

# Statistica I

Unità B: Distribuzione di frequenza

**Tommaso Rigon**

**Università Milano-Bicocca**



## Argomenti affrontati

- Frequenze assolute, relative e cumulate
- Istogramma
- Funzione di ripartizione empirica

## Riferimenti al libro di testo

- §3.1 — §3.5
- **Nota.** Il concetto di densità di frequenza verrà affrontato nell'Unità G.

# Il problema epidemiologico

- Il **DDT** è estremamente efficace contro le zanzare da malaria ed è pertanto largamente usato in zone in cui la malaria è endemica.
- Al tempo stesso, il DDT potrebbe costituire un **rischio per la salute**, specialmente nel caso di **donne in gravidanza**.
- Per un campione di 2312 donne in gravidanza, viene misurato il DDE, ovvero una sostanza connessa al DDT, presente nel siero materno durante il terzo trimestre della gravidanza.
- Osserviamo inoltre che 361 donne hanno **partorito prematuramente**, ovvero prima della conclusione della 37a settimana.
- **Domanda di ricerca**: la quantità di DDE è maggiore tra donne che hanno partorito prematuramente?

# I dati grezzi

DDE (mg/L), parto non prematuro. Numero di osservazioni = 1951

```
[1] 24.56 15.56 15.00 33.54 22.68 25.02 31.85 37.45 32.27 31.43
[11] 15.23 31.23 54.39 18.11 79.70 15.68 29.43 12.62 22.92  9.51
[21] 10.94 23.16 10.51 13.82 26.80 17.91 88.65 23.90 16.42 23.47
[31] 20.42 15.94 38.61 34.17 22.60 24.69 40.34 47.29 14.62 22.53
[41] 19.86 17.40 42.06 10.75 11.14 31.81 21.51 12.52 18.54 24.38
...
```

DDE (mg/L), parto prematuro. Numero di osservazioni = 361

```
[1] 54.80 27.37 28.01  6.34  6.28 25.61 25.02 13.08 54.98 34.86
[11] 46.00 19.30 21.98 22.19 31.24 19.94 15.12 24.64 11.91 70.04
[21] 59.52 94.60 19.89 21.95 15.18 27.94 29.46 60.24 37.82 28.07
[31] 23.71 15.09 23.36 17.11 15.38 33.06 19.76 27.49 12.11 12.66
[41] 15.72 36.04 18.01 25.88 76.84 19.94 25.59 45.22 30.75 31.02
...
```

# Organizzazione dei dati in frequenze

- I dati non sono “tantissimi” rispetto ad altre situazioni.
- Infatti, ci sono “solamente”  $n = 361 + 1951 = 2312$  donne in gravidanza.
- Sono però troppi per capire qualcosa solamente “guardandoli”, come potremmo fare “guardando” i voti in un libretto universitario. Dobbiamo quindi cercare di **sintetizzarli**.
- Potremmo quindi suddividere l'intervallo che contiene tutti i valori osservati (ovvero  $(0, 180]$ ) in un certo numero di **sotto-intervalli** e poi semplicemente nel **contare** quante osservazioni cadono nei vari sotto-intervalli.
- Questa operazione viene fatta nella tabella seguente, utilizzando 10 sotto-intervalli di lunghezza 18, chiusi a destra.

# Frequenze assolute

| DDE (mg/L) | Nascita non prematura | Nascita prematura |
|------------|-----------------------|-------------------|
| (0,18]     | 573                   | 68                |
| (18,36]    | 906                   | 164               |
| (36,54]    | 308                   | 65                |
| (54,72]    | 91                    | 34                |
| (72,90]    | 40                    | 14                |
| (90,108]   | 19                    | 10                |
| (108,126]  | 6                     | 3                 |
| (126,144]  | 5                     | 1                 |
| (144,162]  | 2                     | 1                 |
| (162,180]  | 1                     | 1                 |
| Totale     | 1951                  | 361               |

# Commenti alla tabella

- La prima colonna mostra i sotto-intervalli utilizzati. Le altre mostrano il **numero di donne** la cui dose di DDE appartiene al sotto-intervallo considerato.
- Ad esempio, il 68 che compare nella prima riga alla terza colonna indica che esattamente 68 donne, delle 361 che hanno partorito prematuramente, hanno una dose di DDE (mg/L) strettamente maggiore di 0 e minore o uguale a 18.
- Le ultime due colonne contengono le **frequenze assolute**.
- Le prime due colonne (intervalli + parto non prematuro) mostrano come le donne sono “distribuite” nei vari intervallini. Quando prese congiuntamente, queste due colonne sono chiamate la **distribuzione di frequenza**.
- **Nota**. Dalla tabella delle frequenze assolute non è ancora molto chiaro se, in termini di DDE, ci sia una differenza tra nascite premature e non.

# Frequenze assolute

- Siano  $x_1, \dots, x_n$  i valori assunti da una variabile per tutte le  $n$  unità statistiche.
- Siano  $c_1, \dots, c_k$  invece
  - le  $k$  distinte modalità relative ai dati discreti  $x_1, \dots, x_n$  (**Variabile discreta**);
  - i  $k$  sotto-intervalli in cui abbiamo diviso i valori numerici  $x_1, \dots, x_n$  (**Variabile continua**).
- **Frequenze assolute**. Il numero di volte  $n_1, \dots, n_k$  che i valori distinti  $c_1, \dots, c_k$  compaiono nei dati  $x_1, \dots, x_n$  si chiamano frequenze assolute.
- Le frequenze assolute  $n_1, \dots, n_k$  sono quindi numeri interi non-negativi caratterizzati dalle proprietà

$$n_1 + n_2 + \dots + n_k = \sum_{j=1}^k n_j = n, \quad 0 \leq n_j \leq n, \quad j = 1, \dots, k.$$

- **Nota**. Alcuni libri di testo usano  $x_1, \dots, x_k$  per indicare i valori distinti. Questo potrebbe fare confusione, per questo si è preferito usare  $c_1, \dots, c_k$ .

# Per capire la notazione

| DDE (mg/L)            | Nascita non prematura            |
|-----------------------|----------------------------------|
| $c_1 = (0, 18]$       | $n_1 = 573$                      |
| $c_2 = (18, 36]$      | $n_2 = 906$                      |
| $c_3 = (36, 54]$      | $n_3 = 308$                      |
| $c_4 = (54, 72]$      | $n_4 = 91$                       |
| $c_5 = (72, 90]$      | $n_5 = 40$                       |
| $c_6 = (90, 108]$     | $n_6 = 19$                       |
| $c_7 = (108, 126]$    | $n_7 = 6$                        |
| $c_8 = (126, 144]$    | $n_8 = 5$                        |
| $c_9 = (144, 162]$    | $n_9 = 2$                        |
| $c_{10} = (162, 180]$ | $n_{10} = 1$                     |
| Totale                | $n = \sum_{j=1}^{10} n_j = 1951$ |

- Ricordando i dati riportati in precedenza, abbiamo invece che  $x_1 = 24.56$ ,  $x_2 = 15.56$ ,  $x_3 = 15.00$ ,  $x_4 = 33.54$ , e via dicendo.

# Frequenze relative

- Le frequenze assolute non chiariscono se il DDE sia legato ai tempi di gravidanza. Le **numerosità campionarie** delle due distribuzioni di frequenze sono **diverse**.
- Per un confronto più equo, possiamo usare le **frequenze relative**, ovvero

$$(\text{frequenze relative}) = \frac{(\text{frequenze assolute})}{(\text{numerosità campionaria})}.$$

- Frequenze relative.** Siano  $n_1, \dots, n_k$  delle frequenze assolute, allora le frequenze relative  $f_1, \dots, f_k$  sono pari a:

$$f_j = \frac{n_j}{n}, \quad j = 1, \dots, k.$$

- A volte le frequenze relative vengono moltiplicate per 100, in tal caso parleremo di **frequenze percentuali**.
- Esercizio.** Mostrare che le frequenze relative  $f_1, \dots, f_k$  sono numeri reali non-negativi caratterizzati dalle proprietà

$$f_1 + f_2 + \dots + f_k = \sum_{j=1}^k f_j = 1, \quad 0 \leq f_j \leq 1, \quad j = 1, \dots, k.$$

# Frequenze relative $f_1, \dots, f_k$

| DDE (mg/L) | Nascita non prematura | Nascita prematura |
|------------|-----------------------|-------------------|
| (0,18]     | 0.294                 | 0.188             |
| (18,36]    | 0.464                 | 0.454             |
| (36,54]    | 0.158                 | 0.180             |
| (54,72]    | 0.047                 | 0.094             |
| (72,90]    | 0.021                 | 0.039             |
| (90,108]   | 0.010                 | 0.028             |
| (108,126]  | 0.003                 | 0.008             |
| (126,144]  | 0.003                 | 0.003             |
| (144,162]  | 0.001                 | 0.003             |
| (162,180]  | 0.001                 | 0.003             |
| Totale     | 1                     | 1                 |

- Le distribuzioni di frequenze relative sono confrontabili e mostrano che un **basso** dosaggio di DDE è associato a un **minor** numero di parti prematuri (e viceversa). Infatti, confrontando ad esempio i valori della prima riga si nota che  $0.294 > 0.188$ .
- **Esercizio**. Ricalcolare almeno un paio di queste frequenze relative.

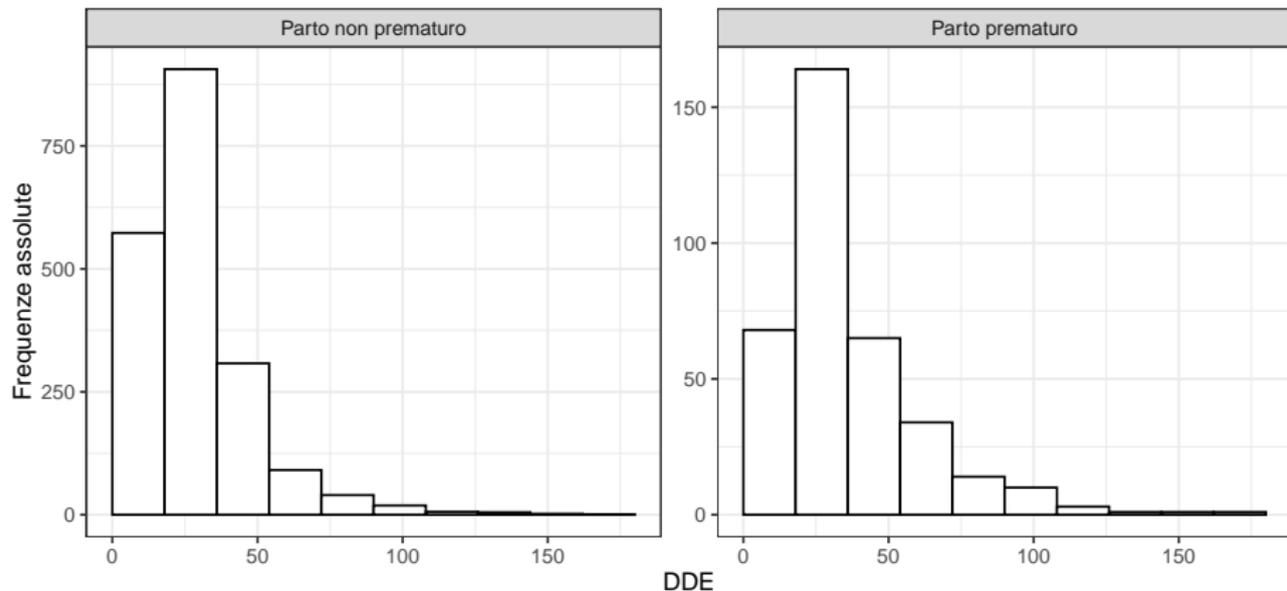
- Le differenze tra le distribuzioni di frequenza sono ancora più evidenti se rappresentate graficamente.
- Una possibilità è utilizzare un **istogramma**. Ne presentiamo qui una versione semplificata (ci torneremo in seguito!)
- Costruiamo il grafico ponendo

(base rettangoli) = (sotto-intervalli)

(altezza rettangoli) = (frequenze assolute)

- Alcuni aspetti che chiariremo in seguito:
  - Quanti intervalli scegliere?
  - Come gestire intervalli di lunghezze diverse?
  - È possibile rappresentare un istogramma che faccia uso di frequenze “relative”?

# Istogramma



- La distribuzione di “parto prematuro” è spostata più a destra, ovvero associata a dosi maggiori di DDE.
- **Esercizio.** Ricostruire “a mano” gli istogrammi rappresentati qui sopra.

# Funzione di ripartizione empirica

- Una seconda rappresentazione grafica di uso frequente è la cosiddetta **funzione di ripartizione empirica**  $F(x)$ , ovvero

$$\left( \begin{array}{l} \text{funzione di ripartizione} \\ \text{empirica calcolata in } x \end{array} \right) = \frac{(\text{numero di osservazioni minori o uguali di } x)}{(\text{numerosità campionaria})}$$

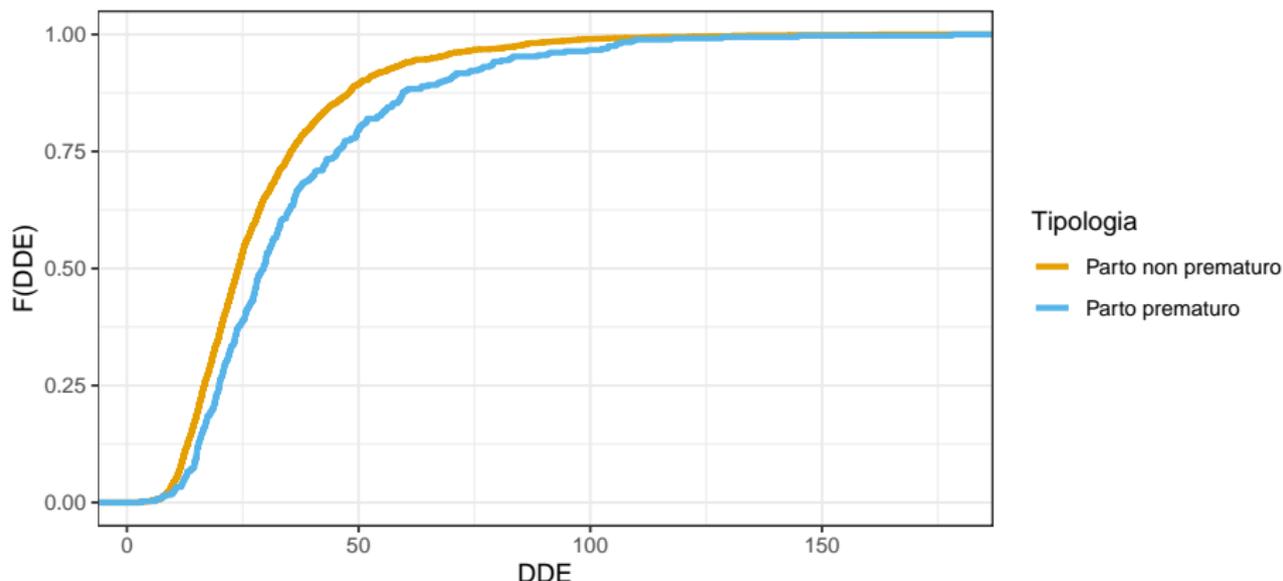
- **Funzione di ripartizione empirica.** Siano  $x_1, \dots, x_n$  una collezione di dati, allora

$$F(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(x_i \leq x),$$

dove  $\mathbb{1}(x_i \leq x)$  si chiama **funzione indicatrice** e vale 1 se  $x_i \leq x$  e 0 se  $x_i > x$ .

- **Esercizio.** Ci si convinca che la definizione informale e quella matematicamente più rigorosa sono equivalenti.

# Funzione di ripartizione empirica



- Il “messaggio” può forse sembrare meno evidente di quello contenuto negli istogrammi.
- Lo studente guardi però la definizione precedente e il grafico fino a che non si convince che il “messaggio” è il medesimo.

# Funzione di ripartizione empirica

- Operativamente, la funzione  $F(x)$  si può ottenere dai dati  $x_1, \dots, x_n$  come segue.
- In primo luogo, si ottengono i **dati ordinati**  $x_{(1)}, \dots, x_{(n)}$  a partire dal valore minimo  $x_{(1)}$  fino al massimo  $x_{(n)}$ .
- Per un certo valore di  $x$  avremo quindi che

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x \leq \dots \leq x_{(n-1)} \leq x_{(n)}.$$

- Il valore di  $F(x)$  è la frazione di dati ordinati “a sinistra”, ovvero più piccoli, di  $x$ .
- **Proprietà** della funzione di ripartizione empirica:

$$0 \leq F(x) \leq 1, \quad \lim_{x \rightarrow -\infty} F(x) = 0, \quad \lim_{x \rightarrow +\infty} F(x) = 1,$$

$F(x)$  è non decrescente,  $F(x)$  è continua a destra.

# Frequenze cumulate

- Le frequenze cumulate si ottengono sommando progressivamente le frequenze e quindi conteggiano il numero (o la frazione) di dati minori di una certa soglia.
- **Frequenze cumulate assolute.** Siano  $n_1, \dots, n_k$  delle frequenze assolute, allora le frequenze cumulate assolute  $N_1, \dots, N_k$  sono pari a:

$$N_j = n_1 + \dots + n_j = \sum_{j'=1}^j n_{j'}, \quad j = 1, \dots, k.$$

- **Frequenze cumulate relative.** Siano  $f_1, \dots, f_k$  delle frequenze relative, allora le frequenze cumulate relative  $F_1, \dots, F_k$  sono pari a:

$$F_j = f_1 + \dots + f_j = \sum_{j'=1}^j f_{j'}, \quad j = 1, \dots, k.$$

- **Esercizio.** Si mostri che  $N_k = n$  e che  $F_k = 1$ .

# Frequenze cumulate relative

| DDE (mg/L) | Nascita non prematura | Nascita prematura |
|------------|-----------------------|-------------------|
| (0,18]     | 0.294                 | 0.188             |
| (18,36]    | 0.758                 | 0.643             |
| (36,54]    | 0.916                 | 0.823             |
| (54,72]    | 0.963                 | 0.917             |
| (72,90]    | 0.983                 | 0.956             |
| (90,108]   | 0.993                 | 0.983             |
| (108,126]  | 0.996                 | 0.992             |
| (126,144]  | 0.998                 | 0.994             |
| (144,162]  | 0.999                 | 0.997             |
| (162,180]  | 1                     | 1                 |

- Le frequenze cumulate relative sono strettamente connesse alla funzione di ripartizione empirica. Si ragioni sulla loro definizione.
- Ad esempio, dalla tabella si ottiene che  $F(18) = 0.294$  e che  $F(36) = 0.758$  nel caso di nascita non prematura.
- **Nota.** Dalla tabella non è tuttavia possibile calcolare ad esempio  $F(20)$ . Come mai?