Introduzione alla data science e al pensiero computazionale
Lezione 8: Revisione dei fondamenti di statistica

Giancarlo Succi
Dipartimento di Informatica – Scienza e Ingegneria
Università di Bologna
g.succi@unibo.it

Probability – History of the term

- Classical probability
- Frequency probability
- Axiomatic probability

Evolution: Classical $\rightarrow$ Frequency $\rightarrow$ Axiomatic

Classical probability. Definition

Definition If a random experiment (process with an uncertain outcome) can result in n *mutually exclusive and equally likely* outcomes, and if n_A of these outcomes has an attribute A, then the probability of A is the fraction n_A/n.

Classical probability. Definition

A basic assumption in the definition of classical probability is that n is a finite number; that is, there is only a finite number of possible outcomes. If there is an infinite number of possible outcomes, the probability of an outcome is not defined **in the classical sense**.

Classical probability. Examples

- Roll of a dice
- Draw a card from a deck
- Toss of a coin

Classical probability. Problems

- (Already mentioned) Infinite events
- Definition of equally likely (circular)
- Management of non physical problems for which there is no experiment

Frequency probability

- We start from considering the case of the outcomes not being equally likely and n being potentially infinite.
- In such cases, how might we define the probability of an outcome that has attribute “a”?

Frequency probability

Definition We might take a random (finite) sample from the population of interest and identify the proportion of the sample with attribute "a".

$$Freq_a = \frac{n_a}{n}$$

We **estimate** $Pr[a]$ with $Freq_a$.

Problems with frequency probability

- We cannot run infinite trials, and our stopping point may induce ambiguities or errors in results, e.g., tossing the coint . . .
- What if we cannot run trials? This definition requires repeatable experiments.

Axiomatic probability – Sample space

Definitions:

- The sample space Ω is the set of possible outcomes of an experiment.

For example:

- Tossing a coin, the sample space is { H, T }
- Tossing a coin twice, the sample space is { HH, HT, TH, TT }
- Rolling a dice, the sample space is {1, 2, 3, 4, 5, 6}
- Football match, the sample space is {W, L, D }
- Selecting a point in the interval (0,1), the sample space is S = (0,1)
- Selecting a salary in the range [50K₽, 200K₽], the sample space is S = [50K₽, 200K₽]

This and the following slides are inspired by:

https://faculty.math.illinois.edu/~kkirkpat/SampleSpace.pdf

Axiomatic probability – Events

Definitions:

- Subsets of Ω are called events.

For example:

- Tossing a coin with the sample space { H, T }, an event E is { H }
- Tossing a coin twice with sample space { HH, HT, TH, TT }, an event E is { HH, TT }
- Rolling a dice with sample space {1, 2, 3, 4, 5, 6}, an event E is {1, 3, 5 }, *the odd sides*
- Football match with sample space {W, L, D }, an event E is {W, D }, *not loosing*
- Selecting a point in the interval (0,1) with the sample space S = (0,1), an event E is (0,1/2), *the first half*
- Selecting a salary in the range [50K₽, 200K₽], an event E is (150K₽, 200K₽], *top salary*

Axiomatic probability – Probability measure

A probability measure is a function $\mathbb{P}$ defined on the σ-algebra[†] of events $\mathcal{E}$ such that:

- $\forall A \in \mathcal{E}, \mathbb{P}(A) \geq 0$,
- $\mathbb{P}(\Omega) = 1$,
- if $A_1, A_2, \ldots \in \mathcal{E}$ are disjoint then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

[†] *The definition of σ-algebra is omitted*

Axiomatic probability – Probability space

The triple $(\Omega, \mathcal{E}, \mathbb{P})$ is called a **probability space**.

For the example of rolling a dice:

- Ω - all possible outcomes: 1, 2, ... , 6
- $\mathcal{E}$ - subsets of Ω:
  - {1 }
  - the odd sides: {1, 3, 5 }
  - the small sides: {1, 2 }
  - any side: {1, 2, 3, 4, 5, 6}
- $\mathbb{P}$ - the probability measure of events from $\mathcal{E}$:
  - $\mathbb{P}(\{1\}) = 1/6$
  - $\mathbb{P}(\text{the odd sides}) = 1/2$
  - $\mathbb{P}(\text{the small sides}) = 1/3$
  - $\mathbb{P}(\text{any side}) = 1$

Axiomatic probability – Random variable

A **random variable** is a function $X : \Omega \to \mathbb{A}$ such that:

- $\mathbb{A}$ is a measurable space,
- for every real a, $\{\omega \in \Omega : X(\omega) \leq a\} \in \mathcal{E}$.

Note that:

- Usually $\mathbb{A}$ is $\mathbb{R}$ or a subset of it
- A random variable is NOT a variable but a function

Example of random variable (1/2)

Tossing a dice three times:

- the sample space Ω is $\{$ 111, 112, 113, 211, $\ldots$ $\}$
- events E_i are:
    - $\{$ 222 $\}$
    - $\{$ 111, 555 $\}$
    - $\{$ 123, 456, 531 $\}$
    - $\ldots$
- the probability measure is the function $\mathbb{P}$
    - $\mathbb{P}(\{\ 222\ \}) = 1/216$
    - $\mathbb{P}(\{\ 111, 555\ \}) = 1/108$
    - $\mathbb{P}(\{\ 123, 456, 531\ \}) = 1/72$
    - $\mathbb{P}(\ldots)$
    - $\ldots$

Example of random variable (2/2)

Tossing a dice three times:

- a random variable X is the sum of the three results
- the sample space Ω_X is $\{ 3, 4, 5, 6, 7, 8, 9, \dots , 18\}$
- events $\mathrm{E}_{X,i}$ are:
  - $\{ 3\}$
  - $\{4, 6 \}$
  - $\{ 16, 17, 18 \}$
  - $\dots$
- the probability measure is the function $\mathbb{P}_X$
  - $\mathbb{P}_X(\{ 3\}) = 1/216$
  - $\mathbb{P}_X(\{4, 6 \}) = 13/216$
  - $\mathbb{P}_X(\{ 16, 17, 18 \}) = 10 \ / \ 216 = 5/108$
  - $\dots$

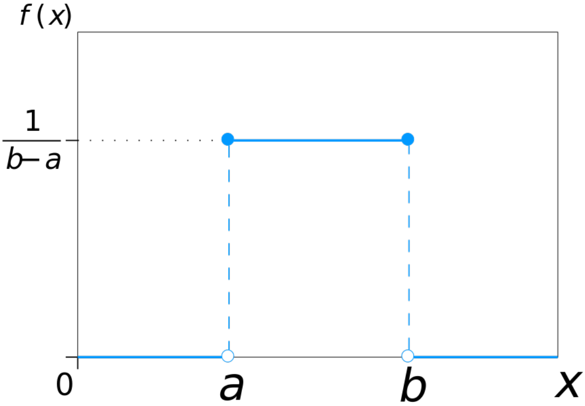
Axiomatic probability – Pdf

Given:

- A random variable X
- A set in which it is defined S (called the support)
- A **probability density function** (pdf) is a function $f(X) \geq 0, X \in S$
- defined for the continuous case as:

$$\mathbb{P}(S) = \int_S f(X) dX$$

Example: Uniform distribution:
pdf:



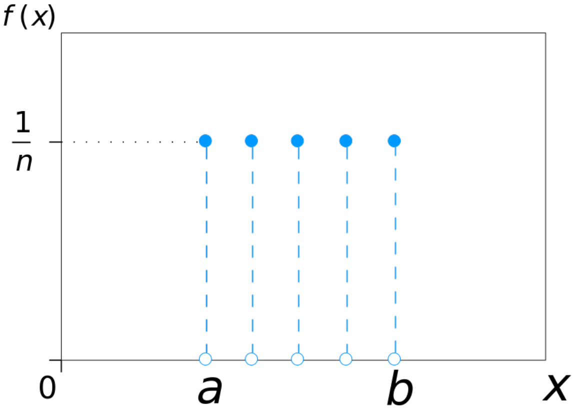
Axiomatic probability – Pmf

Given:

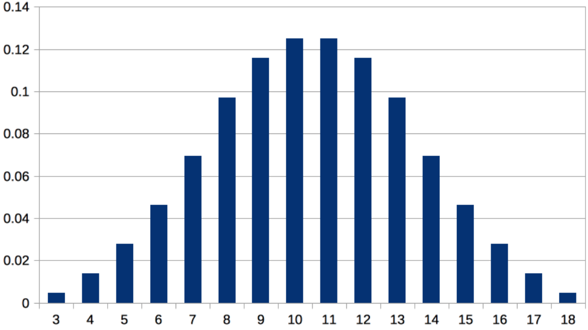
- A random variable X
- A set in which it is defined S (called the support again)
- A **probability mass function** (pdf) is a function $f(X) \geq 0, X \in S$
- defined for the discrete case as:

$$\mathbb{P}(S) = \sum_{S} f(X)$$

Example 1: Discrete Uniform distribution:
pmf:



Example 2: Computing the score of rolling a dice three times: pmf:



Axiomatic probability – Cdf (continuous)

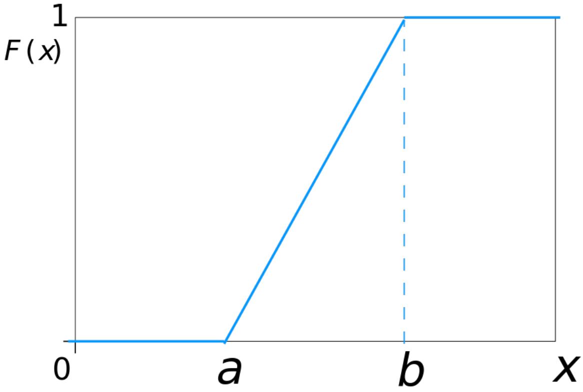
Given:

- A random variable X
- A set in which it is defined S (called the support)
- The probability density function of X, f_X
- We define the **(cumulative) distribution function** of X in the continuous case as:

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(w)dw$$

Note. The subscript X is sometimes omitted, so we often write F instead of F_X.

Example: Uniform distribution:
cdf:



Axiomatic probability – Cdf (discrete)

Given:

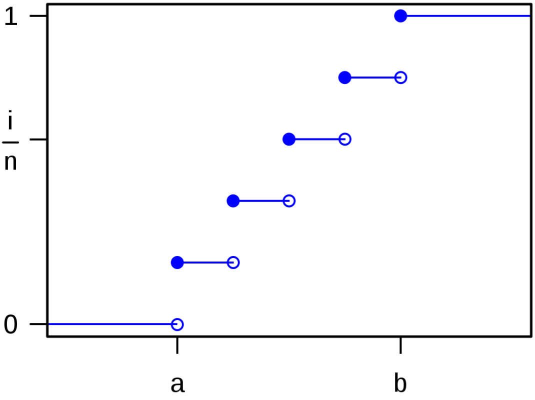
- A random variable X
- A set in which it is defined S (called the support)
- The probability mass function of X, f_X
- We define the **(cumulative) distribution function** of X in the discrete case as:

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{w \leq x} f_X(w)$$

Note. The subscript X is sometimes omitted, so we often write F instead of F_X.

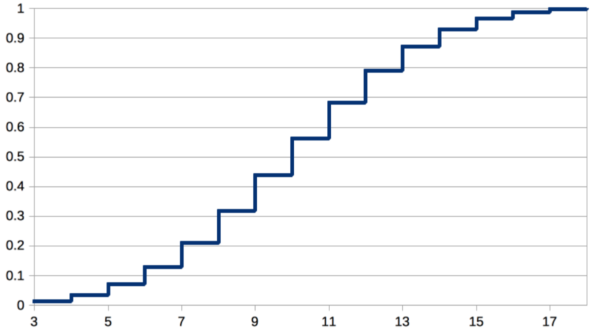
Axiomatic probability – Cdf – Example 1

Example 1: Discrete uniform distribution:
cdf:



Axiomatic probability – Cdf – Example 2

Example 2: Computing the score of rolling a dice three times: cmf:



Axiomatic probability – Expected value

Let:

- x be a random variable
- $a(x)$ be a function of x
- F be its cdf
- f be its pdf or pmf (if discrete)

The **expected value** of $a(X)$ is:

- In continuous:

$$\mathbb{E}(a(X)) = \int_{x \in S} a(x) f(x) dx$$

- In discrete:

$$\mathbb{E}(a(X)) = \sum_{j} a(x_j) f(x_j)$$

Basic definitions – Mean (continuous)

Let $a(X) = X$, then the mean of X is:

- in the continuous case:

$$\mu = \mathbb{E}(X) = \int_{x \in S} x f(x) dx$$

where S is the Support.

Example of mean (continuous)

Example: Uniform distribution on an interval [a,b].
pdf = $\frac{1}{b-a}$ if $x \in [a, b]$ and pdf=0 elsewhere
Mean:

$$\mu = \mathbb{E}(X) = \int_{-\infty}^{+\infty} x \frac{1}{b-a} dx = \frac{1}{b-a} \left(\frac{x^2}{2} \right) \Big|_a^b =$$

$$= \frac{1}{2} * \frac{1}{b-a} * (b^2 - a^2) = \frac{1}{2}(a+b)$$

Basic definitions – Mean (discrete)

Let $a(X) = X$, then the mean of X is:

- in discrete case:

$$\mu = \mathbb{E}(X) = \sum_j x_j f(x_j)$$

Example of mean (discrete)

Example: a dice with 6 edges (6 outcomes: $X = \{1, 2, 3, 4, 5, 6\}$).
PMF equals 1/6 for each outcome.
Mean:

$$\mu = \mathbb{E}(X) = \sum_{x \in X} x\frac{1}{6} =$$

$$= 1 * \frac{1}{6} + 2 * \frac{1}{6} + 3 * \frac{1}{6} + 4 * \frac{1}{6} + 5 * \frac{1}{6} + 6 * \frac{1}{6} =$$

$$= \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5$$

Basic definitions – Variance

Variance of random variable:

$$\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]$$

- for discrete case: $\mathbb{V}(X) = \sum_j (x_j - \mu)^2 f(x_j)$
- for continuous case: $\mathbb{V}(X) = \int_S (x - \mu)^2 f(x) dx$

Example of variance

Example: Uniform distribution on an interval [a,b]. $\mu = \frac{1}{2}(a+b)$

$$\mathbb{V}(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 \frac{1}{b-a} dx =$$

$$= \int_{-\infty}^{+\infty} (x - \frac{1}{2}(a+b))^2 \frac{1}{b-a} dx = \frac{1}{b-a} \left(\frac{(x-\frac{1}{2}(a+b))^3}{3} \right) \Big|_a^b =$$

$$= \frac{1}{3 \cdot 2^3 (b-a)} ((b-a)^3 - (a-b)^3) = \frac{1}{12}(b-a)^2$$

Basic definitions – α-quantile

α-quantile $(\alpha \in (0,1))$:

$$X_\alpha : \mathbb{P}(X \leq X_\alpha) = \alpha;$$

$$\mathbb{P}(X > X_\alpha) = 1 - \alpha.$$

Basic definitions – Median

Median (0.5-**quantile**):

$$X_{0.5} : \mathbb{P}(X \leq X_{0.5}) = 0.5;$$

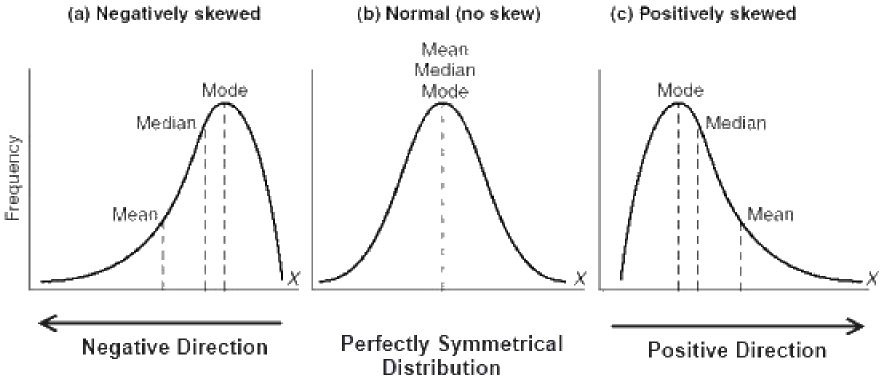
$$\mathbb{P}(X > X_{0.5}) = 0.5$$

Basic definitions – Mode

Mode (the most frequent element):

$$mode = argmax(f(x))$$

Mean. Median. Mode. Examples



Normal distribution

The random variable X has a **normal distribution** with mean μ and variance σ^2 if it has density:

$$\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}};$$

and CDF:

$$\Phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt.$$

We write $X \sim N(\mu, \sigma^2)$.

Standard Normal distribution

The random variable Z has a **standard normal distribution** if $\mu = 0$ and $\sigma = 1$. Hence it has density:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2};$$

and CDF:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.$$

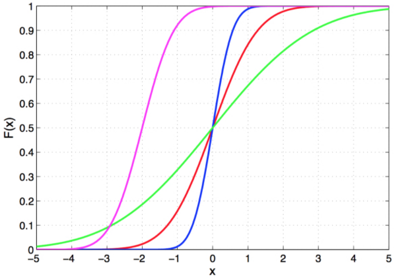
We write $Z \sim N(0, 1)$.
The α upper quantile is denoted by z_α. Thus, if $Z \sim N(0, 1)$, then we write $\mathbb{P}(Z > z_\alpha) = \alpha$.

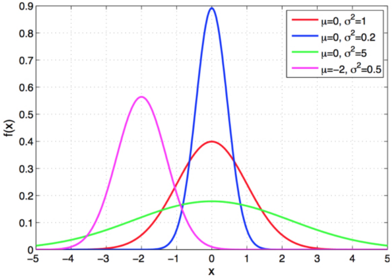
Normal distribution

$$X \in \mathbb{R} \sim N(\mu, \sigma^2), \sigma^2 > 0$$

$$F(x) = \Phi(\frac{x-\mu}{\sigma})$$

$$f(x) = \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})$$





Exercise 1

- A discrete random variable has the following probability mass function defined as:

x	0	2	4	6
$P(X = x)$	$\frac{1}{7}$	$\frac{3}{7}$	$\frac{1}{7}$	$\frac{2}{7}$

- Compute:
  - mean
  - median
  - mode
  - variance
  - standard deviation

Exercise 1 – solution

- A discrete random variable has the following probability mass function defined as:

x	0	2	4	6
$P(X = x)$	$\frac{1}{7}$	$\frac{3}{7}$	$\frac{1}{7}$	$\frac{2}{7}$
$x \times P(X = x)$	0	$0,857$	$0,571$	$1,714$
$x - \overline{x}$	$-3,143$	$-1,143$	$0,857$	$2,857$
$(x - \overline{x})^2$	$9,878$	$1,306$	$0,735$	$8,163$
$(x - \overline{x})^2 \times P(X = x)$	$1,411$	$0,560$	$0,105$	$2,332$

- So we get:
  - $\overline{x} = \sum(x \times P(X = x)) = 3,143$
  - $\mathbb{V}(x) = \sum((x - \overline{x})^2 \times P(X = x)) = 4,408$

Exercise 2

- A discrete random variable has the following probability mass function defined as:

x	1	2	3	4	5
$P(X=x)$	$\frac{2}{10}$	$\frac{2}{10}$	k	$\frac{3}{10}$	$\frac{1}{10}$

- Compute:
  - the value of k
  - mean
  - median
  - mode
  - variance
  - standard deviation

Exercise 3

- A continuous random variable X has probability density function defined as:

$$f(x)=\begin{cases} \frac{x^2}{9}, & \text{if } 0 < x < 3 \\ 0, & \text{elsewhere} \end{cases}$$

- Compute:
  - mean
  - median
  - mode
  - variance
  - standard deviation

Exercise 3 – solution

- A continuous random variable X has probability density function defined as:

$$f(x)=\begin{cases} \frac{x^2}{9}, & \text{if } 0 < x < 3 \\ 0, & \text{elsewhere} \end{cases}$$

- Compute:
  - mean $= 2{,}25$
  - median $\approx 2{,}38$
  - mode $= 3$
  - variance $= 0{,}3375$
  - standard deviation $= 0{,}581$

Exercise 4

- A continuous random variable X has probability density function defined as:

$$f(x)=\begin{cases} \frac{\sin(x)}{2}, & \text{if } 0 \leq x \leq \pi \\ 0, & \text{elsewhere} \end{cases}$$

- Compute:
  - mean
  - median
  - mode
  - variance
  - standard deviation

Exercise 4 - solution

- A continuous random variable X has probability density function defined as:

f(x)=$\begin{cases} \frac{\sin(x)}{2}, & \text{if } 0 \leq x \leq \pi \\ 0, & \text{elsewhere} \end{cases}$

- Compute:
  - mean $= \frac{\pi}{2}$
  - median $= \frac{\pi}{2}$
  - mode $= \frac{\pi}{2}$
  - variance $= \frac{\pi^2}{4}$ - 2 $\approx$ 0,467
  - standard deviation $\approx$ 0,684

Domande?

Fine della lezione otto.