Analisi Univariata

Le funzioni dell’ analisi monovariata

- Lo scopo per cui si raccolgono e si organizzano dati mediante una matrice è di investigare le relazioni fra proprietà (dunque fra variabili).
- Quindi deve essere considerata una fase preliminare per effettuare analisi più complesse, fase obbligatoria in quanto fondamentale.
- L’analisi monovariata ha due funzioni principali:

1. Controllare la plausibilità dei valori – *wild code check*: Controllo dei valori selvaggi ossia dei possibili errori di rilevazione/immissione nella matrice dati.

2. Segnalare squilibri nella distribuzione e opportunità di aggregazione

Tipi di squilibrio

VARIABILE	SQUILIBRIO	SOLUZIONE
Categoriale non ordinata	Frequenze troppo alte/basse	Aggregazione delle modalità secondo il criterio *relazionale*
Categoriale ordinata	Frequenze troppo alte/basse	Aggregazione delle modalità secondo il criterio *monotonico*
Cardinale	Frequenze posizionate asimmetricamente	Trasformazioni matematiche sui valori

Tipo di comune di residenza	Va	%
Urbani	3642	52%
Quasi urbani	298	4%
Semi urbani	349	5%
Semi rurali	562	8%
Quasi rurali	431	6%
Rurali	1352	19%
Imprecisato	390	6%
TOTALE	7024	100%

Tipo di comune di residenza	Va	%
Urbani centrali	2241	32%
Urbani periferici	1208	17%
Quasi urbani	647	9%
Quasi rurali	1186	17%
Rurali	1352	19%
Imprecisato	390	6%
TOTALE	**7024**	**100%**

La distribuzione di frequenza di una variabile

- La distribuzione di frequenza di una variabile è la rappresentazione *sintetica* dei dati in forma tabellare, attraverso la quale ad ogni valore della modalità della variabile viene associata la frequenza (numero dei casi) con la quale essa si presenta.
- Le colonne in cui vengono indicate le etichette numeriche e semantiche vengono denominate *colonne madri.*
- Ciascuna frequenza rappresenta il numero dei casi che ricade nella modalità corrispondente;
- Il totale corrisponde al numero dei casi (**N**).

Valore	Modalità	Frequenza
0	Modalità 1	15
1	Modalità 2	42
2	Modalità 3	191
3	Modalità 4	165
4	Modalità 5	45
5	Modalità 6	22
TOTALE		N

Valore	Modalità	Frequenza assoluta
0	Nessuno	15
1	Lic. Elementare	42
2	Lic. Media Inferiore	191
3	Diploma Media Superiore	165
4	Laurea	45
5	Specializzazione post-laurea	22
TOTALE		**480**

Cinque tipi di frequenze

1. **Freq. Assolute** = numero dei casi che presentano quel valore senza che si effettui alcuna manipolazione (conteggio).

2. **Freq. Relative** = proporzione (rapporto) del numero dei casi che presentano quel valore rapportato con il numero totale dei casi; il totale sarà sempre uguale ad 1

3. Freq. Relative percentuali = proporzione (rapporto) del numero dei casi che presentano quel valore rapportato con il numero totale dei casi moltiplicato per 100; il totale sarà sempre uguale a 100

4. **Freq. Cumulate** = in corrispondenza di ogni valore si riporta la somma delle frequenze di quel valore e dei valori inferiori; il totale dell'ultima categoria sarà sempre uguale ad 1

5. Freq. Cumulate **%**= in corrispondenza di ogni valore si riporta la % di quel valore e dei valori inferiori; il totale dell'ultima categoria sarà sempre uguale a 100.

6. **Freq. Retro-cumulate** = in corrispondenza di ogni valore si riporta la somma delle frequenze di quel valore e dei valori superiori; il totale della prima categoria sarà sempre uguale ad 1

7. Freq. Retro-cumulate **%** = in corrispondenza di ogni valore si riporta la % di quel valore e dei valori superiori; il totale della prima categoria sarà sempre uguale a 100.

Valore	Modalità	Frequenza assoluta	Frequenza relativa	Frequenza relativa %	Frequenza cumulata	Frequenza cumulata %	Frequenza retro-cumulata	Frequenza retro-cumulata %
0	Nessuna	15	0,03	3,13%	15	3,13%	480	100,00%
1	Licenza elementare	42	0,09	8,75%	57	11,88%	465	96,88%
2	Licenza Media Inferiore	191	0,4	39,79%	248	51,67%	423	88,13%
3	Diploma Media Superiore	165	0,34	34.38%	413	86,04%	232	48,34%
4	Laurea	45	0,09	9,38%	458	95,42%	67	13,96%
5	Specializzazione post-laurea/dottorato	22	0,05	4,58%	480	100,00%	22	4,58%
TOTALE		480	1	100,00%	----	----	----	----

Quali frequenze si devono includere in una tabella?

E’ necessario seguire sei regole:

- **Frequenze percentuali**: preferire le frequenze percentuali: questo consente una maggiore leggibilità e confrontabilità di differenti distribuzioni di frequenza.
- **Frequenze assolute**: A volte, si possono ritenere interessanti anche le frequenze assolute, (vedi tabella).
- **Parsimonia**: inserire solo le informazioni indispensabili, indicare solo un tipo di frequenza (assoluta, relativa, percentuale, etc…)
- **Numerosità dei casi**: nel caso si utilizzino le frequenze percentuali (più usate) è **necessario** indicare il numero complessivo dei casi in valore assoluto (N) in questo modo è possibile ricalcolare le frequenze assolute della distribuzione.
- **Utilità delle percentuali**: **non** usare le frequenze percentuali se N è minore di 50 casi (riportare le percentuali se si vuole comparare più distribuzioni di frequenza).
- ***Fallacy of the misplaced precision*****:** evitare la tendenza a riportare percentuali con un numero eccessivo di decimali, ma riportare solo quelli strettamente necessari (vedi lucido successivo).

	Valori assoluti (in migliaia)		Valori percentuali	
	Lombardia	Emilia R.	Lombardia	Emilia R.
Forza Italia	1510	451	23,6%	15,1%
Alleanza Nazionale	575	344	9,0%	11,5%
Ccd-Cdu	298	144	4,6%	4,8%
Lega Nord	1636	216	25,5%	7,2%
Pds	965	1065	15,1%	35,7%
Lista Dini	267	116	4,2%	3,9%
Ppi	398	238	6,2%	8,0%
Verdi	152	75	2,4%	2,5%
Rifond. Com.	437	249	6,8%	8,3%
Altri	168	90	2,6%	3,0%
Totale	6406	2988	100,0%	100,0%

Come si arrotondano i decimali?

Per evitare la *Fallacy of the misplaced precision*, una possibile regola, suggerita da Marradi (2001), è la seguente:

- se N ≥ 1.000 casi 1 cifra decimale
  se 1.001 ≥ N ≤ 10.000 casi 2 cifre decimali

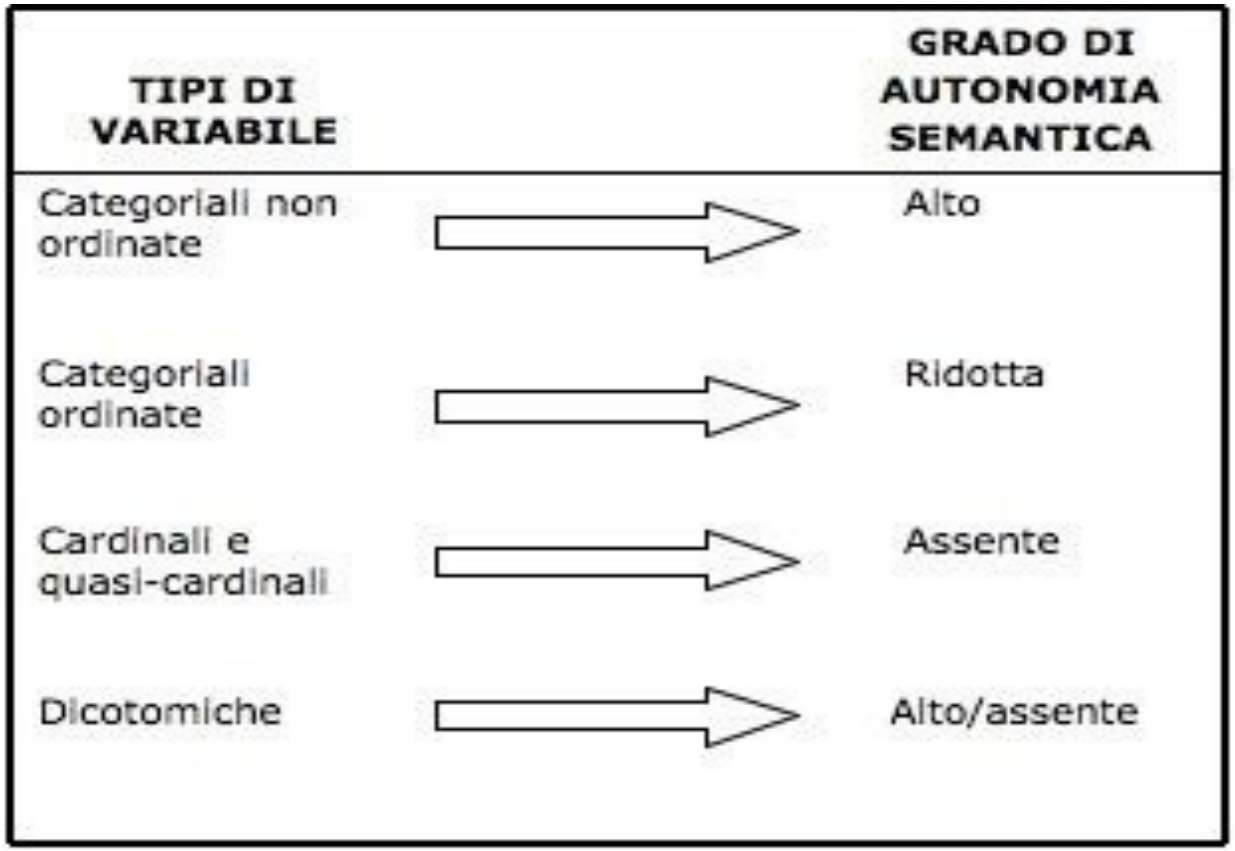
Arrotondamenti corretti:

- da 0 a 4 → arrotondamento per difetto.
  - Es.: 16,73 → 16,7
- da 6 a 9 → arrotondamento per eccesso
  - Es.: 16,78 → 16,8
- se 5 →controllare il decimale successivo o, meglio, troncare

Età (anni)	Frequenza	Percentuale	Percentuale cumulata
15	43	4,3	4,3
16	50	5,0	9,3
17	41	4,1	13,4
18	62	6,2	19,6
19	53	5,3	24,9
20	31	3,1	28,0
21	57	5,7	33,7
22	49	4,9	38,6
23	40	4,0	42,6
24	66	6,6	49,2
25	73	7,3	56,4
26	39	3,9	60,3
27	49	4,9	65,2
28	34	3,4	68,6
29	53	5,3	73,9
30	66	6,6	80,5
31	30	3,0	83,5
32	52	5,2	88,7
33	70	7,0	95,7
34	43	4,3	100,0
Totale	***1001***	***100,0***	

Età (in classi)	Frequenze	Percentuale	Percentuale cumulata
15-17	134	13,4	13,4
18-23	292	29,2	42,6
24-29	314	31,4	73,9
30-34	261	26,1	100,0
Totale	***1001***	***100,0***	

(Autonomia semantica)



L’importanza delle distribuzioni equilibrate

L’alto grado di autonomia semantica delle variabili categoriali ordinate impone che la distribuzione di frequenza sia equilibrata

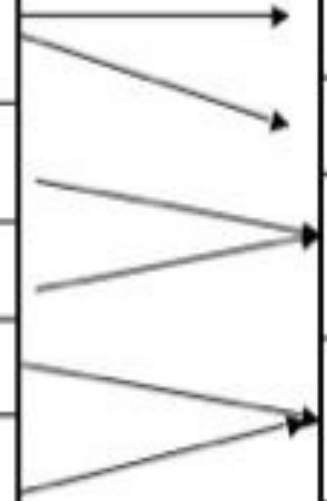
Per due motivi:

- le categorie con frequenze troppo *basse* danno informazioni *distorte*, le categorie con frequenze troppo *alte* danno invece informazioni *scarse* sull’andamento generale;
- quando si vorrà incrociare la variabile considerata con un’altra, se si presenta un numero troppo alto di modalità (distribuzione *sensibile*) e/o un numero troppo basso di casi, si formerà una tabella a doppia entrata con numerose celle, molte delle quali vuote o con un basso numero di casi; questo stato di cose non consente il ricorso a determinate tecniche di analisi bivariata dei dati (es. il Chi-quadro).

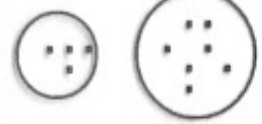
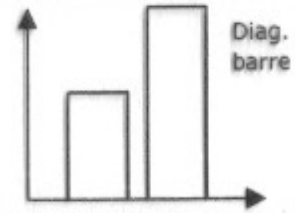
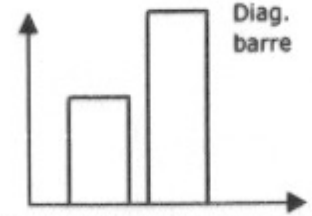
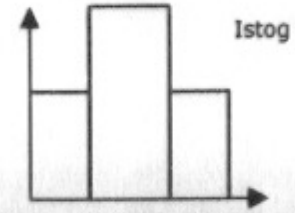
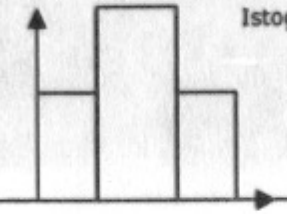
Aggregazione/disaggregazione delle modalità

- Per rendere equilibrate le distribuzioni di frequenza è necessario talvolta procedere ad un'aggregazione/disaggregazione delle frequenze delle modalità di risposta.
  Nella scelta di quali categorie aggregare/disaggregare è necessario seguire alcuni accorgimenti:
  - Riunire categorie di *affine semanticità*;
  - Tenere conto degli obiettivi cognitivi e delle caratteristiche dei fenomeni considerati.

Tipo di comune di residenza	Va	%
Urbani	3642	52%
Quasi urbani	298	4%
Semi urbani	349	5%
Semi rurali	562	8%
Quasi rurali	431	6%
Rurali	1352	19%
Imprecisato	390	6%
TOTALE	7024	100%



Tipo di comune di residenza	Va	%
Urbani centrali	2241	32%
Urbani periferici	1208	17%
Quasi urbani	647	9%
Quasi rurali	1186	17%
Rurali	1352	19%
Imprecisato	390	6%
TOTALE	**7024**	**100%**

	TIPO DI VARIABILE	VALORI CARATTERISTICI DI TENDENZA CENTRALE	NOTE	VALORI CARATTERISTICI DI DISPERSIONE	NOTE	RAPPRESENTAZIONI GRAFICHE (della distribuzione di frequenza)
	CATEGORIALE NON ORDINATA (Es. Genere)	MODA (Categoria più frequente)	M F  MODA	$Sq = \sum_{j=1}^{k} p_j^{\,2}$	$\frac{1}{k}$ Max dispersione 1 (min. dispersione)	 Diag. barre
	CATEGORIALE ORDINATA (Es. Livello di istruzione)	MEDIANA (Valore che divide la distribuzione in due parti uguali)	Non subisce l'influenza dei valori estremi	$d = \frac{4\sum_{h-1}^{k-1} p_h^{(1-ph)}}{K-1}$	0 (MINIMA DISPERSIONE) 1 (MAX DISPERSIONE)	 Diag. barre
CARDINALI	INTERVALLI (Es. Data di nascita)	MEDIA $\bar{x} = \frac{\sum x_i}{n}$	Subisce l'influenza dei valori estremi	SCARTO TIPO $s = \frac{\sqrt{\sum (xi - \bar{x})^2}}{n}$	0 (MINIMA DISPERSIONE + HIGHEST (MAX DISPERSIONE	 Istog
	Metriche/enumerate (Es. Peso/N. libri posseduti)	MEDIA $\bar{x} = \frac{\sum x_i}{n}$	Subisce l'influenza dei valori estremi	COEFFICIENTE DI VARIAZIONE cv= s/ $\bar{x}$	0 (MINIMA DISPERSIONE) + HIGHEST (MAX DISPERSIONE)	 Istog
	QUANTITA' (Es. Reddito)	MEDIA $\bar{x} = \frac{\sum x_i}{n}$	Subisce l'influenza dei valori estremi	COEFFICIENTE DI CONCENTRAZIONE	0 (max dispersione) 1 (minima dispersione)	 Istog

Valori caratteristici delle distribuzioni in categorie non ordinate

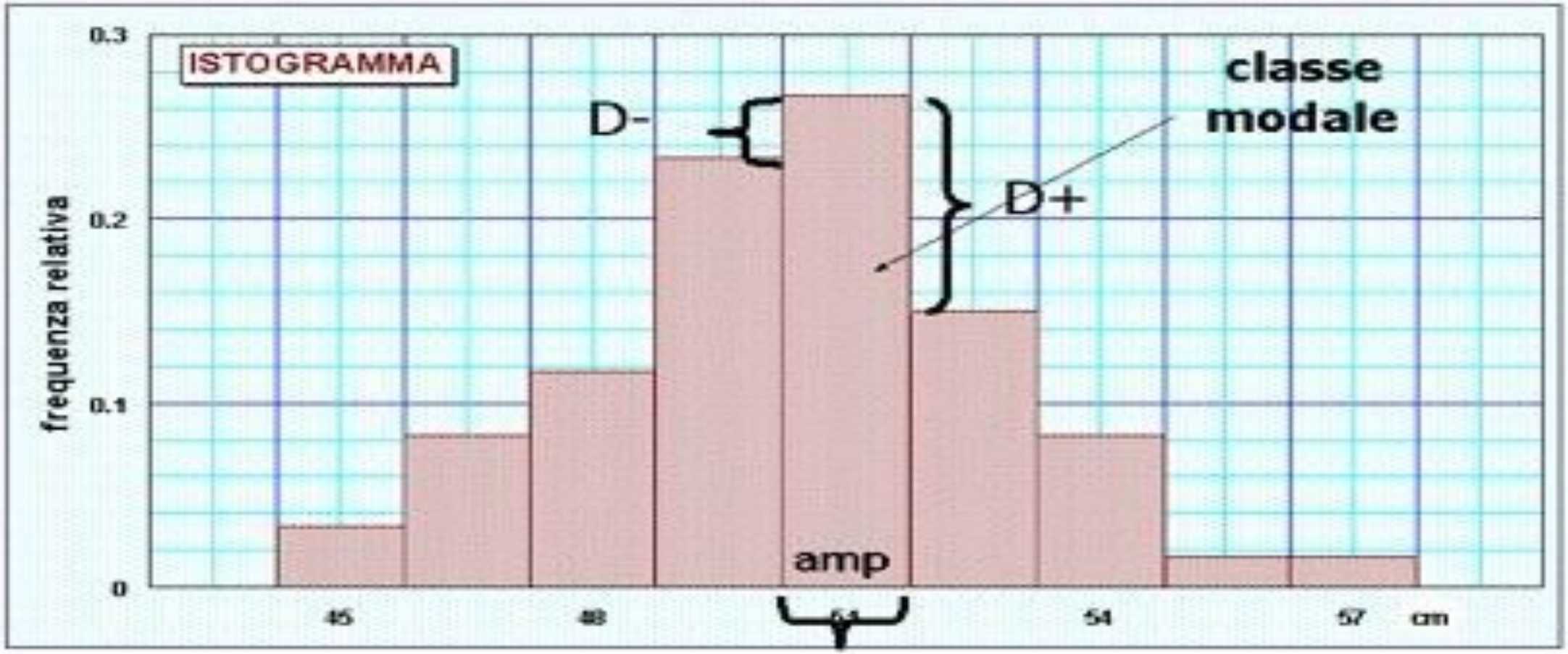
Per analizzare le distribuzioni di dati in categorie non ordinate è possibile ricorrere:

- Misure di tendenza centrale → Moda
- Misure di dispersione → Indice di Galtung (1967)
- Misure di variabilità → Indici di equilibrio/squilibrio (I di Gini)

La moda

- Il valore di tendenza centrale delle distribuzioni di dati in categorie non ordinate è la MODA.
- La moda è la categoria con frequenza (o percentuale) più alta ovvero È la modalità prevalente (con la frequenza più alta) nella distribuzione
- Una distribuzione può presentare più mode: se ce ne sono due, viene detta bimodale; se ve ne sono più di due viene detta multimodale

moda


ISTOGRAMMA
frequenza relativa
0.3
0.2
0.1
0
D-
D+
classe
modale
amp
45
48
54
57
cm

Tipi di distribuzione

Distribuzione unimodale

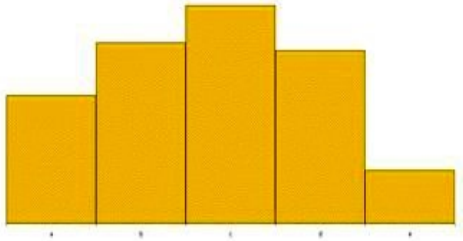
Modalità	%
1. Occupati	50
2. Casalinghe	21
3. Pensionati e inabili	14
4. Disoccupati	8
5. Studenti	7
6. Non accertato	1
totale	100

Distribuzione bimodale

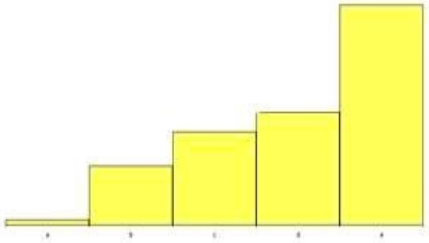
Modalità	%
1. Occupati	35,5
2. Casalinghe	35,5
3. Pensionati e inabili	13,6
4. Disoccupati	7,7
5. Studenti	6,6
6. Non accertato	1,1
totale	100

Alcune specificità delle distribuzioni multi-modali

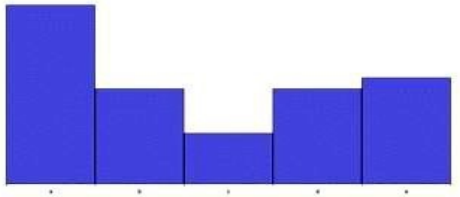
- **Distribuzione unimodale centrale**
  La moda cade nella categoria centrale, mentre le altre frequenze declinano gradatamente dall'uno all'altro lato del valore centrale



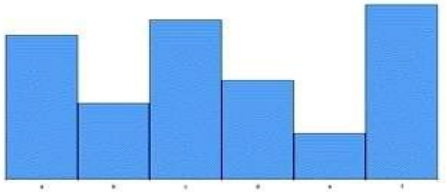
- **Distribuzione unimodale**
  La moda coincide con un valore estremo della distribuzione e le altre frequenze declinano gradatamente fino all'altro estremo



- **Distribuzione bimodale**
  Le frequenze declinano da un estremo all'altro fino ai valori centrali e poi tornano a crescere andando verso l'altro estremo



- **Distribuzione trimodale**
  Distribuzioni con una moda centrale ed una in ciascuno dei valori estremi



Indici di dispersione e equilibrio

- Come si collocano le modalità intorno al centro della distribuzione?
- È opportuno accompagnare le misure di tendenza centrale con misure di *variabilità*
- Una variabile è massimamente omogenea quando concentra tutti i casi in una stessa modalità
- È massimamente eterogenea quando i casi sono distribuiti in tutte le modalità
- Indice omogeneità: $O=p_1^2+p_2^2+\ldots+p_k^2=\sum p_i^2$
- Indice di Gini

$$I_{Gini} = 1 - \sum_{i=1}^{k} f_k^2$$

Distribuzioni di dati con variabili cardinali

Le singole categorie delle variabili cardinali
non hanno alcuna autonomia semantica

↓

Diviene, quindi rilevante, l'andamento globale dell'intera distribuzione.
È comunque possibile ottenere una distribuzione con un numero molto alto di modalità.

Valori caratteristici

I **valori caratteristici** delle distribuzioni di dati con variabili cardinali devono tener conto:

1. delle frequenze di tutte le modalità della distribuzione;
2. del valore "cardinale" delle etichette.

I valori caratteristici delle variabili cardinali si distinguono in:

- Valori di tendenza centrale;
- Valori di dispersione.

I valori di tendenza centrale

Le misure di tendenza centrale che si possono applicare alle variabili cardinali sono anche quelle che si applicano alle variabili categoriali.

Questo perché – come detto nelle precedenti lezioni – le tecniche d'analisi che si possono applicare alle variabili sono **cumulative**:

1. **Moda**
2. **Mediana**
3. **Quantili**
4. **Midrange**
5. **Media aritmetica**

1. La moda

- La moda rappresenta la **categoria** con la frequenza più alta.
- **Esempio**: si voglia calcolare la moda nella distribuzione di frequenza della variabile età di una classe di primo anno del liceo.

2. La mediana

- La mediana di una distribuzione è la **modalità del caso** che lascia dietro di sé il 50% della distribuzione.
- Se **N è dispari** c'è un unico caso centrale: $M_e = x_{\left(\frac{N+1}{2}\right)}$
- Se **N è pari** ci sono due casi centrali che potrebbero generare una distribuzione bimodale qualora i due casi ricadessero in due categorie differenti e la formula è la seguente:

$$M_e = \frac{x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}}{2}$$

La media

La media è il valore che rappresenta la ripartizione di una variabile cardinale tra le unità del collettivo. Si ottiene sommando i valori di tutte le osservazioni presenti nel collettivo e dividendo il totale così ottenuto per il numero di osservazioni.

Esempio: Si voglia calcolare l'età media di un nucleo familiare composto da 5 membri.

$$\bar{x} = \frac{\sum X_i}{N}$$

Casi (xi)	Età
Nonno	82
Padre	58
Madre	60
I figlio	20
II figlio	25
Σ	245

$$\overline{X} = \frac{(82+58+60+20+25)}{5} = 245/5 = 49$$

La media ponderata

- Quando i dati sono organizzati in una *distribuzione di frequenza oppure sono raggruppati in classi*, ciascuna frequenza rappresenta il "peso" di ciascun valor X_i;
- In questi casi per individuare la media è necessario ponderare pesare (ponderare) le X_i associate a ciascuna frequenza.

In questi casi si parla di **media ponderata**

Dove:
n = numero dei valori distinti di X_i
f_i = frequenza (peso) di ciascun valore X_i

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i}$$

- **La media ponderata: distribuzione di frequenza**

Esempio 1:

- Si voglia calcolare la media ponderata dei voti riportati da 40 studenti all'esame di Tecniche di ricerca sociale (N=40).

$$\text{Mediaponderata} = \frac{(18*4)+(25*11)+(26*8)+(28*3)+(29*6)+(30*8)}{40} = 26{,}32$$

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i}$$

Voti (X_i)	Frequenze (f_i)
18	4
25	11
26	8
28	3
29	6
30	8
∑fi	40

5. Il midrange

Altra misura di tendenza centrale che possiamo applicare alle variabili cardinali è il midrange.

- **Midrange**= (valore minimo + valore massimo)/2
- *A cosa serve*?
  Questo valore sintetico ci permette di valutare rapidamente il grado di asimmetria di una distribuzione.

Valenza posizionale*: Se la mediana < midrange allora l'asimmetria sarà* ***positiva****.*
Se la mediana > ***midrange*** *allora l'asimmetria sarà negativa.*

I Valori di dispersione

I valori di dispersione rilevano :

- la dispersione dei dati intorno ad un valore di tendenza centrale (mediana), ovvero la dispersione tra le categorie ordinate attraverso determinati valori.
- la dispersione dei dati in generale, ovvero considerando l'intera distribuzione.

La dispersione dei dati intorno ad un valore di tendenza centrale: i quantili

- I quantili sono i valori di una variabile che ne dividono la **distribuzione** di frequenza in sottogruppi di eguale numerosità; essi, al pari della mediana, non derivano da operazioni sui valori ma dalla posizione dei casi; sono quindi *Valori posizionali.*
- I Quantili si definiscono e, di conseguenza, si calcolano in modo analogo alla mediana.

Quantili

Divisione della distribuzione	Quantili	%
in tre parti	terzili	33%
in quattro parti	quartili	25%
in dieci parti	decili	10%

quartili

	% cum			
I categoria	10%	I decile		
II categoria	20%	II decile		
III categoria	30%	III decile	I quartile (25%)	
IV categoria	40%	IV decile		
V categoria	50%	V decile	II quartile (50%)	*Mediana (50%)*
VI categoria	60%	VI decile		
VII categoria	70%	VII decile		
VIII categoria	80%	VIII decile	III quartile (75%)	
IX categoria	90%	IX decile		
X categoria	100%	X decile	IV quartile (100%)	

Differenze interquartili

	Collocazione politica	Anno	
		2000	2004
1	Estrema sinistra	2,2	10,1
2	Sinistra	24,2	26,9
3	Centro sinistra	55,0	38,9
4	Centro	68,0	64,0
5	Centro destra	89,0	74,3
6	Destra	99,0	88,0
7	Estrema destra	100,0	100,0

Anno 2000: Q1 = 3 (centro sinistra) Q3 = 5 (centro sinistra)
Anno 2004: Q1 = 2 (sinistra) Q3 = 6 (destra)

Q2000 = 5 – 3 = 2 BASSA DISPERSIONE
Q2004 = 6 – 2 = 4 ALTA DISPERSIONE

1. Scarto dalla media

- I valori di dispersione rilevano quanto la distribuzione è dispersa dai valori centrali.

Scarto dalla media

$$X_i - \bar{X} \Longrightarrow x_i$$

- dove x_i è una forma contratta per indicare lo scarto dalla media.
- Lo **scarto**, detto anche *scostamento* o *deviation* rappresenta la distanza di un valore dalla media aritmetica della distribuzione.
- Se X*i* > *X* lo scarto avrà segno positivo
  Se X*i* < *X* lo scarto avrà segno negativo.
- **La somma degli scarti dalla media è sempre UGUALE a 0**.

Scarto semplice medio

$$SSM = \frac{\sum |x_i|}{N}$$

$Xi - \overline{X}$

Esempio: Rileviamo la distribuzione dell'età in un nucleo familiare

Media = (245/5) = 49
∑ scarti = 0
SSM = 106/ 5 = 21,2

Componente (casi)	Età	Scarti dalla media	| xi |
Nonno	82	(82-49)= **+33**	**|33|**
Padre	58	(58-49)= **+9**	**|9|**
Madre	60	(60-49)= **+11**	**|11|**
I figlio	20	(20-49)=**-29**	**|29|**
II figlio	25	(25-49)= **-24**	**|24|**
Σ	**245**	**0**	**|106|**

I valori sintetici delle variabili cardinali

- Devianza
- Varianza
- Scarto tipo (DS)
- Coefficiente di variazione

1. Devianza

La devianza è la somma dei *quadrati* degli scarti dalla media.

- Caratteristiche:

$$dev = \sum xi^2 \quad \Rightarrow \quad xi = \left(Xi - \bar{X}\right)$$

- È influenzata dal numero dei casi; all'aumentare di N la dispersione diminuisce;
- Si utilizza per confrontare due distribuzioni con un N simile;
- È una grandezza quadratica;
- È espressa in valori assoluti.

2. La varianza

La varianza è il rapporto tra devianza e numero dei casi.

Caratteristiche:

- Si utilizza per confrontare differenti distribuzioni aventi media uguale;
- Si utilizza per confrontare distribuzioni con un N significativamente diverso;
- È espressa in valori assoluti;
- È una grandezza quadratica;

$$s^2 = \frac{\sum xi}{N}$$

3. Lo scarto tipo

Lo scarto tipo – detto anche scarto quadratico o deviazione standard – è la radice quadrata della varianza.

Caratteristiche:

- Si utilizza per confrontare differenti distribuzioni aventi media uguale;
- Si utilizza per confrontare distribuzioni con un N significativamente diverso;
- È espresso in valori assoluti;
- È una grandezza lineare;

$$s = \sqrt{\frac{\sum xi^2}{N}}$$

Il coefficiente di variazione

Un discorso a parte merita il coefficiente di variazione che si utilizza come valore sintetico per confrontare due distribuzioni con **medie** significativamente differenti.

Caso	Xi	$xi = (Xi - \overline{X})$	xi^2
A	€ 1.000	-€ 1.364	€ 1.860.496
B	€ 1.800	-€ 564	€ 318.096
C	€ 2.000	-€ 364	€ 132.496
D	€ 2.100	-€ 264	€ 69.696
E	€ 2.300	-€ 64	€ 4.096
F	€ 2.350	-€ 14	€ 196
G	€ 5.000	€ 2.634	€ 6.938
∑	€ 16.550	€ 0	€ 2.392.014

Dev.	$\sum xi^2$		=	€ 2.392.014
Var.	$s^2 = \frac{\sum x_i^2}{N}$	= 2.392.014 / 7	=	€ 341.716
Sc. tipo	$s = \sqrt{\frac{\sum xi^2}{N}}$	= $s = \sqrt{341.716}$	=	€ 584,56
Cv	$V = s / \overline{X}$	= 584,56 / 2.364		= 0,247

Confronti tra valori di distribuzione

analisi monovariata della variabile cardinale "reddito" a partire dai singoli casi

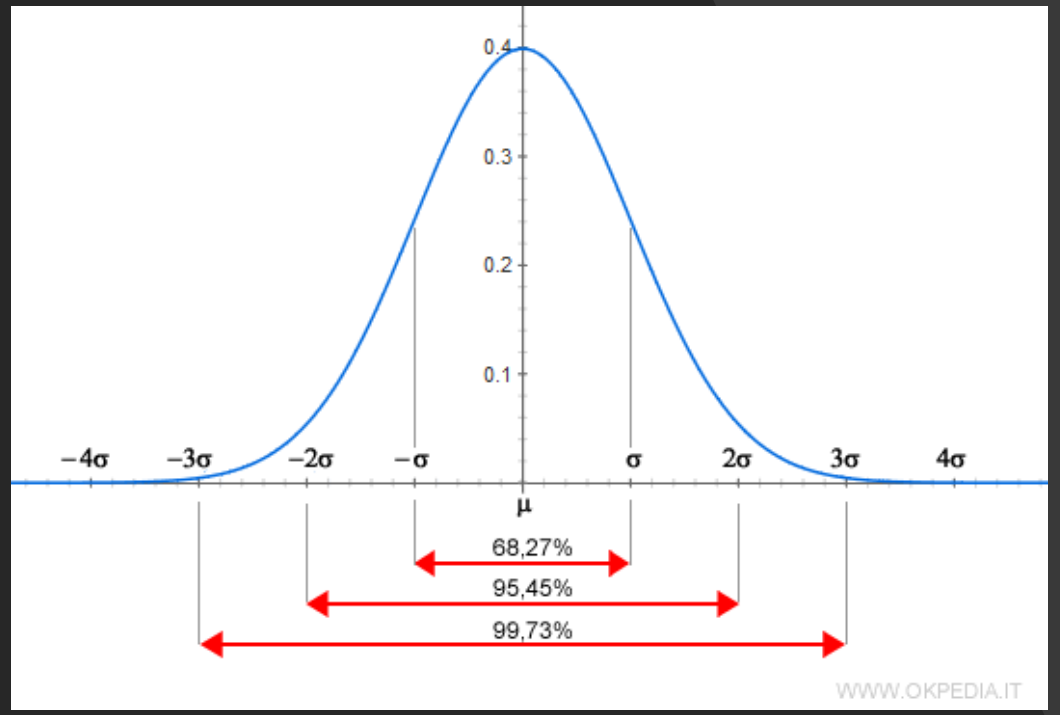
Caso	Xi	$xi - (Xi - \overline{X})$
A	€ 1.000	-€ 1.364
B	€ 1.800	-€ 564
C	€ 2.000	-€ 364
D	€ 2.100	-€ 264
E	€ 2.300	-€ 64
F	€ 2.350	-€ 14
G	€ 5.000	€ 2.636
Σ	€ 16.550	€ 0

N = 7

Moda	*Valore più alto*			=	€ 5.000
Mediana	*N = Dispari (N+1)/2*	=	(7 +1) / 2 = 4° pos.	=	€ 2.100
Midrange	*(val. max + val. min)/2*	=	(5000 + 1000)/2	=	€ 3.000
Range o CV	*val. max - val. min*	=	5000 -1000	=	€ 4.000
Media	$\sum xi/N$	=	16550 / 7	=	€ 2.364
Mediana < Midrange	2000 < 3000		Asimmetria positiva		

Le principali caratteristiche della curva della distribuzione normale

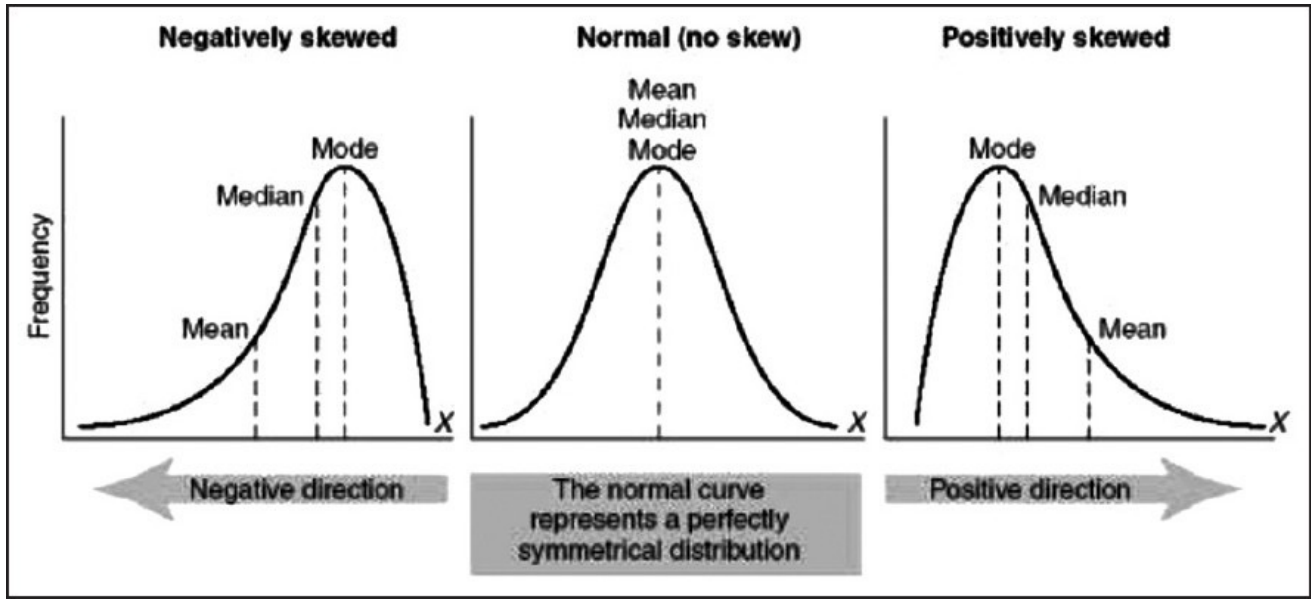
- La frequenza/probabilità più elevata coincide con il valore medio centrale e decresce spostandosi a destra o a sinistra.
- Allontanandosi dalla media la curva si avvicina sempre più all'asse orizzontale delle ascisse, senza mai toccarlo.
- L'area complessiva sotto la curva normale è uguale a uno perché comprende tutte le probabilità dell'evento.



La distribuzione: il coefficiente di asimmetria

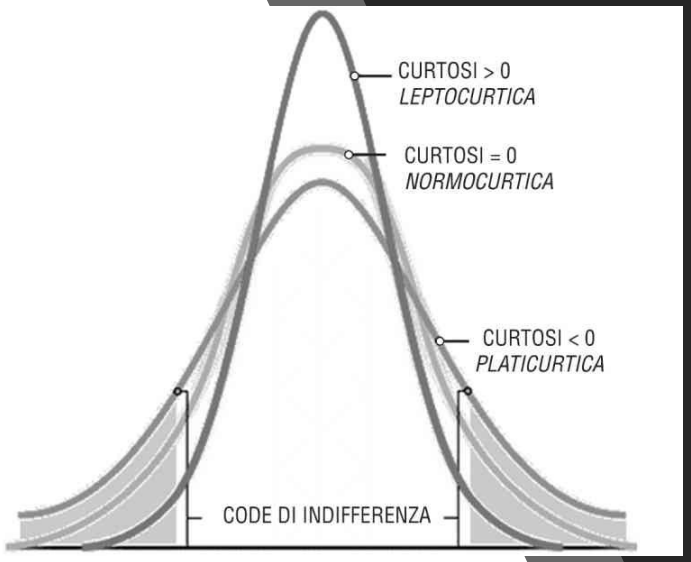
- Distribuzione di Gauss : $\mu = M_e = M0$
- Le distribuzioni reali sono spesso

Asimmetriche:



- skewness positiva (coda della distribuzione verso dx)
- skewness negativa (coda della distribuzione verso sx)

La distribuzione: il coefficiente di curtosi

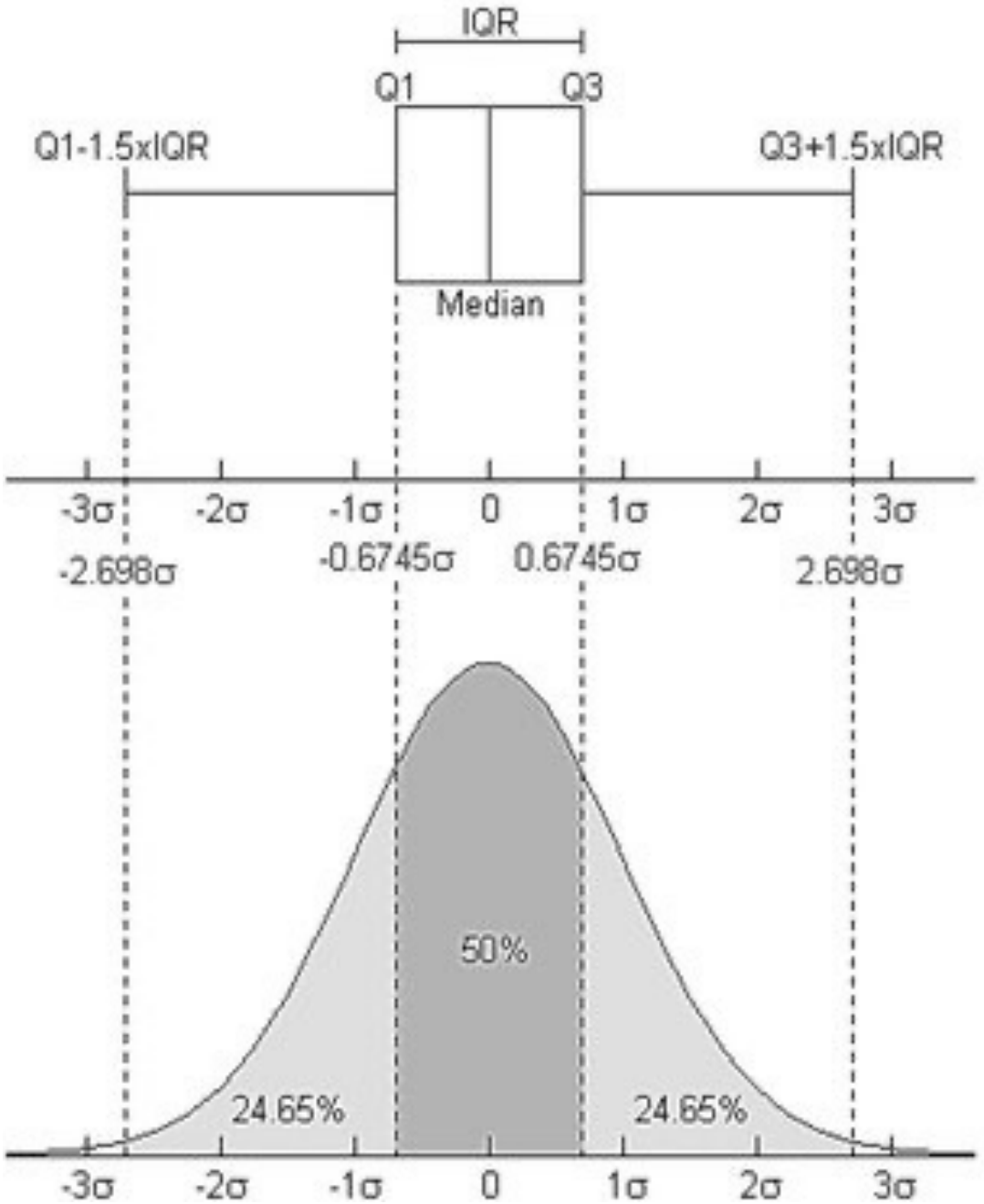


- Questo coefficiente indica il grado di appiattimento di una curva di distribuzione

È Se il coefficiente di curtosi è:

- > 0 la curva si definisce *leptocurtica*, cioè più "appuntita" di una normale.
- < 0 la curva si definisce *platicurtica*, cioè più "piatta" di una normale.
- = 0 la curva si definisce *normocurtica (o mesocurtica)*, cioè "piatta" come una normale.
- Il calcolo del coefficiente di curtosi ha senso solo nelle distribuzioni monomodali.

BoxPlot


IQR
Q1
Q3
Q1-1.5xIQR
Q3+1.5xIQR
Median
-3σ
-2σ
-1σ
0
1σ
2σ
3σ
-2.698σ
-0.6745σ
0.6745σ
2.698σ
50%
24.65%
24.65%

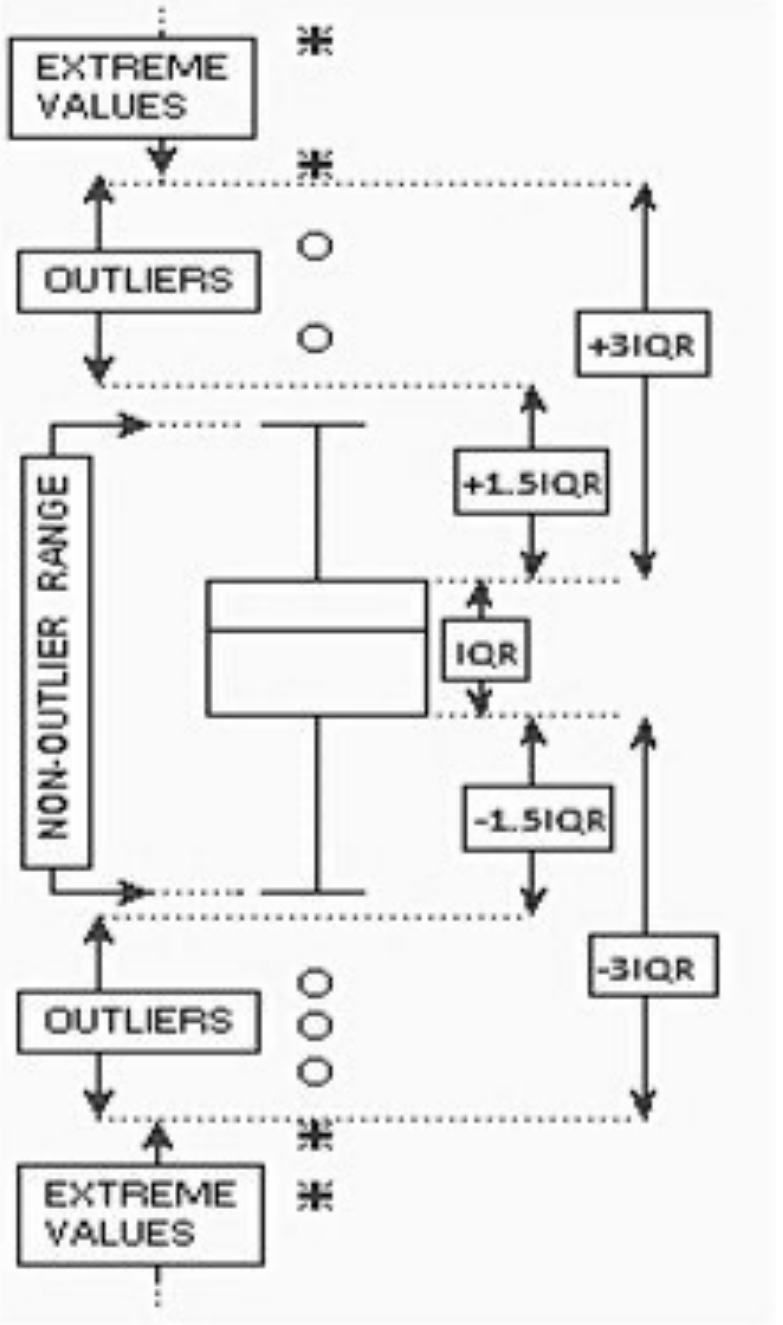

EXTREME VALUES
OUTLIERS
+3IQR
+1.5IQR
NON-OUTLIER RANGE
IQR
-1.5IQR
-3IQR
OUTLIERS
EXTREME VALUES

Tabella 8.18 **Misure di statistica descrittiva riferite alle variabili indice di massa corporea, reddito e salute autopercepita (dati SHP 2015).**

	Imc	Reddito	Salute
N	10.919	14.705	11.181
Media	24,52	66.785,31	7,76
Mediana	24,03	59.440	8
Max	66,79	1.836.000	10
Min	13,52	130	0
Varianza	18,46	1,68E+09	3,11
Asimmetria	1,12	8,37	-1,29
Curtosi	6,58	263,55	5,66

Tab. 19 - Valori caratteristici e forme di rappresentazione relativi ai vari tipi di variabili

	valori caratteristici		*forme*
	posizionali	*sintetici*	*di rappresentazione*
categoriali non ordinate	moda	Sq, Eq, H, M	distrib. di frequenza diagramma a barre diagramma torta grafico a raggi istogramma
categoriali ordinate	moda mediana	d*	distrib. di frequenza istogramma istogramma di composiz. diagramma a bandiera spezzata a gradini
quasi-cardinali	moda mediana **quartili**	media devianza varianza scarto-tipo coeff. di variaz. asimmetria	*distrib. di frequenza* *istogramma* *istogramma di compos.* *diagramma a bandiera* spezzata a gradini poligono di frequenza
cardinali	moda min. & massimo mediana quartili **decili** **percentili**	media valore centrale devianza varianza scarto-tipo coeff. di variaz. asimmetria	*distrib. di frequenza* *diagramma a barre* *istogramma* *istogramma di composiz.* *diagramma a bandiera* spezzata a gradini poligono di frequenza **curva di frequenza**