

## Slide Set 2

# Distribuzioni Univariate

Pietro Coretto  
pcoretto@unisa.it

### Corso di

## Analisi e Visualizzazione dei Dati (Parte I)

Corso di Laurea in "Statistica per i Big Data" (L-41)  
Università degli Studi di Salerno

Versione: 14 marzo 2022 (h08:37)

Pietro Coretto ©

Distribuzioni Univariate

1 / 19

Il punto di partenza dell'analisi è sempre un data set più o meno complesso. Nella maggior parte dei casi esso ha la forma di un 2D-array, detta anche *matrice dei dati*:

faminc	cigtax	cigprice	bwght	fatheduc	motheduc	parity	male	white	cigs	lbwght	bwghtlbs	packs	lfaming
13.5	16.5	122.3	109	12	12	1	1	1	0	4.6913481	6.8125	0	2.60269
7.5	16.5	122.3	133	6	12	2	1	0	0	4.8903489	8.3125	0	2.0149031
0.5	16.5	122.3	129		12	2	0	0	0	4.8598118	8.0625	0	-0.693147
15.5	16.5	122.3	126	12	12	2	1	0	0	4.8362818	7.875	0	2.74084
27.5	16.5	122.3	134	14	12	2	1	1	0	4.89784	8.375	0	3.3141861
7.5	16.5	122.3	118	12	14	6	1	0	0	4.7706852	7.375	0	2.0149031
65	16.5	122.3	140	16	14	2	0	1	0	4.9416418	8.75	0	4.174387
27.5	16.5	122.3	86	12	14	2	0	0	0	4.4543471	5.375	0	3.3141861
27.5	16.5	122.3	121	12	17	2	0	1	0	4.7957911	7.5625	0	3.3141861
37.5	16.5	122.3	129	16	18	2	0	1	0	4.8598118	8.0625	0	3.624341
27.5	16.5	122.3	101	12	16	2	1	0	0	4.6151199	6.3125	0	3.3141861
27.5	16.5	122.3	133	16	15	1	1	1	0	4.8903489	8.3125	0	3.3141861
6.5	16.5	122.3	61		12	3	1	0	0	4.1108742	3.8125	0	1.871802
10.5	16.5	122.3	104	12	12	1	1	1	0	4.6443911	6.5	0	2.3513751
12.5	16.5	122.3	92	7	12	1	1	0	0	4.5217891	5.75	0	2.5257289
17.5	16.5	122.3	122	13	13	1	1	1	0	4.8040209	7.625	0	2.862201
42.5	16.5	122.3	159	18	16	1	1	1	0	5.0689039	9.9375	0	3.7495041
4.5	16.5	122.3	154		11	1	0	0	0	5.036952	8.625	0	1.504077
32.5	16.5	122.3	120	16	16	1	1	0	0	4.7874918	7.5	0	3.48124
6.5	16.5	122.3	138		11	2	0	0	0	4.9272542	8.625	0	1.871802
47.5	16.5	122.3	127	16	17	1	0	1	0	4.8441868	7.9375	0	3.8607299
22.5	16.5	122.3	107	12	12	2	1	1	0	4.6728292	6.6875	0	3.1135149
15.5	16.5	122.3	129	8	11	2	1	1	0	4.8598118	8.0625	0.3	2.74084
13.5	16.5	122.3	129	12	12	1	0	0	0	4.8598118	8.0625	0	2.60269
37.5	16.5	122.3	154	11	11	1	0	1	0	5.0499562	9.75	0	3.624341

Descrizione "bw" data set:

<https://pietro-coretto.github.io/datasets/bw/readme.txt>

Notes

---

---

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

---

---

Pietro Coretto ©

Distribuzioni Univariate

2 / 19

### Analisi univariata

studia le proprietà statistiche dei dati analizzando una variabile/feature alla volta. Essa non prende in considerazione le relazioni tra le variabili.

#### Domande:

- Esiste un valore di  $b_w$  tale da sintetizzare il complesso dei dati osservati?
- Se esiste questo valore, quanto è rappresentativo per tutte le unità nel campione?
- Vi sono campioni con valori “*anomali*” rispetto alla maggioranza?
- ...

Notes

---

---

---

---

---

---

---

---

## Nozione intuitiva di distribuzione

L'oggetto “*atomico*” dell'analisi dei dati, e della statistica, è la nozione di distribuzione

Prima di tutto impareremo a caratterizzare distribuzioni univariate (DU), ovvero iniziamo a guardare il data set studiano le variabili singolarmente

Dal punto di vista matematico la nozione di DU sarà formalizzata in corsi successivi. Qui lavoreremo sulla nozione intuitiva/operativa

Una DU è un oggetto complesso determinato da una collezione di dati osservati. Per DU si intende il modo in cui questi dati si *posizionano* lungo l'intervallo/range dei valori osservati

Notes

---

---

---

---

---

---

---

---

## Notazione

- $n$ : dimensione campionaria, numero di unità statistiche osservate ( $n$  è il numero di righe del data set in forma matriciale)
- $X$ : variabile di interesse osservata. Solitamente una colonna del data set (in forma matriciale)
- $i = 1, 2, \dots, n$ : indice che identifica le unità statistiche (righe della matrice dei dati)
- $x_i$ : è il livello/label di  $X$  sulla  $i$ -ma unità statistica
- Per *raw data* intendiamo il set di dati osservati

$$\{x_1, x_2, \dots, x_n\}$$

essi corrispondono ai valori lungo la colonna  $X$  della matrice dei dati

- Indichiamo con

$$x_{\min} = \min\{x_1, x_2, \dots, x_n\}$$

$$x_{\max} = \max\{x_1, x_2, \dots, x_n\}$$

Il range dei dati (se numerici) è l'intervallo  $[x_{\min}, x_{\max}]$

Notes

---

---

---

---

---

---

---

---

---

---

Iniziamo dalla nozione intuitiva di DU. Per adesso si assume che  $X$  sia numerica

Per DU di  $X$  intendiamo il *modo* in cui questi  $n$  livelli/labels osservati della  $X$  si distribuiscono lungo il range dei dati osservati

Una prima e semplicissima visualizzazione della distribuzione di  $X$  è lo *stripchart*:

- determino il range dei dati
- rappresento ogni punto osservato  $x_i$  su un segmento che copre il range dei dati

Notes

---

---

---

---

---

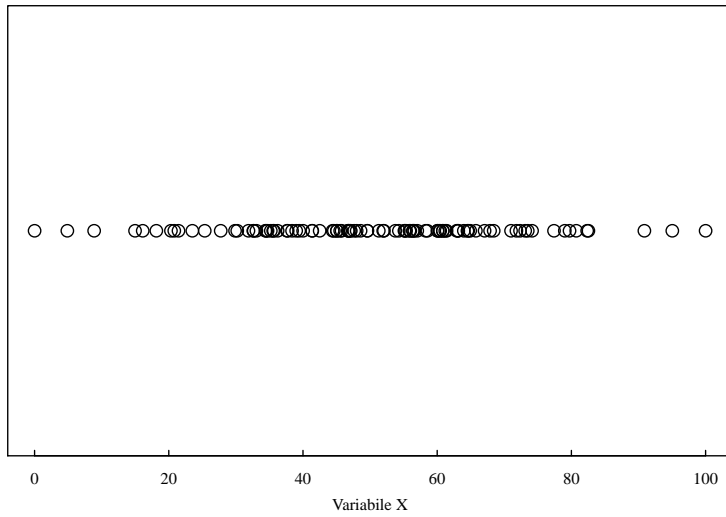
---

---

---

---

---



Notes

---

---

---

---

---

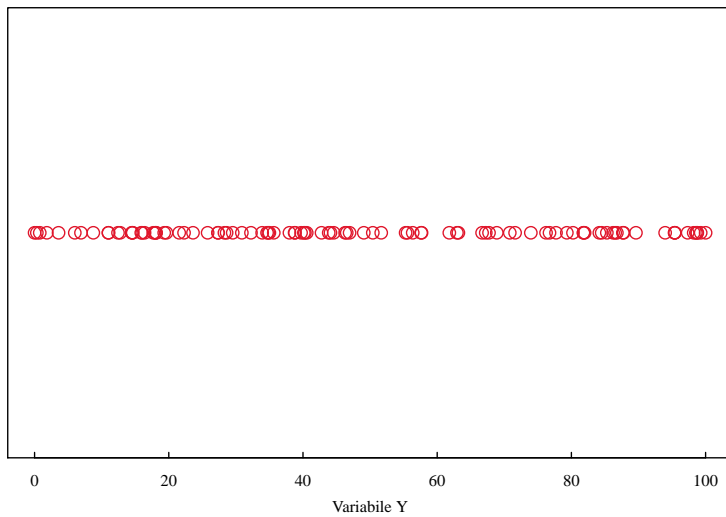
---

---

---

---

---



Notes

---

---

---

---

---

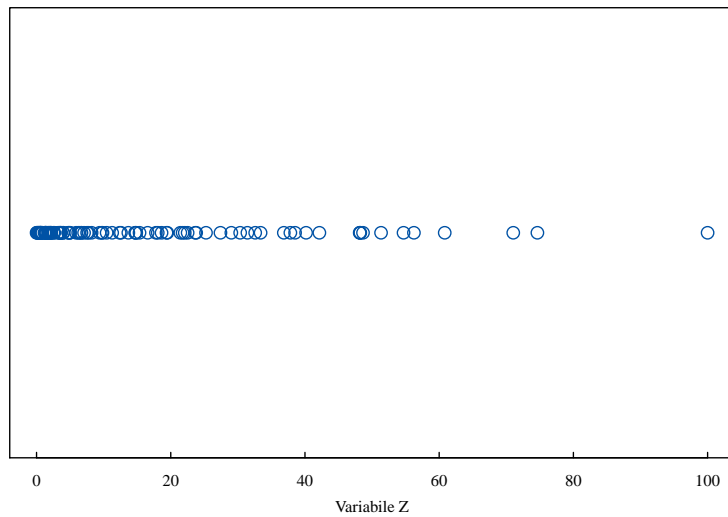
---

---

---

---

---



Notes

---

---

---

---

---

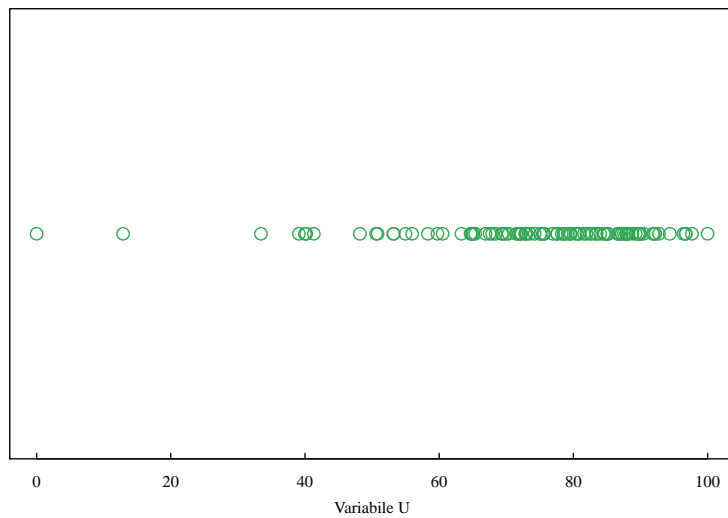
---

---

---

---

---



Notes

---

---

---

---

---

---

---

---

---

---

## Aspetti rilevanti di una DU

- **Locazione/posizione/centralità:** esiste un valore *rappresentativo*? Se questo valore esiste abbiamo identificato la *parte centrale* della distribuzione. I dati lontano dal centro li chiamiamo *code della distribuzione*
- **Dispersione/scatter:** quanto sono dispersi i dati?
- **Simmetria:** esiste un punto all'interno del range rispetto al quale la *massa* di dati si dispone in modo speculare?
- **Curtosi:** vi sono molti dati osservati lontano dalla parte centrale (ovvero nelle code)?
- **Densità:** se concentriamo l'attenzione *localmente* in un punto  $x_0$  del range, la massa di dati in quel punto quanto è *densa* di valori osservati?
- **Outliers:** esistono valori atipici/anomali? Esistono valori molto diversi dalla maggioranza dei punti osservati?

Notes

---

---

---

---

---

---

---

---

## Frequenze relative ed assolute

Sia  $X$  una variabile che rileviamo nel campione con  $K$  livelli/labels distinti  $\{x_1, x_2, \dots, x_K\}$ , definiamo:

$n_k$  è **frequenza assoluta del livello/label**  $x_k$

questo è il numero di volte che  $X = x_k$  compare nei dati osservati

$f_k = \frac{n_k}{n}$  è **frequenza relativa del livello/label**  $x_k$

questo è un numero in  $[0,1]$  che esprime la porzione di campioni osservati per i quali  $X = x_k$ .

Nota: posso esprimere le frequenze relative in termini percentuali come

$$f_k = \frac{n_k}{n} \times 100$$

Per costruzione

- $\sum_{k=1}^K n_k = n$
- $\sum_{k=1}^K f_k = 1$

Notes

---

---

---

---

---

---

---

---

### Esempio 1

Su  $n = 5$  unità si osserva  $X = \text{sex}$ , i dati osservati sono

$$\{F, F, M, F, F\}$$

In questo caso abbiamo  $K = 2$  labels distinti:  $x_1 = F, x_2 = M$ . Le frequenze assolute sono

$$n_1 = 4, \quad n_2 = 1, \quad f_1 = \frac{4}{5} = 0.8, \quad f_2 = \frac{1}{5} = 0.2$$

### Esempio 2

Su  $n = 6$  unità si osserva  $Y = \text{numero di scarpe}$ , i dati osservati sono

$$\{38, 40, 40, 39, 38, 40\}$$

In questo caso abbiamo  $K = 3$  livelli distinti:  $x_1 = 38, x_2 = 39, x_3 = 40$ .  
Le frequenze assolute sono

$$n_1 = 2, \quad n_2 = 1, \quad n_3 = 3$$
$$f_1 = \frac{2}{6} = 0.33, \quad f_2 = \frac{1}{6} = 0.17, \quad f_3 = \frac{3}{6} = 0.5$$

Notes

---

---

---

---

---

---

---

---

Possiamo calcolare le frequenze in questo modo per qualsiasi tipo di variabile? Pensiamo al caso in cui  $X$  è continua

Consideriamo la variabile  $X = \text{faminc}$  nel data set *bw*. ci sono 27 diversi livelli osservati, molti di essi estremamente simili tra di loro. I primi valori osservati sono

$$2.6026900, \quad 2.0149031, \quad -0.6931472, \quad 2.7408400, \quad 3.3141861, \dots$$

Dividiamo il range in  $K$  intervalli di livelli (anche detti classi di livelli o bins), e calcolo le frequenze rispetto al numero di osservazioni in ogni classe di livello.

Questo metodo appartiene ad una classe più generale di *metodi di windowing*: i dati vengono divisi e analizzati in tante piccole *finestre*.

Notes

---

---

---

---

---

---

---

---

Problema: come determino queste classi di valori? Approcci:

- Scelgo gli intervalli in modo discrezionale
- Fisso  $\Delta$ , l'ampiezza dell'intervallo, e determino  $K$  di conseguenza
- Fisso  $K$  e determino l'ampiezza  $\Delta$

Esistono regole *ottimali* per ottenere classi di ampiezza uniforme. Ad esempio il metodo di *Sturge*:

- $K = 1 + \lceil \log_2(n) \rceil$
- $\Delta = \frac{x_{\max} - x_{\min}}{K}$
- Primo intervallo  $\rightarrow [x_{\min}, x_{\min} + \Delta]$   
Secondo intervallo  $\rightarrow (x_{\min} + \Delta, x_{\min} + 2\Delta]$   
...  
k-esimo intervallo  $\rightarrow (x_{\min} + (k-1)\Delta, x_{\min} + k\Delta]$

**Esempi/Applicazioni**  $\rightarrow$  [R script file](#)

Notes

---

---

---

---

---

---

---

---

---

---


Se  $K$  non è troppo grande, le frequenze consentono di ottenere una presentazione compatta della distribuzione dei dati

Distribuzione semplice

$X$	$n_i$	$f_i$
$x_1$	$n_1$	$f_1$
$x_2$	$n_2$	$f_2$
$\vdots$	$\vdots$	$\vdots$
$x_K$	$n_K$	$f_K$
	$n$	1

Distribuzione in classi

$X$	$n_i$	$f_i$
$x_0 \mid - \mid x_1$	$n_1$	$f_1$
$x_1 \mid - \mid x_2$	$n_2$	$f_2$
$\vdots$	$\vdots$	$\vdots$
$x_{K-1} \mid - \mid x_K$	$n_K$	$f_K$
	$n$	1

- $X$  è quantitativo, discreto, ed espresso con troppi livelli distinti: potrebbe avere senso a costruire le classi
-  la suddivisione in classi produce perdita di informazione

Notes

---

---

---

---

---

---

---

---

---

---



**Esempio:** consideriamo la variabile (continua) `faminc` dal data set `bw`.  
Prendiamo  $K = 3$  classi uniformi:

<code>faminc</code> [ $10^3 \times \text{USD}$ ]	$n_i$	$f_i(\%)$
[0.435, 22]	526	37.9
(22, 43.5]	602	43.4
(43.5, 65.1]	260	18.7
	1388	100

**Esempio:** consideriamo la variabile (categoriale/nominale) `male` dal data set `bw`.

<code>male</code>	$n_i$	$f_i(\%)$
Femmina (=0)	665	47.9
Maschio (=1)	723	52.1
	1388	100

**Esempi/Applicazioni** → [R script file](#)

Grafico	Tipo di variabile				
	Continua	In Classi(*)	Discreta(*)	Ordinale	Nominale
stripchart	+	no	no	no	no
barplot	no	no	+	+	+
piechart	no	no	ni	+	+
istogramma	no	+	no	no	no

**Nota (\*):**

- “*In Classi*” include il caso delle variabili continue trasformate in classi, e le variabili discrete che sono espresse con troppi livelli distinti.
- “*Discreta*” include il caso di variabili discrete con un numero ragionevole di livelli distinti.

**Esempi/Applicazioni** → **R script file** (stripchart, barplot, piechart)

[illegible][illegible]