Lecture 13:

Limit theorems

Law of large numbers

Law of large numbers

Assume we have i.i.d. $X_1, X_2, \ldots$ with finite mean μ and finite variance σ^2. Let

$$\overline{X}_n = \frac{X_1 + \ldots + X_n}{n}$$

be the sample mean (of X_1 through X_n) - it itself is a r.v. with mean μ:

$$E(\overline{X}_n) = \frac{1}{n} E(X_1 + \ldots + X_n) = \frac{1}{n}\Big(E(X_1) + \ldots + E(X_n)\Big) = \mu$$

and variance σ^2/n:

$$\text{Var}(\overline{X}_n) = \frac{1}{n^2}\text{Var}(X_1 + \ldots + X_n) = \frac{1}{n^2}\Big(\text{Var}(X_1) + \ldots + \text{Var}(X_n)\Big) = \frac{\sigma^2}{n}$$

The **law of large numbers (LLN)** says that as n grows, the sample mean $\overline{X}_n$ converges to the true mean μ. LLN has two versions – “weak” and “strong”.

Law of large numbers

Theorem 10.2.1 (Strong LLN). The sample mean $\overline{X}_n$ converges to the true mean μ pointwise, with probability 1. Recall that r.v.s are functions from the sample space S to $\mathbb{R}$ – ***pointwise convergence*** says that $\overline{X}_n \to \mu$ for each point $s \in S$, except maybe some set B_0 of points, as long as $P(B_0) = 0$. In short, $P(\overline{X}_n \to \mu) = 1$.

Theorem 10.2.2 (Weak LLN). For all $\varepsilon > 0$, $P(|\overline{X}_n - \mu| > \varepsilon) \to 0$, as $n \to \infty$. (This is called ***convergence in probability***).

Proof: Fix $\varepsilon > 0$. By Chebyshev's ineq., $P(|\overline{X}_n - \mu| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$. As $n \to \infty$, the r.h.s goes to), and so must the l.h.s.

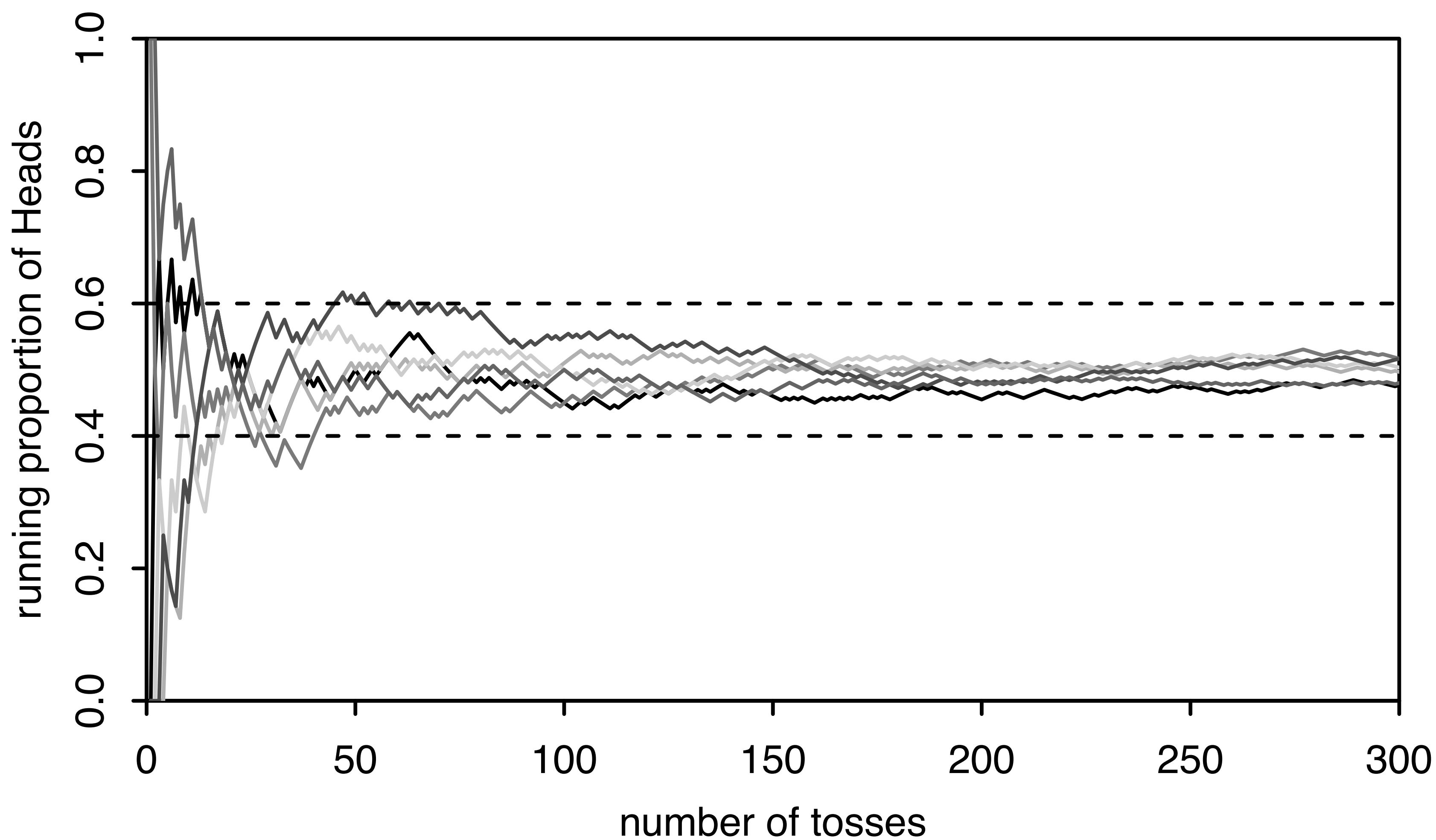
Law of large numbers

LLN is essential for simulations, statistics and science in general – when generating data by replicating an experiment and averaging the result to approximate the theoretical average, we appeal to LLN.

Example 10.2.3 (Running proportion of Heads). Let $X_1, X_2, \dots$ be i.i.d. Bern$(1/2)$ (coin tosses). $\bar{X}_n$ – the proportion of Heads after n tosses, by SLLN, the sequence $\bar{X}_1, \bar{X}_2, \dots$ will converge to 1/2.

Outomes like HHHH… or HHTHHT… are possible, but collectively have zero probability of occurring. The WLLN says that $\forall \varepsilon > 0$, $P(|\bar{X}_n - 1/2| > \varepsilon)$ can be made as small as we like as n grows

Law of large numbers

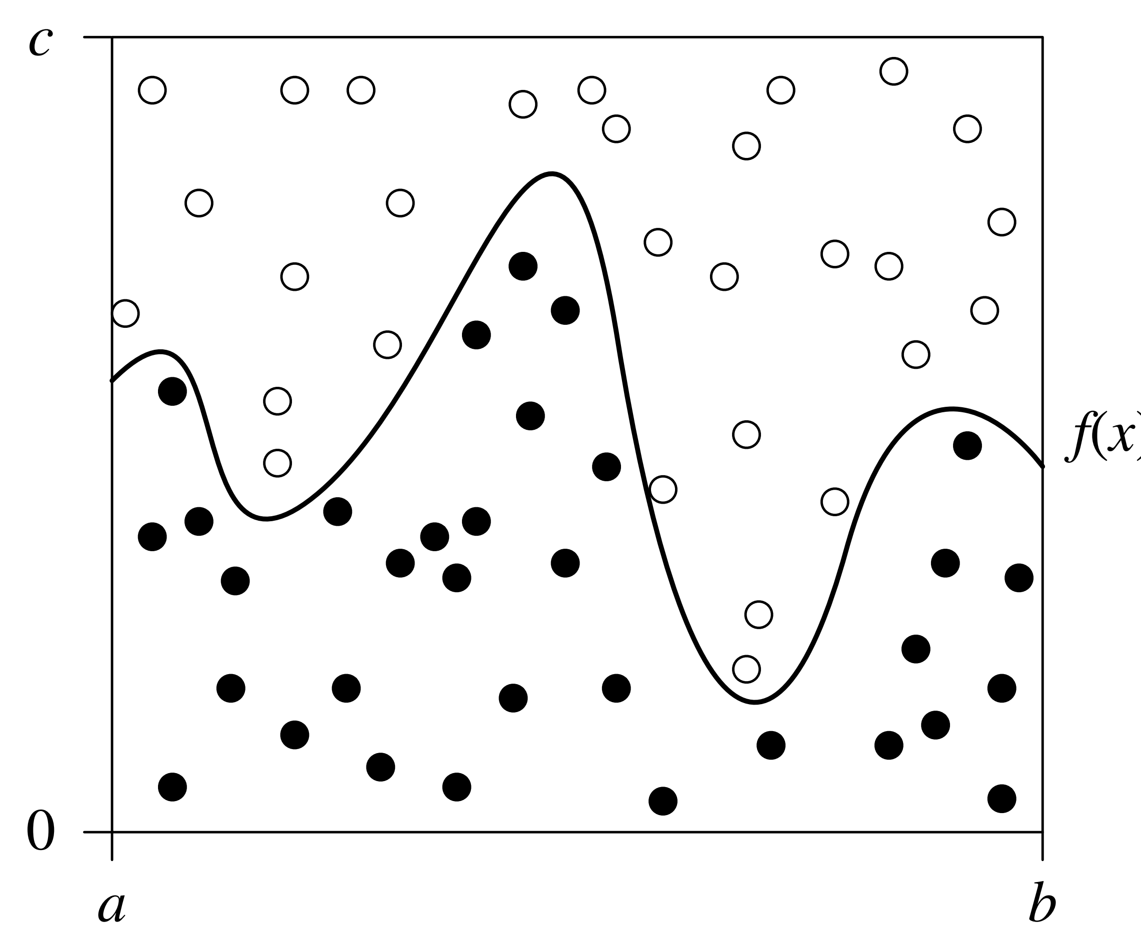


LLN does not contradict the memoryless-ness of the coin: the fact that the proportion of Heads converges to 1/2 does ***not*** imply that after a long string of Heads the coin is “due” for a Tails to balance things – convergence happens by past tosses being swamped away by infinitely many tosses yet to come.

Law of large numbers

Example 10.2.5 (Monte Carlo integration). Let f be some complicated function whose integral $\displaystyle\int_a^b f(x)\,dx$ we'd like to approximate. Assume $0 \leq f(x) \leq c$, so we know the integral is finite.

By randomly (uniformly) generating points from the rectangle $a \leq x \leq b$, $0 \leq y \leq c$, area under $y = f(x)$ (the integral's value) can be approximated by $c \cdot (b - a) \cdot p$, where p is the fraction of points under $y = f(x)$



The Central Limit Theorem

Central Limit Theorem

Assume we have i.i.d. $X_1, X_2, \ldots$ with finite mean μ and finite variance σ^2. LLN says that as $n \to \infty$, $\overline{X}_n$ converges to μ with probability 1. But what is the distribution?

Theorem 10.3.1 (Central limit theorem). As $n \to \infty$,

$$\sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right) \to \mathcal{N}(0,1) \text{ in distribution.}$$

That means that the CDF of the l.h.s converges to Φ, the CDF of the standard Normal distribution.

Central Limit Theorem

Proof: Let $M(t) = E(e^{tX_j})$ – be the moment generating function (MGF) of X_j. Assume $\mu = 0$, $\sigma^2 = 1$. (We would standardise $\overline{X}_n$ for the theorem anyway, so we might as well standardise X_j).
Then $M(0) = 1$, $M'(0) = \mu = 0$ and $M''(0) = \sigma^2 = 1$.

We wish to show that the MGF of $\sqrt{n}\overline{X}_n = (X_1 + \ldots + X_n)/\sqrt{n}$ converges to the MGF of the $\mathcal{N}(0,1)$, which is $e^{t^2/2}$. (This is because convergence of MGFs implies convergence of distributions)

So $E(e^{t(X_1+\ldots+X_n)/\sqrt{n}}) = E(e^{tX_1/\sqrt{n}})\,E(e^{tX_2/\sqrt{n}})\ldots E(e^{tX_n/\sqrt{n}}) =$
$= \left(M(t/\sqrt{n})\right)^n$.

Central Limit Theorem

Proof: $E(e^{t(X_1+\ldots+X_n)/\sqrt{n}}) = \left(M(t/\sqrt{n})\right)^n$ – as $n \to \infty$, this is indeterminate of form 1^∞, so let's look at the logarithm (and then exponentiate):

$$\lim_{n\to\infty} n \log M\left(\frac{t}{\sqrt{n}}\right) = \lim_{y\to 0} \frac{\log M(yt)}{y^2} \quad \text{where } y = 1/\sqrt{n}$$

$$= \lim_{y\to 0} \frac{t\,M'(yt)}{2y\,M(yt)} \quad \text{by L'Hôpital's rule}$$

$$= \frac{t}{2} \lim_{y\to 0} \frac{M'(yt)}{y} \quad \text{since } M(yt) \to 1$$

$$= \frac{t^2}{2} \lim_{y\to 0} M''(yt) \quad \text{by L'Hôpital's rule}$$

$= t^2/2$. So the MGF of $\sqrt{n}\overline{X}_n$ approaches $e^{t^2/2}$, the $\mathcal{N}(0,1)$ MGF.

Central Limit Theorem

CLT gives an ***approximation*** for the distr. of $\overline{X}_n$ for n large, but finite:

Approximation 10.3.2 (CLT approximation). For large n, the distribution of $\overline{X}_n \dot{\sim} \mathcal{N}(\mu, \sigma^2/n)$ – is approximately normal.

Note that the distribution of X_j can be ***any!*** – as long as the mean and variance are finite, and yet large-sample average would be approximately normally distributed!

Example 10.3.3 (Running proportion of Heads). Again $X_1, X_2, \dots$ – i.i.d. Bern$(1/2)$. LLN says $\overline{X}_n \to 1/2$ as $n \to \infty$. Now we can say more: $\overline{X}_n \dot{\sim} \mathcal{N}\left(\frac{1}{2}, \frac{1}{4n}\right)$. For example, when $n = 100$, SD$(\overline{X}_n) = 1/20 = 0.05$, so the 68-95-99.7 rule says there's a 95% chance that $\overline{X}_n$ is in the interval $[0.4, 0.6]$.

Central Limit Theorem

Equivalently, $W_n = X_1 + \ldots + X_n = n\overline{X}_n$ is also approximately Normally distributed: $W_n \dot{\sim} \mathcal{N}(n\mu, n\sigma^2)$.

Example 10.3.4 (Poisson). Let $Y \sim$ Pois(n). We can consider Y to be a sum of n i.i.d. Pois(1) r.v.s. So, for large n:

$$Y \dot{\sim} \mathcal{N}(n, n)$$

Example 10.3.5 (Gamma). Let $Y \sim$ Gamma(n, λ). We can consider Y to be a sum of n i.i.d. Expo(λ) r.v.s. So, for large n,

$$Y \dot{\sim} \mathcal{N}(n/\lambda, n/\lambda^2)$$

Example 10.3.6 (Binomial). Let $Y \sim$ Bin(n, p). We can consider Y to be a sum of n i.i.d. Bern(p) r.v.s. So, for large n,

$$Y \dot{\sim} \mathcal{N}(np, np(1-p))$$

The latter is very widely used in Statistics!

Chi-Square and Student-t

Chi-Square and Student-t

These are two continuous distributions closely related to the Normal

Definition 10.4.1 (Chi-Square) Let $V = Z_1^2 + \ldots + Z_n^2$ where Z_i are all i.i.d. $\mathcal{N}(0,1)$. Then V is said to have the Chi-Square distribution with n degrees of freedom, written as $V \sim \chi_n^2$.

Actually, χ_n^2 is a special case of the Gamma:

Theorem 10.4.2 The χ_n^2 distribution is the Gamma$(\frac{n}{2}, \frac{1}{2})$ distribution

Proof follows from the PDF of $Z_i^2 \sim \chi_1^2$ being the PDF of Gamma$(\frac{1}{2}, \frac{1}{2})$, and so $V = Z_1^2 + \ldots + Z_n^2$ – sum of n independent Gamma$(\frac{1}{2}, \frac{1}{2})$ r.v.s – is $V \sim$ Gamma$(\frac{n}{2}, \frac{1}{2})$.

Chi-Square and Student-t

Expectation of Chi-Square is $E(V) = nE(Z_1^2) = n$ and variance:
Var$(V) = n$Var$(Z_1^2) = n\Big(E(Z_1^4) - (EZ_1^2)^2\Big) = n(3-1) = 2n$.

Chi-Square is important in Statistics because it's related to **sample variance** – used to estimate the true variance of a distribution:

Example 10.4.3 (Distribution of sample variance). For i.i.d. $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, the sample variance is the r.v.

$$S_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2$$

which is $\dfrac{(n-1)\,S_n^2}{\sigma^2} \sim \chi^2_{n-1}$ – distributed.

Chi-Square and Student-t

The Student-t is defined in terms of standard Normal and χ_n^2:

Definition 10.4.4 (Student-t). Let $T = \dfrac{Z}{\sqrt{V/n}}$ where $Z \sim \mathcal{N}(0,1)$, $V \sim \chi_n^2$ and Z is independent of V. Then T is said to have the Student-t distribution with n degrees of freedom, written $T \sim t_n$.

This distribution was introduced in 1908 by William Gosset, a Master Brewer at Guinness, working on quality control of beer. He was required to publish this work under a pseudonym.

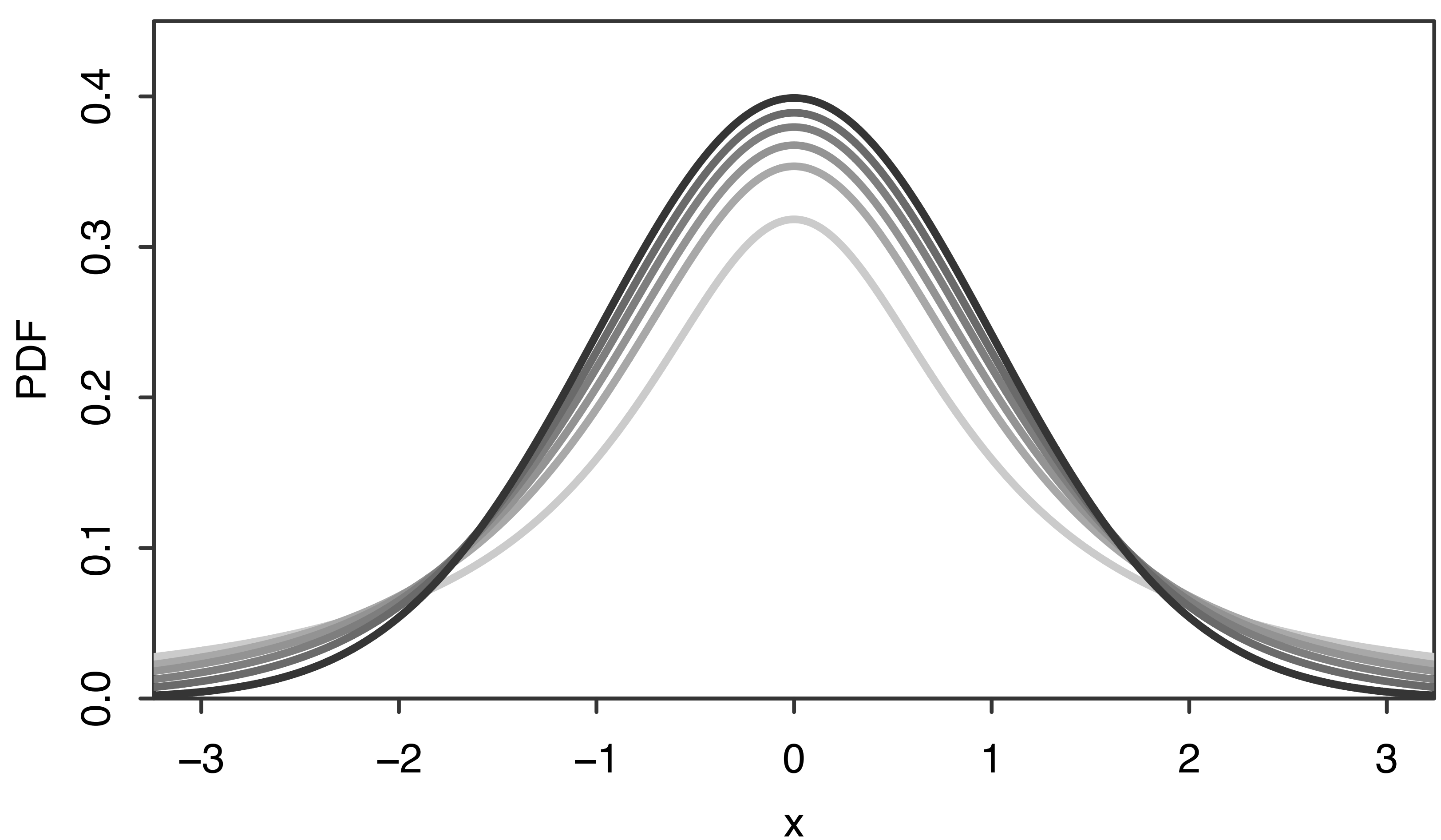
The t distribution forms the basis of hypothesis testing, ***t-tests***, extremely widely used in practice.

Chi-Square and Student-t

The PDF of the Student-t with n d.o.f. looks like standard Normal, except with heavier tails:

$$f_T(t) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\,\Gamma(n/2)}\left(1 + t^2/n\right)^{-(n+1)/2}$$

Here's how it looks for different n (smaller n – heavier tails):



Chi-Square and Student-t

Theorem 10.4.5 (Student-t properties). The Student-t distribution t_n has the following properties:

1. Symmetry: If $T \sim t_n$, then $-T \sim t_n$ as well.

2. Cauchy as special case: The t_1 distribution is the same as the Cauchy distribution!

3. Convergence to Normal: as $n \to \infty$, the t_n distribution converges to the standard Normal $\mathcal{N}(0,1)$