1 Probability

1.1 Basic concepts and results

- Random experiment: set of all possible outcomes is known; impossible to predict the actual outcome of the experiment
- Ω — sample space, contains all possible outcomes of the experiment
- Event: subset of Ω; we say that $A \subset \Omega$ has occurred iff the outcome of the experiment is an element of A
- Formally, the family of events forms a σ-algebra of subsets of Ω that we denote by $\mathcal{A}$

- Probability measure: $P : \mathcal{A} \to \mathbb{R}$ satisfying Kolmogorov's axioms
  - $P(A) \geq 0$
  - $P(\Omega) = 1$
  - if $A_1, A_2, \ldots$ is a collection of pairwise incompatible events, i.e., if $A_i \cap A_j = \varnothing$, $i \neq j$, then $P(\cup_i A_i) = \sum_i P(A_i)$
- It's easy to prove various properties:
  - $P(\varnothing) = 0$
  - $A \subset B \Rightarrow P(A) \leq P(B)$
  - $0 \leq P(A) \leq 1$
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - $P(\bar{A}) = 1 - P(A)$
  - $P(A - B) = P(A \cap \bar{B}) = P(A) - P(A \cap B)$

- **Conditional probability**: if $P(B) > 0$,

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

- Let $\{A_1, A_2, \ldots\}$ denote a partition of Ω: $\cup_i A_i = \Omega$; $A_i \cap A_j = \varnothing$, $i \neq j$. Then, if $P(A_i) > 0$ for all i
  - **Total probability theorem**

$$P(B) = \sum_i P(B \mid A_i)\ P(A_i)$$

  - **Bayes' theorem** (if $P(B) > 0$)

$$P(A_j \mid B) = \frac{P(B \mid A_j)P(A_j)}{\sum_i P(B \mid A_i)P(A_i)}$$

- Remark: $P(\cdot \mid B)$ is a probability measure

- **Random variable**: function defined in Ω and taking values in $\mathbb{R}$

$$X : \Omega \to \mathbb{R}$$
$$\omega \mapsto X(\omega) = x$$

- A random variable X induces a probability measure in $\mathbb{R}$ that we denote by P_X: if $B \subset \mathbb{R}$,

$$P_X(B) = P(A), \quad \text{where } A = X^{-1}(B) = \{\omega \in \Omega : \ X(\omega) \in B\}$$

- Formally, there must be a σ-algebra of subsets of $\mathbb{R}$, $\mathcal{B}$, and we have to verify that for every set $B \in \mathcal{B}$ we have $X^{-1}(B) \in \mathcal{A}$

- Typically, $\mathcal{B}$ is the so-called Borel σ-algebra, and it suffices to make sure that X satisfies

$$X^{-1}(]-\infty, x]) \in \mathcal{A} \qquad \forall x \in \mathbb{R}$$

- **Distribution function of a r.v.** X: for all $x \in \mathbb{R}$

$$F_X(x) = P_X(]-\infty, x]) = P(X \leq x)$$

- It suffices to know $F_X(\cdot)$ to be able to compute $P_X(B)$ for all $B \in \mathcal{B}$ (Borel σ-algebra)

- For all $a > b$,

$$P(a < X \leq b) = F_X(b) - F_X(a)$$

- $F_X(-\infty) = 0$; $F_X(+\infty) = 1$

- F_X is right-continuous and non-decreasing

- Any function F that is non-decreasing, right-continuous and satisfies $F(+\infty) = 1$ and $F(-\infty) = 0$ is a distribution function

- The set of points at which F_X is discontinuous

$$D_X = \{x : F_X(x) - \lim_{y \nearrow x} F_X(y) > 0\} \ ,$$

is either finite or countable.

- **Discrete r.v.**: X is discrete if D_X is such that $P_X(D_X) = 1$
- The probability mass function of X is defined as

$$f_X(x) = F_X(x) - \lim_{y \nearrow x} F_X(y) = \begin{cases} P(X = x) & \text{if } x \in D_X \\ 0 & \text{otherwise} \end{cases}$$

- Any f satisfying
  - $f(x) \geq 0$ for all x
  - $f(x) > 0$ iff $x \in D$, where $D \subset \mathbb{R}$ is finite or countable
  - $\sum_{x \in D} f(x) = 1$

  is a probability mass function
- For any event $B \subset \mathbb{R}$,

$$P(X \in B) = \sum_{x \in B \cap D_X} f_X(x)$$

- $F_X(x) = \sum_{y \leq x} f_X(y)$

- Discrete distributions: Bernoulli, binomial, Poisson, [geometric, negative binomial, multinomial, hypergeometric, etc.]

- **Continuous r.v.**: X is continuous if $P_X(D_X) = 0$ and if additionally there is f_X such that for all $x \in \mathbb{R}$
  - $f_X(x) \geq 0$
  - $F_X(x) = \int_{-\infty}^{x} f_X(u)\, du$
- f_X is known as the probability density function of X
- At the points where F_X is differentiable, we have $F'_X(x) = f_X(x)$
- Any f satisfying
  - $f(x) \geq 0$ for all x
  - $\int_{-\infty}^{+\infty} f(x)\, dx = 1$

  is a probability density function
- Continuous distributions: uniform, exponential, gamma, chi-squared, normal, t-'Student', F-Snedcor, [beta, Pareto, Weibull, log-normal, etc.]

- **Funţions of a r.v.**: Let X be a r.v. and $Y = h(X)$ where $h:\ \mathbb{R} \to \mathbb{R}$
- We know the distribution of X, how do we determine the distribution of Y?
- Example: Suppose $X \sim \mathrm{N}(0,1)$; determine the distribution of $Y = X^2$.

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y})\ , \quad y > 0 \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y})\ , \quad y > 0 \end{aligned}$$

hence, taking derivatives with respect to y,

$$f_Y(y) = \frac{1}{2}\ y^{-1/2}\ f_X(\sqrt{y}) - \frac{1}{2}\ (-y^{-1/2})\ f_X(-\sqrt{y}) = y^{-1/2} f_X(\sqrt{y})$$

where we have used the fact $f_X(x) = f_X(-x)$. Recalling that $f_X(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$ and that $\Gamma(1/2) = \sqrt{\pi}$, it is easy to conclude that

$$f_Y(y) = \frac{1}{2^{1/2}\ \Gamma(1/2)}\ y^{1/2-1}\ e^{-y/2}\ , \quad y > 0$$

that is, $Y \sim \chi^2(1)$.

- In general, if $X = g(Y)$ with g invertible and differentiable, and X continuous, we have

$$f_Y(y) = |g'(y)|\ f_X(g(y))$$

- **Expected value**: Let $Y = h(X)$. The expected value of Y is defined by

$$E[Y] = \begin{cases} \sum_x h(x)\ f_X(x) & \text{if } X \text{ discrete} \\ \int_{-\infty}^{+\infty} h(x)\ f_X(x)\ dx & \text{if } X \text{ continuous} \end{cases}$$

- (Formally, we must additionally verify that the integral or series are absolutely convergent)

- $E[Y]$ may not exist

- Two methods of computing $E[Y]$ with $Y = h(X)$: use the definition above (method 1), or first of obtain the distribution of Y and compute (method 2)

$$E[Y] = \begin{cases} \sum_y y\ f_Y(y) & \text{if } Y \text{ discrete} \\ \int_{-\infty}^{+\infty} y\ f_Y(y)\ dy & \text{if } Y \text{ continuous} \end{cases}$$

  The two methods are equivalent.

- Example: Let X be a r.v. with pdf $f_X(x) = 3x^{-4}$, $x > 1$.
  Determine $E[X^2]$ using both methods.

- **Raw moment of order** k

$$\mu'_k = E[X^k]$$

- **Central moment of order** k

$$\mu_k = E[(X-\mu)^k], \qquad \mu = E[X]$$

- If μ'_k exists, then μ'_r exists for all $r \leq k$; similarly for μ_k

- Important moments: $\mu'_1 = E[X]$, measure of location; $\mu_2 = \mathrm{Var}(X)$, measure of dispersion

- **Moment generating function**:

$$M_X(s) = E[e^{sX}]$$

whenever the expectation exists for s in a neighborhood of the origin

- If $M_X(s)$ exists, then X has moments of all orders and

$$M^{(k)}(0) = E[X^k]$$

- The moment generating function (when it exists) identifies the probability distribution

Properties:

- $E[h_1(X) + h_2(X)] = E[h_1(X)] + E[h_2(X)]$
- If $c \in \mathbb{R}$, then $E[cX] = cE[X]$; $E[c] = c$
- If $c \in \mathbb{R}$, $\mathrm{Var}(cX + b) = c^2 \mathrm{Var}(X)$
- $\mathrm{Var}(X) = E[X^2] - (E[X])^2$
- $\mathrm{Var}(X) \geq 0$; $\mathrm{Var}(X) = 0 \Leftrightarrow P(X = c) = 1$ for some $c \in \mathbb{R}$

- **Bivariate random variables**

$$(X,Y) : \Omega \to \mathbb{R}^2$$
$$\omega \mapsto (X(\omega), Y(\omega)) = (x, y)$$

- If (X,Y) discrete, we define the joint probability mass function

$$f(x,y) = P(X = x, Y = y)$$

- If (X,Y) continuous, then there exists the joint probability density function, $f(x,y)$, such that for all $(x,y) \in \mathbb{R}^2$
  - $f(x,y) \geq 0$
  - $F(x,y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v)\, dv\, du$
- Whenever the derivatives exist $\frac{\partial^2}{\partial x \partial y} F(x,y) = f(x,y)$

- **Marginal distributions**

$$f_X(x) = \begin{cases} \sum_y f(x,y) & \text{if } (X,Y) \text{ discrete} \\ \int_{-\infty}^{+\infty} f(x,y)\, dy & \text{if } (X,Y) \text{ continuous} \end{cases}$$

- **Expectation of** $Z = h(X,Y)$:

$$E[Z] = \begin{cases} \sum_x \sum_y h(x,y)\ f(x,y) & \text{if } (X,Y) \text{ discrete} \\ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h(x,y)\ f(x,y)\ dy\ dx & \text{if } (X,Y) \text{ continuous} \end{cases}$$

- **Conditional distributions:**

$$f_{X|Y=y}(x) = \frac{f(x,y)}{f_Y(y)}\ , \quad y \text{ fixed} : f_Y(y) > 0$$

- **Conditional expectations**: $Z = h(X,Y)$

$$E[Z \mid Y = y] = \begin{cases} \sum_x h(x,y)\ f_{X|Y=y}(x) & \text{if } (X,Y) \text{ discrete} \\ \int_{-\infty}^{+\infty} h(x,y)\ f_{X|Y=y}(x)\ dx & \text{if } (X,Y) \text{ continuous} \end{cases}$$

- **Raw moment of order** (r, s)

$$\mu'_{(r,s)} = E[X^r \; Y^s]$$

- **Central moment of order** (r, s)

$$\mu_{(r,s)} = E[(X - \mu_X)^r \; (Y - \mu_Y)^s]$$

- **Covariance**: $\mathrm{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = \mu_{(1,1)}$

  Properties:

  - $\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$
  - $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
  - $\mathrm{Cov}(cX, Y) = c \; \mathrm{Cov}(X, Y)$, $c \in \mathbb{R}$
  - $\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$
  - $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\mathrm{Cov}(X, Y)$

- **Law of the iterated expectation**: if $Z = h(X, Y)$ then

$$E[Z] = E_X[E[Z \mid X]]$$

- For the marginal variance:

$$\mathrm{Var}(Y) = \mathrm{Var}_X(E[Y \mid X]) + E_X[\mathrm{Var}(Y \mid X)]$$

- $E[h(X)\ Y \mid X = x] = h(x)\ E[Y \mid X = x]$
- $\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, E[Y \mid X])$

- **(Stochastic) independence** X and Y are (stochastically) independent iff

$$\forall (x, y) \in \mathbb{R}^2 \quad f(x, y) = f_X(x)\ f_Y(y)$$

- If X and Y are independent, then
  - $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$
  - $M_{X+Y}(s) = M_X(s)\ M_Y(s)$
  - $\mathrm{Cov}(X, Y) = 0$
  - $E[X^r Y^s] = E[X^r]\ E[Y^s]$
  - $E[Y \mid X = x] = E[Y]$; $E[X \mid Y = y] = E[X]$
  - $f_{X|Y=y}(x) = f_X(x)$; $f_{Y|X=x}(x) = f_Y(y)$

Other forms of independence

- **Mean independence**: Y is mean independent of X iff $E[Y \mid X = x]$ does not depend on x (for all x)
- **Uncorrelatedness**: X and Y are uncorrelated iff $\mathrm{Cov}(X, Y) = 0$

Results

- If X and Y are stochastically independent, then Y is mean-independent of X (and X is mean-independent of Y)
- If Y is mean-independent of X, then X and Y are uncorrelated
- The converses are not true

More results

- If Y is uncorrelated with X, then $E[XY] = E[X]E[Y]$
- If Y is mean-independent of X, then $E[X^k Y] = E[X^k]E[Y]$ for all k

- If Y and X are stochastically independent, then $E[X^kY^r] = E[X^k]E[X^r]$ for all k, r