Probability Rehearsal

Renato Lo Cigno

Simulation and Performance Evaluation 2014-15

Why Probabilities & Uncertainty

- In models we reduce a complex physical phenomena to a simple mathematical description
- The hidden complexity pops-up as random variability
  - The arrival of queries to a web-server represented as a Poisson process
  - Movement of chemical species represented as a Brownian Motion
- In measures there are inherent uncertainties that we model as random errors
  - The noise introduced by electronics in a multimeter
  - The latency introduced by an interrupt in a software-based delay measure

Sample Space & Events

- The **Sample Space** S is the set of all possible outcomes of our experiment
  - Can be finite or infinite, numerable or not numerable
  - $S_d = \{1, 2, 3, 4, 5, 6\}$ is the sample space of a dice throw
  - $S_p = \{x, y, z\}; x \in \mathbf{R}, y \in \mathbf{R}, z \in \mathbf{R}$ is the sample space of a point in space

- $\mathbf{P}[S] = 1$; $\mathbf{P}[\emptyset] = 0$

Sample Space & Events

- An **Event** E is the outcome of an experiment
- Throwing dices
  - $E_1 = \{1, 3, 5\}$ "the dice is odd"
  - $E_2 = \{2, 4, 6\}$ "the dice is even"
  - $E_1 \cup E_2 = S_d$
  - $\mathbf{P}[E_1] = \mathbf{P}[E_2] = 0.5$ (if the dice is fair)
- Finding a point in space
  - $E_s = \{x, y, z \mid (x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2) = r^2\}$ "the point lies within a sphere of radius r centered in (x_0, y_0, z_0)"
  - $\mathbf{P}[E_s] = 0$

Probability Axioms

1. For any event A, $\mathbf{P}[A] \geq 0$
2. $\mathbf{P}[S] = 1$
3. If $A \cap B = \emptyset$ then $\mathbf{P}[A \cup B] = \mathbf{P}[a] + \mathbf{P}[B]$
4. By induction from Axiom 3 given mutually independent events A_i, i.e., $A_j \cap A_k = \emptyset \; \forall j \neq k$, then

$$\mathbf{P}\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \mathbf{P}[A_i]$$

Together with set algebra this is all we need to work with probabilities . . . well, with some manipulation here and there . . .

Probability Rules

Derive directly from the axioms and set algebra

1. $\forall A,\ \mathbf{P}[\overline{A}] = 1 - \mathbf{P}[A]$
2. $\mathbf{P}[\overline{S}] = \mathbf{P}[\emptyset] = 0$
3. Given two generic events A and B,
   $\mathbf{P}[A \cup B] = \mathbf{P}[A] + \mathbf{P}[B] - \mathbf{P}[A \cap B]$
   the rule can be extended easily to any number n of events, but it can be computationally long and cumbersome.
4. A more efficient computation of n given events A_i is

$$\mathbf{P}\left[\bigcup_{i=1}^{n} A_i\right] = \mathbf{P}[A_1] + \mathbf{P}[\overline{A_1} \cap A_2] + \mathbf{P}[\overline{A_1} \cap \overline{A_2} \cap A_3] + \cdots$$

$$+\mathbf{P}[\overline{A_1} \cap \overline{A_2} \cap \cdots \cap \overline{A_{n-1}} \cap A_n]$$

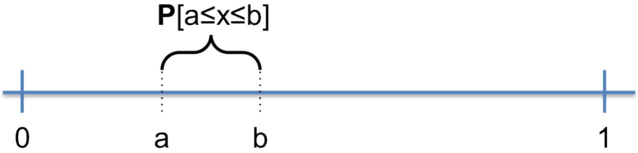
Events, Sample Space and Probabilities

A formal mathematical definition

- Standard set algebra apply to events and their space
- A *measure* is any function that assigns a positive real number to a subset
- Events are subsets of S that are *measurable*; we call this class of subsets $\mathcal{F}$
- A **probability system** or **probability space** is the triple $(S, \mathcal{F}, \mathbf{P})$ where
    - S is a set
    - $\mathcal{F}$ is a σ-field of subsets of S
    - $\mathbf{P}$ is a probability measure on $\mathcal{F}$

Example 1 (continuous space)

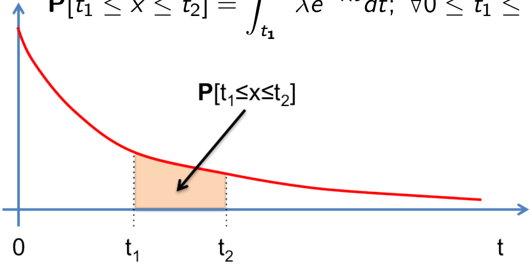
- Uniform measure on the interval $[0, 1]$
  - Let $S = [0, 1]$
  - The class of events $\mathcal{F}$ is defined by all possible segments
  - The probability measure is the length of a segment:
    $\mathbf{P}[a \leq x \leq b] = b - a;\ \forall 0 \leq a \leq b \leq 1$



Example 2 (continuous space)

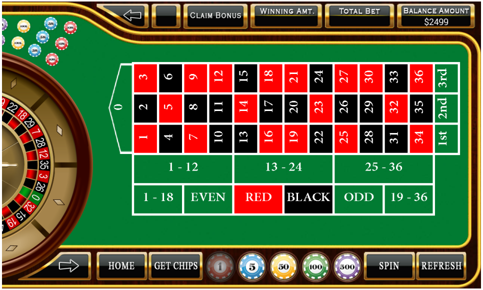
- Exponentially decreasing arrivals in time with parameter λ
  - Let $S = [0, \infty)$s
  - The class of events $\mathcal{F}$ is defined by all possible arrival intervals $[t_1, t_2]$
  - The probability measure is the integral of the exponential function: $f_X(x) = \lambda e^{-\lambda t}$ over the interval $[t_1, t_2]$

$$\mathbf{P}[t_1 \leq x \leq t_2] = \int_{t_1}^{t_2} \lambda e^{-\lambda t} dt; \; \forall 0 \leq t_1 \leq t_2 \leq \infty$$



Example 3 (discrete space)

- Roulette
  - Let $S = \{0, 1, 2, \cdots, 36\}$
  - The class of events $\mathcal{F}$ is defined as each number $(E_i = i)$, red, black, odd, even, $\{1, \cdot, 18\}$, $\{19, \cdots, 36\}$, $\{1, \cdots, 12\}$, $\{13, \cdots, 24\}$, $\{25, \cdots, 36\}$, C1, C2, C3
  - The probability measure is the size of the event divided by the size of S: $\mathbf{P}[E] = |E|/|S|$



0	3	6	9	12	15	18	21	24	27	30	33	36	3rd
	2	5	8	11	14	17	20	23	26	29	32	35	2nd
	1	4	7	10	13	16	19	22	25	28	31	34	1st

1 - 12		13 - 24		25 - 36	
1 - 18	EVEN	RED	BLACK	ODD	19 - 36

Example 3 (discrete space)

- Roulette
  - Let $S = \{0, 1, 2, \cdots, 36\}$
  - The class of events $\mathcal{F}$ is defined as each number ($E_i = i$), red, black, odd, even, $\{1, \cdot, 18\}$, $\{19, \cdots, 36\}$, $\{1, \cdots, 12\}$, $\{13, \cdots, 24\}$, $\{25, \cdots, 36\}$, C1, C2, C3
  - The probability measure is the size of the event divided by the size of S: $\mathbf{P}[E] = |E|/|S|$

Needless to say that the winning ratio is smaller than the fair share:

$$W_r < \frac{1}{\mathbf{P}[E]}$$

Example 3 (discrete space)

- Roulette
  - Let $S = \{0, 1, 2, \cdots, 36\}$
  - The class of events $\mathcal{F}$ is defined as each number ($E_i = i$), red, black, odd, even, $\{1, \cdot, 18\}$, $\{19, \cdots, 36\}$, $\{1, \cdots, 12\}$, $\{13, \cdots, 24\}$, $\{25, \cdots, 36\}$, C1, C2, C3
  - The probability measure is the size of the event divided by the size of S: $\mathbf{P}[E] = |E|/|S|$

But you can actually win at the roulette by playing always the same simple number and doubling each time your bet until you win, and then you stop . . .

. . . the only problem is that you might need and infinite amount of money before you win

Homework: Prove the statement just done

Example 4 (discrete space)

- Repeated binary experiments with biased output p: throwing coins, success or failure, or any other experiment that yield a binary output $\in [0,1]$ with success probability p (output$=1$)
  - Let $S = [0, n]$, where n is the number of experiments
  - The class of events $\mathcal{F}$ is defined by the counter of successes: $E_i =$ "all strings with i 1s"
  - The probability measure derives from combinatorial calculus observing that any given combination with k 1s has probability $p^k(1-p)^{n-k}$ and that there are $\begin{pmatrix} n \\ k \end{pmatrix}$ such combinations
  - $\mathbf{P}[E_k] = \binom{n}{k} p^k (1-p)^{n-k}$

Example 4 (discrete space)

The binomial coefficient

- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
- The binomial coefficient stems from the calculus of the simple combination of n distinct objects taken k at a time or in other words from the number of possible dispositions of k objects into n positions, where position has no meaning (all the k objects are equal one another).
- Given we have n places and k ones, we have n ways of placing the first one, $n-1$ of placing the second one, $\ldots$, $(n-k+1)$ of placing the k-th one.

Example 4 (discrete space)

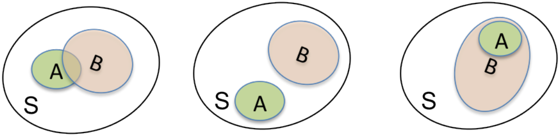
The binomial coefficient (cont.)

- Thus we have $n(n-1)(n-2\cdots(n-k+1) = \dfrac{n!}{(n-k)!}$ possible simple dispositions of objects
- However the k “ones” cannot be distinguished, thus their position is irrelevant; this means that we have to divide the number of simple dispositions by the number of permutations of k elements, which is $k!$, thus we finally find the binomial coefficient

Conditional Probability

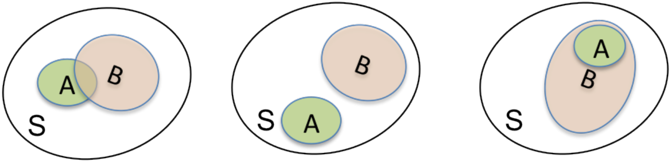
- What is the probability $\mathbf{P}[A|B]$ of event A given event B?
  - $B \neq \emptyset \longrightarrow \mathbf{P}[B] \neq 0$
- $\mathbf{P}[A|B] = \dfrac{\mathbf{P}[A \cap B]}{\mathbf{P}[B]}$
- $\mathbf{P}[A|B] \neq \mathbf{P}[B|A]$
- Some simple algebra also yields

$$\mathbf{P}[A \cap B] = \begin{cases} \mathbf{P}[A] \cdot \mathbf{P}[B|A] & \text{if } \mathbf{P}[A] \neq 0 \\ \mathbf{P}[B] \cdot \mathbf{P}[A|B] & \text{if } \mathbf{P}[B] \neq 0 \\ 0 & \text{otherwise} \end{cases}$$



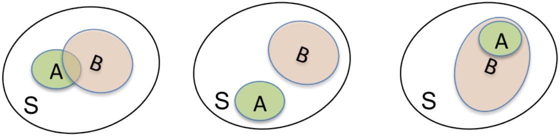
Independent events

- A is independent from B if $\mathbf{P}[A|B] = \mathbf{P}[A]$
- If A is independent from B then B is independent from A
- Disjoint events cannot be independent!
- It is easily shown that for independent events $\mathbf{P}[A \cap B] = \mathbf{P}[A]\mathbf{P}[B]$



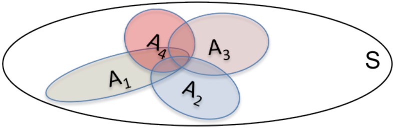
Independent events

- Independence is not transitive
- A is independent from A only if $A = S$
- If A and B are independent then also $\overline{A}$ and $\overline{B}$, A and $\overline{B}$, $\overline{A}$ and B are all independent
- If $A \subset B$ then A and B cannot be independent



Independent events

- Independence can be extended to any set of events $A_i, \; i = 1, \ldots, n$
- Events A_i are mutually independent only if for **any** set of $k, \; 2 \leq k \leq n$ distinct indices $i_1, i_2, \ldots i_k$ then $\mathbf{P}[A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}] = \mathbf{P}[A_{i_1}]\mathbf{P}[A_{i_2}] \cdots \mathbf{P}[A_{i_k}]$
- Pairwise independence $\mathbf{P}[A_i \cap A_j] = \mathbf{P}[A_i]\mathbf{P}[A_j], \; \forall i, j < n, i \neq j$ does not imply mutual independence
- $\mathbf{P}[A_1 \cap A_2 \cap \cdots \cap A_n] = \mathbf{P}[A_1]\mathbf{P}[A_2] \cdots \mathbf{P}[A_n]$ does not imply mutual independence



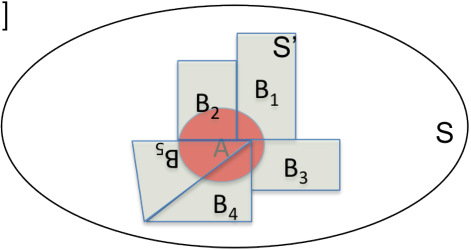
Total Probability

- Observation: $\mathbf{P}[A] = \mathbf{P}[A \cap B] + \mathbf{P}[A \cap \overline{B}]$ ...trivial
- By induction from the observation above it is easy to show that given $A \subset S'$ and a set of events $B_i,\ i = 1, \ldots, n$ such that

$$S' = \bigcup_{i=1}^{n} B_i;\ B_i \cap B_j = \emptyset\ \forall i, j,\ i \neq j$$

- Then

$$\mathbf{P}[A] = \sum_{i=1}^{n} \mathbf{P}[A|B_i]\mathbf{P}[B_i]$$



Bayes' Rule (or Theorem)

- Suppose we have observed and event A, that we know is associated to a set of mutually exclusive events $B_i, \; i = 1, \ldots, n$, but we do not know which of the B_i actually occurred with (or before) A
- However, we know that $\mathbf{P}[B_j|A] = \dfrac{\mathbf{P}[B_j \cap A]]}{\mathbf{P}[A]}$
- Applying the definition of conditional probability to the numerator we get

$$\mathbf{P}[B_j|A] = \frac{\mathbf{P}[A|B_j]\mathbf{P}[B_j]}{\mathbf{P}[A]}$$

Bayes' Rule (or Theorem)

- Often Bayes' Rule is written as applying the theorem or total probability to the denominator

$$\mathbf{P}[B_j|A] = \frac{\mathbf{P}[A|B_j]\mathbf{P}[B_j]}{\sum_{i=1}^{n}\mathbf{P}[A|B_i]\mathbf{P}[B_i]}$$

- In practice this "transforms" apriori probabilities $\mathbf{P}[A|B_i]$ into aposteriori probabilities $\mathbf{P}[B_j|A]$ probabilities, hence gives insight into several class of problems, including classification problems
- Bayes' Rule is the base for machine learning

Probability Mass Function (pmf)

- A RV X is a function on the state space S that maps onto $\mathbf{R}$

$$X : S \mapsto \mathbf{R}$$

- For a discrete RV X we define the pmf $p_X(x)$ as the probability that the outcome of a random experiment with RV X yields as result x

$$p_X(x) = \mathbf{P}[X = x]$$

- Any function that guarantees that
  1. $0 \leq p_X(x) \leq 1$
  2. $\sum_{x(s) \in \mathbf{R}} p_X(x) = 1$

  can be a pmf

Mapping example

- Consider as random experiment throwing a generic coin the events are Head or Tail: $S = \{H, T\}$

Mapping example

- Valid mapping function (depending on the coin “honesty“ are
  1. $H \mapsto x_1 = 1;\ pp[X = x_1] = 0.5$
     $T \mapsto x_2 = -1;\ pp[X = x_2] = 0.5$
  2. $H \mapsto x_1 = \sqrt{2};\ pp[X = x_1] = 0.22$
     $T \mapsto x_2 = 0.1;\ pp[X = x_2] = 0.78$

- Wile the following mappings violate the definition of probability function
  1. $H \mapsto x_1 = 1;\ pp[X = x_1] = 0.33$
     $T \mapsto x_2 = -1;\ pp[X = x_2] = 0.33$
  2. $H \mapsto x_1 = \sqrt{2};\ pp[X = x_1] = -0.1$
     $T \mapsto x_2 = 0.1;\ pp[X = x_2] = 1.1$

Vectors of RVs

- Let $\overrightarrow{X} = \{X_1, X_2, \cdots, X_n\}$ be a vector of n RVs
- In this case the outcome of a random experiment is the vector $\overrightarrow{x} = \{x_1, x_2, \cdots, x_n\}$
- Extending the notion of pmf to vectors is conceptually easy

$$\overrightarrow{X} : S \mapsto \mathbf{R}^n; \ ||\overrightarrow{X}|| = n$$

S is the cartesian product of each RV state space

- The joint pmf of $\overrightarrow{X}$ is

$$p_{\overrightarrow{X}}(\overrightarrow{x}) = \mathbf{P}[X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n]$$

Whenever there is no ambiguity we will drop the notation $\overrightarrow{\cdot}$

Independent RVs

- We can extend the notion of independence from events to RVs
- Two RVs X_1 and X_2, $\overrightarrow{X} = \{X_1, X_2\}$ are said independent if

$$p_{\overrightarrow{X}}(\overrightarrow{x}) = \mathbf{P}[X_1 = x_1, X_2 = x_2] = \mathbf{P}[X_1 = x_1]\mathbf{P}[X_2 = x_2];\ \forall x_1, x_2$$

- A similar notation for independent RVs is

$$p_{X,Y}(x, y) = p_X(x)p_Y(y)$$

- Mutual independence for a generic vector $\overrightarrow{X}$ of RVs means that

$$p_{\overrightarrow{X}}(\overrightarrow{x}) = \prod_i p_{X_i}(x_i)$$

Example: two dices

- We expect that throwing two dices the outcome of the RVs D_1 and D_2 representing the experiment are independent $\mathbf{P}[1,1] = \mathbf{P}[1]\mathbf{P}[1] = 1/36$, $\mathbf{P}[2,1] = \mathbf{P}[2]\mathbf{P}[1] = 1/36$,etc.
- Notice, however, the possibility of another random experiment with outcome $D_s = D_1 + D_2$; D_s is clearly another RV function of D_1 and D_2
  1. What is the state space of D_s?
  2. What is the mapping that leads to a proper representation of D_s as a probability function?
  3. What are the probabilities of the elementary events of D_s?
  4. Does independence of D_1 and D_2 play a role here?

Cumulative Distribution Function (CDF)

- Consider now a continuous RV X
- The probability cannot be associated to single points, as they have a null support
- However, we can consider the event $E = X \in (-\infty, x]$, and we define the CDF as

$$F_X(x) = \mathbf{P}[X \leq x]$$

1. $F_X(x)$ is a continuous non decreasing function of x
2. $\lim_{x \to -\infty} F_X(x) = 0$
3. $\lim_{x \to \infty} F_X(x) = 1$

Probability Density Function (pdf)

- We define pdf the derivative of $F_X(x)$

$$f_X(x) = \frac{d\, F_X(x)}{dx}$$

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\, dx \;;\;\; -\infty \leq x \leq \infty$$

- The CDF must be continuous and derivable
- A mixed continuous/discrete RV require the use of generalized analysis including the Dirac's delta δ function

$$\delta(x_0) = \begin{cases} 0 & \forall x \neq x_0 \\ \infty & x = x_0 \end{cases} \;;\;\; \int_{-\infty}^{\infty} \delta(x_0) = 1$$

Joint Continuous RVs

- Consider two RVs X and Y
- $F_{X,Y}(x,y) = \mathbf{P}[X \leq x, Y \leq y]$
- $f_{X,Y}(x,y) = \dfrac{\partial^2 F_{X,Y}(x,y)}{\partial x\, \partial y}$
- $F_{X,Y}(x,y) = \displaystyle\int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(x,y)\, dx\, dy$
- They are independent if

$$F_{X,Y}(x,y) = F_X(x) F_Y(y)$$

$$f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

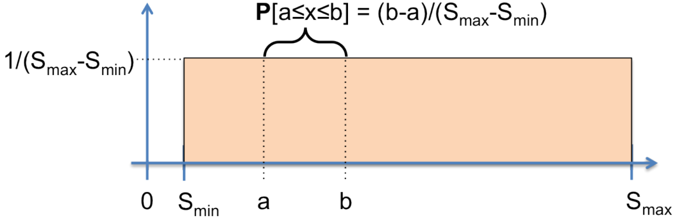
and the partial derivatives becomes standard derivatives

Probabilities are volumes

- For a mono-dimensional continuous RV the probability of an event is the area (integral) of $f_X(x)$ on the interval defining the event

$$\mathbf{P}[a \le x \le b] = \int_a^b f_X(x)\,dx$$

- Example for a uniform distribution



Probabilities are volumes

In general for an m-dimensional random vector **X** the probability of an event $\mathbf{E} = \{x_1 \in e_1, x_2 \in e_2, \cdots, x_m \in e_m\}$ is the volume subtended by the m-dimensional pdf integrated over the intervals that define **E**

$$\mathbf{P}[\mathsf{x} \in \mathbf{E}] = \int_{e_1}\int_{e_2}\cdots\int_{e_m} f_{\mathbf{X}}(\mathsf{x})\, dx_1\, dx_2 \cdots dx_m$$

if **X** is continuous or

$$\mathbf{P}[\mathsf{x} \in \mathbf{E}] = \sum_{x_1 \in e_1}\sum_{x_2 \in e_2}\cdots\sum_{x_m \in e_m} p_{\mathbf{X}}(\mathsf{x})$$

if **X** is discrete

Probabilities are volumes

If the random vector **X** is composed of mutually independent RVs the integrals (sums) are decoupled (product form)

$$\mathbf{P}[\mathbf{x} \in \mathbf{E}] = \int_{e_1} f_{X_1}(x_1)\,dx_1 \int_{e_2} f_{X_2}(x_2)\,dx_2 \cdots \int_{e_m} f_{X_m}(x_m)\,dx_m$$

if **X** is continuous or

$$\mathbf{P}[\mathbf{x} \in \mathbf{E}] = \sum_{x_1 \in e_1} p_{X_1}(x_1) \sum_{x_2 \in e_2} p_{X_2}(x_2) \cdots \sum_{x_m \in e_m} p_{X_m}(x_m)$$

if **X** is discrete

Marginal Distributions

Given an m-dimensional random vector $\mathbf{X} = \{X_1, X_2, \cdots, X_m\}$ we define **marginal probability distribution** of x_i the integral of the joint pdf with respect to all other RVs in the vector

$$f_{X_i}(x_i) = \int_{-infty}^{\infty} \int_{-infty}^{\infty} \cdots \int_{-infty}^{\infty} f_{\mathbf{X}}(\mathbf{x})\, dx_1\, dx_2 \cdots dx_j \cdots dx_m \; ; \; j \neq i$$

$$p_{X_i}(x_i) = \sum_{i_1} \sum_{i_2} \cdots \sum_{i_j} \cdots \sum_{i_m} p_{\mathbf{X}}(\mathbf{x}); \quad j \neq i$$

This is in practice the pdf of the RV X_i without any influence of the other RVs

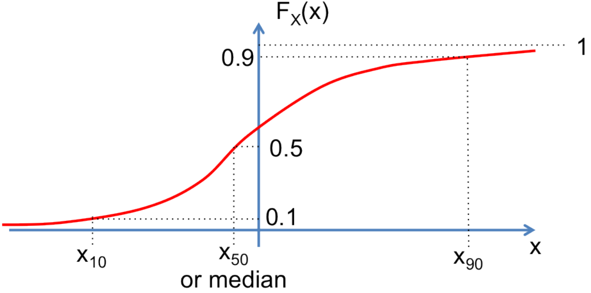
- Given the random vector $\mathbf{D} = \{D_1, D_2\}$ representing the throw of two independent dices, find the marginal distribution $f_{D_1}(x)$.

- Given $f_{X,Y}(x, y) = \dfrac{2}{ab}$; $0 \le x \le a$, $0 \le y \le b$, find the marginal distributions of both X and Y and draw all three distributions

Percentiles

- The n-th percentile of a distribution is the value x_n such that

$$\int_{-\infty}^{x_n} f_X(x)\,dx = \frac{n}{100}$$



Percentiles

- Percentiles are fundamental in reliability analysis
- They can be used for quick hypothesis rejection
  - Given a set or measured points we make an hypothesis on their stochastic distribution, if the percentiles of the distribution are not compatible with the measured points our hypothesis is wrong
- Some performances are related to fraction of "objects" lost/delayed/not met/. . . or achieved/reached/survived/. . .
  - In a video stream the main performance metric is the fraction of video frames that arrive within the playout delay
  - In a real-time system the key performance is the fraction of jobs that do not finish within the deadline

Moments

- We call n-th *moment* $E[X^n]$ of an RV the sum (integral) of the n-th power of the RV value multiplied by its probability (pdf)

$$E[X^n] = \sum_{i=-\infty}^{\infty} x_i^n \, p_X(x_i); \quad \text{for a discrete RV}$$

$$E[X^n] = \int_{i=-\infty}^{\infty} x^n \, f_X(x) \, dx; \quad \text{for a continuous RV}$$

The average

- The first moment of a distribution is its average

$$E[X] = \sum_{i=-\infty}^{\infty} x_i \, p_X(x_i); \quad \text{for a discrete RV}$$

$$E[X] = \int_{\infty}^{\infty} x \, f_X(x) \, dx; \quad \text{for a continuous RV}$$

- $E[X]$ is often indicates as μ_X or simply μ

Examples

- $f_X(x) = \frac{1}{b-a};\ a \le x \le b$

$$E[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, dx = \int_a^b \frac{x}{b-a}\, dx = \\ = \left.\frac{x^2}{2(b-a)}\right|_{[a,b]} = \frac{b+a}{2}$$

- Throwing a dice

$$E[X] = \sum_{i=1}^{6} x_i = \sum_{i=1}^{6} 1/6 = 3.5$$

- $E[X]$ is not necessarily $\in S$!!

Central moments

- We call n-th *central moment* $E[(X-\mu)^n]$ or simply $\mu_n[X]$ of an RV the moment computed on the RV minus its average value

$$\mu_n[X] = \sum_{i=-\infty}^{\infty} (x_i - \mu)^n \, p_X(x_i); \quad \text{for a discrete RV}$$

$$\mu_n[X] = \int_{-\infty}^{\infty} (x - \mu)^n \, f_X(x) \, dx; \quad \text{for a continuous RV}$$

Variance and Standard Deviation

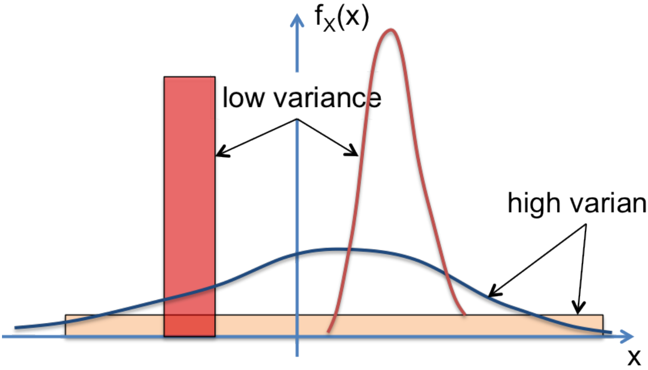
- The second *central moment* is normally $E[(X-\mu)^2]$ or simply σ_X^2 is called variance and is strictly related to the "spread" of the distribution; σ is called standard deviation

$$\sigma_X^2 = \sum_{i=-\infty}^{\infty} (x_i - \mu)^n \, p_X(x_i); \quad \text{for a discrete RV}$$

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu)^n \, f_X(x) \, dx; \quad \text{for a continuous RV}$$

Variance and Standard Deviation

- The second *central moment* is normally $E[(X-\mu)^2]$ or simply σ_X^2 is called variance and is strictly related to the “spread” of the distribution; σ is called standard deviation



Exercise

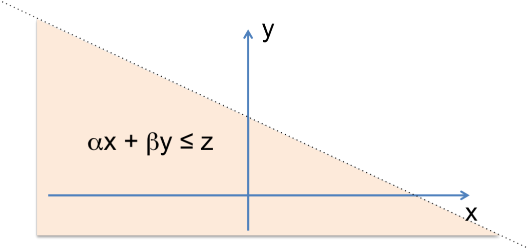
- Compute the variance of the following distributions
- $f_X(x) = \dfrac{1}{b-a};\ a \leq x \leq b$
- Throwing a dice
- $f_X(x) = \lambda e^{-\lambda x};\ 0 \leq x \leq \infty$
- The sum of throwing n dices, increasing n. How does the variance change?

Distribution of sums of RVs

- Consider a RV $Z = \alpha X + \beta Y$.
  What is the pdf $f_Z(z)$ given $f_{X,Y}(x,y)$?
- We can immediately write

$$F_Z(z) = \mathbf{P}[Z \le z] = \int\int_{A_z} f_{X,Y}(x,y)\,dxdy$$

$A_z = \{(x,y) | \alpha x + \beta y \le z\}$ identifies a half plane of (x,y)



pdf of a sum of RVs

- Let's set $\alpha = \beta = 1$ for the sake of simple computations ...
- Since $y = z - x$, then

$$F_Z(z) = \int\int_{A_z} f_{X,Y}(x,y)\,dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{z-x} f_{X,Y}(x,y)\,dydx$$

and with the change of variable $y = t - x$

$$F_Z(z) = \int_{-\infty}^{\infty}\int_{-\infty}^{z} f_{X,Y}(x,(t-x))\,dtdx$$

$$= \int_{-\infty}^{z}\int_{-\infty}^{\infty} f_{X,Y}(x,(t-x))\,dxdt$$

pdf of a sum of RVs

- Integrating in x, by definition of pdf we have

$$F_Z(z) = \int_{-\infty}^{z} f_Z(t)\,dt$$

- Which implies, comparing the last equation of slide 98

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x,(t-x))\,dx$$

pdf of a sum of independent RVs

- Let X and Y be independent: $f_{X,Y}(x,y) = f_X(x) f_Y(y)$.
- Then given $Z = X + Y$

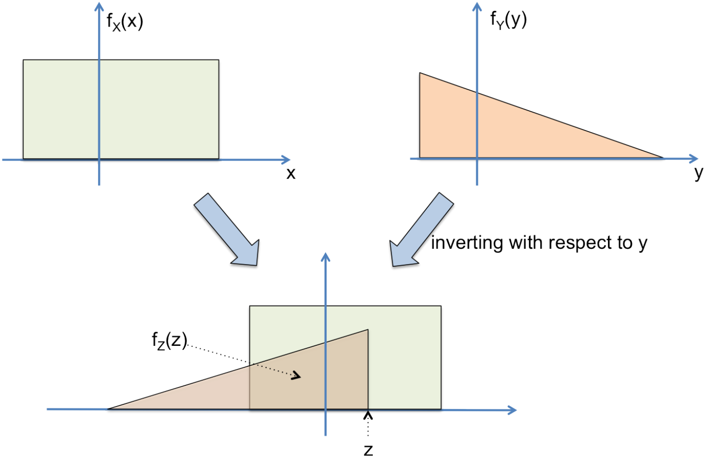
$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx$$

which is known as **convolution** (or convolutional product/integral) of $f_X(x)$ and $f_Y(y)$ often indicated as $f_X(x) * f_Y(y)$

- Furthermore if $x \geq 0,\ y \geq 0$

$$f_Z(z) = \int_{0}^{z} f_X(x) f_Y(z-x)\, dx$$

The process of convolution



Linearity of expectation

Consider the first moment of a sum of RVs $Z = \alpha X + \beta Y$

$$E[Z] = E[\alpha X + \beta Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(\alpha x + \beta y) f_{X,Y}(x,y)\,dxdy =$$

$$= \int_{-\infty}^{\infty}\alpha x \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dydx + \int_{-\infty}^{\infty}\beta y \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dxdy$$

by definition of marginal distribution we have

$$E[\alpha X + \beta Y] = \int_{-\infty}^{\infty}\alpha x\, f_X(x)\,dx + \int_{-\infty}^{\infty}\beta y\, f_Y(y)\,dy$$

$$E[Z] = \alpha E[X] + \beta E[Y]$$

Linearity of expectation

- The linearity of expectation can be extended to any number of RVs

$$E\left[\sum_{i=1}^{n}\alpha_i X_i\right] = \sum_{i=1}^{n}\alpha_i E[X_i]$$

- It is **not** required that the RVs are independent!

Linearity is valid only for the first moment, and not for the others (obviously!!)

Central Limit Theorem (without proof)

- Consider the sum Z of n independent RVs with finite mean and variance and let $n \to \infty$
- The pdf $f_Z(z)$ is a Gaussian distribution with $\mu = \sum_i \mu_i$ and $\sigma^2 = \sum_i \sigma_i^2$

$$f_Z(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$$

- This theorem is fundamental in data analysis, but requires that RVs are independent and they have finite variance
- Convergence speed depends on the underlying distributions, but it is normally fast (10-20 RVs are enough to have a good approximation)

Central Limit Theorem

- A more formal statement of the theorem can be enunciated forming the normalized RV Z_n such that

$$Z_n = \frac{\sum_i^n X_i - \sum_i^n \mu_i}{\sqrt{\sum_i^n \sigma_i^2}}$$

so that $E[Z_n] = 0$ and $\mathrm{Var}[Z_n] = 1$

- Under reasonable regularity conditions of the X_i, we have that Z_n converges to a Gaussian (normal) distribution with zero mean and unitary variance; $Z_n \rightarrow N(0,1)$:

$$\lim_{n\to\infty} F_{Z_n}(z) = \lim_{n\to\infty} \mathbf{P} Z_n < z = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$$

Covariance (mutual variance of RVs)

- We define **covariance** of 2 RVs X and Y the quantity

$$\mathrm{Cov}(X,Y) = E[(X-\mu_x)(Y-\mu_y)] = E[XY] - E[X]E[Y]$$

- The covariance measure how much two RVs are interdependent: the larger the covariance the more correlation the the RVs have
- It follows that in general

$$E[XY] = E[X]E[Y] + \mathrm{Cov}(X,Y)$$

Covariance properties - 1

- By definition if $\text{Cov}(X, Y) = 0$ then X and Y are incorrelated
- For independent variables $\text{Cov}(X, Y) = 0$ so independent variables are incorrelated (prove it)
- The converse is not necessarily true, i.e., $\text{Cov}(X, Y) = 0$ does not imply independence. Example:
  - Consider X uniformly distributed in $(-1, 1)$ and $Y = X^2$. Clearly Y is completely dependent on X
  - All odd moments of X are zero by definition: $E[X^k] = 0\ \forall k$ odd
  - $\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = E[X^3] - E[X]E[Y] = 0$
  - Thus X and Y are uncorrelated even if they are strictly dependent one another

Covariance properties - 2

- Take $Y = aX, \; a \neq 0$, so that Y is linearly dependent from X

$$\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y] = aE[X^2] - aE[X]E[X] =$$

$$= a\mathrm{Var}[X] = \frac{1}{a}\mathrm{Var}[Y]$$

taking the square value we have

$$\mathrm{Cov}^2(X,Y) = \mathrm{Var}[X]\mathrm{Var}[Y]$$

- In general it can be shown that

$$0 \leq \mathrm{Cov}^2(X,Y) \leq \mathrm{Var}[X]\mathrm{Var}[Y]$$

Properties of independent RVs

Given the independent RVs (can be extended to any number of them) the following holds

- $E[XY] = E[X]E[Y] \Longrightarrow \mathrm{Cov}(X, Y) = 0$
- $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$
- Incorrelation is not independence
    - $\mathrm{Cov}(X, Y) = 0$ is sufficient (definition) for RVs to be incorrelated, but as we have see it is only necessary and not sufficient for RVs to be independent

Independence vs Incorrelation

- $\mathrm{Cov}(X, Y)$ measures the degree of linear dependence between RVs, but misses higher order dependencies
- Independence instead requires that also higher order dependencies do not exist, but this is difficult to express in terms of relations of higher order moments.

Coefficient of variation and correlation

It is useful to normalize coefficients to compare different pdfs

- $C_x = \dfrac{\sigma_x}{\mu_x}$ is the coefficient of variation
  For an exponential distribution $C_x = 1$, so it is a convenient means to measure how much a distribution resembles an exponential
- $\rho(X,Y) = \dfrac{\mathrm{Cov}(X,Y)}{\sigma_x\sigma_y}$ is the correlation coefficient (provided σ_x and σ_y exist): $-1 \leq \rho(X,Y) \leq 1$:

$$\rho(X,Y) \begin{cases} -1 & \text{if} Y = -aX, \ \ a > 0 \\ 1 & \text{if} Y = ax, \ \ a > 0 \\ 0 & \text{if} X \text{and} Y \text{are uncorrelated} \\ -1 < c < 1 & \text{otherwise} \end{cases}$$