Probability – Survey Lecture
(sketchy notes for the SFI/CEU Summer School)

Bálint Tóth
Institute of Mathematics
TU Budapest

June 3, 2001

Warning: These are very sketchy notes. I did not care at all about style. **In the two hours of my survey lecture I will concentrate on limit theorems, such as laws of large numbers, central limit (and other limit) theorems. I will not have time for the completely elementary introduction. Please read the first cca. 50 pages of these notes, before the summer school. So, in the two hours survey I will be able to present some more interesting stuff (the content of the second half of these notes).**

1 Introduction

1.1 Sample space + algebra of events + probability = probability space

Here are some very simple 'typical probabilistic' statements (and questions) formulated in ordinary everyday language. Let's try to analyze them:

(A) Tossing a coin the chances for HEAD to turn up are 1:2.

(B) Throwing two dice the chances for getting at least one ace are 11:36.

(C) In the classroom there is a quantity of gas (air) in thermal equilibrium. The probability of finding all the gas molecules gathered in the left half of the classroom, is extremely small.

(D) On the square grid $\mathbb{Z}^2$ (streets of Manhattan) a random walker (drunken person) starts from the origin (leaves the pub) and walks in discrete time units from site to site (corner to corner) so that at each stage (s)he

chooses the next visited site completely randomly ((s)he is completely drunk). What is the probability that (s)he returns to the origin (to the pub) whithin n steps? Does (s)he surely return at all? How about the analogous problem on the cubic lattice $\mathbb{Z}^3$?

(E) Sites of the square grid $\mathbb{Z}^2$ are coloured green (1, open), respectively, red (0, closed) with probability p, respectively $1-p$, independently. What is the probablility that there is an infinite green (open) path starting from the origin of the grid?

In order to describe mathematically these situations first we define the **sample space**, that is, the set of all possible outcomes of the 'experiment'. Denote the sample space by Ω. For the five concrete setups the sample space is:

(A) $\Omega := \{H, T\}$

(B) $\Omega := \{1, 2, \ldots, 6\} \times \{1, 2, \ldots, 6\}$

(C) $\Omega := \{(\underline{r}_1, \underline{r}_2, \ldots, \underline{r}_N) : \underline{r}_i = (x_i, y_i, z_i) \in \Lambda, i = 1, 2 \ldots, N\}$
where $\Lambda = [0, L_1] \times [0, L_2] \times [0, L_3] \subset \mathbb{R}^3$ is the room, in Euclidean coordinates and N is the number of molecules in the room

(D) $\Omega := \{(\omega_i)_{i=1}^{\infty} : \omega_i \in \{\uparrow, \rightarrow, \downarrow, \leftarrow\}\} = \{\uparrow, \rightarrow, \downarrow, \leftarrow\}^{\mathbb{N}}$

(E) $\Omega := \{(\omega_x)_{x \in \mathbb{Z}^2} : \omega_x \in \{0, 1\}\} = \{0, 1\}^{\mathbb{Z}^d}$

Cases (A) and (B) are trivial. In case (C) $\underline{r}_i$ denotes the position of the i-th molecule, in Euclidean coordinates. In caze (D) ω_i denotes the i-th step of the random walker. In case (E) ω_x denotes the state (closed or open) of the site $x \in \mathbb{Z}^2$. The possible outcomes of the experiment are **sample points**, i.e., elements of the sample space.

We ususlly speak about probability of events. **Events** are (almost) arbitrary agregates of possible outcomes. That is: subsets of the sample space.

In our concrete cases we spoke about the following events:

(A) $A = \{H\} \subset \Omega$

(B) $B = \{(6,1),(6,2),\ldots,(6,6),\ldots,(2,6),(1,6)\} \subset \Omega$

(C) $C = \{(\underline{r}_1, \underline{r}_2, \ldots, \underline{r}_N) \in \Omega : x_i \in [0, L_1/2], i = 1, 2 \ldots, N\}$

(D) $D_n = \{\omega \in \Omega : \exists k \leq n \text{ s.t. } \omega_1 + \omega_2 + \cdots + \omega_k = 0\}$

$D_\infty = \{\omega \in \Omega : \exists k \leq \infty \text{ s.t. } \omega_1 + \omega_2 + \cdots + \omega_k = 0\}$

(E) $E = \{\omega \in \Omega :$ there exists an infinite self avoiding lattice path $0 = x_0, x_1, \ldots$ s.t. for all $j \in \mathbb{N} \quad \omega_{x_j} = 1\}$

You notice that in the first three cases the events considered are rather simple to formulate, in the last two cases this is slightly more tricky.

In general, the collection of (meaningful) events $\mathcal{F}$ is a subset of the power set of the sample space: $\mathcal{F} \subset \mathcal{P}(\Omega)$. In simple combinatorial cases, i.e., when the sample space is either finite or countably infifnite (e.g. cases (A) and (B) above) we can consider all possible subsets of Ω, as events (that is: $\mathcal{F} = \mathcal{P}(\Omega)$). When the sample space is uncountable (e.g. cases (C), (D) and (E) above) more care is needed. In the present lecture we shall *completely disregard* these technical subtlties, just warn the student that some nontrivial technical problems may arise here, which are subject of the so called *measure theroetic* foundations of probability theory.

The collection of events $\mathcal{F}$ is closed under the natural set theoretical operations. It satisfies the following axioms:

(Ai) $\Omega \in \mathcal{F}$

(Aii) If $A, B \in \mathcal{F}$ then $A \setminus B \in \mathcal{F}$

(Aiiia) If $A, B \in \mathcal{F}$ then $A \cup B \in \mathcal{F}$

Note that from these axioms it follows that $\mathcal{F}$ is closed under all natural set theroetical operations. In particular, if $A, B \cdots \in \mathcal{F}$ then $A^c := \Omega \setminus A \in \mathcal{F}$, $A \cap B \in \mathcal{F}$, etc. By induction it follows that any finite number of these operations are allowed within $\mathcal{F}$. But it does not follow that $\mathcal{F}$ is closed under formation of countable union. This subtlety is interesting only when the sample space is infinite. In this case we have to replace (iiia) by

(Aiiib) If $A_j \in \mathcal{F}$, $j \in \mathbb{N}$ then $\bigcup_j A_j \in \mathcal{F}$

Elements of $\mathcal{F}$ are called *events* or *measurable subsets* of Ω. The two special event $\Omega \in \mathcal{F}$ and $\emptyset \in \mathcal{F}$ are called the *sure*, respectively, the *impossible* event. $\mathcal{F}$, with the natural set theoretical operations forms a *sigma-algebra*. (the term 'sigma' refers to the fact that countable unions are allowed.)

We spoke about *probability* of events. The **probability** is a weight assigned to each event. That is: $\mathbf{P} : \mathcal{F} \to [0,1]$ is a function defined on the domain $\mathcal{F}$ with range $[0,1]$. It satifies the following very natural axioms.

(Pi) $\mathbf{P}(\Omega) = 1$. That is: the sure event has full weight.

(Piia) If $A, B \in \mathcal{F}$ and $A \cap B = \emptyset$ then $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$. That is, the probability is *additive*.

In case of infinite sample space we have to allow for countable-additivity (or σ-additivity):

(Piib) If $A_j \in \mathcal{F}$, $j = 1, 2, \ldots$ and for any $i \neq j$ $A_i \cap A_j = \emptyset$ then $\mathbf{P}(\bigcup_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} \mathbf{P}(A_j)$. That is: the probability is countably additive (or σ-additive.

The asignment of the weights (probabilities) to events is actually the definition of the problem to be considered. In many cases there is some natural *symmetry consideration* leading to some natural asignment of probabilities. In most combinatorial and geometric problems the probability weight is *uniformly* distributed among the outcomes. In these special cases the probability of the arbitrary event A is given by the simple formula

$$\mathbf{P}(A) = \frac{|A|}{|\Omega|},$$

where $|\cdot|$ denotes the natural measure of size, cardinality in discrete cases, volume in continuous cases. But we should em[phasize that this is not a universally valid formula. We apply the 'uniform assignment of weights' only if we have good reason (symmetry) to do so.

Generally, if the sample space Ω is discrete (finite or countable), we can assign weights to *sample points*:

$$p : \Omega \to [0,1], \quad \text{such that} \quad \sum_{\omega \in \Omega} p(\omega) = 1,$$

and define the probability of the events $A \subset \Omega$ by

$$\mathbf{P}(A) := \sum_{\omega \in A} p(\omega).$$

Clearly, $\mathbf{P}(\cdot)$ defined this way will satisfy the axioms of probability function. And vice versa, on discrete sample spaces, any probability function $\mathbf{P} : \mathcal{P}(\Omega) \to [0,1]$ uniquely defines a weight function $p : \Omega \to [0,1]$, by $p(\omega) := \mathbf{P}(\{\omega\})$. But we should emphasize that this procedure does not work in case of *continuous* sample spaces. Then, typically, each single sample point will have zero weight and the probability function can not be built up from elementary atomic weights.

Back to our concrete cases:

(A) Assuming that the coin to be tossed is unbiased, by symmetry between H and T, we naturally assign equal weights to the elementary outcomes. So, $\mathbf{P}(A) = 1/2$, indeed.

(B) Assuming that the two dice are not leaded, again we assign uniformly, equal weights to all 36 possible outcomes. We get indeed $\mathbf{P}(B) = 11/36$.

(C) A (very much simplified) mathematical model of thermal equilibrium could be: uniform distribution of the sample point in the space $\Omega = \Lambda^N$. Simple computation leads to $\mathbf{P}(C) = 2^{-N}$. Given N of the order 10^{26} this event seems to be really very-very unlikely, though not impossible.

(D) This is much more complicated than the previous cases. One starts with assigning probabilities to the finite random walk trajectories, say of N steps. We asssign equal weight to all 4^N (or: in d-dimensions $(2d)^N$) N-step trajectories. Applying a far non trivial statement from abstract measure theory we conclude that taking the limit $N \to \infty$ this defines uniquely a probability measure on the set Ω of infinitely long trajectories. After some rather involved and beautyful arguments it turns out that the probability that the random walker *ever returns* to the origin is 1 in one and two dimensions and it is less than 1 in three and more dimensions. This is George Pólya's celebrated theorem about recurrence/transience of simple symmetric random walk on $\mathbb{Z}^d$.

Believe it or not, this classical result of probability theory is essentially equivalent to non-existence, respectively existence, of condensation of the ideal Bose gas in two, respectively three and more dimensions.

(E) This is the *site percolation problem* on $\mathbb{Z}^2$. Again, we start with finite boxes $\Lambda_N := [-N, N] \times [-N, N] \cap \mathbb{Z}^2$. On the set of finite configurations $\Omega_N := \{0, 1\}^{\Lambda_N}$ assign the weight function

$$p : \Omega_N \to [0, 1], \qquad p(\omega) = p^{\sum_{x \in \Lambda_N} \omega_x} (1 - p)^{\sum_{x \in \Lambda_N} (1 - \omega_x)}.$$

This means assigning multiplicatively weight p, respectively $(1-p)$, to every 1, respectively, every 0. Note, that this is not uniform weight assignment. Again, by taking the infinite volume limit $N \to \infty$, $\Lambda_N \to \mathbb{Z}^d$, the probability measure extends canonically to the set Ω of infinite configurations. After some very nice combinatorial arguments we can conclude that (in two and more dimensions) there exists a *critical threshold value* $p_c \in (0, 1)$ such that the probability that the origin is connected to infinity by an open path is zero for $p < p_c$ and positive for $p > p_c$. This is Hammersley's observation, with which the story of percolation theory started.

We conclude this subsection with the

Definition 1 *Probability space is the triplet* $(\Omega, \mathcal{F}, \mathbf{P})$. Ω *is the sample space.* $\mathcal{F} \subset \mathcal{P}(\Omega)$ *is the sigma-algebra of events satisfying axioms (Ai),(Aii) and (Aiiib).* $\mathbf{P} : \mathcal{F} \to [0, 1]$ *is the probability function, satisfying axioms (Pi) and (Piib).*

1.2 Some elementary facts

Given three events $A, B, C \in \mathcal{F}$, by additivity of probability and elemntary manipulations with sets we find

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$$

or

$$\begin{aligned} \mathbf{P}(A \cup B \cup C) = \mathbf{P}(A) + \mathbf{P}(B) + \mathbf{P}(C) \\ - \mathbf{P}(A \cap B) - \mathbf{P}(B \cap C) - \mathbf{P}(C \cap A) + \mathbf{P}(A \cap B \cap C). \end{aligned}$$

The following theorem extends these formulas to arbitrary number of events. The proof is elementary combinatorics.

Theorem 2 The Sieve Formula.
Let $A_1, A_2, \ldots, A_n$ be events of the probability space $(\Omega, \mathcal{F}, \mathbf{P})$. For $I \subset \{1, 2, \ldots, n\}$ denote $A_I := \bigcap_{i \in I} A_i$. The following identity holds:

$$\mathbf{P}(\bigcup_{i=1}^{n} A_i) = \sum_{I \subset \{1,\ldots,n\}} (-1)^{|I|+1} \mathbf{P}(A_I). \tag{1}$$

P r o o f.
The proof goes through induction on n. For $n = 1$ the formula is clearly true. The induction step goes as follows:

$$\begin{aligned}
\mathbf{P}(\bigcup_{i=1}^{n+1} A_i) &= \mathbf{P}((\bigcup_{i=1}^{n} A_i) \cup A_{n+1}) \\
&= \mathbf{P}(\bigcup_{i=1}^{n} A_i) + \mathbf{P}(A_{n+1}) - \mathbf{P}((\bigcup_{i=1}^{n} A_i) \cap A_{n+1}) \\
&= \mathbf{P}(\bigcup_{i=1}^{n} A_i) + \mathbf{P}(A_{n+1}) - \mathbf{P}(\bigcup_{i=1}^{n} (A_i \cap A_{n+1})) \\
&= \sum_{I \subset \{1,\ldots,n\}} (-1)^{|I|+1} \mathbf{P}(A_I) + \mathbf{P}(A_{n+1}) \\
&\qquad - \sum_{I \subset \{1,\ldots,n\}} (-1)^{|I|+1} \mathbf{P}((A_I \cap A_{n+1})) \\
&= \sum_{I \subset \{1,\ldots,n,n+1\}} (-1)^{|I|+1} \mathbf{P}(A_I).
\end{aligned}$$

Problem I wrote N letters to N different friends, put them in N envelopes. Unfortunately I sealed the envelopes before writing names and addresses on them. Now, I wrote the N names and addresses on the envelopes *at random.* What is the probability that at least one of my friends will get the letter intended to him? Also, find the asymptotics as $N \to \infty$.

More on sigma-additivity of probabilities. Axiom (Piib) is equivalent to either one of the two statements of the following theorem.

Theorem 3 Monotone limits of measures.
Let A_i, $i = 1, 2, \ldots$ be events of the probability space $(\Omega, \mathcal{F}, \mathbf{P})$.
(i) If $A_i \subset A_{i+1}$ for any i (i.e. the sequence of events is monotone increas-

ing) then

$$\mathbf{P}(\bigcup_{i=1}^{\infty} A_i) = \lim_{i\to\infty} \mathbf{P}(A_i). \tag{2}$$

(ii) If $A_{i+1} \subset A_i$ for any i (i.e. the sequence of events is monotone decreasing) then

$$\mathbf{P}(\bigcap_{i=1}^{\infty} A_i) = \lim_{i\to\infty} \mathbf{P}(A_i). \tag{3}$$

P r o o f.

(i) Apply countable additivity to the *disjoint events* $B_1 := A_1$, $B_j := A_j \setminus A_{j-1}$, $j = 2, 3, \ldots$:

$$\mathbf{P}(\bigcup_{i=1}^{\infty} A_i) = \mathbf{P}(\bigcup_{i=1}^{\infty} B_i) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbf{P}(B_i) = \lim_{i\to\infty} \mathbf{P}(\bigcup_{i=1}^{n} B_i) = \lim_{n\to\infty} \mathbf{P}(A_n).$$

(ii) Apply (2) to the complements, A_i^c:

$$\mathbf{P}(\bigcap_{i=1}^{\infty} A_i) = 1 - \mathbf{P}(\bigcup_{i=1}^{\infty} A_i^c) = \lim_{i\to\infty} \left(1 - \mathbf{P}(A_i^c)]\right) = \lim_{i\to\infty} \mathbf{P}(A_i).$$

2 Classical combinatorial and geometric problems

2.1 Classical combinatorial problems

Coincidence of birthdays.

Problem There are $n = 23$ randomly selected people in a company. What is the probability that no two birthdays coincide? (Disregard leap years. Give equal weights to all possible birthday assignments.)

Problem Distribute n distinguishable balls into r boxes. All distributions equally likely. What is the probability that there will be at most one ball in every box?

Clerely, the birthday problem is special case of the second one, with $r = 23$, $n = 365$. We compute the probability by the classical formula

$$\text{probability} = \frac{\text{number of favourable assignments}}{\text{total number of assignments}}.$$

We get

$$\mathbf{P}(\text{no two balls in any single box}) = \frac{r(r-1)(r-2)\dots(r-n+1)}{r^n}$$
$$= \prod_{j=0}^{n-1}\left(1-\frac{j}{r}\right).$$

Numerically, this gives that the probability that no two birthdays coincide is slightly bigger than $1/2$ for 22 people and slightly less than $1/2$ for 23 people. Sort of surprizing.

Instead of using a calculator let's approximate:

$$\log \mathbf{P}(\dots) = \sum_{j=1}^{n-1} \log\left(1-\frac{j}{r}\right) \approx -\sum_{j=1}^{n-1}\frac{j}{r} = -\frac{n(n-1)}{2r}$$

where we have used $\log(1+x) \approx x$ for small x. Thus we get

$$\mathbf{P}(\text{no two balls in any single box}) \approx \exp(-n(n-1)/(2r)).$$

The approximation is reasonably good if $n(n-1)/(2r) < 1$. It is worth comparing the approximate numerical values gotten this way with the true ones, for the birthday problem.

Sampling without replacement: Lottery

Problem In an urn we have M white and N black balls. We draw n balls at random from the urn, without replacement ($n \le N+M$). What is the probability that among the n balls drawn there will be k white and $n-k$ black ones.

Notation: let $n \in \mathbb{N}$

$$\binom{n}{k} := \begin{cases} \frac{n!}{k!(n-k)!} & \text{if } k \in [0,n] \cap \mathbb{Z} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Again we apply the 'uniform weight assignment' – all possible outcomes are equally likely. The total number of different outcomes of the experiment is $\binom{M+N}{n}$, the number of favourable outcomes is $\binom{M}{k}\binom{N}{n-k}$, so

$$\mathbf{P}(k \text{ white and } n-k \text{ black balls are drawn}) = \frac{\binom{M}{k}\binom{N}{n-k}}{\binom{N+M}{n}} =: h_{M,N,n}(k).$$

This is the so called *hypergeometric distribution.* It arises naturally in all cases of sampling without replacement. E.g., binary opinion polls. In

the Hungarian lottery system (five-out-of-ninety) the distribution of hits is $h_{5,85,5}(k)$, $k = 0, 1, \ldots, 5$. Prize is payed only for two and more hits. Compute the probabilities $h_{5,85,5}(k)$, $k = 0, 1, \ldots, 5$.

As a byproduct we get the combinatorial identity

$$\sum_k \binom{N}{k}\binom{M}{n-k} = \binom{N+M}{n}$$

valid for any $M, N, n \in \mathbb{N}$.

The Maxwell-Boltzmann distributions

The following problem has its origins in classical statistical physics. It has many more applications in other areas.

Problem n distinguishable balls (labelled 1 to n) are distributed at random between n boxes. All distributions are equally likely. Denote by ν_j the (random) number of balls put in the j-th box, $j = 1, 2, \ldots, r$.

(1) Given $k_1, k_2, \ldots, k_r \in \mathbb{N}$ such that $k_1 + k_2 + \cdots + k_r = n$, what is the probability that of the event $\{\nu_1 = k_1, \nu_2 = k_2, \ldots, \nu_r = k_r\}$?

(2) Compute the probability of the event $\{\nu_1 = k\}$, where $k \in [0, n]$. (That is: disregard the rest of the boxes.)

(3) Compute the limit of this last probability, when $n \to \infty$, $r \to \infty$ so that $\frac{n}{r} \to \lambda \in (0, \infty)$. (Many balls, many boxes, finite density.)

(1) The number of all possible distributions of the n balls in the r boxes is r^n. The number of distributions of balls, with k_j balls in the j-th box, $j = 1, 2, \ldots, r$, is $n!/(k_1!\, k_2!\, \ldots\, k_r!)$ — this is simple combinatorics. We get

$$\mathbf{P}(\{\nu_1 = k_1, \nu_2 = k_2, \ldots, \nu_r = k_r\}) = \frac{n!}{k_1!\, k_2!\, \ldots\, k_r!} r^{-n}.$$

As a byproduct we also get the combinatorial identity

$$\sum_{\substack{k_1, k_2, \ldots, k_r \in \mathbb{N} \\ k_1 + \cdots + k_r = n}} \frac{n!}{k_1!\, k_2!\, \ldots\, k_r!} = r^n. \tag{5}$$

valid for any fixed $n, r \in \mathbb{N}$. (Prove it also drectly, by induction on r.)

(2) Obviously,

$$\begin{aligned} \mathbf{P}(\{\nu_1 = k\}) &= \sum_{\substack{k_2, \ldots, k_r \in \mathbb{N} \\ k_2 + \cdots + k_r = n-k}} \mathbf{P}(\{\nu_1 = k, \nu_2 = k_2, \ldots, \nu_r = k_r\}) \\ &= \sum_{\substack{k_2, \ldots, k_r \in \mathbb{N} \\ k_2 + \cdots + k_r = n-k}} \frac{n!}{k_1!\, k_2!\, \ldots\, k_r!} r^{-n}. \end{aligned}$$

Using (5) with $(r-1)$ and $(n-k)$ we readily find

$$\mathbf{P}(\{\nu_1 = k\}) = \binom{n}{k} \left(\frac{1}{r}\right)^k \left(1 - \frac{1}{r}\right)^{n-k}.$$

This is a particular case of the so called *binomial distribution*

$$b_{p,n}(k) := \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, 2, \ldots, n.$$

(3) Straightforward computations yield:

$$\lim_{\substack{r,n\to\infty \\ n/r\to\lambda}} \binom{n}{k} \left(\frac{1}{r}\right)^k \left(1 - \frac{1}{r}\right)^{n-k} = e^{-\lambda}\frac{\lambda^k}{k!} =: p_\lambda(k), \qquad k = 0, 1, 2, \ldots.$$

On the right hand side we see the *Poisson distribution* with parameter λ. This is a special case of the Poisson approximation of the binomial distribution.

2.2 Classical geometric problems

Buffon's needle.

TO BE COMPLETED

Bertrand's paradox.

TO BE COMPLETED

3 Conditional probability and stochastic independence

3.1 Conditional probability

The point is that *partial information changes the probability of events*
Example: In a population of N individuals there are N_F females and N_M males, where $N_F + N_M = N$. N_L of them are lefthanders and N_R righthanders, where again $N_L + N_R = N$. We also know that the number of lefthanders among the females, respectively, among the males is N_{FL}, respectively, N_{ML}, where clearly $N_{FL} + N_{ML} = N_L$ holds. We sample one person at random from the population. What is the probability that the selected person will be lefthander? What is the probability of the same event, *given the partial information that the selected person is female*?

Each person from the population has equal chances to be selected, so $\mathbf{P}(\text{lefthander is selected}) = N_L/N$. Given that the selected person is female, this probablity will change: we compute it as fraction of favourable outcomes within the possibilities reduced by the partial information given. That is:

$$\begin{aligned}\mathbf{P}(\text{lh is selected given that f is selected}) &= \frac{N_{FL}}{N_F} = \frac{N_{FL}/N}{N_F/N} \\ &= \frac{\mathbf{P}(\text{lefthander female is selected})}{\mathbf{P}(\text{female is selected})}\end{aligned}$$

Example: Wimbledon finals are played by Boris Becker and Pete Sampras. The player who wins three sets (out of max. five) is the champion. They are equally good tennis players. The sample space is $\Omega = \Omega_3 \cup \Omega_4 \cup \Omega_5$, where Ω_j contains those posible outcomes (of the full match) where j sets were played. With obvious notations:

$$\begin{aligned}\Omega_3 &= \{bbb, sss\} \\ \Omega_4 &= \{bbsb, bsbb, sbbb, ssbs, sbss, bsss\} \\ \Omega_5 &= \{bbssb, bsbsb, sbbsb, bssbb, sbsbb, ssbbb, \\ &\qquad ssbbs, sbsbs, bssbs, sbbss, bsbss, bbsss\}\end{aligned}$$

The fact that they are equally good players id mathematically modeled by the following assignment of probability weights:

$$\begin{aligned}p(bbb) = p(sss) &= \frac{1}{8} \\ p(bbsb) = \cdots = p(bsss) &= \frac{1}{16} \\ p(bbssb) = \cdots = p(bbsss) &= \frac{1}{32}\end{aligned}$$

Explain why, without relying on the notion of *stochastic independence*, which was not yet introduced. Consider the following events:

$$\begin{aligned}B &:= \{\text{Becker wins the match}\} \\ B_j &:= \{\text{Becker wins the } j\text{-th set}\}.\end{aligned}$$

What is the probability that Becker wins the match? What is the probability of the same event, given that he won the first set of games? Elementary computations show that

$$\mathbf{P}(B) = \frac{1}{2}.$$

(This is not surprising, as we modeled the situation of equally good players.) Now, assume that Becker won already one set. We have to compute *the relative weight* of all outcomes where B occurs, whithin the event B_1. That is:

$$\mathbf{P}(B \text{ given } B_1) = \frac{\mathbf{P}(B \text{ and } B_1)}{\mathbf{P}(B_1)} = \frac{\frac{1}{8} + \frac{2}{16} + \frac{3}{32}}{\frac{1}{8} + \frac{3}{16} + \frac{6}{32}} = \frac{11}{16}$$

We are ready now to define the notion of *conditional probability*

Definition 4 *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and $H \in \mathcal{F}$ an event of positive probability, $\mathbf{P}(H) > 0$. The conditional probability of an arbitrary event $A \in \mathcal{F}$, given H is*

$$\mathbf{P}(A \,|\, H) := \frac{\mathbf{P}(A \cap H)}{\mathbf{P}(H)}.$$

The definition is *very natural*. It says that given H, the sample space is reduced from Ω to H (outcomes falling in $\Omega \setminus H$ must be disregarded) and we have to consider *the relative weight of A within H*.

Example: (American students are very familiar with this one.) Three boxes are given, numbered 1,2,3. In one of the boxes a prize id hidden the other two are empty. You want to win the prize. You point to one of the boxes. I open one of the other two boxes, showing that it is empty. Now, you may stick to your first choice or change your mind and choose the nother one. What should you do? Set up a proper probabilistic (mathematical) model and analyse!

3.2 Three easy statements

Multiplication of conditional probabilities: the tower rule. Let A_1, A_2, $\ldots$, A_n be events in a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Then the probability of their joint occurence is expressed as:

$$\begin{aligned}\mathbf{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbf{P}(A_1)\mathbf{P}(A_2 \,|\, A_1)\mathbf{P}(A_3 \,|\, A_1 \cap A_2)\ldots \\ \mathbf{P}(A_n \,|\, A_1 \cap A_2 \cdots \cap A_{n-1})\end{aligned}$$

The proof goes easily through induction on n.

Example: *Urn models.* Let c and d be two fixed integers satisfying $c \geq -1$, $d \geq 0$. In an urn there are blue and red balls. Initially the number of blue,

respectively, red balls is β_0, respectively, ρ_0. (Assume $\beta_0 + \rho_0 > 0$.) We draw one ball (at random) from the urn and replace it with $c+1$ balls of the same colour and d balls of the other colour. (The total number of balls in the urn changes by $c+d$.) Continue doing this. (If $c+d<0$ stop when the urn is empty.) What is the probability that the first three balls drawn are blue, blue, red (in this order)?
Remark: If $c=d=0$ this is simply sampling with replacement. If $c=-1$, $d=0$ this is sampling withou replacement. If $c>0$, $d=0$ It is Pólya's urn model. If $c=-1$, $d=1$ it is Ehrenfest's urn model. Both have very interesting behaviour and applications.

The theorem of complete probabilities.

Definition 5 *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. A (finite or countable)* partition *of Ω is a collection of events $H_j \in \mathcal{F}$, $j = 1, 2, \ldots$, which*
1. have all positive probability: for all $j \quad \mathbf{P}(H_j) > 0$;
2. are pairwise disjoint: if $j \neq k$ then $H_j \cap H_k = \emptyset$;
3. cover the sample space: $\cup_j H_j = \Omega$.

Given a partition H_j, $j = 1, 2, \ldots$, the probability of any avent $A \in \mathcal{F}$ can be written as:

$$\mathbf{P}(A) = \sum_j \mathbf{P}(A \,|\, H_j)\mathbf{P}(H_j).$$

The proof is completely trivial: it goes through (sigma)additivity of the probabilities and plain application of the definition of conditional probability.

Bayes' Rule. Given a partition H_j, $j = 1, 2, \ldots$, and an event A, assume we know the probabilities $\mathbf{P}(H_j)$ of the partition elements, and the conditional probabilities $\mathbf{P}(A \,|\, H_j)$. It may be natural to ask about the conditional probabilities $\mathbf{P}(H_j \,|\, A)$. Think about H_j-s as possible (disjoint and exhaustive) causes and A as consequence. Given that A occured, what are the *a posteriori* probabilities of the different possible causes? The answer is:

$$\mathbf{P}(H_j \,|\, A) = \frac{\mathbf{P}(A \,|\, H_j)\mathbf{P}(H_j)}{\sum_i \mathbf{P}(A \,|\, H_i)\mathbf{P}(H_i)}.$$

The proof is again straightforward: plainly apply the definition of conditional probability and the 'theorem of complete probabilities'.

Example: A population consists of 60% females and 40% males. 5% of males and 1% of females is colour-blind. A person is selected at random. Given that (s)he is colour-blind, what is the probability that a male was selected?

Applying Bayes' rule we find (with obvious notations):

$$\begin{aligned}\mathbf{P}(\mathrm{M}\,|\,\mathrm{CB}) &= \frac{\mathbf{P}(CB\,|\,M)\mathbf{P}(M)}{\mathbf{P}(CB\,|\,M)\mathbf{P}(M)+\mathbf{P}(CB\,|\,F)\mathbf{P}(F)}\\ &= \frac{0.05\cdot 0.40}{0.05\cdot 0.40+0.01\cdot 0.60}=\frac{10}{13}\end{aligned}$$

3.3 Stochastic independence

Independence of two events. Let A and B be two events in the probability space $(\Omega,\mathcal{F},\mathbf{P})$. Natural candidate for the condition of independence is that occurence of one does not influence the (conditional) probability of the other. I.e.

$$\mathbf{P}(A\,|\,B)=\mathbf{P}(A)\qquad\text{and}\qquad\mathbf{P}(B\,|\,A)=\mathbf{P}(B).$$

It turns out that these two are actually the same condition. Namely:

$$\mathbf{P}(A\cap B)=\mathbf{P}(A)\mathbf{P}(B).$$

Example: Throw an unbiased die. Let $A=\{\text{result is even}\}$ and $B=\{\text{result is multiple of 3}\}$.

Example: (this is slightly more surprizng). Choose a random number ω (uniformly distributed) in the interval $[0,1]$. That is: The sample space is $\Omega=[0,1)$. The sigma-algebra of events is $\mathcal{F}=\mathcal{B}$ the Borel algebra generated by the natural topology (but do not worry about this). The probability is the Lebesgue measure. This means that the probability of ω to fall in any fixed interval $(a,b)\subset[0,1]$ is proportiaonal (equal) to the length of the interval, $\mathbf{P}(\omega\in(a,b))=b-a$. This is exactly what a standard random number generator does. Now, write ω in usual binary expansion:

$$\omega=0.\omega_1\omega_2\omega_3\omega_4\cdots=\sum_{j=1}^{\infty}\omega_j 2^{-j}. \tag{6}$$

This is done in a canonical unique way if, e.g., we do not allow expansions terminating with and infinite sequence of ones. Any other system (e.g.

decimal) would be equally good for this example. Let $A = \{\text{first digit is } 1\}$, $B = \{\text{second digit is } 1\}$. Prove that A and B are independent.

Independence of more than two events. Given three events A, B and C, how should we define their independence? Is it suffitient to require that any two of them are (pairwise) independent? No! In the last example take beside A and B already defined, $C = \{$ sum of the first two digits is even$\}$. You easily check that any two of them are independent, but determine the third!

The good definition is:

Definition 6 *Let A_j, $j = 1, 2, \ldots, n$ be a collection of events in the probability space $(\Omega, \mathcal{F}, \mathbf{P})$. We say thet these are* (completely) independent *if for any subcollection*

$$\forall I \subset \{1, 2, \ldots, n\}: \qquad \mathbf{P}(\bigcap_{j \in I} A_j) = \prod_{j \in I} \mathbf{P}(A_j). \tag{7}$$

Let A_j, $j = 1, 2, \ldots$ be a countable collection of events in the probability space $(\Omega, \mathcal{F}, \mathbf{P})$. We call them completely independent if any finite subcollection of them are completely independent.

Note that (7) contains $2^n - (n+1)$ algebraically independent equations! Condition (7) can be written equivalently as

$$\forall I \subset \{1, 2, \ldots, n\}: \qquad \mathbf{P}((\bigcap_{j \in I} A_j) \cap (\bigcap_{j \notin I} A_j^c)) = \prod_{j \in I} \mathbf{P}(A_j) \prod_{j \notin I} \mathbf{P}(A_j^c).$$

which is more symmetric in the events A_j and their complements A_j^c.

Modeling independent experiments by product spaces. Given n probabilistic experiments modeled mathematically by the probability spaces $(\Omega_1, \mathcal{F}_1, \mathbf{P}_1)$, $(\Omega_2, \mathcal{F}_2, \mathbf{P}_2)$, $\ldots$, $(\Omega_n, \mathcal{F}_n, \mathbf{P}_n)$. We want to model mathematically the situation when these are performed *jointly and completely independently.* The natural choice is:
Sample space:

$$\Omega := \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n.$$

Algebra of events:

$$\mathcal{F} := \sigma(\mathcal{F}_1 \times \mathcal{F}_2 \times \cdots \times \mathcal{F}_n).$$

This means: the smallest (sigma-)algebra of subsets of Ω, which contains all subsets of the form

$$A_1 \times A_2 \times \cdots \times A_n, \qquad \text{where } A_j \in \mathcal{F}_j. \tag{8}$$

(This object is well defined. You easily understand it in the discrete cases. Do not worry too much about it in the continuous cases.)
Probability:

$$\mathbf{P} := \mathbf{P}_1 \times \mathbf{P}_2 \times \cdots \times \mathbf{P}_n.$$

This means that for rectangular subsets of Ω of the form (18) we define

$$\mathbf{P}(A_1 \times A_2 \times \cdots \times A_n) = \mathbf{P}_1(A_1)\mathbf{P}_2(A_2)\ldots\mathbf{P}_n(A_n)$$

and extend this measure (canonically) to the whole (sigma)algebra $\mathcal{F}$.

This procedure of defining product spaces as model of independent experiments is starightforward and very natural.

Remark: Uniform measure on product spaces is the product of uniform measures on the factor spaces. Look at the combinatorial and geometric examples with uniform assignment of weights, with this remark in mind!

'Hidden' independence.

In the previous paragraph we spoke about modeling independence. Now we show some cases where we find independence somewhat unexpectedly.

An example from number theory: For $s \in (1, \infty)$ define *Riemann's ζ-function* by the series:

$$\zeta(s) := \sum_{n=1}^{\infty} n^{-s}.$$

The series is convergent for $s > 1$. Consider a random number X with distribution

$$\mathbf{P}(X = n) = p_s(n) = \frac{n^{-s}}{\zeta(s)}.$$

Given a standard random number generator, which picks random numbers ω, uniformly distributed in the interval $[0, 1]$, we produce a random number with distribution (3.3) by seting

$$X = X(\omega) := \min\{n : \sum_{k=1}^{n} p_s(k) > \omega\}.$$

Check it! Now, for all prime numbers r, we define the event

$$E_r := \{X \text{ is divisible by } r\}.$$

What is the probability of the event E_r?

$$\mathbf{P}(E_r) = \mathbf{P}(X = kr \text{ for some } k \in \mathbb{N}\} = \sum_{k=1}^{\infty} p_s(kr) = \sum_{k=1}^{\infty} \frac{k^{-s} r^{-s}}{\zeta(s)} = r^{-s}.$$

Next we prove that the countable collection of events E_r, r prime, are completely indepentdent. Indeed: given any finitely many primes $r_1 < r_2 < \cdots < r_m$, an identical computation yields

$$\mathbf{P}(E_{r_1} \cap E_{r_2} \cap \cdots \cap E_{r_m}) = (r_1 r_2 \dots r_m)^{-s} = \mathbf{P}(E_{r_1})\mathbf{P}(E_{r_2}) \dots \mathbf{P}(E_{r_m}).$$

This is sort of a surprize: there was no obvious a priori reason for this.

A well known classical result from elementary number theory is *Euler's formula*

$$\zeta(s)^{-1} = \prod_{r:\, r \text{ prime}} (1 - r^{-s}).$$

Let's see what is its probabilistic content.

$$\begin{aligned}
\prod_{r:\, r \text{ prime}} (1 - r^{-s}) &= \prod_{r:\, r \text{ prime}} (1 - \mathbf{P}(E_r)) \\
&= \prod_{r:\, r \text{ prime}} \mathbf{P}(E_r^c) = \mathbf{P}\Big(\bigcap_{r:\, r \text{ prime}} E_r^c \Big) \\
&= \mathbf{P}(X \text{ is not divisible by any prime}) \\
&= \mathbf{P}(X = 1) = p_s(1) = \zeta(s)^{-1}.
\end{aligned}$$

We have used independence in the third equality.

Binary expansions revisited: $(\Omega, \mathcal{F}, \mathbf{P}) = ([0,1], \mathcal{B}, \text{Lebesgue measure})$. That is: ω is a random number uniformly distributed in the interval $[0,1]$. Write ω in binary expansion, as in (6). It turns out that its binary digits ω_j, $j = 1, 2, 3, \dots$ are *completely independent.* (Not only the first two, as already seen in a previous example.) This remark hints towards a deep relation between analysis and probability theory. E.g. limit theorems for sums of independent random variables (laws of large numbers, central limit theorem, law of the iterated logarithm, large deviation theorems) will be valid in real-analytic context.

4 Random variables I.: single random variables and their distributions

A *random variable* is a random number, i.e., a number asociated to the outcome of a probabilistic experiment. Mathematically speaking, let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. A random variable is a (measurable) map

$$\xi : \Omega \to \mathbb{N} \text{ or } \mathbb{Z} \text{ or } \mathbb{R} \text{ or } \mathbb{C}$$

Later we shall speak also about vector valued random variable

$$\xi : \Omega \to \mathbb{R}^n .$$

This is nothing more than n real valued random variables, jointly defined on the same probability space. We always assume that the random variable is measurable. That is: inverse images of reasonable sets from the range (e.g. intervals) are elements of $\mathcal{F}$.

Given a real random variable, its *distribution* characterizes it completely. The distribution of the random variable ξ is the probability measure generated $\xi : (\Omega, \mathcal{F}, \mathbf{P}) \to (\mathbb{R}, \mathcal{B})$: for $A \in \mathcal{B}$

$$\mu_\xi(A) := \mathbf{P}(\xi \in A) = \mathbf{P}(\{\omega \in \Omega : \xi(\omega) \in A\}) \tag{9}$$

If we know for any reasonable subset A of the range this probability then we have a complete characterization of the random number. The sample space does not play an essential rôle: random variables difined a very different sample spaces (associated to very different experiments) but having the same distribution *are the same*.

4.1 Discrete distributions

First we speak about $\mathbb{N}$- and Z-valued random variable. Concretely, we shall formulate things for N-valued variables. Everything is easily tranposed to Z-valued random variables (or to random variables with some other countable range).

Given

$$\xi : \Omega \to \mathbb{N}$$

its *distribution density* is

$$f_\xi : \mathbb{N} \to [0, 1], \qquad f_\xi(k) = \mathbf{P}(\xi = k) = \mathbf{P}(\{\omega \in \Omega : \xi(\omega) = k\}).$$

The discrete distribution density f_ξ has the property

$$\sum_k f_\xi(k) = 1. \tag{10}$$

This is nothing else, but (sigma)addititvity of probabilities. In general, a funcion $f : \mathbb{N} \to [0,1]$ with property (10) is called *discrete distribution density*. Given a discrete distribution density one can easily define a probability space and a discrete random variable defined on it which fhas the given function as distribution density. Indeed: define the (cumulated) distribution function

$$F : \mathbb{N} \to [0,1], \qquad F(n) = \sum_{k=0}^{n-1} f(k),$$

and let $(\Omega, \mathcal{F}, \mathbf{P}) = ([0,1], \mathcal{B}, \text{Leb.})$, that is: let ω be a random number unoformly distributed in the interval $[0,1]$ and define

$$\xi(\omega) := \sup\{n : F(n) < \omega\}.$$

It is easy to check that $\mathbf{P}(\xi = n) = F(n+1) - F(n) = f(n)$.

We continue with a list of most important discrete random variables.

Indicator, $IND(p)$: The random variable takes on two possible values, 0 and 1. Its distribution has one fixed parameter: $p \in [0,1]$. The distribution density is:

$$\mathbf{P}(\xi = 0) = 1 - p, \quad \mathbf{P}(\xi = 1) = p.$$

The random variable indicates whether an event (with probability p) occured or not. Sometimes we shall interpret $\xi = 1$, respectively, $\xi = 0$ as success, respectively failure.

Hypergeometric, $HYG(M,N,n)$: This is the distribution of the result in sampling without replacement. $M, N, n \in \mathbb{N}$ with $n \leq M + N$ are fixed parameters. Let M white and N black balls be in an urn. We draw (without replacement) a random sample of n balls. Let ξ denote the number of white balls drawn. We have seen, see section 2.1, that the distribution density of ξ is

$$\mathbf{P}(\xi = k) = \frac{\binom{M}{k}\binom{N}{n-k}}{\binom{M+N}{n}} := h_{M,N,n}(k), \qquad k = 0, 1, 2, \ldots, n.$$

(The convention (4) for the binomial coefficients is always used.)

Binomial, $BIN(p,n)$**:** The parameters $n \in \mathbb{N}$ and $p \in [0,1]$ are fixed. We perform n identical and independent experiments. Each experiment results in 'success' with probability p, or 'failure' with probability $1-p$. Let ξ be the number of successes among the n independent trials. It is easily checked that the distribution density of ξ is

$$\mathbf{P}(\xi = k) = \binom{n}{k} p^k (1-p)^{n-k} =: b_{p,n}(k), \qquad k = 0, 1, 2, \ldots, n.$$

Relation (10) is easily checked.
Let $M, N \in \mathbb{N}$ be the number of white, respectively black balls in an urn. Put $p = M/(M+N)$, $1-p = N/(M+N)$ then ξ of $BIN(p,n)$ distribution models the random number of white balls drawn in n consecutive trials if after each trial we replace the ball. This experiment is usually calles *sampling with replacement.* Examples abund.

Convergence of $HYG(M,N,n)$ *to* $BIN(p,n)$. If the numbers M, N of white and black balls in the urn is very large compared with the number n of balls drawn then there is no significant difference between sampling with or without replacement. Mathematically speaking, one can easily prove that if $M, N \to \infty$ so that $M/(M+N) \to p$ (and, consequently, $N/(M+N) \to 1-p$) then for any $n \in \mathbb{N}$ and $k \in \{0, 1, \ldots, n\}$ fixed

$$h_{M,N,n}(k) \to b_{p,n}(k).$$

Poisson, $POI(\lambda)$**:** The parameter $\lambda \in (0, \infty)$ is fixed. Take the binomial distribution $BIN(p,n)$ and perform the limit: $n \to \infty$, $p \to 0$ so that $np \to \lambda \in (0, \infty)$, keeping k fixed. Taking this limit.

$$\begin{aligned} b_{p,n}(k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{1}{k!}(np)^k (1-p)^n (1-p)^{-k} \prod_{j=0}^{k-1} (1 - \frac{j}{n}) \\ &\to e^{-\lambda} \frac{\lambda^k}{k!} =: p_\lambda(k), \qquad k = 0, 1, 2, \ldots . \end{aligned}$$

The Poisson distribution describes a situation when the probability of success of every single experiment is negligeably small, but sufficiently many experiments are performed, so that the overall probability of a fixed number of successes is finite.

Examples. Compare the binomial distribution and its Poisson approxiamtion in the following two cases:
(1) We throw a fair die six consecutive times, ξ is the number of aces we get.
(2) We buy one lottery ticket ('5-out-of-90' system) every weak of a whole year. At least two hits is success. Denote by ξ the number of succsses during the year.

Further examples abund!

Geometric (or negative binomial), $GEO(p)$: Independent and identical experiments are performed with probability of success, respectively, failure p, respectively, $1-p$. (Bernoulli trials). Let ξ denote *the number of failures suffered before the first success occurs.* Easy computations show:

$$\mathbf{P}(\xi = k) = (1-p)^k p =: g_p(k), \qquad k = 0, 1, 2, \dots$$

A remarkable characteristic property of the geometric distribution is its *ever-freshness.* This means that, if ξ is a geometrically distributed random variable, then for any $k_0 \geq 0$

$$\mathbf{P}(\xi = k_0 + k \,|\, \xi \geq k_0) = \mathbf{P}(\xi = k).$$

Prove this formula! If we interpret ξ as a random waiting time, then the amount time already spent with waiting does not influence whatsoever the distribution of the remaining time to wait. (Think about waiting for the tram: no matter how long you had already spent in the station, the distribution of the time you still have to wait is the same.) The geometric distributions are the only N-valued discrete distributions with this remarkable property. (See also the ever-freshness of the exponential distributions.)

Examples abund!

4.2 Distribution functions – in general

Given a *real valued* random variable $\xi : \Omega \to \mathbb{R}$, its distribution is the probability measure μ_ξ on $\mathbb{R}$ defined in (9). We define its (cumulative) *distribution function* as

$$F_\xi : \mathbb{R} \to [0,1], \qquad F_\xi(x) := \mathbf{P}(\xi < x) = \mathbf{P}(\{\omega \in \Omega : \xi(\omega) < x\}).$$

As there is a simple one-to-one correspondence between distributions and distribution functions on $\mathbb{R}$ (see below) we shall use interchangeably the two notions.

The distribution function F_ξ has the following primary properties
(i) It is monotone non-decreasing: if $x \leq y$ then $F_\xi(x) \leq F_\xi(y)$ (This property follows from addititvity of probability.)
(ii) $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to+\infty} F(x) = 1$. (These properties follow from (3), respectively, (2) in Theorem 3)
(iii) It is continuous from the left: for any $x \in \mathbb{R}$, $\lim_{y\nearrow x} F_\xi(y) = F_\xi(x)$. (This property also follows from (2), Theorem 3.)

Given a function $F : \mathbb{R} \to [0,1]$ having the properties (1)-(3) above, one can construct a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a random variable $\xi : \Omega \to \mathbb{R}$ which will have the (cumulative) distribution function F. Indeed, let $(\Omega, \mathcal{F}, \mathbf{P}) = ([0,1], \mathcal{B}, \text{Leb.})$, that is, let ω be a random number uniformly distributed in the interval $[0,1]$ and define

$$\xi = \xi(\omega) := \sup\{x : F(x) < \omega\}.$$

It is easily checked that $\mathbf{P}(\xi < x) = F(x)$. A funtion $F : \mathbb{R} \to [0,1]$ satisfying conditions (i)-(iii) above is called *probability distribution function.* There is a one-to-one correspondence between probability distribution functions and regular Borel probability measures on $\mathbb{R}$:
Given a probability distribution function F define for $a < b$

$$\nu_F([a,b)) := F(b) - F(a)$$

and extend μ_F to all Borel measurable sets, by standard procedures.
Given a probability measure μ, define

$$F_\mu(x) := \mu((-\infty, x)).$$

Due to this one-to-one correspondence we are allowed to use the two notions interchangeably.

We call $x \in \mathbb{R}$ an atom of the probability distribution function F, if $\Delta F(x) := F(x+) - F(x) > 0$, or, equivalently if $\mu_F(\{x\}) > 0$. Denote the set of atoms of F

$$\mathcal{A}_F := \{x \in \mathbb{R} : \Delta F(x) > 0\}.$$

It is easy to see that for any probability distribution function F the set $\mathcal{A}_F$ is at most countable. (Prove this!)

Definition 7 *(i) The probability distribution function F is* pure point *if and only if for any $x \in \mathbb{R}$*

$$F(x) := \sum_{x' \in \mathcal{A}_F,\, y<x} \Delta F(x').$$

This is equivalent to the corresponding probability distribution μ_F being totally concentrated in the set of atoms. If the set of atoms does not have points of accumulation we call F (and μ_F) discrete.
(ii) The probability distribution function F is continuous *if it has no atoms at all, i.e., $\mathcal{A}_F = \emptyset$. Equivalently, the corresponding probability distribution μ_F does not give positive weight to any single point in $\mathbb{R}$. In this case F is indeed continuous function.*
(iii) The probability distribution function F is absolutely continuous *if there exists a measurable function $f : \mathbb{R} \to [0, \infty]$ such that for any $x \in \mathbb{R}$*

$$F(x) = \int_{-\infty}^{x} f(y)dy.$$

(This is Lebesge integral.) In terms of the distribution: the measure μ_F is absolutely continuous with respect to Lebesgue measure. The function f is called density function *of F (or μ_F).*
(iv) A probability distribution function which is continuous, (Lebesgue) almost everywhere differentiable with $F'(x) = 0$ (Lebesgue) almost everywhere, is called continuous but singular. *In terms of the distribution: μ_F is continuous but singular with respect to Lebesgue measure.*

Remarks. (1) The set of atoms $\mathcal{A}_F$ is at most countable. (Prove it!)
(2) In the definition of continuous distributions it seems tautological to say that in this case 'F is indeed a continuous finction'. But this is not the case: a priori this is a different notion of continuity.
(3) In the absolutely continuos case: the density function f a fortiori satisfies

$$F(x) = \int_{-\infty}^{\infty} f(y)dy = 1.$$

(4) In the absolutely continuous case: it is a fact (prove it!) that if f_1 and f_2 are density functions of the same distribution then $f_1 = f_2$ (Lebesgue) almost everywhere.

(5) Still in the absolutely continuous case: F is (Lebesgue) almost everywhere differentiable and almost everywhere

$$F'(x) = f(x).$$

(6) Cantor's function (sometimes called 'the Devil's staircase') is a typical example of continuous but singular function.

Theorem 8 Lebesgue's decomposition theorem.
Every probability distribution function F is uniquely decomposable into convex combination of a pure point, an absolutely continuous and a continuous but singular part. I.e., there are three numbers $p, q, r \in [0, 1]$ such that $p + q + r = 1$, a discrete probability distribution function F_d, an absolutely continuous probability distribution function F_{ac} and a continuous but singular probability distribution function F_{cs}, so that $F = pF_{pp} + qF_{ac} + rF_{cs}$. The decomposition is unique.

4.3 Absolutely continuous distributions

The density function (of an absolutely continuous probability distribution function) has the primary property

$$\int_{-\infty}^{\infty} f(y)dy = 1. \tag{11}$$

A measurable function $f : \mathbb{R} \to [0, \infty]$ with this property will be called *probability density function.* A probability density function defines an absolutely continuous probability distribution function by

$$F(x) := \int_{-\infty}^{x} f(y)dy.$$

Linear transformations of random variables. Let ξ be a random variable, $a > 0$ and $b \in \mathbb{R}$ fixed (deterministic) numbers and define $\eta := a\xi + b$. If F is the distribution function of ξ, then the disstribution function G of η is simply expressed as

$$G(y) := \mathbf{P}(\eta < y) = \mathbf{P}(a\xi + b < y) = \mathbf{P}(\xi < (y - b)/a) = F((y - b)/a).$$

If the distribution F is absolutely continuous, with density f, then so is G and is density g is expressed as:

$$g(y) := G'(y) = a^{-1} f((y - b)/a).$$

Definition 9 *We say that two distributions, say F and G, are* of the same type *if there are two numbers, $a \in (0,\infty)$ and $b \in \mathbb{R}$, such that $F(x) = G(ax+b)$. (I.e. if the random variables having these distributions are related by a regular affine transformation.) It is straightforward to check that this defines an equivalence relation in the set of distribution functions.*

The most common absolutely continuous distributions are:

The uniform distribution, $UNI(a,b)$: Fixed parameters are $-\infty < a < b < \infty$. The distribution function is

$$F(x) := \begin{cases} 0 & \text{if } x \le a \\ \dfrac{x-a}{b-a} & \text{if } a < x \le b \\ 1 & \text{if } b < x. \end{cases}$$

The density function is

$$f(x) := \begin{cases} 0 & \text{if } x \notin [a,b] \\ \dfrac{1}{b-a} & \text{if } x \in [a,b]. \end{cases}$$

The standard choice is $a = 0$, $b = 1$. If ξ has distribution $UNI(0,1)$ then $\eta := (b-a)\xi + a$ will have distribution $UNI(a,b)$

The exponential distribution, $EXP(\lambda)$: The parameter $\lambda \in (0,\infty)$ is fixed. The distribution function is

$$F(x) := \begin{cases} 0 & \text{if } x \le 0 \\ 1 - e^{-\lambda x} & \text{if } x > 0 \end{cases}$$

The density function is

$$f(x) := \begin{cases} 0 & \text{if } x < 0 \\ \lambda e^{-\lambda x} & \text{if } x \ge 0 \end{cases}$$

The standard choice is $\lambda = 1$. If ξ has distribution $EXP(1)$ then $\eta := \lambda^{-1}\xi$ will have distribution $EXP(\lambda)$ Remarkable property of the exponential distribution is its *ever-freshness*: let ξ be a random variable distributed according to $EXP(\lambda)$, then for any $x_0 \ge 0$ and $x \ge 0$

$$\mathbf{P}(\xi \ge x_0 + x \,|\, \xi \ge x_0) = \mathbf{P}(\xi \ge x).$$

(Prove this!) The exponential distributions are the only continuous distributions with this remarkable property. (See also 'ever freshness' of the geometric distribution.)

Normal or Gaussian distributions, $N(m,\sigma)$**:** Now $m \in \mathbb{R}$ and $\sigma \in (0,\infty)$ are fixed parameters. The distribution function is

$$F(x) = \int_{-\infty}^{x} f(y)dy$$

with density function is

$$f(x) := \frac{e^{-(x-m)^2/(2\sigma^2)}}{\sqrt{2\pi}\sigma}.$$

The standard choice is $m = 0$, $\sigma = 1$ and we shall denote the standard normal density function

$$\varphi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

and the standard normal distribution function

$$\Phi(x) = \int_{-\infty}^{x} \varphi(y)dy.$$

If ξ is a random variable with distribution $N(0,1)$ and $\eta := \sigma\xi + m$ then η has distribution $N(m,\sigma)$.

One proves $\int_{-\infty}^{\infty} \varphi(x)dx = 1$ by

$$\left(\int_{-\infty}^{\infty} \varphi(x)dx\right)^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi(x)\varphi(y)dxdy = \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\infty} e^{-r^2/2}rdrd\theta = 1.$$

The Cauchy distribution, $CAU(m,\tau)$**:** $m \in \mathbb{R}$ and $\tau \in 0,\infty)$ are fixed parameters. The distribution function is

$$F(x) := \frac{1}{\pi}\operatorname{arctg}\left(\frac{x-m}{\tau}\right) + \frac{1}{2},$$

with density function

$$f(x) := \frac{1}{\pi}\frac{\tau}{\tau^2 + (x-m)^2}$$

The standard choice is $m = 0$, $\tau = 1$. If ξ is a random variable with distribution $CAU(0,1)$ and $\eta := \tau\xi + m$ then η has distribution $CAU(m,\tau)$.

Transformations of random variables. Let ξ be a real valued random variable and $\psi : \mathbb{R} \to \mathbb{R}$ a measurable map. Then $\eta := \psi(\xi)$ is also a random variable. If ξ has absolutely continuous distribution F, with density f and the map ψ

is *piecewise smooth and strictly monotone* then the distribution G of η will be absolutely continuous, too. Its density g is expressed in terms of f and the map ψ as follows:

$$g(y) = \sum_{x:\ \psi(x)=y} \frac{f(x)}{|\psi'(x)|}.$$

This formula is nothing else but the formula for change of variable under an intergral sign. Here is the simple (two lines) proof for ψ strictly increasing:

$$G(y) = \mathbf{P}(\eta < y) = \mathbf{P}(\psi(\xi) < y) = \mathbf{P}(\xi < \psi^{-1}(y)) = F(\psi^{-1}(y))$$

$$g(y) = G'(y) = \frac{d}{dy} F(\psi^{-1}(y)) = F'(\psi^{-1}(y)) \frac{d}{dy} \psi^{-1}(y) = \frac{f(\psi^{-1}(y))}{\psi'(\psi^{-1}(y))}.$$

In the last step we have used the formula of differentiating the inverse function. The general case is done very similarly, piecewise on the strictly monotone (invertible) and smooth pieces of ψ.

The log-normal distribution, $LN(m, \sigma)$**:** Fix the parameters $m \in \mathbb{R}$ and $\sigma \in (0, \infty)$. Let ξ be a random variable of distribution $N(m, \sigma)$ and let $\eta := \exp(\xi)$. The distribution of η is called *log-normal distribution of parameters m and σ*. Applying formula (4.3) with $\psi(x) := e^x$ we find the density function of the log-normal distribution:

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ \dfrac{1}{\sqrt{2\pi}\sigma x} \exp\left\{-(\ln x - m)^2/(2\sigma^2)\right\} & \text{if } x < 0 \end{cases}$$

This distribution also arises in some practical applications.

5 Random variables II.: expectation, variance, moments, etc.

A distribution function is a very complicated object. Try to characterize it with a few characteristic numerical data. The two most important ones are *the centre* or and the *dispersion, diffuseness* of the distribution.

5.1 The expectation.

This is the most natural parameter for characterizing the centre of the distribution. It is analogue to the *centre of mass* for a classical distribution of matter. The expectation is the *weighted average*.

Definition 10 *Let F be a dsitribution function. If*

$$\int_{-\infty}^{\infty} |x| dF(x) < \infty \tag{12}$$

then the centre of mass *or* mean value *of the distribution F is defined as*

$$m_F := \int_{-\infty}^{\infty} x dF(x). \tag{13}$$

(These are Riemann-Stieltjes integrals on $\mathbb{R}$, with respect to the probability distribution function F.)

In case of discrete, respectively, absolutely continuous distributions (12) and (13) are written as

$$\sum_{x \in \mathcal{A}_F} |x| \Delta F(x) < \infty, \qquad m_F := \sum_{x \in \mathcal{A}_F} x \Delta F(x),$$

respectively,

$$\int_{-\infty}^{\infty} |x| f(x) dx < \infty, \qquad m_F := \int_{-\infty}^{\infty} x f(x) dx.$$

Definition 11 *Let ξ be a random variable defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$. If*

$$\int_{\Omega} |\xi(\omega)| d\mathbf{P}(\omega) < \infty, \tag{14}$$

then we define the expected value *or* expectation *of the random variable ξ as*

$$\mathbf{E}(\xi) := \int_{\Omega} \xi(\omega) d\mathbf{P}(\omega).$$

(These are Lebesgue inegrals over Ω, with respect to the probability measure $\mathbf{P}$.)

Now, it is straightforward to see that if ξ is a random variable and F is its distribution function then conditions (12) and (14) are equivalent and $\mathbf{E}(\xi) = m_F$. So, we shall use the two notions exchangeably.

Examples: we compute the expectation/mean value of the various distributions encountered so far.
Indicator:

$$\mathbf{E}(\xi) = 0(1-p) + 1p = p.$$

Binomial, $BIN(p,n)$:

$$\mathbf{E}(\xi) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = \cdots = np.$$

Poisson, $POI(\lambda)$:

$$\mathbf{E}(\xi) = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} = \cdots = \lambda.$$

Geometric, $GEO(p)$:

$$\mathbf{E}(\xi) = \sum_{k=0}^{\infty} k(1-p)^k p = \cdots = \frac{1-p}{p}.$$

Uniform, $UNI(a,b)$:

$$\mathbf{E}(\xi) = \frac{1}{b-a} \int_a^b x dx = \cdots = \frac{a+b}{2}.$$

Exponential, $EXP(\lambda)$:

$$\mathbf{E}(\xi) = \lambda \int_0^{\infty} x e^{-\lambda x} dx = \cdots = \lambda^{-1}.$$

Normal, $N(m,\sigma)$:

$$\mathbf{E}(\xi) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x \exp(-(x-m)^2/(2\sigma^2)) dx = \cdots = m.$$

Cauchy, $CAU(m,\tau)$:

$$\mathbf{E}(|\xi|) = \int_{-\infty}^{\infty} \frac{1}{\pi} \frac{\tau}{\tau^2 + (x-m)^2} |x| dx = \infty \quad !!!!$$

Warning: mean value of the Cauchy distribution (or: the expectation of a random variable having Cauchy distribution) *is not defined*!

Expectation of transformed random variable. We have seen that if ξ is a random variable with distribution function F and $\psi : \mathbb{R} \to \mathbb{R}$ is a measurable map then $\eta := \psi(\xi)$ is also a random variable. Denote its distribution function by G. If (12) holds for G (or, equivalently, (14) holds for η) then the expectation of η is defined by

$$\mathbf{E}(\eta) = \int_{-\infty}^{\infty} y dG(y).$$

By change of variable this turns out to be

$$\mathbf{E}(\eta) = m_G = \int_{-\infty}^{\infty} \psi(x) dF(x),$$

defined if and only if

$$\int_{-\infty}^{\infty} |\psi(x)| dF(x) < \infty.$$

In case of discrete, respectively, absolutely continuous distributions these formulas are

$$\sum_{x \in \mathcal{A}_F} |\psi(x)| \Delta F(x) < \infty, \qquad \mathbf{E}(\eta) = m_G = \sum_{x \in \mathcal{A}_F} \psi(x) \Delta F(x),$$

respectively,

$$\int_{-\infty}^{\infty} |\psi(x)| f(x) dx < \infty, \qquad \mathbf{E}(\eta) = m_G := \int_{-\infty}^{\infty} \psi(x) f(x) dx.$$

Example. *Log-normal, $LN(m, \sigma)$:*

$$\mathbf{E}(\xi) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp(x) \exp(-(x-m)^2/(2\sigma^2)) dx = \cdots =$$

5.2 The variance.

Another most important numerical characteristic of a distribution is its dispersion/diffuseness.

Definition 12 *Given the distribution function F, if*

$$\int_{-\infty}^{\infty} x^2 dF(x) < \infty, \tag{15}$$

then we define the dispersion *of the distribution F, as*

$$\sigma_F^2 := \int_{-\infty}^{\infty} (x - m_F)^2 dF(x) \tag{16}$$

(The integrals are Riemann-Stieltjes.)

In case of discrete, respectively, absolutely continuous distributions (15) and (16) are written as

$$\sum_{x \in \mathcal{A}_F} |x|^2 \Delta F(x) < \infty, \qquad \sigma_F^2 := \sum_{x \in \mathcal{A}_F} (x - m_F)^2 \Delta F(x),$$

respectively,

$$\int_{-\infty}^{\infty} |x|^2 f(x) dx < \infty, \qquad \sigma_F^2 := \int_{-\infty}^{\infty} (x - m_F)^2 f(x) dx.$$

Definition 13 *Let ξ be a random variable defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$. If*

$$\int_\Omega |\xi(\omega)|^2 d\mathbf{P}(\omega) < \infty, \tag{17}$$

then we define the variance *of the random variable ξ as*

$$\mathbf{Var}(\xi) := \int_\Omega (\xi(\omega) - \mathbf{E}(\xi))^2 d\mathbf{P}(\omega).$$

(The integrals are Lebesgue.)

Clearly, if ξ is a random variable and F is its probability distribution function then conditions (15) and (17) are equivalent and $\mathbf{Var}(\xi) = \sigma_F^2$. We shall use the two terms exchangeably.

By simple algebraic manipulation under the integral sign we find the identity

$$\sigma_F^2 = \int_{-\infty}^{\infty} x^2 dF(x) - m_F^2,$$

or, equivalently

$$\mathbf{Var}(\xi) = \mathbf{E}(\xi^2) - \mathbf{E}^2(\xi).$$

Examples: we compute the variance/dispersion of the various distributions encountered so far.

Indicator:

$$\mathbf{Var}(\xi) = 0(1-p) + 1p - p^2 = p(1-p).$$

Binomial, $BIN(p, n)$:

$$\mathbf{Var}(\xi) = \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} - (np)^2 = \cdots = np(1-p).$$

Poisson, $POI(\lambda)$:

$$\mathbf{Var}(\xi) = \sum_{k=0}^{\infty} k^2 e^{-\lambda} \frac{\lambda^k}{k!} - \lambda^2 = \cdots = \lambda.$$

Geometric, $GEO(p)$:

$$\mathbf{Var}(\xi) = \sum_{k=0}^{\infty} k^2 (1-p)^k p - \frac{(1-p)^2}{p^2} = \cdots = \frac{1-p}{p^2}.$$

Uniform, $UNI(a,b)$:

$$\mathbf{Var}(\xi) = \frac{1}{b-a}\int_a^b x^2 dx - \frac{(a+b)^2}{4} = \cdots = \frac{(b-a)^2}{12}.$$

Exponential, $EXP(\lambda)$:

$$\mathbf{Var}(\xi) = \lambda\int_0^\infty x^2 e^{-\lambda x} dx - \lambda^{-2} = \cdots = \lambda^{-2}.$$

Normal, $N(m,\sigma)$:

$$\mathbf{Var}(\xi) = \int_{-\infty}^\infty (x-m)^2 \frac{e^{-(x-m)^2/(2\sigma^2)}}{\sqrt{2\pi}\sigma} dx = \cdots = \sigma^2.$$

Cauchy, $CAU(m,\tau)$: the expectation was already not defined!

Theorem 14 Steiner's Theorem.
Let F be a distribution function for which (15) holds. Define $M : \mathbb{R} \to \mathbb{R}_+$ by

$$M(u) := \int_{-\infty}^\infty (x-u)^2 dF(x).$$

Then $M(u)$ is minimized (uniquely) by $u = m_F$ and $M(m_F) = \sigma_F^2$.

The proof is trivial.

5.3 Higher moments, exponential moments, etc.

Absolute moments and moments. Given a random variable ξ and its distribution function F, the *κ-th absolute moment*, $\kappa > 0$, is:

$$\mathbf{E}(|\xi|^\kappa) = \int_{-\infty}^\infty |x|^\kappa dF(x) \in [0,\infty], \qquad \kappa \in \mathbb{R}_+.$$

By Jensen's inequality, if $\kappa_1 \le \kappa_2$ then

$$\mathbf{E}(|\xi|^{\kappa_1}) \le (\mathbf{E}(|\xi|^{\kappa_2}))^{\kappa_1/\kappa_2}.$$

For $k = 1, 2, \ldots$ if $\mathbf{E}(|\xi|^k) < \infty$ then the *k-th moment* of ξ is

$$\mathbf{E}(\xi^k) = \int_{-\infty}^\infty x^k dF(x), \qquad k = 1, 2 \ldots.$$

Exercise. Compute the absolute moments and moments of the following distributions: $UNI(-1/2, 1/2)$, $EXP(1)$, $N(0,1)$.

Factorial moments of $\mathbb{N}$-valued r.v.-s. If ξ is an $\mathbb{N}$-valued random variable, its k-th *factorial moment*, $k = 1, 2 \ldots$, is

$$\mathbf{E}(\xi(\xi-1)\ldots(\xi-k+1)) = \sum_{n=0}^{\infty} \mathbf{P}(\xi = n) \prod_{j=0}^{k-1} (n-j).$$

Exercise. Compute the factorial moments of the following distributions: $BIN(p,n)$, $POI(\lambda)$, $GEO(p)$.

The exponential moment. The *moment generating function* is

$$Z : \mathbb{R} \to (0, \infty], \qquad Z(s) := \mathbf{E}(\exp(s\xi)) = \int_{-\infty}^{\infty} e^{sx} dF(x).$$

We say that the random variable has exponential moments if there exist $s_* < 0$ and $s^* > 0$ so that $Z(s)$ is finite for $s_* < s < s^*$. In this case the function $s \mapsto Z(s)$ is analytic in (s_*, s^*) and its power series is

$$Z(s) = \sum_{k=0}^{\infty} \frac{s^k}{k!} \mathbf{E}(\xi^k).$$

The *logarithmic moment generating function*

$$J : \mathbb{R} \to (-\infty, \infty], \qquad J(s) := \log Z(s)$$

has important rôle in large deviation theory. Note, that $s \mapsto J(s)$ is analytic and *convex* in (s_*, s^*):

$$J''(s) = \int_{-\infty}^{\infty} x^2 dF_s(x) - (\int_{-\infty}^{\infty} x dF_s(x))^2 > 0,$$

where F_s is the *exponentially biased probability distribution function*

$$F_s(x) := Z(s)^{-1} \int_{-\infty}^{x} e^{sy} dF(y).$$

Remark: Compare with the formalism of statistical physics.

Exercise. Compute $Z(s)$ and $J(s)$ for the following distributions: $BIN(p,n)$, $POI(\lambda)$, $GEO(p)$, $UNI(a,b)$, $EXP(\lambda)$, $N(m,\sigma)$.

The generating function. For $\mathbb{N}$-valued random variable ξ with discrete probability density function $p(n) := \mathbf{P}(\xi = n)$ we define the *generating function*

$$\psi : [0,1] \to [0,1], \qquad \psi(s) := \mathbf{E}(s^{\xi}) = \sum_{n=0}^{\infty} s^n p(n).$$

The function $s \mapsto \psi(s)$ extends analytically to the open complex unit disc. We shall say much more about the generating function in a subsequent section.

Exercise. Compute the generating function of the following N-valued distributions: $BIN(p,n)$, $POI(\lambda)$, $GEO(p)$.

The characteristic function. Given an arbitrary random variable ξ with distribution function F, its *characteristic function* is

$$\phi : \mathbb{R} \to \mathbb{C}, \qquad \phi(t) := \mathbf{E}(e^{it\xi}) = \mathbf{E}(\cos(t\xi)) + i\mathbf{E}(\sin(t\xi))$$
$$= \int_{-\infty}^{\infty} e^{itx} dF(x).$$

It is actually the Fourier-Stieltjes transform of the distribution function F. It is well defined for any F and $t \in \mathbb{R}$. It is probably the most powerful analytic tool of classical probability theory. We shall say much more about it in a subsequent chapter.

Exercise. Compute the characteristic function of all concrete distributions encountered so far.

5.4 Conditional expectation

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probablity space and $\xi : \Omega \to \mathbb{R}$ a random variable defined on it. Let $H \in \mathcal{F}$ be an event of positive probability, $\mathbf{P}(H) > 0$. Denote by $\mathbb{1}_H$ the indicator variable of the event H:

$$\mathbb{1}_H(\omega) = \begin{cases} 1 & \text{if } \omega \in H \\ 0 & \text{if } \omega \notin H. \end{cases}$$

Definition 15 *The* conditional expectation *of ξ, given the condition H is*

$$\mathbf{E}(\xi \mid H) := \frac{\mathbf{E}(\xi \mathbb{1}_H)}{\mathbf{P}(H)} = \frac{\int_H \xi(\omega) d\mathbf{P}(\omega)}{\mathbf{P}(H)}.$$

The theorem of complete probabilties transposed to conditional expectations is easily checked:

Theorem 16 Theorem of complete expectations.
Let H_j, $j = 1, 2, \ldots$ be a finite or countable partition of the probability space $(\Omega, \mathcal{F}, \mathbf{P})$. If the expectation of the random variable $\xi : \Omega \to \mathbb{R}$ is defined then

$$\mathbf{E}(\xi) = \sum_j \mathbf{E}(\xi \mid H_j)\mathbf{P}(H_j).$$

Examples and applicatins abund!

6 Random variables III.: joint distribution of random variables

6.1 Jointly defined random variables and their distributions

Usually we are concerned with more than one single (random) number assigned to a random experiment. Let the n random variables $\xi_1, \ldots, \xi_n$ be jointly defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$:

$$\xi_1, \ldots, \xi_n : \Omega \to \mathbb{R} \quad \text{or, equivalently,} \quad \vec{\xi} := (\xi_1, \ldots, \xi_n) : \Omega \to \mathbb{R}^n.$$

We assume, of course, that each map is measurable from Ω to $\mathbb{R}$, which is the same as saying that the last map is measurable form Ω to $\mathbb{R}^n$.

The *joint distribution* of these random variables is the probability measure on $(\mathbb{R}^n, \mathcal{B})$ defined for $A \in \mathcal{B}$ by

$$\begin{aligned}\mu_{\xi_1,\ldots,\xi_n}(A) &:= \mathbf{P}((\xi_1, \ldots, \xi_n) \in A) \\ &= \mathbf{P}(\{\omega \in \Omega : (\xi_1(\omega), \ldots, \xi_n(\omega)) \in A\}).\end{aligned}$$

Their *joint distribution function* is the function $F_{\xi_1,\ldots,\xi_n} : \mathbb{R}^n \to [0,1]$

$$\begin{aligned}F_{\xi_1,\ldots,\xi_n}(x_1, \ldots, x_n) &:= \mathbf{P}(\xi_1 < x_1, \ldots, \xi_n < x_n) \\ &= \mathbf{P}(\{\omega \in \Omega : \xi(\omega) < x_1, \ldots, \xi_n(\omega) < x_n\}).\end{aligned}$$

We'll see soon that, again, there is a natural one-to-one correspondence between distributions ans distribution functions on $\mathbb{R}^n$, so the two notions will be used interchangeably.

Given the random variables $\xi_1, \ldots, \xi_n$, their joint distribution function $F_{\xi_1,\ldots,\xi_n}$ has the following (easy to check) primary properties:
(i-ε) The map $(x, y, \ldots, z) \mapsto F_{\xi_1,\ldots,\xi_n}(x_1, \ldots, x_n)$ is monotone *nondecreasing in each variable.* This is immediate: it follows from positivity and additivity of the probability.
(ii) For all any $(x_1, \ldots, x_n) \in \mathbb{R}^n$ and $j = 1, 2, \ldots, n$

$$\lim_{x_j' \to -\infty} F_{\xi_1,\ldots,\xi_n}(x_1, \ldots, x_j', \ldots, x_n) = 0,$$

and

$$\lim_{x_1', \ldots, x_n' \to \infty} F_{\xi_1,\ldots,\xi_n}(x_1', \ldots, x_n') = 1.$$

These follow from (3), respectively, (2) of Theorem 3.
(iii) The function $F_{\xi_1,\ldots,\xi_n}$ is *continuous from the left in each variable*: for any $(x_1,\ldots,x_n)\in\mathbb{R}^n$ and $j=1,2,\ldots,n$

$$\lim_{x_j'\nearrow x_j} F_{\xi_1,\ldots,\xi_n}(x_1,\ldots,x_j'\ldots,x_n)=F_{\xi_1,\ldots,\xi_n}(x_1,\ldots,x_n)$$

This follows again from (2) of Theorem 3.

Property (i-ε) is immediate but it is not sharp. Actually the stronger monotonicity property of $F_{\xi_1,\ldots,\xi_n}$ holds:
(i) For any $(x_1,\ldots,x_n)\in\mathbb{R}$ and $(x_1',\ldots,x_n')\in\mathbb{R}$ with $x_1\le x_1',\ldots,x_n\le x_n'$ and $\alpha_1,\ldots,\alpha_n\in\{0,1\}$

$$\begin{aligned}&\textstyle\sum_{\alpha_1,\ldots,\alpha_n=0}^{1}(-1)^{\alpha_1+\cdots+\alpha_n}\times \qquad (18)\\ &F_{\xi_1,\ldots,\xi_n}(\alpha_1x_1+(1-\alpha_1)x_1',\ldots,\alpha_nx_n+(1-\alpha_n)x_n)\ge 0.\end{aligned}$$

This property follows directly from the sieve formula, (1) of Theorem 2 (check it!): it means that

$$\mathbf{P}(\xi_1\in[x_1,x_1'),\ \ldots\ ,\xi_n\in[x_n,x_n'))\ge 0.$$

Of course, it implies the simple monotonicity formulated as property (i-ε).

Given a function $F:\mathbb{R}^n\to[0,1]$, with properties (i), (ii) and (iii) one can construct a probability space $(\Omega,\mathcal{F},\mathbf{P})$ and random variables $(\xi_1,\ldots,\xi_n):\Omega\to\mathbb{R}^n$, so that the joint distribution function of $(\xi_1,\ldots,\xi_n)$ is exactly F. A function $F:\mathbb{R}^n\to[0,1]$ possessing properties (i), (ii) and (iii) will be called probability distribution function (of n variables).

There is natural a one-to-one correspondence between probability distribution functions $F:\mathbb{R}^n\to[0,1]$ and probability measures (probability distributions) on $(\mathbb{R}^n,\mathcal{B})$ realized by the following identifications.
Given the probability measure μ on $(\mathbb{R}^n,\mathcal{B})$, let $F_\mu:\mathbb{R}^n\to[0,1]$,

$$F_\mu(x_1,\ldots,x_n):=\mu((-\infty,x_1)\times\cdots\times(-\infty,x_n))$$

Given the probability distribution function $F:\mathbb{R}^n\to[0,1]$ let the measure of rectangles $(x_1,x_1']\times\cdots\times[x_3,z_3')$, $x_1\le x_1',\ \ldots\ ,\ x_n\le x_n'$, be given by (18) and extend it to all Borel measurable sets $A\in\mathcal{B}$ by standard measure-theoretic procedures. By this one-to-one correspondence we are allowed to interchange freely the two objects.

Marginal distributions. The joint distribution $F_{\xi_1,\dots,\xi_n}$ contains all information about the random variables. In particular we may ask about the distribution of each variable separately. It is straightforward that these (one variable) distribution functions are retrieved from the $F_{\xi_1,\dots,\xi_n}$, by

$$F_{\xi_j}(x_j) = \lim_{\substack{x_i \to \infty \\ i:\, i \neq j}} F_{\xi_1,\dots,\xi_n}(x_1, \dots, x_n)$$

These are called the *one dimensional marginal distributions* of the joint distribution $F_{\xi_1,\dots,\xi_n}$. *Warning:* the full collection of the one-dimensional marginal distributions contains much less information than the joint distribution.

For any *subcollection* $1 \leq i_1 < \cdots < i_r \leq n$ one can obviously define the marginal distribution of the random variables $\xi_{i_1}, \dots, \xi_{i_r}$

Discrete and absolutely continuous distributions on $\mathbb{R}^n$.
The point $(x_1, \dots, x_n) \in \mathbb{R}^n$ is an *atom* of the distribution function F if the limit of the expression in (18), as $x'_j \searrow x_j$, $j = 1, 2, \dots, n$, is positive. This means exactly that the corresponding measure μ_F gives positive weight to the single point set $\{(x_1, \dots, x_n)\}$. As in the one-dimensional case, we shall denote by $\mathcal{A}_F$ the set of atoms of F and by $\Delta F(x_1, \dots, x_n)$ the weight of the atom.

Definition 17 *(i) The probability distribution function $F : \mathbb{R}^n \to [0,1]$ is* pure point *if for any $(x_1, \dots, x_n) \in \mathbb{R}^n$*

$$F(x_1, \dots, x_n) = \sum_{\substack{(x'_1,\dots,x'_n) \in \mathcal{A}_F \\ x'_1 < x_1, \dots, x'_n < x_n}} \Delta F(x'_1, \dots, x'_n).$$

This means exactly that the whole weight of the distribution is concentrated on the set of atoms. We call discrete *a pure point distribution if $\mathcal{A}_F$ does not have points of accumulation.*
(ii) The probability distribution function $F : \mathbb{R}^n \to [0,1]$ is called continuous *if $\mathcal{A}_F = \emptyset$, i.e., no single point gets positive weight. In this case the function F is indeed continuous.*
(iii) The probability distribution function $F : \mathbb{R}^n \to [0,1]$ is absolutely continuous *if there exists a measurable function $f : \mathbb{R}^n \to \mathbb{R}_+$ such that for any $(x_1, \dots, x_n) \in \mathbb{R}^n$,*

$$F(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(x'_1, \dots, x'_n)\, dx'_1 \dots dx'_n.$$

The function f is called the density *of the distribution.*
(iv) The probability distribution function $F : \mathbb{R}^n \to [0,1]$ is continuous but singular *if it is continuous, (Lebesgue) almost everywhere n-times differentiable and $\partial^n F/(\partial x_1 \dots \partial x_n) = 0$ (Lebesgue) almost everywhere.*

Remarks. (1) The set of atoms $\mathcal{A}_F$ is at most countable. (Prove it!)
(2) In the definition of 'continuous' distributions it seems tautological to say that 'F is indeed continuous'. But this is not the case: a priori this is another notion of continuity.
(3) In the absolutely continuous case: the density function f a fortiori satisfies

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1', \dots, x_n')\, dx_1' \, \dots \, dx_n' = 1. \tag{19}$$

A function $f : \mathbb{R}^n \to \mathbb{R}_+$ (19) will be called *probability density function* on $\mathbb{R}^n$.
(4) In the absolutely continuous case: it is a fact (check it!) that if f_1 and f_2 are density functions of the same distribution function F, then $f_1 = f_2$ (Lebesgue) almost everywhere.
(5) Still in the absolutely continuous case: (Lebesgue) almost everywhere

$$\frac{\partial^n F}{\partial x_1 \dots \partial x_n} = f.$$

(6) Examples of continuous but singular distributions are measures concentrated on lower dimensional submanifolds. But wilder examples also abund, see e.g. Cantor's function for $n = 1$.
(7) Lebesgue's decomposition theorem, Theorem (8), is transposed word-by-word to the n dimensional case.

It is evident that all marginal distributions of a pure point, respectively, absolutely continuous distribution, are pure point, respectively, absolutely continuous. But: it may easily happen that the marginal of a discrete distribution is pure point but not discrete, or the marginal of a continuous but singular distribution is absolutely continuous. Construct examples of this kind!

The densities of the one dimensional marginals of an absolutely continuous n-variate distribution function F are expressed from the density f of

F, as follows

$$f_j(x_j) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n) dx_1 \ldots, \widehat{dx_j}, \ldots, dx_n$$

Examples of discrete joint distributions.

The polynomial distribution, $POLYN(p_1, p_2, \ldots, p_r, n)$**:** Fixed parameters are $r \in \mathbb{N}$, $p_1, p_2, \ldots, p_r \in (0,1)$ with $p_1 + p_2 + \cdots + p_r = 1$ and $n \in \mathbb{N}$. Let a random experiment have r possible reults, say $1, 2, \ldots, r$, with probabilities $p_1, p_2, \ldots, p_r \in (0,1)$, respectively. Perform n independent identical trials of this experiment and denote by ξ_j the total number of results j in the sequence of experiments, where $j = 1, 2, \ldots, r$. Clearly, $\xi_1 + \xi_2 + \cdots + \xi_r = n$. Elementary combinatorial considerations yield the joint distribution of these random variables

$$\mathbf{P}(\xi_1 = k_1, \xi_2 = k_2, \ldots, \xi_r = k_r) = \binom{n}{k_1, k_2, \ldots, k_r} p_1^{k_1} p_2^{k_2} \ldots p_r^{k_r}$$

Where we use the conventional notation

$$\binom{n}{k_1, k_2, \ldots, k_r} := \begin{cases} \dfrac{n!}{k_1! k_2! \ldots k_r!} & \text{if } k_1, k_2, \ldots, k_r \in \{0, 1, 2, \ldots, n\} \\ & \text{and } k_1 + k_2 + \cdots + k_r = n \\ 0 & \text{otherwise.} \end{cases}$$

Sampling with replacement: Let $N_1, N_2, \ldots, N_r \in \mathbb{N}$ and denote $p_j := N_j/(N_1 + N_2 + \cdots + N_r)$. In an urn there are $N_1 + N_2 + \cdots + N_r$ balls, N_j balls of colour j, $j = 1, 2, \ldots, r$. We draw (randomly) n consecutive times one ball, record its colour and replace it. Denote by ξ_j the number of results j in this sequence of trials. The joint distribution of $\xi_1, \xi_2, \ldots, \xi_r$ will be polynoomial with the given parameters.

Exercise: Find the one-dimensional marginal distributions of the polynomial distribution. What are the higher dimensional marginals?

The polyhypergeometric distribution, $POLYHYP(N_1, N_2, \ldots, N_r, n)$**:** Fixed parameters are $r \in \mathbb{N}$, $N_1, N_2, \ldots, N_r \in \mathbb{N}$ and $n \in \mathbb{N}$, with $n \leq N_1 + N_2 + \cdots + N_r$.

Sampling without replacement: Let $N_1, N_2, \ldots, N_r \in \mathbb{N}$. In an urn there are $N_1 + N_2 + \cdots + N_r$ balls, N_j balls of colour j, $j = 1, 2, \ldots, r$. This time we draw $n \leq N_1 + N_2 + \cdots + N_r$ balls *without replacing them.* Denote again

by ξ_j the number of balls of colour j drawn in this sequence. Again, by elementary combinatorial considerations we find

$$\mathbf{P}(\xi_1 = k_1, \xi_2 = k_2, \ldots, \xi_r = k_r) = \frac{\binom{N_1}{k_1}\binom{N_2}{k_2}\cdots\binom{N_r}{k_r}}{\binom{N_1+N_2+\cdots+N_r}{n}}.$$

Exercise: Find the marginals of the polyhypergeometric distribution.

Convergence of $POLYHYP(N_1, \ldots, N_r, n)$ to $POLYN(p_1, \ldots, p_r, n)$: If the number of balls of each colour is much larger than the number n of draws then there should be no significant difference between sampling with or without replacement. Intuitively this is very clear. Mathematically speaking: Keeping n and $k_1, k_2, \ldots, k_r$ fixed, let $N_1, N_2, \ldots, N_r \to \infty$ so that $N_j/(N_1+N_2+\cdots+N_r) \to p_j \in (0,1)$ (note that a fortiori: $p_1+p_2+\cdots+p_r = 1$). Then

$$\frac{\binom{N_1}{k_1}\binom{N_2}{k_2}\cdots\binom{N_r}{k_r}}{\binom{N_1+N_2+\cdots+N_r}{n}} \to \binom{n}{k_1, k_2, \ldots, k_r} p_1^{k_1} p_2^{k_2} \ldots p_r^{k_r}.$$

Examples of absolutely continuous joint distributions.

The uniform distribution, $U(D)$: The fixed parameter is a nice compact domain $D \subset \mathbb{R}^n$. By nice we mean that D is equal to the (topological) closure of its interior. Let $|D|$ be the volume of the domain D. The density function is

$$f(x_1, x_2, \ldots, x_n) = \begin{cases} |D|^{-1} & \text{if} \quad (x_1, x_2, \ldots, x_n) \in D \\ 0 & \text{if} \quad (x_1, x_2, \ldots, x_n) \notin D \end{cases}$$

We have seen plenty of examples in the classical geometric problems.

The n dimensional normal (or Gaussian) distribution, $N(\vec{m}, \mathbf{C})$: We use now vectorial notations: $\vec{x} := (x_1, \ldots, x_n)^\dagger \in \mathbb{R}^n$ etc. The fixed parameters are the vector $\vec{m} = (m_1, \ldots, m_n)^\dagger \in \mathbb{R}^n$ and the *positive definite* real matrix $\mathbf{C} = (c_{ij})_{ij=1}^n$. Browse up your linear algebra notes for basic facts about matrices, quadratic forms, positive definiteness etc.). The positive definite matrix $\mathbf{C}$ *is symmetric and regular (invertible).* Denote its inverse

matrix by $\mathbf{A}$, which is itself positive definite. The two-dimensional normal density function, with obvious notations, is

$$f(x_1,\dots,x_n) := \sqrt{\frac{\det \mathbf{A}}{(2\pi)^n}} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{m})^\dagger \mathbf{A}(\vec{x}-\vec{m})\right\} \tag{20}$$

$$= \sqrt{\frac{\det \mathbf{A}}{(2\pi)^n}} \exp\left\{-\frac{1}{2}\sum_{i,j=1}^{n}(x_i-m_i)A_{ij}(x_j-m_j)\right\}.$$

Prove that this is indeed a probability density function, i.e., check (19).
The standard choice is $\vec{m}=\vec{0}$, $\mathbf{C}=\mathbf{I}$, i.e., $C_{ij}=\delta_{i,j}$. Let $\vec{\xi}=(\xi_1,\dots,\xi_n)$ have distribution $N(\vec{0},\mathbf{I})$ $\eta := \mathbf{C}^{1/2}\vec{\xi}+\vec{m}$ then $\vec{\eta}$ will have $N(\vec{m},\mathbf{C})$ distridution.
More later about the multivariate normal distribution.

6.2 Transformation of random variables

Please browse up your notes on multivariate analysis!

Regular transformation of random variables:
Let $D\subset\mathbb{R}^n$ be open and $\vec{\psi}: D\to\mathbb{R}^n$ be a smooth (differentiable) vector field. Its *Jacobian matrix* is

$$\mathbf{J}^{\vec{\psi}}: \mathbb{R}^n \to M_{n\times n}(\mathbb{R}), \qquad J^{\vec{\psi}}_{ij}(x_1,\dots,x_n) := \frac{\partial \psi_i}{\partial x_j}(x_1,\dots,x_n).$$

(Here $M_{n\times n}(\mathbb{R})$ denotes the space of n by n real real matrices.) The point $(x_1,\dots,x_n)\in D$ is called *regular*, respectively, *singular*, if $\det \mathbf{J}^{\vec{\psi}}(x_1,\dots,x_n)\neq 0$, respectively, if $\det \mathbf{J}^{\vec{\psi}}(x_1,\dots,x_n)=0$. The vector field $\vec{\psi}: D\to\mathbb{R}^n$ is *regular* if all $(x_1,\dots,x_n)\in D$ are regular. A regular vector field $\vec{\psi}$ is invertible on its range. The vector field $\vec{\psi}$ is *piecewise regular* if $D=\overline{D}_1\cup\overline{D}_2\cup\cdots\cup\overline{D}_k$ such that $\vec{\psi}|_{D_j}$, $j=1,2,\dots,k$ are regular.

Let $\vec{\xi}:=(\xi_1,\dots,\xi_n)$ be n real valued random variable with *absolutely continuous* joint distribution. Denote the density function of their joint distribution by $f_{\xi_1,\dots,\xi_n}$. Let $\vec{\psi}:\mathbb{R}^n\to\mathbb{R}^n$ be a vectror field and $(\eta_1,\dots,\eta_n)=\vec{\eta}:=\vec{\psi}(\vec{\xi})$. If ψ is piecewise regular then the joint distribution of the random variables $\eta_1,\dots,\eta_n$ will be absolutely continuous, too, and their joint density function $f_{\eta_1,\dots,\eta_n}$ will be

$$f_{\eta_1,\dots,\eta_n}(y_1,\dots,y_n) = \sum_{\vec{x}:\,\vec{\psi}(\vec{x})=\vec{y}} \frac{f_{\xi_1,\dots,\xi_n}(x_1,\dots,x_n)}{|\det \mathbf{J}^{\vec{\psi}}(x_1,\dots,x_n)|}.$$

This formula is exactly Jacobi's Theorem of classical multivariate analysis, formulated for probability density functions.

Particular case: regular linear transformations. Let $\mathbf{A} \in M_{n\times n}(\mathbb{R})$ be a regular n by n matrix and $\vec{b} \in \mathbb{R}^n$ and $\vec{\eta} = \mathbf{A}\vec{\xi} + \vec{b}$. Then the density function transoforms as

$$f_{\vec{\eta}}(\vec{y}) = |\det \mathbf{A}|^{-1} f_{\vec{\xi}}(\mathbf{A}^{-1}(\vec{x} - \vec{b})).$$

Real functions of random variables:

Let $\xi_1, \ldots, \xi_n$ be n real valued random variables, jointly defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and denote by $F_{\xi_1,\ldots,\xi_n}$ their joint distribution function. Let $\psi : \mathbb{R}^n \to \mathbb{R}$ be a measurable function. Then $\xi := \psi(\xi_1, \ldots, \xi_n)$ is a real valued random variable. Its distribution function is expressed:

$$\begin{aligned} F_\xi(x) &:= \mathbf{P}(\xi < x) = \mathbf{P}(\psi(\xi_1, \ldots, \xi_n) < x) \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbb{1}_{\{\psi(x_1,\ldots,x_n)<x\}} dF_{\xi_1,\ldots,\xi_n}(x_1, \ldots, x_n) \\ &= \mu_{\xi_1,\ldots,\xi_n}(\{(x_1, \ldots, x_n) \in \mathbb{R}^n : \psi(x_1, \ldots, x_n) < x\}) \end{aligned} \quad (21)$$

This formula makes perfectly good sense, but it is rather unpleasant.

TO BE COMPLETED

6.3 Independence of random variables

Definition 18 *The random variables $\xi_1, \ldots, \xi_n$ are (completely)* independent *if for any intervals $I_1, \ldots, I_n \subset \mathbb{R}$ the events $\{\xi_1 \in I_1\}, \ldots, \{\xi_n \in I_n\}$ are independent. That is: if for any subcollection $1 \le i_1 < \cdots < i_r \le n$:*

$$\mathbf{P}(\xi_{i_1} \in I_{i_1}, \ldots, \xi_{i_r} \in I_{i_r}) = \mathbf{P}(\xi_{i_1} \in I_{i_1}) \ldots \mathbf{P}(\xi_{i_r} \in I_{i_r})$$

In terms of the joint distribution function:

Theorem 19 Factorization of joint distribution of independent random variables.

The random variables $\xi_1, \ldots, \xi_n$ are independent if and only if their joint distribution function factorizes:

$$F_{\xi_1,\ldots,\xi_n}(x, y, \ldots, z) = F_{\xi_1}(x_1) \ldots F_{\xi_n}(x_n). \quad (22)$$

If the joint distribution of $\xi_1, \ldots, \xi_n$ is absolutely continuous, then they are independent if and only if the joint density function factorizes:

$$f_{\xi_1,\ldots,\xi_n}(x_1,\ldots,x_n) = f_{\xi_1}(x_1)\ldots f_{\xi_n}(x_n).$$

Theorem 20 Functions of independent random variables are independent. *Let $\xi_1, \ldots, \xi_n$ be independent random variables and $\psi_1, \ldots, \psi_n : \mathbb{R} \to \mathbb{R}$ measurable functions. Then the transformed random variables $\eta_1 := \psi_1(\xi_1)$, $\ldots$, $\eta_n := \psi_n(\xi_n)$ are also independent.*

The proofs is these theorems are straightforward.

Examples.
(1) Let $D = [a_1, a_1'] \times \cdots \times [a_n, a_n']$ be a rectangle in $\mathbb{R}^n$. It is straightforward to check that the random variables $\xi_1, \ldots, \xi_n$ with joint uniform distribution on D, are independent.
(2) Let $(\xi_1, \ldots, \xi_n)$ have normal distribution $N(\vec{m}, \mathbf{C})$ with $\mathbf{C}$ *diagonal*: $C_{ij} = \sigma_i^2 \delta_{i,j}$, Simple computations show that in this case (20) reads

$$f(x_1,\ldots,x_n) := \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x_i - m_i)^2}{2\sigma_i^2}\right).$$

That is: the random variables $\xi_1, \ldots, \xi_n$ are independent Gaussians of expectation m_j and variance σ_j^2, $j = 1, 2, \ldots, n$.

6.4 Conditional distribution

TO BE COMPLETED

6.5 Expectation and covariance

Let $\xi_1, \ldots, \xi_n$ be n real valued random variables, jointly defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$, and $\xi := \psi(\xi_1, \ldots, \xi_n)$, where $\psi : \mathbb{R}^n \to \mathbb{R}$ is a measurable function. By definition of the expectation:

$$\mathbf{E}(\xi) := \int_0^\infty x dF_\xi(x),$$

with the distribution function F_ξ given by (21). By change of variable under the integral we actually get

Theorem 21 Expectation of real functional of several random variables.
Let $\xi := \psi(\xi_1, \ldots, \xi_n)$. The expectation $\mathbf{E}(\xi)$ is defined if and only if

$$\begin{aligned}\mathbf{E}(|\xi|) &= \mathbf{E}(|\psi(\xi_1, \ldots, \xi_n)|) \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |\psi(x_1, \ldots, x_n)| dF_{\xi_1,\ldots,\xi_n}(x_1, \ldots, x_n) < \infty.\end{aligned}$$

In this case

$$\begin{aligned}\mathbf{E}(\xi) &= \mathbf{E}(\psi(\xi_1, \ldots, \xi_n)) \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \psi(x_1, \ldots, x_n) dF_{\xi_1,\ldots,\xi_n}(x_1, \ldots, x_n)\end{aligned} \tag{23}$$

The proof relies on change of variable under the integral sign. We omit this proof.

A first consequence of this formula is the simple but important fact is that *the expectation is linear*:

Theorem 22 Linearity of the expectation.
Let $\xi_1, \ldots, \xi_n$ be random variables, jointly defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and $a_1, \ldots, a_n$ real numbers. Consider the random variable $\xi := a_1\xi_1 + \cdots + a_n\xi_n$. If $\mathbf{E}(|\xi_j|) < \infty$, $j = 1, \ldots, n$, then $\mathbf{E}(|\xi|) < \infty$ and

$$\mathbf{E}(\xi) := \mathbf{E}(a_1\xi_1 + \cdots + a_n\xi_n) = a_1\mathbf{E}(\xi_1) + \cdots + a_n\mathbf{E}(\xi_n).$$

This follows immediately from (23) and linearity of integration.

Another simple consequence of formula (23) is thr following:

Theorem 23 Factorization of expectation of product of independent rv-s.
Let $\xi_1, \ldots, \xi_n$ be random variables, jointly defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and define $\xi := \xi_1 \cdot \xi_2 \cdot \cdots \cdot \xi_n$. If $\mathbf{E}(|\xi_j|) < \infty$, $j = 1, \ldots, n$, then $\mathbf{E}(|\xi|) < \infty$ and

$$\mathbf{E}(\xi) := \mathbf{E}(\xi_1 \cdot \xi_2 \cdot \cdots \cdot \xi_n) = \mathbf{E}(\xi_1)\mathbf{E}(\xi_2) \cdot \mathbf{E}(\xi_n).$$

This follows directly from (23) and (**??**) of Theorem **??**.

Theorem 24 Schwarz's inequality.
Let ξ and η be two random variables, jointly defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Assume $\mathbf{E}(\xi^2) < \infty$, $\mathbf{E}(\eta^2) < \infty$. The following inequality holds:

$$|\mathbf{E}(\xi\eta)| \leq \sqrt{\mathbf{E}(\xi^2)\mathbf{E}(\eta^2)}$$

P r o o f.
Define $f : \mathbb{R} \to \mathbb{R}$ by $f(\lambda) = \mathbf{E}((\xi + \lambda\eta)^2)$. Then

$$0 \leq f(\lambda) = \mathbf{E}(\xi^2) + 2\lambda\mathbf{E}(\xi\eta) + \lambda^2\mathbf{E}(\eta^2).$$

This holds for all $\lambda \in \mathbb{R}$ only if the discriminant is non-positive: $(\mathbf{E}(\xi\eta))^2 \leq \mathbf{E}(\xi^2)\mathbf{E}(\eta^2)$, and the inequality is proved.

Definition 25 *(i) Let ξ and η be two random variables, jointly defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Assume $\mathbf{E}(\xi^2) < \infty$, $\mathbf{E}(\eta^2) < \infty$. The* covariance *of ξ and η is*

$$\mathbf{Cov}(\xi, \eta) := \mathbf{E}\big((\xi - \mathbf{E}(\xi))(\eta - \mathbf{E}(\eta))\big) = \mathbf{E}(\xi\eta) - \mathbf{E}(\xi)\mathbf{E}(\eta).$$

The correlation coefficient *of ξ and η is*

$$\mathbf{R}(\xi, \eta) := \frac{\mathbf{Cov}(\xi, \eta)}{\sqrt{\mathbf{Var}(\xi)\mathbf{Var}(\eta)}}.$$

We say that the random variables are positively/negatively correlated, *respectively,* uncorrelated, *iff their covariance is positive/negative, respectively, zero.*
(ii) Let $\xi_1, \ldots, \xi_n$ be n random variables, jointly defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Assume $\mathbf{E}(\xi_j^2) < \infty$, $j = 1, \ldots, n$. Their covariance matrix *is*

$$\mathbf{C} := (\mathbf{Cov}(\xi_i, \xi_j))_{i,j=1}^n.$$

Remarks. (1) The covariance is the most important numerical indicatior of two random variables influence on each other. ... ... Independence of ξ and η imply that they are uncorrelated. The converse is not true: uncorrelated random variables may well depend. Construct examples and counterexamples!
(2) Note, that $\mathbf{Cov}(\xi, \xi) = \mathbf{Var}(\xi)$.
(3) Let $a, b, c, d \in \mathbb{R}$. Then $\mathbf{Cov}((a\xi + b), (c\eta + d)) = ac\mathbf{Cov}(\xi, \eta)$ – this is again straightforward.
(4) Schwarz's inequality implies that

$$|\mathbf{Cov}(\xi, \eta)| \leq \sqrt{\mathbf{Var}(\xi)\mathbf{Var}(\eta)},$$

and, consequently $\mathbf{R}(\xi, \eta) \in [-1, 1]$.

(5) The covariance matrix is, by its definition, real and symmetric.

(6) Also, from Schwarz's inequality it follows that the covariance matrix $\mathbf{C}$ is positive semidefinite and, if there is no deterministic linear relation between the random variables $\xi_1, \ldots, \xi_n$, then it is positive definite. Indeed, given any complex numbers $z_1, \ldots, z_n$

$$\sum_{i,j=1}^{n} z_i \overline{z}_j \mathbf{Cov}(\xi_i, \xi_j) = \cdots = \mathbf{E}\left(\left| \sum_{i=1}^{n} z_i (\xi_i - \mathbf{E}(\xi_i)) \right|^2 \right) \geq 0$$

where equality holds *only if* $\sum_{i=1}^{n} z_i(\xi_i - \mathbf{E}(\xi_i)) \equiv 0$, almost surely.

Theorem 26 Variance of sum of random variables.

Let $\xi_1, \ldots, \xi_n$ be n random variables, jointly defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Assume $\mathbf{E}(\xi_j^2) < \infty$, $j = 1, \ldots, n$. Let $S := \xi_1 + \xi_2 + \cdots + \xi_n$. Then,

$$\mathbf{Var}(S) = \sum_{i=1}^{n} \mathbf{Var}(\xi_i) + 2 \sum_{i=1}^{n} \sum_{j=1}^{i-1} \mathbf{Cov}(\xi_i, \xi_j).$$

If the random variables are uncorrelated (in particular: if they are independent)

$$\mathbf{Var}(S) = \sum_{i=1}^{n} \mathbf{Var}(\xi_i).$$

This is again elementary.

6.6 Sums of independent random variables: the convolution

Discrete convolutions.

Let ξ and η be *independent* Z-valued random variables and let their distribution be

$$f(k) := \mathbf{P}(\xi = k) \qquad g(l) := \mathbf{P}(\eta = l).$$

Let $\zeta := \xi + \eta$. We want to express the distribution

$$h(m) := \mathbf{P}(\zeta = m)$$

in terms of f and g.

$$h(m) = \mathbf{P}(\zeta = m) = \mathbf{P}(\bigcup_{k=-\infty}^{\infty} \{\xi = k,\ \eta = m-k\})$$

$$= \sum_{k=-\infty}^{\infty} \mathbf{P}(\xi = k,\ \eta = m-k) = \sum_{k=-\infty}^{\infty} \mathbf{P}(\xi = k)\mathbf{P}(\eta = m-k)$$

$$= \sum_{k=-\infty}^{\infty} f(k)g(m-k). =: \big(f * g\big)(m).$$

In the third and fourth equalities additivity of the probability and independence of ξ and η were used, respectively. The last equality is the formal definition of *convolution* of two discrete distributions defined over $\mathbb{Z}$. If the random variables ξ and η are actually $\mathbb{N}$-valued then so is ζ and f, g, h vanish for negative values of their respective variables. In this case we get the simpler expression:

$$h(m) = \big(f * g\big)(m) = \sum_{k=0}^{m} f(k)g(m-k).$$

Let $\mathcal{E}_Z$ and $\mathcal{E}_N$ denote the space of all discrete distribution densities over $\mathbb{Z}$, respectively, $\mathbb{N}$:

$$\mathcal{E}_{\mathbb{Z}} := \{f : \mathbb{Z} \to [0,1] : \sum_{k=-\infty}^{\infty} f(k) = 1\}$$

$$\mathcal{E}_{\mathbb{N}} := \{f : \mathbb{N} \to [0,1] : \sum_{k=0}^{\infty} f(k) = 1\}$$

Theorem 27 Convolutions ... .

The convolution is a binary operation in $\mathcal{E}_{\mathbb{Z}}$ and $\mathcal{E}_{\mathbb{N}}$. $(\mathcal{E}_{\mathbb{Z}}, *)$ and $(\mathcal{E}_{\mathbb{N}}*)$ are Abelian (commutative) semigroups with the neutral element e, $e(k) = \delta_{0,k}$.

P r o o f.

Straightforward computations yield associativity and commutativity of the convolution and the fact that e is indeed the neutral element. These are only reflections of the facts that addition of integers is associative and commutative, with 0 as neutral element.

Warning: $(\mathcal{E}_{\mathbb{Z}}, *)$ (respectively, $(\mathcal{E}_{\mathbb{N}}*)$) is by no means a group. There is no inverse element defined! Explain why.

Examples.

Convolution of binomial distributions:

$$BIN(p,m) * BIN(p,n) = BIN(p,n+m).$$

Prove it! In particular,

$$BIN(p,n) = IND(p) * IND(p) * \cdots * IND(p) \qquad (n\text{-fold}).$$

Convolution of Poisson distributions:

$$POI(\lambda) * POI(\mu) = POI(\lambda+\mu).$$

Prove it!

Continuous convolutions.

Let ξ and η be two independent random variables jointly defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Denote their distribution functions F, respectively, G. Define $\zeta := \xi + \eta$ and denote its distribution function by H. Then H is expressed as follows:

$$H(z) = \int_{-\infty}^{\infty} F(z-y)\, dG(y) = \int_{-\infty}^{\infty} G(z-x)\, dF(x). \tag{24}$$

The formulas are intuitively quite clear – by discretization one can reduce them to the discrete formulas presented in the prvious paragraph. Here is a sketchy proof:

$$\begin{aligned}
&H(z) := \mathbf{P}(\zeta < z) = \mathbf{P}(\xi + \eta < z) \\
&= \mathbf{P}\left(\bigcup_{m<\infty} \bigcup_{k=-\infty}^{\infty} \left\{ \eta \in [\frac{k}{2^m}, \frac{k+1}{2^m}),\ \xi < \frac{\lfloor 2^m z - (k+1) \rfloor}{2^m}) \right\} \right) \\
&= \lim_{m\to\infty} \mathbf{P}\left(\bigcup_{k=-\infty}^{\infty} \left\{ \eta \in [\frac{k}{2^m}, \frac{k+1}{2^m}),\ \xi < \frac{\lfloor 2^m z - (k+1) \rfloor}{2^m}) \right\} \right) \\
&= \lim_{m\to\infty} \sum_{k=-\infty}^{\infty} \mathbf{P}\left(\eta \in [\frac{k}{2^m}, \frac{k+1}{2^m}),\ \xi < \frac{\lfloor 2^m z - (k+1) \rfloor}{2^m}) \right) \\
&= \lim_{m\to\infty} \sum_{k=-\infty}^{\infty} \mathbf{P}\left(\eta \in [\frac{k}{2^m}, \frac{k+1}{2^m}) \right) \mathbf{P}\left(\xi < \frac{\lfloor 2^m z - (k+1) \rfloor}{2^m}) \right) \\
&= \lim_{m\to\infty} \sum_{k=-\infty}^{\infty} F(\frac{\lfloor 2^m z - (k+1) \rfloor}{2^m}) \left(G(\frac{k+1}{2^m}) - G(\frac{k}{2^m}) \right) = \dots \\
&= \int_{-\infty}^{\infty} F(z-y)\, dG(y).
\end{aligned}$$

The first two equalities are just definitions. In order to understand the third (and forthcoming) steps, DO make a picture! The fourth step is application of (2) from Theorem 3. The fifth step is sigma additivity of probablities. In the sixth step independence of ξ and η is used. The seventh step is just transcription in terms of the distribution functions F and G. The ... stand for standard analytic procedures: this is actually discrete approximation of the Riemann-Stieltjes integral with respect to G. Here we have to use *continuity from the left* of the distribution function F.

Theorem 27 has its natural extention: denote by $\mathcal{E}_{\mathbb{R}}$ the space of distribution functions on $\mathbb{R}$.

Theorem 28 Convolutions ... .

The convolution defined in (24) is a binary operation in $\mathcal{E}_{\mathbb{R}}$. $(\mathcal{E}_{\mathbb{R}}, *)$ *is Abelian (commutative) semigroup with the neutral element* E, $e(x) = \mathbb{1}_{(0,\infty)}(x)$.

The proof of associativity, commutativity and neutrality of $E(\cdot)$ is immediate.

If F and G are absolutely continuous with densities $f = F'$, respectively, $g = G'$ then so is $H = F * G$ and the density $h = H'$ is

$$h(z) = \int_{-\infty}^{\infty} f(z-y)g(y)dy = \int_{-\infty}^{\infty} f(x)g(z-x)dx.$$

Examples.

Uniform: Let $f(x) = \mathbb{1}_{[-1/2,1/2]}(x)$ be the density of the distribution $UNI(-1/2, 1/2)$. Then

$$\big(f * f\big)(x) = \big(1 - |x|\big)_+.$$

Check it and compute also $f * f * f$.

Normal:

$$N(m_1, \sigma_1^2) * N(m_2, \sigma_2^2) = N(m_1 + m_2, \sigma_1^2 + \sigma_2^2)$$

Check it! Use Gaussian integrals.

Cauchy:

$$CAU(m_1, \tau_1) * CAU(m_2, \tau_2) = CAU(m_1 + m_2, \tau_1 + \tau_2)$$

Check it! Use complex integration and the theorem of residues.

In Definition 9 we defined an equivalence relation on the set $\mathcal{E}_{\mathbb{R}}$.

Definition 29 *A class of distributions of the same type is called* stable *if it is closed under convolutions. I.e., if the sum of two independent random variables of the same type is again of the same type. With slight abuse of terminology we shall call a stable distribution belonging to a stable class.*
A class of equivalent distributions is called symmetric *if it has a representative F satisfying $F(-x) = 1 - F(x)$ (i.e., the representative F is symmetric about 0).*

The last two examples show that the normal (Gaussian) and the Cauchy distributions are symmetric and stable. Symmetric and stable distributions have a very special rôle in the theory limiting distributions of (rescaled) sums of independent and identically distributed (i.i.d.) random variables.

Convolution of exponentials: the gamma distributions. We are going to compute $EXP(\lambda) * EXP(\lambda) * \cdots * EXP(\lambda)$. Let $\lambda \in (0, \infty)$ be fixed and denote the density function of the distribution $EXP(\lambda)$ by f_λ:

$$f(x) := \mathbb{1}_{[0,\infty)}(x)\lambda e^{-\lambda x}$$

Let

$$\gamma_1 := f, \qquad \gamma_{n+1} := \gamma_n * f.$$

That is: γ_n is the density function of the distribution of sum of n i.i.d. $EXP(\lambda)$-distributed random variables.

One can easily prove by induction that

$$\gamma_n(x) = \mathbb{1}_{[0,\infty)}(x)\frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1!}.$$

The distribution with this density function is called *gamma distribution* of parameters λ and n. We shall denote it $GAM(\lambda, n)$. It is the distribution of the sum of n i.i.d. *memoryless waiting times.*
MORE TEXT HERE

Relation between $GAM(\lambda, n)$ and $POI(\lambda)$. Let $\tau_1, \tau_2, \ldots$ be i.i.d. $EXP(\lambda)$ distributed waiting times (e.g., τ_n could be the time elapsed between the $(n-1)$-th and n-th beep of a Geiger-Müller counter. Denote

$$T_n := \sum_{j=1}^{n} \tau_j.$$

In plain words: T_n is the time of the n-th event (beeps of the Geiger-Müller counter) occuring. Define also for $t > 0$

$$\nu_t := \max\{n : T_n < t\}.$$

That is: ν_t is the number of events (beeps) occuring up to time t. Clearly, by definition of ν_t

$$\{\nu_t \geq n\} = \{T_n < t\}.$$

We prove that ν_t is $POI(\lambda t)$-distributed.

This is equivalent to proving

$$\int_0^t \gamma_n(s)\, ds = \sum_{k=n}^{\infty} e^{-\lambda t} \frac{(\lambda t)^k}{k!}, \qquad n > 0.$$

Indeed: for $t = 0$ both sides vanish and differentiating both sides (after straighforward manipulations) we see that the derivatives are identical. This proves the claim.

MORE EXPLANATIONS

MORE ON $GAM(\lambda, \nu)$, WITH $\nu \in \mathbb{R}_+$.

7 The laws of large numbers I.: The weak law

7.1 Bernoulli's weak law of large numbers

Let $p \in (0, 1)$ be fixed and denote $q := 1 - p$. Let S_n be a random variable with distribution $BIN(p, n)$:

$$\mathbf{P}(S_n = k) = \binom{n}{k} p^k q^{n-k} =: b_n(k).$$

Since p is kept fixed throughout this, we section suppress notation of dependence on p. S_n can be written as

$$S_n = \xi_1 + \xi_2 + \cdots + \xi_n$$

where ξ_j, $j = 1, 2, \ldots$ are i.i.d. random variables of distribution $\mathbf{P}(\xi_j = 1) = p = 1 - \mathbf{P}(\xi_j = 0)$. The expectation and varianve of S_n is

$$\mathbf{E}(S_n) = np, \qquad \mathbf{Var}(S_n) = npq.$$

The binomial distribution is unimodal. Note that for any integer $k \in [1, n]$

$$\frac{b_n(k)}{b_n(k-1)} = \cdots = \frac{(n-k+1)p}{kq} \tag{25}$$

We conclude that $b_n(k) \geq b_n(k-1)$ if and only ik $k \leq np+p$. Denote by k^* the unique integer in the unit interval $(np-q, np+p]$ and call it the *modus* of the distribution $BIN(p, n)$. Then $k \mapsto b_n(k)$ is monotone increasing in the interval $[0, k^*-1]$ and monotone decreasing in the interval $[k^*, n]$. If $np+p$ is not an integer then $b_n(k^*-1) < b_n(k^*)$. If $np+p$ is integer (degenerate cases) then $b_n(k^*-1) = b_n(k^*)$. k^* (possibly also k^*-1) is the most probable value of the random variable S_n.

We are going to analyse the asymptotic behaviour of the distribution of S_n, as $n \to \infty$. We are going to prove that the asymptotically, the distribution of S_n/n will be concentrated at its expected value p.

Theorem 30 Bernoulli's law of large numbers.
Fix $p \in (0,1)$. For any $\varepsilon > 0$ and $n > \varepsilon^{-1}$

$$\mathbf{P}\Big(\frac{|S_n - np|}{n} > \varepsilon\Big) \leq \frac{1+4\varepsilon}{2n\varepsilon^2} \tag{26}$$

Note that the right hand side converges to 0 as $n \to \infty$.

P r o o f. Let $r \geq np+p$ be an integer. In particular we also have $r \geq k^*$. Applying (25) we find that for any $k \geq 1$

$$\frac{b_n(r+k)}{b_n(r+k-1)} = \frac{(n-r-k+1)p}{(r+k)q} \leq \frac{(n-r)p}{rq} =: y < 1.$$

Hence, for any $l \geq 0$

$$b_n(r+l) = b_n(r) \prod_{k=1}^{l} \frac{b_n(r+k)}{b_n(r+k-1)} \leq b_n(r) y^k,$$

and

$$\mathbf{P}(S_n \geq r) = \sum_{l=0}^{n-r} b_n(r+l) \leq \sum_{l=0}^{\infty} b_n(r) y^k = b_n(r) \frac{1}{1-y} = b_n(r) \frac{rq}{r-np}. \tag{27}$$

Next we give an upper bound on $b_n(r)$:

$$1 = \sum_{k=0}^{n} b_n(k) \geq \sum_{k=k^*}^{r} b_n(k) \geq (r - k^* + 1) b_n(r)$$

Thus

$$b_n(r) \leq \frac{1}{r - np}. \tag{28}$$

Inserting (28) in (27) we find for any integer $r \geq k^*$

$$\mathbf{P}(S_n \geq r) \leq \frac{rq}{(r - np)^2}.$$

Hence

$$\mathbf{P}(S_n - np \geq n\varepsilon) = \mathbf{P}(S_n \geq \lceil np + n\varepsilon \rceil) \leq \frac{\lceil np + n\varepsilon \rceil q}{(\lceil np + n\varepsilon \rceil - np)^2}$$

$$\leq \frac{(np + n\varepsilon + 1)q}{(n\varepsilon)^2} = \frac{pq + q\varepsilon + n^{-1}q}{n\varepsilon^2}. \tag{29}$$

Similarly (changing the rôles of p and q and applying exactly the same arguments to $\widetilde{S}_n := n - S_n$) we prove

$$\mathbf{P}(S_n - np \leq -n\varepsilon) \leq \frac{pq + p\varepsilon + n^{-1}p}{n\varepsilon^2}. \tag{30}$$

Finally, (29) and (30) together yield

$$\mathbf{P}(|S_n - np| \geq n\varepsilon) \leq \frac{2pq + \varepsilon + n^{-1}}{n\varepsilon^2}$$

and, noting that $pq \leq 1/4$, the bound (26) follows for $n \geq \varepsilon^{-1}$.

Remarks.

(1) Note, that actually a much stronger statement can be proved. Let γ_n be an sequence of positive real numbers increasing faster than $\sqrt{n}$:

$$\lim_{n \to \infty} \sqrt{n} \gamma_n^{-1} = 0.$$

Then, replacing $n\varepsilon$ by γ_n in (29) and (30), we get

$$\mathbf{P}\Big(\frac{|S_n - np|}{\gamma_n} > \varepsilon\Big) \leq \frac{2npq + \gamma_n \varepsilon + 1}{(\gamma_n \varepsilon)^2} \to 0. \tag{31}$$

This result suggests that the fluctuations of $S_n - \mathbf{E}(S_n)$ are actually of order $\sqrt{n}$. Precise formulation of this phenomenon will be the subject of the central limit theorem.

(2) From Bernoulli's law of large numbers we easily derive the weak law of large numbers for sums of i.i.d. random variables *with finite range*.

Theorem 31 The weak law of large numbers – finite range.
Let ξ_i, $i = 1, 2, \ldots$ be i.i.d. random variables with the common distribution

$$\mathbf{P}(\xi = x_k) = p_k, \qquad k = 1, 2, \ldots, r, \qquad \sum_{k=1}^{r} p_k = 1.$$

Denote $S_n = \xi_1 + \xi_2 + \cdots + \xi_n$. Then for any $\varepsilon > 0$ and $n > \sum_{k=1}^{r} |x_k|/\varepsilon$

$$\mathbf{P}\Big(\frac{|S_n - nm|}{n} > \varepsilon\Big) \le \frac{r(\sum_{k=1}^{r} |x_k|)^2 + 4\varepsilon \sum_{k=1}^{r} |x_k|}{n\varepsilon^2}. \tag{32}$$

where

$$m := \mathbf{E}(\xi_i) = \sum_{k=1}^{r} x_k p_k.$$

Note that the right hand side converges to 0, as $n \to \infty$.

P r o o f. Denote by $S_n^{(k)}$ the number of occurences of x_k among $\xi_1, \ldots, \xi_n$:

$$S_n^{(k)} := \sum_{i=1}^{n} \mathbb{1}_{\{\xi_i = x_k\}}, \qquad k = 1, 2, \ldots, r.$$

Clearly, $S_n^{(k)}$ has distribution $BIN(p_k, n)$ and

$$S_n = \sum_{k=1}^{r} x_k S_n^{(k)}.$$

Thus

$$|S_n - nm| = \Big|\sum_{k=1}^{r} x_k (S_n^{(k)} - np_k)\Big| \le \big(\sum_{k=1}^{r} |x_k|\big) \max_{1 \le k \le r} |S_n^{(k)} - np_k|.$$

It follows, that

$$|S_n - nm| \le n\varepsilon \quad \text{if} \quad \max_{1 \le k \le r} |S_n^{(k)} - np_k| \le n\varepsilon / \big(\sum_{k=1}^{r} |x_k|\big)$$

and, consequently,

$$\mathbf{P}(|S_n - nm| \ge n\varepsilon) \le \sum_{k=1}^{r} \mathbf{P}\Big(|S_n^{(k)} - np_k| \ge n\varepsilon / \big(\sum_{k=1}^{r} |x_k|\big)\Big).$$

Now, applying Bernoulli's theorem to $S_n^{(k)}$, $k = 1, 2, \ldots, r$, we find that for $n > \sum_{k=1}^{r} |x_k|/\varepsilon$ the bound (32) holds.

An application: Weierstrass' approximation theorem

This is a theorem in pure real analysis. It states that a continuous real function can be uniformly approximated by polynomials on any fixed compact interval.

Theorem 32 Weierstrass' approximation theorem.
Let $f : [0,1] \to \mathbb{R}$ be a continuous function. Then for any $\varepsilon > 0$ there exists $n \in \mathbb{N}$ and a polynomial P_n of degree n such that

$$\sup_{1 \le x \le 1} |f(x) - P_n(x)| < \varepsilon. \tag{33}$$

P r o o f. (S. Bernstein) We shall denote the variable by $p \in [0,1]$ instead of x. Define

$$B_n(p) := \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f(k/n) = \mathbf{E}(f(S_n/n))$$

where S_n is a $BIN(p,n)$-distributed random variable. These are the Bernstein polynomials associated to the function f. We claim, that given $\varepsilon > 0$, for n siffuciently large B_n will satisfy (33).

We shall use two elementary facts about continuous functions defined on compact intervals. Both are direct consequences of the Heine-Borel Theorem.
(1) f is bounded:

$$M := \sup_{0 \le p \le 1} |f(p)| < \infty. \tag{34}$$

(2) f is *uniformly* continuous:

$$(\forall \varepsilon > 0)\ \ (\exists \delta > 0) : |p - p'| < \delta \Rightarrow |f(p) - f(p')| < \varepsilon/2. \tag{35}$$

Now we turn to the proof of the claim:

$$\begin{aligned} |f(p) - B_n(p)| &= \Big| \sum_k \binom{n}{k} p^k (1-p)^{n-k} (f(p) - f(k/n)) \Big| \\ &\le \sum_k \binom{n}{k} p^k (1-p)^{n-k} |f(p) - f(k/n)|. \end{aligned}$$

First choose δ as in (35) and write the right hand side of the last inequality as

$$\sum_{k:|p-k/n|<\delta} \binom{n}{k} p^k (1-p)^{n-k} |f(p) - f(k/n)|$$

$$+ \sum_{k:|p-k/n|\ge\delta} \binom{n}{k} p^k (1-p)^{n-k} |f(p) - f(k/n)|.$$

Due to (35) the first sum is bounded by $\varepsilon/2$. Due to (34) the second sum is bounded by

$$2M \sum_{k:|p-k/n|\geq\delta} \binom{n}{k} p^k (1-p)^{n-k} = 2M\mathbf{P}(|S_n - np| \geq n\delta) \leq 2M \frac{1+4\delta}{2n\delta^2}$$

where in the last step we have used Bernoulli's theorem. Now, choose $n > 4M(1+4\delta)/(2\varepsilon\delta^2)$ to bound this last term by $\varepsilon/2$. Putting these bounds together, the theorem is proved.

7.2 Markov's and Chebyshev's inequalities

Markov's inequality and its immediate consequences are extremely simple and extremely efficient tools.

Markov's inequality gives an efficient upper bound on the probability of (extremely) large values of a positive random variable.

Theorem 33 *Markov's inequality.*
Let ξ be a non-negative *random variable. For any $\lambda > 0$*

$$\mathbf{P}(\xi \geq \lambda) \leq \frac{\mathbf{E}(\xi)}{\lambda}.$$

P r o o f. Indeed,

$$\mathbf{E}(\xi) \geq \mathbf{E}(\xi \mathbb{1}\{\xi \geq \lambda\}) \geq \lambda \mathbf{E}(\mathbb{1}\{\xi \geq \lambda\}) = \lambda \mathbf{P}(\xi \geq \lambda).$$

The same thing, written in terms of the distribution function F of ξ:

$$m_F = \int_0^\infty x dF(x) \geq \int_\lambda^\infty x dF(x) \geq \lambda \int_\lambda^\infty dF(x) = \lambda(1 - F(\lambda)).$$

Note, thet the condition of *positivity of ξ is essential.*

Chebyshev's inequality gives an upper bound on the probability of extremely large fluctuations (i.e. deviations from the mean value) of a random variable.

Corollary 34 Chebyshev's inequality.
Let ξ be an arbitrary random variable. For any $\lambda > 0$

$$\mathbf{P}(|\xi - \mathbf{E}(\xi)| \geq \lambda) \leq \frac{\mathbf{Var}(\xi)}{\lambda^2}.$$

P r o o f. Apply Markov's inequality to the random variable $\eta := (\xi - \mathbf{E}(\xi))^2$.

Corollary 35 Generalized Markov inequality.
Let ξ be an arbitrary random variable and $f : \mathbb{R} \to \mathbb{R}_+$ a monotone non-decreasing function. For any $\lambda \in \mathbb{R}$

$$\mathbf{P}(\xi \geq \lambda) \leq \frac{\mathbf{E}(f(\xi))}{f(\lambda)}.$$

P r o o f. Apply Markov's inequality to the random variable $\eta := f(\xi)$.

Corollary 36 Generalized Chebyshev inequality.
Let ξ be an arbitrary random variable and $f : \mathbb{R}_+ \to \mathbb{R}_+$ a monotone non-decreasing function. For any $\lambda \in \mathbb{R}$

$$\mathbf{P}(|\xi - \mathbf{E}(\xi)| \geq \lambda) \leq \frac{\mathbf{E}(f(|\xi - \mathbf{E}(\xi)|))}{f(\lambda)}.$$

Remark. All these inequalities are, sharp in the sense that ther exist distributions where the inequalities actually become equalities. Find them!

7.3 The weak law: general case

The WLLN tells that the fluctuations of the arithmetic mean of i.i.d random variables are asymptotically negligeable, for large n.

Theorem 37 The weak law of large numbers.
Let $\xi_1, \xi_2, \ldots$ be independent and identically distributed random variables which have finite second moment. Denote $m := \mathbf{E}(\xi_i)$ and $\sigma^2 := \mathbf{Var}(\xi_i)$ and $S_n := \xi_1 + \xi_2 + \cdots + \xi_n$. Thenfor any $\varepsilon > 0$

$$\mathbf{P}\Big(\frac{|S_n - nm|}{n} \geq \varepsilon\Big) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

Note, that the right hand side converges to zero as $n \to \infty$.

P r o o f. Apply Chebyshev's inequality to the random variable S_n:

$$\mathbf{P}\Big(\frac{|S_n - nm|}{n} \geq \varepsilon\Big) = \mathbf{P}(|S_n - nm| \geq n\varepsilon) \leq \frac{\mathbf{Var}(S_n)}{n^2\varepsilon^2} = \frac{n\sigma^2}{n^2\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

Remark. Actually, the same argument proves a much sharper result. Let γ_n be a sequence of positive numbers which increases faster to infinity then $\sqrt{n}$, that is: $\sqrt{n}/\gamma_n \to 0$ as $n \to \infty$. Then we have

$$\mathbf{P}\Big(\frac{|S_n - nm|}{\gamma_n} \geq \varepsilon\Big) = \mathbf{P}(|S_n - nm| \geq \gamma_n \varepsilon) \leq \frac{\mathbf{Var}(S_n)}{\gamma_n^2 \varepsilon^2} = \frac{n\sigma^2}{\gamma_n^2 \varepsilon^2} \to 0$$

as $n \to \infty$. This argument shows that the fluctuations of the sum $S_n = \xi_1 + \xi_2 + \cdots + \xi_n$ are of order $\sqrt{n}$. The precise formulation of this statement is the content of the central limit theorem.

MORE ON WEAK CONVERGENCE!!!!

8 The central limit theorem I.: DeMoivre-Laplace

8.1 Stirling's formula

Theorem 38 Stirling's formula.
For any $n \in \mathbb{N}$

$$1 < \frac{n!}{\sqrt{2\pi}\, n^{n+\frac{1}{2}}\, e^{-n}} < e^{1/(12n)}. \tag{36}$$

Remarks.
(1) In particular it follows that

$$n! = \sqrt{2\pi}\, n^{n+\frac{1}{2}}\, e^{-n}\, (1 + \mathcal{O}(1/n)),$$

as $n \to \infty$. With some extra work the sharper asymptotics

$$n! = \sqrt{2\pi}\, n^{n+\frac{1}{2}}\, e^{-n}\, (1 + 1/(12n) + \mathcal{O}(1/n^2))$$

can be derived.
(2) In the forthcoming proof we shall prove (36) with some unidentified positive constant c_∞ in the place of $\sqrt{2\pi}$. The identification $c_\infty = \sqrt{2\pi}$ will drop out from the proof of DeMoivre's theorem.

P r o o f. Denote

$$d_n := \log n! - \big((n + \frac{1}{2}) \log n - n\big)$$

We show that the sequence d_n is strictly decreasing while the sequence $d_n - \frac{1}{12n}$ is strictly increasing. From these two facts it follows that the sequences

d_n, respectively $d_n - \frac{1}{12n}$, converge downwards, respectively upwards, to the same (finite) limit d_∞. This proves (36) with the constant $c_\infty := \exp(d_\infty)$ in the place of $\sqrt{2\pi}$.

$$d_n - d_{n+1} = \cdots = \left(n + \frac{1}{2}\right) \log \frac{n+1}{n} - 1 = (2n+1)\frac{1}{2} \log \frac{1 + \frac{1}{2n+1}}{1 + \frac{1}{2n+1}} - 1.$$

Let

$$t := \frac{1}{2n+1} < 1$$

and use the Taylor series

$$\log(1+t) = -\sum_{k=1}^{\infty} \frac{1}{k}(-t)^k$$

valid for $|t| < 1$, to get

$$\frac{1}{2t} \log \frac{1+t}{1-t} = \frac{1}{2t} \left(-\sum_{k=1}^{\infty} \frac{1}{k}(-t)^k - \sum_{k=1}^{\infty} \frac{1}{k} t^k \right) = \sum_{k=0}^{\infty} \frac{1}{2k+1} t^{2k}.$$

Thus,

$$d_n - d_{n+1} = \sum_{k=0}^{\infty} \frac{1}{2k+1} \cdot \frac{1}{(2n+1)^{2k}} - 1 = \sum_{k=1}^{\infty} \frac{1}{2k+1} \cdot \frac{1}{(2n+1)^{2k}}.$$

Hence

$$0 < d_n - d_{n+1} < \frac{1}{3} \sum_{k=1}^{\infty} \frac{1}{(2n+1)^{2k}} = \frac{1}{12n} - \frac{1}{12(n+1)}$$

Hence we conclude that the sequence d_n is strictly decreasing, while the sequence $d_n - \frac{1}{12n}$ is strictly increasing. From these two facts (36) directly follows.

8.2 The normal distribution revisited

TO BE COMPLETED

8.3 DeMoivre-Laplace CLT

Before reading this section, read carefully (and understand) Bernoulli's law of large numbers (with comments). We analyze finer asymptotics of the distribution of $BIN(p.n)$, for large n.

For the rest of this section fix $p \in (0,1)$ and denote $q := 1-p$. Let S_n be a random variable with distribution $BIN(p.n)$:

$$\mathbf{P}(S_n = k) = \binom{n}{k} p^k q^{n-k} =: b_n(k).$$

(As p will be kept fixed, we supress the notation of dependence on p.) Note, that

$$S_n = \xi_1 + \xi_2 + \cdots + \xi_n$$

where ξ_i, $i = 1, 2, \ldots, n$ are i.i.d. with distribution $\mathbf{P}(\xi_i = 1) = p = 1 - \mathbf{P}(\xi_i = 0)$, and

$$\mathbf{E}(S_n) = np, \qquad \mathbf{Var}(S_n) = npq.$$

We have seen (see the remark after Bernoulli's law of large numbers) that the fluctuations (i.e. random deviations from the mean value) of S_n are of order $\sqrt{n}$, that is, of order $\sqrt{\mathbf{Var}(S_n)}$. So, it is rather natural to ask about the $n \to \infty$ asymptotics of the distribution of

$$S_n^* := \frac{S_n - \mathbf{E}(S_n)}{\sqrt{\mathbf{Var}(S_n)}} = \frac{S_n - np}{\sqrt{npq}}.$$

Theorem 39 CLT for the binomial distribution — global form.
Let $p \in (0,1)$ be fixed. Then for any fixed interval $[a,b] \subset \mathbb{R}$, with $-\infty \leq a \leq b \leq +\infty$

$$\lim_{n\to\infty} \mathbf{P}(S_n^* \in [a,b]) \to \Phi(b) - \Phi(a), \tag{37}$$

where

$$\Phi(x) = \int_{-\infty}^{x} \varphi(y)dy = \int_{-\infty}^{x} \frac{\exp(-y^2/2)}{\sqrt{2\pi}} dy$$

is the standard normal distribution function.

This theorem will follow from the sharper result due to Abraham De-Moivre.

Theorem 40 DeMoivre's theorem.
Fix $p \in (0,1)$. Let M_n be a sequence increasing to infinity slower than $n^{2/3}$:

$$\lim_{n\to\infty} M_n = \infty, \qquad \lim_{n\to\infty} \frac{M_n}{n^{2/3}} = 0. \tag{38}$$

Then there exists a threshold index N_0 (depending on the sequence M_n), such that for any $n > N_0$ the following bound holds

$$\sup_{k:\,|k-np|<M_n} \left| \frac{\sqrt{npq}\, b_n(k)}{\varphi\big((k-np)/\sqrt{npq}\big)} - 1 \right| \le 4 \max\{\frac{M_n}{npq}, \frac{M_n^3}{(npq)^2}\} \tag{39}$$

Remark. This is the sharpest formulation of DeMoivre's theorem. Typically, we use it for sequences $M_n = M\sqrt{npq}$, with some fixed M. Then we have:

$$\sup_{k:\,|k-np|<M\sqrt{npq}} \left| \frac{\sqrt{npq}\, b_n(k)}{\varphi\big((k-np)/\sqrt{npq}\big)} - 1 \right| \le \frac{4M^3}{\sqrt{npq}}. \tag{40}$$

P r o o f o f t h e CLT.
Let $-\infty < a < b < \infty$ be fixed.

$$\begin{aligned}
\mathbf{P}(S_n^* \in [a,b]) &= \sum_{k=np+a\sqrt{npq}}^{np+b\sqrt{npq}} b_n(k) \\
&= \frac{1}{\sqrt{npq}} \sum_{k=np+a\sqrt{npq}}^{np+b\sqrt{npq}} \varphi\big(\frac{k-np}{\sqrt{npq}}\big) \\
&\quad + \sum_{k=np+a\sqrt{npq}}^{np+b\sqrt{npq}} \left(b_n(k) - \frac{1}{\sqrt{npq}} \varphi\big(\frac{k-np}{\sqrt{npq}}\big) \right) \\
&=: I_n + J_n
\end{aligned}$$

From (40), using the fact that $\sup_x \varphi(x) < 1/\sqrt{2\pi}$, it follows easily that

$$\sup_{k:\,|k-np|<M\sqrt{npq}} \left| b_n(k) - \frac{1}{\sqrt{npq}} \varphi\big(\frac{k-np}{\sqrt{npq}}\big) \right| < \frac{4\sqrt{2\pi}M^3}{npq}.$$

Hence

$$|J_n| \le 4\sqrt{2\pi}\, \max\{|a|^3, |b|^3\}\, (b-a)\, (npq)^{-1/2} \to 0$$

as $n \to \infty$.

On the other hand, I_n is exactly the Riemann approximation of $\int_a^b \varphi(y)dy$. By standard calculus we have:

$$\left| \int_a^b \varphi(y)dy - I_n \right| \le (b-a) \sup_x |\varphi'(x)|\, (npq)^{-1/2} \to 0$$

as $n \to \infty$. So, we got (37) for compact intervals $[a,b]$

TO BE COMPLETED FOR UNBOUNDED $[a,b]$!!

P r o o f o f D e M o i v r e' s t h e o r e m.
Denote $z := k - np$ and assume $|z| < M_n$. In terms of z we have

$$b_n(k) = \frac{n!}{(np+z)!(nq-z)!} p^{np+z} q^{nq-z}$$

Step 1: use Stirling's formula.
We denote by $b'_n(k)$ the expression gotten from this one, by replacing all factorials with Stirling's asymptotics

$$b'_n(k) := \frac{c_\infty n^{n+1/2} e^{-n}\ p^{np+z} q^{nq-z}}{c_\infty (np+z)^{np+z+1/2} e^{-np-z} c_\infty (nq-z)^{nq-z+1/2} e^{-nq+z}}$$

After starightforward manipulations we get:

$$b'_n(k) = \frac{1}{c_\infty \sqrt{npq}} \left(1 + \frac{z}{np}\right)^{-(np+z+1/2)} \left(1 - \frac{z}{nq}\right)^{-(nq-z+1/2)}$$

Using (36) we have

$$\exp(-\frac{1}{12(np+z)}) \exp(-\frac{1}{12(nq-z)}) < \frac{b_n(k)}{b'_n(k)} < \exp(\frac{1}{12n}). \qquad (41)$$

Now, choose N_1 so, that for $n > N_1$

$$\frac{M_n}{npq} < \frac{1}{2}. \qquad (42)$$

(Note that by (38) $M_n/n \to 0$.) The bound (42) also implies

$$\min\{np+z, nq-z\} > npq/2$$

and from (41) we get

$$\exp(-\frac{1}{3npq}) < \frac{b_n(k)}{b'_n(k)} < \exp(\frac{1}{12n}). \qquad (43)$$

Step 2: Get rid of some unpleasent factors.
Next, we define

$$\begin{aligned} b''_n(k) &:= b'_n(k) \sqrt{\left(1 + \frac{z}{np}\right)\left(1 - \frac{z}{nq}\right)} \\ &= \frac{1}{c_\infty \sqrt{npq}} \left(1 + \frac{z}{np}\right)^{-(np+z)} \left(1 - \frac{z}{nq}\right)^{-(nq-z)} \end{aligned}$$

Note that due to (42), for $|z| < M_n$ we have

$$\max\{|z|/np, |z|/nq\} < 1/2. \tag{44}$$

Using

$$\exp(x - x^2) < 1 + x < \exp(x)$$

valid for any $x > -1/2$ we find

$$\exp(\frac{z(p-q)}{2npq} - \frac{z^2(p^2+q^2)}{2(npq)^2}) < \frac{b_n(k)b'}{b''_n(k)} < \exp(\frac{z(p-q)}{2npq}).$$

Using the bound (44), and $z| < M_n$ we have

$$\exp(-\frac{3M_n}{4npq}) < \frac{b'_n(k)}{b''_n(k)} < \exp(\frac{M_n}{2npq}) \tag{45}$$

Remark. Note that in the special case $p = q = 1/2$ these bounds become much more friendly:

$$\exp(-\frac{4M_n^2}{n^2}) < \frac{b'_n(k)}{b''_n(k)} < 1.$$

Step 3: Use Taylor expansion of $\log(1+x)$.
For $|x| < 1/2$ the following bounds hold

$$-|x|^3 < \log(1+x) - x + \frac{x^2}{2} < |x|^3. \tag{46}$$

Using (46), after elementary computations we find

$$\left|\log\left\{\left(1+\frac{z}{np}\right)^{-(np+z)}\left(1-\frac{z}{nq}\right)^{-(nq-z)}\right\} - \frac{z^2}{2npq}\right| < \frac{2|z|^3}{(npq)^2}$$

This, together with $|z| < M_n$ implies

$$\exp(-\frac{2M_n^3}{(npq)^2}) < \frac{c_\infty\sqrt{npq}\,b''_n(k)}{\sqrt{2\pi}\,\varphi(z/\sqrt{npq})} < \exp(\frac{2M_n^3}{(npq)^2}) \tag{47}$$

Putting the bounds (43), (45) and (47) together we find

$$\exp(-\frac{M_n}{npq} - \frac{2M_n^3}{(npq)^2}) < \frac{c_\infty\sqrt{npq}\,b_n(k)}{\sqrt{2\pi}\,\varphi(z/\sqrt{npq})} < \exp(\frac{M_n}{npq} + \frac{2M_n^3}{(npq)^2})$$

Now, choose N_2 so that for $n > n_2$

$$\frac{M_n}{npq} + \frac{2M_n^3}{(npq)^2} < 1$$

and use the fact that for $0 < x < 1$

$$e^x - 1 < 2x$$

to get

$$\left| \frac{c_\infty \sqrt{npq}\, b_n(k)}{\sqrt{2\pi}\, \varphi(z/\sqrt{npq})} - 1 \right| < 4 \max\{\frac{M_n}{npq}, \frac{M_n^3}{(npq)^2}\}. \tag{48}$$

which, modulo $c_\infty = \sqrt{2\pi}$ is identical to (39).

It remains to identify the constant c_∞. Fix $M < \infty$, then from (48) we have (see the argument in the proof of the CLT):

$$\sum_{k=np-M\sqrt{npq}}^{np+M\sqrt{npq}} b_n(k) \to \frac{\sqrt{2\pi}}{c_\infty} \int_{-M}^{M} \varphi(y) dy.$$

On the other hand, the argument used in Bernoulli's law of large numbers (see (31) implies (one could also simply use Chebyshev's inequality here):

$$1 - \frac{2}{M^2} - \frac{1}{M\sqrt{npq}} - \frac{1}{M^2 npq} < \sum_{k=np-M\sqrt{npq}}^{np+M\sqrt{npq}} b_n(k) < 1.$$

Taking the limit $n \to \infty$ these bounds imply

$$1 - \frac{2}{M^2} \leq \frac{\sqrt{2\pi}}{c_\infty} \int_{-M}^{M} \varphi(y) dy \leq 1.$$

Now take the limit $M \to \infty$ to get

$$1 \leq \frac{\sqrt{2\pi}}{c_\infty} \leq 1$$

which proves the claim.

APPLICATIONS: TO BE COMPLETED

8.4 Local CLT for γ distributions

Let ξ_i, $i = 1, 2 \ldots$, be independent and identically distributed random variables with the common exponential distribution

$$\mathbf{P}(\xi_i < x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 - e^{-\lambda x} & \text{if } x > 0. \end{cases}$$

Denote Then, as seen in section (**??**), S_n has gamma distribution with density function

$$f_n(x) = \begin{cases} 0 & \text{if } x \le 0 \\ \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!} & \text{if } x > 0 \end{cases}$$

The expectation and variance of S_n is

$$\mathbf{E}(S_n) = n\lambda^{-1} \qquad \mathbf{Var}(S_n) = n\lambda^{-2}$$

Denote by S_n^* the standardized version of S_n:

$$S_n^* := \frac{S_n - \mathbf{E}(S_n)}{\sqrt{\mathbf{Var}(S_n)}} = \frac{S_n - n\lambda^{-1}}{\sqrt{n}\lambda^{-1}}.$$

The density function of the distribution of S_n^* is

$$f_n^*(x) = \sqrt{n}\lambda^{-1} f_n(n\lambda^{-1} + \sqrt{n}\lambda^{-1}x).$$

Theorem 41 Local central limit theorem for gamma distributions.
For any $x \in \mathbb{R}$ fixed

$$f_n^*(x) \to \varphi(x)$$

as $n \to \infty$.

P r o o f. Indeed,

$$f_n^*(x) = \cdots = \frac{n^{n+\frac{1}{2}}\, e^{-n}}{n!} \left(1 + \frac{x}{\sqrt{n}}\right)^{n-1} e^{-\sqrt{n}x}$$

Applying Stirling's formula and standard expansion we get

$$f_n^*(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \left(1 + \mathcal{O}(1/\sqrt{n})\right).$$

9 Analytic tools I.: The generating function

TO BE COMPLETED

10 Analytic tools II.: The characteristic function

10.1 Definition and primary properties

Definition 42 *The* characteristic function *of the real-valued random variable ξ (or: of the distribution function F) is*

$$\phi : \mathbb{R} \to \mathbb{C}, \qquad \phi(t) := \mathbf{E}(\exp(it\xi)) = \int_{-\infty}^{\infty} \exp(itx) dF(x).$$

I.e., the characteristic function is the Fourier-Stieltjes transform of the distribution function F. If F is absolutely continuous and $f := F'$ is its density function, then

$$\phi(t) = \int_{-\infty}^{\infty} \exp(itx) f(x) dx$$

is the Fourier transform of the density function f.

The primary properties of the characteristic function are formulated in the following theorem:

Theorem 43 Primary properties of the characteristic function
Let ξ be an arbitrary real-valued random variable, F its distribution function and ϕ its characteristic function. Then ϕ has the folloing properties:
(1) Bound:

$$(\forall\, t \in \mathbb{R}) \;\; |\phi(t)| \leq 1, \qquad \textit{and} \qquad \phi(0) = 1.$$

(2) Uniform continuity: $t \mapsto \phi(t)$ is uniformly continuous on $\mathbb{R}$.
(3) Positive type: for any $n \in \mathbb{N}$, $t_1, \ldots, t_n \in \mathbb{R}$, and $z_1, \ldots, z_n \in \mathbb{C}$

$$\sum_{k,l=1}^{n} z_k \bar{z}_l \phi(t_k - t_l) \geq 0. \tag{49}$$

(A function with this last property is called of positive type.*)*

P r o o f.
(1) This is straightforward:

$$|\phi(t)| = |\int_{-\infty}^{\infty} e^{itx} dF(x)| \leq \int_{-\infty}^{\infty} |e^{itx}| dF(x) = \int_{-\infty}^{\infty} dF(x).$$

The inequality is valid, since F is nondecreasing.
(2) Fix $M < \infty$ and use the following bounds valid for any $a, b \in \mathbb{R}$:

$$|e^{ia} - e^{ib}| \leq \min\{2, |a - b|\}.$$

$$|\phi(t)-\phi(s)| \le |\int_{-\infty}^{\infty} |e^{itx}-e^{isx}|dF(x)$$
$$= \int_{\{|x|\le M\}} |e^{itx}-e^{isx}|dF(x) + \int_{\{|x|>M\}} |e^{itx}-e^{isx}|dF(x)$$
$$\le M|t-s| + 2(F(-M)+1-F(M)).$$

We prove that for any $\varepsilon > 0$ we can find a $\delta > 0$ such that if $|t-s| < \delta$ then the right hand side of this inequality is less then ε. First choose M large enough to have $F(-M)+1-F(M) < \varepsilon/4$. Second: choose $\delta < \varepsilon/(2M)$.
(3)

$$\sum_{k,l=1}^{n} z_k \bar{z}_l \phi(t_k - t_l) = \sum_{k,l=1}^{n} z_k \bar{z}_l \mathbf{E}(e^{i(t_k-t_l)\xi})$$
$$= \mathbf{E}\left(\sum_{k,l=1}^{n} z_k \bar{z}_l e^{i(t_k-t_l)\xi}\right) = \mathbf{E}\left|\sum_{k=1}^{n} z_k e^{it_k\xi}\right|^2 \ge 0.$$

Remark. On property (1). A random variable ξ (or its distribution function F) is called *of lattice type* if it takes its values from (or the distribution is fully concentrated on) an arithmetic progression:

$$(\exists d \in \mathbb{R}_+, r \in [0,d)) : \mathbf{P}(\xi \in \{kd+r : k \in \mathbb{Z}\}) = 1.$$

It is easy to see that
(1) If the random variable ξ is *not* of lattice type then for any $t \neq 0$ $|\phi(t)| < 1$.
(2) If the random variable ξ is of lattice type with period d and shift r then

$$\phi(2\pi k/d) = \exp(i2\pi kr/d), \quad k \in \mathbb{Z},$$

and for $t \notin \{2\pi k/d : k \in \mathbb{Z}\}$ $|\phi(t)| < 1$.

The converse of Theorem 43 is the following:

Theorem 44 Bochner's theorem.
Let the function $\phi : \mathbb{R} \to \mathbb{C}$ have the following three properties
(1) $\phi(0) = 1$.
(2) phi is continuous at $t = 0$.
(3) ϕ is of positive type, in the sense of (49).
Then ϕ is the characteristic function of a probability distribution function. I.e., there exists a probability distribution function F such that $\phi(t) = \int_{-\infty}^{\infty} exp(itx)dF(x)$.

Remarks.
(1) This is a beautiful 'structural' theorem identifying the range of Fourier-Stieltjes trasform defined for probability distribution functions. Practically, if you have to decide about a function whether it is the characteristic function of a probability distribution function or not, it is not of much use. Usually: properties (1) and (2) are easy to check but property (3) is typically hopeless.
(2) Note that the properties stated in Bochner's theorem are formally weaker than those in Theorem 43.

Two further straightforward properties of the characteristic function: let ξ be a random variable and $\phi(t)$ its characteristic function. Then

$$\phi(-t) = \overline{\phi(t)}$$
$$\phi_{a\xi+b}(t) := \mathbf{E}(\exp(it(a\xi+b))) = e^{itb}\phi(at), \qquad a, b \in \mathbb{R}.$$

Examples. (Compute them! Use complex integrals, theorem of residues etc.)
Uniform, $U(a,b)$:

$$\phi(t) = \frac{1}{b-a}\int_a^b e^{itx}dx = \frac{e^{itb}-e^{ita}}{it(b-a)}$$

Exponential, $EXP(\lambda)$:

$$\phi(t) = \int_0^\infty e^{itx}\lambda e^{-\lambda x}dx = \frac{\lambda}{\lambda - it}$$

Gamma, $GAM(\lambda,\nu)$:

$$\phi(t) = \int_0^\infty e^{itx}\frac{\lambda^\nu\, x^{\nu-1}e^{-\lambda x}}{\Gamma(\nu)}dx = \left(\frac{\lambda}{\lambda - it}\right)^\nu.$$

Normal, $N(m,\sigma)$:

$$\phi(t) = \int_{-\infty}^\infty e^{itx}\frac{\exp(-(x-m)^2/2\sigma^2)}{\sqrt{2\pi}\sigma}dx = \exp(imt - \frac{\sigma^2t^2}{2})$$

Cauchy, $CAU(m,\tau)$:

$$\phi(t) = \int_{-\infty}^\infty e^{itx}\frac{1}{\pi}\,\frac{\tau}{\tau^2+(x-m)^2}dx = \exp(imt - \tau|t|)$$

10.2 Moments of ξ and derivatives of the characteristic function

Theorem 45 Taylor expansion of the characteristic function.
If for some $k \in \mathbb{N}$ the k-th absolute moment of ξ exists, $\mathbf{E}(|\xi|^k) < \infty$, then the characteristic function ϕ is at least k times continuously differentiable and

$$\phi^{(k)}(0) = i^k \mathbf{E}(\xi^k).$$

P r o o f.

$$\phi^{(k)}(t) = \frac{d^k}{dt^k} \int_{-\infty}^{\infty} e^{itx} dF(x) = \int_{-\infty}^{\infty} \frac{d^k}{dt^k} e^{itx} dF(x) = \int_{-\infty}^{\infty} (ix)^k e^{itx} dF(x).$$

The differentiation under the integral is allowed due to the absolute integrability of the right hand side.

Remark. Warning! For k odd the converese statement is bot true in general. It is possible that due to cancellations $\phi^{(2k+1)}(t)$ exists, but $\mathbf{E}(|\xi|^{2k+1}) = \infty$. For k even, the converse is true.

From the last theorem it follows, that if for some $m \in \mathbb{N}$ $\mathbf{E}(|\xi|^m) < \infty$ then the characteristic function has the following Taylor expansion at $t_0 = 0$:

$$\phi(t) = \sum_{k=0}^{m} \mathbf{E}(\xi^k) \frac{(it)^k}{k!} + o(t^m).$$

Analiticity of the characteristic function.

Theorem 46 Covergent power series expansion of the characteristic function.
If for any $m \in \mathbb{N}$ we have $\mathbf{E}(|\xi|^m) < \infty$ and

$$\limsup_{m \to \infty} \left(\frac{|\mathbf{E}(\xi^m)|}{m!} \right)^{1/m} =: R^{-1} < \infty$$

then $\phi(t)$ extends analitically to the strip

$$\{z \in \mathbb{C} : |Imz| < R\}.$$

In this case the characteristic function and, consequently, the distribution (see Theorem 49 below) is determined by the moments $\mathbf{E}(\xi^m)$.

P r o o f.

Let $t \in \mathbb{R}$ be fixed. Then

$$\left|\phi^{(2k)}(t)\right| = \left|\int_{-\infty}^{\infty} (ix)^{2k} dF(x)\right| \leq \mathbf{E}(\xi^{2k}),$$
$$\left|\phi^{(2k+1)}(t)\right| = \left|\int_{-\infty}^{\infty} (ix)^{2k+1} dF(x)\right| \leq \sqrt{\mathbf{E}(\xi^{2k})\mathbf{E}(\xi^{2k+2})}$$

Hence it follows that for any $t \in \mathbb{R}$

$$\limsup_{m\to\infty} \left(\frac{|\phi^{(m)}(t)|}{m!}\right)^{1/m} \leq R^{-1}$$

and the statement follows.

10.3 Smoothness of the distribution function and decay of the characteristic function at $\pm\infty$

Theorem 47 Decay of the characteristic function at $\pm\infty$.

Let the distribution function be absolutely continuous with density function f. If f is n times differentiable and

$$\int_{-\infty}^{\infty} |f^{(k)}(x)| dx < \infty, \qquad k = 1, 2, \ldots, n \tag{50}$$

then

$$\lim_{|t|\to\infty} |t|^k \, |\phi(t)| = 0.$$

P r o o f.

By integration by parts, induction on $k = 1, 2, \ldots, n$ shows

$$\phi(t) = \left(\frac{i}{t}\right)^k \int_{-\infty}^{\infty} e^{itx} f^{(k)}(x) dx, \qquad k = 1, 2, \ldots, n.$$

In the induction step (50) is used. So we have

$$|t|^n \, |\phi(t)| = \left|\int_{-\infty}^{\infty} e^{itx} f^{(n)}(x) dx\right|.$$

Due to (50) and the Riemann-Lebesgue lemma the right hand side of (50) goes to 0 as $|t| \to \infty$.

10.4 Characteristic function of sum of independent random variables

Let ξ and η be *independent* random variables and $\zeta := \xi + \eta$. Let the characteristic functions of ξ, η and ζ be

$$\phi_\xi(t) := \mathbf{E}(\exp(it\xi)), \quad \phi_\eta(t) := \mathbf{E}(\exp(it\eta)), \quad \phi_\zeta(t) := \mathbf{E}(\exp(it\zeta)).$$

Then

$$\phi_\zeta(t) = \mathbf{E}(\exp(it(\xi+\eta))) = \mathbf{E}(\exp(it\xi)\exp(it\eta))$$

$$= \mathbf{E}(\exp(it\xi))\mathbf{E}(\exp(it\eta)) = \phi_\xi(t)\phi_\eta(t).$$

Written in terms of the distribution functions:

$$\phi_\zeta(t) = \int_{-\infty}^{\infty} e^{itz} d(F*G)(z) = \int_{-\infty}^{\infty} e^{itz} d\{\int_{-\infty}^{\infty} F(z-y)dG(y)\}$$

$$= \cdots = \int_{-\infty}^{\infty} e^{itx} dF(x) \int_{-\infty}^{\infty} e^{ity} dG(y) = \phi_\xi(t)\phi_\eta(t).$$

This is nothing more than the well known fact that Fourier transform of convolutions is equal to the product of the Fourier transforms of the factors.

By induction on the number of summands: if $\xi_1, \xi_2, \ldots, \xi_n$ are independent random variables then, with obvious notations

$$\phi_{\sum_i \xi_i}(t) = \prod_i \phi_{\xi_i}(t).$$

Forecasting the CLT. Let ξ_i, $i = 1, 2, 3 \ldots$, be independent and identically distributed random variables with finite second moment. Denote their expectation and variance by

$$m := \mathbf{E}(\xi_i), \qquad \sigma^2 := \mathbf{Var}(\xi_i).$$

Let $S_n := \xi_1 + \xi_2 + \cdots + \xi_n$. Then

$$\mathbf{E}(S_n) = nm, \qquad \mathbf{Var}(S_n) = n\sigma^2.$$

We have seen (see the comment after the weak law of large numbers) that the typical (normal) fluctuations of S_n about its mean value, are of order $\sqrt{\mathbf{Var}(S_n)}$. So it is very natural to ask about the asymptotics of the distribution of

$$S_n^* := \frac{S_n - \mathbf{E}(S_n)}{\sqrt{\mathbf{Var}(S_n)}} = \frac{S_n - nm}{\sigma\sqrt{n}}.$$

We have seen already a number of examples where the distribution S_n^* converges (in some sense) to the standard normal distribution. Is this a coincidence or something deeper?

Without loss of generality we may assume $m = 0$. Let's compute the characteristic functions. Denote the characteristic function of ξ by ϕ and the characteristic function of S_n^* by ϕ_n^*. Then we have for $t \in \mathbb{R}$ fixed

$$\phi_n^*(t) = \mathbf{E}(\exp(itS_n/\sigma\sqrt{n})) = \mathbf{E}(\exp(\frac{it}{\sigma\sqrt{n}}\sum_{i=1}^{n}\xi_i))$$
$$= \prod_{i=1}^{n}\mathbf{E}(\exp(\frac{it}{\sigma\sqrt{n}}\xi_i)) = \phi(it/\sigma\sqrt{n})^n.$$

As t is kept fixed and $n \to \infty$, we use the Taylor expansion of ϕ around 0:

$$\phi_n^*(t) = \left(1 + 0\frac{it}{\sigma\sqrt{n}} - \frac{\sigma^2}{2}\frac{t^2}{\sigma^2 n} + o(n^{-1})\right)^n \to \exp(-t^2/2).$$

The limit is exactly the characteristic function of the standard normal distribution. So, we can conclude that if one could prove that given a sequence of distribution functions, pointwise convergence of the sequence of their characteristic functions to some limit characteristic function, implies convergence of the distribution functions (in some sense), then the CLT in its most general form would follow. This will be the subject of the next section.

10.5 Reconstruction of the distribution function from the characteristic function

Is the characteristic function *characteristic* indeed? That is: is the map $F \mapsto \phi_F$ injective? Can we reconstruct the distribution function from the characteristic function?

This is the problem of inverting the Fourier transform. Let's consider first the absolutely continuous distributions. If f is a density function and ϕ its Fourier transform,

$$\phi(t) = \int_{-\infty}^{\infty} e^{itx} f(x)dx, \tag{51}$$

then *formally*

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi(t)dt. \tag{52}$$

The question is: when is this formula indeed valid in a strict mathematical sense and what is the truth if this formula is not valid.

There is a first obstacle here: while the integral in (51) is well defined (since f is integrable), the integral in (52) may not be. Indeed, $\phi(t)$ in general is bounded by 1 but not integrable. Theorem 47 at least tells us that if f is twice continuously differentiable and $|f^{(k)}|$, $k = 1, 2$, are integrable then at least the integral in (52) is well defined.

Theorem 48 Fourier inversion.
Let f be a probability density function and ϕ its Fourier transform. If f is continuous and its Fourier transform is absolutely integrable then the inversion formula (52) holds.

P r o o f.
For $\sigma > 0$ let f_σ be the convolution of f with a normal density function of mean 0 and variance σ^2:

$$f_\sigma(x) = \int_{-\infty}^{\infty} f(y) \frac{\exp(-(x-y)^2/(2\sigma^2))}{\sqrt{2\pi}\sigma} dy.$$

The Fourier transform of f_σ is

$$\phi_\sigma(t) = \cdots = \phi(t) e^{\frac{\sigma^2 t^2}{2}}. \tag{53}$$

Let's compute the inverse Fourier transform of ϕ_σ.

$$\begin{aligned}
\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi_\sigma(t) dt &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} e^{-\sigma^2 t^2/2} \phi(t) dt \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} e^{-\sigma^2 t^2/2} \left(\int_{-\infty}^{\infty} e^{ity} f(y) dy \right) dt \\
&= \int_{-\infty}^{\infty} \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-it(x-y)} e^{-\sigma^2 t^2/2} dt \right) f(y) dy \\
&= \int_{-\infty}^{\infty} \frac{\exp(-(x-y)^2/(2\sigma^2))}{\sqrt{2\pi}\sigma} f(y) dy = f_\sigma(x). \qquad (54)
\end{aligned}$$

In the third step we have used Fubini's theorem. So, we can conclude that the inversion formula (52) at least applies to f_σ. Now, we let $\sigma \to 0$. Due to continuity of f, for any fixed $x \in \mathbb{R}$

$$\lim_{\sigma \to 0} f_\sigma(x) = f(x).$$

On the other hand: if ϕ is absolutely integrable then, by Lebesgue's dominated convergence theorem we have

$$\lim_{\sigma\to 0}\int_{-\infty}^{\infty} e^{-itx}e^{-\sigma^2t^2/2}\phi(t)dt = \int_{-\infty}^{\infty} e^{-itx}\phi(t)dt$$

Hence the theorem.

Assuming that the the distribution function F is absolutely continuous with density function f and the conditions of the previous theorem hold, we find

$$\begin{aligned} F(b)-F(a) &= \int_a^b f(x)dx = \int_a^b \left(\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi(t)dt\right) f(x)dx \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\phi(t)\frac{e^{-itb}-e^{-ita}}{-it}dt. \end{aligned} \tag{55}$$

Now, can this formula be extended to the general case? The answer is given in the following theorem

Theorem 49 Reconstruction of the distribution from the characteristic function.
Let F be an arbitrary probability distribution function and ϕ its characteristic function. Let $a < b$ be points of continuity of F. Then the following inversion formulas hold:

$$\begin{aligned} F(b)-F(a) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{1}{2}\left(\phi(t)\frac{e^{-itb}-e^{-ita}}{-it}+\phi(-t)\frac{e^{itb}-e^{ita}}{it}\right) \\ &= \lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\phi(t)\frac{e^{-itb}-e^{-ita}}{-it}dt. \\ &= \lim_{\sigma\to 0}\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\sigma^2t^2/2}\phi(t)\frac{e^{-itb}-e^{-ita}}{-it}. \end{aligned}$$

P r o o f.
We prove the last formula. We adapt the proof of the previous theorem. Denote by F_σ the convolution of F with a normal distribution of mean 0 and variance σ^2:

$$\begin{aligned} F_\sigma(x) &= \int_{-\infty}^{\infty}\frac{\exp(-(x-y)^2/(2\sigma^2))}{\sqrt{2\pi}\sigma}dF(y) \\ &= \int_{-\infty}^{\infty} F(x-y)\frac{\exp(-y^2/(2\sigma^2))}{\sqrt{2\pi}\sigma}dy. \end{aligned}$$

The distribution F_σ is absolutely continuous. Denote its density functionby f_σ. The characteristic function ϕ_sigma of F_σ is the Fourier transform (51) of f_σ. Formula (53) again holds and the arguments form (54) can be repeated to show that the inversion foromula (52) applies to the density function f_σ and its Fourier transform ϕ_sigma. Applying (55) to F_σ we get

$$F_\sigma(b) - F_\sigma(a) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_\sigma(t)\frac{e^{-itb} - e^{-ita}}{-it}dt.$$

We have to prove that

$$\lim_{\sigma\to 0} F_\sigma(x) = F(x)$$

at points of continuity of F. This could be proved purely analitically. We give a probabilistic proof. Let ξ and η_σ be two independent random variables (jointly defined on the same probability space), ξ with distribution $\mathbf{P}(\xi < x) = F(x)$ and η_sigma of distribution $N(0,\sigma)$. Then we have

$$\begin{aligned}|F_\sigma(x) - F(x)| &= |\mathbf{P}(\xi + \eta_\sigma < x) - \mathbf{P}(\xi < x)| \\ &\le \max\{\mathbf{P}(\xi < x,\ \ \xi + \eta_\sigma \ge x), \mathbf{P}(\xi \ge x,\ \ \xi + \eta_\sigma < x)\}\end{aligned}$$

We prove that $\mathbf{P}(\xi \ge x,\ \ \xi + \eta_\sigma < x) \to 0$ as $\sigma \to 0$, the other one is proved identically.Let $\varepsilon > 0$ be fixed. For any $\delta > 0$

$$\begin{aligned}&\mathbf{P}(\xi \ge x,\ \ \xi + \eta_\sigma < x) \\ &\qquad = \mathbf{P}(\xi \in [x, x+\delta),\ \ \xi + \eta_\sigma < x) + \mathbf{P}(\xi \ge x + \delta,\ \ \xi + \eta_\sigma < x) \\ &\qquad\qquad \le \mathbf{P}(\xi \in [x, x+\delta)) + \mathbf{P}(\eta_\sigma < -\delta)\end{aligned}$$

Now, first choose δ sufficiently small to have $\mathbf{P}(\xi \in [x, x+\delta)) < \varepsilon/2$. This can be done since x is point of continuity of F_σ. Next choose σ sufficiently small to have $\mathbf{P}(\eta_\sigma < -\delta) < \varepsilon/2$. The theorem is proved.

11 Weak convergence of distributions and the central limit theorem II.

11.1 Weak convergence of probability measures on metric spaces

We give the general definitions and primary properties of weak convergence of probability measures on metric spaces. Later we shall be concerned with the special case of $\mathbb{R}$.

Let (S, d) be a metric space. Denote $\mathcal{B}$ the Borel algebra generated by the metric topology (i.e. $\mathcal{B}$ is the smallest sigma algebra containing all open balls of (S, d)). We first define the natural notion of convergence of probability measures on $(S, \mathcal{B})$.

Definition 50 *Let μ_n, $n = 1, 2, \ldots$ and μ be probability measures on $(S, \mathcal{B})$. We say that the sequence μ_n* converges weakly *to μ, denoted $\mu_n \Rightarrow \mu$, if for any bounded and continuous function $f : \S \to \mathbb{R}$*

$$\lim_{n\to\infty} \int_S f d\mu_n = \int_S f d\mu.$$

(Lebesgue integrals are meant.)

Remark. Note that limits of weakly convergent sequences of probability measures are unique: if $\mu_n \Rightarrow \nu$ and $\mu_n \Rightarrow \nu'$ then $\mu = \mu'$.

There are a number of equivalent formulations of weak convergence:

Theorem 51 Various formulations of weak convergence of probability measures.
The following formulations are equivalent:
(1) $\mu_n \Rightarrow \mu$, as $n \to \infty$.
(2) For any open *subset $A \subset S$, $\liminf_{n\to\infty} \mu_n(A) \geq \mu(A)$.*
(3) For any closed *subset $A \subset S$, $\limsup_{n\to\infty} \mu_n(A) \leq \mu(A)$.*

We shall be concerned with the special case $S = \mathbb{R}$ (or possibly $S = \mathbb{R}^n$) with the usual (Euclidean) metric. The measures μ_n will be the distributions of random variables Y_n. I.e., $\mu_n(A) := \mathbf{P}(Y_n \in A)$.

As already mentioned the space $S = \mathbb{R}$ (or $S = \mathbb{R}^n$) is very special: due to the natural linear order of $\mathbb{R}$ (or natural partial order of $\mathbb{R}^n$) probability measures are in one-to-one correspondence with the conceptually simpler distribution functions, realised by:

$$F(x) := \mu((-\infty, x)).$$

In terms of distribution functions: weak convergence is essentially pointwise convergence of the sequence of distribution functions.

Theorem 52 Weak convergence of probability distribution functions on $\mathbb{R}$.
Let $(\mu_n)_{n=1}^{\infty}$, ν be probability measures of $(\mathbb{R}, \mathcal{B})$ and $(F_n)_{n=1}^{\infty}$, F their distribution functions. Then the (1) and (2) below are equivalent.

(1) $\mu_n \Rightarrow \mu$, *as* $n \to \infty$.
(2) $\lim_{n\to\infty} F_n(x) = F(x)$ *at points of continuity of* F.

P r o o f.
(1) implies (2): For $x \in \mathbb{R}$ and $\varepsilon > 0$ fixed, define the functions

$$g_{x,\varepsilon} : \mathbb{R} \to [0,1], \qquad g_{x,\varepsilon}(y) := \begin{cases} 1 & \text{if} \quad -\infty < y \le x \\ (x+\varepsilon-y)/\varepsilon & \text{if} \quad x < y \le x+\varepsilon \\ 0 & \text{if} \quad x+\varepsilon < y < \infty \end{cases}$$

Then, for any $x \in \mathbb{R}$, we have

$$\begin{aligned} F_n(x) \quad &\le \textstyle\int_{\mathbb{R}} g_{x,\varepsilon}(y)\, d\mu_n(y) \le \quad F_n(x+\varepsilon), \\ F(x) \quad &\le \textstyle\int_{\mathbb{R}} g_{x,\varepsilon}(y)\, d\mu(y) \le \quad F(x+\varepsilon). \end{aligned}$$

But, since $\mu_n \Rightarrow \mu$ and $g_{x,\varepsilon}$ is bounded and continuous, we have

$$\lim_{n\to\infty} \int_{\mathbb{R}} g_{x,\varepsilon}(y)\, d\mu_n(y) = \int_{\mathbb{R}} g_{x,\varepsilon}(y)\, d\mu(y).$$

These imply that for any $\varepsilon > 0$

$$F(x-\varepsilon) \le \liminf_{n\to\infty} F_n(x) \le \limsup_{n\to\infty} F_n(x) \le F(x+\varepsilon).$$

letting now $\varepsilon \to 0$ yields the result.
(2) implies (1): Let $f : \mathbb{R} \to \mathbb{R}$ be bounded and continuous and $\varepsilon > 0$. Denote $||f||_\infty = \sup_x |f(x)|$. We prove that given the pointwise convergence $F_n(x) \to F(x)$ on points of continuity of F, we can find $N_0 \in \mathbb{N}$ so that for $n > N_0$

$$\left| \int_{\mathbb{R}} f\, d\mu_n - \int_{\mathbb{R}} f\, d\mu \right| < \varepsilon.$$

In the following arguments, for sake of economy, we use the subscript $n = \infty$ for the limit objects $F_\infty = F$, $\mu_\infty = \mu$.
First step. Choose $M < \infty$ so that $\pm M$ are points of continuity of F and for all $n = 1, 2, \ldots, \infty$

$$F_n(-M) + 1 - F_n(M) < \frac{\varepsilon}{6||f||_\infty}. \tag{56}$$

Note that for any single n the left hand side is monotone decreasing with M and goes to zero as $M \to \infty$. We want the bound to hold uniformly for

all $n \leq \infty$. This can be easily done: first choose M' so that $\pm M'$ are points of continuity of F and $F(-M') + 1 - F(M') < \varepsilon/(12||f||_\infty)$ then choose N_1 so that for $n > N_1$, $|F_n(\pm M') - F(\pm M')| < \varepsilon/(24||f||_\infty)$. It follows that for $n > N_1$ (56) will hold, with $\pm M'$ in place of $\pm M$. Finally, choose $M > M'$ so that $\pm M$ be points of continuity of F and (56) hold also for the finitely many indices $n = 1, \ldots, N_1$. This choice of M will imply that for all $n = 1, \ldots, \infty$

$$\left| \int_R f d\mu_n - \int_{[-M,M]} f d\mu_n \right| < \frac{\varepsilon}{6}. \tag{57}$$

Second step. Using the fact that the continuous function f is *uniformly continuous* in the compact interval $[-M, M]$, choose a subdivision $-M = x_0 < x_1 < \cdots < x_{k-1} < x_k = M$ so that $|f(x_j) - f(y)| < \varepsilon/6$ for all $j = 1, 2, \ldots, k$ and $y \in [x_{j-1}, x_j]$. We have for all $n = 1, \ldots, \infty$

$$\left| \int_{[-M,M]} f d\mu_n - \sum_{j=1}^{k} f(x_j)\big(F_n(x_j) - F_n(x_{j-1})\big) \right| < \frac{\varepsilon}{6}.$$

By rearrangement of the sum, this reads

$$\left| \int_{[-M,M]} f d\mu_n - \sum_{j=1}^{k} \big(f(x_{j-1}) - f(x_j)\big) F_n(x_j) \right| < \frac{\varepsilon}{6}. \tag{58}$$

Third step. Choose N_0 so large that for any $n > \mathbb{N}_0$, $|F_n(x_j) - F(x_j)| < 2/k$ for all $j = 1, 2, \ldots, k$. This yields

$$\left| \sum_{j=1}^{k} \big(f(x_{j-1}) - f(x_j)\big)\big(F_n(x_j) - F(x_j)\big) \right| < \frac{\varepsilon}{3}. \tag{59}$$

Putting together (57), (58) and (59) we find

$$\left| \int_{\mathbb{R}} f \, d\mu_n - \int_{\mathbb{R}} f \, d\mu \right| < 2\frac{\varepsilon}{6} + 2\frac{\varepsilon}{6} + \frac{\varepsilon}{3}.$$

Notation. With this theorem in mind, we shall denote by $F_n \Rightarrow F$ pointwise convergence of the sequence F_n *at points of continuity of* F. Furthermore with slight *abuse* we shall sometimes use the notation $\xi_n \Rightarrow \xi$ if the distributions μ_n of the random variables ξ_n converge weakly to the distribution μ of ξ. In this case the terminology will be 'the sequence of random variables ξ_n

converges in distribution to the random variable ξ. We emphasize that this is indeed abuse of terminology, since it is not the sequence of measurable maps $\xi_n : \Omega \to \mathbb{R}$ which converges in some sense, but the sequence of their distributions. So, if ξ_n and ξ are random variables, μ_n and μ their distributions, F_n and F theire distribution functions then the notation $\xi_n \Rightarrow \xi$, $\mu_n \Rightarrow \mu$ and $F_n \Rightarrow F$ will be used interchangeably.

Examples.
(1) Let ξ_n, $n = 1, 2, \ldots$ and ξ be random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$and let μ_n, $n = 1, 2, \ldots$ and μ be their distributions. If $\xi_n \overset{\mathbf{P}}{\to} \xi$ then $\mu_n \Rightarrow \mu$, as $n \to \infty$. However this is not the typical occurence of weak convergence of distributions. But: it is worth noting that if $\mu_n \Rightarrow \mu$ on $\mathbb{R}$ then one can construct a probability space $(\Omega, \mathcal{F}, \mathbf{P})$and jointly defined random variables $\xi_n, \xi : \Omega \to \mathbb{R}$ such that $\xi_n \overset{\mathbf{P}}{\to} \xi$. (Find it!)
(2) In DeMoivre's theorem

$$F_n^*(x) := \mathbf{P}\left(\frac{S_n - np}{\sqrt{npq}} < x\right) \to \Phi(x)$$

where the limit distribution function Φ is the standard normal. This is a typical weak convergence of distributions.

It is important to know that weak convergence is conserved under continuous maps. More generally:

Theorem 53 weak convergence of transformed probability measures.
Let μ_n, $n = 1, 2, \ldots$ and μ be probability measures on the metric space S and assume $\mu_n \Rightarrow \mu$. Let $\widetilde{S}$ be a second metric space and $\psi : S \to \widetilde{S}$ a measurable map. Define the probability measures $\widetilde{\mu}_n$, $n = 1, 2, \ldots$ and $\widetilde{\mu}$ on $\widetilde{S}$ by $\widetilde{\mu}_n(B) := \mu_n(\psi^{-1}B)$ and $\widetilde{\mu}(B) := \mu(\psi^{-1}B)$. Denote by $\mathcal{J}_\psi := \{x \in S : \psi$ is not continuous at $x\}$. If $\mu(\mathcal{J}_\psi) = 0$ then $\widetilde{\mu}_n \Rightarrow \widetilde{\mu}$ on $\widetilde{S}$.

In particular, let $S = \widetilde{S} = \mathbb{R}$ and $\psi : \mathbb{R} \to \mathbb{R}$ continuous function. Given the random variables ξ_n. $n = 1, 2, \ldots$ and ξ define $\eta_n := \psi(\xi_n)$ and $\eta := \psi(\xi)$. If $\xi_n \Rightarrow \xi$ then $\eta_n \Rightarrow \eta$.
Example. The asymptotic distribution of 'peebles of random size'. A simple-minded model for the formation of peebles on the sea-shore is that pieces are formed by a long sequence of random halfing ... . Let ξ_i, $i = 1, 2, \ldots$ be i.i.d. random variables with the common distribution $\mathbf{P}(\xi_i =$

1) $= q = 1 - \mathbf{P}(\xi_i = 1/2)$. Let $P_n := \prod_{i=1}^n \xi_i$ and $P_n^* := 2^{p\sqrt{n}} P_n^{1/\sqrt{n}}$. Then the sequence of random variables P_n^* converges in distribution to a log-normal, $LN(0, \sqrt{pq})$. Apply DeMoivre and the previous theorem.

From the first step of the proof of the second part of Theorem 51 one can see that if a sequence of probability measures μ_n converges weakly to the probability measure μ then

$$(\forall \varepsilon > 0)(\exists M < \infty) \text{ such that } (\forall n): \ \mu_n([-M, M]) > 1 - \varepsilon. \tag{60}$$

Definition 54 *The sequence of probability measures μ_n, $n = 1, 2, \ldots$ on $(\mathbb{R}, \mathcal{B})$ is called* tight *if (60) holds.*

For a general outlook we give the definition of tightness on metric spaces, though it will not be used later in the present notes:

Definition 55 *The sequence of probability measures μ_n, $n = 1, 2, \ldots$ on the metric space $(S, \mathcal{B})$ is called* tight *if*

$$(\forall \varepsilon > 0)(\exists K \subset S, \ K \text{ compact}) \text{ such that } (\forall n): \ \mu_n(K) > 1 - \varepsilon.$$

The crucial fact about tight sequences of probability measures is, that one can extract weakly convergent subsequences from them.

Theorem 56 Prohorov's theorem on $\mathbb{R}$.
Let μ_n, $n = 1, 2, \ldots$ be a tight sequence of probability measures on $\mathbb{R}$. Then there exists a subsequence n_k, $k = 1, 2, \ldots$ and a probability measure μ on $\mathbb{R}$, so that $\mu_{n_k} \Rightarrow \mu$, as $k \to \infty$.

P r o o f.
Let F_n be the distribution function of the measure μ_n. Using the standard 'diagonal extraction' trick, we find a subsequence n_k, $k = 1, 2 \ldots$, such that for any $q \in \mathbb{Q}$ the sequence $F_{n_k}(q)$, $k = 1, 2 \ldots$ is convergent. Denote $\lim_{k\to\infty} F_{n_k}(q) =: L(q)$.
From monotonicity of the functions $x \to F_n(x)$ it follows that

$$q \le q' \quad \text{implies} \quad L(q) \le (q)'. \tag{61}$$

From tightness of the sequence ν_n it follows that

$$\lim_{q\to-\infty} L(q) = 0, \qquad \lim_{q\to\infty} L(q) = 1. \tag{62}$$

Define for $x \in \mathbb{R}$

$$F(x) := \sup\{L(q) : q \in \mathbb{Q} \cap (-\infty, x)\}.$$

By this definition $x \mapsto F(x)$ will be continuous from the left. Furthermore, (61), respectively, (62) imply that F is monotone non-decreasing and $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$. So we conclude that F is a probability distribution function. It remains to be proved that $F_{n_k} \Rightarrow F$, as $k \to \infty$. Let $x - \varepsilon < q < x$ with $q \in \mathbb{Q}$. Then

$$\liminf_{k\to\infty} F_{n_k}(x) \geq \lim_{k\to\infty} F_{n_k}(q) = L(q) \geq F(x - \varepsilon).$$

Similarly, let $x < q < x + \varepsilon$ with $q \in \mathbb{Q}$. Then

$$\limsup_{k\to\infty} F_{n_k}(x) \leq \lim_{k\to\infty} F_{n_k}(q) = L(q) \leq F(x + \varepsilon).$$

Thus, for any $x \in \mathbb{R}$ and $\varepsilon > 0$

$$F(x - \varepsilon) \leq \liminf_{k\to\infty} F_{n_k}(x) \leq \limsup_{k\to\infty} F_{n_k}(x) \leq F(x + \varepsilon)$$

and hence $F_{n_k} \Rightarrow F$. The theorem is proved.

Remarks.
(1) We shall speak about tighness of the sequence of random variables ξ_n, their distributions μ_n and their distribution functions F_n interchangeably.
(2) The same theorem holds for arbitrary metric spaces, but the proof becomes considerably more technical. The theorem formupated for metric spaces is of crucial importance in the theory of weak convergence of stochastic processes.

11.2 Convergence of the characteristic functions and weak convergence of distributions

Let F_n, $n = 1, 2, \ldots$ and F be probability distribution functions and ϕ_n, $n = 1, 2, \ldots$ and ϕ their characteristic functions. From the definition of weak convergence it follows that $F_n \Rightarrow F$ implies pointwise convergence of the characteristic functions. Indeed, for any fixed $t \in \mathbb{R}$, $x \mapsto \exp(itx)$ is continuous and bounded, so that

$$\phi_n(t) = \int_{-\infty}^{\infty} e^{itx} dF_n(x) \to \int_{-\infty}^{\infty} e^{itx} dF(x) = \phi_n(t),$$

as $n \to \infty$. The question is: is the converse statement true? does poinwise convergence of the sequence of characteristic functions imply weak convergence of the sequence of distributions? For the importance of this question see the forecasting of the central limit theorem et the end of section 10.4.

Theorem 57 Pointwise convergence of characteristic functions implies weak convergence of the distributions.
Let F_n, $n = 1, 2, \ldots$ be a sequence of probability distribution functions on $\mathbb{R}$ and ϕ_n their characteristic functions. If for any fixed $t \in \mathbb{R}$ $\phi_n(t) \to \phi(t)$ as $n \to \infty$ and the limit function ϕ is continuous at $t = 0$ then ϕ is the characteristic function of a probability distribution function F and $F_n \Rightarrow F$.

Remark. Note that pointwise convergence of a sequence of (uniformly) continuous functions does not imply continuity of the limit function. The assumption of continuity of ϕ is crucial.

P r o o f.
The main point is that from continuity of ϕ at $t = 0$ tightness of the sequence F_n follows.

Lemma 58 *Let ξ be a random variable, F its distribution function and ϕ its characteristic function. Then for any $M < \infty$ the following inequality holds*

$$\mathbf{P}(|\xi| \geq M) \leq \frac{M}{2} \int_{-2/M} 2/M(1 - \phi(t))dt.$$

P r o o f o f t h e l e m m a.

$$\begin{aligned}
\frac{M}{2} \int_{-2/M}^{2/M} (1 - \phi(t))dt &= \frac{M}{2} \int_{-2/M}^{2/M} \left\{ \int_{-\infty}^{\infty} (1 - e^{itx} dF(x) \right\} dt \\
&= 2 \int_{-\infty}^{\infty} \left\{ 1 - \frac{\sin(2x/M)}{(2x/M)} \right\} dF(x) \\
&\geq 2\{ \int_{-\infty}^{-M} + \int_{M}^{\infty} \} \left\{ 1 - \frac{\sin(2x/M)}{(2x/M)} \right\} dF(x) \\
&\geq 2\{ \int_{-\infty}^{-M} + \int_{M}^{\infty} \} \left\{ 1 - \frac{M}{2|x|} \right\} dF(x) \\
&\geq \{ \int_{-\infty}^{-M} + \int_{M}^{\infty} \} dF(x) \\
&= \mathbf{P}(|X| \geq M) - \mathbf{P}(X = M).
\end{aligned}$$

In the second equality we have used Fubini's theorem. In the first two inequalities $\frac{\sin y}{y} \leq \max\{1, \frac{1}{|y|}\}$ are used.
The lemma follows.

Now back to the proof of the theorem. First we prove tightness of the sequence F_n. Let $\varepsilon > 0$. From continuity at $t = 0$ of the limit function ϕ it follows that we can find $M < \infty$, such that

$$\frac{M}{2} \int_{-2/M}^{2/M} (1 - \phi(t))dt < \frac{\varepsilon}{2}.$$

Since $\phi_n(t) \to \phi(t)$ pointwise, using Lebesgue's dominated convergence theorem we find $N_0 \in \mathbb{N}$ such that for all $n > N_0$

$$\frac{M}{2} \int_{-2/M}^{2/M} (1 - \phi_n(t))dt < \varepsilon.$$

Using the lemma we conclude that the sequence of probability distribution functions F_n is tight.

By applying Prohorov's theorem we conclude that there exists a subsequence n_k, $k = 1, 2, \ldots$ and a probability distribution function F such that $F_{n_k} \Rightarrow F$. Then, due fact that weak convergence of distributions implies pointwise convergence of the characteristic functions (see the beginning of this section) the characteristic function of the distribution F must be the limit function ϕ. Assume now that $F_n \not\Rightarrow F$. Then, aplying Prohorov's theorem once again, we find another subsequence $\widetilde{n}_k$ and a different distribution function $\widetilde{F}'$ such that $F_{\widetilde{n}_k} \Rightarrow \widetilde{F}$, and forcibly $\phi_{\widetilde{n}_k}(t) \to \widetilde{\phi}(t)$, pointwise. But, by assumption of the theorem $\phi_n(t) \to \phi(t)$ pointwise. So, $\widetilde{\phi} = \phi$ and, by the inversion theorem, $\widetilde{F} = F$, which contradicts our starting assumption.
The theorem is proved.

11.3 The central limit theorem in its full generality

Before reading this section read again section 8.3.

Theorem 59 The central limit theorem for sums of i.i.d. random variables. *Let $\xi_1, \xi_2, \ldots$ be i.i.d. random variables with finite second moment, $m := \mathbf{E}(\xi_i)$ and $\sigma^2 := \mathbf{Var}(\xi_i)$. Denote $S_n := \xi_1 + \xi_2 + \cdots + \xi_n$ and*

$$S_n^* := \frac{S_n - \mathbf{E}(S_n)}{\sqrt{\mathbf{Var}(S_n)}} = \frac{S_n - mn}{\sigma\sqrt{n}}.$$

Then, for any $x \in \mathbb{R}$

$$\mathbf{P}(S_n^* < x) \to \Phi(x), \tag{63}$$

as $n \to \infty$, *where* Φ *is the standard normal distribution function. That is: the sequence of distribution of the random variables* S_n^* *converges weakly to the standard normal* $N(0,1)$.

P r o o f.
The argument presented at the end of section 10.4 is completed by reference to the previous theorem.

The condition for the summands ξ_i to be *identically distributed* is not necessary. Actually, the central limit theorem is valid for sums of independent random variables if every summand ξ_i is *small, compared with the sum.* I.e., essentially no one of the summands dominates. The sharpest formulation is:

Theorem 60 Lindeberg's central limit theorem for sums of independent random variables.
Let $\xi_1, \xi_2, \ldots$ *be independent random variables with finite second moment,* $m_i := \mathbf{E}(\xi_i)$ *and* $\sigma_i^2 := \mathbf{Var}(\xi_i)$. *Denote* $S_n := \xi_1 + \xi_2 + \cdots + \xi_n$,

$$B_n^2 := \mathbf{Var}(S_n) = \sum_{i=1}^{n} \sigma_i^2$$

and

$$S_n^* := \frac{S_n - \mathbf{E}(S_n)}{\sqrt{\mathbf{Var}(S_n)}} = \frac{S_n - \sum_{i=1}^n m_i}{B_n}.$$

If for any $\varepsilon > 0$

$$\lim_{n\to\infty} \frac{1}{B_n^2} \max_{1\le k\le n} \int_{-\infty}^{\infty} \mathbb{1}_{\{|x|>\varepsilon B_n\}} x^2 dF_k(x) = 0 \tag{64}$$

Then, for any $x \in \mathbb{R}$

$$\mathbf{P}(S_n^* < x) \to \Phi(x), \tag{65}$$

as $n \to \infty$, *where* Φ *is the dtandard normal distribution function. That is: the sequence of distribution of the random variables* S_n^* *converges weakly to the standard normal* $N(0,1)$.

Remarks.

(1) Lindeberg's condition (64) is written alternatively

$$\lim_{n\to\infty} \frac{1}{B_n^2} \max_{1\le k\le n} \mathbf{E}(\mathbb{1}_{\{|\xi_k|>\varepsilon B_n\}}\xi_k^2) = 0 \tag{66}$$

(2) From condition (64) (or (66)) it follows indeed that the random variables ξ_i are one-by-one small, compared with the sum S_n, in the following sense:

$$\lim_{n\to\infty} B_n = \infty \quad \text{and} \quad \lim_{n\to\infty} \max_{1\le k\le n} \frac{\sigma_k^2}{B_n^2} = 0.$$

(3) This is indeed the sharpest form of the CLT for sums of independent random variables. One can prove that if ξ_i, $i = 1, 2\ldots$ are independent random variables, and (65) holds then (64) (or, equivalently, (66)) also holds.

P r o o f.

TO BE COMPLETED

Once (63) is established the most natural question to ask is the *speed of convergence*. The following theorem gives an upper estimate.

Theorem 61 Cramer-Berry-Essén theorem on the rate of convergence in the CLT.

Let $\xi_1, \xi_2, \ldots$ be i.i.d random variables. Beside the conditions of Theorem 11.3 we also assume finite third absolute moments. There exists a universal constant $C < \infty$, so that for all $n \ge 1$

$$\sup_{-\infty<x<\infty} |\mathbf{P}(S_n^* < x) - \Phi(x)| \le \frac{C\mathbf{E}(|\xi_i|^3)}{\sqrt{n}}. \tag{67}$$

Remark. With stronger assumptions on the existence higher moments of the random variables ξ_i, i.e., the upper bound in (67) can be sharpened.

P r o o f.

TO BE COMPLETED

Theorem 62 Local central limit theorem, uniform convergence of the density functions.

Let the i.i.d. random variables $\xi_1, \xi_2, \ldots$ have absolutely continuous *common distribution with density function f and denote by f_n^* the density function of the rescaled sum S_n^*. (We use the notation of Theorem 11.3.) If the density function f is bounded then*

$$\lim_{n\to\infty} \sup_{-\infty<x<\infty} |f_n^*(x) - \varphi(x)| = 0,$$

where $\varphi(x) = \exp(-x^2/2)/\sqrt{2\pi}$ *is the standard normal density function.*

The proof relies on a finer Fourier analysis. We omit it.

11.4 Symmetric stable distributions and weak convergence to them

TO BE COMPLETED

12 The laws of large numbers II.: The strong law

TO BE COMPLETED