# Basic knowledge of Probability Theory

Prerequisite for Information Theory

Marco Lops

# Probability Spaces

# Experiment

- An experiment (or trial) is any procedure that can be infinitely repeated and has a well defined set of possible outcomes.
- An experiment is random if it has more than one possible outcome.
- The set of all possible outcomes of an experiment is a sample space, denoted $\Omega$ hereafter.

# Examples

- Coin Tossing:

$$\Omega=\{\text{Head, Tail}\};$$

- Die tossing:

$$\Omega=\{1,2,3,4,5,6\};$$

- Dice pair tossing:

$$\Omega=\{1,2,3,4,5,6\}\times\{1,2,3,4,5,6\}=\{1,2,3,4,5,6\}^2;$$

- Cars lining up at a traffic light:

$$\Omega=\{0,1,2,3,....\};$$

# Events

- An event is any subset of the sample space defined by a proposition.

## Examples

- Die Tossing: $E=\{2,4,6\}\subset\Omega$ ({The outcome is even})
- Dice Tossing: $E=\{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\}\subset\Omega$ ({The dice yield the same outcome})
- Cars at a traffic light: $E=\{0,2,4,6,8\}\subset\Omega$ ({The number of cars lining up is even and smaller than 10})

# Algebra of Events

- Given a set ($\Omega$ in our context), an algebra is such a collection $\mathscr{E}$ of subsets satisfying the conditions:

$$A, B \in \mathscr{E} \rightarrow A \cup B \in \mathscr{E}$$

$$A \in \mathscr{E} \rightarrow \bar{A} = \Omega \setminus A \in \mathscr{E}$$

Consequences

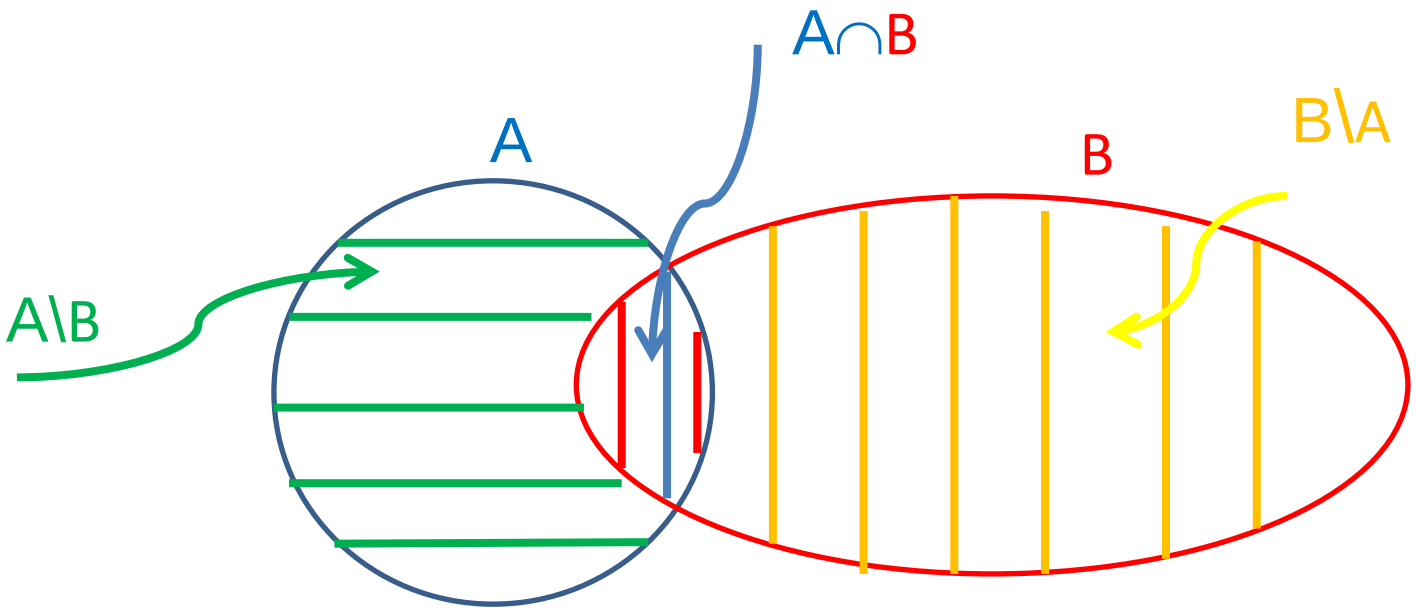
$$A, B \in \mathscr{E} \rightarrow A \cap B \in \mathscr{E}$$

$$\Omega, \varnothing \in \mathscr{E}$$

$$A, B \in \mathscr{E} \rightarrow A \setminus B = A \cap \bar{B} \in \mathscr{E}$$

**Remark:** An algebra is a sigma-algebra if it is closed with respect to countable unions of its subsets.

# Operations on Events




Example: Die Tossing

A={2,4,6}: The result is even  B={5,6}: The result is ≥ 5

A\B={2,4}; B\A={5}; A∩B={6}; A∪B={2,4,5,6}

De Morgan

$$\overline{A \cup B} = \{1,3\} = \{1,3,5\} \cap \{1,2,3,4\} = \overline{A} \cap \overline{B}$$

# Probability law (or measure)

This is any function from a sigma-algebra $\mathscr{E}$ to [0,1], i.e.:

$P: A \in \mathscr{E} \rightarrow P(A) \in [0,1]$ such that:

1. $P(A) \geq 0 \ \forall \ A \in \mathscr{E}$
2. $P(\Omega)=1$
3. $P(A \cap B)=P(A)+P(B) \ \forall \ A,B \in \mathscr{E}: A \cap B=\varnothing$
4. $P(\cup A_n)= \Sigma_n P(A_n)$ if $\forall \ A_i, A_j \in \mathscr{E}: A_i \cap A_j= \varnothing \ \forall \ i \neq j$

These are the Kolmogorov's axioms.

# Properties of probability laws

1. $P(\varnothing)=0$
2. $P(\overline{A})=1-P(A) \quad \forall A \in \mathscr{E}$
3. $P(A \setminus B)=P(A)-P(A \cap B) \quad \forall A, B \in \mathscr{E}$
4. $P(A \cup B)=P(A)+P(B)-P(A \cap B) \quad \forall A, B \in \mathscr{E}$
5. $A \subseteq B \Rightarrow P(A) \leq P(B) \quad \forall A, B \in \mathscr{E}$

# Independent Events

Two events, A, B $\in \mathscr{E}$ are independent iff:

$$P(A \cap B)=P(A)P(B)$$

N events, $A_1, A_2,\ldots,A_N$ are independent iff any subset thereof

$$\{A_{i1},\ldots.,A_{ik}\} \subseteq \{A_1, A_2,\ldots,A_N\}$$

is such that

$$P(A_{i1} \cap A_{i2} \cap \ldots. \cap A_{ik}) = P(A_{i1})P(A_{i2})\ldots.P(A_{ik}) \quad k=2,\ldots., N$$

# Conditional Probability

Let $B \in \mathscr{E}$ be an event with non-zero probability. We define the conditional probability of an event $A \in \mathscr{E}$ given B as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Remarks:

- P(A|B), as A varies in $\mathscr{E}$ and B stays constant, satisfies the Kolmogorov's axioms, i.e. it is a probability law.
- If A and B are independent P(A|B)=P(A).
- P(A|B)=0 A and B are mutually exclusive.

# Alternative definition

- Assume that we undertake N repeated (independent) trials of a given experiment;
- Assume that an event A occurs $N_A$ times;
- Assume that another event B occurs $N_B$ times.
- In $N_{AB}$ trials both A and B occur.

$$P(A) = \lim_{N\to\infty} \frac{N_A}{N}$$

$$P(A|B) = \lim_{N\to\infty} \frac{N_{AB}}{N_B}$$

# Counting

- If an experiment has a finite number of outcomes, the probability of an event A can be computed as

$$P(A) = \frac{\#\text{ cases where A occurs}}{\#\text{ possible outcomes}}$$

- It is thus important to develop effective counting techniques for both the numerator and the denominator.

# Combinatorial Analysis

Given *k* sets $\Omega_1,..,\Omega_k$ with cardinalities $|\Omega_i|=n_i$, we have the basic formula of combinatorial analysis

$$|\Omega_1 \times \Omega_2 \times \ldots \times \Omega_k| = \prod_{i=1}^{k} |\Omega_i| = \prod_{i=1}^{k} n_i$$

where $\Omega_1 \times \Omega_2 \times \ldots \times \Omega_k$ denotes the *k*-fold set product, i.e. the set of (ordered) *k*-tuples whose first element is drawn from $\Omega_1$, the second from $\Omega_2$, the *k*-th from $\Omega_k$.

# Combinatorial Analysis: Permutations

Assume we draw *k* elements from one and the same set of cardinality *n*. We have two cases:

a) Drawing with replacements (element repetition is allowed).
b) Drawing without replacements (i.e., no repetition allowed).

Case a: $|\Omega_1|=|\Omega_2|=...=|\Omega_k|=n$;
Permutation number=$n^k$.

Case b: $|\Omega_1|=n$; $|\Omega_2|=n-1$;...$|\Omega_k|=n-k+1$
Permutation number= $n(n-1)...(n-k+1)= \frac{n!}{(n-k)!}$

Special case: $n=k$
Permutation number of *k* distinct elements= $k!$

# Combinatorial Analysis: Combinations

Permutations of *k* items drawn from a set with cardinality *n* where repetitions are not allowed and order is immaterial are defined combinations.

Thus *k!* permutations with no repetitions are equivalent to one combination, whereby

Number of combinations= $C_{n,k} = \frac{n!}{k!(n-k)!} = \left( \begin{array}{c} n \\ k \end{array} \right)$

Example: the number of combinations of cards taken 5 at a time from a deck is

$$\left( \begin{array}{c} 52 \\ 5 \end{array} \right) = \frac{52!}{5!47!} = 2598960$$

# Fundamental laws: Bayes

Since

$$P(A|B)\,P(B) = P(A \bigcap B) = P(B|A)\,P(A)$$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

# Fundamental laws: The law of total probability

Let A1, A2, ...,An,.. be a countable partition of Ω, i.e.:

$$\Omega = \bigcup A_n \qquad A_i \bigcap A_j = \emptyset \ , \quad \forall\, i \neq j$$

$$P(E) = \sum_n P\left(E \middle| A_n\right) P\left(A_n\right) \ , \qquad \forall\, E \in \mathcal{E}$$

# Random Variables

# Random variables: Definitions

A random variable X is a measurable map between a probability space and a subset of the real set, i.e.:

$(\Omega, \mathcal{E}, P)$ → Parent probability space

RV: Map definition

$$X : \omega \in \Omega \longrightarrow X(\omega) \in \mathcal{A}_X \subseteq R$$

$X^{-1}(A) \in \mathcal{E} \quad \forall A \in \mathcal{E}$ → Measurability

# Discrete versus Continuous RV's

An RV X is discrete if $\mathcal{A}_X = X(\Omega)$ is finite or countable.

An RV X is continuous if $\mathcal{A}_X = X(\Omega)$ has uncountably many points.

$\mathcal{A}_X$ is in both cases referred to as alphabet of the RV X.

# The Cumulative Distribution Function (CDF)

Definition:

$$F_X: x \in R \rightarrow F_X(x) = P\{X \leq x\} \in [0,1]$$

Properties:

1) $F_X(-\infty)=0 \quad F_X(+\infty)=1$
2) $F_X(x)$ non-decreasing in x;
3) $F_X(x)$ is a right-continuous function.

Remark: for discrete RV's the CDF is a staircase function

# Probability Density Functions (pdf)

Formally, a pdf is the derivative of a probability measure with respect to the standard measure, i.e. the Lebesgue and the counting measure for continuous and discrete RV respectively.

In practice we have:

a) X discrete on $\mathcal{A}_X \rightarrow$ pdf=p(x)=$f_X$(x)=P{X=x}

b) X continuous on $\mathcal{A}_X$:

$$f_X(x) = \frac{dF_X(x)}{dx}$$

Remark: densities of discrete variables are also known as probability mass functions (pmf)

# Properties of densities

## Discrete RV's

1) $0 \leq f_X(x) \leq 1 \ \forall \ x \in \mathcal{A}_X$;
2) $\Sigma_x \ f_X(x) = 1$

## Continuous RV's

1) $f_X(x) \geq 0 \ \forall \ x \in \mathcal{A}_X$;
2) $\int f_X(x) \, dx = 1$

# Noticeable discrete RV's

Bernoulli counting: $X \in \{0, 1, \ldots, n\}$ $\quad 0 \leq p \leq 1$

$$P\{X = k\} = f_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Poisson: $X \in \{0, 1, 2, \ldots\} = N_0$ $\quad \lambda > 0$

$$P\{X = k\} = f_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Geometric $X \in \{1, 2, \ldots\} = N$ $\quad 0 \leq p \leq 1$

$$P\{X = k\} = f_X(k) = (1-p)^{k-1} p$$

# Noticeable continuous RV's

Uniform in $[a,b]$, $X \sim \mathcal{U}(a,b)$

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{elsewhere} \end{cases}$$

Gaussian with parameters $\mu \in R$ and $\sigma > 0$, $X \sim \mathcal{N}(\mu,\sigma)$

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \qquad x \in R$$

Exponential with parameter $\lambda > 0$, $X \sim \mathcal{E}(\lambda)$

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & x \ge 0 \\ 0 & x < 0 \end{cases}$$

# Random Vectors

$$(\Omega, \mathcal{E}, P) \longrightarrow \text{Probability Space}$$

$$\mathbf{X} : \omega \in \Omega \longrightarrow \mathbf{X} = (X_1, \ldots, X_n) \in R^n$$

$$\text{CDF:} \quad F_{\mathbf{X}}(\mathbf{x}) = F_{\mathbf{X}}(x_1, \ldots, x_n) = \Pr\{X_1 \leq x_1, \ldots, X_n \leq x_n\}$$

$$\text{pdf:} \quad f_{\mathbf{X}}(\mathbf{x}) = \begin{cases} \Pr\{X_1 = x_1, \ldots, X_n = x_n\} & \text{discrete case} \\ \dfrac{\partial^n F_{\mathbf{x}}(\mathbf{x})}{\partial x_1 \ldots \partial x_n} & \text{continuous case} \end{cases}$$

# Conditional densities

Consider first a bi-dimensional vector $\mathbf{X} = [X_1, X_2]^T$

Conditional density

$$f_{X_2|X_1}(x_2|x_1) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)}$$

More generally, given two random vectors $\mathbf{X}$ and $\mathbf{Y}$:

Conditional density

$$f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = \frac{f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})}{f_{\mathbf{Y}}(\mathbf{y})}$$

# Transformations

Let $X \sim f_X(x)$

Let $Y = g(X)$

How can we characterize $Y$ ?

a) g(·) invertible for all points of the alphabet of X;

b) g(·) non-invertible for all points of the alphabet of X;

# Invertible transformations and continuous variables

$$F_Y(y) = \begin{cases} F_X\left[g^{-1}(y)\right] & g(\cdot) \text{ increasing} \\ 1 - F_X\left[g^{-1}(y)\right] & g(\cdot) \text{ decreasing} \end{cases}$$

$$f_Y(y) = \frac{f_X\left[g^{-1}(y)\right]}{\left|g'\left[g^{-1}(y)\right]\right|}$$

# Non-Invertible transformations and continuous variables

Let $\{x_i = x_i(y)\}_{i=1}^{n}$ be the solutions of the equation $y = g(x)$

$$f_Y(y) = \sum_{i=1}^{n} \left. \frac{f_X(x)}{|g'(x)|} \right|_{x=x_i(y)}$$

# Moments of a Random Variable

# Moments of a random variable

Non-central moment of order $m$

$$\mathrm{E}\left[X^m\right] = \begin{cases} \sum_x x^m f_X(x) & X \text{ discrete} \\ \int x^m f_X(x)\, dx & X \text{ continuous} \end{cases}$$

Central moment of order $m$

$$\mathrm{E}\left[(X - \mathrm{E}\left[X\right])^m\right] = \begin{cases} \sum_x \left(x - \mathrm{E}\left[X\right]\right)^m f_X(x) & X \text{ discrete} \\ \int \left(x - \mathrm{E}\left[X\right]\right)^m f_X(x)\, dx & X \text{ continuous} \end{cases}$$

# Mean, Variance and Standard Deviation

mean: $\mu_X = \mathrm{E}[X]$ Variance: $\mathrm{E}\left[(X-\mu_X)^2\right] = \sigma_X^2$

$\sigma_X$: Standard Deviation

Tchebyshev Inequalities:

$$\Pr\{|X-\mu_X| > b\} = 1 - \Pr\{\mu_X - b \le X \le \mu_X + b\} \le \frac{\sigma_X^2}{b^2}$$

$$\Pr\{|X-\mu_X| > k\sigma_X\} = 1 - \Pr\{\mu_X - k\sigma_X \le X \le \mu_X + k\sigma_X\} \le \frac{1}{k^2}$$

Probability of observing values spaced from the mean more than $k\sigma_X$

# Moments of Random Vectors

Let $\mathbf{X} = [X_1, \ldots, X_n]^T$ be an RV, $\mathbf{X} \sim f_{\mathbf{X}}(\mathrm{x})$

$$\text{Mean: } \mathrm{E}[\mathbf{X}] = \begin{cases} \sum_{\mathrm{x}} \mathrm{x} f_{\mathbf{X}}(\mathrm{x}) & \mathbf{X} \text{ discrete} \\ \int \mathrm{x} f_{\mathbf{X}}(\mathrm{x})\, d\mathrm{x} & \mathbf{X} \text{ continuous} \end{cases}$$

$$\text{Covariance Matrix: } \mathbf{C} = \mathrm{E}\left[(\mathbf{X} - \mathrm{E}[\mathbf{X}])(\mathbf{X} - \mathrm{E}[\mathbf{X}])^T\right]$$

Remark: $\mathbf{C}$ is non-negative definite

# Gaussian Vectors

Let $\mathbf{X} \in R^n$ a Random Vector with

$$\mathrm{E}[\mathbf{X}] = \mathbf{m}\,, \qquad \mathrm{E}\left[(\mathbf{X}-\mathbf{m})(\mathbf{X}-\mathbf{m})^T\right] = \mathbf{C}$$

$\mathbf{X}$ is a Gaussian vector (short-hand notation: $\mathbf{X} \sim \mathcal{N}(\mathbf{m}, \mathbf{C})$) iff

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^n |\mathbf{C}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})\right]$$

Example $(n = 2)$

$$\mathbf{m} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} \qquad \mathbf{C} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \qquad \rho = \frac{\mathrm{E}[(X_1-m_1)(X_2-m_2)]}{\sigma_1\sigma_2}$$

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{(2\pi)\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}} \exp\left[-\frac{\sigma_2^2(x_1-m_1)^2+\sigma_1^2(x_2-m_2)^2-2\sigma_1\sigma_2\rho(x_1-m_1)(x_2-m_2)}{2\sigma_1^2\sigma_2^2(1-\rho^2)}\right]$$

# Properties of Gaussian vectors

Incorrelation implies independence. Indeed:

$$\mathbf{C} = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix} \longrightarrow f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x_i - m_i)^2}{2\sigma_i^2}\right]$$

Closure under linear transformations

If $\mathbf{A}$ is a full-rank (tall or square) matrix and $\mathbf{X} \sim \mathcal{N}(\mathbf{m}, \mathbf{C})$, then

$$\mathbf{Y} = \mathbf{AX} + \mathbf{k} \sim \mathcal{N}\left(\mathbf{Am} + \mathbf{k}, \mathbf{ACA}^T\right)$$