



Continuous Random Variables

Densities, the exponential, and inverse-transform sampling

Probability Theory • Lecture 6

Applied Data Analysis — Module I

September 2026

Roadmap & learning objectives

Everything so far lived on a countable set. Today the outcome can be any real number, and sums become integrals.

- Define a **density**; prove the basic properties of F_X and see why $\mathbb{P}(X = x) = 0$.
- Compute **expectation and variance** by integration; state LOTUS in its continuous form.
- Define **quantiles** and the median, and say when they beat the mean.
- Master the **uniform** and the **exponential**; prove memorylessness and its converse.
- Prove the **change-of-variables** formula, including the non-monotone case.
- Prove the **probability integral transform** and use it for **inverse-transform sampling**.
- ML tie-in: generating data, the reparameterization trick, hazard models, and calibration.

Where we stand after Lectures 3–5

$$p_X(x) = \mathbb{P}(X = x); \quad F_X(x) = \mathbb{P}(X \leq x); \quad \mathbb{E}[g(X)] = \sum_x g(x)p_X(x); \quad \text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}X)^2.$$

- The CDF was defined for *every* random variable — that is the bridge we cross today.
- Expectation, variance, covariance, Markov, Chebyshev, generating functions: none of the proofs used discreteness in an essential way. Replace $\sum_x$ by $\int dx$ and they all survive.

Today's question

A waiting time, a price, a measurement error can take any value in an interval, so every individual value has probability zero. How do we describe such a variable, and how do we simulate one on a computer?

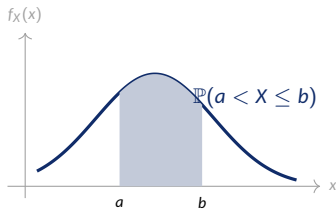


1. Densities

From mass to density

Pick a point uniformly on $[0, 1]$. By symmetry every point is equally likely, so $\mathbb{P}(X = x) = c$ for all x ; but uncountably many equal masses cannot sum to 1 unless $c = 0$.

- So $\mathbb{P}(X = x) = 0$ for every x , and yet $\mathbb{P}(X \in [0, 1]) = 1$. Probability is carried by *intervals*, not points.
- Physical analogy: a wire of unit mass has zero mass at each point, but a well-defined mass *per unit length*.
- That linear density is f_X , and probability is its area.



Probability = area under the density.

Definition — density and distribution function

Definition (PDF)

X is **continuous** with **density** $f_X \geq 0$ if $F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f_X(u) du$ for all x , and $\int_{\mathbb{R}} f_X = 1$.

Theorem

(1) $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X$; (2) $\mathbb{P}(X = x) = 0$, so open and closed intervals have the same probability; (3) F_X is continuous and non-decreasing, from 0 to 1; (4) $F'_X(x) = f_X(x)$ wherever f_X is continuous.

Proof.

(1) additivity, as in Lecture 3. (2) $\mathbb{P}(X = x) = F_X(x) - F_X(x^-) = 0$: the integral has no jumps. (3) $f_X \geq 0$ is integrable. (4) the fundamental theorem of calculus. $\square$

The CDF is the common language: discrete variables have a staircase, continuous ones a smooth curve.

A density is not a probability

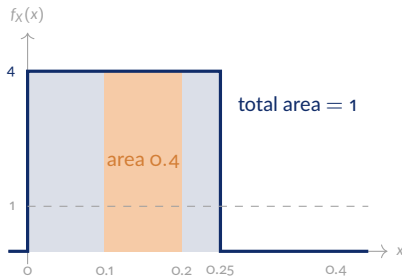
- $f_X(x)$ may exceed 1: for $X \sim \text{Unif}(0, \frac{1}{2})$ the density equals 2 on that interval. Only the *area* is capped at 1.
- The right reading is infinitesimal: $\mathbb{P}(x < X \leq x + dx) \approx f_X(x) dx$. Density has units of “probability per unit of x ”, so it changes when you change units — prices in kopecks give a density 100 times smaller than prices in roubles.
- Comparing $f_X(x_1) > f_X(x_2)$ is meaningful (more mass nearby); reading $f_X(x)$ as a probability is not.

Mixed types exist

Insurance losses are 0 with probability 0.7 and otherwise spread over $(0, \infty)$: a jump plus a density. Such variables have neither a PMF nor a PDF, but they do have a CDF — one more reason F_X is the primary object.

Example — a uniform whose density exceeds 1

Let $X \sim \text{Unif}(0, \frac{1}{4})$: a point chosen at random from an interval of length $\frac{1}{4}$.



- The density is constant on the interval and must integrate to 1: $f_X(x) = \frac{1}{1/4} = 4$ for $0 \leq x \leq \frac{1}{4}$, and 0 elsewhere.
- So $f_X(0.15) = 4$ — a value no probability could ever take.
- A probability is an *area*:
 $\mathbb{P}(0.1 < X \leq 0.2) = \int_{0.1}^{0.2} 4 \, dx = 4 \cdot 0.1 = 0.4.$
- And $\mathbb{P}(X = 0.15) = \int_{0.15}^{0.15} 4 \, dx = 0$: the height at a point carries no mass.

A density is not a probability. $f_X(x)$ is probability *per unit length*; it can exceed 1, and only its integral over an interval is a probability.

Expectation and variance by integration

Definition

$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x) dx$ whenever $\int_{\mathbb{R}} |x| f_X(x) dx < \infty$, and $\text{Var}(X) = \mathbb{E}[(X - \mu)^2]$.

Theorem (LOTUS, and what carries over)

$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x) f_X(x) dx$. *Linearity, monotonicity, $\text{Var}(aX + b) = a^2 \text{Var}(X)$, $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}X)^2$, Markov, Chebyshev and MGFs hold verbatim.*

Theorem (Tail formula)

If $X \geq 0$ then $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > x) dx$ (the continuous tail sum).

Proof.

$\int_0^\infty \mathbb{P}(X > x) dx = \int_0^\infty \int_x^\infty f_X(u) du dx = \int_0^\infty \int_0^u dx f_X(u) du$, by exchanging the order of integration. $\square$

Worked example — mean and variance of a uniform

$X \sim \text{Unif}(0, 1)$, so $f_X(x) = 1$ on $[0, 1]$ and 0 elsewhere. Every integral runs over $[0, 1]$ only.

$$\mathbb{E}[X] = \int_0^1 x \cdot 1 \, dx = \left[\frac{x^2}{2} \right]_0^1 = \frac{1}{2},$$

$$\mathbb{E}[X^2] = \int_0^1 x^2 \cdot 1 \, dx = \left[\frac{x^3}{3} \right]_0^1 = \frac{1}{3},$$

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}, \quad \sigma_X = \frac{1}{\sqrt{12}} \approx 0.29.$$

Same recipe, a sloped density

$f_X(x) = 2x$ on $[0, 1]$ (check: $\int_0^1 2x \, dx = 1$). Then $\mathbb{E}[X] = \int_0^1 x \cdot 2x \, dx = \frac{2}{3}$, $\mathbb{E}[X^2] = \int_0^1 x^2 \cdot 2x \, dx = \frac{1}{2}$, $\text{Var}(X) = \frac{1}{2} - \frac{4}{9} = \frac{1}{18}$. The mass leans right, so the mean moves from $\frac{1}{2}$ to $\frac{2}{3}$.

The pattern never changes: multiply by the density, integrate over the support, subtract the squared mean.

Worked example — mean and variance of an exponential

$X \sim \text{Exp}(\lambda)$, $f_X(x) = \lambda e^{-\lambda x}$ on $[0, \infty)$. Integrate by parts with $u = x$, $dv = \lambda e^{-\lambda x} dx$:

$$\mathbb{E}[X] = \int_0^\infty x \lambda e^{-\lambda x} dx = \left[-x e^{-\lambda x} \right]_0^\infty + \int_0^\infty e^{-\lambda x} dx = 0 + \left[-\frac{e^{-\lambda x}}{\lambda} \right]_0^\infty = \frac{1}{\lambda},$$

$$\mathbb{E}[X^2] = \int_0^\infty x^2 \lambda e^{-\lambda x} dx = \left[-x^2 e^{-\lambda x} \right]_0^\infty + \int_0^\infty 2x e^{-\lambda x} dx = 0 + \frac{2}{\lambda} \mathbb{E}[X] = \frac{2}{\lambda^2},$$

$$\text{Var}(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}, \quad \sigma_X = \frac{1}{\lambda} = \mathbb{E}[X].$$

- The boundary terms vanish because $e^{-\lambda x}$ kills any polynomial as $x \rightarrow \infty$.
- In the second line the remaining integral is $\frac{2}{\lambda} \int_0^\infty x \lambda e^{-\lambda x} dx$ — the mean again. Each extra power of x costs one more integration by parts: $\mathbb{E}[X^k] = k!/\lambda^k$.
- With $\lambda = 0.5$ (mean wait 2 minutes): $\mathbb{E}[X] = 2$, $\text{Var}(X) = 4$, $\sigma_X = 2$.

Quantiles and the median

Definition

For $\alpha \in (0, 1)$ the α -**quantile** is $q_\alpha = \inf\{x : F_X(x) \geq \alpha\}$; if F_X is continuous and strictly increasing it is simply $q_\alpha = F_X^{-1}(\alpha)$. The **median** is $q_{0.5}$.

- The median minimizes $\mathbb{E}|X - c|$, just as the mean minimizes $\mathbb{E}[(X - c)^2]$ (Lecture 3). Absolute loss gives you a quantile; squared loss gives you a mean. The loss chooses the statistic.
- **Quantile regression** and **Value-at-Risk** are the same idea: predict q_α by minimizing the pinball loss $\rho_\alpha(u) = u(\alpha - \mathbf{1}\{u < 0\})$.
- Quantiles survive heavy tails: the St. Petersburg payoff of Lecture 3 has no mean but a median of 2.

Exponential

$F(x) = 1 - e^{-\lambda x}$ gives $q_\alpha = -\ln(1 - \alpha)/\lambda$; the median $\ln 2/\lambda \approx 0.69/\lambda$ is well below the mean $1/\lambda$ — a right-skewed law.



2.

Two basic families

The uniform distribution

Definition

$X \sim \text{Unif}(a, b)$ has $f_X(x) = \frac{1}{b-a}$ on $[a, b]$ and 0 elsewhere.

Theorem

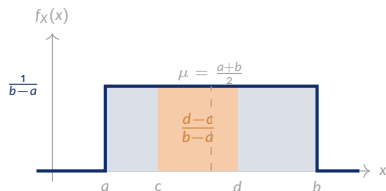
$$F_X(x) = \frac{x-a}{b-a} \text{ on } [a, b]; \quad \mathbb{E}[X] = \frac{a+b}{2}, \quad \text{Var}(X) = \frac{(b-a)^2}{12}.$$

Proof.

$$\mathbb{E}[X] = \frac{1}{b-a} \int_a^b x \, dx = \frac{a+b}{2} \text{ and } \mathbb{E}[X^2] = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3}; \text{ subtract } \left(\frac{a+b}{2}\right)^2. \quad \square$$

- $\text{Unif}(0, 1)$ is the raw material of simulation: every random number generator produces it, and the last section turns it into anything else.

The uniform in pictures



Flat density: every sub-interval of the same length gets the same probability, $\mathbb{P}(c < X < d) = \frac{d-c}{b-a}$ — length over length.

Where it shows up

- **Arrival in a window.** A bus comes at a uniform time in a 10-minute window; your wait $W \sim \text{Unif}(0, 10)$: $\mathbb{P}(W > 7) = \frac{3}{10}$, $\mathbb{E}[W] = 5$, $\sigma_W = 10/\sqrt{12} \approx 2.9$.
- **Rounding error.** Measurements rounded to the nearest gram have error $\text{Unif}(-\frac{1}{2}, \frac{1}{2})$: mean 0, variance $\frac{1}{12}$ — the “1/12” in every quantization-noise formula.

Compute one

$X \sim \text{Unif}(2, 8)$: $f_X = \frac{1}{6}$ on $[2, 8]$,
 $\mathbb{P}(3 < X < 5) = \frac{2}{6} = \frac{1}{3}$, $\mathbb{E}[X] = 5$,
 $\text{Var}(X) = \frac{36}{12} = 3$.

The exponential distribution

Definition

$X \sim \text{Exp}(\lambda)$, $\lambda > 0$, has $f_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$.

Theorem

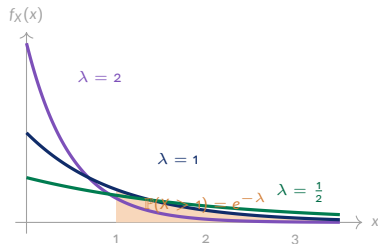
$F_X(x) = 1 - e^{-\lambda x}$ and $\mathbb{P}(X > x) = e^{-\lambda x}$ for $x \geq 0$; $\mathbb{E}[X] = \frac{1}{\lambda}$, $\text{Var}(X) = \frac{1}{\lambda^2}$, $M_X(t) = \frac{\lambda}{\lambda - t}$ for $t < \lambda$.

Proof.

The CDF is an immediate integral. Mean by the tail formula: $\int_0^\infty e^{-\lambda x} dx = 1/\lambda$. For the second moment, $\mathbb{E}[X^2] = \int_0^\infty x^2 \lambda e^{-\lambda x} dx = 2/\lambda^2$ (integrate by parts twice), so $\text{Var}(X) = 2/\lambda^2 - 1/\lambda^2$. The MGF is $\int_0^\infty e^{tx} \lambda e^{-\lambda x} dx$, finite only for $t < \lambda$. $\square$

λ is a **rate** (events per unit time); $1/\lambda$ is a mean waiting time. Note $\sigma = \mu$: the exponential is as spread out as it is large.

The exponential in pictures



The larger the rate λ , the taller the start and the faster the decay: events come sooner. The shaded tail is

$$\mathbb{P}(X > 1) = e^{-\lambda} \text{ for } \lambda = 1: \text{ about } 0.37.$$

Where it shows up

- **Time between events** that occur at a steady rate: calls to a help desk, clicks on a site, radioactive decays, packet arrivals.
- **Lifetimes without ageing**: a component that fails from random shocks, not wear.
- **Time to the next purchase** of a customer in a simple churn model (constant hazard).

Compute one

Calls arrive at $\lambda = 2$ per minute, so the gap $T \sim \text{Exp}(2)$: mean $\frac{1}{2}$ min, $\mathbb{P}(T > 1) = e^{-2} \approx 0.135$, $\mathbb{P}(T \leq 0.1) = 1 - e^{-0.2} \approx 0.18$, median $\ln 2/2 \approx 0.35$ min — half of all gaps are shorter than 70% of the mean.

Memorylessness

Theorem

$X \sim \text{Exp}(\lambda)$ satisfies $\mathbb{P}(X > s + t | X > s) = \mathbb{P}(X > t)$ for all $s, t \geq 0$, and it is the only continuous distribution on $[0, \infty)$ with this property.

Proof.

Direct: $\frac{\mathbb{P}(X > s + t)}{\mathbb{P}(X > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t}$. Conversely, write $\bar{F}(x) = \mathbb{P}(X > x)$; the property is the functional equation $\bar{F}(s + t) = \bar{F}(s)\bar{F}(t)$. A monotone solution must satisfy $\bar{F}(x) = \bar{F}(1)^x$, i.e. $\bar{F}(x) = e^{-\lambda x}$ with $\lambda = -\ln \bar{F}(1)$. □

- Exactly the geometric theorem of Lecture 4, with the same proof: the exponential is the geometric in continuous time.
- Consequence to internalize: a memoryless component does not age. If your bus has waited 20 minutes' worth of exponential, the expected further wait is still $1/\lambda$.

Exponential waits and the Poisson process

Theorem

Let events arrive as a Poisson process of rate λ (Lecture 4): the count in a window of length t is $\text{Pois}(\lambda t)$, and counts in disjoint windows are independent. Then the waiting time T to the first event is $\text{Exp}(\lambda)$, and the gaps between consecutive events are i.i.d. $\text{Exp}(\lambda)$.

Proof.

$\mathbb{P}(T > t) = \mathbb{P}(\text{no event in } [0, t]) = e^{-\lambda t} \frac{(\lambda t)^0}{0!} = e^{-\lambda t}$, which is the exponential tail. Independence of the gaps follows from independence over disjoint windows. $\square$

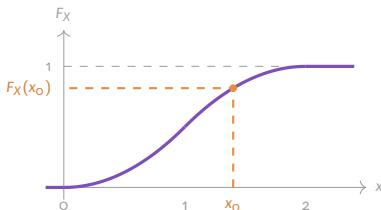
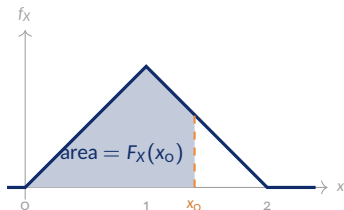
- Counting and waiting are two views of one model: Poisson counts, Exponential waits, and memorylessness is what glues them.
- The time to the r -th event is a sum of r i.i.d. exponentials — a **Gamma**(r, λ), the continuous negative binomial, which is Lecture 7.

3.

The CDF of a continuous variable, in pictures

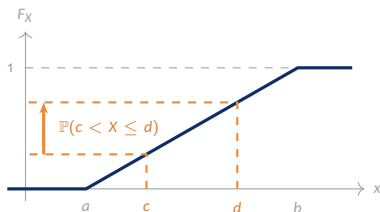
The CDF is accumulated area

Take the triangular density $f_X(x) = x$ on $[0, 1]$, $2 - x$ on $[1, 2]$. Its CDF $F_X(x) = \int_{-\infty}^x f_X$ is the area swept out so far.

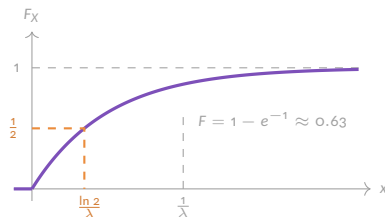


- F_X rises fastest where the density is tallest, is flat wherever $f_X = 0$, and climbs from 0 to 1: its slope is f_X .
- Here $F_X(x) = x^2/2$ on $[0, 1]$ and $1 - (2 - x)^2/2$ on $[1, 2]$; at $x_0 = 1.4$ the area is $F_X(1.4) = 1 - 0.18 = 0.82$.

Two CDFs you should recognize



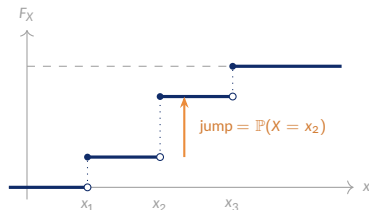
Uniform(a, b): a ramp, $F_X(x) = \frac{x - a}{b - a}$. Constant slope
= constant density.



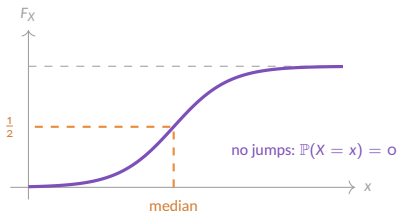
Exponential(λ): $F_X(x) = 1 - e^{-\lambda x}$, steep at first, then flattening. The median sits at $\ln 2 / \lambda$, well left of the mean.

- Probabilities are **vertical differences**: $\mathbb{P}(c < X \leq d) = F_X(d) - F_X(c)$. No integral needed once you have F .
- Quantiles are **horizontal cuts**: draw $y = \alpha$, read off where it meets the curve.

Reading a CDF — discrete versus continuous



Discrete: a staircase. Flat between atoms, jumps at them; the jump size is the PMF.



Continuous: a smooth curve. No jumps, so no atom carries mass; the slope is the density.

In both worlds $F_X(x) = \mathbb{P}(X \leq x)$ and $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$. The CDF is the one object that never changes its definition — which is why it is the right place to start when a variable is neither purely discrete nor purely continuous.



4.

Transformations

Change of variables

Theorem

Let X have density f_X and let g be differentiable and strictly monotone on the range of X . Then $Y = g(X)$ has density

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|.$$

Proof.

Take g increasing. Then $F_Y(y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$; differentiate in y by the chain rule. For decreasing g the inequality flips, giving $F_Y(y) = 1 - F_X(g^{-1}(y))$ and a derivative of the opposite sign — the absolute value covers both cases. $\square$

- Always work through the **CDF**: it is one line and never loses a Jacobian.
- The factor $|(g^{-1})'|$ is the one-dimensional Jacobian; in several dimensions it becomes $|\det J|$, which is exactly the term normalizing flows compute in modern generative models.

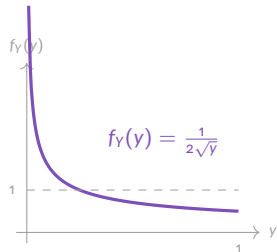
Change of variables — example 1: $Y = X^2$

$X \sim \text{Unif}(0, 1)$ and $Y = X^2$. Work through the CDF; the density is its derivative. Since $X \geq 0$, $X^2 \leq y$ means $X \leq \sqrt{y}$:

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X^2 \leq y) = \mathbb{P}(X \leq \sqrt{y}) = F_X(\sqrt{y}) = \sqrt{y}, \quad 0 \leq y \leq 1,$$

$$f_Y(y) = \frac{d}{dy} \sqrt{y} = \frac{1}{2\sqrt{y}}, \quad 0 < y \leq 1.$$

- Same answer from the formula: $g^{-1}(y) = \sqrt{y}$,
 $|(g^{-1})'(y)| = \frac{1}{2\sqrt{y}}$, $f_X(\sqrt{y}) = 1$.
- Sanity check: $\int_0^1 \frac{dy}{2\sqrt{y}} = [\sqrt{y}]_0^1 = 1$ — unbounded at 0, still integrates to 1.
- Mean two ways: $\int_0^1 y \cdot \frac{1}{2\sqrt{y}} dy = \frac{1}{3}$ and LOTUS
 $\mathbb{E}[X^2] = \int_0^1 x^2 dx = \frac{1}{3}$. They must agree.



Squaring pushes mass toward 0: small x become smaller.

Change of variables — example 2: scaling and a square root

Scaling an exponential: $X \sim \text{Exp}(\lambda)$, $Y = cX$ with $c > 0$

$$F_Y(y) = \mathbb{P}(cX \leq y) = \mathbb{P}\left(X \leq \frac{y}{c}\right) = 1 - e^{-\lambda y/c}, \quad f_Y(y) = \frac{d}{dy} \left(1 - e^{-\lambda y/c}\right) = \frac{\lambda}{c} e^{-(\lambda/c)y}.$$

So $Y \sim \text{Exp}(\lambda/c)$: waiting times in seconds instead of minutes are still exponential, with the rate divided by 60. Formula check: $g^{-1}(y) = y/c$, $|g^{-1}| = 1/c$, $f_X(y/c) = \lambda e^{-\lambda y/c}$.

A square root: $X \sim \text{Exp}(1)$, $Y = \sqrt{X}$

$$F_Y(y) = \mathbb{P}(\sqrt{X} \leq y) = \mathbb{P}(X \leq y^2) = 1 - e^{-y^2}, \quad f_Y(y) = 2ye^{-y^2}, \quad y \geq 0.$$

A new shape (the Rayleigh density): it starts at 0, rises, then decays. Check: $\int_0^\infty 2ye^{-y^2} dy = [-e^{-y^2}]_0^\infty = 1$.

Recipe: write $\{Y \leq y\}$ as an event about $X \rightarrow$ evaluate with $F_X \rightarrow$ differentiate in $y \rightarrow$ check the integral is 1.

Worked transformations

Affine

$Y = aX + b$ with $a \neq 0$: $f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$. Shifting moves the density, scaling stretches it and divides the height — so the area stays 1.

Exponentiating a uniform

$X \sim \text{Unif}(0, 1)$, $Y = e^X$: $g^{-1}(y) = \ln y$, so $f_Y(y) = 1 \cdot \frac{1}{y}$ on $(1, e)$. Check: $\int_1^e \frac{dy}{y} = \ln e - \ln 1 = 1$. A log-uniform law — the shape behind “sample the learning rate on a log scale”.

Making an exponential out of a uniform

$U \sim \text{Unif}(0, 1)$, $X = -\frac{1}{\lambda} \ln U$. Then $\mathbb{P}(X > x) = \mathbb{P}(U < e^{-\lambda x}) = e^{-\lambda x}$, so $X \sim \text{Exp}(\lambda)$. This is the next section in miniature.

When g is not monotone

Theorem

If g is piecewise monotone with branches $g_1, \dots, g_m$, then $f_Y(y) = \sum_{j=1}^m f_X(g_j^{-1}(y)) \left| \frac{d}{dy} g_j^{-1}(y) \right|$: add the contributions of all pre-images.

Proof.

The event $\{Y \leq y\}$ splits into disjoint pieces, one per branch; differentiate each piece as before and add. $\square$

Two branches

$X \sim \text{Unif}(-1, 1)$ with $f_X = \frac{1}{2}$, and $Y = X^2$. The pre-images of y are $\pm\sqrt{y}$, each contributing $\frac{1}{2} \cdot \frac{1}{2\sqrt{y}}$, so $f_Y(y) = \frac{1}{2\sqrt{y}}$ on $(0, 1)$. Forgetting one branch halves the answer.

The discrete analogue is Lecture 3's $p_Y(y) = \sum_{x:g(x)=y} p_X(x)$: sums replaced by Jacobians.



5.

Simulation: turning uniforms into anything

The probability integral transform

Theorem (PIT)

If F_X is continuous and strictly increasing, then $U = F_X(X) \sim \text{Unif}(0, 1)$.

Proof.

For $u \in (0, 1)$: $\mathbb{P}(F_X(X) \leq u) = \mathbb{P}(X \leq F_X^{-1}(u)) = F_X(F_X^{-1}(u)) = u$, which is the CDF of $\text{Unif}(0, 1)$. □

- Every continuous distribution is a relabelling of the uniform. The CDF is that relabelling.
- Read backwards, it says a percentile rank is uniform — the basis of rank transforms, copulas, and of the $[-1, 1]$ ranking of characteristics used by the ML asset-pricing papers.
- Continuity matters: for a discrete X the transform lands on atoms and is not uniform.

Inverse-transform sampling

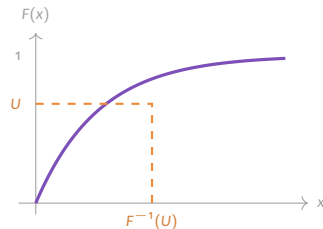
Theorem

If $U \sim \text{Unif}(0, 1)$ then $X = F^{-1}(U)$ has CDF F .

Proof.

$\mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x)$, using that F is increasing and U is uniform. $\square$

- Exponential: $X = -\ln(1 - U)/\lambda$ (and $1 - U$ is uniform too, hence the shorter $-\ln U/\lambda$).
- Discrete version: walk up the cumulative probabilities until they exceed U — this is how a categorical sampler works.



Draw U on the vertical axis, read x off the horizontal one.

When you cannot invert the CDF

Rejection sampling

Suppose $f \leq c q$ for a density q you can sample. Draw $X \sim q$ and $U \sim \text{Unif}(0, 1)$ independently; accept X if $U \leq f(X)/(c q(X))$, otherwise start again. The accepted values have density exactly f , and the acceptance probability is $1/c$.

- Proof idea: the pair $(X, U c q(X))$ is uniform under the curve $c q$; keeping the points below f leaves a uniform sample under f , whose horizontal coordinate has density f .
- The normal is sampled by neither method in practice — Box-Muller and the ziggurat are faster (Lecture 8).
- In high dimensions c explodes and rejection dies; that failure is what MCMC was invented to fix.

Reproducibility

Every sampler starts from a pseudo-random uniform stream. Fix and report the seed.





6.

Machine learning connections

Sampling inside a model

The reparameterization trick

A VAE needs $\nabla_{\theta} \mathbb{E}_{z \sim q_{\theta}} [\ell(z)]$, but you cannot differentiate through a sampler. Write the sample as a transformation of a parameter-free noise variable: $z = \mu_{\theta} + \sigma_{\theta} \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, 1)$. Now the randomness is outside the parameters and the gradient passes through — this is the change-of-variables slide, used in reverse.

- **Normalizing flows** go further: apply a chain of invertible maps to a simple density and track $\log |\det J|$ at every step. The exact same Jacobian factor we proved today.
- **Data augmentation and simulation studies** rest on inverse-transform sampling: you can only test an estimator on data whose truth you generated yourself.
- **Dropout, diffusion noise, bootstrap resampling** — all are transformations of uniforms.

Time-to-event models

Definition (Hazard)

For a positive X with density f and survival $S(x) = \mathbb{P}(X > x)$, the **hazard rate** is $h(x) = \frac{f(x)}{S(x)}$: the instantaneous failure rate given survival to x .

Theorem

h determines the law: $S(x) = \exp\left(-\int_0^x h(u) du\right)$. Constant hazard $h \equiv \lambda$ is exactly the exponential; increasing hazard means ageing, decreasing hazard means burn-in.

- Churn, equipment failure, time to conversion, credit default: the target is a duration, and often **censored** — the customer has not churned yet. Ordinary regression cannot use such a row; survival models can.
- Weibull ($h(x) = \alpha \lambda x^{\alpha-1}$) is the standard flexible alternative; Cox regression models $h(x | z) = h_0(x) e^{\beta^\top z}$ without committing to h_0 .

Calibration — the PIT as a diagnostic

A probabilistic forecast outputs a whole predictive CDF $\hat{F}_i$ for each observation y_i , not just a point.

If the forecasts are calibrated, the values $u_i = \hat{F}_i(y_i)$ are $\text{Unif}(0, 1)$. Plot their histogram.

- A **flat** histogram means honest uncertainty. A **U-shape** means the intervals are too narrow — the model is overconfident, the most common failure. A **hump** in the middle means they are too wide.
- Systematic drift to one side is bias, not miscalibration of spread: different diagnoses, different fixes.

Why this matters

Two models with identical RMSE can have completely different PIT histograms. If a downstream decision uses a quantile — inventory at the 95th percentile, VaR, a staffing threshold — calibration is what you are selling.



7. **Wrap-up**

Pitfalls & summary

- f_X is a density, not a probability: it can exceed 1, and it changes when you change units.
- $\mathbb{P}(X = x) = 0$ does not mean the value is impossible — only that it carries no mass.
- Never drop the Jacobian $|(g^{-1})'|$, and never forget the second branch when g is not monotone.
- The exponential is the only memoryless continuous law. Do not use it for anything that ages.
- Sample by transforming a uniform; fix the seed and say so.

$$F_X(x) = \int_{-\infty}^x f_X; \quad \mathbb{E}[g(X)] = \int g f_X; \quad f_Y(y) = f_X(g^{-1}(y)) |(g^{-1})'(y)|; \quad F_X(X) \sim \text{Unif}(0, 1).$$

Seminar problems & next lecture

- 1 $f(x) = cx^2$ on $[0, 2]$. Find c , F , the mean, the variance, the median, and $\mathbb{P}(X > 1 \mid X > 0.5)$.
- 2 $X \sim \text{Exp}(2)$. Find $\mathbb{P}(X > 1)$, $\mathbb{E}[X \mid X > 1]$, and verify memorylessness numerically.
- 3 Prove that the median minimizes $\mathbb{E}|X - c|$ over c .
- 4 $U \sim \text{Unif}(0, 1)$. Find the density of $Y = -\ln(1 - U)$, of $Y = U^{1/3}$, and of $Y = \min(U, 1 - U)$.
- 5 Write the inverse-transform sampler for the Pareto law $S(x) = (x_0/x)^\alpha$, $x \geq x_0$, and explain why its MGF does not exist (Lecture 5).
- 6 Let X, Y be independent $\text{Exp}(\lambda)$. Show $\min(X, Y) \sim \text{Exp}(2\lambda)$ and find $\mathbb{P}(X < Y)$.
- 7 Simulate 10^4 draws from $\text{Exp}(1)$ by inverse transform, plot the empirical CDF against the true one, and use Chebyshev and Hoeffding to bracket the error at $x = 1$.

Next: the Gamma and Beta families — waiting for several events, and modelling a probability.

Further reading

Core textbooks — densities, the uniform and exponential, transformations

- S. Ross, *A First Course in Probability*, 10th ed., Pearson, 2019 — Ch. 5 (continuous random variables, distribution of a function of a random variable).
- J. Blitzstein, J. Hwang, *Introduction to Probability*, 2nd ed., CRC Press, 2019 — Ch. 5; their “universality of the uniform” is our probability integral transform. Free at probabilitybook.net.
- D. Bertsekas, J. Tsitsiklis, *Introduction to Probability*, 2nd ed., Athena Scientific, 2008 — Ch. 3.
- G. Casella, R. Berger, *Statistical Inference*, 2nd ed., Duxbury, 2002 — Ch. 2, the cleanest treatment of the change-of-variables formula.

Simulation, survival and calibration

- L. Devroye, *Non-Uniform Random Variate Generation*, Springer, 1986 — inverse transform, rejection and everything else; free on the author's web page.
- D. Kroese, T. Taimre, Z. Botev, *Handbook of Monte Carlo Methods*, Wiley, 2011.
- J. Klein, M. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd ed., Springer, 2003 — hazard functions and censoring.
- T. Gneiting, F. Balabdaoui, A. Raftery, “Probabilistic forecasts, calibration and sharpness”, *J. Royal Statistical Society B* 69(2), 2007 — the paper behind PIT histograms.