cdots + \xi_N) &= \mathbb{E}\left[(\xi_1 + \cdots + \xi_N)^2\right] \\ &= \sum_{i,j} \mathbb{E}\,(\xi_i \xi_j) = \sum_{i \neq j} \mathbb{E}\,(\xi_i)\,\mathbb{E}\,(\xi_j) + \sum_{i=1}^N \mathbb{E}\left(\xi_i^2\right) = \sum_{i=1}^N \mathbb{E}\left(\xi_i^2\right). \end{aligned} \tag{1.23}$$

1.2.4 Characteristic function

The characteristic function Φ of a random variable ξ is defined by

$$\Phi_\xi(t) = \mathbb{E}\left(\mathrm{e}^{it\xi}\right). \tag{1.24}$$

If the random variable has a probability density, this amounts to

$$\Phi_\xi(t) = \int_{-\infty}^{\infty} \mathrm{d}x\ \mathrm{e}^{itx} f x, \tag{1.25}$$

with the probability density given by the characteristic function through an inverse Fourier transform:

$$f_\xi(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}t\ \mathrm{e}^{-itx} Phit, \tag{1.26}$$

Some general properties of the characteristic functions are useful to keep in mind:

- $\Phi_\xi(0) = 1$. The total probability is normalized to 1.
- $\Phi_\xi(t) = \Phi^*_\xi(-t)$. This is because the probability density function is real.
- $|\Phi_\xi(t)| \leq 1$. This is because the absolute value of the integral is bounded from above by the integral of the absolute value, which is $\Phi_\xi(0)$.
- $\Phi_{a\xi+b}(t) = \mathrm{e}^{ibt}\Phi_\xi(at)$. Under a change of variables $\xi' = f(\xi)$, with $f(\xi)$ monotonous, the probability distribution transforms as follows

$$\pi_{\xi'}(x) = \frac{\pi_\xi(f^{-1}(x)}{|f'(f^{-1}(x))|}\,. \tag{1.27}$$

In particular, under a linear transformation we have

$$\pi_{a\xi+b}(x) = |a|^{-1}\pi_\xi((x-b)/a) \tag{1.28}$$

and hence

$$\Phi_{a\xi+b}(t) = \int \mathrm{d}x\ \mathrm{e}^{ixt}|a|^{-1}\pi_\xi((x-b)/a) = \mathrm{e}^{ibt}\Phi_\xi(at)\,. \tag{1.29}$$

Let ξ_1 and ξ_2 two independent random variables. The characteristic function of their sum of n independent random variables can be computed as follows: Since ξ_1 and ξ_2 are independent, so are $\mathrm{e}^{it\xi_1}$ and $\mathrm{e}^{it\xi_2}$.

$$\Phi_{\xi_1+\xi_2}(t) = \mathbb{E}\left(\mathrm{e}^{it(x_1+x_2}\right) = \mathbb{E}\left(\mathrm{e}^{itx_1}\mathrm{e}^{itx_2}\right) = \mathbb{E}\left(\mathrm{e}^{itx_1}\right)\mathbb{E}\left(\mathrm{e}^{itx_2}\right) = \Phi_{\xi_1}(t)\Phi_{\xi_2}(t) \tag{1.30}$$

This can be readily generalized to N variables

$$\Phi_{\sum_i^N \xi_i}(t) = \prod_{i=1}^{N} \Phi_{\xi_i}(t)\,. \tag{1.31}$$

The first cumulant κ_1 is the expectation. The second cumulant is the variance. The cumulants are proportional to the coefficients of the series expansion of the logarithm of the characteristic function. Since the logarithm of a product is the sum of the logarithms, the n-th cumulant of the sum of two independent random variables is the sum of the n-th cumulant of the random variables. In particular, this applies to $n = 2$, *i.e.* to the variance.

Sum of random variables with uniform distribution. [EASY] Compute the characteristic function of the sum of n random variables ξ_j with uniform distribution $f_{\xi_j}(x) = \frac{1}{2a}\theta_H(x+a)\theta_H(a-x)$, where $\theta_H(x+a)$ is the Heaviside theta

[EASY] Let ξ_1 and ξ_2 two independent random variables, what is the characteristic function of their sum? What about the sum of n independent random variables? Since ξ_1 and ξ_2 are independent, so are $\mathrm{e}^{it\xi_1}$ and $\mathrm{e}^{it\xi_2}$.

$$\Phi_{\xi_1+\xi_2}(t) = \mathbb{E}\left(\mathrm{e}^{it(x_1+x_2)}\right) = \mathbb{E}\left(\mathrm{e}^{itx_1}\mathrm{e}^{itx_2}\right)) = \mathbb{E}\left(\mathrm{e}^{itx_1}\right)\mathbb{E}\left(\mathrm{e}^{itx_2}\right) = \Phi_{\xi_1}(t)\Phi_{\xi_2}(t)$$

This can be readily generalized to N variables

$$\Phi_{\sum_i^N \xi_i}(t) = \prod_{i=1}^{N} \Phi_{\xi_i}(t). \tag{1.32}$$

The characteristic function of f_{ξ_j} is simply given by $\frac{\sin(at)}{at}$, therefore the characteristic function of the sum of n uniform variables is $(\frac{\sin(at)}{at})^n$.

Show that the characteristic function of $\xi^{(n)} = \sum_{j=1}^{n} \xi_j$ can be written in the following form:

$$\Phi_{\xi^{(n)}} = \frac{1}{(2ia)^n} t^{-n} \sum_{k=0}^{n} \binom{n}{k} (-1)^k \mathrm{e}^{i(n-2k)at} \tag{1.33}$$

[Hint 1:] Express the sine functions using complex exponentials ($\sin x = \frac{\mathrm{e}^{ix}-\mathrm{e}^{-ix}}{2i}$) and use the binomial theorem $(a+b)^j = \sum_{k=0}^{j} \binom{j}{k} a^j b^{j-k}$.

$$\left(\frac{\sin(at)}{at}\right)^n = \frac{t^{-n}}{(2ia)^n}(\mathrm{e}^{iat} - \mathrm{e}^{-iat})^n = \frac{t^{-n}}{(2ia)^n} \sum_{k=0}^{n} \binom{n}{k} [\mathrm{e}^{iat}]^{n-k}[-\mathrm{e}^{-iat}]^k = \frac{t^{-n}}{(2ia)^n} \sum_{k=0}^{n} \binom{n}{k} (-1)^k \mathrm{e}^{i(n-2k)at} \tag{1.34}$$

[HARD] Compute the inverse Fourier transform of the characteristic function and show that the distribution of $\xi^{(n)}$ can be written as [**?**]

$$f_{\xi^{(n)}} = \frac{1}{(n-1)!(2a)^n} \sum_{k=0}^{n} \binom{n}{k} (-1)^k \max\Big((n-2k)a - x, 0\Big)^{n-1}. \tag{1.35}$$

[Hint 1:] Move the sum outside of the integral of the inverse Fourier transform. *Warning*: the resulting integrals are divergent, but the divergencies have to simplify, so don't worry too much! The finite part of the integrals can be extracted using the *Cauchy principal value*, usually denoted by P.V., which, in the case of a singularity at zero, reads as

$$\mathrm{P.V.} \int_{-\infty}^{\infty} \mathrm{d}t\ f(t) = \lim_{\epsilon \to 0^+} \Big[\int_{-\infty}^{-\epsilon} \mathrm{d}t\ f(t) + \int_{\epsilon}^{\infty} \mathrm{d}t\ f(t) \Big]. \tag{1.36}$$

[Hint 2:] Compute the (finite part of the) integrals by integrating by parts $n-1$ times (note that the original product of sin functions has a zero of order n at $t = 0$).

[Hint 3:] P.V. $\int_{-\infty}^{\infty} \mathrm{d}t\; t^{-1}\mathrm{e}^{itb} = i\pi\mathrm{sgn}(b)$. [Hint 4:] $\sum_{k=0}^{n}\binom{n}{k}(-1)^k(x+k)^j = 0$ for any x and integer $j = 1,\dots,n-1$.

We must compute

$$f_{\xi^{(n)}}(x) = \frac{1}{2\pi}\int \mathrm{d}t\; \mathrm{e}^{-ixt}\frac{1}{(2ia)^n}t^{-n}\sum_{k=0}^{n}\binom{n}{k}(-1)^k\mathrm{e}^{i(n-2k)at}\,. \tag{1.37}$$

First, we move the sum outside of the integral and take the principal value

$$f_{\xi^{(n)}}(x) = \frac{1}{2\pi}\sum_{k=0}^{n}\mathrm{P.V.}\int \mathrm{d}t\; \frac{1}{(2ia)^n}t^{-n}\binom{n}{k}(-1)^k\mathrm{e}^{i((n-2k)a-x)t}\,. \tag{1.38}$$

Since $(\sin(at))^n$ has a zero of order n at $t=0$, the boundary parts which come from the integration by parts (taking the integral of t^{-n} and the derivative of the rest) and which could have given contribution from $t=0$ are in fact zero for $n-1$ consecutive integration by parts. Thus we find

$$f_{\xi^{(n)}}(x) = \frac{1}{2\pi}\sum_{k=0}^{n}\binom{n}{k}(-1)^k\frac{((n-2k)a-x)^{n-1}}{(2a)^n(n-1)!}\mathrm{P.V.}\int \mathrm{d}t\;\; t^{-1}\mathrm{e}^{i((n-2k)a-x)t}\,. \tag{1.39}$$

The integral can be easily evaluated and gives

$$\begin{aligned} f_{\xi^{(n)}}(x) = &\sum_{k=0}^{n}\binom{n}{k}(-1)^k\frac{((n-2k)a-x)^{n-1}}{2(2a)^n(n-1)!}\mathrm{sgn}((n-2k)a-x) = \\ &\sum_{k=0}^{n}\binom{n}{k}(-1)^k\frac{((n-2k)a-x)^{n-1}}{2(2a)^n(n-1)!}[2\theta_H((n-2k)a-x)-1] = \\ &\sum_{k=0}^{n}\binom{n}{k}(-1)^k\frac{((n-2k)a-x)^{n-1}}{(2a)^n(n-1)!}\theta_H((n-2k)a-x)\,, \end{aligned} \tag{1.40}$$

where in the last step we used the identity in Hint 4 to keep only the term multiplied by the step function. The proof is concluded noting that $x\theta_H(x) = \max(x,0)$.

1.3 Inequalities and their meaning

Inequalities provide important tools in probabilities and statistics. They often provide simple yet exact statements for finite sums of independent random variables $\xi_1 + \cdots + \xi_n$, rather than for their $N \to \infty$ limit. We discuss the truly fundamental Markov (Section 1.3.1 and Chebychev (Section 1.3.2) inequalities that apply to very general distributions and to sums of independent random variables, for which the existence of a finite expectation, or of a finite variance allows one to bound the tail probabilities. Hoeffding's inequality (Section 1.3.3) is an example of an inequality for finite sums bounded random variables, whereas Mill's inequality (see Section 1.3.4) is distribution-specific. It applies only to the Gaussian random variable.

As mentioned just above, probabilistic inequalities are for a single random variable or a finite sum of n iid random variables. They allow to reach strong mathematical results, in probability theory and in statistics, without invoking the $n \to \infty$ limit.

1.3.1 Markov's inequality

Markov's inequality, for a non-negative random variable with finite mean states that

$$\mathbb{P}(\xi > t) \leq \frac{\mathbb{E}(\xi)}{t} \tag{1.41}$$

The proof of Markov's inequality is as follows:

$$\mathbb{E}(\xi) = \int_0^\infty dx x f(x) = \underbrace{\int_0^t \mathrm{d}x\ x f(x)}_{\geq 0} + \int_t^\infty \mathrm{d}x\ x f(x) \geq \underbrace{t \int_t^\infty \mathrm{d}x\ f(x)}_{\mathbb{P}(\xi > t)}. \tag{1.42}$$

In eq. (1.41), the difference between the expectation $\mathbb{E}(\xi)$ and 0 (the lower limit of the distribution) is a scale, and the probability to be more than k times above this scale is less than $1/k$ (for $k \geq 1$). Let's suppose that household income is non-negative, then Markov's inequality states that less than one percent of households have more than one hundred times the average income. This may appear as a meager consolation. For Markov's inequality to hold, the expectation must exist (be finite).

1.3.2 Chebychev inequality

The Chebychev inequality:

$$\mathbb{P}\left(|\xi - \mathbb{E}\xi| \geq \epsilon\right) \leq \frac{\mathrm{Var}\,(\xi)}{\epsilon^2} \tag{1.43}$$

is one of the fundamental achievements in probability theory, and it is of great importance in statistics. In eq. (1.43), $\mathrm{Var}\,(\xi)$ denotes the variance of the distribution and $\mathbb{E}\,(\xi)$ its expectation.

To prove Chebychev's inequality, we may restrict ourselves to a random variable with zero expectation. Then

$$\mathrm{Var}\,(\xi) = \int dx x^2 f(x) \geq \int_{|x| \geq \varepsilon} dx x^2 f(x) \geq \varepsilon^2 \int_{|x| \geq \varepsilon} dx f(x) = \varepsilon^2 \mathbb{P}(|\xi| \geq \varepsilon) \tag{1.44}$$

It is just as easy to prove Chebychev's inequality using Markov's inequality.

For the Chebychev inequality to hold for a given random variable, there is really only a single condition, the finiteness of its variance. This is one of the very many examples in probability theory and statistics where the variance has a meaning that other quantities, for example the mean absolute distance, simply do not have. Naturally, a finite variance requires a finite expectation.

For many distributions, the Chebychev inequality gives a much worse bound for the tail probability than either the exact tail probability itself, or than sharper inequalities (see for example Hoeffding's inequality. Nevertheless, it cannot be replaced by a better inequality without making additional assumptions. To show that the Chebychev inequality is sharp, we consider the "double δ-peak" distribution:

$$\pi(x) = \frac{1}{2}\delta(-1) + \frac{1}{2}\delta(1) \tag{1.45}$$

whose variance equals 1. The correct way to refer to this distribution is through the cumulative distribution function (cdf), which equals

$$F_\xi(x) = \begin{cases} 0 & \text{for } x < -1 \\ \frac{1}{2} & \text{for } -1 < x \leq 1 \\ 1 & \text{for } x > 1 \end{cases} \tag{1.46}$$

The variance of this random variable is $\mathrm{Var}\,(\xi) = 1$,

For this random variable, we have

$$\mathbb{P}(|\xi| > 1 - \epsilon) = 1 \quad \forall \epsilon > 0 \tag{1.47}$$

and at the same time, Chebychev's inequality yields

$$\mathbb{P}(|\xi| > 1 - \epsilon) \leq \frac{1}{(1-\epsilon)^2} \tag{1.48}$$

In the limit $\epsilon \to 1$, this inequality thus becomes sharp at two points, namely at $x = \pm 1$.

These properties cannot really be improved. It is for example impossible for the Chebychev inequality to be sharp at all values of x. If this was true, then we would have that $\mathbb{P}\,(|\xi| > x) = c/x^2$, which would for the probability density function to be $f_\xi(x) = c/|x|^3$. However, this random variable has no variance (in other words, the variance is infinite).

The Chebychev inequality is concerned with the two-sided tail probability of a random variable, but there exists a one-sided variant of it, called Cantelli inequality. It is given by

$$\mathbb{P}\,[\xi - \mathbb{E}\,(\xi) > \epsilon] \leq \frac{\mathrm{Var}\,(\xi)}{\mathrm{Var}\,(\xi) + \epsilon^2} \tag{1.49}$$

and there are now now absolute values on the left-hand side of the equation.

1.3.3 Hoeffding's inequality

Hoeffding's inequality considers bounded independent random variables $\xi_1, \xi_2, \ldots, \xi_n$ that all have zero expectation and satisfy $a_i \leq \xi_i < b_i$. We discuss this particular inequality (out of a great number of other ones, certainly also worthy of study), because it provides an exponential tail probability for a finite sum of bounded random variable, that we will need later.

For every $t > 0$, Hoeffding's inequality states:

$$\mathbb{P}(\sum_{i=1}^{n} \xi_i \geq \epsilon) \leq \exp\,(-t\epsilon) \prod_{i=1}^{n} \exp\left[t^2 (b_i - a_i)^2 / 8\right]. \tag{1.50}$$

The inequality is easy to prove, see [1, p. xxx]). It holds for any $t \geq 0$, and in usual applications one has to find the value of t that provides the best bound. Note also that there is no absolute value on the left-hand side of eq. (1.50).

Hoeffding's inequality provides an exponential bound on the tail probability of the sample mean, but it is much sharper than the Chebychev inequality only if we go deep into the tails of the distribution, that is, consider large values of ϵ.

Hoeffding's inequality for sums of Bernoulli random variables

An inequality about tail probabilities is not very useful, if we know beforehand that the random variable is Bernoulli-distributed, that is that it takes on only values 0 (with probability $1-\theta$) and 1 (with probability θ. The situation is much more interesting if we look at the sum of n Bernoulli-distributed random variables, but where we may ignore the value of θ. Nevertheless it is possible to obtain a powerful estimate, that we will use a simplified expression of Hoeffding's inequality, for the special case where the random variables X_i are i.i.d Bernoulli distributed, and instead of the the sum of random variables, we consider the sample-mean random variable

$$\overline{\xi}_n = 1/n \sum_{i=1}^{n} \xi_i \tag{1.51}$$

We formulate Hoeffding's inequality for $\mathbb{P}\left[|\overline{\xi}_n - \theta| \geq \epsilon\right]$ (although Hoeffding's inequality is for random variables of zero mean, but this can be arranged).

One finds $\mathbb{P}\left[|\overline{\xi}_n - \theta| \geq \epsilon\right] \leq 2\mathrm{e}^{-2n\epsilon^2}$

We note that $\xi_k - \theta$ is a random variable with zero mean, as required. We furthermore note that Hoeffding inequality is for the sum of random variables, not for the mean value. However, the $1/n$ is innocuous:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i \geq \frac{1}{n}\epsilon\right) \leq \exp\left(-t\frac{1}{n}\epsilon\right)\prod_{i=1}^{n}\exp\left[t^2(b_i-a_i)^2/8\right]. \tag{1.52}$$

leads for $\epsilon/n \to \epsilon$ and Bernoulli variables, for which $b_i - a_i = 1$, to the probability:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i \geq \epsilon\right) \leq \exp\left(-t\epsilon\right)\exp\left[nt^2/8\right]. \tag{1.53}$$

The rhs is $\exp\left(-t\epsilon + nt^2/8\right)$. We find the best value of t by deriving. This gives $(n/4)t = \epsilon$ and for the r.h.s. exponentiall the term $-2\epsilon^2/n$. Therefore, we find that the tail probability from Hoeffding's inequality is $2\exp\left(-2n\epsilon^2\right)$, that is, exponential in $n\epsilon^2$. Chebychev gives $\theta(1-\theta)/(n\epsilon^2)$. We need $n \sim 1/\epsilon^2$ in both cases, but for $n \gg 1/\epsilon^2$, Hoeffding is much sharper.

1.3.4 Mill's inequality, asymptotic expansion of the error function

For the normal distribution (that is for a random variable $\eta \sim N(0,1)$:

$$f_\eta(z) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right), \tag{1.54}$$

the much stronger Mill's inequality holds. It is given by:

$$\mathbb{P}(|\eta| > t) < \sqrt{\frac{2}{\pi}}\,\frac{\mathrm{e}^{-t^2/2}}{t} \quad \text{(Mill's inequality; normal distribution)}$$

Prove Mill's inequality and generalize it for a Gaussian distribution with zero mean, but standard deviation σ (Hint: note that $\mathbb{P}(|z|>t)=2\mathbb{P}(z>t)$).

From the definition, we have that

$$\begin{aligned}\mathbb{P}(|\eta|>t)=2\mathbb{P}(\eta>t)=\sqrt{\frac{2}{\pi}}\int_t^\infty dz\exp\left(-z^2/2\right)=\\ \sqrt{\frac{2}{\pi}}\exp\left(-t^2/2\right)\int_t^\infty dz\exp\left(-(z^2/2-t^2/2)\right)=\\ \sqrt{\frac{2}{\pi}}\exp\left(-t^2/2\right)\int_t^\infty dz\exp\left[-\frac{1}{2}(z+t)(z-t)\right]<\\ \sqrt{\frac{2}{\pi}}\exp\left(-t^2/2\right)\int_t^\infty dz\exp\left[-t(z-t)\right]=\\ \sqrt{\frac{2}{\pi}}\exp\left(-t^2/2\right)\int_0^\infty du\exp\left[-tu\right]\end{aligned}$$

and Mill's inequality follows immediately. For a Gaussian with standard deviation σ, one simply changes $\mathbb{P}(|z|>t)$ into $\mathbb{P}(|z|>t\sigma)$. The rhs of the above eqs remains unchanged.

Here is a derivation of Mill's inequality that gives the exact value of the corrective term:

We use:

$$\exp\left(-z^2/2\right)=-\frac{1}{z}\frac{\mathrm{d}}{\mathrm{d}z}\exp\left(-z^2/2\right) \tag{1.55}$$

from which it follows that

$$\frac{\mathrm{d}}{\mathrm{d}z}\left[\frac{1}{z}\exp\left(-z^2/2\right)\right]=\underbrace{\left[\frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{1}{z}\right)\right]\exp\left(-z^2/2\right)}_{-\frac{1}{z^2}\exp(-z^2/2)}+\underbrace{\frac{1}{z}\frac{\mathrm{d}}{\mathrm{d}z}\exp\left(-z^2/2\right)}_{-\exp(-z^2/2)} \tag{1.56}$$

Integrating both sides from t to ∞ and multiplying with $\sqrt{2/\pi}$ gives

$$\underbrace{\sqrt{\frac{2}{\pi}}\int_t^\infty dz\mathrm{e}^{-z^2/2}}_{\mathbb{P}(|z|>t)}=\sqrt{\frac{2}{\pi}}\frac{1}{t}\mathrm{e}^{-z^2/2}-\underbrace{\sqrt{\frac{2}{\pi}}\int_t^\infty dz\frac{1}{z^2}\mathrm{e}^{-z^2/2}}_{>0} \tag{1.57}$$

To expand the final integral in its turn, we simply multiply eq. (**??**) on both sides by $1/z^2$ to obtain

$$-\int_t^\infty\frac{1}{z^2}\mathrm{e}^{-z^2/2}=-\frac{1}{t^3}\mathrm{e}^{-t^2/2}+3\int_t^\infty\frac{1}{z^4}\mathrm{e}^{-z^2/2} \tag{1.58}$$

from which follows the asymptotic expansion

$$\underbrace{\sqrt{\frac{2}{\pi}}\int_t^\infty dz\mathrm{e}^{-z^2/2}}_{\mathbb{P}(|z|>t)}\sim\sqrt{\frac{2}{\pi}}\frac{\mathrm{e}^{-t^2/2}}{t}\left[1+\sum_{n=1}^\infty(-1)^n\frac{1\cdot3\cdots(2n-1)}{t^{2n}}\right]. \tag{1.59}$$

We note that for fixed t the sum on the right-hand side diverges, because for large n, $(2n-1)$ eventually gets bigger than t^2.

confidence the

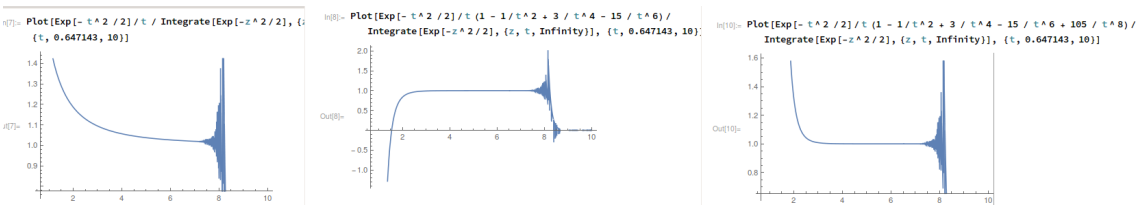

Figure 1.2: Error Asymptotic (several terms of the Mill's inequality)

1.4 Convergence of random variables

We consider sequences of random variables $\xi_1, \xi_2, \ldots$, and address the question of their convergence towards a random variable ξ.

Two cases are the most important to consider. First, one considers the convergence of the random variables where ξ_n is the sample mean of iid random variables. This is called the law of large numbers, and one usually studies the weak law of large numbers. This was sufficient when not all researchers did not spend all their time watching data on their computer. In this section, we will try to explain that any sequence of random variables satisfies the strong law of large numbers (rather than the weak law), and why we should understand this in our daily work with data.

The second point that we want to make is about the central limit theorem. This theorem (that we often know from highschool) states that the sum of random variables (for iid random variables with the only condition, that their variances exist) converge to a Gaussian random variable. The central limit theorem thus appears as an asymptotic statement, valid only in the $n \to \infty$ limit. However, we will discuss a very simple additional condition (on the third absolute moment) which allows to quantify the difference with the Gaussian (that the distribution will converge to) for any finite value of n. This is the Berry–Esseen theorem (with many extensions) that we will discuss in Section **??**

1.4.1 Different types of convergence of random variables

There are different ways in which

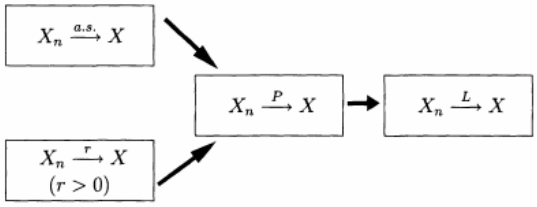

Figure 1.3: modes of convergence

Convergence in probability and almost sure convergence can be discussed in ex-

ample:

1. Convergence in probability: This is, as we have already discussed for the Chebychev inequality, $\xi_n \xrightarrow{\text{in prob.}} \xi$:

$$\mathbb{P}(|\xi_n - \xi| > \epsilon) \to 0 \tag{1.60}$$

   for every $\epsilon > 0$ as $n \to \infty$. Convergence in probability is defined for a sequence of random variables that converge towards another random variable, not for the distributions converging towards another probability distribution. These are complicated concepts, as the ξ_n and the ξ are (normally) by no means independent. Situation is simpler if the distribution is concentrated on a point. Then the convergence in probability means that ξ_n becomes more and more concentrated around this point.

2. Almost sure convergence (we restrict ourselves here to almost sure convergence towards a constant-mass distribution): Suppose that the sequence ξ_n of partial sums almost surely converges towards a constant μ. Just as for convergence in analysis, this means that for an individual sequence ξ_n with $n = 1, 2, \ldots$, and for every ϵ, there exists an n', such that ξ_n remains within the window $\mu \pm \epsilon$ for all $n \geq n'$.

   In frequentist interpretation, this means that the number of paths that go out of a window of size ϵ for all later k goes to zero.

3. Convergence in quadratic mean: Expectation value $\mathbb{E}\left[(X_n - X)^2\right] \to\to 0 \quad \text{for } n \to \infty$

4. Convergence in distribution:

$$\lim_{n\to\infty} F_n(t) = F(t) \tag{1.61}$$

   for all t for which F is continuous.

   Convergence in distribution is much easier to visualize for an extended distribution than the three other concepts.

Convergence in probability considers constant n and compares the probability distribution of sample means

Almost sure convergence makes a strong statement, not only on what happens at

The different types of convergence (almost sure, in probability, in distribution) have been discussed in detail in the literature (see [3]).

1.4.2 Law(s) of large numbers

Socalled weak law of large numbers: Consider $\xi_1, \ldots, \xi_N$ iid (independent and identially distributed) random variables $\xi_1, \ldots, \xi_N$ with finite mean $\mathbb{E}\xi$. It satisfies:

$$\overline{\xi}_n = \frac{1}{n}\sum_i^n \xi_i \xrightarrow{\text{in prob.}} \mathbb{E}\xi \tag{1.62}$$

By the definition of convergence in probability, this means

$$\mathbb{P}(|\overline{\xi}_n - \mu| > \epsilon) \to 0 \forall \epsilon \tag{1.63}$$

Proof by Chebychev inequality (for finite variance of ξ, which is not necessary).

$$\mathrm{Var}(\sum_i \xi_i) = n\sigma^2 \tag{1.64}$$

$$\mathrm{Var}(\frac{1}{n}\sum_i \xi_i = \frac{1}{n}\sigma^2 \to 0 \tag{1.65}$$

is actually a distribution with zero variance in the $n \to \infty$ limit

$$\mathbb{P}(|\overline{\xi_n} > \epsilon) < \mathrm{Var}\,/(n\varepsilon^2) \tag{1.66}$$

Now, in order to have finite precision $\mathrm{Var}\,/(n\varepsilon^2) = 0.05$ from which follows $\epsilon \sim 1/\sqrt{n}$.

It was originally believed that (pairwise) independence was necessary for the LLN to hold. Markov saw that this was not correct, and proceeded to construct a counterexample. This was the first "Markov" chain [4, p23].

The weak law of large numbers is independent of the interpretation of probabilities (frequentist, Bayesian), but it appears that without the weak law of large numbers, the frequentist interpretation would make little sense.

In this context, it is of great importance that the law of large numbers, discovered by Borel, applies to practically all the sequences that are described by the weak law For iid variables, the strong law of large numbers tells us that Let $\xi_1, \xi_2, \ldots$ be IID. if $\mu = \mathbb{E}|X_1| < \infty$, then $\langle X_n \rangle \to \mu$ almost surely. What this means is best discussed in terms of partial sums:

- The weak law states that the probability distribution of partial sums, at time n, become more and more peaked. It makes no statement on the behavior of each individual partial sum, as a function of n.

- The strong law of large numbers states that, for any ϵ there is an n, so that the partial sums do not move out of the window $\mathbb{E}(\xi) \pm \epsilon$ for all $n' > n$.

Just about all partial sum of iid random variables satisfy the strong law of large numbers. It is the author's firm opinion that anyone analyzing data (which are often time series) must understand the subtle difference between weak law and strong law.

1.4.3 Central limit theorem

$$\xi_n = \frac{1}{\sqrt{n}}(\xi_1 + \cdots + \xi_n) \tag{1.67}$$

Famous theorem by Gnedenko and Kolmogorov: iid finite variance necessary and sufficient convergence in distribution to gaussian. Proof is easy if we suppose furthermore that $\mathbb{E}(\xi_i) = 0$ and all powers finite. (see [6, p. 54]).

Central limit theorem for uniform random variables

FIXME {See the TD for detailed text}

1. Verify the validity of the central limit theorem for the sum of variables with uniform distribution (you can work with the characteristic function).

Hint 1: $\log \frac{\sin t}{t} = \sum_{n=1}^{\infty} \frac{(-1)^n B_{2n}}{2n(2n)!}(2t)^{2n}$, where the coefficients B_n are known as "Bernoulli numbers", $B_0 = 1$, $B_2 = \frac{1}{6}$, $B_4 = -\frac{1}{30}$, et cetera.

The characteristic function of $\tilde{\xi} = \frac{\xi^{(n)}}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \xi_j$ is given by

$$\phi_{\tilde{\xi}}(t) = \phi_{\xi^{(j)}}(t/\sqrt{n}) = \Big(\sqrt{n}\frac{\sin(at/\sqrt{n})}{at}\Big)^n = \exp\Big(n\log\Big(\sqrt{n}\frac{\sin(at/\sqrt{n})}{at}\Big)\Big) = \exp\Big(\sum_{j=1}^{\infty} n^{1-j}\frac{(-1)^j B_{2j}}{2j(2j)!}(2at)^{2j}\Big) = \exp\Big(-\frac{(at)^2}{6} + O(1/n)\Big). \quad (1.68)$$

In the limit $n \to \infty$ this approaches the characteristic function of a Gaussian with mean zero and variance $a^2/3$.

2. **Stable distributions. Definition**: A non-degenerate distribution π_ξ is a stable distribution if it satisfies: let ξ_1 and ξ_2 be independent copies of a random variable ξ (they have the same distribution π_ξ). Then π_ξ is said to be stable if for any constants $a > 0$ and $b > 0$ the random variable $a\xi_1 + b\xi_2$ has the distribution $\pi_{c\xi+d}$ for some constants $c > 0$ and d.

3. [MEDIUM] Prove that the Gaussian $\pi_\xi(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$ is a stable distribution. Remember the propriety of the generating functions

$$\Phi_{a\xi+b}(t) = \mathrm{e}^{ibt}\Phi_\xi(at) \quad (1.69)$$

The generating function $\phi^{(G)}(t)$ of a Gaussian distribution is given by

$$\Phi^{(G)}(t) = \mathrm{e}^{-(2\sigma^2)t^2} \quad (1.70)$$

We consider the characteristic function of the sum of the two random variables $a\xi + b\xi$, which we know to be given by the product of the characteristic functions of the single (Gaussian) distributions

$$\Phi_{a\xi+b\xi}(t) = \Phi^{(G)}_{a\xi}(t)\Phi^{(G)}_{b\xi}(t) = \mathrm{e}^{-(2\sigma^2)t^2(a^2+b^2)} = \Phi^{(G)}_{\sqrt{a^2+b^2}\xi}(t) \quad (1.71)$$

We have then shown that the distribution of the sum of the two Gaussian random variables $a\xi + b\xi$ is a Gaussian distribution of the variable $c\xi$ with $c = \sqrt{a^2+b^2}$. Therefore the Gaussian distribution is stable.

4. [EASY] Consider a characteristic function of the form

$$\Phi_\xi(t) = \exp\big(it\mu - (c_0 + ic_1 f_\alpha(t))|t|^\alpha)\big), \quad (1.72)$$

with $1 \leq \alpha < 2$. Show that $f_\alpha(t) = \mathrm{sgn}(t)$, for $\alpha \neq 1$, and $f_1(t) = \mathrm{sgn}(t)\log|t|$ produce stable distributions. These are also known as *Lévy distributions*, after Paul Lévy, the first mathematician who studied them.

As before we consider the conbination of two random variables with a Lévi distribution has the characteristic function

$$\Phi_{a\xi_1+b\xi_2}(t) = \exp(it(a+b)\mu - (c_0 + ic_1 f_\alpha(t))|(a+b)^{1/\alpha}t|^\alpha)\,. \tag{1.73}$$

If $\alpha \neq 1$ this is mapped into the same distribution by the transformation $t \to (a+b)^{-1/\alpha}t$ and $\mu \to (a+b)^{1/\alpha-1}\mu$. For $\alpha = 1$ the transformation is $t \to (a+b)^{-1}t$ and $\mu \to \mu - \frac{c_1}{\alpha}\log(a+b)$.

[EASY] *Find a distinctive feature of the Lévy distributions.*

The second cumulant

$$\kappa_2 = (-i)^2\partial_t^2 \log\Phi(t)\Big|_{t=0} \sim sign(t)|t|^{\alpha-2} + \ldots\Big|_{t=0} = \infty \tag{1.74}$$

as $\alpha < 2$, it diverges.

5. [EASY] Assumes $\alpha \neq 1$ and show that, in order to be $\Phi_\xi(t)$ the Fourier transform of a probability distribution, the coefficient c_1 can not be arbitrarily large; determine its maximal value.

Hint 1: One can show (MEDIUM-HARD) that the inverse Fourier transform of (1.84) has the tails

$$\pi_\xi(x) \xrightarrow{|x|\gg 1} \frac{\Gamma(1+\alpha)}{2\pi|x|^{1+\alpha}}\left(c_0 \sin\frac{\pi\alpha}{2} - c_1 \mathrm{sgn}(x)\cos\frac{\pi\alpha}{2}\right). \tag{1.75}$$

The probability distribution must be positive or equal to zero, therefore the coefficients of the tails of the Lévi distributions must be positive. Since

$$\pi_\xi(x) \xrightarrow{|x|\gg 1} \frac{\Gamma(1+\alpha)}{2\pi|x|^{1+\alpha}}\left(c_0 \sin\frac{\pi\alpha}{2} - c_1 \mathrm{sgn}(x)\cos\frac{\pi\alpha}{2}\right) \tag{1.76}$$

we find

$$|c_1| < c_0\left|\tan\frac{\pi\alpha}{2}\right| \tag{1.77}$$

FIXME { Compute the difference between the asymptotic value and the finite-n value for a random variable with finite third moment. This can bring out the Berry–Esseen theorem [7, 8]}

Berry–Esseen theorem

We know that for iid random variables, the existence of the second moment (that is, of the variance) is paramount to the convergence of the sum of the random variables divided by $\sqrt{n}$ towards a Gaussian (see eq. (1.67)). The natural question about the speed of this convergence is answered by the Berry–Esseen theorem, a classic result in the theory of probabilities:

$$\sup_z |\mathbb{P}(z_n \leq z) - \Phi(z)| < c_E \frac{\mathbb{E}|\xi - \mathbb{E}\xi|^3}{\sqrt{n}\sqrt{\mathrm{Var}\,\xi^3}} \tag{1.78}$$

where c_E is a constant. We see that the convergence is controlled by the third absolute moment of the random variable: $\mathbb{E}|\xi - \mathbb{E}\xi|^3$, divided by a power of its variance.

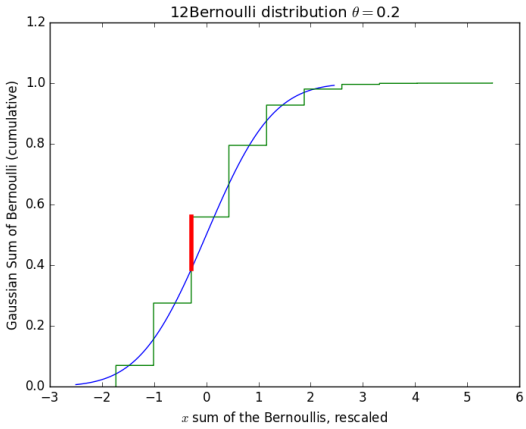

*Figure 1.4: Cumulative distribution function for the sum of $n = 12$ Bernoulli-distributed iid random variables with $\theta = 0.2$, cumulative distribution function of a Gaussian random variable with the same mean value and variance, and their Kolmogorov distance, that is, the maximum of the differences between the two (shown in red). This maximum scales as $1/\sqrt{n}$ for $n \to \infty$, because the Bernoulli distribution has a finite third absolute moment (see the Berry-Esseen theorem of eq. (**??**)).*

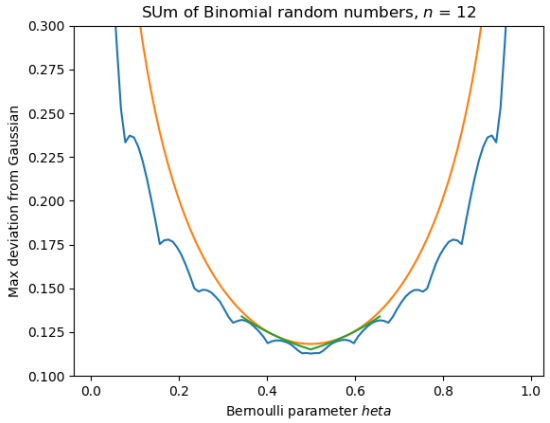

*Figure 1.5: Berry-Esseen probability for a sum of Bernoulli random variables with parameter θ and for $n = 12$. It is plotted the length of the maximum distance (the length of the red interval, plotted in Fig. **??** for $\theta = 0.2$, for all θ between 0 and 1. This is the blue curve. It is compared with the theoretical orange curve, which corresponds to eq. (1.79) with the constant eq. (1.80) proven in [**?**]. The other curve is a tighter limit which is valid only for $1/3 < \theta < 2/3$, also proven in [**?**].*

We will take a look at the Berry–Esseen theorem for a sum of Bernoulli random variables. In this case, the third absolute moment is: $(1-\theta)^3\theta+\theta^3(1-\theta)$.

For the Bernoulli variables, the above equation becomes

$$\sup_z|\mathbb{P}(z_n\leq z)-\Phi(z)|<c_E\frac{\theta^2+(1-\theta)^2}{\sqrt{n\theta(1-\theta)}}, \tag{1.79}$$

where it is clear that the right-hand side of eq. (1.78) is unbounded for θ approaching zero or one. In contrast, the constant c_E can be bounded. Decade-long research has culminated in strict bounds for general distributions, and a tight result for the case of Bernoulli variables. Theorem 1 of [**?**], proves for example that

$$c_E=\frac{\sqrt{10}+3}{6\sqrt{2\pi}} \tag{1.80}$$

This is a statement valid for all n, and it is illustrated for $n=12$ in Fig. 1.4 and Fig. 1.5.

All this means that the Gnedenko–Kolmogorov theorem assures of convergence to the Gaussian distribution for iid variables under the condition that the variance is finite. But then the rate of convergence can be arbitrarily bad. On the other hand, if there is one more power of moment, namely the absolute third moment, than the convergence is for all n at a speed $1/\sqrt{n}$. So the situation is analogous to the situation for the weak law of large numbers, and the Chebychev inequality. The Bernoulli distribution appears as a bad case, already.

1.5 Stable distributions

Definition: A non-degenerate distribution f_ξ is a stable distribution if it satisfies: let ξ_1 and ξ_2 be independent copies of a random variable ξ (they have the same distribution f_ξ). Then π_ξ is said to be stable if for any constants $a>0$ and $b>0$ the random variable $a\xi_1+b\xi_2$ has the distribution $\pi_{c\xi+d}$ for some constants $c>0$ and d.

[MEDIUM] Prove that the Gaussian $\pi_\xi(x)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$ is a stable distribution. Remember the propriety of the generating functions

$$\Phi_{a\xi+b}(t)=\mathrm{e}^{ibt}\Phi_\xi(at) \tag{1.81}$$

The generating function $\phi^{(G)}(t)$ of a Gaussian distribution is given by

$$\Phi^{(G)}(t)=\mathrm{e}^{-(2\sigma^2)t^2} \tag{1.82}$$

We consider the characteristic function of the sum of the two random variables $a\xi+b\xi$, which we know to be given by the product of the characteristic functions of the single (Gaussian) distributions

$$\Phi_{a\xi+b\xi}(t)=\Phi^{(G)}_{a\xi}(t)\Phi^{(G)}_{b\xi}(t)=\mathrm{e}^{-(2\sigma^2)t^2(a^2+b^2)}=\Phi^{(G)}_{\sqrt{a^2+b^2}\xi}(t) \tag{1.83}$$

We have then shown that the distribution of the sum of the two Gaussian random variables $a\xi+b\xi$ is a Gaussian distribution of the variable $c\xi$ with $c=\sqrt{a^2+b^2}$. Therefore the Gaussian distribution is stable.

[EASY] Consider a characteristic function of the form

$$\Phi_\xi(t) = \exp\left(it\mu - (c_0 + ic_1 f_\alpha(t))|t|^\alpha)\right), \tag{1.84}$$

with $1 \leq \alpha < 2$. Show that $f_\alpha(t) = \mathrm{sgn}(t)$, for $\alpha \neq 1$, and $f_1(t) = \mathrm{sgn}(t)\log|t|$ produce stable distributions. These are also known as *Lévy distributions*, after Paul Lévy, the first mathematician who studied them. As before, we consider the combination of two random variables with a Lévi distribution has the characteristic function

$$\Phi_{a\xi_1+b\xi_2}(t) = \exp\left(it(a+b)\mu - (c_0 + ic_1 f_\alpha(t))|(a+b)^{1/\alpha}t|^\alpha)\right).) \tag{1.85}$$

If $\alpha \neq 1$ this is mapped into the same distribution by the transformation $t \to (a+b)^{-1/\alpha}t$ and $\mu \to (a+b)^{1/\alpha-1}\mu$. For $\alpha = 1$ the transformation is $t \to (a+b)^{-1}t$ and $\mu \to \mu - \frac{c_1}{\alpha}\log(a+b)$.

[EASY] Find a distinctive feature of the Lévy distributions. The second cumulant

$$\kappa_2 = (-i)^2\partial_t^2 \log\Phi(t)\Big|_{t=0} \sim \mathrm{sign}(t)|t|^{\alpha-2} + \ldots\Big|_{t=0} = \infty \tag{1.86}$$

as $\alpha < 2$, it diverges.

[EASY] Assumes $\alpha \neq 1$ and show that, in order to be $\Phi_\xi(t)$ the Fourier transform of a probability distribution, the coefficient c_1 can not be arbitrarily large; determine its maximal value.

[Hint 1:] One can show (MEDIUM-HARD) that the inverse Fourier transform of (1.84) has the tails

$$f_\xi(x) \xrightarrow{|x|\gg 1} \frac{\Gamma(1+\alpha)}{2\pi|x|^{1+\alpha}}\left[c_0 \sin\frac{\pi\alpha}{2} - c_1\mathrm{sgn}(x)\cos\frac{\pi\alpha}{2}\right],. \tag{1.87}$$

The probability distribution must be positive or equal to zero, therefore the coefficients of the tails of the Lévi distributions must be positive. Since

$$f_\xi(x) \xrightarrow{|x|\gg 1} \frac{\Gamma(1+\alpha)}{2\pi|x|^{1+\alpha}}\left[c_0 \sin\frac{\pi\alpha}{2} - c_1\mathrm{sgn}(x)\cos\frac{\pi\alpha}{2}\right], \tag{1.88}$$

we find

$$|c_1| < c_0\left|\tan\frac{\pi\alpha}{2}\right| \tag{1.89}$$

In lecture 1 and tutorial 1, we discussed and derived Lévy distributions: Universal (stable) distributions that have infinite variance. A good example for producing such random variables is from uniform random numbers between 0 and 1, $\mathrm{ran}(0,1)$ taken to a power $-1 < \gamma < -0.5$. Such random numbers are distributed according to a distribution

$$f_\xi(x) = \begin{cases} \frac{\alpha}{x^{1+\alpha}} & \text{for } 1 < x < \infty \\ 0 & \text{else} \end{cases} \tag{1.90}$$

where $\alpha = -1/\gamma$ (you may check this by doing a histogram, and read up on this in SMAC book).

1. Is the probability distribution of eq. (1.90) normalized for $\gamma = -0.8$ (that is $\alpha = 1.25$), is it normalized for $\gamma = -0.2$ (that is $\alpha = 5$)?

2. What is the expectation of the probability distribution for the above two cases, and what is the variance?

3. Write a (two-line) computer program for generating the sum of 1000 random numbers with $\gamma = -0.2$, and plot the empirical histogram of this distribution (that is, generate 1000 times the sum of 1000 such random numbers. Interpret what you observe. For your convenience, you may find a closely related program on WK's website. Modify it so that it solves the problem at hand, and adapt the range in the drawing routine. Produce output and discuss it.

4. Write a (two-line) computer program for generating the sum of 1000 random numbers with $\gamma = -0.8$, and plot the empirical histogram of this distribution. Interpret what you observe. For your convenience, please take the closely related program from WK's website. Modify it so that it solves the problem at hand, and adapt the range in the drawing routine. Produce output and discuss it.

1.6 Homework 1

1.6.1 Rényi's formula for the sum of uniform random numbers, variations

In tutorial 1, you derived Rényi's formula for the sum of uniform random numbers between -1 and 1:

$$\pi_n(x) = \begin{cases} \frac{1}{2^n(n-1)!} \sum_{k=0}^{\left[\frac{n+x}{2}\right]} (-1)^k \binom{n}{k} (n+x-2k)^{n-1} & \text{for } |x| < n \\ 0 & \text{else} \end{cases} \tag{1.91}$$

1. Compute the variance of the distribution of eq. (1.91) for $n = 1$, that is for uniform random numbers between -1 and 1.

2. Compute the variance of Rényi's distribution for general n (Hint: this can be computed in 1 minute, if you use a result presented in the lecture).

3. Implement eq. (1.91) in a computer program for general n. For your convenience, you will find such a computer program on WK's website. This program also computes $P(X > \epsilon)$. Download this program and run it (in Python2). Notice that you may change the value of n in this program.

4. Modify the program (plot) so that it compares $P_n(X > \epsilon)$ to the upper limit given by the Chebychev inequality (Attention: you may modify Chebychev's inequality to take into account that $\pi_n(x)$ is symmetric around $x = 0$). Comment.

5. Modify the program (plot) so that it compares $\mathbb{P}_n(\xi > \epsilon)$ to the Cantelli inequality:

$$\mathbb{P}\left[\xi - \mathbb{E}(\xi) > \epsilon\right] \leq \frac{\text{Var } \xi}{\text{Var } \xi + \epsilon^2} \tag{1.92}$$

(note that there are now no absolute values). Comment.

6. Modify the program so that it compares $\mathbb{P}(\xi > \epsilon)$ to Hoeffding's inequality. Hoeffding's inequality considers a probability distribution with zero expectation and $a_i \leq \xi_i < b_i$ (we will later take constant bounds a and b, but in fact, they may depend on i). For every $t > 0$, it states:

$$\mathbb{P}(\sum_{i=1}^{n} \xi_i \geq \epsilon) \leq \exp(-t\epsilon) \prod_{i=1}^{n} \exp\left[t^2(b_i - a_i)^2/8\right]. \tag{1.93}$$

Is Hoeffding's inequality *always* sharper than the Chebychev inequality, that is, is Hoeffding with the best value of t better than Chebychev for all ϵ? What is the asymptotic behavior for $\epsilon \to \infty$ behavior of Hoeffding's inequality, and why does it satisfy such a stringent bound if the Chebychev inequality does not achieve it? Return a plot that contains, next to $\pi_n(x)$ and its integral $P_n(X > \epsilon)$, the comparison with Chebychev, Cantelli, and Hoeffding.

Lecture 2

Statistics

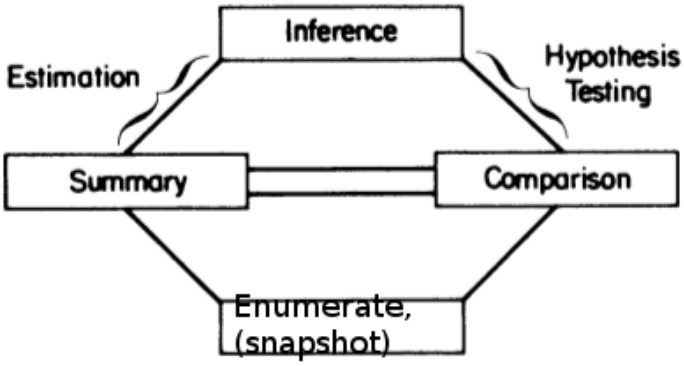

Figure 2.1: Four operations in statistics: "Enumeration" (snap-shotting), summary, comparison, and inference.

We remember from Lecture 1 that probability theory attempts to describe the properties of outcome given a data generating process, while statistics attempts to make statements about the process that generates the data. However, this is not all that statistics is about.

Statistics: "Given the outcomes, what can we say about the process that generated them?" (see Fig. **??**) but, as pointed out by Efron, it is more precise to state statistics is the science of (finite amounts) of data, and it can be broken up into four concepts.

As stressed by Efron 1982[9], we can often add to this statement: "Given the outcomes, how can we summarize data?"

The simplest operation that can be done with data is to keep them all, and look at them. This naive treatment, that one can call "enumeration", is very common. It consists in looking at individual data (or sets of data). Many research papers present a snapshot of "typical" data, and an explanation of the phenomena that one sees in them.

simply consists in obtaining the data and carrying them around with us. What appears a very basic approach to data is part of many publications (which love to show "snapshots".

Summarization is an often overlooked aspect of data treatment. It consists in describing N data not through N pictures of them, but through a summary, often a finite (and small) set of parameters of a statistical model. The inference problem is the question of how well the summary that worked for N data points would fare if we had a much larger number of data points. Also: What would be the parametes of the infinite data set, if we only have a finite number of them. "Comparison" is the opposite of the summary: We try to take the data and try to separate them (for example into two sets) to make differences appear. Again with reference to an infinite pool, we pose the question (of hypothesis testing) whether the two sets are indeed different.

Probability theory and statistics are the same when there is an infinite number of data (having all the data allows to construct the probability distribution), and oftentimes, statistics can be understood as an attempt (sometimes a quite desperate attempt) to make an infinite number of data from a finite number ofdata.

Finally, it is useful

Note: Mathematical definition: "A statistic" means "a function of data".

2.1 Parametric statistics

2.1.1 Models in statistics

In mathematics, a (statistical) model is defined as a set $\mathcal{F}$ of distributions. If this set can be described by a finite number of parameters $\theta = \{\theta_1, \ldots, \theta_k\}$:

$$\mathcal{F} = \{f(x, \theta) : \theta \in \Theta\}, \tag{2.1}$$

(where Θ is some parameter space) then $\mathcal{F}$ is a parametric model. Otherwise it is termed nonparametric.

We notice that in mathematics, the definition of what constitues a model is clear-cut. In physics, we have a much harder time with this: We can define precisely the Ising model, the XY model, the standard model of fundamental interactions, the hard-sphere model, etc, but generally are inprecises about what a "model in physics" actually means. For statistical-physics models, the understanding of the role of models has come from the renormalization group (see Section **??**). In a dramatic change to what was believed, say, before the 1970s, where the role of models in statistical physics was likened to what can be called "toy-models", statistical-physics model are now understood to exactly describe the long-wavelength behavior of entire classes of physical systems. One given class can comprise very different physical systems as, for example, classical spin systems and superconductors.

Parametric statistics poses the question of what justifies the use of parametric models (see [1, p. 119]). The response to this (good) question is surprisingly complex. Efron [9] (see Fig. 2.1)

One of the reasons is that much of statistics is "descriptive". This means that it is

- In some cases, it can be proven, or is otherwise reasonable. For binary data, we can for example sometimes prove that the data come from a Bernoulli distribution. We can also have some information on the data, for example we might

know that it decays exponentially. For the velocity distribution in a gas, we may know that the distribution is a Gaussian, but we do not know the temperature, that is, the width of the distribution

- As stressed by Efron, we have to do something with data: After enumeration (keeping all the data in our pocket), we need to summarize them.
- If we have summarized them, we can actually use the summarization to generate infinite amounts of data. This means that the parametric model can always be used for estimation.
- Non-parametric models don't constrain the data as the parametric model, but they also do not add any supplementary information that we might have.

If maximum likelihood estimation acts as if the maximum likelihood summary is exactly true, how can the theory provide an assessment of estimation error? The answer is simple but ingenious (Efron 1982, section 5).

2.1.2 Method of moments

This is a point estimate, even though the very idea of a point estimate is not as strict as it may seem: As soon as you have your parameter θ, you immediately have an infinite number of samples (or of sets of samples).

This is what the physicist knows: Let us suppose we $\Theta : \Theta_1, \ldots, \Theta_k$, which makes k parameters. From these parameters, we now go to the moments of the distribution

$$\alpha_j = \mathbb{E}\left(\xi^j\right) \quad \text{for } j = 1, \ldots k \tag{2.2}$$

and we compare these moments of the distribution to the

$$\hat{\alpha}_j = \frac{1}{n}\sum_i \xi_i^j. \tag{2.3}$$

The method-of-moment estimator is that which satisfies:

$$\alpha_1(\Theta_1, \ldots, \Theta_k) = \hat{\alpha}_1 \tag{2.4}$$

$$\vdots \qquad \vdots$$

$$\alpha_k(\Theta_1, \ldots, \Theta_k) = \hat{\alpha}_k \tag{2.5}$$

Example: Bernoulli distribution

The Bernoulli distribution has a single parameter, namely θ, so that:

$$\alpha_1 = \mathbb{E}(\xi) = \theta, \tag{2.6}$$

$$\hat{\alpha}_1 = \frac{1}{n}\sum_i \xi_i, \tag{2.7}$$

from which follows $\theta_n = \frac{1}{n}\sum_i \xi_i$. As we know that Bernoulli-distributed random variables have a finite expectation value (in addition to a finite variance), so that they satisfy the weak law of large numbers, we see automatically that the method of moments is consistent in this case (that is $\theta_n \xrightarrow{\text{in prob.}} \theta$).

We will get back to this interesting case for two reasons: One to show what we actually learn from p_n for finite n.

Example: Gaussian distribution

Properties of the method of moments

Properties of the method of moments. If the stuff actually works, then

1. $\hat{\Theta}_n$ exists
2. $\hat{\Theta}_n \to \Theta$, which means that it is consistent.
3. It is asymptotically normal, which means that it can be used, in principle principle principle, to obtain confidence intervals.

With all these nice properties, the method of moments has problems. Although it is consistent, it is not optimal. It requires higher moments of the distribution to exist. Also, it is not reparametrization independent: that is, if we describe our probability distribution through angles and radii rather than (x, y) pairs, we will get different results.

FIXME {I would suppose that the theorem of the lazy statistician makes that the above statement is not true for a function that only depends on a single parameter, that is, for the Bernoulli distribution).}

2.1.3 Maximum likelihood estimation

The maximum likelihood estimation, created single-handledy by the statistician/biologist Ronald Fisher (1890-1962) addresses the problem of estimating parameters of the family of ... The method is a point estimate, and it is quite fishy.

Let us suppose we have 15 data points.

FIXME {Avoid examples} Suppose that we have n data points $x_1, x_2, x_3, , ..., x_n$ (these points are real numbers between $-\infty$ and ∞), and we know that they are drawn from a Gaussian distribution with unknown values of the variance σ^2 and the mean value $\langle x \rangle$:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-(x - \langle x \rangle)^2/(2\sigma^2)). \tag{2.8}$$

What is the maximum likelihood estimator for the mean value $\langle x \rangle$ and variance σ^2 of this Gaussian distribution from the data?

Hint1 Remember that the likelihood function is given by $p(x_1)p(x_2)....p(x_n)$.

Hint2 If you use the log likelihood function, explain why this can be done.

Carefully explain your calculation.

Note: $L(\alpha)$ is not a probability distribution.

Maximum likelihood and data summarization

Fisher information

2.2 Non-parametric statistics

2.2.1 Glivenka–Cantelli theorem

This theorem states, very simply and in full generality, that if $\xi_1, \ldots, \xi_n \sim F$, then

$$\sup_x |\hat{F}_n(x) - F(x)| \to 0 \tag{2.9}$$

where the convergence is in probability. This theorem is just the tip of the iceberg.

2.2.2 DKW inequality

$$\mathbb{P}\left[\sup_x |F(x) - \hat{F}_n(x)| > \epsilon\right] \leq 2\exp\left(-2n\epsilon^2\right) \tag{2.10}$$

FIXME {Please take a look at the fact sheet of week 3, that discusses a paper by Massart [10].}

2.2.3 The bootstrap

A bootstrap is a loop of sturdy material, often leather, fixed to the top line of shoes, above their heels. Holding on to both your bootstraps, you will lift yourself up into the air. This is called "boot-strapping" [1]. More generally, boot-strapping describes super-simple solutions to problems that appear hopeless: Lifting yourself over a river when there is no bridge (simply pull on the bootstraps and up you go), extracting yourself and your horse out of a swamp (simply hold on to your horse and pull on your pony-tails), starting a computer—that needs an operating system to run—after the operating system crashed (simply "reboot" your computer).

As we discussed repeatedly, the separation of probability theory and of statistics lies in the $1/\sqrt{n}$ factor related to the finite number of data. The bootstrap method turns a finite number of samples into an infinite number, by simply using the cumulative density $\hat{F}(x)$ instead of the unknown $F(x)$. The strategy consists in placing all the samples in a barrel, and in taking data. It finds its mathematical justification in the DWK inequality (and many other statements). However it is even more clever.

More seriously, the bootstrap is a non-parametric method for going from $N =$finite to $N = \infty$ (Wasserman: For estimating standard errors and computing confidence intervals). The method is so general that it can also be used with parametric methods.

[1] Advice: don't wave to your friends back on the ground. With your hands off your shoes, you will fall down instantly

Let

$$T_n = g(X_1, \ldots, X_n) \tag{2.11}$$

be a "statistic" (that is, a function of the data). Suppose we want to know the variance $\mathrm{Var}(T_n)$. Then simply replace $F \to \hat{F}_n$, that is, replace the true but unknown distribution F by the empirical function $\hat{F}_n$ that adds $1/n$ at each data point. The real ingenious trick here is that we always draw n points.

2.2.4 The jackknife

The jackknife is a T

2.3 Bayes' Statistics and the interpretation of probabilities

2.4 The children on the beach problem

Interval from Chebychev: $\mathbb{P}(|\Delta > \epsilon| < \mathrm{Var}/\epsilon^2$. This leads to $p = \mathrm{Var}/(N\epsilon^2)$ or to $\epsilon = \sigma/\sqrt{Np}$ and using the estimate $\sigma < 1/2$ leads to the conservative estimate $\epsilon = \frac{1}{\sqrt{N}} 1/(2\sqrt{p})$.

For Hoeffding, we have $p = 2\exp\left(-2N\epsilon^2\right)$, which sounds a lot sharper then Chebychev. It likewise leads to the estimate $\epsilon = \frac{1}{\sqrt{N}}\sqrt{-\log(p/2)/2}$. As shown in Fig. 2.2, we see that Chebychev is better than Hoeffing for large p, as for example the 68% confidence interval that is used in physics, but it is of course much better for smaller p.

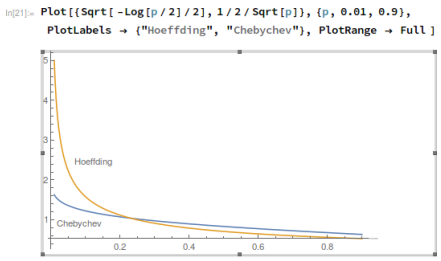

Figure 2.2: Comparison of Hoeffding and Chebychev for the children's game.

Brown Cai DasGupta discuss the standard interval. This means that one supposes that the binomial distribution is described approximately by a Gaussian normal distribution, and that the variance is indeed given by $\hat{p}(1-\hat{p})$ This leads to what is called the standard error bar $z(...)/\sqrt{n}\sqrt{\hat{p}\hat{q}}$. We can now do two plots

- Brown Cai DasGupta
- Wasserman p 65 (with Hoeffding)

2.5 Homework 02

In lecture 02, we treated the maximum likelihood approach as one of the key methods for estimating the parameters of a distribution. Here we treat two examples. The second one was of great historical importance on the battlefields of WW2 although it proved necessary to go one step farther than we will do here.

2.6 Uniform distribution

Preparation

What follows, a preparation for Section 2.7, is described in Wasserman as a hard example "that many people find confusing". It will not confuse You!

Application

Suppose a uniform distribution between 0 and θ, and consider k samples drawn from this distribution.

- What is the likelihood function $L(\theta)$ given $x_1, \ldots x_k$? (Hint: suppose that "the probability to sample x_i" is $1/\theta$. The $1/\theta$ factor is "physically" clear). Plot $L(\theta)$.
- What is the maximum-likelihood estimator of θ given $x_1, \ldots, x_k$, that is, the samples?
- Comment your finding.

2.7 German tank problem (Frequentist approach)

This example has been of considerable importance, first in WW2, then in the theory of statistics. It is the discrete version of Section 2.6.

2.7.1 Preparation

Consider N balls numbered $1, 2, 3, ..., N$, and take k out them (urn problem without putting them back). What is the probability $p_N(m)$ that the largest of them is equal to m?

Hint0 How many ways are there to pick k (different) balls out of N?

Hint1 To solve this simple combinatorial problem, consider that m must be contained in $k, k+1, k+2, ...N$.

Hint2 Count the number of ways to pick $(k-1)$ balls so that they are smaller than m.

Carefully explain your calculation.

2.7.2 Application

From an urn with an unknown number N of balls (tanks), the following $k = 4$ balls were drawn (without putting them back):

$$1, 2, 4, 14$$

What is the maximum likelihood estimator of the total number N of balls (tanks) (based on the probability distribution of the sample maximum m, here 14) that are contained in the urn (destroyed tanks left on the battlefield)?

The (disappointing) result of the maximum likelihood estimator (here in the famous "German tank problem") points to one of the limitations of the maximum likelihood method, namely that it presents a bias. Comment on this property. A trick allows to arrange the situation. In simple terms it consists in supposing that the mean of the intervals between the first ball and zero, the second and the first ball, the third and the second... etc is probably as large as the interval between the largest ball and N.

2.8 German tank problem (Bayesian approach)

The Bayesian approach treats the total number N of the balls (tanks) as a random variable, and it has been much studied in the literature. But to start, simply write a program for $k = 4$ that samples N from a discretized exponential distribution with parameter λ. Then sample *k different balls* from this distribution, if possible.

2.8.1 Maximum

Numerically compute the probability distribution of all the N for which the largest of the $k = 4$ balls is equal to 14 (see previous example). Do this by sampling: Sample N, then sample $k = 4$ balls, and enter N into a histogram if the largest of the 4 balls is equal to 14.

Plot this distribution (histogram), its expectation and variance for different values of λ. For your convenience you find the Python program, already 95% written, on the course website. Modify it (to compute the expectation and variance as a function of λ), and run it.

2.8.2 Total sample

Numerically compute the probability distribution of all the N for which the $k = 4$ balls exactly equal $1, 2, 4, 14$. Plot this distribution (histogram), its mean value and variance for different values of λ. Do these distribution differ (empirically) from the ones in Section 2.8.1? For your convenience, the Python program on the course website already contains the crucial modification. Is the outcome different from the one of the maximum version (Section 2.8.1)?

2.9 Midterm2017

2.9.1 Method of moments and Maximum likelihood

Consider n independent samples $X_1, \ldots, X_n$ drawn from a uniform distribution with bounds a and b, where a and b are unknown parameters and $a < b$.

$$X_1, \ldots, X_n \sim \text{Uniform}(a, b) \tag{2.12}$$

1. Explain what a method-of-moments estimator is. For n samples $X_1, \ldots, X_n$ and a probability distribution π depending on k parameters $(\theta_1, \ldots, \theta_k)$, the method-of-moments estimator is the value $\hat{\theta}$ such that the k lowest moments $\alpha_j = \int x^j \pi(x, \theta) dx$ agree with the sample moments $\hat{\alpha_j} = \sum_{i=1}^{n} X_i^j$. This is a system of k equations with k unknowns. The method of moments is not optimal, and sometimes the moments of the distribution do not exist, although the sample moments always exist. But the method of moments is easy to use.

2. Find the method-of-moments estimator for a and b. We need to solve

$$\frac{1}{b-a}\int_a^b x dx = \hat{\alpha}_1 \qquad \frac{1}{b-a}\int_a^b x^2 dx = \hat{\alpha}_2 \tag{2.13}$$

   equationequationequationequatioequationequation equation equation equation equation equation equation equation equation equation equationn Therefore we have $(a+b)/2 = \hat{\alpha_1}$ and $\frac{1}{3}(b^3 - a^3)/(b-a) = \frac{1}{3}(a^2 + ab + b^2) = \hat{\alpha}_2$, one finds $b = \hat{\alpha_1} + \sqrt{3}\sqrt{\hat{\alpha_2} - \hat{\alpha_1}^2}$ and $a = \hat{\alpha_1} - \sqrt{3}\sqrt{\hat{\alpha_2} - \hat{\alpha_1}^2}$.

3. Explain the essence of maximum-likelihood estimation, and discuss the difference between a likelihood and a probability. Using the same definitions as above, the likelihood function is defined as the product over the probabilities $\mathcal{L}(\theta) = \prod_{i=1}^{n} \pi(X_i, \theta)$, and the maximum likelihood estimator is the value of the parameters θ that maximizes this value, as a function of the data.

4. Find the maximum-likelihood estimator $\hat{a}$ and $\hat{b}$. Because of the normalization $(b-a)$, the maximum likelihood estimator is largest if the interval $(b-a)$ is smallest. Therefore, $b = \max(X_i)$ and $a = \min(X_i)$.

2.10 Interval tests for the Bernoulli distribution

Statistics is the science of data and their data-procuding processes. As mentioned, the goal of descriptive statistics is often "summarization", that is, the characterization of often vast amounts of data with few numbers. Examples are the maximum likelihood summary of data (see [9]) and the five-figure summary (the minumum, the maximum, the median, the first quartile (25%) and the last quartile (75%) of data).

In contrast to descriptive statistics, the theorems of statistical inference rely on conditions for the data-producing process. All these conditions translate, on a one-to-one basis, into corresponding assumptions about real data. These conditions are usually:

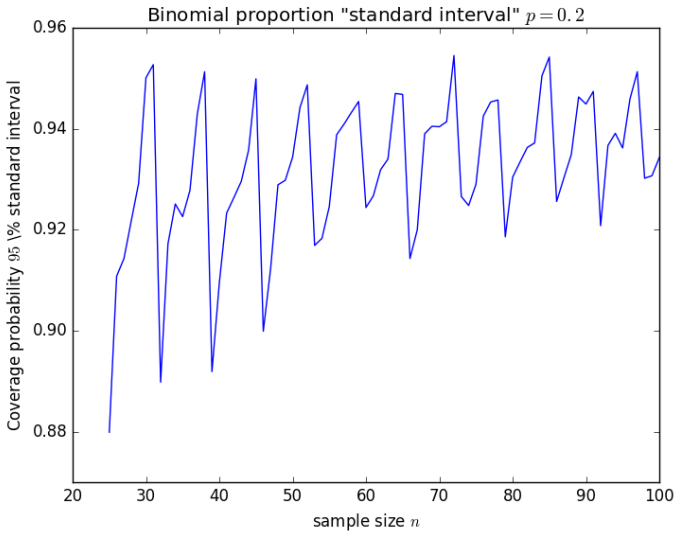

Figure 2.3: Standard interval.. Connect to the reference [11] [12]

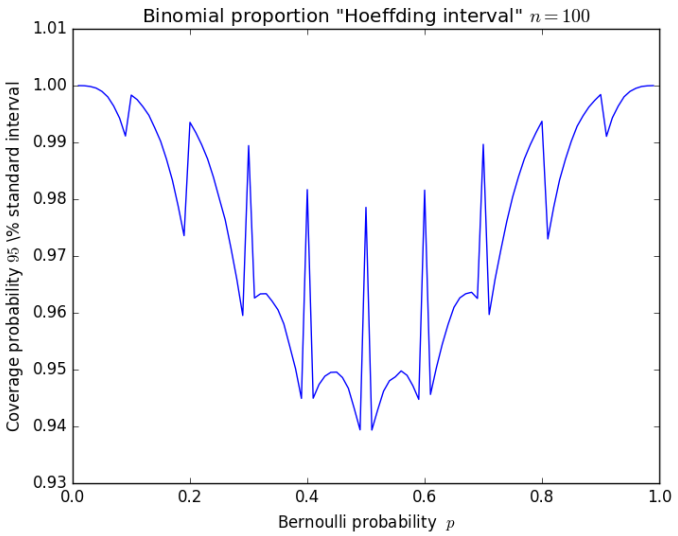

Figure 2.4: Hoeffding interval ... connect to the reference...

Independence of random variable

Identity of distribution distribution This is the question of whether the random variables come from the same distribution

Existence of moments of random variables

Type of distribution Because of the existence of limit theorems, this condition is sometimes confounded with the one of how close one is to the limit distribution, at is, a Gaussian or one of Lévy's stable laws.

In statistical physics, we in addition have the problem of the finite-size scaling towards the thermodynamic limit. This means that while the distribution F_ξ^N of the random variables of an finite system (for example, with N particles or spins) is sampled, one is interested in the properties of the data-generating process (that is, the physical system) for $N \to \infty$.

2.10.1 Frequentist vs. Bayesian statistics

We now consider an example of the sample space Ω consisting of the square (x, y) with $x \in [-1, 1]$ and $y \in [-1, 1]$. Pebbles (see [6]) are randomly thrown into this square producing the samples ω. Bernoulli-distributed random variable ξ equal one if $|(x, y)_\xi| < 1$ and 0 otherwise. Clearly, the parameter θ equals $\pi/4$, but the trouble is that we totally forgot the value of Pythagoras' number, but threw 4000 independent random variables.

$$x_{\xi_i} = \begin{cases} 1 & \text{with probability } \theta \\ 0 & \text{with probability } 1-\theta \end{cases} \tag{2.14}$$

The result of 4000 independent throws was 3156 "hits", where the pebble landed inside the circle, and the remaining 844 "no-hits", where it ended up outside. We suppose that the data ξ_i are i.i.d random variables, we also know that they are Bernoulli, and that their variance $\theta(1-\theta) \leq 1/4$.

If we insist on a 0.95 covering interval, we will have a covering interval of width 0.282 (that is 0.141 to the left and to the right).

If we now use the standard normal approximation of the Bernoulli distribution and the maximum-likelihood approximation of the standard deviation $1/\sqrt{\hat{\theta}(1-\hat{\theta})}$ we suppose that in our example we have $\sqrt{\hat{\theta}(1-\hat{\theta})} = 0.408$ and the Gaussian error interval for $p = 0.05$ equals $1.96 \times 0.408/\sqrt{4000} = 0.0126444$, which is a whole lot smaller than the 0.141 covering element, and this, for the same data. Now, let us check, for some data how good the covering interval of all these approximations really is. To do so, we use numerical simulation in the following way: For a very fine grid of values of θ, we do a Bernoulli experiment, and compute the value $\hat{\theta}$ as well as the error bar (confidence interval) computed by each of the methods. For added generality, we do not simply test this outcome for a given value of n, but rather as a function of n.

We can also use the Hoeffding estimate, where we have $p = 2\exp\left(-2N\epsilon^2\right)$, which sounds a lot sharper then Chebychev. It likewise leads to the estimate $\epsilon = \frac{1}{\sqrt{N}}\sqrt{-\log\left(p/2\right)/2}$.

Bibliography

[1] L. Wasserman, *All of Statistics*. New York: Springer, 2004.

[2] A. Tversky and D. Kahneman, "Judgment under Uncertainty: Heuristics and Biases," vol. 185, no. 4157, pp. 1124–1131, 1974.

[3] P. L. de Micheaux and B. Liquet, "Understanding Convergence Concepts: A Visual-Minded and Graphical Simulation-Based Approach," *The American Statistician*, vol. 63, no. 2, pp. 173–178, 2009.

[4] E. Seneta, "A Tricentenary history of the Law of Large Numbers," *Bernoulli*, vol. 19, no. 4, pp. 1088–1121, 2013.

[5] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Times*. American Mathematical Society, 2008.

[6] W. Krauth, *Statistical Mechanics: Algorithms and Computations*. Oxford University Press, 2006.

[7] V. Y. Korolev and I. G. Shevtsova, "On the Upper Bound for the Absolute Constant in the Berry–Esseen Inequality," *Theory of Probability & Its Applications*, vol. 54, no. 4, pp. 638–658, 2010.

[8] V. Korolev and I. Shevtsova, "An improvement of the Berry–Esseen inequality with applications to Poisson and mixed Poisson random sums," *Scandinavian Actuarial Journal*, vol. 2012, no. 2, pp. 81–105, 2012.

[9] B. Efron, "Maximum Likelihood and Decision Theory," *Ann. Statist.*, vol. 10, no. 2, pp. 340–356, 1982.

[10] P. Massart, "The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality," *The Annals of Probability*, vol. 18, no. 3, pp. 1269–1283, 1990.

[11] L. D. Brown, T. T. Cai, and A. DasGupta, "Interval Estimation for a Binomial Proportion," *Statistical Science*, vol. 16, no. 2, pp. 101–117, 2001.

[12] A. Agresti and B. A. Coull, "[Interval Estimation for a Binomial Proportion]: Comment," *Statistical Science*, vol. 16, no. 2, pp. 117–120, 2001.

[13] M. E. Fisher, S.-k. Ma, and B. G. Nickel, "Critical Exponents for Long-Range Interactions," *Phys. Rev. Lett.*, vol. 29, pp. 917–920, 1972.

[14] F. J. Dyson, "Existence of a phase-transition in a one-dimensional Ising ferromagnet," *Communications in Mathematical Physics*, vol. 12, no. 2, pp. 91–107, 1969.

[15] D. J. Thouless, "Long-Range Order in One-Dimensional Ising Systems," *Physical Review*, vol. 187, no. 2, pp. 732–733, 1969.

[16] J. M. Kosterlitz, "Phase Transitions in Long-Range Ferromagnetic Chains," *Phys. Rev. Lett.*, vol. 37, pp. 1577–1580, 1976.

[17] R. Peierls, "On Ising's model of ferromagnetism," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 32, pp. 477–481, 1936.

[18] L. Onsager, "Crystal Statistics. I. A Two-Dimensional Model with an Order-Disorder Transition," *Physical Review*, vol. 65, pp. 117–149, 1944.

[19] T. Schultz, E. Lieb, and D. Mattis, "Two-Dimensioal Ising Model as a Soluble Problem of Many Fermions," *Reviews of Modern Physics*.

[20] M. Plischke and B. Bergersen, *Equilibrium Statistical Physics*. Hong Kong ?: World Scientific, 2006.

[21] H. A. Kramers and G. H. Wannier, "Statistics of the Two-Dimensional Ferromagnet. Part I," *Physical Review*, vol. 60, pp. 252–262, 1941.

[22] B. Kaufman, "Crystal Statistics. II. Partition Function Evaluated by Spinor Analysis," *Physical Review*, vol. 76, pp. 1232–1243, 1949.

[23] A. E. Ferdinand and M. E. Fisher, "Bounded and Inhomogeneous Ising Models. I. Specific-Heat Anomaly of a Finite Lattice," *Physical Review*, vol. 185, pp. 832–846, 1969.