Review of Probability

David I. Inouye

Why probability? Probability is useful for handling *uncertainty*

- Inherent stochasticity
  - Quantum mechanics
  - Card games
- Incomplete observability
  - "Let's Make a Deal" game show of three doors (called "Monty Hall" problem)
- Incomplete modeling
  - Discretization of space for object locations

Why probability?
Sometimes more practical than deterministic

- “Most birds fly”

- “Birds fly, except for very young birds that have not yet learned to fly, sick or injured birds that have lost the ability to fly, flightless species of birds including the cassowary, ostrich and kiwi…”
    - (Example from Deep Learning, Goodfellow et al., 2016, Ch. 3)

Why probability?
An extension of formal logic rules

- Original AI systems based on formal logic and reasoning
  - Chess
  - TurboTax

- Many AI applications based on deterministic logic were too brittle and failed often
  - Traditional linguistic approaches to natural language processing

- Modern AI systems almost always rooted in probability
  - Computer vision
  - Speech recognition
  - Natural language processing

How are these statements similar or different?

- A boardgame player: “The probability of getting a heads when flipping a fair coin is 50%.”
- The weather forecaster: “The probability of rain tomorrow is 50%.”
- Your doctor after examining your symptoms: “The probability of you having the flu is 50%.”

Frequentist and Bayesian interpretations lead to the same set of axioms

- Frequentist
  - Related to rates that events occur under repeated experimentation
- Bayesian interpretation
  - "Degree of belief"
- Pragmatic interpretation
  - They lead to the same math and are useful in similar circumstances
  - Use whichever interpretation is most useful

Big Picture Overview

Introduction to probability

- Motivation
- Interpretation

Random variables

- Outcomes and events

Probability distributions

- PMF, PDF, CDF

Multivariate distributions

- Joint, marginal and conditional distributions
- Chain rule and Bayes rule
- Independence

Expectations of random variables

- Linearity of expectations
- Covariance
- Empirical expectations

A **random variable** maps outcomes/events of a random/uncertain process to numbers

- Flipping a coin
  - Outcomes: {“Heads”, “Tails”}
  - Possible random variable: “Heads”-> 0, “Tails”-> 1

- Flipping two coins
  - Outcomes: {(H,H), (H,T), (T, H), (T, T)}
  - Possible random variables: # heads, # tails, same, different

- Flipping coins until you get one tails
  - Outcomes: ?
  - Random variables: ?

A **random variable** maps outcomes to numbers: Defining a random variable is the first step

- Random Tweet
  - Outcomes: ?
  - Random variables: ?

- Random Instagram image
  - Outcomes: ?
  - Random variables: ?

Random variables can be <u>discrete</u> or <u>continuous</u>

- Discrete
  - Values are in some finite set or countably infinite set
  - $\{-1,\ \ 1\}, \{5, 10, -20, 3\}, \{0, 1, 2, \dots\}, \mathbb{Z},$
- Continuous
  - Values associated with intervals of $\mathbb{R}$
  - $[0,1], [-1, 1], [0.5, 1] \cup [-1, 0.5], \mathbb{R}_+ \equiv [0, \infty)$
- *Note: Random variables by themselves do not provide any probability information.*

An event is a set of possible outcomes

- For discrete RV such as $\mathrm{X} \in \{0,1,2,\ldots\}$, then events could be:
  - $E = \{0,5,1\}$
  - $E = \{0,2,4,6,\ldots\}$ (i.e., all even numbers)
- For continuous random variables $X \in \mathbb{R}$, events are sets of the real numbers:
  - $E = [0,0.5)$
  - $E = [4,5] \cup [8,9]$

Big Picture Overview

Introduction to probability

- Motivation
- Interpretation

Random variables

- Outcomes and events

Probability distributions

- PMF, PDF, CDF

Multivariate distributions

- Joint, marginal and conditional distributions
- Chain rule and Bayes rule
- Independence

Expectations of random variables

- Linearity of expectations
- Covariance
- Empirical expectations

Probability distributions attach probabilities to all possible events of a random variable

- **Probability mass function (PMF)** is used for *discrete* random variables
- A PMF P for random variable X that satisfies the following:
  1. Domain of P must include all possible states of X
  2. Unit domain: $\forall\, x \in X, 0 \leq P(x) \leq 1$
  3. Sum to 1: $\sum_{x \in X} P(x) = 1$

Probability distributions attach probabilities to all possible events of a random variable

- **Probability density function (PDF)** is used for *continuous* random variables
- A PDF p for random variable X that satisfies the following:
  1. Domain of p must include all possible states of X
  2. Non-negative: $\forall\, x \in X, p(x) \geq 0$ ** $p(x)$ **could be greater than 1**
  3. Integrate to 1: $\int_{x \in X} p(x) = 1$
- $p(x)$ is NOT a probability, rather ***integrating*** the PDF gives probabilities over **sets**

Suppose $X \in (0, 1)$ (note: 0 is not included)

- Are the following functions valid PDFs? Why?
- $\forall x \in (0, 0.5), p(x) = 2;\ \ \forall x \notin (0, 0.5), p(x) = 0$
- $p(x) = 3x^2$
- $p(x) = -\log x$

Integrate PDF to get probabilities that random variable lies within a set (usually a range)

- The probability that X is less than q

$$\Pr(X \le q) = \int_{-\infty}^{q} p(x)dx$$

- The probability that X lies between a and b

$$\Pr(a \le X \le b) = \int_{a}^{b} p(x)dx$$

- The probability that X lies between (a and b) or between (c and d)

$$\Pr(a \le X \le b \text{ OR } c \le X \le d) = \int_{a}^{b} p(x)dx + \int_{c}^{d} p(x)dx$$

<u>Cumulative distribution function (CDF)</u> is the integral of the PDF from the left up to query point q

- The CDF is the probability that X is less than q

$$F(q) \equiv \Pr(X \leq q) = \int_{-\infty}^{q} p(x)dx$$

- What does $F(\infty)$ equal?

- The probability between a and b can be written as:

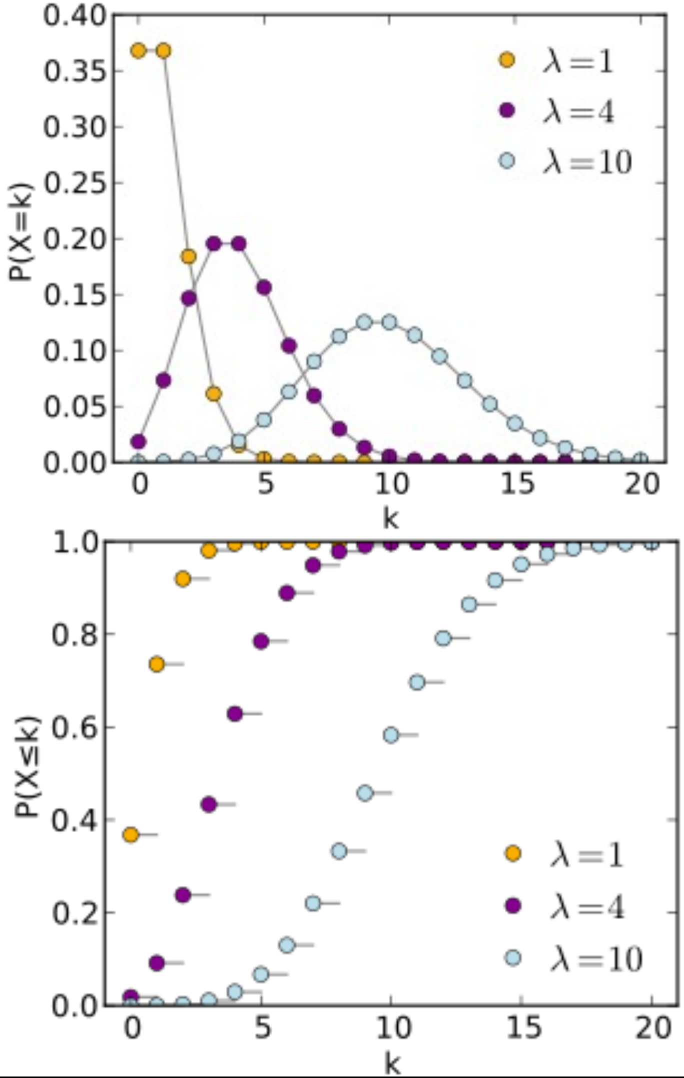
$$\Pr(a < X \leq b) = F(b) - F(a)$$

- The PDF is the derivative of CDF:

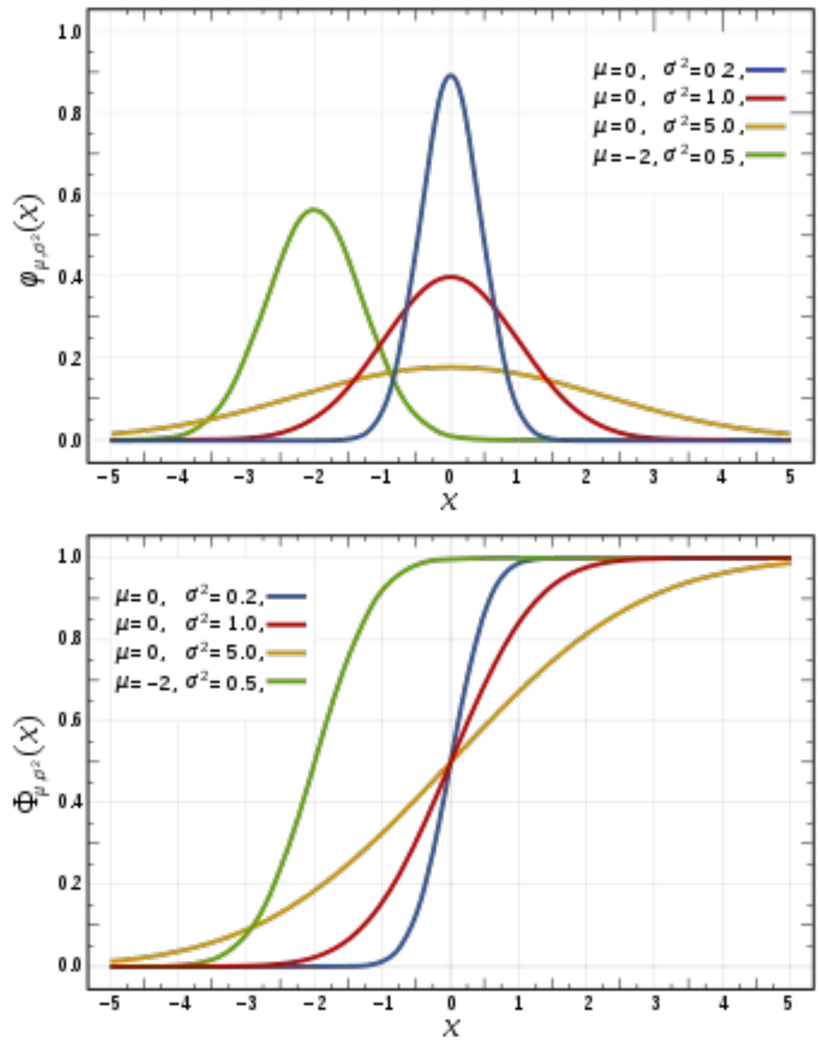
$$p(x) = \frac{dF(x)}{dx}$$

Examples of PMF/PDF and corresponding CDF

Discrete PMF/CDF



Continuous PDF/CDF



Notation: Tilde used to specify distribution of random variable ($\sim$ in LaTeX)

- $X \sim \mathcal{N}(\mu = 0, \sigma = 1)$
  - "Random variable X is **distributed** as a normal distribution with mean of zero and standard deviation of 1."
- $X \sim \text{Uniform}(\alpha, \beta)$
  - "Random variable X is **distributed** as a uniform distribution with parameters α and β (parameters may be unknown)."
- $X \sim P(x)$ or $X \sim \mathbb{P}(x)$
  - "Random variable X is **distributed** as the distribution represented by PMF/PDF $P(x)$ or $\mathbb{P}(x)$."

Notation: Semicolon ";" (or sometimes bar "|") often used to specify parameters

- $p(x\,;\alpha,\beta)=\frac{1}{\beta-\alpha}$
  - "The PDF of X is <u>parameterized by</u> α and β."
  - This is the uniform distribution between α and β.
- $P(x\,;\lambda)$
  - "The PMF of X is <u>parameterized by</u> λ."
- $p(x\mid\mu,\sigma)$
  - "The PDF of X is <u>parameterized by</u> μ and σ."

Big Picture Overview

Introduction to probability

- Motivation
- Interpretation

Random variables

- Outcomes and events

Probability distributions

- PMF, PDF, CDF

Multivariate distributions

- Joint, marginal and conditional distributions
- Chain rule and Bayes rule
- Independence

Expectations of random variables

- Linearity of expectations
- Covariance
- Empirical expectations

<u>Joint distribution</u> of multiple variables

- Joint PDF/PMF is a function of two or more random variables (or a random vector)

- Joint PDF/PMF can be written as:

$$p(x, y), \qquad p(x_1, x_2), \qquad p(\boldsymbol{x})$$

- If $X \in [-1, 1]$ and $Y \in [-1,1]$ is the following a valid PDF?

$$p(x, y) = xy$$

- If $X \in [0, 1]$ and $Y \in [0,1]$ is the following a valid PDF?

$$p(x, y) = 4xy$$

Marginal distribution is sum/integral over other variables

- Example: Height and weight, "What is the distribution of height regardless of weight?"
- Given joint distribution $P(x, y)$ the marginal is:

$$P(x) = \sum_{y \in \mathcal{Y}} P(x, y) \text{ and } P(y) = \sum_{x \in \mathcal{X}} P(x, y)$$

- Given joint distribution $P(x, y)$ the marginal is:

$$p(x) = \int_{y \in \mathcal{Y}} p(x, y) dy \text{ and } p(y) = \int_{x \in \mathcal{X}} p(x, y) dx$$

- Example: $P(x, y) = \begin{bmatrix} y = 1 & 0.1 & 0.4 \\ y = 0 & 0.3 & 0.2 \\ & x = 0 & x = 1 \end{bmatrix}$
- Example: $p(x, y) = 4xy$

<u>Conditional distribution</u> is the distribution *given* some other event

- What is the distribution of weight *given* that a person is x inches tall?
- Conditional density is the joint PDF/PMF *renormalized* by marginal density of event:

$$p(y \mid x) \equiv \frac{p(x, y)}{p(x)}$$

- Example: $P(x, y) = \big[[0.1, 0.4], [0.3, 0.2]\big]$
- Example: $p(x, y) = 4xy$

Note: Conditional and marginal distributions exist for *any set of variables*

- Suppose $p(\boldsymbol{x}) = p(x_1, x_2, x_3, x_4)$

$$p(x_1, x_3) = \int_{x_2, x_4} p(\boldsymbol{x}) dx_2 dx_4$$

$$p(x_1, x_2 | x_3) = \frac{p(x_1, x_2, x_3)}{p(x_3)}$$

$$= \frac{\int_{x_4} p(\boldsymbol{x}) dx_4}{\int_{x_1, x_2, x_4} p(\boldsymbol{x}) dx_1 dx_2 dx_4}$$

<u>Chain rule</u> (or <u>product rule</u>) of probability

- The joint distribution can be written as product of conditional PDFs/PMFs:

$$p(x_1, x_2) = p(x_1)p(x_2|x_1)$$
$$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)$$

- This can be written as:

$$p(x_1, x_2, \ldots, x_d) = \prod_{i=1}^{d} p(x_i|x_1, \ldots, x_{i-1})$$

- Consequence (order doesn't matter):

$$p(x)p(y|x) = p(y)p(x|y)$$

Bayes rule: Enables conversion between one conditional and the other (they are ***different***)

$$p(x|y) = \frac{p(y|x)\, p(x)}{p(y)}$$

(derive on board)

When are $p(x|y)$ and $p(y|x)$ equal?

<u>Independence</u> means that one variable is not affected by the other variable

- Example: Flip two coins, X and Y are 0 or 1.
- Counterexample: Roll dice for number X; then flip that number of coins and count the number of heads Y.
- Formally, PDF/PMF can be written as product of functions that only involve x or y (but not both)
$$p(x, y) = f(x)g(y)$$
- Usually, these are the marginal densities:
$$p(x, y) = p(x)p(y)$$
- Equivalent definition:
$$p(x|y) = p(x) \text{ and } p(y|x) = p(y)$$

Big Picture Overview

Introduction to probability

- Motivation
- Interpretation

Random variables

- Outcomes and events

Probability distributions

- PMF, PDF, CDF

Multivariate distributions

- Joint, marginal and conditional distributions
- Chain rule and Bayes rule
- Independence

Expectations of random variables

- Linearity of expectations
- Covariance
- Empirical expectations

An **expectation** (or **expected value**) of a function of a random variable is the average or mean value with respect to its distribution

- Formal definitions

$$\mathbb{E}_{X\sim P(x)}[f(x)] \equiv \sum_{x\in X} f(x)P(x)$$

$$\mathbb{E}_{X\sim p(x)}[f(x)] \equiv \int_{x\in X} f(x)p(x)dx$$

- Sometimes drop notation to $\mathbb{E}_X[f(x)]$ or just $\mathbb{E}[f(x)]$ if clear from context
- Common: Mean of the distribution $\mu = \mathbb{E}[x]$
- Examples: $P(x) = [0.4,0.3,0.1,0.3]$, $p(x) = 3x^2$

Expectation is a *linear operator* (i.e. splits on summation and scale can come out)

- A linear operator H must satisfy two properties:

$$H\big(f(x) + g(x)\big) = H\big(f(x)\big) + H\big(g(x)\big)$$

$$H\big(\alpha f(x)\big) = \alpha H\big(f(x)\big)$$

- Exercise: Derive for expectations, i.e. $H = \mathbb{E}$

$$\mathbb{E}[af(x) + bg(x)] = a\mathbb{E}[f(x)] + b\mathbb{E}[g(x)]$$

<u>Variance</u> measures the "spread" of a distribution

- Definition

$$\mathrm{Var}[x] = \sigma^2 \equiv \mathbb{E}_X[(x-\mu)^2]$$
$$= \mathbb{E}_X[(x - \mathbb{E}_X[x])^2]$$

- Intuitively, recenter and then measure expected value of $f(x) = x^2$

- **<u>Standard deviation</u>** is square root of variance

$$\sigma = \sqrt{\sigma^2} = \sqrt{\mathbb{E}_X[(x-\mu)^2]}$$

Covariance and correlation measure *linear* relationship between two variables

- Covariance definition

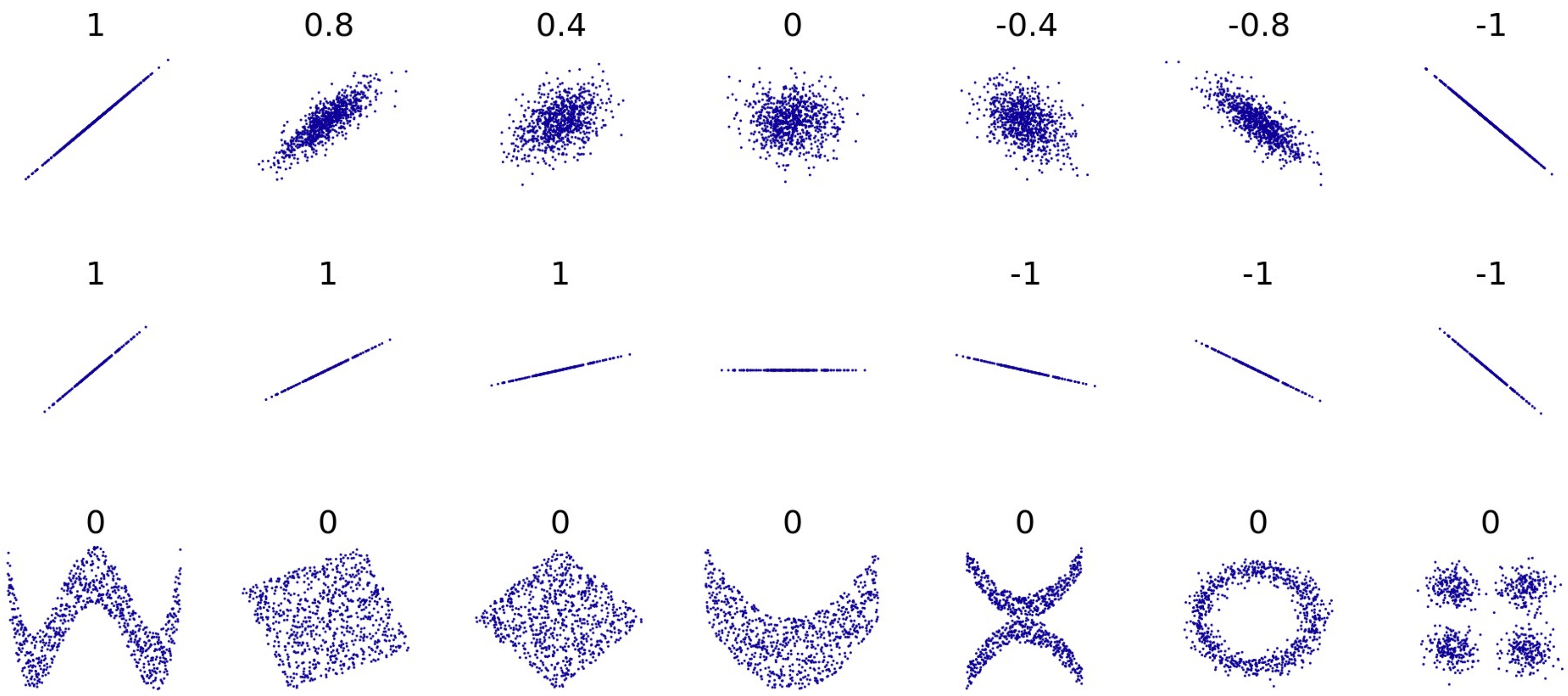
$$\mathrm{Cov}[x,y] \equiv \sigma_{X,Y}^2 \equiv \mathbb{E}_{X,Y}\big[(x-\mu_X)\big(y-\mu_y\big)\big]$$

- Correlation is a normalized covariance

$$\rho_{X,Y} \equiv \frac{\sigma_{X,Y}^2}{\sigma_X \sigma_Y}$$

- Example: $P(x,y) = \begin{array}{cc} \begin{matrix} y=1 \\ y=0 \end{matrix} & \begin{bmatrix} 0.4 & 0.1 \\ 0.1 & 0.4 \end{bmatrix} \\ & \begin{matrix} x=0 & x=1 \end{matrix} \end{array}$
- Solution: $\mu_X = \mu_Y = 0.5, \sigma_X^2 = \sigma_Y^2 = 0.25$
  - $\sigma_{X,Y}^2 = -\frac{3}{20}, \rho_{X,Y} = -\frac{3}{5}$

Uncorrelated ($\rho_{X,Y} = 0$) is **NOT** the same as independence (because only measures *linear* relationship)



Covariance and correlation matrix

are generalizations for vectors

- Covariance matrix has covariance of every pair of random variables

$$\Sigma = \begin{bmatrix} \sigma^2_{X_1,X_1} & \cdots & \sigma^2_{X_1,X_d} \\ \vdots & \ddots & \vdots \\ \sigma^2_{X_d,X_1} & \cdots & \sigma^2_{X_d,X_d} \end{bmatrix}$$

- Matrix has variance along diagonal $\sigma^2_{X_i,X_i} = \sigma^2_{X_i}$
- Correlation matrix is similar but with 1s on diagonal

$$\mathrm{R} = \begin{bmatrix} 1 & \cdots & \rho_{X_1,X_d} \\ \vdots & \ddots & \vdots \\ \rho_{X_d,X_1} & \cdots & 1 \end{bmatrix}$$

- Both matrices are symmetric $\Sigma = \Sigma^T$ and $R = R^T$

The *empirical* expectation is a *sample* version of the population-level expectation

- **Empirical expectation** is the average over i.i.d. samples from the distribution
$$\hat{\mathbb{E}}_n[f(x)] = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$
  - where $x_1, x_2 \dots, x_n$ are i.i.d. samples from the true distribution $p(x)$
- **Law of large numbers** ensures this approaches the population expectation as the number of samples grows
$$\lim_{n\to\infty} \hat{\mathbb{E}}_n[f(x)] \to \mathbb{E}_{p(x)}[f(x)]$$