Slide Set 3

Densità, ECDF e Quantili

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 (h08:39)

Notes

Nozioni locali: massa e densità

Massa

- In un particolare punto $X = x_0$ ci sono più osservazioni, diremo che in x_0 esiste una massa di dati
- tipicamente con dati discreti X assume $K \leq n$ livelli distinti $\{x_1, x_2, \ldots, x_K\}$, ognuno dei quali viene osservato più di una volta.
- la grandezza di questa massa in $X = x_k$ misura n_k oppure f_k

Densità

- se X è continua finisce per esprimere troppi livelli distinti. Al limite, quando $n \to \infty$, ci saranno tanti livelli distinti ciascuno con frequenza $n_k \approx 1/n \to 0$
- non ha senso misurare la massa di dati in un punto $X = x'$
- la densità dei dati in $X = x'$ misura quanta la massa di dati in un intervallo contenente x'

Notes

Come possiamo misurare la densità per dati univariati? Costruiamo una funzione di densità:

$$\text{densità}(x') \quad = \quad \frac{\text{massa di dati intorno a } x'}{\text{grandezza dell'intervallo che contiene } x'}$$

Scegliendo opportunamente numeratore e denominatore otteniamo funzioni di densità alternative.

Definizione (Istogramma)

Sia X una variabile continua, e siano $(x_{k-1}, x_k]$ per $k = 1, 2, \ldots, K$ le classi di livelli di X. L'istrogramma è la funzione di densità dei dati che associa ad ogni punto x'

$$h(x') := \begin{cases} \frac{f_k}{x_k - x_{k-1}} & \text{se } x' \in (x_{k-1}, x_k] \text{ per qualche } k \\ 0 & \text{altrimenti} \end{cases}$$

Notes

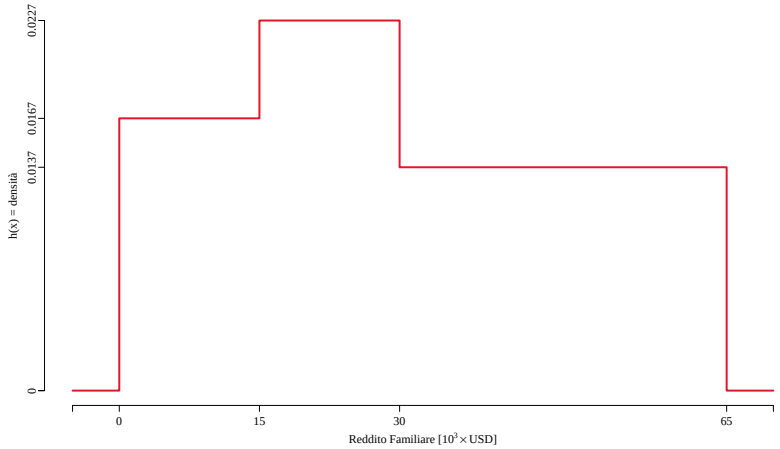
Esempio: consideriamo X=`faminc` dal data set `bw.csv`

- $n = 1388$, `range(fminc)` $= [0.5, 65]$
- windowing dei dati con $K = 3$ classi non uniformi
- $\Delta_k = x_k - x_{k-1}$: ampiezza dell'intervallo della k-ma classe

`famic` [$10^3\times$ USD]	n_i	f_i	Δ_k	Densità$_k$
$[0, 15]$	348	0.25	15 (=15-0)	0.0167 (=0.25/15)
$(15, 30]$	466	0.34	15 (=30-15)	0.0227 (=0.34/15)
$(30, 65]$	574	0.41	35 (=65-30)	0.0137 (=0.41/35)

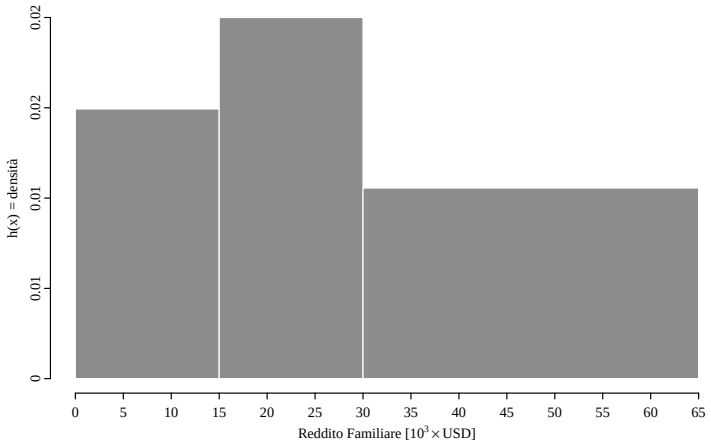
Notes

Rappresentazione della funzione $h(x)$: è un grafico comunemente chiamato istogramma



Notes

... solitamente lo rappresentiamo così



⚠ In generale: la scala verticale del grafico = densità (non frequenze)

Notes

Ogni classe k è rappresentata da un rettangolo che ha

$$\begin{aligned} \text{Base} &= \Delta_k = x_k - x_{k-1} \\ \text{Altezza} &= f_k/\Delta_k \\ \text{Area} &= \text{Base}\times\text{Altezza} = \Delta_k \times \tfrac{f_k}{\Delta_k} = f_k \end{aligned}$$

Quindi, la somma delle aree dei rettangoli è data da

$$\sum_{k=1}^{K} \text{Area}_k = \sum_{k=1}^{K} f_k = 1$$

Interpretazione corretta (esempio precedente):

- l'intervallo $(30, 65]$ è meno denso di dati di $(0, 15]$
- la classe $(30, 65]$ contiene più dati di tutte le altre... anche se questo non è facile da *vedere*, non trovi?

Notes

Caso particolare: classi uniformi

$\Delta_k = \Delta$ è uguale per tutte le classi. Per ogni classe k la densità di classe è misurata da

$$\frac{f_k}{\Delta} \implies \text{Area}_k = f_k\times\Delta$$

i valori di densità, e le aree dei rettangoli, sono tutti proporzionali alle frequenze

In tal caso si usa spesso riscalare l'asse verticale dell'istogramma in modo da ottenere le frequenze assolute n_k.

Infatti moltiplicando tutti valori di densità (asse verticale del grafico) per la costante $n\Delta$

$$\frac{f_k}{\Delta}\times n\Delta \;=\; f_k\times n \;=\; n_k$$

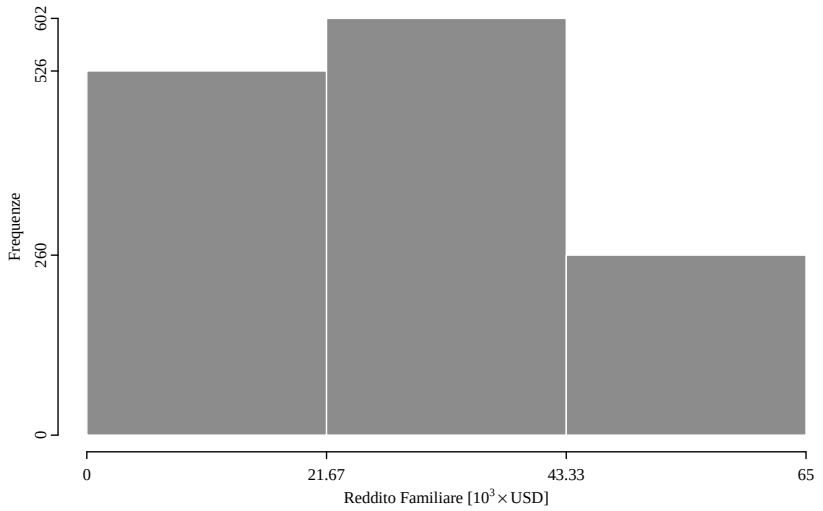
Notes

Esempio: consideriamo X=`faminc` dal data set `bw.csv`.

- fissiamo $K = 3$, con classi uniformi
- $\Delta = (x_{\max} - x_{\min})/3 = 21.667$
- $n\Delta = 1388 \times 21.667 = 30073.33$

`famic` $[10^3\times$ USD]	n_i	f_i	Δ	Densità$_k$	$n\Delta\times$Densità$_k$
$[0,\ 21.67]$	526	0.38	21.667	0.0175	526
$(21.67,\ 43.33]$	602	0.43	21.667	0.0198	602
$(43.33,\ 65]$	260	0.19	21.667	0.0088	260

Notes



Notes

Osservazioni finali sull'istogramma/densità

⚠ Lo scaling verticale dell'istogramma in termini di frequenze assolute n_k, ha senso solo quando le classi sono uniformi

Quando non si ha nessuna intuizione circa le classi, è meglio affidarsi ad un metodo di windowing "*automatico*". Ci sono molti metodi (es: Sturges) che determinano un numero "*ottimale*" K (in seguito ne vedremo altri)

L'istogramma è una delle tante possibili misure di densità:

- **(+)**: facile da costruire, l'interpretazione grafica è semplice (soprattutto con classi uniformi)
- **(−)**: • approssima la densità in modo omogeneo/costante per tutti i punti all'interno della classe • è una funzione discontinua che produce una transizione "*brusca*" tra una classe e quelle adiacenti (fra qualche lezione introdurremo la densità kernel).

Esempi/Applicazioni $\longrightarrow$ `R` **script file**

Notes

Funzione indicatrice e conteggio

Data una variabile X definiamo la seguente funzione indicatrice

$$\mathbf{1}\{X \leq t\} = \begin{cases} 1 & \text{se } X \leq t \\ 0 & \text{altrimenti} \end{cases}$$

Ad esempio fissato $t = 10$, la funzione indicatrice vale 1 ogni per ogni $X \in (-\infty, 10]$, e vale 0 per ogni $X \in (10, +\infty)$

Dati $\{x_1, x_2, \ldots, x_n\}$, la seguente somma conta il numero di osservazioni minori o uguali di t

$$\sum_{i=1}^{n} \mathbf{1}\{x_i \leq t\}$$

Esempio: si osserva $\{x_1, x_2, x_3, x_4\} = \{-3, 4, 0, 10\}$, fissiamo $t = 7.5$

$$\sum_{i=1}^{n} \mathbf{1}\{x_i \leq 7.5\} = 1 + 1 + 1 + 0 = 3$$

Notes

Frequenze cumulate

Nel nostro campione esistono $K \leq n$ livelli distinti $\{x_1, x_2, \ldots, x_K\}$.
Fissiamo l'attenzione su un particolare livello x_k e definiamo:

Frequenza cumulata assoluta di x_k

$$N_k = (\text{numero di osservazioni } \leq x_k) = \sum_{i=1}^{n} \mathbf{1}\{x_i \leq x_k\}$$

Frequenza cumulata relativa di x_k

$$F_k = (\text{proporzione di osservazioni } \leq x_k) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{x_i \leq x_k\} = \frac{N_k}{n}$$

Osserva
per ottenere le F_k in scala percentuale calcoliamo $F_k \times 100$

Notes

Esempio: consideriamo nuovamente il data set `bw.csv`, sia
X=`parity`="*numero di figli*"

X	n_k	$f_k(\%)$	N_k	$F_k(\%)$
1	795	57.3	795	57.3
2	389	28.0	1184	85.3
3	146	10.5	1330	95.8
4	39	2.8	1369	98.6
5	15	1.1	1384	99.7
6	4	0.3	1388	100.0

Interpretazioni

- 1330 famiglie hanno al massimo 3 figli
- l'85.3% delle famiglie campionate hanno non più di 2 figli
- il 95.8% delle osservazioni esprime valori tra 1 e 3
- etc ...

Notes

Esempio: consideriamo `faminc` dal data set `bw`.

`famic` $[10^3\times$ USD]	n_k	$f_k(\%)$	N_k	$F_k(\%)$
$[0.435, 22]$	526	37.9	526	37.9
$(22, 43.5]$	602	43.4	1128	81.3
$(43.5, 65.1]$	260	18.7	1388	100

Interpretazioni

- l'81.3% del campione ha un reddito famigliare inferiore o uguale a 43500 USD
- il 37.9% del campione più povero dispone di un reddito non più grande di 22000 USD
- etc ...

Notes

Uno strumento potentissimo per descrivere moltissimi aspetti di una DU:

Definizione (ECDF)

Sia X una variabile quantitativa, e sia $\{x_1, x_2, \ldots, x_n\}$ una campione osservato. Fissato $t \in \mathbb{R}$ la *Funzione di Distribuzione Empirica Cumulata* (ECDF) nel punto t è

$$\mathbb{F}(t) = (\text{proporzione delle osservazioni } \leq t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{x_i \leq t\} \quad (3.1)$$

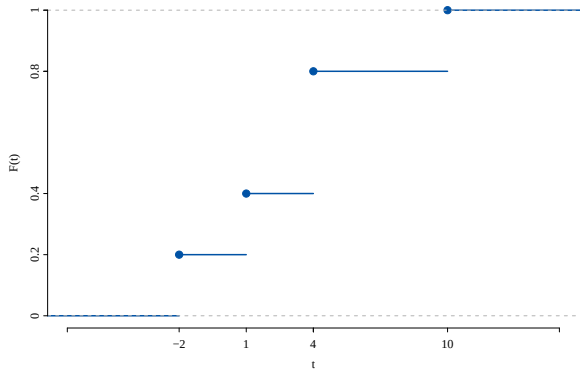
Osserva

- nella (3.1) l'argomento t non è necessariamente un valore osservato
- se x_k è uno dei livelli osservati, per $t = x_k$ otteniamo $\mathbb{F}(x_k) = F_k$

Notes

Rappresentazione grafica dell'ECDF

Esempio: osserviamo $\{-2, 1, 4, 10, 4\}$. Immaginiamo di calcolare $\mathbb{F}(t)$ facendo *scorrere* t da $-\infty$ verso $+\infty$



Osserva
per costruire $\mathbb{F}(t)$ basta calcolare $\mathbb{F}(x_k) = F_k$ per tutti i livelli distinti osservati

Notes

Alcune proprietà di $\mathbb{F}(t)$

- è non decrescente
- il suo grafico ha una struttura a gradini: è costante tra ogni coppia di livelli distinti
- è continua a destra, "*salta*" su ogni valore distinto del campione
- $\lim_{t\to-\infty} \mathbb{F}(t) = 0$ e $\lim_{t\to+\infty} \mathbb{F}(t) = 1$

A cosa serve

- da informazioni circa le frequenze cumulate (...e quantili, vedi dopo)
- descrive come la massa di dati si accumula lungo il range
- nei tratti in cui l'ECDF ha una pendenza forte (≈derivata grande), $\Longrightarrow$ zona del range ad alta densità di dati
- nei tratti in cui l'ECDF è piuttosto piatta (≈derivata piccola) $\Longrightarrow$ zona del range a bassa densità di dati
- ...molto altro ancora! In pratica
  ECDF in statistica ≡ "*sega e martello in falegnameria*"

Esempi/Applicazioni $\longrightarrow$ `R` script file

Notes

Quantili e loro approssimazioni numeriche

Definizione (Quantile)

Sia X una variabile quantitativa, e siano $\{x_1, x_2, \ldots, x_n\}$ i suoi valori osservati. Fissato $\alpha \in [0, 1]$, il quantile di livello α è il valore X_α tale che

(i) $\alpha\%$ dei valori osservati sono $\leq X_\alpha$

(ii) $(1 - \alpha)\%$ dei valori osservati sono $> X_\alpha$

Esempio: si osservano $n = 5$ valori per la variabile Y: $\{10, -1, 7, 2, 0\}$. Fissiamo $\alpha = 0.4$ (40%). Chi è $Y_{0.4}$?
Cerchiamo un valore osservato $Y_{0.4}$ tale che il 40% delle osservazioni $\leq Y_{0.4}$. Il $(40\%$ di $n) = (0.4 \times 5) = 2$. Ordiniamo i dati

$$-1\ ,\ 0\ ,\ 2\ ,\ 7\ ,\ 10$$

Quindi $Y_{0.4} = 0$, infatti esattamente 2 osservazioni (pari al $40\%n$) sono ≤ 0, mentre 3 osservazioni $(60\%n)$ sono >0

Notes

Esempio: consideriamo gli stessi dati, ma questa volta fissiamo $\alpha = 25\%n$, e vogliamo calcolare $Y_{0.25}$.
Cerchiamo $Y_{0.25}$ tale che *un quarto delle osservazioni* (il 25%) è $\leq Y_{0.25}$.
⚠ $25\%n = 0.25 \times 5 = 1.25$

$$-1,\ \downarrow\ 0\ ,\ 2\ ,\ 7\ ,\ 10$$

Teoricamente $Y_{0.25}$ sarebbe collocato tra -1 e 0. Impossibile applicare la *definizione esatta* di Y_α.

Approssimazioni numeriche dei quantili

La funzione `quantile()` di R consente di lavorare con circa 9 diversi metodi di approssimazione. Metodi più comuni:

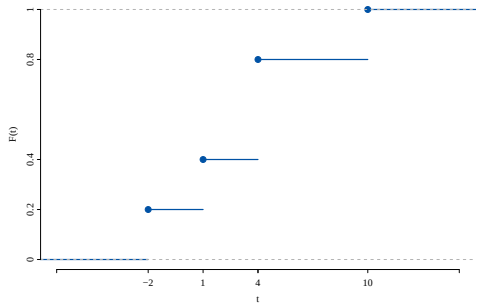
- quantili empirici
  $\rightarrow$ in R è `quantile(..., type = 1)`
- *smoothing* (interpolazione) basata sui *ranks*
  $\rightarrow$ in R è `quantile(..., type = 7)`, dove `type=7` è default

Notes

Quantili empirici

I quantili empirici sono ottenuti costruendo una sorta di *inversa* dell ECDF (l'ECDF non è invertibile)

Dati precedenti: $\{-2, 1, 4, 10, 4\}$, l'ECDF era



$\alpha = 0.2$, quanto vale $X_{0.2}$? $\mathbb{F}(-2) = 0.2 \implies$ il 20% dei valori osservati $\leq -2 \implies X_{0.2} = -2$

Se $\alpha = 0.5$, quanto vale $X_{0.5}$?

Notes

Definizione (Approssimazione 1: quantili empirici)

Sia X una variabile quantitativa, e siano $\{x_1, x_2, \ldots, x_K\}$ i suoi $K \leq n$ livelli distinti osservati. Sia $\mathbb{F}(t)$ la ECDF per il campione osservato. Il quantile empirico di livello α è

$$X_\alpha := \min \{x_k : \mathbb{F}(x_k) \geq \alpha\} \tag{3.2}$$

L'informazione necessaria per calcolare i quantili empirici è tutta contenuta nell'ECDF. È necessario calcolcolare $\mathbb{F}(t)$ per ogni $t \in \mathbb{R}$?... NO!

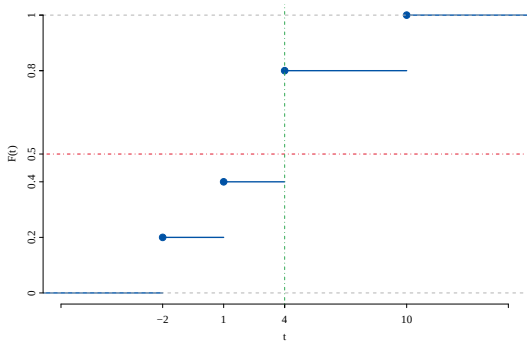
Poichè $\mathbb{F}(x_k) = F_k \implies$ basta trovare il più piccolo valore osservato tale che la corrispondente frequenza cumulata è $\geq \alpha$

Notes

Esempio: dati precedenti: $\{-2, 1, 4, 10, 4\}$, vogliamo calcolare $X_{0.5}$ usando l'approssimazione.

Applichiamo la definizione *alla lettera*:

1. troviamo l'insieme dei valori osservati tali che $F_k \geq 50\%$
2. prendiamo il minimo di questo insieme



$$X_{0.5} = \min\{x_k : F_k \geq 0.5\} = \min\{4, 10\} = 4$$

Notes

Esempio: consideriamo il data set `bw.csv`, e X=`parity`="*numero di figli*"

X	$F_k(\%)$
1	57.3
2	85.3
3	95.8
4	98.6
5	99.7
6	100.0

Calcoliamo il quantile empirico $X_{0.75}$. Applichiamo la definizione (3.2)

$$X_{0.75} = \min\{x_k : F_k \geq 0.75\} = \min\{2, 3, 4, 5, 6\} = 2$$

Calcoliamo il quantile empirico $X_{0.1}$. Applichiamo la (3.2)

$$X_{0.1} = \min\{x_k : F_k \geq 0.1\} = \min\{1\} = 1$$

Esempi/Applicazioni $\longrightarrow$ `R` script file

Notes

Questo tipo di approssimazione funziona bene quando

- n sufficientemente grande
- dati sono misurati in modo "*poco discretizzato*", ovvero non ci sono troppe ripetizioni
- α non troppo vicino al limite inferiore/superiore ($\alpha \to 0$ e $\alpha \to 1$)

In tutti gli altri casi, i quantili empirici introducono delle *distorsioni* che saranno chiare nei corsi successivi

Per ridurre questi effetti distorsivi, sono state introdotte una serie di varianti. La più comune è la seguente approssimazione (3.3)

Esempi/Applicazioni $\longrightarrow$ **R script file**

Notes

Ordinamento dei dati e *ranks*

Date le osservazioni $\{x_1, x_2, \ldots, x_n\}$ consideriamo l'ordinamento (non-decrescente)

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

- $x_{(j)}$ è il valore osservato di X che occupa il j-mo posto nella lista ordinata (non decrescente) dei dati
- $x_{(j-1)} \leq x_{(j)}$ per ogni $j = 1, 2, \ldots, n$
- gergo tecnico: la posizione nell'ordinamento si chiama rank.

Esempio: si rileva la temperatura Z (in C°) in $n = 6$ celle, dati osservati: $\{2, 3, 0, -3, 9, 3\}$,

Ordinamento	-3	0	2	3	3	9
Posizione/rank	(1)	(2)	(3)	(4)	(5)	(6)

Notes

I valori osservati ripetuti in gergo si chiamano ties.

In questi casi il rank non è unico, esistono opportune correzioni (vedi esempi in R)

L'ordinamento dei dati è un task molto rilevante nell'analisi numerica indipendentemente dal calcolo di ranks, quantili, etc

Esempi/Applicazioni $\longrightarrow$ R script file

Notes

L'ordinamento dei dati ed i ranks sono alla base di una delle approssimazioni numeriche dei quantili più popolari

Definizione (Approssimazione 2: *smoothing* su dati ordinati)

Sia X una variabile quantitativa, e siano $\{x_1, x_2, \ldots, x_n\}$ i suoi valori osservati. Il quantile di livello α ottenuto per interpolazione (smoothing) basata sui ranks è

$$X_\alpha = (1-\gamma)\, x_{(j)} + \gamma\, x_{(j+1)} \tag{3.3}$$

dove γ e j sono definiti come segue

$$\begin{aligned} j^* &= 1 + \alpha(n-1) && \text{(rank ottimale)} \\ j &= \lfloor j^* \rfloor && \text{(parte intera di } j^*\text{)} \\ \gamma &= (j^* - j) && \text{(parte decimale di } j^*\text{)} \end{aligned}$$

Notes

Esempio: ritorniamo alle temperature Z. Vogliamo calcolare $Z_{\frac{1}{3}}$. Dati ordinati:

Ordinamento	-3	0	2	3	3	9
Posizione	(1)	(2)	(3)	(4)	(5)	(6)

- $j^* = 1 + \alpha(n-1) = 1 + \frac{1}{3} \times 5 = 2.67$
- $j = \lfloor j^* \rfloor = \lfloor 2.67 \rfloor = 2$
- $\gamma = j^* - j = 0.67$

$$Z_{\frac{1}{3}} = 0.33 \times x_{(2)} + 0.67 \times x_{(3)} = 0.33 \times 0 + 0.67 \times 2 = 1.34$$

Interpretazione del risultato

- "*in circa un terzo delle celle di rilevazione abbiamo temperature non superiori a* $1.34 C^\circ$"
- "*approssimativamente il 66.66% (pari a 2/3) delle celle nelle posizioni più calde, riportano temperature superiori ai* $1.34 C^\circ$"
- "*circa un terzo delle celle nelle posizioni più fredde misurano temperature al massimo pari a* $1.34 C^\circ$"

Notes

L'approssimazione (3.3)

- può essere considerata una correzione dei quantili empirici
- corregge, almeno in parte, gli artifici introdotti dalla discontinuità (salti) dell'ECDF
- introduce un'approssimazione/smoothing chiara e semplice da implementare
- è il metodo di approssimazione default di R

Osservazione finale

Se i dati sono tutti distinti, ovvero in assenza di ties (es: X continua e misurata senza discretizzare eccessivamente), allora è facile verificare che dalla lista ordinata

$$x_{(1)} < x_{(2)} < \cdots < x_{(j)} < \cdots < x_{(n)}$$

Possiamo ottenere facilmente $\mathbb{F}(x_{(j)}) = \frac{j}{n}$

Esempi/Applicazioni $\longrightarrow$ `R` **script file**

Notes