Probabilità e Statistica[1]

Isadora Antoniano-Villalobos
isadora.antoniano@unive.it

Laurea in Informatica
(Data science/ Tecnologie e scienze dell'informazione)
Università Ca' Foscari di Venezia

Anno academico 2023/2024

[1]Materiale didattico redatto da: Isadora Antoniano-Villalobos & Federica Giummolè

Variabili Casuali

Introduzione

Abbiamo già visto vari esempi di **fenomeni aleatori** e **spazi campionari** ad essi collegati.

- Uno spazio campionario relativo ad un esperimento o ad un fenomeno casuale può essere di varia natura.

  ☞ Non è detto che sia un insieme numerico!

- In molte situazioni, anziché essere interessati allo specifico risultato di un esperimento, siamo interessati ad una sua **funzione numerica**.

Introduzione

➔ **Esempio:** si fa una scommessa in cui si vincono 1000 euro se lanciando una moneta si ottiene testa e se ne perdono 1500 se il risultato è croce.

Chiaramente ciò che ci preme di più è l'importo della vincita (negativo se si tratta della perdita) e non il risultato esatto del lancio della moneta.

$$\Omega = \{\ T,\ \ C\,\} \qquad \{1000, \quad -1500\ \} = \,?$$

Variabili aleatorie

Variabile aleatoria

Una **variabile aleatoria** o **casuale** X è una funzione che assume valori numerici determinati dall'esito di un certo fenomeno aleatorio.
Formalmente, se Ω è lo spazio campionario relativo al fenomeno di interesse, X è una particolare funzione

$$X : \Omega \to \mathbb{R} \; .$$

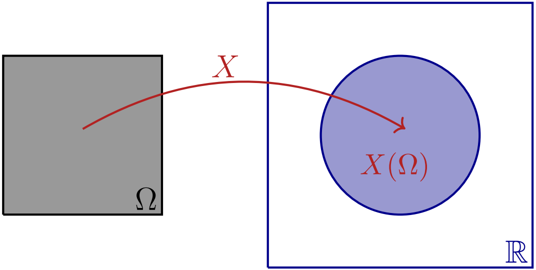
Le variabili aleatorie si indicano di solito con una lettera maiuscola.

Esempi:

1. $S_4 =$ numero di teste in 4 lanci consecutivi di una moneta; i possibili valori di S_4 sono: 0, 1, 2, 3 e 4.
2. $X =$ vincita nella scommessa descritta precedentemente. I possibili valori di X sono 1000 e -1500.
3. $T =$ tempo di vita di un componente elettronico prodotto da una ditta. I possibili valori di T sono tutti i numeri reali maggiori di 0.

Spazio campionario indotto

Una variabile aleatoria associata ad un esperimento definisce dunque un nuovo spazio campionario numerico, costituito da tutti i possibili valori assunti dalla variabile stessa.



In questo modo si passa da un generico spazio campionario ad un sottoinsieme di $\mathbb{R}$.

Spazio campionario indotto

Bisogna ora assegnare le probabilità agli eventi del nuovo spazio campionario.

Vedremo due modi differenti per fare quest'assegnazione, a seconda che lo spazio campionario indotto sia:

- un sottoinsieme numerabile di $\mathbb{R}$ (esempi 1 e 2) ➔ variabile **discreta**
- un sottoinsieme non numerabile di $\mathbb{R}$ (esempio 3) ➔ variabile **continua**

Variabili aleatorie discrete

Variabile aleatoria discreta

Una **variabile aleatoria discreta** X assume valori in un insieme numerabile (o finito) di punti, $\{x_1, x_2, \ldots, x_i, \ldots\}$.

Un modello probabilistico per X è un'assegnazione di probabilità ad ogni suo possibile valore:

$$\mathbb{P}[X = x_i] = p_i, \quad \forall i = 1, 2, \ldots \ .$$

Le probabilità p_i sono tali che:

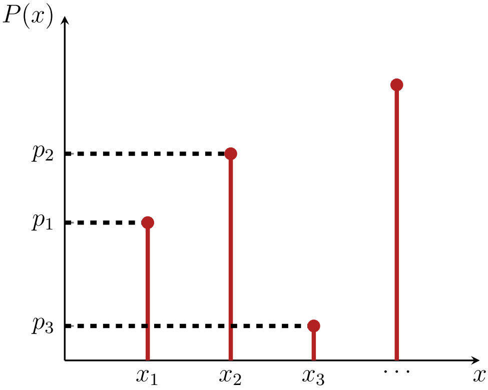
1. $0 \leq p_i \leq 1, \quad \forall i = 1, 2, \ldots$
2. $\sum_i p_i = 1$.

Per calcolare $\mathbb{P}[X \in A]$, si sommano le probabilità dei singoli valori che appartengono ad A:

$$\mathbb{P}[X \in A] = \sum_{i : x_i \in A} p_i \ .$$

Funzioni di probabilità

Un'assegnazione di probabilità per X, $P(x) = \mathbb{P}[X = x]$ viene chiamata **funzione di probabilità** e può essere rappresentata graficamente tramite un diagramma a bastoncini:



Funzioni di probabilità

- **Esempio:** Si ricordi l'esempio dell'urna composta da quattro palle bianche numerate da 1 a 4 e tre palle nere numerate da 1 a 3, dalla quale si campiona casualmente una palla
  - Sia $X =$ numero estratto dall'urna.

$$X \in \{1, 2, 3, 4\} = X(\Omega) \subset \mathbb{R}$$

$$\mathbb{P}[X = 1] = 2/7, \quad \mathbb{P}[X = 2] = 2/7,$$

$$\mathbb{P}[X = 3] = 2/7, \quad \mathbb{P}[X = 4] = 1/7$$

Si scrive anche

$$P_X(x) = \begin{cases} 2/7 & \text{se } x = 1, 2, 3; \\ 1/7 & \text{se } x = 4; \\ 0 & \text{altrimenti} \end{cases}$$

Funzioni di probabilità

👉 Se $A = \{1, 2\}$, allora

$$\mathbb{P}[X \in A] = \mathbb{P}[X \leq 2] = P_X(1) + P_X(2) == \frac{2}{7} + \frac{2}{7} = \frac{4}{7}$$

- Supponiamo di vincere 30 euro se la pallina estratta è bianca e di perderne 20 se è nera. La variabile casuale che rappresenta questo scenario è: $Y =$ importo vinto nel gioco.

$$P_Y(y) = \begin{cases} 4/7 & \text{se } y = 30, \\ 3/7 & \text{se } y = -20, \\ 0 & \text{altrimenti} \end{cases}$$

Funzioni di probabilità

- Consideriamo ora un'estrazione di due palline senza reinserimento. Sia $S =$ somma dei due numeri estratti.

$$P_S(s) = \begin{cases} 2/42 & \text{se } x = 2, \\ 8/42 & \text{se } x = 3, \\ 10/42 & \text{se } x = 4, \\ 12/42 & \text{se } x = 5, \\ 6/42 & \text{se } x = 6, \\ 4/42 & \text{se } x = 7, \\ 0 & \text{altrimenti} \end{cases}$$

Infatti, ad esempio,

$$\begin{aligned} \mathbb{P}[S = 2] &= \mathbb{P}[X_1 = 1 \cap X_2 = 1] \\ &= \mathbb{P}[X_1 = 1]\,\mathbb{P}[X_2 = 1 | X_1 = 1] \\ &= \frac{2}{7} \times \frac{1}{6} \end{aligned}$$

➔ **Esercizio:** Cosa cambia se l'estrazione avviene con reinserimento?

Variabili aleatorie continue

Variabile aleatoria continua

Una **variabile aleatoria continua** X assume valori in un insieme continuo di punti (un sottoinsieme di $\mathbb{R}$ non numerabile).

Un modello probabilistico per X è un'assegnazione di probabilità ad ogni sottoinsieme di suoi possibili valori:

$$\mathbb{P}\left[X \in A\right] = \text{ area su A sottesa ad una curva.}$$

La curva è il grafico di una funzione $f(x)$ tale che:

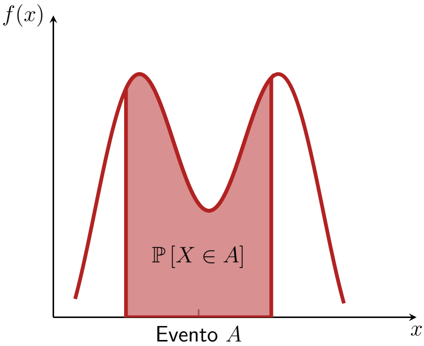
1. $f(x) \geq 0\ , \quad \forall x \in \mathbb{R}$;
2. $\int_{\mathbb{R}} f(x)\, dx = 1$, cioè l'area totale sotto il grafico di $f(x)$ è 1.

Densità di probabilità

Una funzione $f(x)$ con le proprietà precedenti viene chiamata **densità di probabilità**.

Una volta assegnata una densità di probabilità alla variabile X, si può scrivere, per ogni evento A di $\mathbb{R}$:

$$\mathbb{P}[X \in A] = \int_A f(x)\, dx \ ,$$



Importante: $\mathbb{P}[X = x] = P(x) = 0, \ \forall x \in \mathbb{R}$.

Densità di probabilità

➔ **Esempio:** Sia X una variabile aleatoria continua con densità

$$f(x) = 2\mathrm{e}^{-2x}\mathbf{1}_{(0,+\infty)}(x).$$

- $f(x)$ è davvero una densità:
  1. $f(x) \geq 0, \ \forall x \in \mathbb{R}$,
  2. $\int_{\mathbb{R}} f(x)\,dx = \int_0^{+\infty} 2\mathrm{e}^{-2x}\,dx = 1.$
- Considerando due eventi: $A = (1,2)$ e $B = (-1,1)$, è possibile calcolare le probabilità:

$$\mathbb{P}[X \in A] = \mathbb{P}[X \in (1,2)] = \int_1^2 2\mathrm{e}^{-2x}\,dx = \mathrm{e}^{-2} - \mathrm{e}^{-4}$$

$$\mathbb{P}[X \in B] = \mathbb{P}[X \in (-1,1)] = \int_{-1}^0 0\,dx + \int_0^1 2\mathrm{e}^{-2x}\,dx = 1 - \mathrm{e}^{-2}.$$

Funzione di ripartizione

Funzione di ripartizione

Si dice **funzione di ripartizione** (o di distribuzione cumulativa) di una variabile aleatoria X la funzione $F : \mathbb{R} \to [0,1]$ così definita:

$$F(x) = \mathbb{P}\left[X \leq x\right], \quad \forall x \in \mathbb{R}.$$

- La funzione di ripartizione in un dato punto x è semplicemente la probabilità che la variabile X assuma valori minori o al più uguali a x. Per questo il suo codominio è $[0,1]$
- La funzione di ripartizione F ha le seguenti proprietà:
  1. F è non decrescente,
  2. F è continua a destra,
  3. $\lim_{x\to-\infty} F(x) = 0$ e $\lim_{x\to+\infty} F(x) = 1$.

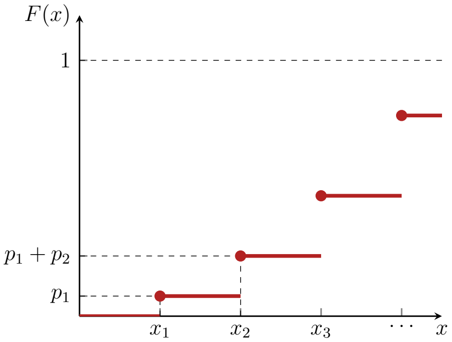
Funzione di ripartizione di una v.a. discreta

Se X è una v.a. discreta con valori $\{x_1, x_2, \ldots\}$ e funzione di probabilità $P(x_i) = p_i$, allora

$$F(x) = \sum_{i:x_i \leq x} \mathbb{P}\left[X = x_i\right] = \sum_{i:x_i \leq x} p_i.$$

☞ La funzione di ripartizione di una v.a. discreta è una funzione costante a tratti, con salti in corrispondenza dei punti massa $x_1, x_2, \ldots$.

Funzione di ripartizione di una v.a. discreta



Dalla funzione di ripartizione si può risalire alla funzione di probabilità della variabile così:

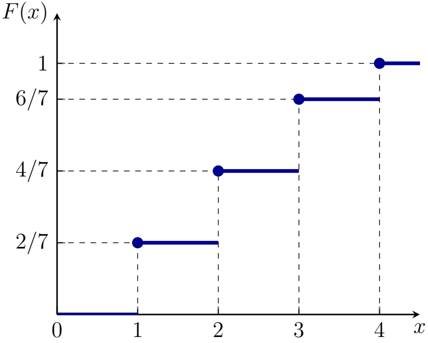
$$\mathbb{P}[X = x] = F(x) - F(x^-), \quad \forall x \in \mathbb{R},$$

cioè la probabilità che X assuma il valore x è uguale al salto della funzione di ripartizione nel punto x.

Funzione di ripartizione di una v.a. discreta

➔ **Esempio:** Per la variabile $X=$ numero estratto dalla solita urna, si ha:

$$F(x) = \begin{cases} 0 & x < 1 \\ 2/7 & 1 \leq x < 2 \\ 4/7 & 2 \leq x < 3 \\ 6/7 & 3 \leq x < 4 \\ 1 & x \geq 4 \end{cases}$$



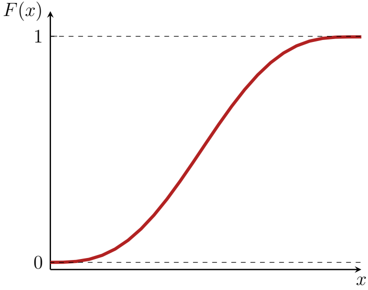
➔ **Esercizio:** trovare la funzione di probabilità di X a partire dalla funzione di ripartizione.

Funzione di ripartizione di una v.a. continua

Se X è una v.a. continua con densità $f(x)$, allora

$$F(x) = \int_{-\infty}^{x} f(t)\,dt.$$

La funzione di ripartizione di una v.a. continua è una funzione continua.



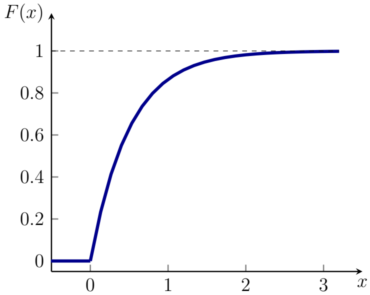
Dalla funzione di ripartizione si può risalire alla densità di probabilità della variabile in tutti i punti in cui $F(x)$ è derivabile:

$$f(x) = \frac{dF(x)}{dx},$$

Funzione di ripartizione di una v.a. continua

- **Esempio:** Per la variabile X con densità $f(x) = 2\mathrm{e}^{-2x}\mathbf{1}_{(0,+\infty)}(x)$, si ha:

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - \mathrm{e}^{-2x} & x \geq 0 \end{cases}$$



- **Esercizio:** trovare la densità di X a partire dalla funzione di ripartizione.

Costanti caratteristiche

Una **costante caratteristica** o indice è un numero associato ad una variabile aleatoria o alla sua distribuzione di probabilità e sintetizza informazione d'interesse sul fenomeno rappresentato dalla variabile.

☞ Conoscere il valore di un indice significa avere informazione sulla distribuzione stessa.

Definiremo ora alcune importanti costanti caratteristiche:

1. Il **valore atteso**, che è un indice di posizione
2. La **varianza**, che è un indice di dispersione
3. I quantili di una variabile aleatoria, che contengono informazione sia sulla posizione che sulla forma di una distribuzione

Valore atteso di una variabile aleatoria discreta

Valore atteso

Se X è una v.a. discreta con valori $\{x_1, x_2, \ldots\}$ e funzione di probabilità $P(x_i) = p_i$, allora il **valore atteso** o **media** di X è

$$\mathbb{E}\left[X\right] = \sum_i x_i p_i = x_1 p_1 + x_2 p_2 + \ldots \ .$$

➔ **Esempio:** Se A è un evento qualsiasi, $\mathbf{1}_A$ è una variabile aleatoria che vale 1 con probabilità $\mathbb{P}\left[A\right]$ e 0 con probabilità $\mathbb{P}\left[\bar{A}\right]$.

- Il suo valore atteso è:

$$\mathbb{E}\left[\mathbf{1}_A\right] = \mathbb{P}\left[A\right]$$

Valore atteso di una variabile aleatoria discreta

➔ **Esempio:** Torniamo alla solita urna.

- Per la variabile $X=$ numero estratto, si ha:

$$P_X(x) = \begin{cases} 2/7 & \text{se } x = 1, 2, 3; \\ 1/7 & \text{se } x = 4; \\ 0 & \text{altrimenti} \end{cases}$$

Allora, $\mathbb{E}[X] = 1\frac{2}{7} + 2\frac{2}{7} + 3\frac{2}{7} + 4\frac{1}{7} = \frac{16}{7}$

- Per la variabile casuale che rappresenta l'importo vinto nel gioco (vincere 30 euro se la pallina estratta è bianca e perderne 20 se è nera), abbiamo

$$P_Y(y) = \begin{cases} 4/7 & \text{se } y = 30, \\ 3/7 & \text{se } y = -20, \\ 0 & \text{altrimenti} \end{cases}$$

Allora, $\mathbb{E}[Y] = 30\frac{4}{7} - 20\frac{3}{7} = \frac{60}{7}$

☞ **Esercizio:** Ricordiamo che $S =$ somma dei due numeri sulle palline estratte senza reinserimento. $\mathbb{E}[S] =$?

Valore atteso di una variabile aleatoria discreta

- **Esempio:** Consideriamo la seguente scommessa. Tu scegli un prezzo p da pagare per giocare, io lo pago e poi guadagno 1 euro se l'evento A si verifica o 0 euro se A non si verifica. Qual è il prezzo giusto della scommessa?
  - Sia X è il mio guadagno (aleatorio) nella scommessa. Allora,

$$X(\omega) = -p + 1 \cdot \mathbf{1}_A(\omega) + 0 \cdot \mathbf{1}_{\bar{A}}(\omega).$$

$$P_X(x) = \begin{cases} \mathbb{P}[A] & \text{se } x = 1 - p, \\ \mathbb{P}[\bar{A}] & \text{se } x = -p, \\ 0 & \text{altrimenti} \end{cases}$$

Un prezzo ragionevole per giocare è quello che mi dà un guadagno atteso nullo, $\mathbb{E}[X] = 0$ (altrimenti nessuno accetterebbe la scommessa). Cioè

$$(1-p)\,\mathbb{P}[A] - p\,\mathbb{P}[\bar{A}] = 0 \Leftrightarrow p = \mathbb{P}[A].$$

Valore atteso di una variabile aleatoria continua

Valore atteso

Sia X una v.a. continua con densità $f(x)$. Allora il **valore atteso** o **media** di X è:

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f(x)\, dx.$$

➔ **Esempio:** per la variabile X con densità $f(x) = 2\mathrm{e}^{-2x}\mathbf{1}_{(0,+\infty)}(x)$, si ha

$$\mathbb{E}[X] = \int_0^{+\infty} 2x\mathrm{e}^{-2x}\, dx = \frac{1}{2}.$$

Proprietà del valore atteso

Il valore atteso ha le seguenti proprietà:

1. $\mathbb{E}[a] = a$, dove a è una costante;
2. $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$, dove a e b sono costanti.

☞ In particolare $\mathbb{E}[X - \mathbb{E}[X]] = 0$

Varianza di una variabile aleatoria discreta

Varianza

Se X è una v.a. discreta con valori $\{x_1, x_2, \ldots\}$ e funzione di probabilità $P(x_i) = p_i$, allora la **varianza** di X è

$$\mathrm{Var}\left[X\right] = \sum_i (x_i - \mathbb{E}\left[X\right])^2 p_i$$

☞ Una formula pratica per il calcolo della varianza è:

$$\mathrm{Var}\left[X\right] = \sum_i x_i^2 p_i - \left[\mathbb{E}\left[X\right]\right]^2 \ .$$

Varianza di una variabile aleatoria discreta

➔ **Esempio:** Per la variabile $X =$ numero estratto dalla solita urna, abbiamo:

$$P_X(x) = \begin{cases} 2/7 & \text{se } x = 1, 2, 3; \\ 1/7 & \text{se } x = 4; \\ 0 & \text{altrimenti} \end{cases}$$

Allora, $\text{Var}\,[X] = 1\frac{2}{7} + 2^2\frac{2}{7} + 3^2\frac{2}{7} + 4^2\frac{1}{7} - \left(\frac{16}{7}\right)^2.$

☞ **Esercizio:** calcolare la varianza per le altre due variabili dell'esempio.

Varianza di una variabile aleatoria continua

Varianza

Sia X una v.a. continua con densità $f(x)$. Allora la **varianza** di X è:

$$\mathsf{Var}[X] = \int_{\mathbb{R}} (x - \mathbb{E}[X])^2 f(x)\, dx.$$

- Una formula pratica per il calcolo è

$$\mathsf{Var}[X] = \int_{\mathbb{R}} x^2 f(x)\, dx - [\mathbb{E}[X]]^2 \; .$$

- **Esempio:** per la variabile già considerata in precedenza abbiamo:

$$\begin{aligned} \mathsf{Var}[X] &= \int_0^{+\infty} 2x^2 \mathrm{e}^{-2x}\, dx - \left(\frac{1}{2}\right)^2 \\ &= \frac{1}{2} - \frac{1}{4} = \frac{1}{4}. \end{aligned}$$

Proprietà della varianza

La varianza ha le seguenti proprietà:

1. $\text{Var}[a] = 0$, dove a è una costante;
2. $\text{Var}[aX + b] = a^2\text{Var}[X]$, dove a e b sono costanti.

Valore atteso di $g(X)$

Sia $Y = g(X)$ una v.a. ottenuta trasformando la v.a. X tramite la funzione $g : \mathbb{R} \to \mathbb{R}$. Il valore atteso di Y si può calcolare anche senza conoscere direttamente la distribuzione di probabilità di Y:

- X discreta: $\mathbb{E}[Y] = \sum_i g(x_i) p_i$.
- X continua: $\mathbb{E}[Y] = \int_{\mathbb{R}} g(x) f(x)\, dx$.

☞ In particolare, la varianza di X è il valore atteso di una trasformata della X tramite la funzione $g(x) = (x - \mathbb{E}[X])^2$ e si può scrivere come

$$\text{Var}[X] = \mathbb{E}\left[X^2\right] - [\mathbb{E}[X]]^2 .$$

Moda di una variabile aleatoria

Moda

La **moda** di una variabile aleatoria X è il punto (o i punti) in cui la funzione di probabilità (o di densità) assume valore massimo.

Esempi:

- La variabile X dell'esempio dell'urna ha tre mode in 1, 2 e 3, mentre la Y ne ha una in 30. E la S?
- La variabile X con densità $f(x) = 2\mathrm{e}^{-2x}\mathbf{1}_{(0,+\infty)}(x)$, ha moda in 0.

La moda, così come il valore atteso di una v.a., è un **indice di posizione**. Altri importanti indici di posizione sono la mediana e, in generale, i quantili di una v.a.

Mediana di una variabile aleatoria

Mediana

La **mediana** di una variabile aleatoria X è il minimo valore m per cui

$$F(m) = \mathbb{P}[X \leq m] \geq \frac{1}{2}.$$

☞ Per una v.a. continua (cioè con f.r. F continua) la mediana è l'unico punto m in cui

$$F(m) = \mathbb{P}[X \leq m] = \mathbb{P}[X \geq m] = 1/2$$

Mediana di una variabile aleatoria

- **Esempio:** La variabile X dell'esempio dell'urna ha mediana in 2, mentre la Y ce l'ha in 30.
  - **Esercizio:** E la S? Disegnate i grafici delle relative f.r. e verificate quanto appena detto.

- **Esempio:** Per la variabile X con densità $f(x) = 2\mathrm{e}^{-2x}\mathbf{1}_{(0,+\infty)}(x)$, la mediana è tale che

$$1 - \mathrm{e}^{-2m} = \mathrm{e}^{-2m} = \frac{1}{2},$$

cioè $m = \log(2)/2$.

Quantili di una variabile aleatoria

Quantili

Fissato un valore $\alpha \in (0,1)$, il **quantile di livello α** di una variabile aleatoria X è il minimo valore q_α per cui

$$F(q_\alpha) = \mathbb{P}\left[X \leq q_\alpha\right] \geq \alpha.$$

- I quantili sono **indici di posizione** che generalizzano il concetto di mediana di una distribuzione.
- Per una v.a. continua (con f.r. F continua) il quantile di livello α è l'unico punto q_α in cui

$$F(q_\alpha) = \mathbb{P}\left[X \leq q_\alpha\right] = \alpha$$

Quantili di una variabile aleatoria

- La **mediana** non è altro che il quantile di livello $1/2$.
- I **percentili** sono i quantili di livello $k/100$, con $k = 1, 2, \ldots, 99$.
- I **decili** sono i quantili di livello $k/10$, con $k = 1, 2, \ldots, 9$.
- I **quartili** sono i quantili di livello 0.25 (primo quartile), 0.5 (mediana) e 0.75 (terzo quartile).

☞ **Esercizio:** Calcolare i quartili delle variabili considerate nei diversi esempi.