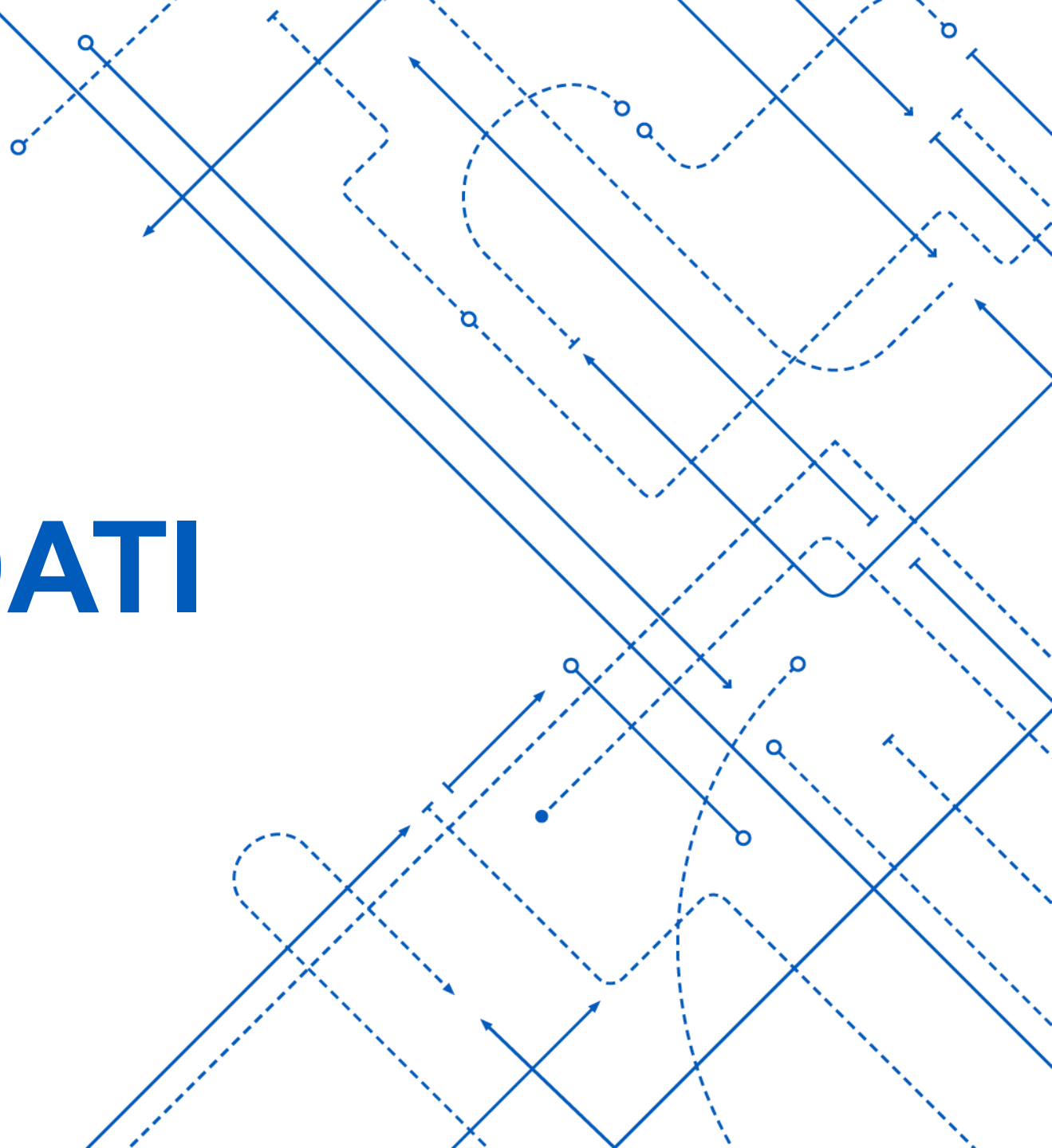
STATISTICA E ANALISI DEI DATI

Capitolo 4 – Statistica Descrittiva

Dott. Stefano Cirillo
Dott. Luigi Di Biasi

a.a. 2025-2026



STATISTICA DESCRITTIVA

Con metodi di natura logica e matematica

- Osservare il mondo in modo **formale**…
  - Il mondo è pieno di **fenomeni osservabili**!

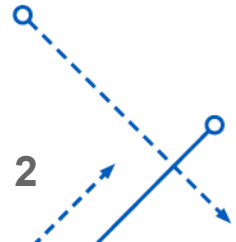
- natura economica, industriale, sociale
- comportamenti o situazioni riguardanti singoli o insiemi individui

  - Abbiamo imparato ad analizzare questi fenomeni **mediante l'osservazione del comportamento** dalle **caratteristiche «visibili (osservabili appunto)»** degli stessi.
    - Come **osserviamo *«come si comportano»*** le **caratteristiche** di un fenomeno?

Possiamo **misurare/contare/osservare** i valori assunti da X

Possiamo associare una variabile ad ogni caratteristica osservabile di un fenomeno **(ad esempio … X)**

In statistica descrittiva **le variabili** possono essere di **tre tipi**: Qualitative, Quantitative, Ordinabili

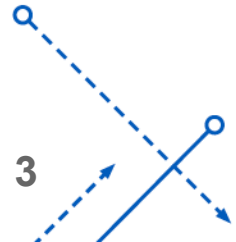


STATISTICA DESCRITTIVA

Uno degli obiettivi della statistica descrittiva è quello di **fornire una comprensione chiara e sintetica** dei dati

- Cos'è la **statistica descrittiva**?
  - è una branca della statistica che si concentra sulla **raccolta**, **elaborazione**, **analisi**, **interpretazione**, **organizzazione** e sulla **presentazione dei dati** in modo da **riassumere e descrivere le principali caratteristiche** di un insieme di dati.
  - **Inoltre**, permette di **estendere la descrizione di certi fenomeni** osservati **ad altri fenomeni dello stesso tipo non ancora osservati**!

- Se uno studente ha la media del «25», che voto «possiamo aspettarci» potrebbe raggiungere al prossimo esame?
- Esiste una «regolarità» nell'andamento dei voti degli studenti nei corsi di programmazione, rispetto ai voti dei corsi di architettura?



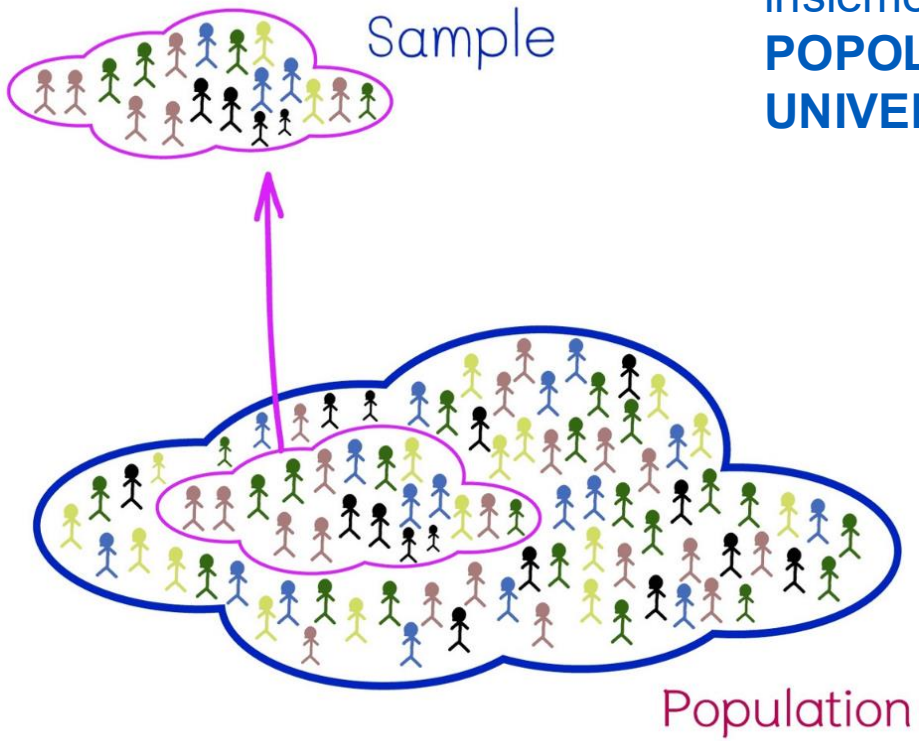
STATISTICA DESCRITTIVA

- Il mondo è pieno di fenomeni osservabili!
  - Tuttavia, un'analisi statistica è sempre effettuato su un **insieme di entità** (individui, oggetti, voti) in cui **si manifesta il fenomeno che si vuole studiare**.

Formalmente questo insieme si chiama **POPOLAZIONE** o **UNIVERSO**.

Questo insieme può essere **FINITO** o **INFINITO**.

Possiamo dunque avere una popolazione **FINITA** o una popolazione **ILLIMITATA**.

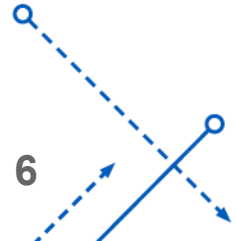


POPOLAZIONE FINITA O LIMITATA

- Popolazione limitata: Una popolazione è considerata limitata quando è composta da un numero finito di elementi
  - La conoscenza delle caratteristiche di una **popolazione finita** può essere ottenuta osservando la totalità delle entità della popolazione **(CASO OTTIMO),** oppure un sottoinsieme di questa, detto **CAMPIONE** estratto dalla popolazione
- Esempio:
  - **Numero di automobili prodotte in un anno da un'azienda**: Si sa esattamente quante automobili sono state prodotte (un numero finito), quindi si tratta di una popolazione limitata.
  - **Studenti iscritti a un corso universitario**: Il numero di studenti iscritti a un corso è finito, quindi è un esempio di popolazione limitata.
  - **Numero di famiglie in un quartiere**: Se si vuole analizzare il reddito di tutte le famiglie di un determinato quartiere, il numero di famiglie è finito, rendendo la popolazione limitata.
  - **Atleti che partecipano a una competizione**: Se si vuole analizzare le prestazioni di tutti gli atleti in una competizione, il numero di partecipanti è noto e finito.

POPOLAZIONE INFINITA O ILLIMITATA

- **Popolazione illimitata**: Una popolazione è illimitata quando si ritiene che il numero di elementi sia infinito o teoricamente infinito
- Cosa accade con una **popolazione illimitata?**
  - Una popolazione illimitata può essere studiata **tramite un CAMPIONE estratto dalla popolazione**
  - **VINCOLI sul Campione:**
    - Il campione deve rappresentare statisticamente TUTTA la popolazione!
    - Occorre scegliere gli elementi in modo completamente casuale poiché ogni criterio di selezione non casuale rischia di produrre campioni sbilanciati verso particolari valori



POPOLAZIONE INFINITA O ILLIMITATA

- **Popolazione illimitata**: Una popolazione è illimitata quando si ritiene che il numero di elementi sia infinito o teoricamente infinito
- Cosa accade con una **popolazione illimitata?**
  - Una popolazione illimitata può essere studiata **tramite un CAMPIONE estratto dalla popolazione**
  - Esempio:
    - **Risultati del lancio di un dado:** Se si vuole studiare la distribuzione dei risultati in infiniti lanci di un dado, la popolazione è teoricamente illimitata
    - **Misurazioni della temperatura in un determinato luogo:** Se si vuole analizzare le temperature registrate ogni giorno in una città senza limiti di tempo, il numero di misurazioni potenziali è illimitato
    - **Produzione giornaliera di un macchinario:** Se si studia la produzione potenziale di un macchinario che funziona senza interruzioni, il numero di unità prodotte può essere considerato illimitato nel tempo

STATISTICA DESCRITTIVA

- **<u>Prima di iniziare</u>** una elaborazione statistica e descrittiva di dati…

Continuo? Discreto?

  - Lo **scopo** dell'analisi che si intende eseguire ☺
  - Bisogna conoscere almeno a grandi linee **la natura del fenomeno** in esame;
  - E' necessario conoscere il numero di osservazioni del fenomeno disponibili (**ampiezza del campione**);
  - E' necessario conoscere **il numero di variabili osservabili**
  - il numero di variabili utilizzate per rappresentare i diversi **aspetti** del fenomeno in esame
  - Dobbiamo conoscere il **tipo di ogni variabile osservata** (Continua? Discreta? Quantitativa? Qualitativa?)

STATISTICA DESCRITTIVA

- Quali strumenti offre la **statistica descrittiva**?
  - **Funzioni di distribuzione;**
  - **Misure di centralità e Indici di Sintesi**
  - Misure di dispersione
  - **Grafici e tabelle;** ✓
  - Misura della forma della distribuzione;
  - **Percentili e quantili**;✓
  - **Frequenze e percentuali**;✓

Li vedremo nelle prossime lezioni!

STATISTICA E ANALISI DEI DATI

Capitolo 4 – Funzione di Distribuzione Empirica

Dott. Stefano Cirillo
Dott. Luigi Di Biasi

a.a. 2025-2026

FUNZIONE DI DISTRIBUZIONE EMPIRICA

- Per i fenomeni quantitativi è spesso utile definire la **funzione di distribuzione empirica (FDE)**
- La **funzione di distribuzione empirica (FDE)** è uno strumento utile per rappresentare graficamente e quantitativamente come un insieme di dati si distribuisce, senza fare assunzioni su una particolare distribuzione teorica
  - Distribuzione normale (o gaussiana)
  - Distribuzione uniforme
  - Distribuzione esponenziale
  - Distribuzione binomiale
  - …
- A cosa serve?
  - Ci permette di capire come si distribuiscono i dati osservati nel **"fenomeno"** per studiarne le "caratteristiche" ed il comportamento delle modalità dei caratteri del fenomeno osservato
- La funzione di distribuzione si può occupare di variabili **discrete** e **continue**
  - Le funzioni di **distribuzione empirica** sia **discreta** (**FdDD**) che **continua** (**FdDC**) da usare a seconda del fenomeno da studiare

FdD Discreta

- Nel caso discreto questa funzione è definita a partire dalle **frequenze relative cumulative**
  - Consideriamo una generica variabile quantitativa X
  - Assumiamo che i valori assunti da X siano $\{z_1, z_2, \dots, z_k\}$
  - Assumiamo anche che i valori siano ordinati ovvero che $z_1 < z_2 < \dots < z_k$
  - Consideriamo poi un **campione** C costituito da n osservazioni di X
  - Indichiamo con $C = (x_1, x_2, \dots, x_n)$ i valori del campione
  - Denotiamo con n_i il **numero di volte (frequenza assoluta)** in cui ciascun valore z_i è presente nel campione
  - Denotiamo con $f_i = \frac{n_i}{n}$ la frequenza relativa con cui compare nel campione z_i
- Definiamo $F_i = f_1, f_2, \dots, f_i$ con il nome di **frequenza relativa comulativa:**

$$F_i = f_1 + f_2 + \ldots + f_i = \frac{n_1 + n_2 + \ldots + n_i}{n} \qquad (i = 1, 2, \ldots, k),$$

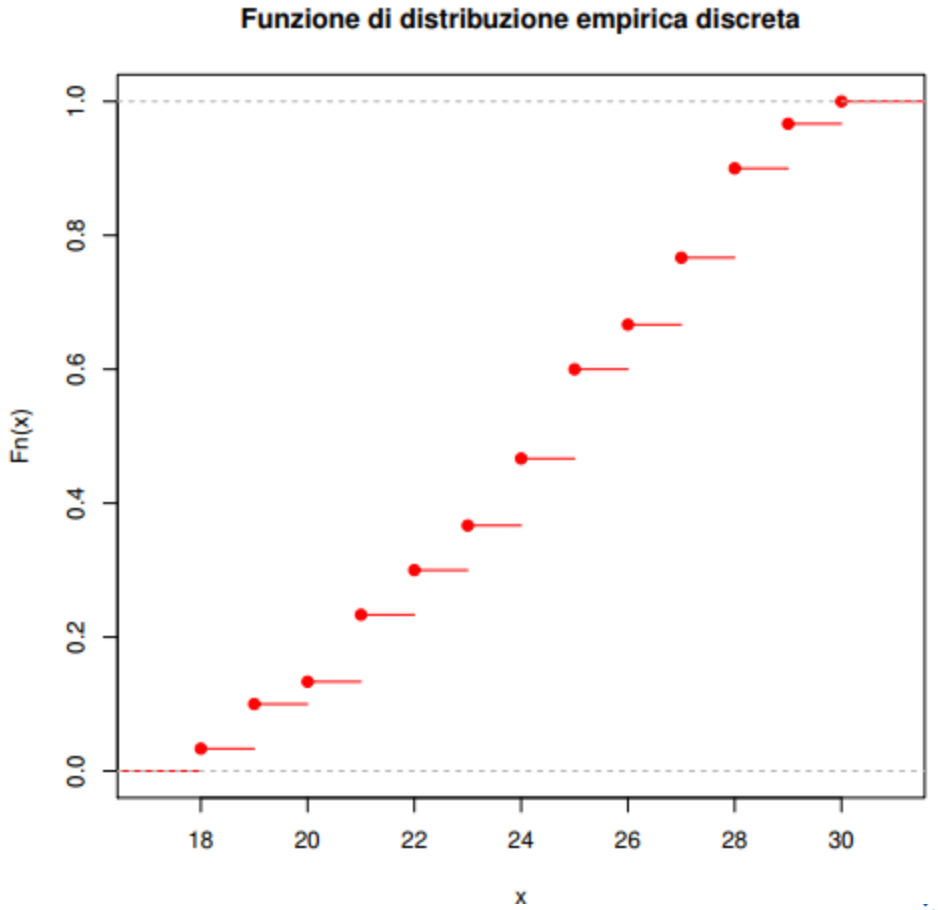
- La generica F_i rappresenta **la proporzione dei dati del campione C che assumono valori minori o uguali a** z_i $(P(F_i \leq z_i))$

FdD Discreta

- Poiché abbiamo assunto che i valori assunti da X siano ordinati $z_1 < z_2 < \ldots < z_k$
- Definiamo la **FdDD** come:

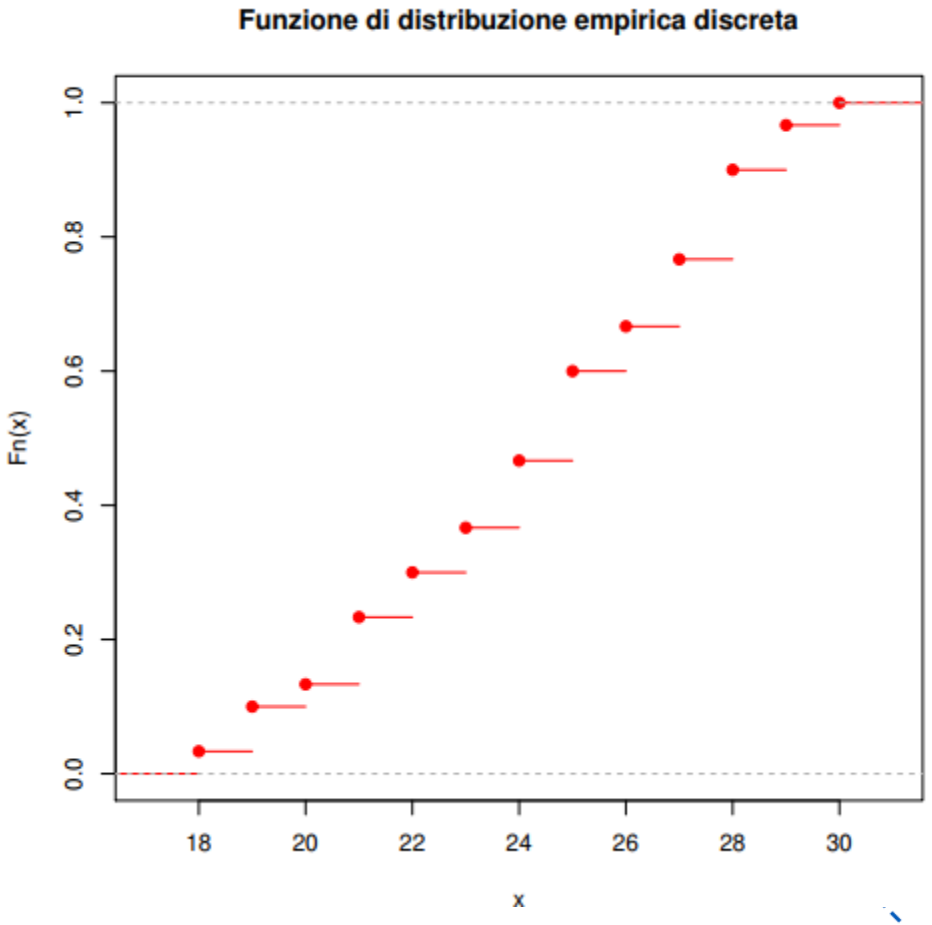
$$F(x) = \frac{\#\{x_i \leq x, i = 1, 2, \ldots, n\}}{n} = \begin{cases} 0, & x < z_1 \\ F_1, & z_1 \leq x < z_2 \\ \ldots & \\ F_i, & z_i \leq x < z_{i+1} \\ \ldots & \\ 1, & x \geq z_k \end{cases}$$

- Dove # indica la cardinalità dell'insieme



PROPRIETÀ DELLA FDD DISCRETA

- Caratteristiche della Funzione di Distribuzione Discreta:
  - E' una funzione a gradino
  - E' una funzione non decrescente
  - La funzione assume il valore a sinistra in corrispondenza di ogni punto di salto;
  - La funzione vale 0 per ogni valore minore dell'osservazione minima
  - La funzione vale 1 per ogni valore maggiore o uguale dell'osservazione massima



ESEMPIO FDD DISCRETA

Esempio: Consideriamo i voti conseguiti da 30 studenti all'esame di SAD nel 2022

X = {18,19,19,20,21,21,21,22,22,23,23,24,24,24,25,25,25,25,26,26,27,27,27,28,28, 28, 28,29,29,30}

i
1
2
3
4
5
6
7
8
9
10
11
12
13

Indici

ESEMPIO FDD DISCRETA

Esempio: Consideriamo i voti conseguiti da 30 studenti all'esame di SAD nel 2022

X = {18,19,19,20,21,21,21,22,22,23,23,24,24,24,25,25,25,25,26,26,27,27,27,28,28, 28, 28,29,29,30}

i	z_i
1	18
2	19
3	20
4	21
5	22
6	23
7	24
8	25
9	26
10	27
11	28
12	29
13	30

Indici

Valori Voti

ESEMPIO FDD DISCRETA

Esempio: Consideriamo i voti conseguiti da 30 studenti all'esame di SAD nel 2022

X = {18,19,19,20,21,21,21,22,22,23,23,24,24,24,25,25,25,25,26,26,27,27,27,28,28, 28, 28,29,29,30}

i	z_i	n_i
1	18	1
2	19	2
3	20	1
4	21	3
5	22	2
6	23	2
7	24	3
8	25	4
9	26	2
10	27	3
11	28	4
12	29	2
13	30	1

Frequenze Assolute

Valori Voti

Indici

ESEMPIO FDD DISCRETA

Esempio: Consideriamo i voti conseguiti da 30 studenti all'esame di SAD nel 2022

X = {18,19,19,20,21,21,21,22,22,23,23,24,24,24,25,25,25,25,26,26,27,27,27,28,28, 28, 28,29,29,30}

i	z_i	n_i	f_i
1	18	1	1/30
2	19	2	2/30
3	20	1	1/30
4	21	3	3/30
5	22	2	2/30
6	23	2	2/30
7	24	3	3/30
8	25	4	4/30
9	26	2	2/30
10	27	3	3/30
11	28	4	4/30
12	29	2	2/30
13	30	1	1/30

Indici

Valori Voti

Frequenze Assolute

Frequenze Relative

ESEMPIO FDD DISCRETA

Esempio: Consideriamo i voti conseguiti da 30 studenti all'esame di SAD nel 2022

X = {18,19,19,20,21,21,21,22,22,23,23,24,24,24,25,25,25,25,26,26,27,27,27,28,28, 28, 28,29,29,30}

i	z_i	n_i	f_i	F_i
1	18	1	1/30	1/30
2	19	2	2/30	3/30
3	20	1	1/30	4/30
4	21	3	3/30	7/30
5	22	2	2/30	9/30
6	23	2	2/30	11/30
7	24	3	3/30	14/30
8	25	4	4/30	18/30
9	26	2	2/30	20/30
10	27	3	3/30	23/30
11	28	4	4/30	27/30
12	29	2	2/30	29/30
13	30	1	1/30	30/30

Indici

Valori Voti

Frequenze Assolute

Frequenze Relative

$$F(x) = \begin{cases} 0, & x < z_1 \\ F_1, & z_1 \le x < z_2 \\ \dots & \\ F_i, & z_i \le x < z_{i+1} \\ \dots & \\ 1, & x \ge z_k \end{cases}$$

$$F(x) = \begin{cases} 0, & x < 18 \\ 1/30, & 18 \le x < 19 \\ 3/30, & 19 \le x < 20 \\ 4/30, & 20 \le x < 21 \\ 7/30, & 21 \le x < 22 \\ 9/30, & 22 \le x < 23 \\ 11/30, & 23 \le x < 24 \\ 14/30, & 24 \le x < 25 \\ 18/30, & 25 \le x < 26 \\ 20/30, & 26 \le x < 27 \\ 23/30, & 27 \le x < 28 \\ 27/30, & 28 \le x < 29 \\ 29/30, & 29 \le x < 30 \\ 1, & x \ge 30 \end{cases}$$

ESEMPIO FDD DISCRETA

Esempio: Consideriamo i voti conseguiti da 30 studenti all'esame di SAD nel 2022

X = {18,19,19,20,21,21,21,22,22,23,23,24,24,24,25,25,25,25,26,26,27,27,27,28,28, 28, 28,29,29,30}

```
# Dati discreti: votazioni esame universitario
dati_discreti <- c(18,19,19,20,21,21,21,22,22,23,23,24,24,24,25,25,25,25,26,
                   26,27,27,27,28,28, 28, 28,29,29,30)

# Calcolo della FDE empirica
FDE_discreta <- ecdf(dati_discreti)
```

- La funzione ecdf() calcola la funzione di distribuzione empirica cumulativa per i dati forniti
  - FDE_discreta diventa una funzione che rappresenta la proporzione di valori che sono minori o uguali a un determinato valore

ESEMPIO FDD DISCRETA

Esempio: Consideriamo i voti conseguiti da 30 studenti all'esame di SAD nel 2022

X = {18,19,19,20,21,21,21,22,22,23,23,24,24,24,25,25,25,25,26,26,27,27,27,28,28, 28, 28,29,29,30}

```r
# Dati discreti: votazioni esame universitario
dati_discreti <- c(18,19,19,20,21,21,21,22,22,23,23,24,24,24,25,25,25,25,26,
                   26,27,27,27,28,28, 28, 28,29,29,30)

# Calcolo della FDE empirica
FDE_discreta <- ecdf(dati_discreti)

# Valori unici dei dati
unique_vals <- sort(unique(dati_discreti))

# Stampa della FDE per ogni valore
for (x in unique_vals) {
  cat("FDE(", x, ") =", FDE_discreta(x), "\n")
}
```

```
FDE( 18 ) = 0.03333333
FDE( 19 ) = 0.1
FDE( 20 ) = 0.1333333
FDE( 21 ) = 0.2333333
FDE( 22 ) = 0.3
FDE( 23 ) = 0.3666667
FDE( 24 ) = 0.4666667
FDE( 25 ) = 0.6
FDE( 26 ) = 0.6666667
FDE( 27 ) = 0.7666667
FDE( 28 ) = 0.9
FDE( 29 ) = 0.9666667
FDE( 30 ) = 1
```

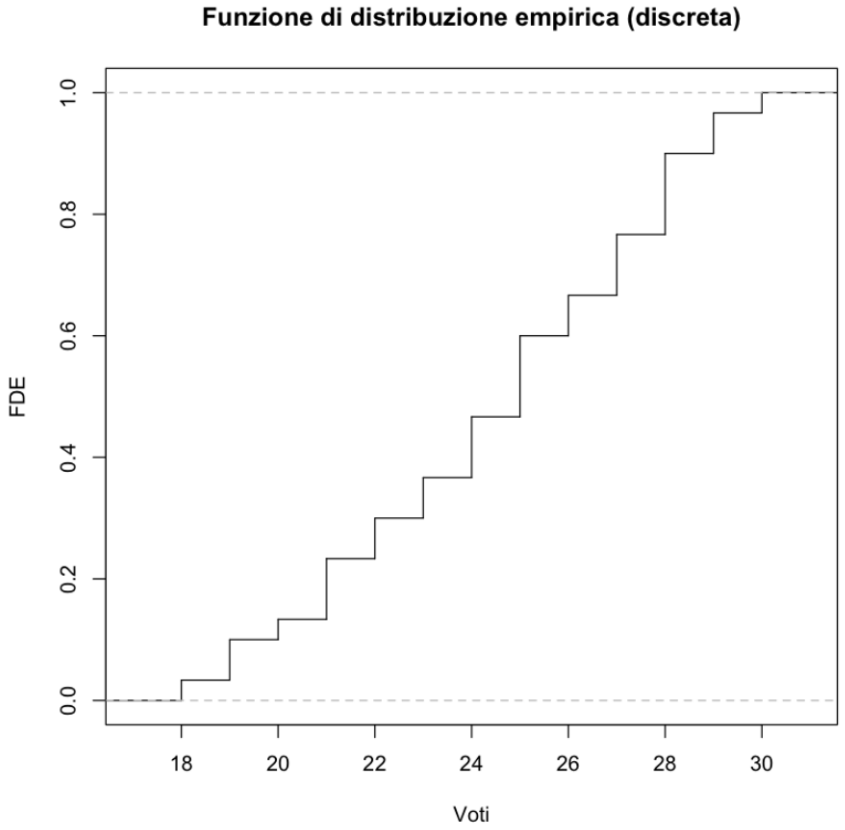
- Estrazione dei valori unici dai dati ed ordinamento in ordine crescente
  - unique_vals conterrà i voti distinti presenti nel vettore dati_discreti

ESEMPIO FDD DISCRETA

Esempio: Consideriamo i voti conseguiti da 30 studenti all'esame di SAD nel 2022

X = {18,19,19,20,21,21,21,22,22,23,23,24,24,24,25,25,25,25,26,26,27,27,27,28,28, 28, 28,29,29,30}

```r
# Dati discreti: votazioni esame universitario
dati_discreti <- c(18,19,19,20,21,21,21,22,22,23,23,24,24,24,25,25,25,25,26,
                   26,27,27,27,28,28, 28, 28,29,29,30)

# Calcolo della FDE empirica
FDE_discreta <- ecdf(dati_discreti)

# Valori unici dei dati
unique_vals <- sort(unique(dati_discreti))

# Stampa della FDE per ogni valore
for (x in unique_vals) {
  cat("FDE(", x, ") =", FDE_discreta(x), "\n")
}

# Grafico della FDE discreta
plot(FDE_discreta, main = "Funzione di distribuzione empirica (discreta)",
     xlab = "Voti", ylab = "FDE", verticals = TRUE, do.points = FALSE)
```



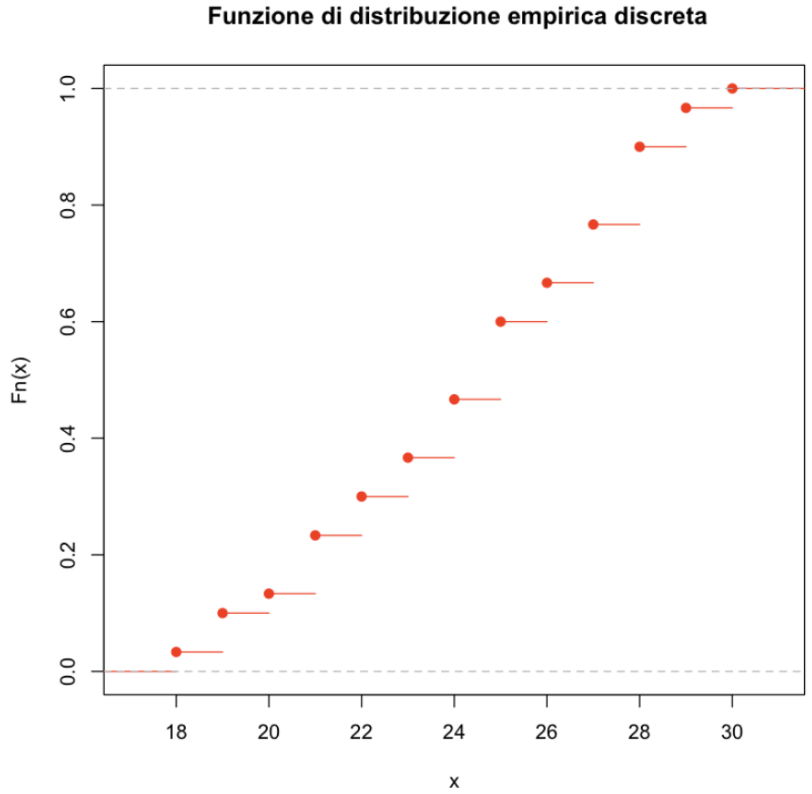
- Il metodo plot() genera un grafico della FdD Discreta
  - do.points = FALSE impedisce la visualizzazione dei punti sui gradini, risultando in un grafico più pulito
  - verticals = TRUE mostra le linee verticali nel grafico a gradini

ESEMPIO FDD DISCRETA

Esempio: Consideriamo i voti conseguiti da 30 studenti all'esame di SAD nel 2022

X = {18,19,19,20,21,21,21,22,22,23,23,24,24,24,25,25,25,25,26,26,27,27,27,28,28, 28, 28,29,29,30}

```r
# Dati discreti: votazioni esame universitario
dati_discreti <- c(18,19,19,20,21,21,21,22,22,23,23,24,24,24,25,25,25,25,26,
                   26,27,27,27,28,28, 28, 28,29,29,30)

# Calcolo della FDE empirica
FDE_discreta <- ecdf(dati_discreti)

# Valori unici dei dati
unique_vals <- sort(unique(dati_discreti))

# Stampa della FDE per ogni valore
for (x in unique_vals) {
  cat("FDE(", x, ") =", FDE_discreta(x), "\n")
}

plot(FDE_discreta,main=" Funzione di distribuzione empirica discreta",
     verticals=FALSE, col ="red")
```



ESEMPIO FDD DISCRETA

- **Esempio:**
- Immaginiamo di avere un campione di altezze di 10 persone:

160,165,170,175,180,185,190,195,200,205

- La FdDD ci permette di rispondere a domande del tipo:

Qual è la proporzione di persone che hanno un'altezza inferiore a 180 cm?

- Costruisci la FdDD:
  - $F(160) = \frac{1}{10} = 0.1$
  - $F(180) = \frac{5}{10} = 0.5$ (5 persone su 10 sono alte al massimo 180 cm)
  - $F(200) = \frac{9}{10} = 0.9$
  - $F(205) = \frac{10}{10} = 1$ (tutti sono ≤ 205 cm)

ESEMPIO FDD DISCRETA

- **Esempio:**
- Immaginiamo di avere un campione di altezze di 10 persone:

160,165,170,175,180 185,190,195,200 205

- La FdDD ci permette di rispondere a domande del tipo:

Qual è la proporzione di persone che hanno un'altezza inferiore a 180 cm?

- Costruisci la FdDD:
  - $F(160) = \frac{1}{10} = 0.1$
  - $F(180) = \frac{5}{10} = 0.5$ (5 persone su 10 sono alte al massimo 180 cm)
  - $F(200) = \frac{9}{10} = 0.9$
  - $F(205) = \frac{10}{10} = 1$ (tutti sono ≤ 205 cm)

STATISTICA E ANALISI DEI DATI

Capitolo 4 – Funzione di Distribuzione Empirica Continua

Dott. Stefano Cirillo
Dott. Luigi Di Biasi

a.a. 2025-2026

FDD CONTINUA

- Per fenomeni quantitativi continui occorre considerare la **funzione di distribuzione empirica continua**, ossia una funzione di distribuzione empirica strutturata in classi

- Supponiamo di organizzare i dati numerici in k distinte classi:

$$C_1 = [z_0, z_1), C_2 = [z_1, z_2) \ldots, C_k = [z_{k-1}, z_k], \text{ con } z_0 < z_1 < \ldots < z_{k-1} < z_k$$

- Dove z_0 corrisponde al minimo delle osservazioni e z_k al massimo delle osservazioni
- La funzione di distribuzione empirica continua è così definita:

$$F(x) = \begin{cases} 0, & x < z_0 \\ \ldots\ldots & \\ F_{i-1}, & x = z_{i-1} \\ \dfrac{F_i - F_{i-1}}{z_i - z_{i-1}}\, x + \dfrac{z_i\, F_{i-1} - z_{i-1}\, F_i}{z_i - z_{i-1}}, & z_{i-1} < x < z_i \\ F_i, & x = z_i \\ \ldots\ldots & \\ 1, & x \geq z_k, \end{cases}$$

- dove F_i denota la frequenza relativa cumulativa della classe C_i con $(i = 1, 2, \ldots, k)$

FDD CONTINUA

- La funzione di distribuzione empirica continua è così definita:

$$F(x) = \begin{cases} 0, & x < z_0 \\ \cdots\cdots & \\ F_{i-1}, & x = z_{i-1} \\ \dfrac{F_i - F_{i-1}}{z_i - z_{i-1}}\, x + \dfrac{z_i\, F_{i-1} - z_{i-1}\, F_i}{z_i - z_{i-1}} & z_{i-1} < x < z_i \\ F_i, & x = z_i \\ \cdots\cdots & \\ 1, & x \geq z_k, \end{cases}$$

- dove
  - F_i denota la frequenza relativa (probabilità) cumulativa della classe C_i con ($i = 1, 2, \ldots, k$)
  - $F(x) = 0$ per $x < z_0$ mentre $F(x) = 1$ per $x \geq z_k$
  - Se $z_{i-1} < x < z_i$ la funzione di distribuzione empirica continua coincide con il segmento che passa per i punti (z_{i-1}, F_{i-1}) e (z_i, F_i)

$$\frac{y - F_{i-1}}{x - z_{i-1}} = \frac{F_i - F_{i-1}}{z_i - z_{i-1}} \quad \text{Da cui} \quad y = F_{i-1} + \frac{F_i - F_{i-1}}{z_i - z_{i-1}}\,(x - z_{i-1}) = \frac{F_i - F_{i-1}}{z_i - z_{i-1}}\, x + \frac{F_{i-1}(z_i - z_{i-1}) - z_{i-1}\,(F_i - F_{i-1})}{z_i - z_{i-1}}$$

$$= \frac{F_i - F_{i-1}}{z_i - z_{i-1}}\, x + \frac{z_i\, F_{i-1} - z_{i-1}\, F_i}{z_i - z_{i-1}}.$$

Esempio FdD Continua

- Consideriamo il vettore voti contenente i voti dei 30 studenti, introduciamo le seguenti classi

$$C_1 = [18, 21),\ C_2 = [21, 24),\ C_3 = [24, 27),\ C_4 = [27, 30]$$

- Dato che 18 è il minimo e 30 è il massimo dei voti considerati, una possibile scelta per gli intervalli è

$$z_0 = 18,\ z_1 = 21,\ z_2 = 24,\ z_3 = 27,\ z_4 = 30.$$

i	C_i	n_i	f_i
1	$[18, 21)$	4	$4/30$
2	$[21, 24)$	7	$7/30$
3	$[24, 27)$	9	$9/30$
4	$[27, 30]$	10	$10/30$

Esempio FdD Continua

- Consideriamo il vettore voti contenente i voti dei 30 studenti, introduciamo le seguenti classi

$$C_1 = [18, 21),\ C_2 = [21, 24),\ C_3 = [24, 27),\ C_4 = [27, 30]$$

- Dato che 18 è il minimo e 30 è il massimo dei voti considerati, una possibile scelta per gli intervalli è

$$z_0 = 18,\ z_1 = 21,\ z_2 = 24,\ z_3 = 27,\ z_4 = 30.$$

i	C_i	n_i	f_i
1	$[18, 21)$	4	$4/30$
2	$[21, 24)$	7	$7/30$
3	$[24, 27)$	9	$9/30$
4	$[27, 30]$	10	$10/30$

Indici

Intervalli Voti

Frequenze Assolute

Frequenze Relative

ESEMPIO FDD CONTINUA

- Consideriamo il vettore voti contenente i voti dei 30 studenti, introduciamo le seguenti classi

$$C_1 = [18, 21),\ C_2 = [21, 24),\ C_3 = [24, 27),\ C_4 = [27, 30]$$

- Dato che 18 è il minimo e 30 è il massimo dei voti considerati, una possibile scelta per gli intervalli è

$$z_0 = 18,\ z_1 = 21,\ z_2 = 24,\ z_3 = 27,\ z_4 = 30.$$

i	C_i	n_i	f_i	F_i
1	$[18, 21)$	4	$4/30$	$4/30$
2	$[21, 24)$	7	$7/30$	$11/30$
3	$[24, 27)$	9	$9/30$	$20/30$
4	$[27, 30]$	10	$10/30$	$30/30$

Indici

Intervalli Voti

Frequenze Assolute

Frequenze Relative

Frequenze Cumulate

$$F(x) = \begin{cases} 0, & x < 18 \end{cases}$$

ESEMPIO FDD CONTINUA

- Consideriamo il vettore voti contenente i voti dei 30 studenti, introduciamo le seguenti classi

$$C_1 = [18, 21),\ C_2 = [21, 24),\ C_3 = [24, 27),\ C_4 = [27, 30]$$

- Dato che 18 è il minimo e 30 è il massimo dei voti considerati, una possibile scelta per gli intervalli è

$$z_0 = 18,\ z_1 = 21,\ z_2 = 24,\ z_3 = 27,\ z_4 = 30.$$

i	C_i	n_i	f_i	F_i
1	$[18, 21)$	4	$4/30$	$4/30$
2	$[21, 24)$	7	$7/30$	$11/30$
3	$[24, 27)$	9	$9/30$	$20/30$
4	$[27, 30]$	10	$10/30$	$30/30$

$$F(x) = \begin{cases} 0, & x < 18 \\ \frac{4}{90}(x - 18), & 18 \le x < 21 \end{cases}$$

$F_i = \frac{4}{30}$
$F_{i-1} = 0$
$Z_{i-1} = 18$
$Z_i = 21$

$$\frac{F_i - F_{i-1}}{z_i - z_{i-1}} x + \frac{z_i F_{i-1} - z_{i-1} F_i}{z_i - z_{i-1}}$$

$$\frac{\frac{4}{30} - 0}{21 - 18} x + \frac{21 * 0 - 18\frac{4}{30}}{21 - 18}$$

Esempio FdD Continua

- Consideriamo il vettore voti contenente i voti dei 30 studenti, introduciamo le seguenti classi

$$C_1 = [18, 21),\ C_2 = [21, 24),\ C_3 = [24, 27),\ C_4 = [27, 30]$$

- Dato che 18 è il minimo e 30 è il massimo dei voti considerati, una possibile scelta per gli intervalli è

$$z_0 = 18,\ z_1 = 21,\ z_2 = 24,\ z_3 = 27,\ z_4 = 30.$$

i	C_i	n_i	f_i	F_i
1	$[18, 21)$	4	$4/30$	$4/30$
2	$[21, 24)$	7	$7/30$	$11/30$
3	$[24, 27)$	9	$9/30$	$20/30$
4	$[27, 30]$	10	$10/30$	$30/30$

$$F(x) = \begin{cases} 0, & x < 18 \\ \frac{4}{90}(x - 18), & 18 \le x < 21 \\ \frac{1}{90}(7x - 135), & 21 \le x < 24 \\ \frac{1}{90}(9x - 183), & 24 \le x < 27 \\ \frac{1}{90}(10x - 210), & 27 \le x < 30 \\ 1, & x \ge 30 \end{cases}$$

ESEMPIO FDD CONTINUA

- Consideriamo il vettore voti contenente i voti dei 30 studenti, utilizziamo R per il calcolo della FdD Continua

i	C_i
1	$[18, 21)$
2	$[21, 24)$
3	$[24, 27)$
4	$[27, 30]$

```
voti <- c(18, 19, 20, 30, 29, 28, 21, 22, 23, 27, 26, 25, 24, 25,
          26, 24, 23, 22, 27, 28, 21, 24, 25, 25, 27, 19, 21, 28, 29, 28)
freqrel <- table(voti)/length(voti)
round(freqrel,3)
```

```
voti
   18    19    20    21    22    23    24    25    26    27    28    29    30
0.033 0.067 0.033 0.100 0.067 0.067 0.100 0.133 0.067 0.100 0.133 0.067 0.033
```

```
m <- length(freqrel) #visualizza la lunghezza del vettore frequenza
m
```

13

```
classi <- c(18,21,24,27,30)
intervalli <- cut(voti,breaks=classi,right=FALSE)
intervalli
```

[18,21) · [18,21) · [18,21) · <NA> · [27,30) · [27,30) · [21,24) · [21,24) · [21,24) · [27,30) · [24,27) · [24,27) · [24,27) · [24,27) · [24,27) · [24,27) · [21,24) · [21,24) · [27,30) · [27,30) · [21,24) · [24,27) · [24,27) · [24,27) · [27,30) · [18,21) · [21,24) · [27,30) · [27,30) · [27,30)

▼ **Levels**:

'[18,21)' · '[21,24)' · '[24,27)' · '[27,30)'

ESEMPIO FDD CONTINUA

- Consideriamo il vettore voti contenente i voti dei 30 studenti, utilizziamo R per il calcolo della FdD Continua

```
voti <- c(18, 19, 20, 30, 29, 28, 21, 22, 23, 27, 26, 25, 24, 25,
          26, 24, 23, 22, 27, 28, 21, 24, 25, 25, 27, 19, 21, 28, 29, 28)
freqrel <- table(voti)/length(voti)
round(freqrel,3)
```

```
voti
   18    19    20    21    22    23    24    25    26    27    28    29    30
0.033 0.067 0.033 0.100 0.067 0.067 0.100 0.133 0.067 0.100 0.133 0.067 0.033
```

i	C_i	n_i	f_i
1	$[18, 21)$	4	$4/30$
2	$[21, 24)$	7	$7/30$
3	$[24, 27)$	9	$9/30$
4	$[27, 30]$	10	$10/30$

```
m <- length(freqrel) #visualizza la lunghezza del vettore frequenza
m
```

13

```
classi <- c(18,21,24,27,30)
intervalli <- cut(voti,breaks=classi,right=FALSE)
intervalli
```

[18,21) · [18,21) · [18,21) · <NA> · [27,30) · [27,30) · [21,24) · [21,24) · [21,24) · [27,30) · [24,27) · [24,27) · [24,27) · [24,27) · [24,27) · [24,27) · [21,24) · [21,24) · [27,30) · [27,30) · [21,24) · [24,27) · [24,27) · [24,27) · [27,30) · [18,21) · [21,24) · [27,30) · [27,30) · [27,30)

▼ **Levels**:

'[18,21)' · '[21,24)' · '[24,27)' · '[27,30)'

```
frelclassi <- table(intervalli)/length(voti)
frelclassi
```

```
intervalli
  [18,21)   [21,24)   [24,27)   [27,30)
0.1333333 0.2333333 0.3000000 0.3000000
```

ESEMPIO FDD CONTINUA

- Consideriamo il vettore voti contenente i voti dei 30 studenti, utilizziamo R per il calcolo della FdD Continua

```
Fcum <- cumsum(frelclassi)
Fcum
```

[18,21): 0.133333333333333 **[21,24):** 0.366666666666667 **[24,27):** 0.666666666666667 **[27,30):** 0.966666666666667

- Poiché in precedenza con la funzione cut() abbiamo specificato l'opzione right = FALSE, l'ultimo intervallo non è corretto in quando dovrebbe essere [27,30], cioè chiuso a destra

i	C_i	n_i	f_i	F_i
1	$[18, 21)$	4	$4/30$	$4/30$
2	$[21, 24)$	7	$7/30$	$11/30$
3	$[24, 27)$	9	$9/30$	$20/30$
4	$[27, 30]$	10	$10/30$	$30/30$

ESEMPIO FDD CONTINUA

- Consideriamo il vettore voti contenente i voti dei 30 studenti, utilizziamo R per il calcolo della FdD Continua

```
Fcum <- cumsum(frelclassi)
Fcum
```

[18,21): 0.133333333333333 **[21,24):** 0.366666666666667 **[24,27):** 0.666666666666667 **[27,30):** 0.966666666666667

```
Fcum[4] <- Fcum[4]+freqrel[m]
Fcum
```

[18,21): 0.133333333333333 **[21,24):** 0.366666666666667 **[24,27):** 0.666666666666667 **[27,30):** 1

- Poiché in precedenza con la funzione cut() abbiamo specificato l'opzione right = FALSE, l'ultimo intervallo non è corretto in quando dovrebbe essere [27,30], cioè chiuso a destra
- Sommiamo la frequenza relativa di 30 ed otteniamo 1

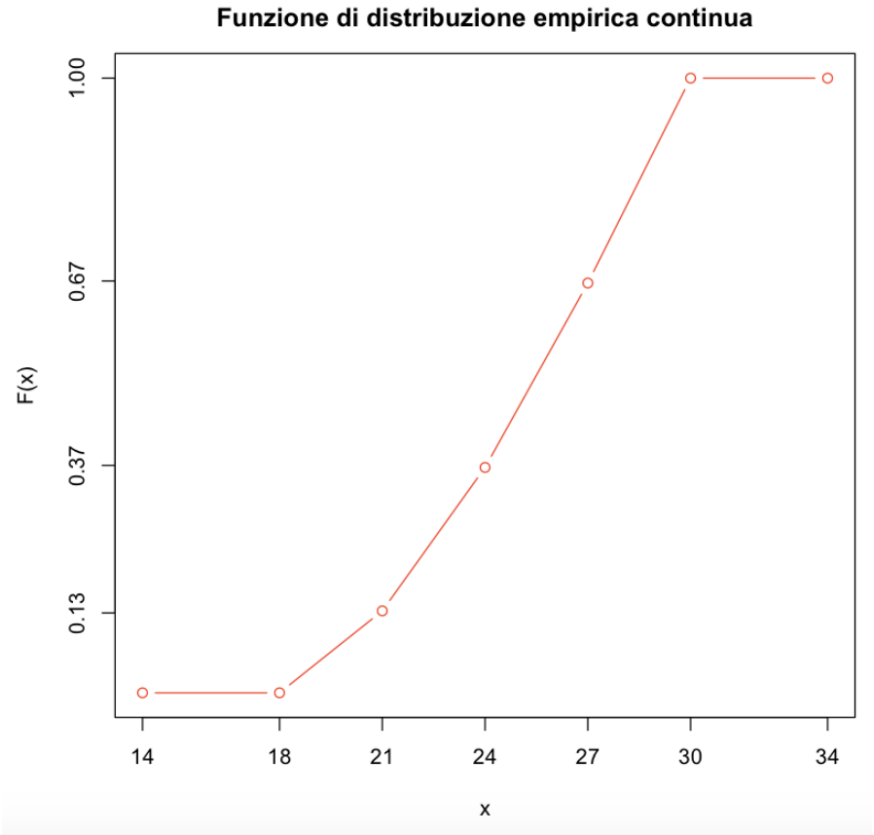
i	C_i	n_i	f_i	F_i
1	$[18, 21)$	4	$4/30$	$4/30$
2	$[21, 24)$	7	$7/30$	$11/30$
3	$[24, 27)$	9	$9/30$	$20/30$
4	$[27, 30]$	10	$10/30$	$30/30$

ESEMPIO FDD CONTINUA

- Consideriamo il vettore voti contenente i voti dei 30 studenti, utilizziamo R per il calcolo della FdD Continua

```
ascisse<-c(14,18,21,24,27,30,34)
ordinate<-c(0,0,Fcum[1:4],1)
plot(ascisse,ordinate,type="b",axes=FALSE,main="
Funzione di distribuzione empirica continua",
col="red",ylim=c(0,1),xlab="x",ylab="F(x)")
axis(1,ascisse)
axis(2,format(Fcum,digits=2))
box()
```

- Sono stati aggiunti 2 intervalli $C_0 = [14,18)$ e $C_5 = [30,34)$ per tracciare la linea y = 0 nell'intervallo C_0 e y = 1 in C_5
  - Di conseguenza anche i valori nelle ordinate sono stati aggiornati aggiungendo due 0 e un 1 alla fine



FUNZIONE ECDF IN R

- La funzione **ecdf(x)** in R serve per **costruire la funzione di distribuzione empirica cumulata** (Empirical Cumulative Distribution Function) a partire da un insieme di dati osservati x
- **Cosa fa ecdf(x) in pratica**

F_emp <- ecdf(x)

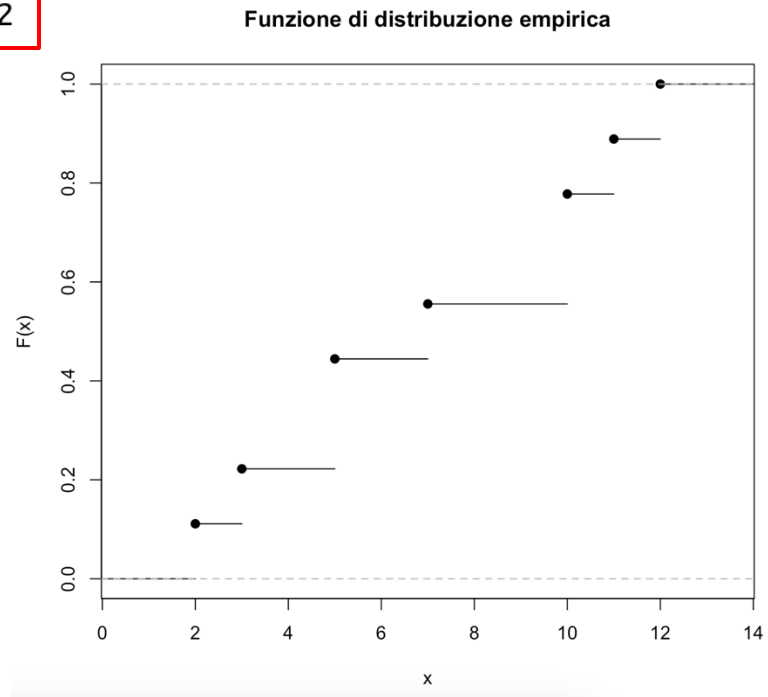
  - **Crea una funzione** (non un **vettore**!) che rappresenta la distribuzione empirica.
  - Questa funzione può poi essere **valutata** in qualsiasi punto o **tracciata** con plot()

```
x <- c(2, 3, 5, 5, 7, 10, 10, 11, 12)
F_emp <- ecdf(x)

# Valuto la funzione in alcuni punti
F_emp(4)   # proporzione di valori <= 4
F_emp(5)   # proporzione di valori <= 5
F_emp(8)   # proporzione di valori <= 8

plot(F_emp, main="Funzione di distribuzione empirica", xlab="x", ylab="F(x)")
```

```
Empirical CDF
Call: ecdf(x)
 x[1:7] =      2,      3,      5,  ...,     11,     12
```

```
> F_emp(3)
[1] 0.2222222
> F_emp(1)
[1] 0
> F_emp(100)
[1] 1
```

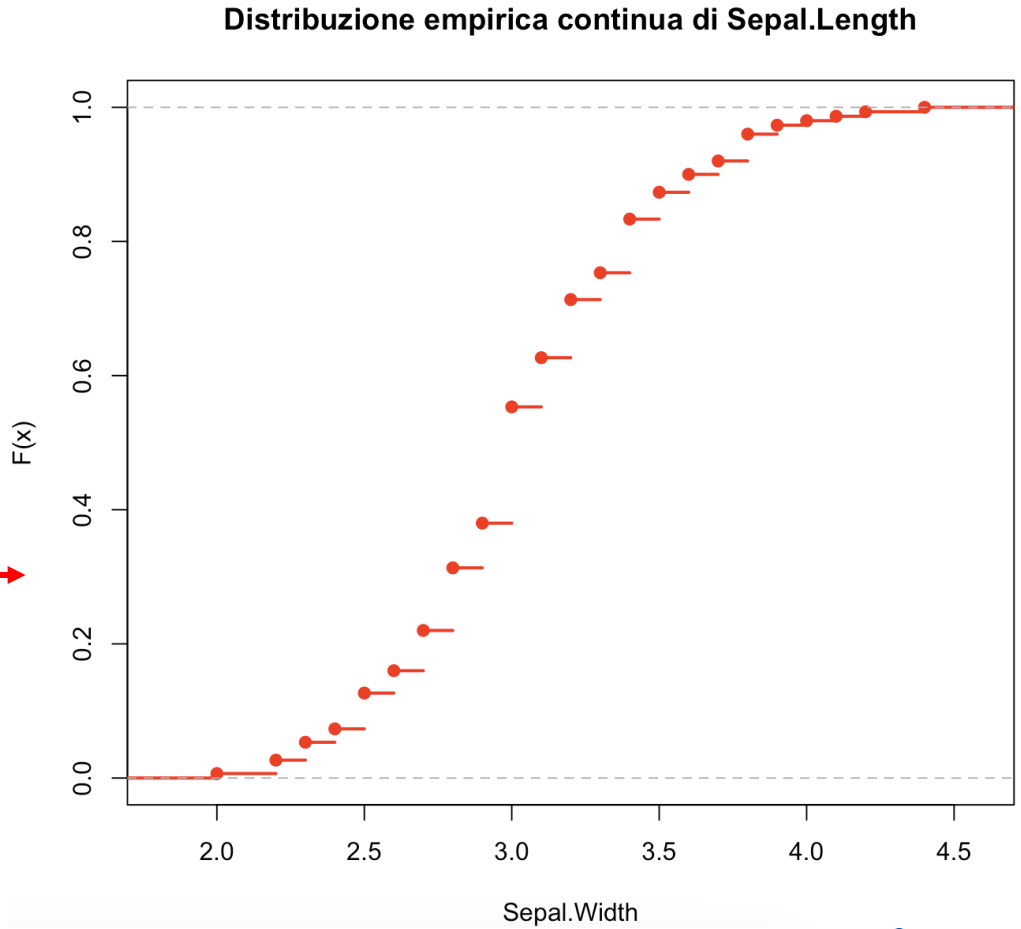
Poiché **ecdf(x)** restituisce una **funzione**, possiamo passarle **qualsiasi valore reale** e otterremo la proporzione di osservazioni **minori o uguali** a quel valore **(probabilità empirica cumulata)**



ESEMPIO ECDF IN IRIS

- Costruire e rappresentare la distribuzione empirica continua della variabile continua Sepal.Width nel dataset *Iris*

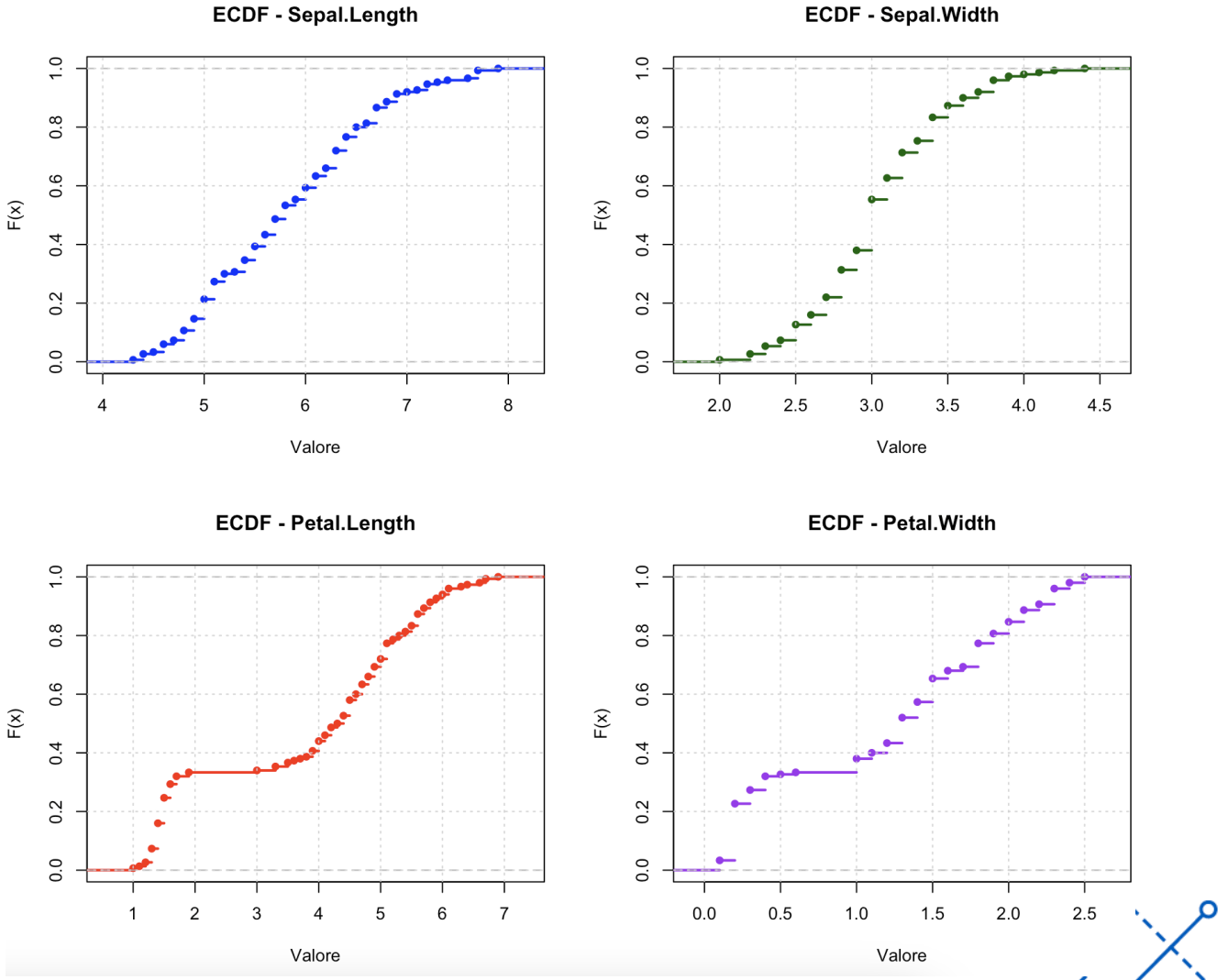
```
data(iris)

x <- iris$Sepal.Width

# Costruisco la funzione di distribuzione empirica
F_emp <- ecdf(x)

# Grafico della distribuzione empirica continua
plot(F_emp,
     main = "Distribuzione empirica continua di Sepal.Width",
     xlab = "Sepal.Width",
     ylab = "F(x)",
     col = "red", lwd = 2)
```



ESEMPIO ECDF IN IRIS

- Costruire e rappresentare la distribuzione empirica continua della variabile continua Sepal.Width nel dataset *Iris*

```r
par(mfrow = c(2, 2))

# Calcolo e plotto la ECDF per ciascuna variabile
plot(ecdf(iris$Sepal.Length),
     main = "ECDF - Sepal.Length",
     xlab = "Valore",
     ylab = "F(x)",
     col = "blue", lwd = 2)
grid()

plot(ecdf(iris$Sepal.Width),
     main = "ECDF - Sepal.Width",
     xlab = "Valore",
     ylab = "F(x)",
     col = "darkgreen", lwd = 2)
grid()

plot(ecdf(iris$Petal.Length),
     main = "ECDF - Petal.Length",
     xlab = "Valore",
     ylab = "F(x)",
     col = "red", lwd = 2)
grid()

plot(ecdf(iris$Petal.Width),
     main = "ECDF - Petal.Width",
     xlab = "Valore",
     ylab = "F(x)",
     col = "purple", lwd = 2)
grid()
```



FDD DISCRETA E CONTINUA

- Le **distribuzioni empiriche**, sia **discrete** che **continue**, forniscono informazioni chiave sulla distribuzione di un insieme di dati
  - **Probabilità empirica**: Fornisce una stima della probabilità associata a ciascun valore discreto nel dataset
  - **Frequenza cumulata**: Indica la frazione di dati che è minore o uguale a un certo valore
  - **Pattern e ripetizioni**: Se un valore si ripete frequentemente (ad esempio, molti studenti hanno preso 25), questo si riflette nel grafico della distribuzione, indicando che c'è una concentrazione su quel valore specifico
- La distribuzione empirica continua si applica a dati quantitativi continui, ossia dati che possono assumere qualsiasi valore in un certo intervallo (es. peso, altezza, tempo)
  - **Stima della distribuzione di variabili continue**: La FdD Continua offre una rappresentazione grafica di come i dati quantitativi continui si distribuiscono lungo un intervallo. Non richiede che i dati appartengano a una particolare distribuzione teorica
  - **Andamento delle classi**: Quando i dati sono suddivisi in classi (come ad esempio intervalli di peso: 60-65, 65-70, ecc.), la distribuzione empirica continua mostra come si distribuiscono i dati tra le diverse classi

DOMANDE?