CORSO DI STATISTICA (parte 1) - ESERCITAZIONE 1

Dott.ssa Antonella Costanzo

a.costanzo@unicas.it

Un breve riepilogo: caratteri, unità statistiche e collettivo

- UNITA' STATISTICA: oggetto dell'osservazione di un fenomeno individuale che costituisce il fenomeno collettivo. Le unità statistiche rappresentano gli elementi della popolazione.
- CARATTERE: caratteristica di ciascuna unità statistica oggetto di studio. Il carattere è una variabile che si vuole misurare.
- MODALITA': modo con cui si presenta il carattere. La modalità rappresenta il valore che può assumere la variabile oggetto di studio.
- FREQUENZA: numero di volte in cui osservo una certa modalità del carattere.

Classificazione dei caratteri statistici: alcune definizioni

I caratteri (o variabili) si distinguono in due grandi gruppi:

- **carattere qualitativo**: sconnesso (definito su scala nominale), ordinale (definito su scala ordinale).

  Ad esempio, *Stato Civile* è un carattere qualitativo sconnesso definito su scala nominale. Le modalità sono gli attributi del carattere *Stato Civile* (celibe/nubile, coniugato, separato). Un particolare tipo di carattere qualitativo sconnesso è la variabile dicotomica cosi chiamata perché può assumere solo due modalità (vero/falso, si/no, maschio/femmina).

  Il carattere *Titolo di Studio* è qualitativo ordinale le corrispondenti modalità (Licenza Elementare, Licenza Media, Diploma, Laurea) sono ordinabili

- **carattere quantitativo**: discreto, continuo

  Ad esempio, *Numero di figli* è un carattere quantitativo discreto, assume un numero finito di modalità definite nell'insieme dei numeri naturali. *Peso* è un carattere quantitativo continuo, assume potenzialmente un qualunque valore definito nell'insieme dei numeri reali.

(alcune) Prime risposte a due domande fondamentali:

1. **CHE TIPO DI VARIABILE HO A DISPOSIZIONE?**

   -Che cosa ha senso calcolare?

2. **COME POSSO RAPPRESENTARE I DATI?**

   -rappresentazione tabellare

   -quale rappresentazione grafica è appropriata per i dati?

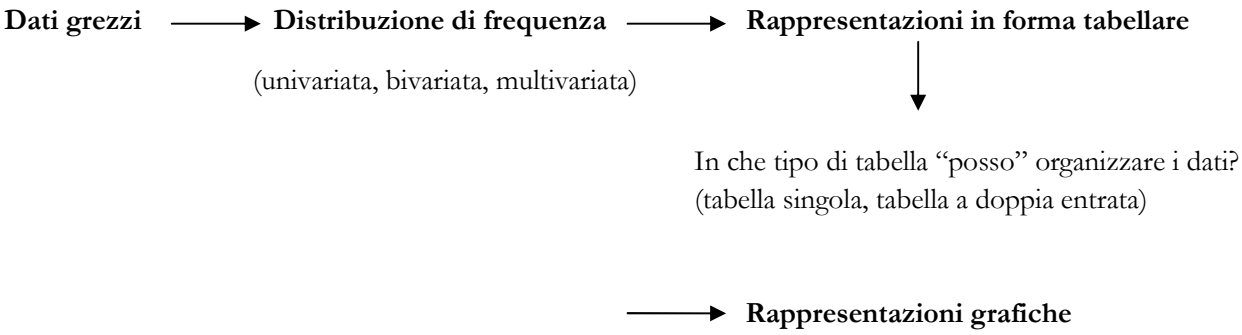
   -quali informazioni (preliminari) posso ottenere dai dati?

Esercizio 1. Che tipo do variabile ho a disposizione?

Definire la tipologia dei seguenti tipi di caratteri

Caratteri	Descrizione
Reddito	Quantitativo continuo
Sesso	Qualitativo sconnesso; nominale
Settore di attività di un'azienda	Qualitativo sconnesso; nominale
Città di residenza	Qualitativo sconnesso; nominale
Stato civile	Qualitativo sconnesso; nominale
Distanza	Quantitativo continuo
Numero di figli	Quantitativo discreto
Voto di laurea	Quantitativo discreto
Numero di pezzi prodotti	Quantitativo discreto
Numero di risposte esatte	Quantitativo discreto
Titolo di studio	Qualitativo ordinale; scala ordinale
Anni di studio	Quantitativo continuo
Peso	Quantitativo continuo

Esercizio 2. Come posso rappresentare i dati?



Definizione: la distribuzione di frequenza descrive come si distribuisce un carattere rispetto alla sua modalità. Le distribuzioni di frequenza possono essere assolute, relative, percentuali e cumulate. Utilizzando le distribuzioni di frequenza è possibile, in genere, rappresentare graficamente i caratteri oggetto di studio. Ogni rappresentazione grafica deve essere appropriata rispetto al tipo di carattere.

Notazione:

Frequenze assolute $\boldsymbol{n_i}$: numero di volte che osservo la modalità del carattere; totale frequenze N

Frequenze relative $\boldsymbol{f_i} = {}^{\boldsymbol{n_i}}/_{\boldsymbol{n}}$: frequenza assoluta associata alla i-esima modalità diviso il totale delle frequenze osservate

Frequenze cumulate, assoluta $\boldsymbol{N_i}$, relativa $\boldsymbol{F_i}$:

La frequenza cumulata assoluta corrispondente ad una certa modalità di un carattere, indica il numero di unità della popolazione considerata che presentano un valore del carattere minore o uguale a quella modalità. Analogamente le frequenze cumulate relative (e percentuali) si riferiscono a frazioni del collettivo considerato. Le frequenze cumulate si ottengono "cumulando" progressivamente le frequenze assolute o relative associate a ciascuna modalità del carattere

Classificazione dei caratteri , distribuzioni di frequenza e rappresentazioni grafiche

Da un'intervista di 10 capifamiglia si sono rilevati i seguenti dati:

Capifamiglia	Stato civile	Titolo di Studio	Numero figli	Peso (kg)
1	Celibe	Laurea	0	72
2	Separato	Laurea	2	65
3	coniugato	Lic.media	1	70
4	Coniugato	Laurea	3	82
5	Celibe	Lic.elementare	2	80
6	Celibe	Lic.elementare	4	78
7	Separato	Diploma	1	77
8	Separato	Diploma	2	64
9	Coniugato	Lic.media	0	89
10	Celibe	Diploma	2	92

Quesiti:

1. Indicare il numero di unità statistiche
2. Indicare quali sono i caratteri e la tipologia (qualitativo nominale/ ordinale, quantitativo discreto/ continuo)
3. Costruire le distribuzioni di frequenza assolute, relative, percentuali e cumulate relative per le variabili *Stato Civile, Titolo di studio, Numero di figli, Peso* e fornire un'opportuna rappresentazione grafica.

Soluzione Q.1

Le unità statistiche sono i singoli individui, il collettivo è formato da 10 capifamiglia

Soluzione Q.2

I caratteri sono le variabili Stato Civile (qualitativo nominale), Titolo di Studio (qualitativo ordinale), Numero di figli (quantitativo discreto) e Peso (quantitativo continuo).

Soluzione Q.3

Distribuzioni uni variate per le variabili *Stato Civile*, *Titolo di Studio*, *Numero di figli e Peso*

Stato Civile	n_i	$f_i = {n_i}/{n}$	p_i	N_i	F_i
Celibe	4	0.4	40%	4	0.4
Coniugato	3	0.3	30%	7	0.7
Separato	3	0.3	30%	10	1
Totale (n)	10	1	100%		

Titolo di studio	n_i	$f_i = {n_i}/{n}$	p_i	N_i	F_i
Lic.elementare	2	0.2	20%	2	0.2
Lic.media	2	0.2	20%	4	0.4
Diploma	3	0.3	30%	7	0.7
Laurea	3	0.3	30%	10	1
Totale (n)	10	1	100%		

Numero figli	n_i	$f_i = {n_i}/{n}$	p_i	N_i	F_i
0	2	0.2	20%	2	0.2
1	2	0.2	20%	4	0.4
2	4	0.4	40%	8	0.8
3	1	0.1	10%	9	0.9
4	1	0.1	10%	10	1
Totale (n)	10	1	100%		

Come già visto, la variabile Peso è un carattere quantitativo continuo. Per fornire una rappresentazione tabellare di una variabile continua si ricorre generalmente ad una suddivisione in classi delle modalità di risposta:

Distribuzione organizzata in classi equifrequenti per la variabile *Peso*

Peso (Kg)	n_i	$f_i = {n_i}/{n}$	p_i	N_i	F_i
[60,70)	2	0.2	20%	2	0.2
[70, 80)	4	0.4	40%	6	0.6
[80,100)	4	0.4	40%	10	1
Totale (N)	10	1	100%		

Le rappresentazioni grafiche appropriate: diagramma a barre (e/o diagramma a torta), diagramma di Pareto per il carattere Titolo di Studio, diagramma a barre (e/o diagramma a torta) per il carattere Stato Civile, diagramma a bastoncini per la variabile Numero Figli, istogramma per la variabile Peso.

A titolo di esempio riportiamo le rappresentazioni grafiche delle seguenti variabili: Titolo di Studio, Numero di figli e Peso. Le rimanenti altre sono lasciate per esercizio.

Diagramma a barre per la variabile Titolo di Studio

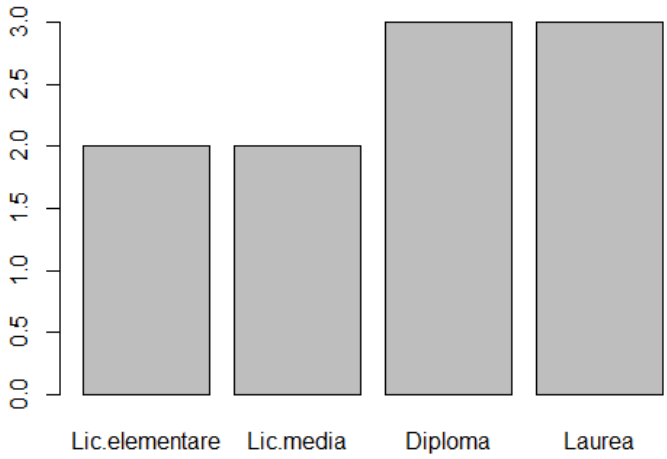


Diagramma a torta per la variabile Titolo di Studio

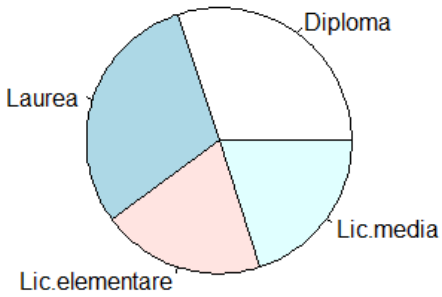
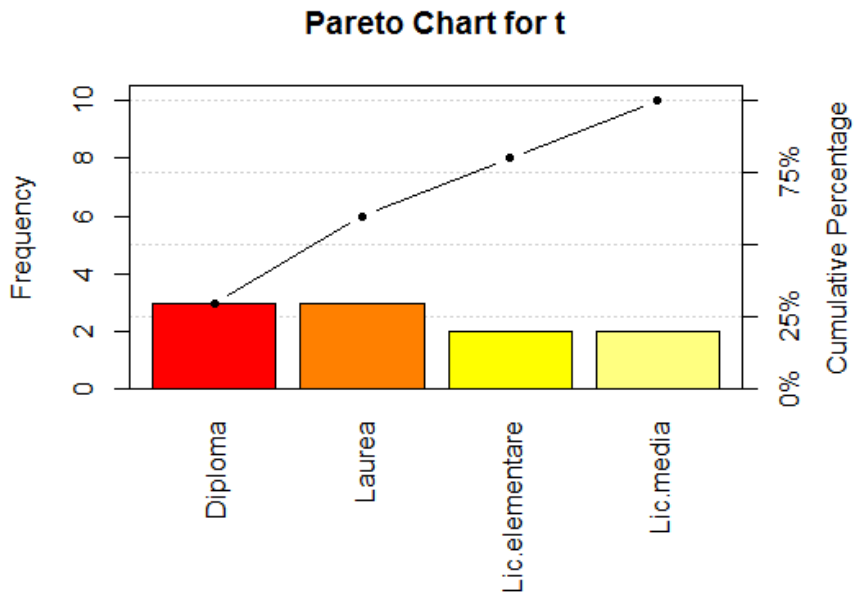


Diagramma di Pareto per la variabile Titolo di Studio

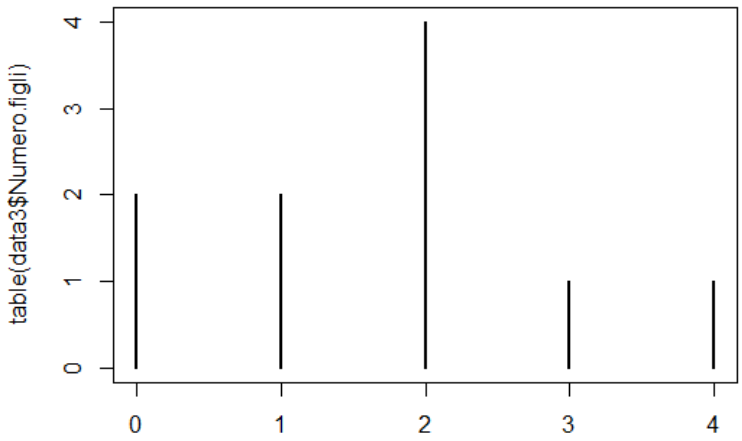


Costruzione e interpretazione:

Il diagramma di Pareto consiste in un diagramma a barre della distribuzione percentuale di un fenomeno, ordinato in senso decrescente, affiancato al grafico delle frequenze cumulate (la linea). Sull'asse delle ascisse si trovano le modalità assunte dalla variabile ordinate in senso decrescente. Sull'asse delle ordinate (a sinistra) si trovano le frequenze assolute mentre a destra si trovano le frequenze cumulate relative associate a ciascuna modalità del carattere.

Nota: il diagramma di Pareto è utile soprattutto per rappresentare l'importanza delle differenze causate da un certo fenomeno. Questo tipo di grafico può aiutare a stabilire quali sono i maggiori fattori che hanno influenza su un dato fenomeno, ed è quindi un utile strumento nelle analisi, nei processi decisionali, nella gestione della qualità ed in numerosi altri settori.

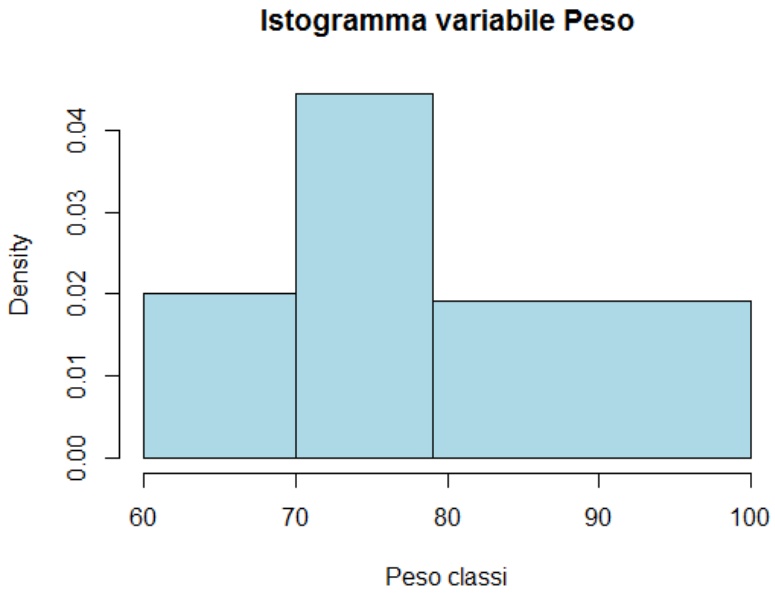
Diagramma a bastoncini per la variabile Numero di Figli



Per la variabile Peso, per disegnare l'istogramma occorrono le densità di frequenza d_i (assolute o relative) definite come il rapporto tra la frequenza (assoluta o relativa) associata a ciascuna classe e l'ampiezza della classe a_i. Per ogni classe abbiamo un rettangolo di base a_i e di altezza d_i. Le barre possono avere larghezza diversa tra loro (a differenza del bar-plot in cui tutte le barre hanno la stessa larghezza). L'area del rettangolo è proporzionale alle frequenze. **ATTENZIONE**: non si può costruire l'istogramma con le sole frequenze tranne nel caso di classi equiampie.

Peso (Kg)	n_i	$f_i = {n_i}/{n}$	p_i	F_i	a_i	$d_{i(\text{ass})} = {n_i}/{a_i}$	$d_{i(\text{rel})} = {f_i}/{a_i}$
[60,70)	2	0.2	20%	0.2	10	0.2	0.02
[70, 80)	4	0.4	40%	0.6	10	0.4	0.04
[80,100)	4	0.4	40%	1	20	0.4	0.02
Totale (n)	10	1	100%				

Istogramma per la variabile Peso



°°°

Esercizio 4.

Da un collettivo di 20 individui si è rilevata la seguente distribuzione unitaria multipla relativa ai caratteri *Sesso, Età, Numero di auto possedute*

Unità	*Età*	*Sesso*	*Nr. auto possedute*
1	35	m	1
2	37	m	2
3	59	f	1
4	54	m	0
5	44	f	2
6	38	m	1
7	62	f	1
8	71	f	0
9	56	m	3
10	60	m	2
11	33	m	2
12	46	f	4
13	41	f	3
14	53	m	1
15	38	f	1
16	55	m	2
17	50	m	3
18	63	m	1
19	35	f	0
20	51	m	2

Quesiti:

1. Costruire le distribuzioni di frequenza semplici per i caratteri *Sesso* e *Nr. auto possedute*.
2. Si consideri il carattere *Età* suddiviso in classi [30,40), [40,50), [50,60), [60,80) e si costruiscano le corrispondenti distribuzioni di frequenza assolute, relative e percentuali.
3. Rappresentare mediante i grafici ritenuti più idonei le distribuzioni di frequenza assolute di *Sesso, Nr. Automobili possedute* e *Età*.
4. Costruire le distribuzioni richieste al punto 2 in forma cumulata e rappresentare graficamente la funzione di ripartizione empirica.

°°°

Soluzione Q.1 Distribuzioni di frequenza unitarie

Distribuzione di frequenza Variabile Sesso

Sesso	n_i	f_i	p_i
m	12	0.6	60%
f	8	0.4	40%
Totale	20	1	100%

Distribuzione di frequenza Variabile Nr.auto possedute

Nr.auto possedute	n_i	f_i	p_i	N_i	F_i
0	3	0.15	15%	3	0.15
1	7	0.35	35%	10	0.50
2	6	0.3	30%	16	0.80
3	3	0.15	15%	19	0.95
4	1	0.05	5%	20	1
Totale	20	1	100%		

Soluzione Q.2 e Q.3 Distribuzioni di frequenza per la variabile Età organizzata in classi e rappresentazioni grafiche

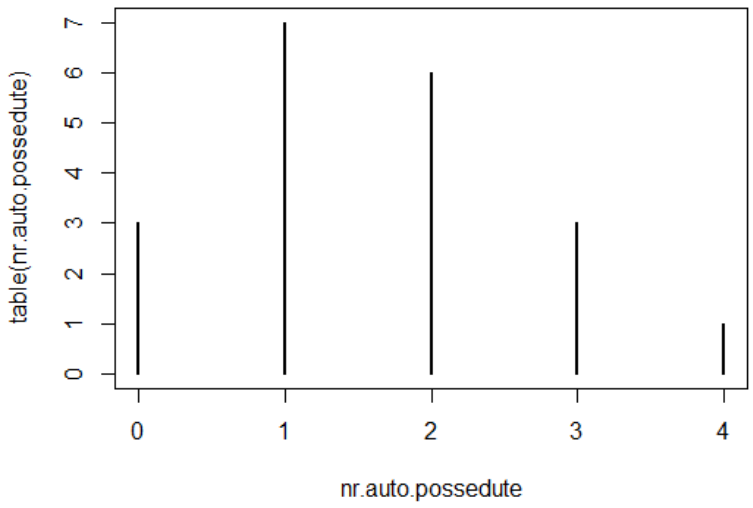
Distribuzione di frequenza della variabile *Età*

Età (classi)	$\boldsymbol{n_i}$	$\boldsymbol{f_i}$	$\boldsymbol{p_i}$	$\boldsymbol{a_i}$	$\boldsymbol{d_{i(ass)} = {n_i}/{a_i}}$	$\boldsymbol{d_{i(rel)} = {f_i}/{a_i}}$
[30,40)	6	0.30	30%	10	0.6	0.03 (=3%)
[40,50)	3	0.15	15%	10	0.3	0.015 (=1.5%)
[50,60)	7	0.35	35%	10	0.7	0.035 (=3.5%)
[60,80)	4	0.20	20%	20	0.2	0.01 (=1%)
Totale (n)	20	1	100%			

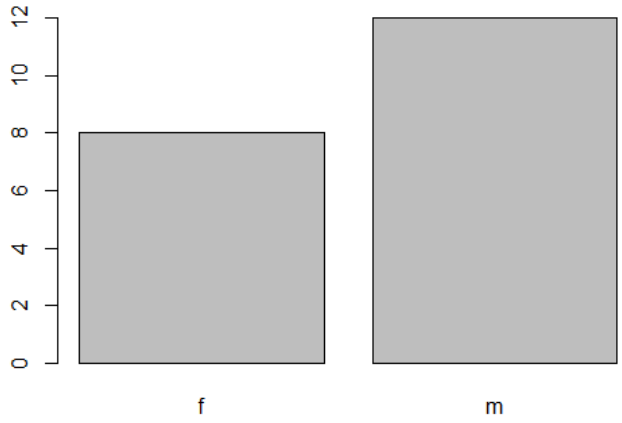
Rappresentazioni grafiche appropriate:

Il diagramma a barre (bar-plot) si utilizza per il carattere *Sesso,* il diagramma a bastoncini si usa per rappresentare il carattere Nr. auto possedute. Per la variabile *Età* (carattere quantitativo continuo organizzato in classi) la corretta rappresentazione grafica è l'istogramma.

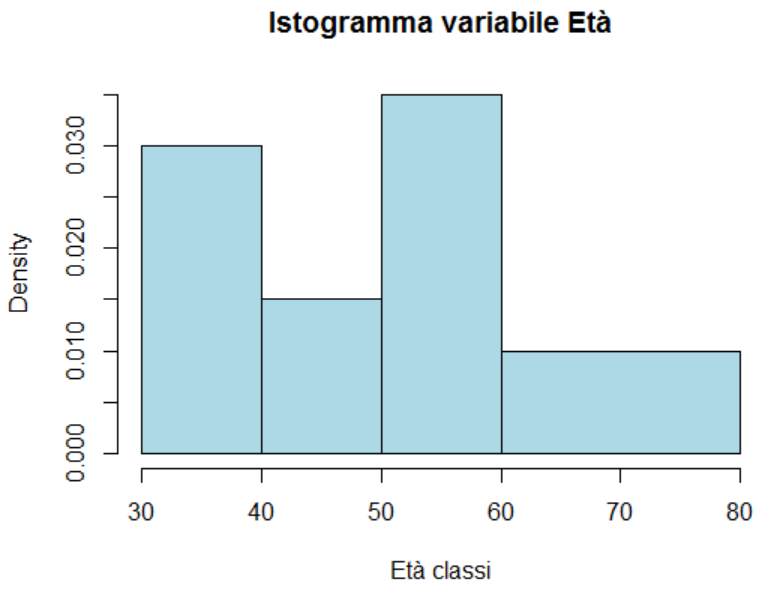
Diagramma a bastoncini per la variabile Nr. auto possedute



Bar-plot variabile sesso



Istogramma variabile Età



Soluzione Q.5 Frequenze cumulate e funzione di ripartizione empirica

Frequenze cumulate per la variabile *Età*

Età (classi)	n_i	f_i	p_i	N_i	F_i
[30,40)	6	0.30	30%	6	0.3
[40,50)	3	0.15	15%	9	0.45
[50,60)	7	0.35	35%	16	0.80
[60,80)	4	0.20	20%	20	1
Totale (n)	20	1	100%		

Nota utile: a partire dalla distribuzione di frequenza cumulata associata ad una certa modalità di un carattere è possibile ottenere la frequenza associata a quella certa modalità. Basta sottrarre alla frequenza cumulata associata alla j-esima modalità la frequenza cumulata corrispondente alla j-1 esima modalità. Ad esempio per la classe di Età [40,50), la corrispondente frequenza assoluta cumulata è 9; sottraendo la frequenza cumulata associata alla classe precedente, 6 si ottiene 9-6=3 che rappresenta proprio il valore della frequenza assoluta associata alla classe d'età [40,50). Stesso procedimento per le frequenze cumulate relative (o percentuali).

Le frequenze cumulate consentono di costruire la funzione di ripartizione empirica, $F(x)$ che rappresenta il numero di osservazioni del fenomeno minori o uguali del valore x. E' costruita ponendo sull'asse delle ascisse i valori della variabile e sull'asse delle ordinate le frequenze cumulate assolute (o relative).

Se $\{x_1 \dots x_n\}$ sono le osservazioni (ordinate in senso crescente), con frequenze relative $f_1 \dots f_n$ la funzione di ripartizione ha espressione analitica

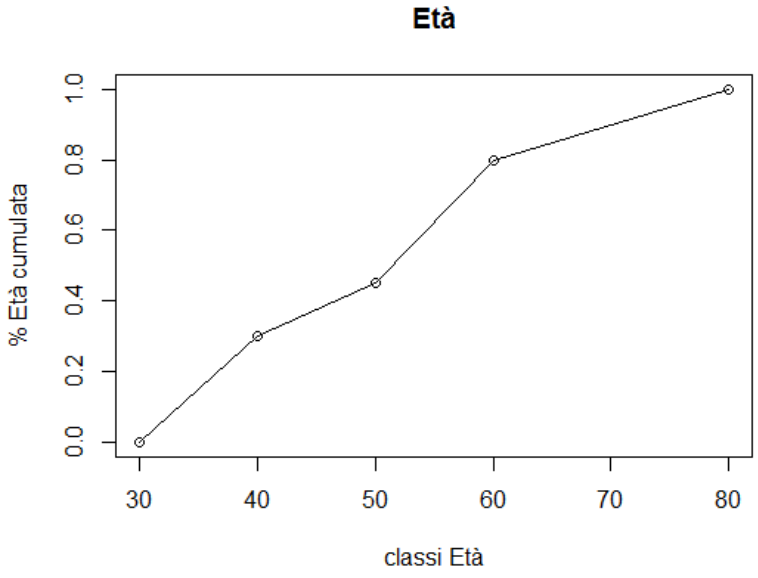
$$F(x) = \left\{ \begin{array}{ll} 0 & x < x_i \\ F_i = \sum_{j \le i} f_i & x_i \le x \le x_{i+1} \\ 1 & x \ge x_n \end{array} \right\}$$

In particolare per la variabile Età, la funzione di ripartizione empirica corrispondente è:

$$F(x) = \left\{ \begin{array}{r} 0 \; per \; x < 30 \\ 0.3 \; per \; 30 \le x < 40 \\ 0.45 \; per \; 40 \le x < 50 \\ 0.80 \; per \; 50 \le x < 60 \\ 1 \; per \; 60 \le x < 80 \end{array} \right\}$$

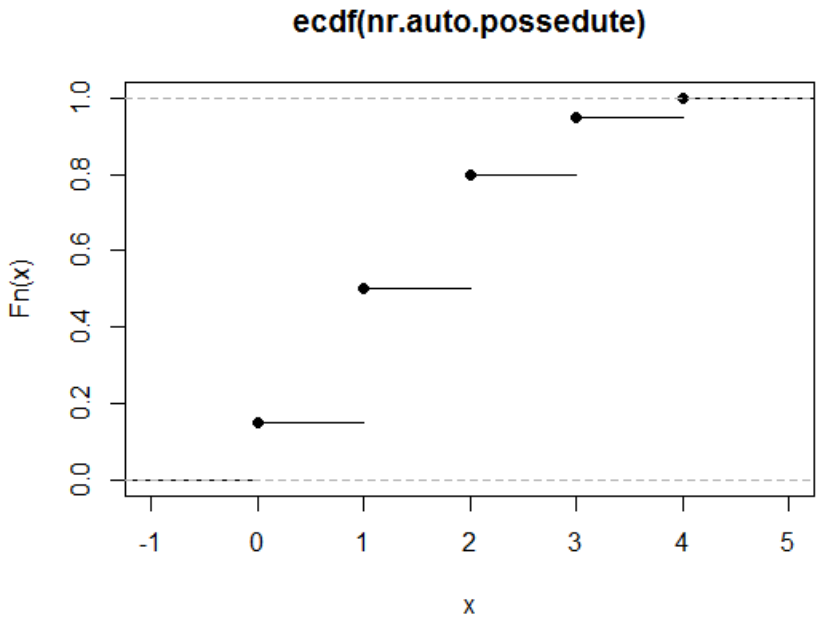
Graficamente, in caso di variabili quantitative continue organizzate in classi (es. Età) la funzione di ripartizione è rappresentata da una spezzata che unisce i punti di coordinate $(x, F(x))$.

Grafico 6. Funzione di ripartizione empirica per la variabile Età



Per caratteri discreti, il grafico della funzione di ripartizione empirica si presenta come una funzione a gradini in cui i "salti" sono in corrispondenza delle modalità del carattere. Ad esempio, per la variabile Nr.auto possedute:

Funzione di ripartizione empirica per la variabile Nr.auto possedute



La funzione di ripartizione empirica per variabili quantitative (o anche per qualitative ordinabili) e con campo di variazione $[x_o, x_k]$ gode delle seguenti proprietà:

- $F(x)$ monotòna non decrescente
- $F(x)$ =0 per $x < x_0$
- $F(x)$ =1 per $x > x_k$

La frequenza retrocumulata assoluta corrispondente ad una certa modalità di un carattere, indica il numero di unità della popolazione considerata che presentano un valore del carattere maggiore o uguale a quella modalità. Analogamente le frequenze retrocumulate relative (e percentuali) si riferiscono a frazioni del collettivo considerato.

Esempio:

Età (classi)	n_i	N_i	RN_i	$f_i = {n_i}/{\mathrm{n}}$	F_i	RF_i
[30,40)	6	6	20	0.30	0.3	1
[40,50)	3	9	14	0.15	0.45	0.70
[50,60)	7	16	11	0.35	0.80	0.55
[60,80)	4	20	4	0.20	1	0.20
Totale (n)	20			1		

Appendice 1

Rappresentazione di caratteri quantitativi continui

Nel caso di una variabile quantitativa continua non è possibile far corrispondere ad ogni modalità la rispettiva frequenza in quanto il carattere potrebbe assumere infinite distinte modalità (ognuna delle quali con frequenza assoluta pari a 1). Per fornire una rappresentazione tabellare di una variabile continua si ricorre quindi ad una suddivisione in classi delle modalità di risposta.

La definizione di classi di modalità deve rispondere ai seguenti requisiti:

- Il numero di classi deve essere abbastanza piccolo da fornire un'adeguata sintesi ma abbastanza grande da mantenere un livello accettabile di dettaglio dell'informazione.
- Le classi devono essere disgiunte (ogni modalità del carattere deve poter essere assegnata ad una e una sola classe).
- Le classi devono comprendere tutte le modalità del carattere.
- Le classi, se possibile, devono avere la stessa ampiezza.

Esercizio. Criteri di suddivisione in classi

Si considerino le temperature minime in gradi centigradi previste per sabato 13 ottobre:

Città	***Temperature minime***
Torino	13
Milano	12
Venezia	8
Genova	15
Bologna	10
Firenze	9
Ancona	13
Perugia	7
Roma	14
Pescara	13
Napoli	12
Palermo	18
Cagliari	19

Quesiti:

1. Suddividere la distribuzione in k=4 classi equiampie di ampiezza pari a tre.
2. Suddividere la distribuzione in k=4 classi equifrequenti

Soluzione Q.1. Suddivisione in classi equiampie

L'ampiezza delle classi è data dalla differenza tra il valore massimo osservato e il valore minimo diviso il numero di classi che si vogliono creare (k), ossia : $\frac{V_{max}-V_{min}}{k}$

Nel nostro caso: $\frac{19-7}{4} = 3$

Suddivisione in classi equiampie

Classi	n_i	f_i	F_i
(7, 10]	4	0.308	0.308
(10, 13]	5	0.385	0.692
(13, 16]	2	0.154	0.846
(16, 19]	2	0.154	1
	13	1	

Soluzione Q.2 Suddivisione in classi equifrequenti

Gli elementi necessari al calcolo delle classi equifrequenti sono:

- Ordinamento crescente dei valori delle modalità del carattere
- Frequenza associata a ciascuna di esse

Nel nostro caso, dapprima si ordinano i dati: (7; 8; 9; 10; 12; 12; 13; 13; 13; 14; 15; 18; 19); poi si costruiscono delle classi in modo che tutte abbiano la stessa frequenza (se possibile)

- Calcolo numero di classi che abbiano la frequenza desiderata (es.freq. desiderata= 3)

  Nr.classi= $\frac{N}{freq.desiderata} = \frac{13}{3} \approx 4$

- Definisco la frequenza associata al nr. di classi desiderate, Freq.classi = $\frac{N}{Nr.classi} = \frac{13}{4} = 3$

Suddivisione in classi equifrequenti

Classi	n_i	f_i	F_i
[7, 9]	3	0.231	0.231
(9, 12]	3	0.231	0.462
(12, 14]	4	0.308	0.769
(14, 19]	3	0.231	1
	13	1	

ooo