PART A: INDEPENDENCE

Contents

1 Long independent sequences

1a Independent events

1a1 Reminder. *Fair coin:* finite probability space $\Omega = \Omega_n = \{0,1\}^n$ with $p(\omega) = p_n(\omega) = 2^{-n}$ for all $\omega \in \Omega$; the number of "heads" — random variable $H = H_n : \Omega_n \to \mathbb{R}$, $H(\omega) = a_1 + \dots + a_n$ for $\omega = (a_1, \dots, a_n) \in \Omega$; its distribution

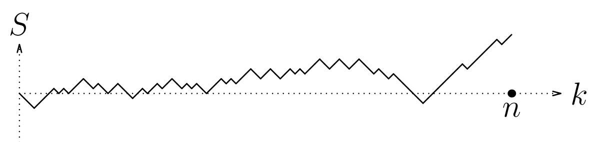
$$\mathbb{P}\big(H = k\big) = P_n\big(\{\omega : H(\omega) = k\}\big) = \sum_{\omega : H(\omega) = k} p(\omega)\,.$$

1a2 Reminder. *Binomial distribution (the fair case):*

$$\mathbb{P}\big(H = k\big) = 2^{-n}\binom{n}{k}\,, \quad \binom{n}{k} = \frac{n!}{k!(n-k)!}\,; \quad H_n \sim \operatorname{Binom}(n, 0.5)\,.$$

1a3 Reminder. *Random signs*: random variables $X_1, \dots, X_n : \Omega_n \to \mathbb{R}$, $X_k(\omega) = 2a_k - 1$ for $\omega = (a_1, \dots, a_n) \in \Omega$.

Simple random walk: random variables $S_0, \dots, S_n : \Omega \to \mathbb{R}$, $S_k = X_1 + \dots + X_k$.



1a4 Remark.

$$S_n = 2H_n - n\,;$$

$$\mathbb{P}\big(S_n = k\big) = \mathbb{P}\Big(H_n = \frac{n+k}{2}\Big) = \frac{1}{2^n}\frac{n!}{\big(\frac{n-k}{2}\big)!\big(\frac{n+k}{2}\big)!} \quad \text{for } k = -n, -n+2, \dots, n\,.$$

1a5 Proposition. [1] For every $\varepsilon > 0$,

$$\mathbb{P}\Big(\tfrac{1}{n}|S_n| \le \varepsilon \Big) \to 1 \quad \text{as } n \to \infty \, .$$

Thus, $\mathbb{P}\big(|\frac{1}{n}H_n - \frac{1}{2}| \le \varepsilon \big) \to 1$ as $n \to \infty$. The frequency is close to the probability...

This is a special case of the Weak Law of Large Numbers, see 1c1. See also 1a24.

How to prove 1a5? Cumbersome sums of binomial coefficients? No, this is the old way. The newer way: via Pythagorean theorem in the (2^n-dimensional) Euclidean space of random variables!

The Euclidean space $L_2(\Omega) = L_2(\Omega_n, P_n)$ consists of all functions $X : \Omega \to \mathbb{R}$ and is endowed with the norm and scalar product

$$\|X\| = \sqrt{\sum_{\omega\in\Omega} |X(\omega)|^2 p(\omega)} = \sqrt{\langle X, X \rangle} \, , \qquad (1a6)$$
$$\langle X, Y \rangle = \sum_{\omega\in\Omega} X(\omega) Y(\omega) p(\omega) \, .$$

Its dimension is equal to the number of points in Ω (think, why; any restriction on $p(\cdot)$?).

Recall the *expectation* $\mathbb{E}\, X$ of a random variable X:

$$\mathbb{E}\, X = \sum_{\omega\in\Omega} X(\omega) p(\omega) = \langle X, 1\!\!1 \rangle \, , \qquad (1a7)$$

an important linear functional on $L_2(\Omega)$. Note that

$$\|X\|^2 = \mathbb{E}\, X^2 \, , \qquad (\text{that is, } \mathbb{E}\,(X^2)) \qquad (1a8)$$
$$\langle X, Y \rangle = \mathbb{E}\, XY \, . \qquad (\text{that is, } \mathbb{E}\,(XY))$$

1a9 Exercise. The random signs $X_1, \dots, X_n$ are orthonormal, that is,

$$\mathbb{E}\, X_i X_j = \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

Prove it. Are they a basis?

1a10 Exercise. $\|S_k\| = \sqrt{k}$ for $k = 0, 1, \dots, n$.

Prove it.

---

[1] [KS, Sect. 2.1, Th. 2.5].

1a11 Exercise. $\mathbb{P}\big(|X| \ge \varepsilon\big) \le \big(\frac{1}{\varepsilon}\|X\|\big)^2$ for all $\varepsilon > 0$ and $X \in L_2(\Omega)$.
Prove it.

1a12 Exercise. Prove Proposition 1a5.

Here is the normal approximation to the binomial distribution.

1a13 Proposition. [1]

$$\mathbb{P}\big(S_n = k\big) = \frac{2}{\sqrt{2\pi n}} \exp\Big(-\frac{k^2}{2n}\Big) \cdot \bigg(1 + \alpha_n\Big(\frac{k}{\sqrt{n}}\Big)\bigg) \quad \text{for } k+n \text{ even}\,,$$

where $\alpha_n(\cdot) \to 0$ uniformly on bounded intervals.

(Clearly, $\mathbb{P}\big(S_n = k\big)$ vanishes if $n + k$ is odd.)

1a14 Remark. The convergence $\alpha_n(\cdot) \to 0$ cannot be uniform on $\mathbb{R}$, since $\alpha_n\big(\sqrt{n} + \frac{1}{\sqrt{n}}\big) = -1$ (think, why). What about $\alpha_n(\sqrt{n})$? Well, it is $-1 + \frac{1}{2}\sqrt{2\pi n}\,\mathrm{e}^{n/2}2^{-n} \to -1$ (think, why).

How to prove Prop. 1a13? Some calculations with binomial coefficients (but not their sums...) are needed.

1a15 Reminder.

$$n! = n^n \mathrm{e}^{-n}\sqrt{2\pi n}\,\beta(n)\,, \quad \beta(n) \to 1\,; \quad \beta(n) = 1 + O(1/n)\,. \quad \text{(Stirling)}$$

Thus,

$$\ln n! = n \ln n - n + \tfrac{1}{2}\ln n + \tfrac{1}{2}\ln(2\pi) + \underbrace{\ln\beta(n)}_{O(1/n)}\,. \tag{1a16}$$

1a17 Exercise.

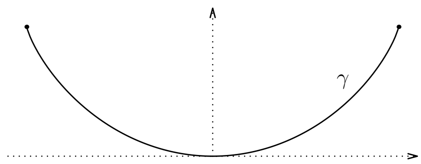
$$\ln \mathbb{P}\big(S_n = cn\big) = -\frac{1}{2}\ln n - \frac{n}{2}\big((1-c)\ln(1-c) + (1+c)\ln(1+c)\big) -$$
$$-\frac{1}{2}\ln(1-c^2) + \ln 2 - \frac{1}{2}\ln(2\pi) + \ln\beta(n) - \ln\beta\Big(\frac{n(1-c)}{2}\Big) - \ln\beta\Big(\frac{n(1+c)}{2}\Big)$$

whenever $cn \in \{-n, -n+2, \dots, n\}$.
Prove it. (Combine 1a4 with 1a15 and enjoy many cancellations!)

[1] [KS, Sect. 2.2, Th. 2.10]; [D, Sect. 2.1, Th. (14)].

We introduce a function $\gamma : [-1,1] \to \mathbb{R}$ by

$$\gamma(c) = \frac12 (1+c)\ln(1+c) + \frac12 (1-c)\ln(1-c) \quad \text{for } c \in (-1,1)\,, \tag{1a18}$$
$$\gamma(-1) = \gamma(+1) = \ln 2\,.$$



Now 1a17 becomes

$$\ln \mathbb{P}\big(S_n = cn \big) = -n\gamma(c) - \frac12 \ln n - \frac12 \ln(1-c^2) + \ln 2 - \frac12 \ln(2\pi) + o(1)\,,$$

if $n(1\pm c) \gg 1$; moreover,

$$\mathbb{P}\big(S_n = cn \big) = \frac{2}{\sqrt n} \frac{1}{\sqrt{1-c^2}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-n\gamma(c)} \cdot \bigg(1 + O\Big(\frac{1}{n(1-|c|)} \Big) \bigg)\,. \tag{1a19}$$

A numerical example: $n = 200$, $c = -0.9$, $\mathbb{P}\big(S_{200} = -180\big)$ =? Really $1.397 \cdot 10^{-44}$; approximately (as above) $1.409 \cdot 10^{-44}$; by normal approximation: $3.7 \cdot 10^{-37}$ (oops...).

About the function γ:

$$\gamma(-c) = \gamma(c)\,; \quad \gamma'(0) = 0;$$
$$\gamma''(0) = 1 \quad \text{since} \quad (x \ln x)' = 1 + \ln x\,, \ \ (x\ln x)'' = \frac1x\,;$$
$$\gamma'''(0) = 0\,; \quad \text{thus} \quad \gamma(c) = \frac12 c^2 + O(c^4) \ \text{ as } c \to 0\,.$$

Proof of 1a13

$$\bigg| \ln \mathbb{P}\big(S_n = k \big) - \ln \Big(\frac{2}{\sqrt{2\pi n}} \exp\Big(-\frac{k^2}{2n} \Big) \Big) \bigg| \le \qquad (k = cn)$$
$$\le \bigg| \ln \mathbb{P}\big(S_n = cn \big) - \ln \Big(\frac{2}{\sqrt n} \frac{1}{\sqrt{1-c^2}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-n\gamma(c)} \Big) \bigg| +$$
$$+ \bigg| \ln \Big(\frac{2}{\sqrt n} \frac{1}{\sqrt{1-c^2}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-n\gamma(c)} \Big) - \ln \Big(\frac{2}{\sqrt{2\pi n}} \mathrm{e}^{-nc^2/2} \Big) \bigg| \le$$
$$\le O\Big(\frac{1}{n(1-|c|)} \Big) + \underbrace{\ln \frac{1}{\sqrt{1-c^2}}}_{O(c^2)} + \underbrace{|n\gamma(c) - nc^2/2|}_{O(nc^4)}\,.$$

Claim:

$$O\Big(\frac{1}{n(1-|c|)}\Big) + O(c^2) + O(nc^4) = O\Big(nc^4 + \frac{1}{n}\Big) \, .$$

Proof of the claim. If $nc^4 + \frac1n \le \delta \le \frac14$ then: $nc^4 \le \delta$; $\frac1n \le \delta$; $c^4 \le \frac1n \delta \le \delta^2$; $c^2 \le \delta$; $|c| \le \frac12$; $\frac{1}{n(1-|c|)} \le \frac2n \le 2\delta$.

Thus,

$$\mathbb{P}\big(S_n = k \big) = \frac{2}{\sqrt{2\pi n}} \mathrm{e}^{-\frac{k^2}{2n}} \cdot \Big(1 + \underbrace{O\Big(\frac{k^4}{n^3} + \frac1n \Big)}_{\alpha_n(k/\sqrt n)} \Big) \, .$$

If $|k| = O(\sqrt n)$ then $\frac{k^4}{n^3} = O\big(\frac{n^2}{n^3}\big) = O(\frac1n)$. Thus, $\sup_{[a,b]} |\alpha(\cdot)| = O(1/n)$ for all a,b.

End of proof of 1a13

Proposition 1a13 is a special case of the Local Limit Theorem. In contrast, the next result is global.

1a20 Theorem. [1]

$$\mathbb{P}\big(a\sqrt n < S_n < b\sqrt n \big) \to \frac{1}{\sqrt{2\pi}} \int_a^b \mathrm{e}^{-x^2/2} \, \mathrm{d}x \quad \text{as } n \to \infty$$

whenever $-\infty \le a \le b \le \infty$.

This is the De Moivre-Laplace theorem, a special case of the Central Limit Theorem.

Proof of 1a20

First, assume that $-\infty < a < b < \infty$. Then:

$$\mathbb{P}\big(a\sqrt n < S_n < b\sqrt n \big) = \sum_{k \in (a\sqrt n, b\sqrt n),\, k+n \text{ even}} \mathbb{P}\big(S_n = k \big) \, ;$$

$$\text{let} \quad \varphi(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2} \, ;$$

$$\mathbb{P}\big(S_n = k \big) = \frac{2}{\sqrt n} \varphi\Big(\frac{k}{\sqrt n} \Big) \Big(1 + \alpha_n\Big(\frac{k}{\sqrt n} \Big) \Big) \, ;$$

[1] [KS, Sect. 2.2]; [D, Sect. 2.1, (1.5)].

$$\frac{2}{\sqrt{n}}\varphi\Big(\frac{k}{\sqrt{n}}\Big)\Big(1+\inf_{[a,b]}\alpha_n\Big) \le \mathbb{P}\big(S_n = k\big) \le \frac{2}{\sqrt{n}}\varphi\Big(\frac{k}{\sqrt{n}}\Big)\Big(1+\sup_{[a,b]}\alpha_n\Big)\,;$$

$$\underbrace{\Big(1+\inf_{[a,b]}\alpha_n\Big)}_{\to 1}\sum_{k\in(a\sqrt{n},b\sqrt{n}),\,k+n\text{ even}}\frac{2}{\sqrt{n}}\varphi\Big(\frac{k}{\sqrt{n}}\Big) \le \mathbb{P}\big(a\sqrt{n} < S_n < b\sqrt{n}\big) \le$$

$$\le \underbrace{\Big(1+\sup_{[a,b]}\alpha_n\Big)}_{\to 1}\sum_{k\in(a\sqrt{n},b\sqrt{n}),\,k+n\text{ even}}\frac{2}{\sqrt{n}}\varphi\Big(\frac{k}{\sqrt{n}}\Big)\,.$$

It remains to prove that the sum converges to the integral. We divide $[a, b]$ into intervals of length $2/\sqrt{n}$ and get an integral sum; only the first and last terms differ, but contribute only $O(1/\sqrt{n})$ anyway.

The case $-\infty < a < b < \infty$ is done. The case $a = b$ is trivial. It is sufficient (think, why) to consider the case $-\infty < a < b = \infty$. We could do it via the equality $\int_{-\infty}^{\infty} \mathrm{e}^{-x^2/2}\,\mathrm{d}x = \sqrt{2\pi}$ well-known in analysis, but it is instructive to do it differently, and get the integral equality as a by-product.[1] (The argument introduced below will be reused in the proof of 1a21.)

We note that

$$\liminf_{n\to\infty} \mathbb{P}\big(a\sqrt{n} < S_n\big) \ge \frac{1}{\sqrt{2\pi}}\int_a^\infty \mathrm{e}^{-x^2/2}\,\mathrm{d}x\,,$$

since it exceeds $\liminf_{n\to\infty}\mathbb{P}\big(a\sqrt{n} < S_n < b\sqrt{n}\big) = \frac{1}{\sqrt{2\pi}}\int_a^b \mathrm{e}^{-x^2/2}\,\mathrm{d}x$ for every $b \in (a,\infty)$. It remains to prove that

$$\limsup_{n\to\infty} \mathbb{P}\big(a\sqrt{n} < S_n\big) \le \frac{1}{\sqrt{2\pi}}\int_a^\infty \mathrm{e}^{-x^2/2}\,\mathrm{d}x\,.$$

We have

$$\mathbb{P}\big(a\sqrt{n} < S_n\big) = \mathbb{P}\big(S_n = k_n\big) + \mathbb{P}\big(S_n = k_n + 2\big) + \dots$$

where $k_n = \min(\mathbb{Z} \cap (a\sqrt{n},\infty))$; and

$$\mathbb{P}\big(S_n = k+2\big) = 2^{-n}\frac{n!}{\big(\frac{n-k-2}{2}\big)!\big(\frac{n+k+2}{2}\big)!} = \frac{n-k}{n+k+2}\mathbb{P}\big(S_n = k\big) \le$$

$$\le \frac{1-\frac{k}{n}}{1+\frac{k}{n}}\mathbb{P}\big(S_n = k\big) \le \frac{1-a/\sqrt{n}}{1+a/\sqrt{n}}\mathbb{P}\big(S_n = k\big)$$

[1] The doubt is: maybe a part $1 - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathrm{e}^{-x^2/2}\,\mathrm{d}x > 0$ of the distribution of S_n escapes to infinity when $n \to \infty$?

for $k = k_n, k_n + 2, \dots$; therefore

$$\frac{\mathbb{P}\big(S_n > a\sqrt{n}\big)}{\mathbb{P}\big(S_n = k_n\big)} \le 1 + \frac{1 - a/\sqrt{n}}{1 + a/\sqrt{n}} + \Big(\frac{1 - a/\sqrt{n}}{1 + a/\sqrt{n}}\Big)^2 + \dots = \frac{1 + a/\sqrt{n}}{2a/\sqrt{n}} = \frac{\sqrt{n}}{2a} + \frac{1}{2}\,,$$

and we get

$$\limsup_{n\to\infty} \mathbb{P}\big(a\sqrt{n} < S_n\big) \le \limsup_{n\to\infty} \frac{\sqrt{n}}{2a}\mathbb{P}\big(S_n = k_n\big) = \frac{1}{a}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-a^2/2}$$

by 1a13, since $k_n/\sqrt{n} \to a$. It follows that

$$\limsup_{n\to\infty} \mathbb{P}\big(a\sqrt{n} < S_n\big) \le \limsup_{n\to\infty} \mathbb{P}\big(a\sqrt{n} < S_n < b\sqrt{n}\big) + \limsup_{n\to\infty} \mathbb{P}\big(b\sqrt{n} \le S_n\big) \le$$
$$\le \frac{1}{\sqrt{2\pi}} \int_a^b \mathrm{e}^{-x^2/2}\,\mathrm{d}x + \frac{1}{b}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-b^2/2}$$

for every $b \in (a, \infty)$; we take $b \to \infty$.

End of proof of 1a20

1a21 Proposition. [1] For every $c \in (0, 1)$,

$$\frac{1}{n} \ln \mathbb{P}\big(S_n > cn\big) \to -\gamma(c) \quad \text{as } n \to \infty\,,$$

where $\gamma(c) = \frac{1}{2}(1+c)\ln(1+c) + \frac{1}{2}(1-c)\ln(1-c)$.

This is a special case of the Large Deviations Principle.[2]

Proposition 1a21 suggests the approximation (for large c and n)

$$\mathbb{P}\big(S_n > c\big) \approx \mathrm{e}^{-n\gamma(c/n)} = \frac{n^n}{\sqrt{(n-c)^{n-c}(n+c)^{n+c}}}\,.$$

However, Theorem 1a20 suggests another approximation,

$$\mathbb{P}\big(S_n > c\big) \approx \frac{1}{\sqrt{2\pi}} \int_{c/\sqrt{n}}^{\infty} \mathrm{e}^{-x^2/2}\,\mathrm{d}x \approx \exp\Big(-\frac{c^2}{2n}\Big)\,.$$

A paradox! What do you think? A clue: for $n = 200$,

c	0	30	60	90	120	150	180
$2^{-n}\binom{n}{(n+c)/2}$	$6\cdot10^{-2}$	$6\cdot10^{-3}$	$6\cdot10^{-6}$	$5\cdot10^{-11}$	$1\cdot10^{-18}$	$3\cdot10^{-29}$	$1\cdot10^{-44}$
$\frac{2}{\sqrt{2\pi n}}\exp(-\frac{c^2}{2n})$	$6\cdot10^{-2}$	$6\cdot10^{-3}$	$7\cdot10^{-6}$	$9\cdot10^{-11}$	$1\cdot10^{-17}$	$2\cdot10^{-26}$	$4\cdot10^{-37}$
$\exp(-n\gamma(\frac{c}{n}))$	1	$1\cdot10^{-1}$	$1\cdot10^{-4}$	$8\cdot10^{-10}$	$2\cdot10^{-17}$	$3\cdot10^{-28}$	$1\cdot10^{-43}$

[1] [D, Sect. 2.1, Exercise 1.3].
[2] [KS, Sect. 10.4]; [D, Sect. 1.9].

Proof of 1a21

We reuse the argument of the last part of the proof of 1a20: $\mathbb{P}\big(S_n > cn\big) = \mathbb{P}\big(S_n = k_n\big) + \mathbb{P}\big(S_n = k_n + 2\big) + \dots$ where $k_n = \min(\mathbb{Z} \cap (cn, \infty))$;

$$\mathbb{P}\big(S_n = k + 2\big) \le \frac{1-c}{1+c} \mathbb{P}\big(S_n = k\big)$$

for $k = k_n, k_n + 2, \dots$; therefore

$$1 \le \frac{\mathbb{P}\big(S_n > cn\big)}{\mathbb{P}\big(S_n = k_n\big)} \le 1 + \frac{1-c}{1+c} + \Big(\frac{1-c}{1+c}\Big)^2 + \dots = \frac{1+c}{2c} \,,$$

and we get $\frac{1}{n} \ln \mathbb{P}\big(S_n > cn\big) - \frac{1}{n} \ln \mathbb{P}\big(S_n = k_n\big) \to 0$. By (1a19),

$$\frac{1}{n} \ln \mathbb{P}\big(S_n = k_n\big) = \frac{1}{n} \ln \bigg(\frac{2}{\sqrt{n}} \frac{1}{\sqrt{1-c^2}} \frac{1}{\sqrt{2\pi}} \bigg) - \gamma\Big(\frac{k_n}{n}\Big) + \frac{1}{n} O\Big(\frac{1}{n(1-|c|)}\Big)\bigg) = \\ = -\gamma\Big(\frac{k_n}{n}\Big) + o(1) \to -\gamma(c) \,,$$

since $\frac{k_n}{n} \to c$.

End of proof of 1a21

1a22 Reminder. *Unfair coin:* The same Ω and H as in 1a1, but different probabilities

$$p_n(a_1, \dots, a_n) = p^{a_1 + \dots + a_n} (1-p)^{n - (a_1 + \dots + a_n)} = \prod_{k=1}^n p_1(a_k) \quad \text{for } a_1, \dots, a_n \in \{0, 1\} \,.$$

It is convenient to write $H = H_{n,p}$; this function on Ω does not depend on p, but its distribution depends on p.

1a23 Reminder. *Binomial distribution:*

$$\mathbb{P}\big(H_{n,p} = k\big) = \binom{n}{k} p^k (1-p)^{n-k} \,; \quad H_n \sim \operatorname{Binom}(n, p) \,.$$

1a24 Proposition. [1] For every $p \in [0, 1]$ and $\varepsilon > 0$,

$$\mathbb{P}\big(|\tfrac{1}{n} H_{n,p} - p| \le \varepsilon \big) \to 1 \quad \text{as } n \to \infty \,.$$

[1][KS, Sect. 2.1, Th. 2.5].

This is more general than (the corollary of) 1a5, but still, a special case of the Weak Law of Large Numbers, see 1c1.

It is easy to prove 1a24 similarly to 1a5, but anyway, 1a24 will follow from 1c1.

Rather unexpectedly, 1a24 can be used for proving Weierstrass's approximation theorem: polynomials are dense in $C[0,1]$.[1]

Here is the idea of the probabilistic proof of Weierstrass's approximation theorem. Consider the distribution $\mu_{n,p}$ of $\frac1n H_{n,p}$,

$$\mu_{n,p} = \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} \delta_{k/n} \, ;$$

it belongs to the $(n+1)$-dimensional linear space of signed measures on $\{0, \frac1n, \frac2n, \dots, 1\}$, and the vector-function $p \mapsto \mu_{n,p}$ is polynomial (of degree n). By 1a24, $\mu_{n,p}$ is close to δ_p (the unit mass at p). Thus, the map $p \mapsto \delta_p$ is approximately polynomial! Now, given a continuous function $f : [0,1] \to \mathbb{R}$ we have $f(p) = \int f \, \mathrm{d}\delta_p \approx \int f \, \mathrm{d}\mu_{n,p} = P_n(p)$, P_n being a polynomial. Namely,

$$|f(p) - \mathbb{E}\, f(\tfrac1n H_{n,p})| \le \mathbb{E}\, |f(\tfrac1n H_{n,p}) - f(p)| \le \\ \le \max_{[p-\varepsilon,p+\varepsilon]} |f(\cdot) - f(p)| + \mathbb{P}\big(|\tfrac1n H_{n,p} - p| > \varepsilon \big) \cdot 2 \max_{[0,1]} |f(\cdot)| \, ;$$

the former summand is made small using uniform continuity of f, the latter summand — using $\mathbb{E}\, |\frac1n H_{n,p} - p|^2 = \frac{p(1-p)}{n} \le \frac{1}{4n}$.

1a25 Proposition. [2] For every $\lambda \in (0,\infty)$ and $k = 0,1,2,\dots$

$$\mathbb{P}\big(H_{n,\lambda/n} = k \big) \to \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda} \quad \text{as } n \to \infty \, .$$

This is the Poisson Limit Theorem.

Proof.

$$\mathbb{P}\big(H_{n,\lambda/n} = k \big) = \frac{n!}{k!(n-k)!} \Big(\frac{\lambda}{n} \Big)^k \Big(1 - \frac{\lambda}{n} \Big)^{n-k} = \\ = \underbrace{\frac{n}{n} \frac{n-1}{n} \dots \frac{n-k+1}{n}}_{\to 1} \frac{\lambda^k}{k!} \underbrace{\Big(1 - \frac{\lambda}{n} \Big)^n}_{\to \mathrm{e}^{-\lambda}} \underbrace{\Big(1 - \frac{\lambda}{n} \Big)^{-k}}_{\to 1} \to \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda} \, .$$

□

[1] [KS, Sect. 2.1, Th. 2.7]; [D, Sect. 1.5, Example 5.1].

[2] [KS, Sect. 2.3]; [D, Sect. 2.6(a)].

1b Measure-theoretic foundations

Some measure theory

1b1 Reminder. An *algebra* of sets (on X): a set $\mathcal{E}$ of subsets of X such that

$$\emptyset, X \in \mathcal{E}\,; \quad \forall E \in \mathcal{E}\;\; X \setminus E \in \mathcal{E}\,; \quad \forall E, F \in \mathcal{E}\;\; E \cap F, E \cup F \in \mathcal{E}\,.$$

A σ-*algebra* of sets (on X): an algebra $\mathcal{A}$ such that

$$\forall A_1, A_2, \dots \in \mathcal{A}\;\; \bigcap_n A_n, \bigcup_n A_n \in \mathcal{A}\,.$$

A *measurable space:* $(X, \mathcal{A})$.

A *probability measure* (on $\mathcal{A}$, or on $(X, \mathcal{A})$): a map $\mu : \mathcal{A} \to [0,1]$ such that $\mu(X) = 1$ and

$$\mu\Big(\bigcup_n A_n\Big) = \sum_n \mu(A_n)$$

whenever $A_1, A_2, \dots \in \mathcal{A}$ are pairwise disjoint. (In such case we may write $\biguplus_n A_n$.)

A *probability space:* $(X, \mathcal{A}, \mu)$.

1b2 Reminder. A *box* (in $\mathbb{R}^d$): a set of the form $I_1 \times \dots \times I_d$ where $I_1, \dots, I_d \subset \mathbb{R}$ are bounded intervals (open, closed, or neither).

An *elementary set* (in $\mathbb{R}^d$): a finite union of boxes.

The *elementary algebra* (on $\mathbb{R}^d$): the algebra generated by all boxes; consists of all elementary sets and their complements ("co-elementary sets").

The *Borel σ-algebra* $\mathcal{B}(\mathbb{R}^d)$: the σ-algebra generated by all boxes, or equivalently, by all open sets.

1b3 Theorem. [1] (Hahn-Kolmogorov) Let $\mathcal{E}$ be an algebra on X, $\mathcal{A} = \sigma(\mathcal{E})$ the σ-algebra generated by $\mathcal{E}$, and $\mu_0 : \mathcal{E} \to [0,1]$ a map. Then the following are equivalent:

(a) there exists one and only one probability measure μ on $\mathcal{A}$ such that $\mu|_{\mathcal{E}} = \mu_0$;

(b) there exists at least one such μ;

(c) $\mu_0(X) = 1$, and $\mu_0\big(\cup_n E_n\big) = \sum_n \mu_0(E_n)$ whenever $E_1, E_2, \dots \in \mathcal{E}$ are pairwise disjoint and $\cup_n E_n \in \mathcal{E}$.

[1] [Tao, Th. 1.7.8].

Clearly, (a)$\Longrightarrow$(b)$\Longrightarrow$(c). In order to prove[1] (c)$\Longrightarrow$(a) we assume (c); define the *outer measure*

$$\mu^*(Z) = \inf \Big\{ \sum_n \mu_0(E_n) : E_1, E_2, \dots \in \mathcal{E}, \cup_n E_n \supset Z \Big\}$$

for arbitrary $Z \subset X$; and call a set $A \subset X$ μ-measurable if

$$\inf_{E \in \mathcal{E}} \mu^*(A \triangle E) = 0 \,.$$

1b4 Exercise. $\mu^*(\cup_n Z_n) \le \sum_n \mu^*(Z_n)$ for arbitrary $Z_1, Z_2, \dots \subset X$.
Prove it. (Do you need (c)?)

1b5 Exercise. $\mu^*(E) = \mu_0(E)$ for all $E \in \mathcal{E}$.
Prove it. (Do you need (c)?)

1b6 Exercise. μ-measurable sets are a σ-algebra.
Prove it.

Taking into account that all sets of $\mathcal{E}$ are μ-measurable we conclude that all sets of $\mathcal{A} = \sigma(\mathcal{E})$ are μ-measurable.
We define μ as the restriction of μ^* to $\mathcal{A}$.

1b7 Exercise. $|\mu^*(Z) - \mu^*(W)| \le \mu^*(Z \triangle W)$ for arbitrary $Z, W \subset X$.
Prove it.

1b8 Exercise. $\mu^*(A \cup B) + \mu^*(A \cap B) = \mu^*(A) + \mu^*(B)$ for all μ-measurable A, B.
Prove it.
Hint: $\mu_0(E \cup F) + \mu_0(E \cap F) = \mu_0(E) + \mu_0(F)$ for all $E, F \in \mathcal{E}$.

Thus, $\mu(A \uplus B) = \mu(A) + \mu(B)$.

1b9 Exercise. $\mu^*\big(\uplus_n A_n\big) = \sum_n \mu^*(A_n)$ for μ-measurable A_n.
Prove it.
Hint: $\mu^*\big(\uplus_{n=1}^\infty A_n\big) - \mu^*\big(\uplus_{n=1}^N A_n\big) \le \mu^*\big(\uplus_{n=N+1}^\infty A_n\big)$.

Thus, μ is a probability measure, which completes the proof of existence.
Here is uniqueness. Let μ_1 be another such measure. Then $\mu_1(A) \le \mu^*(A) = \mu(A)$ for all $A \in \mathcal{A}$ (since $A \subset \cup_n E_n$ implies $\mu_1(A) \le \sum_n \mu_1(E_n) = \sum_n \mu(E_n)$). The same holds for $X \setminus A$, thus, $\mu_1(A) = 1 - \mu_1(X \setminus A) \ge 1 - \mu(X \setminus A) = \mu(A)$ and finally $\mu_1(A) = \mu(A)$.
Theorem 1b3 is proved.

[1]Following Terry Tao, "An alternate approach to the Carathéodory extension theorem" (blog) and Jun Tanaka & Peter F. McLoughlin, "A Realization of Measurable Sets as Limit Points", The American Mathematical Monthly **117**:3, 261–266 (also arXiv:0712.2270).

1b10 Corollary. If two probability measures are equal on an algebra $\mathcal{E}$ then they are equal on the generated σ-algebra $\sigma(\mathcal{E})$.

1b11 Corollary. A probability measure on $\big(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\big)$ is uniquely determined by its values on boxes. The same holds for closed boxes, and for open boxes.

1b12 Exercise. A probability measure μ on $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$ is uniquely determined by its *cumulative distribution function* (CDF)

$$F_\mu(x) = \mu\big((-\infty, x]\big) \quad \text{for } x \in \mathbb{R}\,.$$

Prove it.

1b13 Exercise. (*Lebesgue-Stieltjes measure*) Let $F : \mathbb{R} \to [0,1]$ be an increasing function, $F(-\infty) = 0$, $F(+\infty) = 1$. Then

(a) there exists one and only one additive function μ_0 on the elementary algebra $\mathcal{E}$ (of $\mathbb{R}$) such that

$$\begin{aligned} \mu_0\big((a,b)\big) &= F(b-) - F(a+) \quad \text{for } -\infty < a < b < \infty\,, \\ \mu_0\big(\{a\}\big) &= F(a+) - F(a-) \quad \text{for } -\infty < a < \infty\,; \end{aligned}$$

(b) for every $E \in \mathcal{E}$ and $\varepsilon > 0$ there exists a compact elementary set $K \subset E$ such that $\mu_0(K) \ge \mu_0(E) - \varepsilon$;

(c) μ_0 satisfies Condition 1b3(c).

Prove it.

1b14 Exercise. The correspondence $\mu \longleftrightarrow F_\mu$ is a bijective correspondence between all probability measures on $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$ and all increasing functions $F : \mathbb{R} \to [0,1]$ such that $F(-\infty) = 0$, $F(+\infty) = 1$ and $\forall x \in \mathbb{R}\;\; F(x+) = F(x)$.

Prove it.

In fact, 1b12 generalizes readily to $\mathbb{R}^d$,

$$F_\mu(x_1, \dots, x_d) = \mu\big((-\infty, x_1] \times \dots \times (-\infty, x_d]\big)\,;$$

unfortunately, 1b14 does not.

Discrete measures:

$$\mu = \sum_k p_k \delta_{x_k}\,; \quad \mu(B) = \sum_{k : x_k \in B} p_k = \sum_{x \in B} p(x)\,; \quad \int f \, \mathrm{d}\mu = \sum f(x_k) p_k\,.$$

Absolutely continuous measures:

$$\mu = \int p(x)\delta_x \, \mathrm{d}x \, ; \quad \mu(B) = \int_B p(x) \, \mathrm{d}x \, ; \quad \int f \, \mathrm{d}\mu = \int f(x)p(x) \, \mathrm{d}x \, .$$

Singular measures: nonatomic but concentrated on a set of zero Lebesgue measure.

The *product* of two measurable spaces $(X_1, \mathcal{A}_1)$ and $(X_2, \mathcal{A}_2)$ is defined to be $(X_1 \times X_2, \mathcal{A}_1 \times \mathcal{A}_2)$ where $\mathcal{A}_1 \times \mathcal{A}_2$ is the σ-algebra generated by $A_1 \times A_2$ for all $A_1 \in \mathcal{A}_1$, $A_2 \in \mathcal{A}_2$.

1b15 Exercise. (a) A probability measure μ on $(X_1 \times X_2, \mathcal{A}_1 \times \mathcal{A}_2)$ is uniquely determined by $\mu(A_1 \times A_2)$ for all $A_1 \in \mathcal{A}_1$, $A_2 \in \mathcal{A}_2$.

(b) The same holds for $\mu(E_1 \times E_2)$ for all $E_1 \in \mathcal{E}_1$, $E_2 \in \mathcal{E}_2$ provided that an algebra $\mathcal{E}_1$ generates $\mathcal{A}_1$, and an algebra $\mathcal{E}_2$ generates $\mathcal{A}_2$.

Prove it.

Hint: (b) the σ-algebra generated by $E_1 \times E_2$ contains $E_1 \times A_2$.

In particular, $\mathcal{B}(\mathbb{R}^{d_1}) \times \mathcal{B}(\mathbb{R}^{d_2}) = \mathcal{B}(\mathbb{R}^{d_1+d_2})$.

1b16 Reminder. Let $(X_1, \mathcal{A}_1, \mu_1)$, $(X_2, \mathcal{A}_2, \mu_2)$ be two probability spaces. The formula

$$\mu(A) = \int_{X_1} \biggl(\int_{X_2} \mathbb{1}_A(x_1, x_2) \, \mu_2(\mathrm{d}x_2) \biggr) \mu_1(\mathrm{d}x_1)$$

defines correctly a measure μ on $(X_1 \times X_2, \mathcal{A}_1 \times \mathcal{A}_2)$. Clearly, μ satisfies

(1b17) $$\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2) \quad \text{for } A_1 \in \mathcal{A}_1, \, A_2 \in \mathcal{A}_2 \, .$$

By 1b15, μ is the only measure satisfying (1b17). It follows that $\int_{X_1} \int_{X_2} \dots = \int_{X_2} \int_{X_1} \dots$. We write $\mu = \mu_1 \times \mu_2$, say that μ is the product measure, and $(X_1 \times X_2, \mathcal{A}_1 \times \mathcal{A}_2, \mu_1 \times \mu_2) = (X_1, \mathcal{A}_1, \mu_1) \times (X_2, \mathcal{A}_2, \mu_2)$ is the product of probability spaces.

1b18 Reminder. Let $(X, \mathcal{A}, \mu) = (X_1, \mathcal{A}_1, \mu_1) \times (X_2, \mathcal{A}_2, \mu_2)$.

(a) (Tonelli)

$$\int_X f \, \mathrm{d}\mu = \int_{X_1} \mu_1(\mathrm{d}x_1) \int_{X_2} \mu_2(\mathrm{d}x_2) f(x_1, x_2) \in [0, +\infty]$$

for every measurable $f : X \to [0, \infty]$.

(b) (Fubini)

$$\int_X f \, \mathrm{d}\mu = \int_{X_1} \mu_1(\mathrm{d}x_1) \int_{X_2} \mu_2(\mathrm{d}x_2) f(x_1, x_2) \in \mathbb{R}$$

for every integrable $f : X \to \mathbb{R}$.

In particular,

$$\int_X f_1(x_1)f_2(x_2)\,\mu(\mathrm{d}x_1\mathrm{d}x_2) = \Big(\int_{X_1} f_1\,\mathrm{d}\mu_1 \Big)\Big(\int_{X_2} f_2\,\mathrm{d}\mu_2 \Big)\,. \tag{1b19}$$

The same holds for the product of three, four, ... probability spaces.

1b20 Remark. *Associativity of the multiplication:* the space $(X_1,\mathcal{A}_1,\mu_1) \times (X_2,\mathcal{A}_2,\mu_2) \times (X_3,\mathcal{A}_3,\mu_3) \times (X_4,\mathcal{A}_4,\mu_4)$ is the same as $\big((X_1,\mathcal{A}_1,\mu_1) \times (X_2,\mathcal{A}_2,\mu_2)\big) \times \big((X_3,\mathcal{A}_3,\mu_3) \times (X_4,\mathcal{A}_4,\mu_4)\big)$. That is, $(\mu_1\times\mu_2)\times(\mu_3\times\mu_4) = \mu_1\times\mu_2\times\mu_3\times\mu_4$, which follows from the uniqueness; both measures satisfy $\mu(A_1\times A_2\times A_3\times A_4) = \mu_1(A_1)\mu_2(A_2)\mu_3(A_3)\mu_4(A_4)$ (think, why). The same holds for more than two factors in a group, and more than two groups.

1b21 Reminder. A *measurable map* from a measurable space $(X,\mathcal{A})$ to a measurable space $(Y,\mathcal{B})$ is $f : X \to Y$ such that $\forall B\in\mathcal{B}\;\; f^{-1}(B)\in\mathcal{A}$.

The composition of measurable maps, $(X,\mathcal{A}) \to (Y,\mathcal{B}) \to (Z,\mathcal{C})$, is again a measurable map.

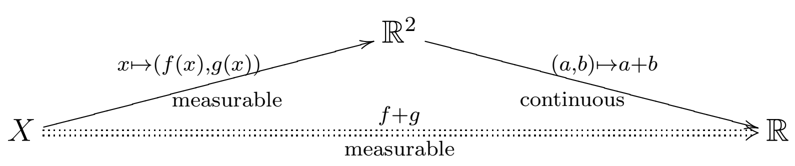
It is sufficient to check the condition $f^{-1}(B)\in\mathcal{A}$ for all B of a set that generates $\mathcal{B}$.

When $Y=\mathbb{R}^d$, the Borel σ-algebra $\mathcal{B}=\mathcal{B}(\mathbb{R}^d)$ is meant by default. Thus, a map $f : X\to\mathbb{R}^d$ is measurable iff $f^{-1}(B)\in\mathcal{A}$ for every box $B\subset\mathbb{R}^d$, or equivalently, for every open set $B\subset\mathbb{R}^d$. A real-valued function $f : X\to\mathbb{R}$ is measurable iff $\{x : f(x)\le a\}\in\mathcal{A}$ for all $a\in\mathbb{R}$.

If $(Y,\mathcal{B}) = (Y_1,\mathcal{B}_1)\times(Y_2,\mathcal{B}_2)$ then a map $f : X\to Y$ boils down to $f_1 : X\to Y_1$ and $f_2 : X\to Y_2$; $f(x) = \big(f_1(x),f_2(x)\big)\in Y_1\times Y_2$. In this case f is measurable iff f_1,f_2 are measurable (think, why). In particular: $Y_1=Y_2=\mathbb{R}$, $Y=\mathbb{R}^2$.

Every continuous map $\mathbb{R}^{d_1}\to\mathbb{R}^{d_2}$ is measurable. (The Borel σ-algebras are meant!)

If $f,g : (X,\mathcal{A})\to\mathbb{R}$ are measurable then $f+g$ is measurable. Here is a short and general proof:



The same holds for $f,g : (X,\mathcal{A})\to\mathbb{R}^d$.

If $f,f_1,f_2,\dots : \mathbb{R}^{d_1}\to\mathbb{R}^{d_2}$, $f_n(\cdot)\to f(\cdot)$ pointwise, and f_n are measurable then f is measurable. Also $\sup_n f_n(\cdot)$, $\limsup_n f_n(\cdot)$ etc.

1b22 Reminder. Given a measurable map $\varphi : (X, \mathcal{A}) \to (Y, \mathcal{B})$ and a probability measure μ on $(X, \mathcal{A})$, the formula

$$\nu(B) = \mu\big(\varphi^{-1}(B)\big) \quad \text{for } B \in \mathcal{B}$$

defines a probability measure ν on $(Y, \mathcal{B})$, — the induced measure. We have

$$\int_X f \circ \varphi \, \mathrm{d}\mu = \int_Y f \, \mathrm{d}\nu \in [0, \infty]$$

for every measurable $f : (Y, \mathcal{B}) \to [0, \infty]$. For $f : (Y, \mathcal{B}) \to \mathbb{R}$ consider $f = f^+ - f^-$.

The σ-algebra $\sigma(\varphi) = \{\varphi^{-1}(B) : B \in \mathcal{B}\}$ generated by $\varphi : X \to (Y, \mathcal{B})$ is the least σ-algebra (on X) that makes φ measurable. The σ-algebra $\sigma(\varphi_1, \varphi_2)$ generated by $\varphi_1 : X \to (Y_1, \mathcal{B}_1)$ and $\varphi_2 : X \to (Y_2, \mathcal{B}_2)$ is, by definition, the least σ-algebra that makes φ_1, φ_2 measurable. It is the same as $\sigma\big(\sigma(\varphi_1) \cup \sigma(\varphi_2)\big)$.

About convergence theorems (monotone, dominated) I give no reminder; I just assume that you never forget them!

Random variables

By a *random variable* we mean a measurable function on a given probability space $(\Omega, \mathcal{F}, P)$. (By default all random variables — on a single probability space.) Usually it maps Ω to $\mathbb{R}$, but can also map Ω to a given measurable space; then it may be called a *random element* of that space. A random element of $\mathbb{R}^d$ (endowed with the Borel σ-algebra) may be called a d-dimensional random vector, or just a d-dimensional random variable, — basically the same as d one-dimensional random variables (the coordinates).

Subsets of Ω belonging to $\mathcal{F}$ are called *events;* $P(A)$ is called the *probability* of an event A.

Random variables generate σ-algebras (of events): $\sigma(X)$, $\sigma(X, Y)$ etc.

By the *distribution* of a random variable $X : \Omega \to \mathbb{R}$ we mean the induced measure P_X on $\mathbb{R}$,

$$P_X(B) = P\big(X^{-1}(B)\big) = P\big(\{\omega : X(\omega) \in B\}\big) = \mathbb{P}\big(X \in B\big)$$

for Borel sets $B \subset \mathbb{R}$.

Random variables $X, Y : \Omega \to \mathbb{R}$ are called *identically distributed* if $P_X = P_Y$; that is, $\mathbb{P}\big(X \in B\big) = \mathbb{P}\big(Y \in B\big)$ for every Borel $B \subset \mathbb{R}$, or equivalently, every interval $B \subset \mathbb{R}$; still equivalently, if $F_X = F_Y$, where F_X is the CDF,

$$F_X(x) = \mathbb{P}\big(X \le x\big) \, .$$

For random elements, $P_X = P_Y$ still applies (in contrast to intervals and CDF). For random vectors CDF applies, and boxes may be used instead of the intervals.

A 2-dim random vector $\omega \mapsto \big(X(\omega), Y(\omega)\big)$ has a 2-dim distribution $P_{X,Y}$, called also the *joint distribution* of X and Y. Usually $P_{X,Y}$ is far from being uniquely determined by P_X, P_Y. Two such vectors (X, Y) and (U, V) are identically distributed iff $P_{X,Y} = P_{U,V}$. Then $P_X = P_U$ and $P_Y = P_V$ (but the converse fails).

If $\varphi : \mathbb{R} \to \mathbb{R}$ is a Borel function and $X : \Omega \to \mathbb{R}$ a random variable then their composition $\varphi(X) : \Omega \to \mathbb{R}$ is another random variable. Likewise, if $\varphi : \mathbb{R}^2 \to \mathbb{R}$ is a Borel function and $X, Y : \Omega \to \mathbb{R}$ random variables then $\varphi(X, Y) : \Omega \to \mathbb{R}$ is a random variable.

The *expectation* $\mathbb{E}\, X$ of a random variable $X : \Omega \to \mathbb{R}$ is, by definition, the Lebesgue integral

$$\mathbb{E}\, X = \int_\Omega X \,\mathrm{d}P$$

provided that X is integrable, that is, $\mathbb{E}\,|X| < \infty$. (Otherwise $\mathbb{E}\, X = \mathbb{E}\, X^+ - \mathbb{E}\, X^-$; the four cases...)

By 1b22,

$$\mathbb{E}\, X = \int_\Omega X \,\mathrm{d}P = \int_\mathbb{R} x \, P_X(\mathrm{d}x) \,;$$

$$\mathbb{E}\, \varphi(X) = \int_\Omega \varphi(X) \,\mathrm{d}P = \int_\mathbb{R} \varphi \,\mathrm{d}P_X = \int_\mathbb{R} z \, P_{\varphi(X)}(\mathrm{d}z) \,;$$

$$\mathbb{E}\, \varphi(X, Y) = \int_\Omega \varphi(X, Y) \,\mathrm{d}P = \int_{\mathbb{R}^2} \varphi \,\mathrm{d}P_{X,Y} = \int_\mathbb{R} z \, P_{\varphi(X,Y)}(\mathrm{d}z)$$

etc.

If X, Y are identically distributed then $\mathbb{E}\, X = \mathbb{E}\, Y$.

Using Tonelli's theorem on $\Omega \times \mathbb{R}$ (or alternatively, approximation) we get

$$\mathbb{E}\, X = \int_0^\infty \mathbb{P}\big(X > a\big) \,\mathrm{d}a - \int_0^\infty \mathbb{P}\big(X < -a\big) \,\mathrm{d}a$$

(if integrable... four cases...)

Some examples of random variables with distributions of different kind (Ω is $(0, 1)$ with Lebesgue measure):

$$X\Big(\sum\nolimits_1^\infty 2^{-k}\beta_k\Big) = \sum\nolimits_1^{10} 2^{-k}\beta_k \qquad \text{— discrete;}$$

$$X\Big(\sum\nolimits_1^\infty 2^{-k}\beta_k\Big) = \sum\nolimits_1^\infty 2^{-k}\beta_{2k} \qquad \text{— absolutely continuous;}$$

$$X\Big(\sum\nolimits_1^\infty 2^{-k}\beta_k\Big) = \sum\nolimits_1^\infty 2^{-2k}\beta_k \qquad \text{— singular.}$$

Independence

Random variables $X, Y : \Omega \to \mathbb{R}$ are called *independent* if

$$P_{X,Y} = P_X \times P_Y \,;$$

that is, $\mathbb{P}\big(X \in A, Y \in B\big) = \mathbb{P}\big(X \in A\big)\mathbb{P}\big(Y \in B\big)$ for all Borel sets $A, B \subset \mathbb{R}$, or equivalently, all intervals $A, B \subset \mathbb{R}$; still equivalently, if $F_{X,Y}(x, y) = F_X(x)F_Y(y)$ for all $X, Y \in \mathbb{R}$. For random elements, $P_{X,Y} = P_X \times P_Y$ still applies (in contrast to intervals and CDF). For random vectors CDF applies, and boxes may be used instead of the intervals. If X, Y are independent then $f(X), g(Y)$ are independent, for arbitrary Borel functions f, g. Two 2-dim random vectors (X, Y) and (U, V) are independent iff $P_{X,Y,U,V} = P_{X,Y} \times P_{U,V}$, that is, $\mathbb{P}\big((X, Y) \in A, (U, V) \in B\big) = \mathbb{P}\big((X, Y) \in A\big)\mathbb{P}\big((U, V) \in B\big)$ for $A, B \subset \mathbb{R}^2$ (Borel sets, or only boxes). Then, $f(X, Y)$ and $g(U, V)$ are independent for all Borel $f, g : \mathbb{R}^2 \to \mathbb{R}$. In particular, X and U are independent; also X and V; Y and U; Y and V. (But the converse fails.)

Random variables $X_1, \dots, X_n$ are called independent, if

$$P_{X_1,\dots,X_n} = P_{X_1} \times \dots \times P_{X_n} \,;$$

that is, $\mathbb{P}\big(X_1 \in B_1, \dots, X_n \in B_n\big) = \mathbb{P}\big(X_1 \in B_1\big) \dots \mathbb{P}\big(X_n \in B_n\big)$. (This is much stronger than the pairwise independence. A counterexample: random signs conditioned by $X_1 \dots X_n = +1$.)

Events $A_1, \dots, A_n$ are called independent, if their indicators $\mathbb{1}_{A_1}, \dots, \mathbb{1}_{A_n}$ are independent random variables. For $n = 2$ this boils down to $\mathbb{P}\big(A_1, A_2\big) = \mathbb{P}\big(A_1\big)\mathbb{P}\big(A_2\big)$, but for $n > 2$ it does not.

One says that σ-algebras $\mathcal{A}_1, \dots, \mathcal{A}_n \subset \mathcal{F}$ are independent if

$$\forall A_1 \in \mathcal{A}_1 \, \dots \, \forall A_n \in \mathcal{A}_n \ \big(A_1, \dots, A_n \text{ are independent events}\big) \,.$$

Random variables $X_1, \dots, X_n$ are independent iff their σ-algebras $\sigma(X_1), \dots, \sigma(X_n)$ are independent. The same holds for events (the σ-algebra generated by an event A being just $\{\emptyset, A, \Omega \setminus A, \Omega\}$).

If random variables X_1, X_2, X_3, X_4 are independent then random vectors (X_1, X_2) and (X_3, X_4) are independent, since

$$\begin{aligned} P_{(X_1,X_2),(X_3,X_4)} = P_{X_1,X_2,X_3,X_4} = P_{X_1} \times P_{X_2} \times P_{X_3} \times P_{X_4} = \\ = (P_{X_1} \times P_{X_2}) \times (P_{X_3} \times P_{X_4}) = P_{X_1,X_2} \times P_{X_3,X_4} \,. \end{aligned}$$

Thus, $f(X_1, X_2)$ and $g(X_3, X_4)$ are independent for all Borel $f, g : \mathbb{R}^2 \to \mathbb{R}$. (Pairwise independence of X_1, X_2, X_3, X_4 is not sufficient! The same counterexample: $(X_1 X_2)(X_3, X_4) = +1$.) The same holds for more than two factors in a group, and more than two groups.

If X, Y are independent then

$$\mathbb{E}\, f(X,Y) = \int_{\mathbb{R}^2} f \,\mathrm{d}(P_X \times P_Y) = \iint_{\mathbb{R}^2} f(x,y) P_X(\mathrm{d}x) P_Y(\mathrm{d}y) \,;$$
$$\mathbb{E}\, f(X) g(Y) = \Big(\int f \,\mathrm{d}P_X \Big) \Big(\int g \,\mathrm{d}P_Y \Big) = \big(\mathbb{E}\, f(X) \big) \big(\mathbb{E}\, g(Y) \big) \,;$$
$$\mathbb{E}\,(XY) = (\mathbb{E}\, X)(\mathbb{E}\, Y) \,.$$

The same holds for more than two random variables.

Independence for discrete:

$$p_{X_1,\dots,X_n}(x_1,\dots,x_n) = p_{X_1}(x_1) \dots p_{X_n}(x_n) \,;$$
$$\mathbb{E}\, f(X_1,\dots,X_n) = \sum_{x_1,\dots,x_n} f(x_1,\dots,x_n) p_{X_1}(x_1) \dots p_{X_n}(x_n) \,.$$

Independence for absolutely continuous:

$$p_{X_1,\dots,X_n}(x_1,\dots,x_n) = p_{X_1}(x_1) \dots p_{X_n}(x_n) \,;$$
$$\mathbb{E}\, f(X_1,\dots,X_n) = \int \dots \int f(x_1,\dots,x_n) p_{X_1}(x_1) \dots p_{X_n}(x_n) \,\mathrm{d}x_1 \dots \mathrm{d}x_n .$$

1c Independent random variables

Let $X_1, X_2, \dots$ be independent identically distributed random variables. Their sums $S_k = X_1 + \dots + X_k$ are a (one-dimensional) random walk.

1c1 Theorem. [1] If $\mathbb{E}\,|X_1| < \infty$ then

$$\mathbb{P}\big(|\tfrac{1}{n} S_n - \mathbb{E}\, X_1| \le \varepsilon \big) \to 1 \quad \text{as } n \to \infty \,.$$

This is the Weak Law of Large Numbers.

Clearly, 1c1 implies 1a5 and 1a24.

Interestingly, 1c1 helps to integrate numerically functions of many (say, 20 or 200) variables ("Monte-Carlo method").[2]

[1] [KS, Sect. 7.1, Th. 7.2]; [D, Sect. 1.5, Corollary (5.8)].
[2] [KS, Sect. 3.8]; [D, Sect. 1.5, Exercise 5.3].

Proof of 1c1

Two main ideas: orthogonality (as for 1a5) and approximation by L_2; L_2 is dense in L_1, but we need also independence...

Let Borel functions $\varphi_1, \varphi_2, \dots : \mathbb{R} \to \mathbb{R}$ satisfy $\forall x \; \varphi_i(x) \to x$ as $i \to \infty$, and $\forall i \, \forall x \; |\varphi_i(x)| \le |x|$. Let $X \in L_1 = L_1(\Omega)$ and $Y_i = \varphi_i(X)$. Then $Y_i \in L_1$ and $Y_i \to X$ in L_1 (that is, $\|Y_i - X\|_1 = \mathbb{E}\,|Y_i - X| \to 0$) by the dominated convergence theorem. For example, we may take $Y_i = \mathbb{1}_{(-i,i)}(X) \cdot X$ or $Y_i = \operatorname{mid}(-i, X, i)$.

Let X_k be as in 1c1; however, only pairwise independence will be used. Given $\varepsilon > 0$ we define $Y_k = \varphi(X_k)$ with a bounded φ such that $\|Y_1 - X_1\|_1 \le \varepsilon$. Then

$$\Big\| \frac{Y_1 + \dots + Y_n}{n} - \mathbb{E}\, Y_1 \Big\|_2 = \frac{1}{n} \| (Y_1 - \mathbb{E}\, Y_1) + \dots + (Y_n - \mathbb{E}\, Y_n) \|_2 = \frac{1}{n} \sqrt{n} \|Y_1 - \mathbb{E}\, Y_1\|_2 \to 0 \quad \text{as } n \to \infty$$

by orthogonality ensured by the pairwise independence: $\langle Y_k - \mathbb{E}\, Y_k, Y_l - \mathbb{E}\, Y_l \rangle = \mathbb{E}\,\big((Y_k - \mathbb{E}\, Y_k)(Y_l - \mathbb{E}\, Y_l) \big) = \big(\mathbb{E}\,(Y_k - \mathbb{E}\, Y_k) \big) \big(\mathbb{E}\,(Y_l - \mathbb{E}\, Y_l) \big) = 0 \cdot 0 = 0$ for $k \ne l$.

We have

$$\Big\| \frac{X_1 + \dots + X_n}{n} - \mathbb{E}\, X_1 \Big\|_1 \le \Big\| \frac{X_1 + \dots + X_n}{n} - \frac{Y_1 + \dots + Y_n}{n} \Big\|_1 + \\ + \Big\| \frac{Y_1 + \dots + Y_n}{n} - \mathbb{E}\, Y_1 \Big\|_1 + |\mathbb{E}\, Y_1 - \mathbb{E}\, X_1| \le \\ \le \|X_1 - Y_1\|_1 + \Big\| \frac{Y_1 + \dots + Y_n}{n} - \mathbb{E}\, Y_1 \Big\|_2 + \|X_1 - Y_1\|_1 \le 2\varepsilon + o(1) \, ;$$

$\limsup_n (\dots) \le 2\varepsilon$ for every ε. Convergence in L_1 is proved; convergence in probability follows (by the Markov inequality).

End of proof of 1c1

Recall the cumulative distribution function F defined by

$$F(t) = \mathbb{P}\big(X_1 \le t \big) \quad \text{for } t \in \mathbb{R} \, .$$

The empirical distribution function is the *random* function F_n defined by

$$F_n(t) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{(-\infty,t]}(X_k) \, .$$

1c2 Theorem. [1] For every $\varepsilon > 0$,

$$\mathbb{P}\big(\sup_t |F_n(t) - F(t)| \le \varepsilon\big) \to 1 \quad \text{as } n \to \infty\,.$$

This is the weak form of the Glivenko-Cantelli theorem.

1c3 Lemma. For every probability measure[2] μ on $\mathbb{R}$ and every $\varepsilon > 0$ there exist m and $t_1 < \dots < t_m$ such that

$$\mu\big((-\infty, t_1)\big) \le \varepsilon\,,\; \mu\big((t_1, t_2)\big) \le \varepsilon\,,\; \dots\,,\; \mu\big((t_{m-1}, t_m)\big) \le \varepsilon\,,\; \mu\big((t_m, +\infty)\big) \le \varepsilon\,.$$

Proof. We take $t_1 = \sup\{t : \mu((-\infty, t)) \le \varepsilon\}$ (the set is not empty!), then $\mu((-\infty, t_1)) = \lim_k \mu((-\infty, t_1 - \frac1k)) \le \varepsilon$ but $\mu((-\infty, t_1]) = \lim_k \mu((-\infty, t_1 + \frac1k)) \ge \varepsilon$. If $\varepsilon \ge 0.5$ then we are done. Otherwise, $t_2 = \sup\{t : \mu((t_1, t)) \le \varepsilon\}$, then $\mu((t_1, t_2)) \le \varepsilon$ but $\mu((t_1, t_2]) \ge \varepsilon$, thus $\mu((-\infty, t_2]) \ge 2\varepsilon$. And so on... ☐

Proof of 1c2

Let $F(t) = \mu\big((-\infty, t]\big)$, random functions $F_n(t) = \mu_n\big((-\infty, t]\big)$, and ε be as in 1c2. Lemma 1c3 gives us $t_1, \dots, t_m$. By 1a24,

$$\mathbb{P}\big(|\mu_n\big((-\infty, t_k]\big) - \mu\big((-\infty, t_k]\big)| > \varepsilon\big) \to 0 \quad \text{as } n \to \infty$$

for each $k = 1, \dots, m$. Sum it over k:

$$\mathbb{P}\big(\max_k |\mu_n\big((-\infty, t_k]\big) - \mu\big((-\infty, t_k]\big)| > \varepsilon\big) \to 0 \quad \text{as } n \to \infty\,.$$

The same holds for open intervals $(-\infty, t_k)$. Assuming that $|\mu_n\big((-\infty, t_k]\big) - \mu\big((-\infty, t_k]\big)| \le \varepsilon$ and $|\mu_n\big((-\infty, t_k)\big) - \mu\big((-\infty, t_k)\big)| \le \varepsilon$ we have for every k and[3] every $t \in (t_{k-1}, t_k)$

$$\begin{aligned}\mu_n\big((-\infty, t]\big) \in [\mu_n\big((-\infty, t_{k-1}]\big), \mu_n\big((-\infty, t_k)\big)] \subset \\ \subset [\mu\big((-\infty, t_{k-1}]\big) - \varepsilon, \mu\big((-\infty, t_k)\big) + \varepsilon] \ni \mu\big((-\infty, t]\big)\end{aligned}$$

and therefore[4] $|\mu_n\big((-\infty, t]\big) - \mu\big((-\infty, t]\big)| \le 3\varepsilon$. Thus,

$$\mathbb{P}\big(\sup_t |F_n(t) - F(t)| \le 3\varepsilon\big) = \mathbb{P}\big(\sup_t |\mu_n\big((-\infty, t] - \mu\big((-\infty, t]| \le 3\varepsilon\big) \to 0$$

as $n \to \infty$.

End of proof of 1c2

[1] [KS, Sect. 2.1, Th. 2.9].

[2] The measure may have both atoms and a continuous part, of course.

[3] The two unbounded intervals are treated similarly.

[4] You can easily improve 3ε to 2ε.