

Incontri di promozione della cultura statistica
Iniziativa Utenti
Maggio, 2024

Misure di tendenza centrale e variabilità

LORENA VIVIANO

DCCI - CIC - Servizio di coordinamento della promozione della cultura statistica e della diffusione dei dati a livello territoriale

L'uso di simboli generici

In questa presentazione viene utilizzato il seguente simbolismo:

- N : ampiezza complessiva della popolazione
- $X_n \rightarrow$ Variabili oggetto di studio da 1 a n
- $x_n \rightarrow$ Modalità della variabile n
- $n_i \rightarrow$ frequenza osservata in corrispondenza di un generico soggetto i -esimo, ovvero numero di soggetti che possiedono una certa modalità di una variabile X

Misura di tendenza centrale con variabile qualitativa nominale: la moda

Moda: modalità della variabile alla quale corrisponde la frequenza osservata di valore più elevato

- Generalmente indicata come M_o
- Distribuzione unimodale o plurimodale: in presenza di un unico valore modale o più valori modali
- La moda viene individuata osservando la distribuzione dei dati e non richiede alcun calcolo matematico

Misura di tendenza centrale con variabile qualitativa ordinabile: la mediana

Nel caso di variabili qualitative ordinabili la tendenza centrale viene rilevata attraverso la **mediana**, cioè il valore che la variabile assume in corrispondenza del caso che occupa la posizione che ripartisce esattamente in due parti uguali la successione ordinata dei casi stessi.

Generalmente indicata con Me

Se N è dispari: la mediana è la modalità associata alla posizione $k = \frac{N}{2}$

Se N è pari: la mediana è la modalità associata alla posizione $k_1 = \frac{N}{2}$ e alla posizione $k_2 = \frac{N}{2} + 1$

L'individuazione della mediana può avvenire osservando la distribuzione di frequenze cumulate: si tratta della modalità in corrispondenza della quale le frequenze cumulate superano il 50%

Per individuarla i dati vengono ordinati dal più piccolo al più grande

La mediana è una misura robusta, ovvero non risente della presenza degli outlier (valori anomali)

I quartili

- Sempre per variabili ordinali è possibile calcolare i **quartili**, che dividono una distribuzione di frequenza in quattro parti uguali, dopo avere ordinato i dati.
- **Primo quartile:** Q_1 corrisponde al valore al di sotto del quale si trova il 25% dei casi e al di sopra il 75%;
- **Secondo quartile:** Q_2 coincide con la mediana;
- **Terzo quartile:** Q_3 corrisponde al valore al di sotto del quale si trova il 75% dei casi e al di sopra il 25%.

Misure di tendenza centrale con variabile quantitativa: la media

- In presenza di variabili quantitative, si possono calcolare sia la moda sia la mediana
- Tuttavia, la misura più informativa è la media aritmetica che tiene conto di tutte le osservazioni
- La media aritmetica è quel valore che, sostituito al valore del carattere di ogni unità statistica, non muta il valore totale del carattere
- La media permette di sintetizzare in un unico valore l'insieme dei valori da cui è stata calcolata
- Media aritmetica: somma dei valori assunti da tutti i casi diviso per il numero dei casi stessi
- Il suo simbolo è M o μ

$$M = \frac{\sum_{i=1}^N x_i}{N}$$

- È sensibile agli outlier, cioè valori molto alti o molto bassi possono influenzarla significativamente

Media

In presenza di dati in tabella, quando a ogni modalità è associata la frequenza, la media si calcola come:

$$M = \frac{\sum_{i=1}^N x_i n_i}{\sum_{i=1}^N n_i}$$

Ciò significa che il valore assunto da ogni modalità va moltiplicato per il numero di soggetti che la possiedono

Nel caso di classi di valore, si calcola il valore centrale della classe: $\bar{x}_i$

$$M = \frac{\sum_{i=1}^N \bar{x}_i n_i}{\sum_{i=1}^N n_i}$$

La formula della media è analoga a quanto visto prima, dove al posto della modalità va sostituito il valore centrale della classe

Misure di variabilità 1/2

- La maggior parte dei fenomeni che ci circondano sono caratterizzati dalla diversità tra le unità (es. reddito percepito da un individuo)
- Anche se i dati raccolti attraverso le indagini possono dare luogo di popolazioni a valori di sintesi uguali (es. popolazioni con la stessa media) quello che potrebbe variare è il grado di addensamento diverso intorno al valore centrale, ossia una diversa variabilità o eterogeneità dei dati
- Per tale ragione a un valore di sintesi va affiancata una misura di variabilità

Misure di variabilità 2/2

- Una prima misura della variabilità della distribuzione è espressa dalla differenza tra il valore massimo e il valore minimo assunto dalla variabile sui casi. Si tratta di una misura grossolana che risente significativamente della presenza di valori anomali
- Per calcolare la dispersione o variabilità nel caso di variabili ordinali si calcola la differenza interquartile, che misura la distanza che separa Q1 da Q3
- La varianza (σ^2), rappresentata dalla somma degli scarti dei valori assunti dai casi dalla media elevati al quadrato, diviso per il numero dei casi stessi
- La deviazione standard o scarto quadratico medio (σ), calcolato come radice quadrata della varianza.
- Coefficiente di variazione $CV = \sigma / M \rightarrow$ esprime la variabilità relativa, utile per fare confronti tra due distribuzioni di uguale tipologia. CV è un numero puro

Varianza

La varianza (σ^2), rappresentata dalla somma degli scarti dei valori assunti dai casi dalla media elevati al quadrato, diviso per il numero dei casi stessi

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$$

Si tratta, quindi, di una misura sintetica che è funzione delle distanze dei valori della variabile dalla propria media

In presenza di dati in tabella, quando a ogni modalità è associata la frequenza, ogni scarto va moltiplicato per la frequenza

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2 * n_i}{N}$$

Per avere un indicatore espresso nella stessa unità di misura del fenomeno di interesse si utilizza lo scarto quadratico medio, vale a dire la radice quadrata della varianza

Un cenno all'asimmetria

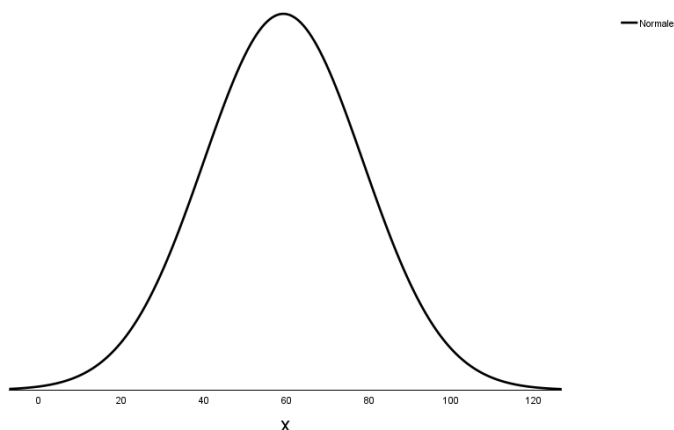
Se si volesse indagare sulla forma assunta dai dati si può introdurre il concetto di asimmetria della distribuzione

Nel caso di una distribuzione simmetrica unimodale moda, media e mediana coincidono

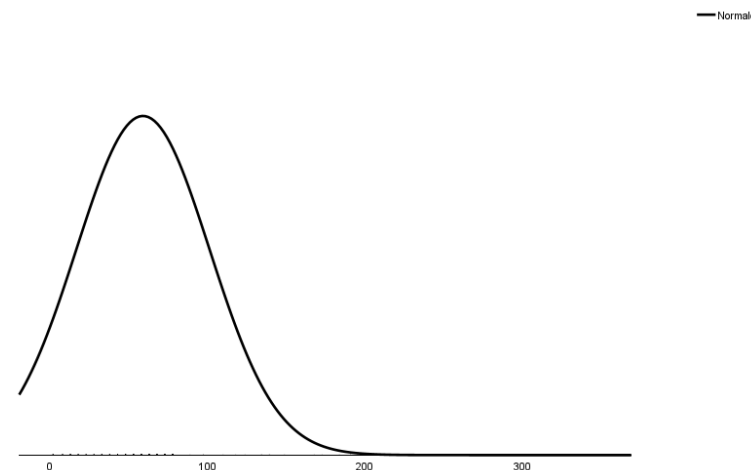
Ciò non avviene per una distribuzione asimmetrica

Nel caso di una distribuzione unimodale asimmetrica a destra (asimmetria positiva), si ha sempre $Mo > Me > \mu$

Il contrario accade per una distribuzione unimodale asimmetrica a sinistra (asimmetria negativa)



Simmetria



Asimmetria positiva

Esempio: variabile qualitativa nominale

Istituzioni non profit per settore di attività. Anno 2021 (valori assoluti) (dati definitivi)

SETTORI DI ATTIVITA'	ISTITUZIONI NON PROFIT
Attività culturali e artistiche	55.326
Attività sportive	121.402
Attività ricreative e di socializzazione	50.708
Istruzione e ricerca	13.533
Sanità	12.244
Assistenza sociale e protezione civile	35.297
Ambiente	6.461
Sviluppo economico e coesione sociale	6.426
Tutela dei diritti e attività politica	6.432
Filantropia e promozione del volontariato	4.058
Cooperazione e solidarietà internazionale	4.479
Religione	17.129
Relazioni sindacali e rappresentanza di interessi	25.027
Altre attività	2.103
Totale	360.625



La misura di tendenza da utilizzare è la **moda**, che in questo esempio è la categoria «attività sportive» associata alla frequenza più alta (circa 121mila istituzioni non profit)

Esempio: variabile qualitativa ordinabile

Soddisfazione per il lavoro			
Misura: Valori in migliaia			
Età: 25-44 anni			
Tempo: 2022			
Soddisfazione per il lavoro	Frequenza assoluta	Frequenza relativa	Frequenza relativa cumulata
per niente	262	2,9	2,9
poco	1.416,00	15,7	18,6
abbastanza	5.615,00	62,4	81,0
molto	1.709,00	19,0	100,0
Totale	9002	100	

Dopo avere ordinato le modalità dal livello di soddisfazione più basso al più alto, la **mediana** corrisponde alla modalità corrispondente alla frequenza relativa cumulata all'interno della quale ricade il 50% delle frequenze, vale a dire «abbastanza».

Anche la moda, in questo esempio, corrisponde alla categoria «abbastanza»

Esempio: variabile continua – Età Media

Popolazione residente al 1° gennaio 2024 per età. Ripartizione: Nord-ovest

Età	ni	Età*ni	(Età-Media)^2*ni
0	100691	0	217688111
1	104258	104258	215808713
2	107513	215026	212870938
3	109233	327699	206664680
4	114603	458412	206969417
5	119067	595335	205030421
6	124918	749508	204863249
7	129191	904337	201536467
8	133041	1064328	197166106
9	136808	1231272	192352279
10	139233	1392330	185459513
11	144386	1588246	181928516
12	146351	1756212	174160844
13	150330	1954290	168674496
14	154255	2159570	162898648
15	155155	2327325	153920186
16	153889	2462224	143124163
17	154360	2624120	134301638
18	152333	2741994	123703732
19	153905	2924195	116362631
20	151792	3035840	106569297

La tabella, tratta da demo.istat.it, va da 0 a 100 anni e oltre (attribuiamo a quest'ultima categoria il valore di 100)

Il calcolo della media e della varianza deve tenere conto del numero degli individui di quell'età (frequenza assoluta o n_i , dove i è l'individuo i -esimo)

Nella ripartizione del Nord- Ovest l'età media è pari a

$$M = \frac{\sum_{i=1}^N x_i n_i}{\sum_{i=1}^N n_i}$$

vale a dire:

$$\frac{739.528.630}{15.904.974} = 46,5 \text{ anni} \rightarrow \text{età media in Italia nel 2024}$$

Esempio: variabile continua – Varianza e scarto quadratico medio

Popolazione residente al 1° gennaio 2024 per età. Ripartizione: Nord-ovest

Età	ni	Età*ni	(Età-Media)^2*ni
0	100691	0	217688111
1	104258	104258	215808713
2	107513	215026	212870938
3	109233	327699	206664680
4	114603	458412	206969417
5	119067	595335	205030421
6	124918	749508	204863249
7	129191	904337	201536467
8	133041	1064328	197166106
9	136808	1231272	192352279
10	139233	1392330	185459513
11	144386	1588246	181928516
12	146351	1756212	174160844
13	150330	1954290	168674496
14	154255	2159570	162898648
15	155155	2327325	153920186
16	153889	2462224	143124163
17	154360	2624120	134301638
18	152333	2741994	123703732
19	153905	2924195	116362631
20	151792	3035840	106569297

Ricordiamo che dobbiamo applicare questo calcolo a tutte le età fino all'età di 100 e oltre (che nella tabella riportata non si visualizza per ragioni di spazio)

Nella ripartizione del Nord- Ovest la varianza è pari la somma di tutti gli scarti delle età dall'età media al quadrato moltiplicati per il numero degli individui. Dopo avere calcolato questi valori per tutte le età, la loro somma va divisa per il numero totale di individui:

$$\frac{8.876.832.785}{15.904.974} = 558$$

Lo scarto quadratico medio è pari a 23,6
I dati oscillano mediamente in più e in meno attorno alla media di circa 24 anni

Esempio: variabile continua

Regione		2014	2015	2016	2017	2018	2019	2020
Piemonte		123,8459	122,0549	126,2502	125,1299	127,8124	137,6144	138,5266
Valle d'Aosta/Vallée d'Aoste		264,7471	247,3153	228,1844	207,769	209,515	219,885	218,1456
Lombardia		120,4194	122,1059	130,4686	127,3041	134,7527	136,6823	144,2112
Trentino-Alto Adige/Südtirol		368,0875	376,8483	384,0966	422,4656	391,0579	398,6912	412,6297
Veneto		106,8756	104,7184	104,7116	98,48224	109,4726	104,4889	109,3822
Friuli-Venezia Giulia		241,3964	252,0979	281,4677	287,1939	277,9215	276,3863	267,6636
Liguria		126,8512	134,3107	134,5758	135,5555	141,9354	147,6992	158,8773
Emilia-Romagna		162,0525	162,4422	160,3666	163,563	173,1183	173,0956	186,4365
Toscana		128,1207	127,3531	125,9214	131,7596	138,0588	138,6908	142,3081
Umbria		86,16962	87,7847	85,78777	89,21337	95,03993	96,4901	101,9844
Marche		110,5475	107,712	99,28951	103,0079	108,4856	113,6192	113,7296
Lazio		146,092	137,9656	130,308	148,3377	153,2217	149,8153	152,4977
Abruzzo		62,75712	65,99687	67,4483	72,69605	73,07134	70,55682	70,7866
Molise		46,06747	58,32577	56,87975	59,11069	70,23769	71,55073	83,01587
Campania		41,82058	43,45209	50,3429	57,12686	56,52258	56,89435	66,48082
Puglia		66,91605	68,6347	68,70907	71,60731	73,50856	77,24669	82,52478
Basilicata		67,25671	62,81662	61,89837	66,20135	59,92586	56,89625	54,22583
Calabria		18,39869	21,25898	23,27513	22,95516	22,09648	23,83246	28,37926
Sicilia		73,2753	74,33181	79,89893	84,04068	83,10064	79,72216	81,6313
Sardegna		238,6841	231,3614	237,0421	232,2288	245,6965	256,2296	283,1383
	M	130,0191	130,4444	131,8461	135,2874	137,2276	139,3044	144,8288
	Dev.standard	85,51556	84,849	86,87586	91,04097	85,90428	88,00846	90,13608
	CV	0,657715	0,650461	0,658918	0,672945	0,625999	0,631771	0,622363

In questo esempio, la spesa per interventi e servizi sociali per abitante nelle regioni italiane, viene presentato in serie storica

Attraverso il foglio di calcolo, sono stati calcolati media, deviazione standard e coefficiente di variazione per ogni anno

Si osserva che la spesa media per abitante è più elevata nel 2020, anno in cui la deviazione standard è di 90 euro: i dati oscillano attorno alla media di circa 90 euro rispetto al valore medio pari 144 euro

Per un confronto sulla base della variabilità relativa, i dati sono più variabili nel 2017, quando il CV raggiunge il valore più elevato (0,67), mentre sono meno variabili nel 2020 (0,62)

Ritorniamo a un esempio

Ritornando all'esempio visto all'interno delle slide introduttive. Avevamo considerato i dati raccolti su tutti i volontari di un'associazione di 200 individui (N=200).

Su di essi (singole unità statistiche) abbiamo raccolto delle informazioni (variabili): sesso, età, titolo di studio, anni dedicati al volontariato

Individuo	Sesso	Età	Titolo di studio	Anni dedicati al volontariato
1	F	31	diploma di qualifica di istruzione professionale	10
2	F	24	licenza media	9
3	F	22	licenza media	8
4	F	23	licenza media	11
5	F	41	licenza media	7
6	F	44	diploma di istruzione secondaria superiore	7
7	F	37	diploma di istruzione secondaria superiore	13
8	F	29	diploma di istruzione secondaria superiore	8
9	F	56	diploma di istruzione secondaria superiore	15
10	F	47	diploma di istruzione secondaria superiore	8
11	F	54	diploma di istruzione secondaria superiore	14
12	F	58	diploma di istruzione secondaria superiore	1
13	F	49	diploma di istruzione secondaria superiore	7
14	M	53	diploma di istruzione secondaria superiore	9
15	F	28	diploma di istruzione secondaria superiore	3
16	F	41	diploma di istruzione secondaria superiore	12
17	F	56	diploma di istruzione secondaria superiore	5
18	F	34	licenza media	14
19	F	52	licenza media	10
20	F	43	licenza media	12
21	F	48	licenza elementare	4
22	F	59	licenza elementare	6
23	F	56	diploma di laurea	5
24	F	21	diploma di istruzione secondaria superiore	8
25	F	33	diploma di laurea	4
26	F	56	diploma terziario extra-universitario	6
27	F	44	diploma terziario extra-universitario	8
28	F	56	diploma di istruzione secondaria superiore	2

Analisi descrittiva delle variabili

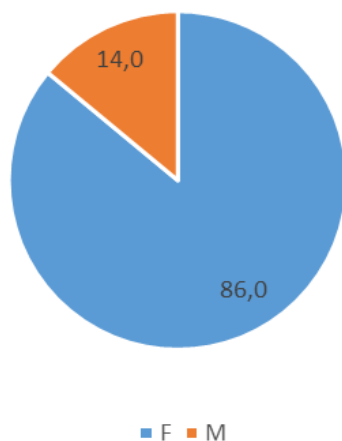
- Abbiamo visto negli interventi precedenti che i dati raccolti attraverso un'indagine possono essere presentati efficacemente attraverso l'uso di tabelle e opportuni grafici
- In questa presentazione sono state introdotte le misure di sintesi, che variano in funzione delle diverse tipologie di variabili
- Nelle slide seguenti verranno analizzate le variabili sesso, età, titolo di studio, anni dedicati al volontariato
- Per completezza, per alcune di queste variabili, vengono riportate le tabelle e i grafici
- Le tabelle sono state calcolate attraverso un foglio elettronico
- Sono riportate tutte le colonne utilizzate per arrivare a determinare media e varianza delle due variabili quantitative, tenendo conto anche delle frequenze associate alle modalità

Analizziamo la variabile «sesso»

Sesso	n	Percentuale
F	172	86,0
M	28	14,0
Totale	200	100

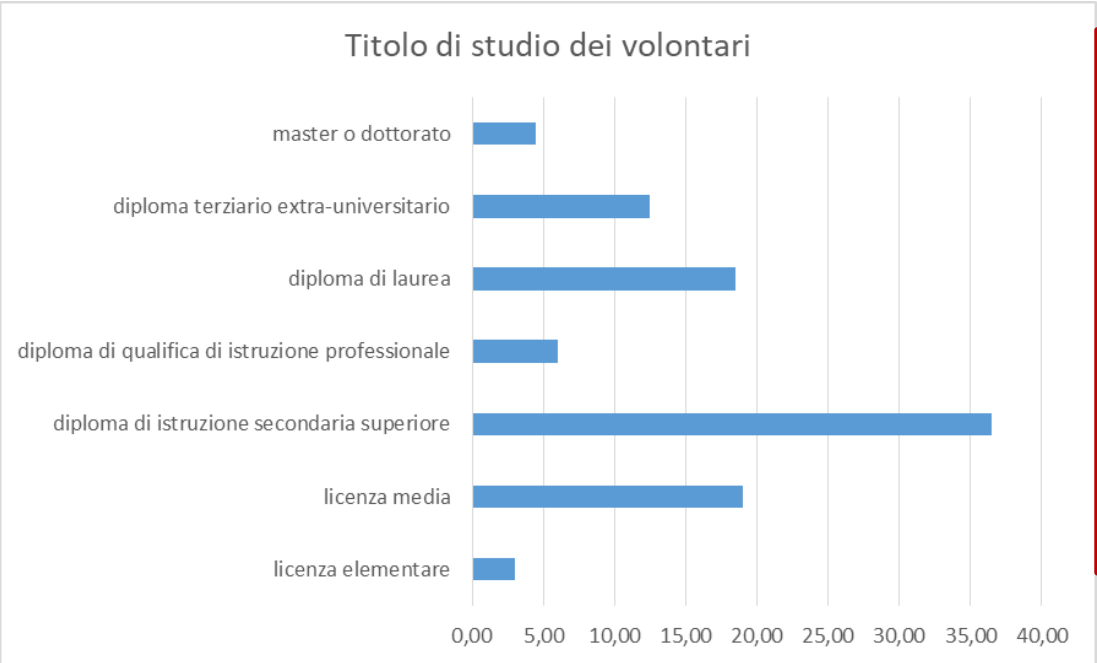
Moda: Sesso femminile

Distribuzione per sesso dei volontari



Analizziamo la variabile «Titolo di studio»

Titolo di studio	n	Percentuale	Freq cumulata
licenza elementare	6	3,0	3,0
licenza media	38	19,0	22,0
diploma di istruzione secondaria superiore	73	36,5	58,5
diploma di qualifica di istruzione professionale	12	6,0	64,5
diploma di laurea	37	18,5	83,0
diploma terziario extra-universitario	25	12,5	95,5
master o dottorato	9	4,5	100,0
Totale	200	100	



Pur trattandosi di modalità individuate da categorie/etichette nominali, si può immaginare di ordinarle facendo riferimento al livello di istruzione dal più basso al più alto.

Si arriva così ad analizzare il titolo di studio come variabile ordinabile

- Moda: diploma di istruzione secondaria superiore
- Primo quartile: diploma di istruzione secondaria superiore
- Secondo quartile-mediana: diploma di istruzione secondaria superiore
- Terzo quartile: diploma di laurea

Analizziamo la variabile «età» -singoli valori

Età	n	Età *n		Età -media	(Età - media)^2	(Età - media)^2*n
24	7	168		-19,4	375,8	2630,4
25	5	125		-18,4	338,0	1690,0
26	8	208		-17,4	302,2	2417,9
27	8	216		-16,4	268,5	2147,7
28	3	84		-15,4	236,7	710,1
29	2	58		-14,4	206,9	413,9
30	8	240		-13,4	179,2	1433,3
31	4	124		-12,4	153,4	613,6
32	5	160		-11,4	129,6	648,1
33	3	99		-10,4	107,8	323,5
34	6	204		-9,4	88,1	528,5
35	7	245		-8,4	70,3	492,2
36	8	288		-7,4	54,5	436,3
37	5	185		-6,4	40,8	203,8
38	3	114		-5,4	29,0	87,0
39	5	195		-4,4	19,2	96,1
40	3	120		-3,4	11,5	34,4

Media $M=43,4$ anni

Varianza= $30665,3$ anni²

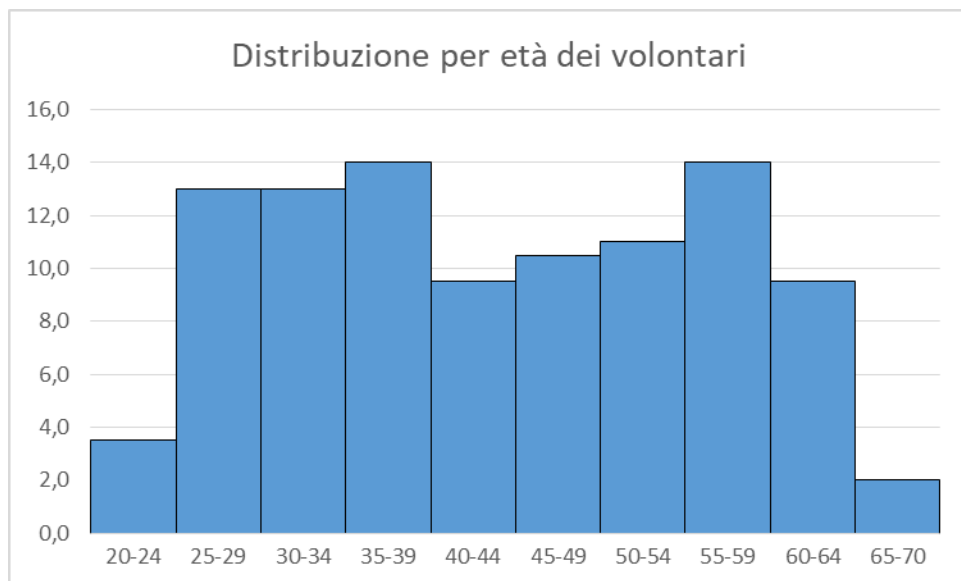
Scarto quadratico medio $12,4$ anni

Si possono inoltre individuare la moda, i quartili e la mediana

NB I valori delle singole colonne vanno calcolati fino all'ultima età registrata sui dati

Analizziamo la variabile «età» - classi di valore

Età	n	Percentuale	Percentuale cumulata	Valore medio classe	Valore medio classe*n	Valore medio classe - media	Scarti^2	Scarti^2*n
20-24	7	3,5	3,5	22	154	-21,5	461,2	3228,2
25-29	26	13,0	16,5	27	702	-16,5	271,4	7057,1
30-34	26	13,0	29,5	32	832	-11,5	131,7	3423,6
35-39	28	14,0	43,5	37	1036	-6,5	41,9	1173,9
40-44	19	9,5	53,0	42	798	-1,5	2,2	41,3
45-49	21	10,5	63,5	47	987	3,5	12,4	260,9
50-54	22	11,0	74,5	52	1144	8,5	72,7	1598,9
55-59	28	14,0	88,5	57	1596	13,5	182,9	5121,9
60-64	19	9,5	98,0	62	1178	18,5	343,2	6520,3
65-70	4	2,0	100,0	67	268	23,5	553,4	2213,7
Totale	200	100			8695			30639,9



Media $M=43,5$ anni

Varianza= $153,2$ anni²

Scarto quadratico medio= $12,4$ anni

Classi modali: 35-39 anni e 55-59 anni

Mediana: 35-39 anni

Primo quartile: 30-34 anni

Terzo quartile: 55-59 anni

Analizziamo la variabile «anni dedicati al volontariato»

Anni	n	Percentuale	Percentuale cumulata	Anni*n	(Anni- media)	(Anni-media) ²	Scarti ² *n
1	11	5,5	5,5	11	-6,9	48,1	529,0
2	16	8	13,5	32	-5,9	35,2	563,6
3	13	6,5	20,0	39	-4,9	24,4	316,6
4	11	5,5	25,5	44	-3,9	15,5	170,3
5	12	6	31,5	60	-2,9	8,6	103,4
6	17	8,5	40,0	102	-1,9	3,7	63,7
7	17	8,5	48,5	119	-0,9	0,9	14,9
8	12	6	54,5	96	0,1	0,0	0,1
9	18	9	63,5	162	1,1	1,1	20,4
10	13	6,5	70,0	130	2,1	4,3	55,4
11	10	5	75,0	110	3,1	9,4	93,9
12	11	5,5	80,5	132	4,1	16,5	181,8
13	12	6	86,5	156	5,1	25,7	307,9
14	11	5,5	92,0	154	6,1	36,8	404,6
15	16	8	100,0	240	7,1	49,9	798,6
Totale	200	100	200,0	1587			3624,2

Media $M=7,9$ anni

Varianza= $18,1$ anni²

Scarto quadratico medio= $4,3$ anni

Moda: 9 anni

Mediana: 8 anni

Primo quartile: 4 anni

Terzo quartile: 11 anni

Scarto interquartile: 7 anni

Un possibile commento ai dati

L'associazione oggetto di studio è costituita da 200 volontari.

Predominano le donne (86%).

Il titolo di studio più ricorrente e mediano è il diploma di scuola media superiore. Solamente il 3% possiede un titolo di licenza elementare e il 4,5% un titolo superiore alla laurea (master o dottorato).

L'età media dei volontari è di circa 43 anni, con una dispersione di circa 12 anni.

Mediamente i soggetti prestano l'attività di volontariato da 8 anni circa con un'oscillazione di circa 4 anni attorno a tale valore.

Per un'analisi più approfondita, si potrebbe:

- ampliare lo studio a più associazioni per ottenere una maggiore rappresentatività della popolazione di riferimento;
- incrociare i dati su titolo di studio ed età con altre variabili, come il genere, il tipo di volontariato e altri fattori socio-demografici;
- indagare le motivazioni che spingono persone di diverse fasce d'età e con differenti livelli di istruzione a fare volontariato.

Va sottolineato che i dati presentati in questo report sono basati su una sola popolazione e potrebbero non essere rappresentativi dell'intera popolazione di volontari.

Sitografia

- Corbetta P. e altri, *Statistica per la ricerca sociale*. Il Mulino, 2001
- Statistica per esempi: <https://www.istat.it/it/files/2012/03/statisticaxesempi2.pdf>
- Sito Istat- Pacchetti didattici: <https://www.istat.it/it/informazioni-e-servizi/per-studenti-e-docenti/pacchetti-didattici>
- Censimento delle Istituzioni non profit, dati 2021 definitiva: <https://www.istat.it/it/archivio/296339> (Tavola 5)
- Popolazione residente al 1° gennaio 2024 per età - Demografia in cifre: <https://demo.istat.it/app/?i=POS&l=it>
- Spese sociali – A misura di Comune: <https://www.istat.it/it/archivio/279229> (Tavola 9a)

grazie

utenti.territorio@istat.it