National Ph.D. Program in *Artificial Intelligence for Society*

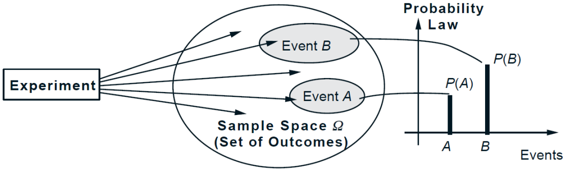
Statistics for Machine Learning

Lesson 02 - Random Variables

Andrea Pugnana, Salvatore Ruggieri

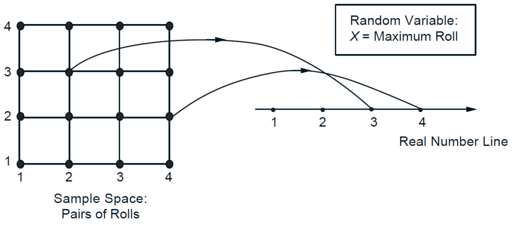
Department of Computer Science
University of Pisa, Italy
andrea.pugnana@di.unipi.it salvatore.ruggieri@unipi.it

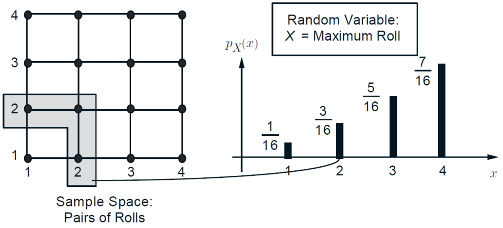
Experiments



- **Experiment**: roll two independent 4 sided die.
- We are interested in probability of the *maximum of the two rolls*.
- Modeling so far
  - $\Omega = \{1,2,3,4\} \times \{1,2,3,4\} = \{(1,1),(1,2),(1,3),(1,4),(2,1),\ldots,(4,4)\}$
  - $A = \{\text{maximum roll is } 2\} = \{(1,2),(2,1),(2,2)\}$
  - $P(A) = P(\{(1,2),(2,1),(2,2)\}) = 3/16$

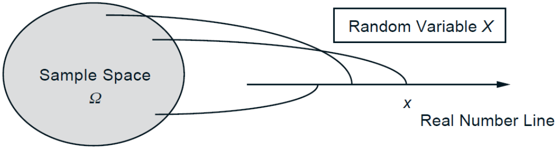
Random variables





- Modeling $X : \Omega \to \mathbb{R}$
  - $X((a,b)) = max(a,b)$
  - $A = \{\text{maximum roll is } 2\} = \{(a,b) \in \Omega \mid X((a,b)) = 2\} = X^{-1}(2)$
  - $P(A) = P(X^{-1}(2)) = {}^{3}/_{16}$
  - We write $P_X(X = 2) \overset{\text{def}}{=} P(X^{-1}(2))$ *[Induced probability]*

(Discrete) Random variables



- A random variable is a function $X : \Omega \to \mathbb{R}$
  - it transforms Ω into a more tangible sample space $\mathbb{R}$
    - from (a, b) to $min(a, b)$
  - it decouples the details of a specific Ω from the probability of events of interest
    - from $\Omega = \{\text{H, T}\}$ or $\Omega = \{\text{good, bad}\}$ or $\Omega = \ldots$ to $\{0, 1\}$
  - it is not 'random' nor 'variable'

> DEFINITION. Let Ω be a sample space. A *discrete random variable* is a function $X : \Omega \to \mathbb{R}$ that takes on a finite number of values $a_1, a_2, \ldots, a_n$ or an infinite number of values $a_1, a_2, \ldots$.

Probability Mass Function (PMF)

DEFINITION. The *probability mass function* p of a discrete random variable X is the function $p : \mathbb{R} \to [0, 1]$, defined by

$$p(a) = \mathrm{P}(X = a) \quad \text{for } -\infty < a < \infty.$$

- Support or domain of X is $dom(X) = \{a \in \mathbb{R} \mid P(X = a) > 0\} = \{a_1, a_2, \ldots a_i, \ldots\}$
  - $p(a_i) > 0$ for $i = 1, 2, \ldots$
  - $p(a_1) + p(a_2) + \ldots = 1$
  - $p(a) = 0$ if $a \notin dom(X)$

Cumulative Distribution Function (CDF) and CCDF

DEFINITION. The *distribution function* F of a random variable X is the function $F : \mathbb{R} \to [0,1]$, defined by

$$F(a) = \mathrm{P}(X \leq a) \quad \text{for } -\infty < a < \infty.$$

- $F(a) = P(X \in \{a_i \mid a_i \leq a\}) = P(X \leq a) = \sum_{a_i \leq a} p(a_i)$
- if $a \leq b$ then $F(a) \leq F(b)$ *[Non-decreasing]*
- $P(a < X \leq b) = F(b) - F(a) = \sum_{a < a_i \leq b} p(a_i)$

Complementary cumulative distribution function (CCDF)

$$\bar{F}(a) = P(X > a) = 1 - P(X \leq a) = 1 - F(a)$$

- $\bar{F}(a) = P(X \in \{a_i \mid a_i > a\}) = P(X > a) = \sum_{a_i > a} p(a_i)$

$X \sim U(m, M)$

Uniform discrete distribution

A discrete random variable X has the *uniform distribution* with parameters $m, M \in \mathbb{Z}$ such that $m \leq M$, if its pmf is given by

$$p(a) = \frac{1}{M - m + 1} \quad \text{for } a = m, m+1, \ldots, M$$

We denote this distribution by $U(m, M)$.

- **Intuition:** all integers in $[m, M]$ have equal chances of being observed.

$$F(a) = \frac{\lfloor a \rfloor - m + 1}{M - m + 1} \quad \text{for } m \leq a \leq M$$

- **Example:** classic 6-faces (fair) die $(m = 1,\ M = 6)$

$X \sim Ber(p)$

DEFINITION. A discrete random variable X has a *Bernoulli distribution* with parameter p, where $0 \leq p \leq 1$, if its probability mass function is given by

$$p_X(1) = \mathrm{P}(X = 1) = p \quad \text{and} \quad p_X(0) = \mathrm{P}(X = 0) = 1 - p.$$

We denote this distribution by $Ber(p)$.

- X models success/failure
- **Example:** getting head (H,T) when tossing a coin, testing for a disease (infected, not infected), membership in a set (member, non-member), etc.
- p_X is the *pmf* (to distinguish from parameter p)
- Alternative definition: $p_X(a) = p^a \cdot (1 - p)^{1-a}$ for $a \in \{0, 1\}$

Identically distributed (i.d.) random variables

Identically distributed random variables

Two random variables X and Y are said *identically distributed* (in symbols, $X \sim Y$), if $F_X = F_Y$, i.e.,

$$F_X(a) = F_Y(a) \quad \text{for } a \in \mathbb{R}$$

- Identically distributed does **not** mean equal
- Toss a fair coin
  - let X be 1 for H and 0 for T
  - let Y be $1 - X$
- $X \sim Ber(0.5)$ and $Y \sim Ber(0.5)$
- Thus, $X \sim Y$ but are clearly always different.

Joint p.m.f.

- For a same Ω, several random variables can be defined
  - Random variables related to the same experiment often influence one another
  - $\Omega = \{(i,j) \mid i,j \in 1,\ldots,6\}$ rolls of two dies
    - $X((i,j)) = i + j$ and $Y((i,j)) = max(i,j)$
    - $P(X = 4, Y = 3) = P(X^{-1}(4) \cap Y^{-1}(3)) = P(\{(3,1),(1,3)\}) = 2/36$
- In general:

$$P_{XY}(X = a, Y = b) = P(\{\omega \in \Omega \mid X(\omega) = a \text{ and } Y(\omega) = b\}) = P(X^{-1}(a) \cap Y^{-1}(b))$$

DEFINITION. The *joint probability mass function* p of two discrete random variables X and Y is the function $p : \mathbb{R}^2 \rightarrow [0,1]$, defined by

$$p(a,b) = \mathrm{P}(X = a, Y = b) \quad \text{for } -\infty < a, b < \infty.$$

Joint and marginal p.m.f.

- **Joint distribution function** $F : \mathbb{R} \times \mathbb{R} \to [0,1]$:

$$F_{XY}(a,b) = P(X \leq a, Y \leq b) = \sum_{a_i \leq a, b_i \leq b} p(a_i, b_i)$$

- By generalized additivity, the **marginal p.m.f.**'s can be derived: *[Tabular method]*

$$p_X(a) = P_X(X = a) = \sum_b P_{XY}(X = a, Y = b) \quad p_Y(b) = P_Y(Y = b) = \sum_a P_{XY}(X = a, Y = b)$$

  and the marginal distribution function of X as:

$$F_X(a) = P_X(X \leq a) = \lim_{b \to \infty} F_{XY}(a,b) \qquad F_Y(b) = P_Y(Y \leq b) = \lim_{a \to \infty} F_{XY}(a,b)$$

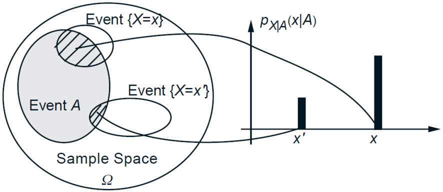
- Deriving the joint p.m.f. from marginal p.m.f.'s is not always possible!
- Deriving the joint p.m.f. from marginal p.m.f.'s is possible for independent events!
  - $\Omega = \{1,2,3,4\} \times \{1,2,3,4\}$, $X((a,b)) = a$, $Y((a,b)) = b$
  - $P(X = 1, Y = 2) = 1/16 = 1/4 \cdot 1/4 = P(X = 1) \cdot P(Y = 2)$

Conditional distribution

Conditional distribution

Consider the joint distribution P_{XY} of X and Y. The conditional distribution of X given $Y \in B$ with $P_Y(Y \in B) > 0$, is the function $F_{X|Y\in B} : \mathbb{R} \to [0,1]$:

$$F_{X|Y\in B}(a) = P_{X|Y}(X \leq a | Y \in B) = \frac{P_{XY}(X \leq a, Y \in B)}{P_Y(Y \in B)} \quad \text{for } -\infty < a < \infty$$



- Distribution of X after knowing $Y \in B$.
- Chain rule: $P_{XY}(X \leq a, Y \in B) = P_{X|Y}(X \leq a | Y \in B) P_Y(Y \in B)$
- What if the distribution does not change w.r.t. the prior P_X?

Independence of two random variables

Independence $X \perp\!\!\!\perp Y$

A random variable X is independent from a random variable Y, if for all $P_Y(Y \leq b) > 0$:

$$P_{X|Y}(X \leq a | Y \leq b) = P_X(X \leq a) \quad \text{for } -\infty < a < \infty$$

- Properties
  - $X \perp\!\!\!\perp Y$ iff $P_{XY}(X \leq a, Y \leq b) = P_X(X \leq a) \cdot P_Y(Y \leq b) \quad \text{for } -\infty < a, b < \infty$
  - $X \perp\!\!\!\perp Y$ iff $Y \perp\!\!\!\perp X$ *[Symmetry]*
- For X, Y **discrete** random variables:
  - $X \perp\!\!\!\perp Y$ iff $P_{XY}(X = a, Y = b) = P_X(X = a) \cdot P_Y(Y = b) \quad \text{for } -\infty < a, b < \infty$
  - $X \perp\!\!\!\perp Y$ iff $P_{XY}(X \in A, Y \in B) = P_X(X \in A) \cdot P_Y(Y \in B) \quad \text{for } A, B \subseteq \mathbb{R}$

Sum of independent discrete random variables

ADDING TWO INDEPENDENT DISCRETE RANDOM VARIABLES. Let X and Y be two independent discrete random variables, with probability mass functions p_X and p_Y. Then the probability mass function p_Z of $Z = X + Y$ satisfies

$$p_Z(c) = \sum_j p_X(c - b_j) p_Y(b_j),$$

where the sum runs over all possible values b_j of Y.

- **Proof (sketch).**

$$\begin{aligned} P(Z = c) &= \sum_j P(Z = c | Y = b_j) \cdot P(Y = b_j) \\ &= \sum_j P(X = c - b_j | Y = b_j) \cdot P(Y = b_j) \\ &= \sum_j P(X = c - b_j) P(Y = b_j) \end{aligned}$$

Independence of multiple random variables

Independence (factorization formula)

Random variables $X_1, \ldots, X_n$ are independent, if:

$$P_{X_1,\ldots,X_n}(X_1 \leq a_1, \ldots, X_n \leq a_n) = \prod_{i=1}^{n} P_{X_i}(X_i \leq a_i) \quad \text{for} \quad -\infty < a_1, \ldots, a_n < \infty$$

- $X_1, \ldots, X_n$ **discrete** random variables are independent iff:

$$P_{X_1,\ldots,X_n}(X_1 = a_1, \ldots, X_n = a_n) = \prod_{i=1}^{n} P_{X_i}(X_i = a_i) \quad \text{for} \quad -\infty < a_1, \ldots, a_n < \infty$$

- **Definition:** $X_1, \ldots, X_n$ are **i.i.d.** (independent and identically distributed) if $X_1, \ldots, X_n$ are independent and $X_i \sim F$ for $i = 1, \ldots, n$ for some distribution F

$X \sim Bin(n, p)$

DEFINITION. A discrete random variable X has a *binomial distribution* with parameters n and p, where $n = 1, 2, \ldots$ and $0 \leq p \leq 1$, if its probability mass function is given by

$$p_X(k) = \mathrm{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, \ldots, n.$$

We denote this distribution by $Bin(n, p)$.

- X models the number of successes in n Bernoulli trials (How many H's when tossing n coins?)
- **Intuition**: for $X_1, X_2, \ldots, X_n$ such that $X_i \sim Ber(p)$ and independent (**i.i.d.**):

$$X = \sum_{i=1}^{n} X_i \sim Bin(n, p)$$

- $p^k \cdot (1-p)^{n-k}$ is the probability of observing first k H's and then $n-k$ T's
- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ number of ways to choose the first k variables *[Binomial coefficient]*
- $p_X(k)$ computationally expensive to calculate (no closed formula, but approximation/bounds)

$X \sim Geo(p)$

DEFINITION. A discrete random variable X has a *geometric distribution* with parameter p, where $0 < p \leq 1$, if its probability mass function is given by

$$p_X(k) = \mathrm{P}(X = k) = (1-p)^{k-1} p \quad \text{for } k = 1, 2, \ldots .$$

We denote this distribution by $Geo(p)$.

- X models the number of Bernoulli trials before a success (how many tosses to have a H?)
- **Intuition**: for $X_1, X_2, \ldots$ such that $X_i \sim Ber(p)$ i.i.d.:

$$X = min_i\,(X_i = 1) \sim Geo(p)$$

- $\bar{F}(a) = P(X > a) = (1-p)^{\lfloor a \rfloor}$
- $F(a) = P(X \leq a) = 1 - \bar{F}(a) = 1 - (1-p)^{\lfloor a \rfloor}$

You cannot always lose

- H is 1, T is 0, $0 < p < 1$
- $B_n = \{\text{T in the first } n\text{-th coin tosses}\}$
- $P(\cap_{n\geq 1} B_i) = ?$
- $X \sim Geom(p)$
- $P(B_n) = P(X > n) = (1-p)^n$
- $P(\cap_{n\geq 1} B_n) = lim_{n\to\infty} P(B_n) = lim_{n\to\infty}(1-p)^n = 0$
- $P(\cap_{n\geq 1} B_n) = lim_{n\to\infty} P(B_n)$ for B_n non-increasing *[σ-additivity, see Lesson 01]*

But if you lost so far, you can lose again

Memoryless property

For $X \sim Geo(p)$, and $n, k = 0, 1, 2, \ldots$

$$P(X > n + k | X > k) = P(X > n)$$

Proof

$$\begin{aligned}
P(X > n + k | X > k) &= \frac{P(\{X > n + k\} \cap \{X > k\})}{P(\{X > k\})} \\
&= \frac{P(\{X > n + k\})}{P(\{X > k\})} \\
&= \frac{(1-p)^{n+k}}{(1-p)^k} \\
&= (1-p)^n = P(X > n)
\end{aligned}$$

Sum of independent random variables (repetita iuvant)

ADDING TWO INDEPENDENT DISCRETE RANDOM VARIABLES. Let X and Y be two independent discrete random variables, with probability mass functions p_X and p_Y. Then the probability mass function p_Z of $Z = X + Y$ satisfies

$$p_Z(c) = \sum_j p_X(c - b_j) p_Y(b_j),$$

where the sum runs over all possible values b_j of Y.

- Example:
  - For $X \sim Bin(n, p)$ and $Y \sim Bin(m, p)$, $Z \sim Bin(n + m, p)$
  - For $X \sim Geo(p)$ (days radio 1 breaks) and $Y \sim Geo(p)$ (days radio 2 breaks):

$$p_Z(X + Y = k) = \sum_{l=1}^{k-1} p_X(l) \cdot p_Y(k - l) = (k-1)p^2(1-p)^{k-2}$$

$X \sim NBin(n, p)$

Negative binomial (or Pascal distribution)

A discrete random variable X has a negative binomial with parameters n and p, where $n = 0, 1, 2, \ldots$ and $0 < p \leq 1$, if its probability mass function is given by

$$p_X(k) = P(X = k) = \binom{k+n-1}{k}(1-p)^k \cdot p^n \quad \text{for } k = 0, 1, 2, \ldots$$

- X models the number of failures before the n-th success in Bernoulli trials (how many T's to have n H's?)
- **Intuition**: for $X_1, X_2, \ldots, X_n$ such that $X_i \sim Geo(p)$ i.i.d.:

$$X = \sum_{i=1}^{n} X_i - n \sim NBin(n, p)$$

- $(1-p)^k \cdot p^n$ is the probability of observing first k T's and then n H's
- $\binom{k+n-1}{k} = \frac{(k+n-1)!}{k!(n-1)!}$ number of ways to choose the first k variables among $k+n-1$ (the last one must be a success!)

$X \sim Poi(\mu)$

DEFINITION. A discrete random variable X has a *Poisson distribution* with parameter μ, where $\mu > 0$ if its probability mass function p is given by

$$p(k) = \mathrm{P}(X = k) = \frac{\mu^k}{k!} \mathrm{e}^{-\mu} \quad \text{for } k = 0, 1, 2, \ldots.$$

We denote this distribution by $Pois(\mu)$.

- X models the number of events in a fixed interval if these events occur with a known constant mean rate μ and independently of the last event
  - telephone calls arriving in a system
  - number of patients arriving at an hospital
  - customers arriving at a counter
- μ denotes the mean number of events
- $Bin(n, \mu/n)$ is the number of successes in n trials, assuming $p = \mu/n$, i.e., $p \cdot n = \mu$
- When $n \to \infty$: $Bin(n, \mu/n) \to Poi(\mu)$ *[Law of rare events]*
  - Number of typos in a book, number of cars involved in accidents, etc.

The discrete Bayes' rule

BAYES' RULE. Suppose the events $C_1, C_2, \dots, C_m$ are disjoint and $C_1 \cup C_2 \cup \cdots \cup C_m = \Omega$. The conditional probability of C_i, given an arbitrary event A, can be expressed as:

$$P(C_i \mid A) = \frac{P(A \mid C_i) \cdot P(C_i)}{P(A \mid C_1)P(C_1) + P(A \mid C_2)P(C_2) + \cdots + P(A \mid C_m)P(C_m)}.$$

Definition. Conditional p.m.f. of X given $Y = b$ with $P_Y(Y = b) > 0$

$$p_{X|Y}(a|b) = \frac{p_{XY}(a, b)}{p_Y(b)} \quad \text{i.e.,} \quad P_{X|Y}(X = a|Y = b) = \frac{P_{XY}(X = a, Y = b)}{P_Y(Y = b)}$$

Discrete Bayes' rule:

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)} = \frac{p_{Y|X}(y|x)p_X(x)}{\sum_{a \in dom(X)} p_{Y|X}(y|a)p_X(a)}$$

From Discrete to Continuous

- Let $X \sim U(0, 1)$
  - $p(0) = p(1) = 1/2$
- Expand the support: let to $X \sim U(0, n)$
  - $p(0) = \ldots = p(i) = \ldots p(n) = 1/(n+1)$
- Ok for $n \in \mathbb{N}$, but for $n \to \infty$, we have:

$$p(a) = P(X = a) = 0 \quad \text{for all } a$$

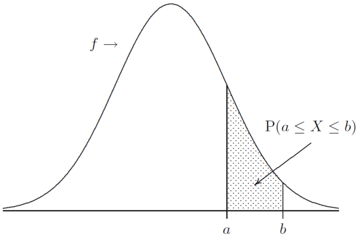
which breaks the properties of p.m.f.! *[Trascurable but possible events]*

- Since $|\mathbb{R}| = 2^{\aleph_0} > \aleph_0 = |\mathbb{N}|$, $n = \infty$ is reached when considering the continuum!

Conclusion: the idea of probability mass function does not extend to the continuum!

Continuous random variables

- We cannot assign a positive "mass" to a real number, but we can assign it to an interval!



DEFINITION. A random variable X is *continuous* if for some function $f : \mathbb{R} \rightarrow \mathbb{R}$ and for any numbers a and b with $a \leq b$,

$$\mathrm{P}(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x.$$

The function f has to satisfy $f(x) \geq 0$ for all x and $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$. We call f the *probability density function* (or *probability density*) of X.

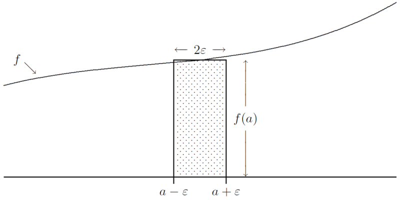
- Support of X is $dom(X) = \{x \in \mathbb{R} \mid f(x) > 0\}$
- $F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx$ *[Cumulative Distribution Function]*
- $P(X \in A) = \int_{x \in A} f(x)dx$ for $A \subseteq \mathbb{R}$ measurable
  - There exist non-measurable subsets of $\mathbb{R}$, i.e., for which **we cannot assign a mass**
  - **Borel sets** are measurable: intervals over $\mathbb{R}$ closed under countable union and complement

Density function

$$P(X=a) \leq P(a-\epsilon \leq X \leq a+\epsilon) = \int_{a-\epsilon}^{a+\epsilon} f(x)dx = F(a+\epsilon) - F(a-\epsilon)$$

- for $\epsilon \to 0$, $P(a-\epsilon \leq X \leq a+\epsilon) \to 0$, hence $P(X=a)=0$

- What is the meaning of the density function $f(x)$ then?
  - $f(a)$ is a (relative to other points) measure of how likely X will be near a
  - "probability mass per unit length" around a: $f(a) \cdot 2\epsilon$



- Discrete vs Continuous Random Variables *[F(x) is a continuous function for continuous r.v.]*

$$F(a) = \sum_{a_i \leq a} p(a_i) \quad p(a_i) = F(a_i) - F(a_{i-1}) \qquad F(x) = \int_{-\infty}^{x} f(y)dy \quad f(x) = \frac{d}{dx}F(x)$$

$X \sim U(\alpha, \beta)$

DEFINITION. A continuous random variable has a *uniform distribution* on the interval $[\alpha, \beta]$ if its probability density function f is given by $f(x) = 0$ if x is not in $[\alpha, \beta]$ and

$$f(x) = \frac{1}{\beta - \alpha} \quad \text{for } \alpha \leq x \leq \beta.$$

We denote this distribution by $U(\alpha, \beta)$.

- $F(x) = \int_{-\infty}^{x} f(x)dx = \frac{1}{\beta-\alpha}\int_{\alpha}^{x} 1dx = \frac{x-\alpha}{\beta-\alpha}$ for $\alpha \leq x \leq \beta$
- Differently from p.m.f.'s, densities can be larger than 1 (and arbitrarily large)
    - E.g., for $U(0, 0.5)$ we have $f(x) = 2$

$X \sim Exp(\lambda)$

- For $X \sim Geo(p)$, we have: $\bar{F}(x) = P(X > x) = (1-p)^{\lfloor x \rfloor}$ for $x \geq 0$
- extend to reals: $\bar{F}(x) = P(X > x) = (1-p)^x = e^{x \cdot log(1-p)} = e^{-\lambda x}$ for $\lambda = -log(1-p)$
- $f(x) = \frac{dF}{dx}(x) = -\frac{d\bar{F}}{dx}(x) = \lambda e^{-\lambda x}$ $F(x) = P(X \leq x) = 1 - e^{-\lambda x}$

> DEFINITION. A continuous random variable has an *exponential distribution* with parameter λ if its probability density function f is given by $f(x) = 0$ if $x < 0$ and
>
> $$f(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0.$$
>
> We denote this distribution by $Exp(\lambda)$.

- λ is the rate of events in a Poisson point process, i.e., a process in which events occur continuously and independently at a constant average rate, e.g.,
  - $\lambda = 1/10$ number of bus arrivals per minute, or $1/\lambda = 10$ minutes to wait for bus arrival
  - $P(X > 1) = e^{-\lambda} = 0.9048$ probability of waiting more than 1 minute.

$X \sim Exp(\lambda)$

DEFINITION. A continuous random variable has an *exponential distribution* with parameter λ if its probability density function f is given by $f(x) = 0$ if $x < 0$ and

$$f(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0.$$

We denote this distribution by $Exp(\lambda)$.

- Plausible and empirically adequate model for:
  - time until a radioactive particle decays, time it takes before your next telephone call, ...
  - time until default (on payment to company debt holders) in reduced-form credit risk modeling, ...
  - time between animal roadkills, time between bank teller serves customers, ...
  - monthly and annual maximum values of daily rainfall, (some types of) surgery duration, ...
- Exponential is *memoryless*: $P(X > s + t | X > s) = e^{-\lambda(s+t)}/e^{-\lambda s} = e^{-\lambda t} = P(X > t)$

$X \sim \mathcal{N}(\mu, \sigma^2)$

DEFINITION. A continuous random variable has a *normal distribution* with parameters μ and $\sigma^2 > 0$ if its probability density function f is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for } -\infty < x < \infty.$$

We denote this distribution by $N(\mu, \sigma^2)$.

- "Normal" means "typical" or "common"
- Also called Gaussian distribution, after **Carl Friedrich Gauss**, but introduced by **De Moivre**
- Standard Normal/Gaussian is $\mathcal{N}(0, 1)$
  - $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ sometimes written as $\phi(x)$
  - No closed form for $F(a) = \Phi(a) = \int_{-\infty}^{a} \phi(x) dx$
- Binomial approximation by a Normal distribution
  - $Bin(n, p) \approx \mathcal{N}(np, np(1-p))$ for n large and $0 \ll p \ll 1$ [**De Moivre–Laplace theorem**]

Quantiles

DEFINITION. Let X be a continuous random variable and let p be a number between 0 and 1. The pth *quantile* or 100pth *percentile* of the distribution of X is the smallest number q_p such that

$$F(q_p) = \mathrm{P}(X \leq q_p) = p.$$

The *median* of a distribution is its 50th percentile.

- Median m_X is $q_{0.5}$
- If $F()$ is *strictly* increasing, $q_p = F^{-1}(p)$
- E.g., for $Exp(\lambda)$, $F(a) = 1 - e^{-\lambda x}$, hence $F^{-1}(p) = \frac{1}{\lambda} \log \frac{1}{(1-p)}$
- General definition (also for discrete r.v.):

$$q_p = \inf_x \{P(X \leq x) \geq p\}$$

Joint distributions: continuous random variables

DEFINITION. Random variables X and Y have a *joint continuous distribution* if for some function $f : \mathbb{R}^2 \to \mathbb{R}$ and for all numbers a_1, a_2 and b_1, b_2 with $a_1 \leq b_1$ and $a_2 \leq b_2$,

$$\mathrm{P}(a_1 \leq X \leq b_1, a_2 \leq Y \leq b_2) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x, y)\, \mathrm{d}x\, \mathrm{d}y.$$

The function f has to satisfy $f(x, y) \geq 0$ for all x and y, and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, \mathrm{d}x\, \mathrm{d}y = 1$. We call f the *joint probability density function* of X and Y.

- The marginal density functions of X and Y are:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

- Moreover, as in the univariate case:

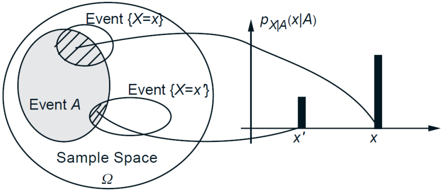
$$F(a, b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f(x, y) dx dy \qquad f(x, y) = \frac{d}{dx}\frac{d}{dy} F(x, y) = \frac{d^2}{dxdy} F(x, y)$$

Recalling conditional distribution: it applies to continuous r.v.'s

Conditional distribution

Consider the joint distribution P_{XY} of X and Y. The conditional distribution of X given $Y \in B$ with $P_Y(Y \in B) > 0$, is the function $F_{X|Y\in B} : \mathbb{R} \to [0, 1]$:

$$F_{X|Y\in B}(a) = P_{X|Y}(X \le a | Y \in B) = \frac{P_{XY}(X \le a, Y \in B)}{P_Y(Y \in B)} \quad \text{for } -\infty < a < \infty$$



- Distribution of X after knowing $Y \in B$.
- Chain rule: $P_{XY}(X \le a, Y \in B) = P_{X|Y}(X \le a | Y \in B) P_Y(Y \in B)$
- What if the distribution does not change w.r.t. the prior P_X?

Independence of two random variables

Independence $X \perp\!\!\!\perp Y$

A random variable X is independent from a random variable Y, if for all $P(Y \leq b) > 0$:

$$P_{X|Y}(X \leq a | Y \leq b) = P_X(X \leq a) \quad \text{for } -\infty < a < \infty$$

- Properties
  - $X \perp\!\!\!\perp Y$ iff $P_{XY}(X \leq a, Y \leq b) = P_X(X \leq a) \cdot P_Y(Y \leq b) \quad \text{for } -\infty < a, b < \infty$
  - $X \perp\!\!\!\perp Y$ iff $Y \perp\!\!\!\perp X$ *[Symmetry]*
- For X, Y **continuous** random variables:
  - $X \perp\!\!\!\perp Y$ iff $f_{XY}(x, y) = f_X(x) \cdot f_Y(y) \quad \text{for } -\infty < x, y < \infty$
  - $X \perp\!\!\!\perp Y$ iff $P_{XY}(X \in \mathcal{A}, Y \in \mathcal{B}) = P_X(X \in \mathcal{A}) \cdot P_Y(Y \in \mathcal{B}) \quad \text{for } \mathcal{A}, \mathcal{B} \subseteq \mathbb{R}$ measurable

Independence of multiple random variables

Independence (factorization formula)

Random variables $X_1, \ldots, X_n$ are independent, if:

$$P_{X_1,\ldots,X_n}(X_1 \leq a_1, \ldots, X_n \leq a_n) = \prod_{i=1}^{n} P_{X_i}(X_i \leq a_i) \quad \text{for} \quad -\infty < a_1, \ldots, a_n < \infty$$

- $X_1, \ldots, X_n$ **continuous** random variables are independent iff:

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i) \quad \text{for} \quad -\infty < x_1, \ldots, x_n < \infty$$

- **Definition:** $X_1, \ldots, X_n$ are **i.i.d.** (independent and identically distributed) if $X_1, \ldots, X_n$ are independent and $X_i \sim F$ for $i = 1, \ldots, n$ for some distribution F

Sum of independent continuous random variables

ADDING TWO INDEPENDENT CONTINUOUS RANDOM VARIABLES. Let X and Y be two independent continuous random variables, with probability density functions f_X and f_Y. Then the probability density function f_Z of $Z = X + Y$ is given by

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\, \mathrm{d}y$$

for $-\infty < z < \infty$.

- The integral is called the **convolution** of $f_X()$ and $f_Y()$
- $X, Y \sim Exp(\lambda)$, $Z = X + Y$, $\quad X, Y, Z \geq 0$ implies $0 \leq Y \leq Z$

$$f_Z(z) = \int_{-\infty}^{\infty} \lambda e^{-\lambda(z-y)} \lambda e^{-\lambda y} \mathbb{I}_{\{0 \leq y \leq z\}} dy = \lambda^2 e^{-\lambda z} \int_0^z 1 dy = \lambda(\lambda z) e^{-\lambda z}$$

- $Z = X_1 + \ldots + X_n$ for $X_i \sim Exp(\lambda)$ independent: *[Earlang $Erl(n, \lambda)$ distribution]*

$$f_Z(z) = \frac{\lambda(\lambda z)^{n-1} e^{-\lambda z}}{(n-1)!}$$

$Gam(\alpha, \lambda)$

- Let λ be the average rate of an event, e.g., $\lambda = 1/10$ number of buses in a minute
  - The waiting time to see **one** event is exponentially distributed. E.g., probability of waiting x minutes to see one bus.
  - The waiting time to see n **events** is Erlang distributed. E.g., probability of waiting x minutes to see n buses.

DEFINITION. A continuous random variable X has a *gamma distribution* with parameters $\alpha > 0$ and $\lambda > 0$ if its probability density function f is given by $f(x) = 0$ for $x < 0$ and

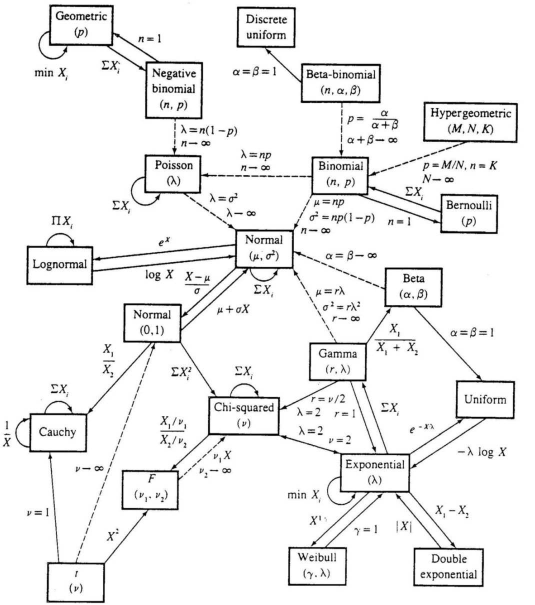
$$f(x) = \frac{\lambda (\lambda x)^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} \qquad \text{for } x \geq 0,$$

where the quantity $\Gamma(\alpha)$ is a normalizing constant such that f integrates to 1. We denote this distribution by $Gam(\alpha, \lambda)$.

- Extends $Erl(n, \lambda)$ from $n \in \mathbb{N}^+$ to $\alpha \in \mathbb{R}^+$ by **Euler's $\Gamma(\alpha)$**
  - The waiting time to see α **quantities** is Gamma distributed. E.g., probability of waiting x minutes to see α volume of rain.

Common distributions

- **Probability distributions at Wikipedia**
- **Probability distributions in R**
- C. Forbes, M. Evans,
  N. Hastings, B. Peacock (2010)
  Statistical Distributions, 4th Edition
  Wiley



Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

The continuous Bayes' rule

BAYES' RULE. Suppose the events $C_1, C_2, \ldots, C_m$ are disjoint and $C_1 \cup C_2 \cup \cdots \cup C_m = \Omega$. The conditional probability of C_i, given an arbitrary event A, can be expressed as:

$$P(C_i \mid A) = \frac{P(A \mid C_i) \cdot P(C_i)}{P(A \mid C_1)P(C_1) + P(A \mid C_2)P(C_2) + \cdots + P(A \mid C_m)P(C_m)}.$$

- **Definition.** Conditional density of X given $Y = y$ with $f_Y(y) > 0$:

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

- Continuous Bayes' rule:

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)} = \frac{f_{Y|X}(y|x)f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y|t)f_X(t)dt}$$