

Statistica descrittiva univariata

La rilevazione è di tipo **esaustivo** e prendiamo in esame **un solo carattere** alla volta, ovvero una sola colonna della matrice dei dati

MATRICE DEI DATI

unità	caratteri					
	A	B	...	X	Y	Z
1	a ₁	b ₁		x ₁	y ₁	z ₁
2	a ₂	b ₂		x ₂	y ₂	z ₂
...
j	a _j	b _j		x _j	y _j	z _j
...
N	a _N	b _N		x _N	y _N	z _N

RACCOLTA DEI DATI

Il risultato della rilevazione di un carattere X su N unità statistiche si presenta come una successione

$$(x_1, x_2, \dots, x_j, \dots, x_N) \rightarrow (x_j)_{j=1,2,\dots,N}$$

TRATTAMENTO DEI DATI

Rappresentazione

Tabellare (distribuzioni)

Grafica

Sintetica (indici)

Analitica

DISTRIBUZIONE DI FREQUENZA

È il risultato di un processo di classificazione

- Individuazione delle modalità distinte $x_1, \dots, x_i, \dots, x_k$ (k =numero di modalità diverse $<$ N =numerosità popolazione)
- Attribuzione delle unità statistiche alle classi
- Conteggio delle unità che risultano assegnate ad ogni classe (frequenza della classe)

<i>modalità del carattere X</i>	<i>frequenza</i>
x_1	n_1
x_2	n_2
...	...
x_i	n_i
...	...
x_k	n_k
	N

Notazione $\rightarrow (x_i, n_i)_{i=1,2,\dots,k}$ con $\sum_{i=1}^k n_i = N$

La distribuzione di frequenza è un modo di organizzare i dati per facilitarne la lettura

Occupati in Provincia di Firenze per sezioni di attività economica

Fonte: Istat, Censimento della popolazione 2011

La successione delle modalità osservate contiene oltre 420mila etichette (o codici) e non è leggibile

Sezioni attività economica	occupati
agricoltura, silvicoltura e pesca	12.122
industria	111.224
commercio, alberghi e ristoranti	85.952
trasporto, magazzinaggio, ecc	26.559
altri servizi	185.196
totale	421.053

Carattere misurato su scala nominale

Popolazione residente con oltre 10 anni di età per titolo di studio – Italia

Fonte: Censimento 2011

Grado di istruzione	residenti >10 anni
analfabeta	590.074
alfabeta privo di titolo di studio	1.800.972
licenza di scuola elementare	11.021.155
licenza di scuola media inferiore o simile	16.706.880
diploma di scuola secondaria super. o simile	17.157.345
titoli universitari	6.064.549
totale	53.340.975

scala ordinale gerarchia naturale

Distribuzione delle famiglie italiane per numero di componenti 2015 – Fonte: Istat

Tutte le modalità diverse sono indicate (salvo la classe 6 e+)

numero componenti	numero famiglie (migliaia)
1	7.910
2	6.833
3	5.116
4	4.051
5	1.088
6 e+	303
totale	25.301

scala di rapporto

Reddito complessivo dichiarato nel 2015 (anno imposta 2014)

Fonte: Ministero dell'economia e delle finanze

*Non è possibile
elencare tutte le
modalità
osservate
È necessario
raggruppare in
classi*

Classi di reddito €	Frequenza
minore di zero	196.196
da 0 a 1000	2.234.286
da 1000 a 1500	612.343
da 1500 a 2000	535.423
da 2000 a 2500	478.121
da 2500 a 3000	448.283
da 3000 a 3500	401.256
da 3500 a 4000	391.474
da 4000 a 5000	769.734
da 5000 a 6000	749.011
da 6000 a 7500	2.816.050
da 7500 a 10000	2.821.276
da 10000 a 12000	2.421.347
da 12000 a 15000	3.341.557
da 15000 a 20000	6.104.263
da 20000 a 26000	6.224.701
da 26000 a 29000	2.215.340
da 29000 a 35000	2.864.038
da 35000 a 40000	1.265.383
da 40000 a 50000	1.280.775
da 50000 a 55000	363.838
da 55000 a 60000	275.036
da 60000 a 70000	393.471
da 70000 a 75000	150.781
da 75000 a 80000	126.131
da 80000 a 90000	186.460
da 90000 a 100000	127.398
da 100000 a 120000	155.153
da 120000 a 150000	113.381
da 150000 a 200000	77.244
da 200000 a 300000	46.696
oltre 300000	31.772
TOTALE	40.218.218

scala di rapporto

DISTRIBUZIONI DERIVATE

Distribuzioni di frequenza relativa (o percentuale)

*Distribuzione delle famiglie italiane per numero di componenti
2015 – Fonte: Istat*

numero componenti	numero famiglie (migliaia)	Valori percentuali
1	7.910	31,3
2	6.833	27,0
3	5.116	20,2
4	4.051	16,0
5	1.088	4,3
6e+	303	1,2
totale	25.301	100,0

Notazione $\rightarrow \left(x_i, \frac{n_i}{N} \cdot 100 \right)_{1=1,2,\dots,k}$

Utile nel confrontare due distribuzioni che hanno stesso argomento ma sono riferite a popolazioni di numerosità diversa.

Morti per causa nel 2013 in Lombardia e Sardegna. Fonte: Istat

cause di morte	valori assoluti		valori percentuali	
	Lombardia	Sardegna	Lombardia	Sardegna
malattie del sistema circolatorio	30.817	4.863	34,0	32,2
tumori	30.409	4.780	33,5	31,6
malattie del sistema respiratorio	6.516	967	7,2	6,4
traumatismi e avvelenamento	3.258	764	3,6	5,1
malattie del sistema nervoso	4.209	673	4,6	4,5
malattie dell'apparato digerente	3.252	651	3,6	4,3
malattie endocrine, nutrizionali e metaboliche	2.941	622	3,2	4,1
disturbi psichici e comportamentali	2.623	527	2,9	3,5
cause anomale e mal definite	1.290	373	1,4	2,5
alcune malattie infettive e parassitarie	2.265	323	2,5	2,1
malattie dell'apparato genitourinario	1.676	293	1,8	1,9
altre cause	1.392	283	1,5	1,9
totale	90.648	15.119	100,0	100,0

Occupati per sezioni di attività economica – Fonte: Censimento 2011

Sezioni di attività economica	valori assoluti		valori percentuali	
	Firenze	Grosseto	Firenze	Grosseto
agricoltura, silvicoltura e pesca	12.122	9.378	2,9	10,6
industria	111.224	18.239	26,4	20,6
commercio, alberghi e ristoranti	85.952	21.345	20,4	24,1
trasporto, magazzinaggio, ecc	26.559	4.351	6,3	4,9
altri servizi	185.196	35.433	44,0	39,9
totale	421.053	88.746	100,0	100,0

Distribuzioni di frequenza cumulata (assoluta o relativa)

Se il carattere è misurato su scala ordinale o di intervallo/rapporto (non nominale) è possibile indicare, accanto a ciascuna modalità, il numero di unità che presentano una modalità uguale o “gerarchicamente inferiore” a quella data.

Notazioni

$$\left(x_i, \sum_{s=1}^i n_s \right)_{i=1,2,\dots,k} \quad \text{oppure} \quad \left(x_i, \sum_{s=1}^i \frac{n_s}{N} \cdot 100 \right)_{i=1,2,\dots,k}$$

**Popolazione residente con oltre 10 anni di età per titolo di studio – Italia
Fonte: Censimento 2011**

Grado di istruzione	frequenza	%	freq cumulata	freq % cum
analfabeta	590.074	1,1	590.074	1,1
alfabeta privo di titolo di studio	1.800.972	3,4	2.391.046	4,5
licenza di scuola elementare	11.021.155	20,7	13.412.201	25,1
licenza di scuola media inferiore o simile	16.706.880	31,3	30.119.081	56,5
diploma di scuola secondaria super. o simile	17.157.345	32,2	47.276.426	88,6
titoli universitari	6.064.549	11,4	53.340.975	100,0
totale	53.340.975	100,0		

30119081 residenti in età superiore a 10 anni (ovvero il 56,5% del totale) hanno un titolo di studio che non supera la licenza media inferiore

Distribuzioni di quantità (intensità)

Anziché contare le unità che ricadono in ciascuna classe, si sommano i (*si fanno elaborazioni sui*) valori relativi ad un carattere rilevato su quelle unità

Reddito complessivo dichiarato nel 2015 (anno imposta 2014)

Classi di reddito €	Frequenza	valore centrale classe	stima reddito dichiarato	Ammontare (migliaia €)	ammont. vero - stimato	Media (migliaia €)
minore di zero	196.196			-1.989.555		-10,14
da 0 a 1000	2.234.286	500	1.117.143	956.394	-160.749	0,43
da 1000 a 1500	612.343	1250	765.429	760.480	-4.949	1,24
da 1500 a 2000	535.423	1750	936.990	933.470	-3.520	1,74
da 2000 a 2500	478.121	2250	1.075.772	1.077.166	1.394	2,25
da 2500 a 3000	448.283	2750	1.232.778	1.231.765	-1.013	2,75
da 3000 a 3500	401.256	3250	1.304.082	1.303.330	-752	3,25
da 3500 a 4000	391.474	3750	1.468.028	1.469.114	1.087	3,75
da 4000 a 5000	769.734	4500	3.463.803	3.472.581	8.778	4,51
da 5000 a 6000	749.011	5500	4.119.561	4.126.665	7.105	5,51
da 6000 a 7500	2.816.050	6750	19.008.338	18.894.841	-113.497	6,71
da 7500 a 10000	2.821.276	8750	24.686.165	24.661.851	-24.314	8,74
da 10000 a 12000	2.421.347	11000	26.634.817	26.590.132	-44.685	10,98
da 12000 a 15000	3.341.557	13500	45.111.020	45.180.558	69.539	13,52
da 15000 a 20000	6.104.263	17500	106.824.603	107.445.327	620.725	17,60
da 20000 a 26000	6.224.701	23000	143.168.123	141.641.787	-1.526.336	22,75
da 26000 a 29000	2.215.340	27500	60.921.850	60.780.191	-141.659	27,44
da 29000 a 35000	2.864.038	32000	91.649.216	90.707.170	-942.046	31,67
da 35000 a 40000	1.265.383	37500	47.451.863	47.148.534	-303.329	37,26
da 40000 a 50000	1.280.775	45000	57.634.875	56.804.654	-830.221	44,35
da 50000 a 55000	363.838	52500	19.101.495	19.056.998	-44.497	52,38
da 55000 a 60000	275.036	57500	15.814.570	15.785.413	-29.157	57,39
da 60000 a 70000	393.471	65000	25.575.615	25.446.995	-128.620	64,67
da 70000 a 75000	150.781	72500	10.931.623	10.922.011	-9.612	72,44
da 75000 a 80000	126.131	77500	9.775.153	9.764.434	-10.719	77,42
da 80000 a 90000	186.460	85000	15.849.100	15.782.982	-66.118	84,65
da 90000 a 100000	127.398	95000	12.102.810	12.067.189	-35.621	94,72
da 100000 a 120000	155.153	110000	17.066.830	16.897.454	-169.376	108,91
da 120000 a 150000	113.381	135000	15.306.435	15.092.973	-213.462	133,12
da 150000 a 200000	77.244	175000	13.517.700	13.195.052	-322.648	170,82
da 200000 a 300000	46.696	250000	11.674.000	11.161.576	-512.424	239,03
oltre 300000	31.772			18.893.998		594,67
TOTALE	40.218.218			817.263.529		20,32

Un esempio rilevante distribuzione cumulata del reddito

Reddito complessivo dichiarato nel 2015 (anno imposta 2014)

Classi di reddito €	Frequenza	Freq %	Freq % cumul	Ammontare (migliaia €)	Ammont %	ammont % cumul
da 0 a 1000	2.234.286	5,6	5,6	956.394	0,1	0,1
da 1000 a 1500	612.343	1,5	7,1	760.480	0,1	0,2
da 1500 a 2000	535.423	1,4	8,5	933.470	0,1	0,3
da 2000 a 2500	478.121	1,1	9,6	1.077.166	0,1	0,4
da 2500 a 3000	448.283	1,2	10,8	1.231.765	0,2	0,6
da 3000 a 3500	401.256	1,0	11,8	1.303.330	0,2	0,8
da 3500 a 4000	391.474	1,0	12,8	1.469.114	0,2	1,0
da 4000 a 5000	769.734	1,9	14,7	3.472.581	0,4	1,4
da 5000 a 6000	749.011	1,9	16,6	4.126.665	0,5	1,9
da 6000 a 7500	2.816.050	7,0	23,6	18.894.841	2,3	4,2
da 7500 a 10000	2.821.276	7,0	30,6	24.661.851	3,0	7,2
da 10000 a 12000	2.421.347	6,1	36,7	26.590.132	3,2	10,4
da 12000 a 15000	3.341.557	8,3	45,0	45.180.558	5,5	15,9
da 15000 a 20000	6.104.263	15,3	60,3	107.445.327	13,1	29,0
da 20000 a 26000	6.224.701	15,6	75,9	141.641.787	17,3	46,3
da 26000 a 29000	2.215.340	5,5	81,4	60.780.191	7,4	53,7
da 29000 a 35000	2.864.038	7,2	88,6	90.707.170	11,1	64,8
da 35000 a 40000	1.265.383	3,2	91,8	47.148.534	5,8	70,6
da 40000 a 50000	1.280.775	3,1	94,9	56.804.654	6,9	77,5
da 50000 a 55000	363.838	0,9	95,8	19.056.998	2,3	79,8
da 55000 a 60000	275.036	0,7	96,5	15.785.413	1,9	81,7
da 60000 a 70000	393.471	1,0	97,5	25.446.995	3,1	84,8
da 70000 a 75000	150.781	0,4	97,9	10.922.011	1,3	86,1
da 75000 a 80000	126.131	0,3	98,2	9.764.434	1,2	87,3
da 80000 a 90000	186.460	0,5	98,7	15.782.982	1,9	89,2
da 90000 a 100000	127.398	0,2	98,9	12.067.189	1,5	90,7
da 100000 a 120000	155.153	0,4	99,3	16.897.454	2,1	92,8
da 120000 a 150000	113.381	0,3	99,6	15.092.973	1,8	94,6
da 150000 a 200000	77.244	0,2	99,8	13.195.052	1,6	96,2
da 200000 a 300000	46.696	0,1	99,9	11.161.576	1,4	97,6
oltre 300000	31.772	0,1	100,0	18.893.998	2,4	100,0
TOTALE	40.022.022	100,0		819.253.085	100,0	

Il 30,6% degli individui dichiara meno di 10mila euro e detiene il 7,2% del reddito complessivamente dichiarato

Il 2,5% (100-97,5) degli individui dichiara oltre 70mila euro e detiene il 15,2% (100-84,8) del reddito complessivamente dichiarato

Riepilogo delle notazioni

$(X_j)_{j=1,2,\dots,N}$ successione

$(x_i, n_i)_{i=1,2,\dots,k}$ distrib di frequenza *modalità singole*

$(x_i - x_{i+1}, n_i)_{i=1,2,\dots,k}$ distrib di frequenza *modalità in classi*

$\left(x_i, \frac{n_i}{N} \cdot 100 \right)_{i=1,2,\dots,k}$	}	distrib di frequenza percentuale
$\left(x_i - x_{i+1}, \frac{n_i}{N} \cdot 100 \right)_{i=1,2,\dots,k}$		

$\left(x_i, \sum_{s=1}^i n_s \right)_{i=1,2,\dots,k}$	}	distrib di frequenza cumulata <i>assoluta o percentuale</i>
$\left(x_i, \sum_{s=1}^i \frac{n_s}{N} \cdot 100 \right)_{i=1,2,\dots,k}$		

$(x_i, X_i^{tot})_{i=1,2,\dots,k}$	}	distrib di intensità
$(x_i, Y_i^*)_{i=1,2,\dots,k}$		

Con

X_i^{tot} = somma del carattere X per tutte le unità che presentano la modalità i-esima

Y_i^* = un qualunque dato statistico (somma, media, ...) riferibile alle unità che presentano la modalità i-esima del carattere X.

ADDIZIONE e SOMMA: UN BREVE RICHIAMO

Si abbia una successione di modalità quantitative (numeri) osservati su N unità statistiche

$$(x_1, x_2, \dots, x_j, \dots, x_N) \rightarrow (x_j)_{j=1,2,\dots,N}$$

dove x_j è la modalità osservata sulla j-esima unità statistica

Il risultato dell'addizione ($x_1 + x_2 + \dots + x_j + \dots + x_N$), ovvero la somma, si scrive

$$\sum_{j=1}^N x_j$$

dove il simbolo Σ (lettera sigma) sta per somma(toria) e si legge: "somma(toria) per j che va da 1 ad N di x_j "

PROPRIETA' DELL'ADDIZIONE

Associativa
$$\sum_{j=1}^N (x_j + y_j) = \sum_{j=1}^N x_j + \sum_{j=1}^N y_j$$

Commutativa
$$\sum_{j=1}^N x_j + \sum_{j=1}^N y_j = \sum_{j=1}^N y_j + \sum_{j=1}^N x_j$$

Distributiva
$$\sum_{j=1}^N \alpha x_j = \alpha \sum_{j=1}^N x_j$$

Somma di una costante
$$\sum_{j=1}^N \alpha = N\alpha$$

ATTN

$$\sum_{j=1}^N x_j \cdot y_j \neq \left(\sum_{j=1}^N x_j \right) \cdot \left(\sum_{j=1}^N x y_j \right)$$

$$(1 \times 3) + (2 \times 2) + (3 \times 5) = 22 \neq (1+2+3) \times (3+2+5) = 60$$