4
Misure sintetiche di una distribuzione

4.1 Medie

La raccolta, l'organizzazione dei dati in tabelle e la relativa rappresentazione grafica costituiscono solo il primo passo nell'analisi statistica. Una efficace sintesi dei dati è ottenuta dal calcolo di particolari valori che caratterizzano una serie o una distribuzione. Questi valori vengono chiamati "medie" e appartengono a un'ampia classe di misure che usualmente distinguiamo in due categorie:

- *Medie analitiche o algebriche*, dette anche medie fisse perché al variare di uno solo dei valori del collettivo esse cambiano inesorabilmente valore.
- *Medie lasche o di posizione*, così chiamate perché possono rimanere invariate per piccoli cambiamenti nella distribuzione essendo individuate tenendo conto solo di alcuni valori della serie o della distribuzione.

Pertanto, secondo *Cauchy* una media è qualunque valore reale M interno a un intervallo entro cui variano i valori di x_i tale che:

$$x_i < M < x_n$$

Secondo *Chisini* una media è, inoltre, quel valore che rispetto a una funzione sintetica delle osservazioni ne lascia invariato il valore:

$$f(x_1, x_2, ...x_N) = f(M, M...M)$$

4.2 Medie analitiche

Analizziamo ora le principali medie analitiche, prendendo in considerazione:

- media aritmetica;
- media geometrica;
- media armonica.

4.2.1 Media aritmetica

La misura più nota tra le medie analitiche e più comunemente utilizzata è la *media aritmetica*. Utilizzando la *somma* come criterio di trasferibilità della variabile definiamo la media aritmetica (indicata con μ) come quel valore che sostituito a tutti i termini della distribuzione ne lascia invariata la funzione somma. Ne deriva che:

$$\mu = \frac{x_1 + x_2 + \ldots\ldots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N} \qquad (4.1)$$

ovvero il calcolo della media aritmetica nel caso di *distribuzioni semplici* si ottiene sommando tutti i dati osservati x_i e dividendoli per il numero totale di osservazioni.

Nel caso di *distribuzioni di frequenza* avremo la media aritmetica *ponderata* data dalla somma dei prodotti tra le singole modalità e le rispettive frequenze, diviso per le frequenze totali:

$$\mu = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + \ldots\ldots + x_s \cdot n_s +}{\sum_{i=1}^{s} n_i} = \frac{\sum_{i=1}^{s} x_i \cdot n_i}{N} \qquad (4.2)$$

La media tiene conto di tutti i dati della distribuzione e per questo è particolarmente influenzata dalla presenza di valori estremi anomali. Il valore della media aritmetica può non coincidere con nessuno dei valori delle osservazioni presenti: se per esempio si deve calcolare la media aritmetica dell'età di un gruppo di individui, dalla divisione si ricaverà molto probabilmente un numero decimale.

Nel caso in cui i dati siano raggruppati in classi la soluzione più comune è quella di sostituire alle classi di modalità (x_{i-1}, x_i) il corrispondente "valore centrale" calcolato attraverso la semisomma dei valori estremi. Tale ipotesi è semplice e immediata ma non è sempre corretta in quanto attribuisce un valore ipotetico a un intervallo di valori.

La statistica distingue con l'uso di simboli differenti, la media aritmetica calcolata su un campione da quella calcolata su di una popolazione. È norma comune in tutta la statistica utilizzare le lettere greche quando ci si riferisce alla popolazione nel suo complesso e le lettere latine se l'analisi riguarda un campione. Si userà quindi:

- media aritmetica di un *campione* $\bar{x}$
- media aritmetica della *popolazione* μ

Nel caso della media aritmetica non c'è differenza di calcolo tra la media del campione rispetto a quella della popolazione, si utilizzano soltanto dei simboli differenti.

Chiariamo con alcuni esempi queste prime definizioni.

Esempio 4.1

A 14 studenti di un corso universitario si chiede l'età e la si annota in forma ordinata:

18, 19, 19, 19, 20, 20, 21, 2
metica sommando le singo
di studenti, come indicato

$$\mu = \frac{(18 + 19 + 19 + 19 +}{}$$

L'età media degli studenti
rielaborati e classificati in

Tabella 4.1 Calcolo della

Età
18
19
20
21
22
23
Totale

In tal caso applicando la fo

$$\mu = \frac{18 \times 1 + 19 \times}{}$$

Anche in tal caso l'età med

Esempio 4.2

Consideriamo una distribuz
di clienti in un dato giorno.

Tabella 4.2 Calcolo della m

Classe di età	F as
15 ⊢ 25	
25 ⊢ 35	
35 ⊢ 45	
45 ⊢ 55	
55 ⊢ 65	
Totale:	

18, 19, 19, 19, 20, 20, 21, 21, 21, 21, 21, 22, 23, 23 e si passa la calcolo della media aritmetica sommando le singole modalità e dividendo il totale per il numero complessivo di studenti, come indicato nella formula (4.1):

$$\mu = \frac{(18+19+19+19+20+20+21+21+21+21+21+22+23+23)}{14} = \frac{288}{14} = 20,57$$

L'età media degli studenti considerati è pari a 20,57 anni. Gli stessi dati possono essere rielaborati e classificati in una distribuzione di frequenza come mostrato in Tabella 4.1.

Tabella 4.1 Calcolo della media aritmetica.

Età	Frequenze assolute	Età x Frequenze
18	1	18
19	3	57
20	2	40
21	5	105
22	1	22
23	2	46
Totale	*14*	*288*

In tal caso applicando la formula (4.2) si avrà:

$$\mu = \frac{18\times1+19\times3+20\times2+21\times5+22\times1+23\times2}{14} = \frac{288}{14} = 20,57$$

Anche in tal caso l'età media degli studenti è pari a 20,57 anni. □

Esempio 4.2

Consideriamo una distribuzione di dati raggruppati in classi rappresentativa del numero di clienti in un dato giorno. Calcoliamo il valore centrale delle classi:

Tabella 4.2 Calcolo della media aritmetica per dati suddivisi in classi.

Classe di età	Frequenza assoluta (n_i)	Valore centrale della classe (m_i)	Età × Frequenze ($n_i \times m_i$)
15 ⊢ 25	25	20	500
25 ⊢ 35	32	30	960
35 ⊢ 45	18	40	720
45 ⊢ 55	27	50	1350
55 ⊢ 65	31	60	1860
Totale:	*133*		*5390*

Avremo quindi:

$$\mu = \frac{20 \times 25 + 30 \times 32 + 40 \times 18 + 50 \times 27 + 60 \times 31}{133} = \frac{5.390}{133} = 40,53$$

L'età media è quindi pari a 40,53 anni.

Esempio 4.3

L'imprenditore di una piccola azienda con 22 dipendenti vuole conoscere la retribuzione media annua lorda erogata dalla sua impresa. I dati vengono raccolti e tabulati sul foglio Excel "Esempio 4.3". Calcolare la media aritmetica delle retribuzioni con l'uso della funzione `MEDIA()`.

Proposta di soluzione in Excel

Per la media aritmetica, così come per altri tipi di medie, Excel fornisce numerose funzioni come si vedrà in dettaglio nel paragrafo 4.8. La funzione da utilizzare in questo esercizio è `=MEDIA(<intervallo>)` dove al posto del parametro `<intervallo>` andranno inserite le celle che contengono i dati. Osservando la Figura 4.1 l'intervallo sarà `B2:B23`. Dopo aver tolto o ridotto a una sola cifra i decimali con il pulsante il risultato apparirà come in Figura 4.1. □

	A	B	C	D	E	F	G
1	Dipendente	Retribuzione annua lorda		Calcolo della media aritmetica			
2	1	€ 21.982		Media:	€ 25.255	< =MEDIA(B2:B23)	
3	2	€ 16.928					
4	3	€ 23.054					
5	4	€ 18.299					
6	5	€ 25.629					
7	6	€ 25.978					
8	7	€ 22.867					
9	8	€ 24.931					
10	9	€ 22.084					
11	10	€ 17.117					
12	11	€ 16.961					
13	12	€ 21.170					
14	13	€ 22.122					
15	14	€ 18.799					
16	15	€ 17.632					
17	16	€ 18.133					
18	17	€ 22.517					
19	18	€ 34.595					
20	19	€ 37.555					
21	20	€ 31.614					
22	21	€ 41.152					
23	22	€ 54.497					
24							

Figura 4.1 Esempio di applicazione della funzione MEDIA().

Esempio 4.4

Un *call center* che impiega più di 400 giovani ai propri terminali, per verificare la qualità del servizio ha raccolto i tempi di attesa di 5000 utenti nella fascia oraria 9.00-12.00 in alcuni giorni feriali non consecutivi. I dati sono stati tabulati sul foglio "Esempio 4.4", mentre in Figura 4.2 sono mostrate le prime dieci righe della tabella. Calcolare le frequenze dei dati suddivisi nelle sei classi: 0 (risposta immediata, nessuna attesa), poi fino a 15, 30, 60, 120, 180 secondi (estremo superiore incluso). Calcolare la media aritmetica in Excel per i dati ripartiti in classi.

Proposta di soluzione in Excel

Il calcolo delle frequenze per i dati ripartiti in classi può essere svolto in modo simile a quanto visto negli esercizi del primo capitolo. Riassumiamo per semplicità i passi:

1. Calcolare con le funzioni MIN() e MAX() i valori minimo e massimo della distribuzione. Si troveranno i valori 0 e 221.
2. Nella colonna `M` a partire da `M7` inserire l'estremo superiore di ogni classe: 0, 15, 30, 60, 120, 180. Osservando che esistono alcuni valori superiori a 180 (il massimo trovato è 221) è necessario inserire una ulteriore classe, (221) per tener conto di questi ultimi.
3. Selezionare le celle `N7:N13` ed inserire la funzione Excel
   `= FREQUENZA (A2:J501;M7:M13)` e confermare con i tre tasti `Ctrl + Maiusc + Invio`.
4. Nella colonna `O` compariranno le frequenze.

Non esiste in Excel una funzione o un comando che calcoli la media per classi di frequenza in modo automatico, occorre svolgere i calcoli a mano. Quindi si deve procedere secondo lo schema illustrato nell'Esempio 4.2:

1. Nella colonna L inserire i limiti inferiori delle classi: 0, 1, 16, 31, 61, 121, 181.
2. Nella colonna O, a partire da O7 calcolare il valore centrale della classe con la formula =(L7 + M7)/2.

	A	B	C	D	E	F	G	H	I	J
1	Tempi di attesa degli utenti di un call center (in secondi)									
2	66	8	59	36	41	4	137	48	58	94
3	23	2	124	43	163	116	54	100	45	48
4	98	69	0	30	114	42	45	0	82	61
5	103	0	18	144	49	109	111	44	46	0
6	66	68	0	46	61	12	86	16	80	142
7	148	141	87	71	65	0	116	4	106	62
8	54	83	44	110	121	151	14	18	85	99
9	0	71	47	136	40	51	74	64	89	11
10	76	14	0	35	0	137	15	0	100	80
11	53	51	96	137	104	154	94	71	78	79

Figura 4.2 La tabella dei dati da raggruppare in classi.

	L	M	N	O	P	Q	R
1							
2	Minimo:	0					
3	Massimo:	221					
4				=FREQUENZA(A2:J501;M7:M13)			
5	Classi di frequenza						
6	Inferiore	Superiore	Frequenza f_i	Valore centrale m_i	$m_i \times f_i$		
7	0	0	346	0	0		
8	1	15	329	8	2632		
9	16	30	429	23	9867		
10	31	60	1259	45,5	57284,5		
11	61	120	2155	90,5	195027,5	< =O11*N11	
12	121	180	460	150,5	69230	< =O12*N12	
13	181	221	22	201	4422	< =O13*N13	
14					338463	< =SOMMA(P7:P13)	
15							
16	=SOMMA(N7:N13)		→ 5000	Media:	67,69	< =P14/N16	
17							
18			=(L10+M10)/2				
19							

Figura 4.3 Prospetto e formule per il calcolo della media di dati raggruppati in classi.

3. Calcolare nella cella N16 con la funzione =SOMMA(), la somma di tutte le frequenze.
4. Calcolare nella cella P14 con la funzione =SOMMA(), la somma dei valori della colonna.
5. Nella cella P16 calcolare la media per classi con la formula =P14/N16.

Seguendo le istruzioni il risultato dovrebbe apparire come in Figura 4.3. □

4.2.2 Media geometrica

Utilizzando il *prodotto* come criterio di trasferibilità della variabile definiamo la *media geometrica* (indicata con M_g) come quel valore che sostituito a tutti i termini della distribuzione ne lascia invariata la funzione prodotto.

Ne deriva che:

$$M_g = \left(\prod_{i=1}^{N} x_i \right)^{\frac{1}{N}} = \sqrt[N]{\prod_{i=1}^{N} x_i} \tag{4.3}$$

Ovvero il calcolo della media geometrica nel caso di *distribuzioni semplici* si ottiene attraverso la radice *n-sima* del prodotto delle modalità x_i. Per calcolarla si ricorre ai logaritmi:

$$LogM_g = \frac{Logx_1 + Logx_2 + ...Logx_N}{N} \tag{4.4}$$

Esempio 4.12

Date tre serie ordinate di voti riportati all'esame di Statistica da tre gruppi di 10 studenti:

A) 21, 21, 22, 23, 23, 24, 24, 24, 24, 24
B) 20, 21, 21, 21, 23, 24, 24, 25, 25, 26
C) 18, 19, 21, 21, 23, 23, 24, 25, 27, 29

la media aritmetica nei tre casi è sempre pari a 23 mentre il campo di variazione presenta risultati differenti a seconda della presenza di valori più o meno elevati negli estremi:

A) $R = 24 - 21 = 3$
B) $R = 26 - 20 = 6$
C) $R = 29 - 18 = 11$ □

4.5.2 Differenza interquartilica

Meno sensibile ai valori estremi è la differenza interquartilica data dalla differenza tra il terzo quartile e il primo quartile, ovvero da:

$$d_q = Q_3 - Q_1 \tag{4.14}$$

Tale misura tiene conto solo dei valori osservati concentrati nella parte centrale della distribuzione.

Esempio 4.13

Riprendendo i dati rilevati nell'esempio 4.11 dove:

$$Q_1 = 25 + \frac{35-25}{32}\left(\frac{147}{4} - 25\right) = 28{,}67 \quad \text{anni}$$

$$Q_3 = 55 + \frac{65-55}{30}\left(\frac{3\times 147}{4} - 102\right) = 57{,}75 \quad \text{anni}$$

avremo che la differenza interquartilica sarà data da

$$d_q = 57{,}75 - 28{,}67 = 29{,}08 \text{ anni}$$ □

4.5.3 Misure di dispersione

Una misura più adeguata per il calcolo della variabilità deve però tenere conto di tutti i valori assunti dalla distribuzione e non solo di una parte di essi. Si ricorre, quindi, per una misura che rilevi di quanto, in media, le diverse quantità rilevate differiscano dalla media aritmetica, assunta come rappresentativa del carattere. Calcolati gli *scarti dalla*

media, $(x_i - \mu)$ delle singole modalità, poiché gli scarti sommati tra di loro danno sempre come risultato zero[2], bisognerà elevare al quadrato tutti gli scarti per renderli positivi e prendere quindi la media aritmetica degli scarti al quadrato. Così si ottiene la *varianza,* uno degli indici più frequentemente utilizzati per il calcolo della variabilità assoluta:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \tag{4.15}$$

dove μ indica la media aritmetica della popolazione. Nel caso in cui i dati siano presentati sotto forma di distribuzioni di frequenza basterà moltiplicare il numeratore per n_i:

$$\sigma^2 = \frac{\sum_{i=1}^{s} (x_i - \mu)^2 \times n_i}{N} \tag{4.16}$$

Effettuare questi calcoli è particolarmente semplice disponendo di un personal computer e uno strumento versatile come Excel, ma fino a non molto tempo fa, prima della diffusione dei pc, gli statistici utilizzavano metodi di calcolo adatti alle calcolatrici in grado di far risparmiare sul numero di operazioni da effettuare e, quindi, sul tempo totale impiegato. A titolo di esempio viene presentato un metodo alternativo per il calcolo della varianza:

$$\sigma^2 = \frac{\sum_{i=1}^{s} x_i^2 \times n_i}{N} - \mu^2 \tag{4.17}$$

La varianza fornisce la misura sintetica di quanto le unità differiscono dalla media aritmetica ed assume valori sempre positivi ($\sigma^2 > 0$). Possiede però un limite: è espressa nell'unità di misura del fenomeno *al quadrato*. Mentre la media è espressa nella stessa unità di misura dei dati, la media delle età è un'età, la media dei prezzi in euro è un prezzo in euro o la media dei libri letti sono libri, la varianza è calcolata elevando al quadrato i dati ottenendo così età al quadrato, euro al quadrato o libri al quadrato (cosa sono dei libri al quadrato?). La varianza non possiede, quindi, un significato "fisico", ma è soltanto un importante indice utile per fini statistici.

Per la varianza si utilizza la lettera greca sigma σ quando il calcolo è effettuato sull'intera popolazione. Si usa invece la lettera s quando i dati provengono da un campione. In questo caso al posto di n al denominatore si pone $n - 1$:

[2] Danno sempre zero poiché il "peso" degli scarti a sinistra della media bilancia il peso di quelli a destra, essendo la media aritmetica il baricentro di un insieme di dati.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{4.18}$$

Il numeratore della varianza viene definito *devianza* ed è dato dalla somma degli scarti dalla media al quadrato:

$$Dev(x) = \sum_{i=1}^{N}(x_i - \mu)^2 \quad \text{nel caso semplice} \tag{4.19}$$

$$Dev(x) = \sum_{i=1}^{s}(x_i - \mu)^2 \times n_i \quad \text{nel caso di distribuzioni} \tag{4.20}$$

Utilizzando la formula semplificata, come nel caso della Formula 4.17 avremo:

$$Dev(x) = \sum_{i=1}^{s} x_i^2 \times n_i - N\mu^2 \tag{4.21}$$

Esempio 4.14

La Pilsen Bier è presente presso l'October fest con $N = 11$ chioschi di degustazione della birra. Le quantità vendute da ogni chiosco in una certa sera sono mostrate nella Tabella 4.7. Calcolare varianza e devianza delle quantità vendute.

Si calcoli innanzi tutto la media delle quantità vendute:

$$\mu = \frac{3.822}{11} = 347,45$$

Si proceda con il calcolo degli scarti dalla media la cui somma sarà uguale a zero, quindi con il quadrato degli scarti. Sommando gli scarti al quadrato ed utilizzando la formula (4.15) si otterrà la varianza:

$$\sigma^2 = \frac{72.864,73}{11} = 6.624,07$$

mentre la devianza si ricava dalla formula (4.19):

$$Dev(x) = 72.864,7$$

Proposta di soluzione in Excel

Lo stesso problema si può risolvere immediatamente in Excel utilizzando le funzioni dirette per il calcolo della media, `MEDIA()` e della varianza `VAR.POP()`. Excel offre due

Tabella 4.7 Calcolo della varianza e della devianza.

Boccali di birra venduti da ogni chiosco			
	x_i	$(x_i - \mu)$	$(x_i - \mu)^2$
	207	-140,5	19.727,5
	254	-93,5	8.733,8
	261	-86.5	7.474,4
	300	-47,5	2.251,9
	327	-20,5	418,4
	353	5,5	30,8
	379	31,5	995,1
	404	56,5	3.197,4
	419	71,5	5.118,8
	455	107,5	11.566,0
	463	115,5	13.350,8
Totale	*3822*	*0,0*	*72.864,7*
$\mu = 347,45$			

funzioni, `VAR.POP()` e `VAR()` per il calcolo della varianza: la prima effettua il calcolo sulla base di una popolazione ed applica la formula (4.15) mentre la seconda considera i dati estratti da un campione e, quindi, pone al denominatore la quantità N-1 per aumentare la variabilità. Nell'esempio dei chioschi di birra, poiché stiamo considerando la totalità delle birre vendute in una sera dovremo utilizzare la funzione `VAR.POP()`. Per il calcolo della Devianza in Excel si utilizza la funzione `DEV.Q()`. La Figura 4.8 mostra il semplice prospetto con i calcoli. □

	A	B	C	D
1	x_i			
2	207	μ =	347,45	<=MEDIA(A2:A12)
3	254	σ^2 =	6.624,07	<=VAR.POP(A2:A12)
4	261	Dev(x) =	72.864,73	<=DEV.Q(A2:A12)
5	300			
6	327			
7	353			
8	379			
9	404			
10	419			
11	455			
12	463			
13				

Figura 4.8 Funzioni per il calcolo della media e della varianza di una popolazione.

Per completezza diamo anche la soluzione di calcolo in Excel che segue gli stessi passaggi[3] della Tabella 4.7. In Figura 4.9 si possono osservare i passaggi per calcolare le colonne degli scarti e degli scarti al quadrato, e in basso il calcolo semplice ed immediato della media e della varianza.

Esempio 4.15

Presso un Policlinico, nel 2004 è stata effettuata una rilevazione sui pazienti dimessi in una certa settimana allo scopo di indagare la durata dei ricoveri presso un reparto di chirurgia. I dati sono mostrati nella Tabella 4.8. Calcolare media, varianza e devianza relativi ai giorni di ricovero.

In questo esempio la durata del ricovero in giorni costituisce la variabile statistica x_i, la frequenza con cui si manifesta è data dal numero di pazienti ricoverati (n_i) e infine la somma dei pazienti rappresenta la numerosità totale $N = \sum_{i=1}^{k} n_i$ La variabile statistica x_i andrà ponderata con le frequenze n_i. Per calcolare la media ponderata si dovrà impiegare la formula (4.2):

$$\mu = \frac{5.346}{378} = 14,14$$

	A	B	C	D	E
1	x_i	$x_i - \mu$		$(x_i - \mu)^2$	
2	207	-140,45	<=A2-B14	19.727,48	<=B2^2
3	254	-93,45	<=A3-B14	8.733,75	<=B3^2
4	261	-86,45		7.474,39	
5	300	-47,45		2.251,93	
6	327	-20,45		418,39	
7	353	5,55		30,75	
8	379	31,55		995,12	
9	404	56,55		3.197,39	
10	419	71,55		5.118,75	
11	455	107,55		11.566,02	
12	463	115,55		13.350,75	
13	3822	<=SOMMA(A2:A12)		72.864,73	<=SOMMA(D2:D12)
14	μ =	347,45	<=A13/11		
15	σ^2 =	6.624,07	<=D13/11		

Figura 4.9 Calcolo di media e varianza senza l'utilizzo delle funzioni di Excel.

[3] Questi calcoli verranno utili quando si dovranno calcolare le misure per dati aggregati in classi o ponderati, come nel successivo esempio.

Tabella 4.8 Calcolo della varianza e della devianza per dati ponderati.

giorni ricovero x_i	numero pazienti n_i	$x_i \times n_i$	$x_i - \mu$	$(x_i - \mu)^2$	$(x_i - \mu)^2 \times n_i$
2	7	14	-12,14	147,45	1.032,14
4	17	68	-10,14	102,88	1.748,92
5	7	35	-9,14	83,59	585,14
6	1	6	-8,14	66,31	66,31
7	5	35	-7,14	51,02	255,10
8	74	592	-6,14	37,73	2.792,37
12	2	24	-2,14	4,59	9,18
13	16	208	-1,14	1,31	20,90
16	163	2608	1,86	3,45	562,18
19	51	969	4,86	23,59	1.203,18
22	33	726	7,86	61,73	2.037,24
28	1	28	13,86	192,02	192,02
33	1	33	18.86	355,59	355.59
175	*378*	*5346*			*10.860,29*
		μ = 14,14			

Si passa poi a calcolare gli scarti, elevarli al quadrato e moltiplicarli per le rispettive frequenze n_i. Dopo aver calcolato la somma con la formula (4.16) si otterrà la varianza:

$$\sigma^2 = \frac{10.860,29}{378} = 28,73$$

mentre applicando la formula (4.20) si avrà la devianza:

$$Dev(\text{x}) = 10.860,29$$

In conclusione si può affermare che la durata media del ricovero per i pazienti considerati è pari a μ = 14,14 giorni con una varianza di σ^2 = 28,73. Ricordiamo che la varianza non ha un significato fisico (sono "giorni al quadrato") ma, come si vedrà nei capitoli successivi, potrà essere utilizzata per ottenere ulteriori importanti informazioni.

Applicando le formule semplificate 4.17 e 4.21 per il calcolo della varianza e della devianza si perviene naturalmente allo stesso risultato:

$$\sigma^2 = \frac{86.468}{378} - 14,14^2 = 28,73$$

$$Dev(x) = 86.468 - 378 \times 14,14^2 = 10.860,29$$

Nella Tabella 4.9 è mostrato lo schema di calcolo.

Tabella 4.9 Calcolo della varianza e della devianza con le formule abbreviate.

giorni ricovero x_i	numero pazienti n_i	$(x_i)^2$	$(x_i)^2 \times n_i$
2	7	4	28
4	17	16	272
5	7	25	175
6	1	36	36
7	5	49	245
8	74	64	4736
12	2	144	288
13	16	169	2704
16	163	256	41728
19	51	361	18411
22	33	484	15972
28	1	784	784
33	1	1089	1089
175	*378*	*3481*	*86468*

Proposta di soluzione in Excel

Questo esempio non può essere svolto in Excel utilizzando le funzioni per il calcolo diretto della varianza dato che siamo in presenza di dati ponderati. Lo schema proposto nella Tabella 4.9 deve essere ripreso in modo puntuale in Excel, creando una formula per ogni colonna, come si può osservare nella Figura 4.10 con l'unica osservazione di aver collocato alla destra di ciascuna colonna la formula impiegata per ottenere il risultato. □

	A	B	C	D	E	F	G	H	I
1	giorni ricovero x_i	numero pazienti n_i	$xi \times ni$	$x_i - \mu$		(xi - μ)2		(xi - μ)2 × ni	
2	2	7	14	-12,14	<=A2-C16	147,45	<=D2^2	1.032,14	<=F2*B2
3	4	17	68	-10,14	<=A3-C16	102,88	<=D3^2	1.748,92	<=F3*B3
4	5	7	35	-9,14	<=A4-C16	83,59	<=D4^2	585,14	<=F4*B4
5	6	1	6	-8,14		66,31		66,31	
6	7	5	35	-7,14		51,02		255,10	
7	8	74	592	-6,14		37,73		2.792,37	
8	12	2	24	-2,14		4,59		9,18	
9	13	16	208	-1,14		1,31		20,90	
10	16	163	2608	1,86		3,45		562,18	
11	19	51	969	4,86		23,59		1.203,18	
12	22	33	726	7,86		61,73		2.037,24	
13	28	1	28	13,86		192,02		192,02	
14	33	1	33	18,86		355,59		355,59	
15	175	378	5346					10.860,29	
16		μ =	14,14	<=C15/B15					
17		σ^2 =	28,73	<=H15/B15					

Figura 4.10 Calcolo della varianza in Excel per dati ponderati .

Le formule impiegate sono molto semplici, eseguono prodotti tra celle appartenenti alla stessa riga ad eccezione di quelli della colonna `E` che utilizzano un riferimento assoluto alla cella `C16` (si dovrà pertanto scrivere `$C$16`) dove si trova la media da sottrarre. Dopo aver calcolato le somme delle colonne utilizzando semplicemente la funzione `=SOMMA()`, nelle righe `16` e `17` sono state calcolate la media e la varianza. □

Come si è visto, l'interpretazione della varianza pone dei limiti poiché è espressa nell'unità di misura del fenomeno al quadrato. Per ovviare a tale inconveniente Pearson[4] introdusse il concetto di *scarto quadratico medio*[5] (abbreviato in *sqm*) quale radice quadrata della varianza:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \tag{4.22}$$

Nel caso di una distribuzione di frequenza occorre prendere gli scarti ponderati (ovvero gli scarti moltiplicati per la frequenza n_i). Si avrà:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{s}(x_i - \mu)^2 \times n_i}{N}} \tag{4.23}$$

Così facendo l'indice di variabilità ritorna a essere espresso nella stessa unità di misura del fenomeno ed è sempre positivo. Lo scarto quadratico medio sarà tanto maggiore quanto maggiore è la variabilità dei valori di un insieme di dati, mentre assumerà valore nullo nel caso in cui tutti i valori siano uguali tra di loro. Anche per lo scarto quadratico medio si utilizza la lettera greca sigma (σ) quando il calcolo è effettuato sull'intera popolazione, mentre si adopera la lettera s quando lo scarto quadratico medio viene calcolato su un campione. Come spiegato all'inizio del paragrafo, al posto di n al denominatore si pone $n - 1$[6]:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{4.24}$$

Poiché il denominatore di s è minore del denominatore di σ, s sarà maggiore di σ. Un altro indice di variabilità, espresso nella stessa unità di misura del fenomeno, è lo *scar-*

[4] K. Pearson (1857-1936), matematico inglese, ha dato numerosi contributi allo sviluppo della moderna statistica.
[5] In inglese Standard deviation.
[6] Ricordiamo che per campioni con numerosità $n>30$ la differenza tra s e σ diventa trascurabile.

to semplice medio (che verrà abbreviato con *ssm*) che rappresenta la media aritmetica degli scarti presi in valore assoluto:

$$\delta = \frac{\sum_{i=1}^{N} |x_i - \mu|}{N} \tag{4.25}$$

Nel caso di una distribuzione di frequenza basterà moltiplicare gli scarti in valore assoluto per n_i ottenendo:

$$\delta = \frac{\sum_{i=1}^{s} |x_i - \mu| \times n_i}{N} \tag{4.26}$$

Tale misura ha il particolare vantaggio di non richiedere calcoli particolarmente complessi in quanto non vi sono funzioni quadratiche. Rispetto allo scarto quadratico medio si verifica sempre la seguente relazione:

$$\delta \leq \sigma$$

Esempio 4.16

Riprendendo l'Esempio 4.14 sui chioschi di birra per calcolare lo scarto quadratico medio è sufficiente estrarre la radice quadrata della varianza:

$$\sigma = \sqrt{\frac{72.864,73}{11}} = 81,39$$

Per concludere l'esempio si può affermare che gli undici chioschi hanno venduto in una sera una media $\mu = 347,45$ boccali di birra con uno sqm $\sigma = 81,39$ boccali. In altre parole e senza avere la pretesa del rigore[7], la maggior parte dell'oscillazione delle quantità vendute tra un chiosco e l'altro si deve aggirare, in più o in meno, intorno al valore dello sqm. Osservando la colonna delle x_i nella Tabella 4.10, si può notare infatti che aggiungendo o togliendo 81 a 347 si ottiene una intervallo che va da 266 a 428 boccali che contiene circa il 60% dei valori.

Per calcolare lo scarto semplice medio utilizzando la formula (4.25) sarà necessario calcolare gli scarti in valore assoluto organizzando i dati come appaiono nella Tabella 4.10:

[7] Nei capitoli successivi verrà definito in termini rigorosi il significato di intervallo di variabilità.

Tabella 4.10 Calcolo dello scarto semplice medio.

x_i	$x_i - \mu$	$\lvert x_i - \mu \rvert$
207	-140,45	140,45
254	-93,45	93,45
261	-86.45	86.45
300	-47,45	47,45
327	-20,45	20,45
353	5,55	5,55
379	31,55	31,55
404	56,55	56,55
419	71,55	71,55
455	107,55	107,55
463	115,55	115.55
3822		*776,55*
	μ = 347,45	

Dividendo la somma dei valori assoluti degli scarti per N avremo lo scarto semplice medio:

$$\delta = \frac{776,55}{11} = 70,60$$

Proposta di soluzione in Excel

Analogamente a quanto visto nel caso della varianza, per il calcolo diretto dello sqm si possono utilizzare le funzioni `=DEV.ST.POP()` nel caso di popolazioni e `=DEV.ST()` per lo scarto quadratico medio di campioni. In questo caso si applicherà la prima delle due ottenendo una tabella simile a quella della Figura 4.8. □

	A	B	C	D	E
1	x_i				
2	207	μ =	347,45	<=MEDIA(A2:A12)	
3	254	σ =	81,39	<=DEV.ST.POP(A2:A12)	
4	261				
5	300				
6	327				
7	353				
8	379				
9	404				
10	419				
11	455				
12	463				
13					

Figura 4.11 Calcolo dello sqm con la funzione Excel.

	A	B	C	D	E
1	x_i	$\lvert x_i - \mu \rvert$			
2	207	140,45	<=ASS(A2-B14)		
3	254	93,45	<=ASS(A3-B14)		
4	261	86,45			
5	300	47,45			
6	327	20,45			
7	353	5,55			
8	379	31,55			
9	404	56,55			
10	419	71,55			
11	455	107,55			
12	463	115,55			
13					
14	μ =	347,45	<=MEDIA(A2:A12)		
15	σ =	81,39	<=DEV.ST.POP(A2:A12)		
16	ssm =	70,60	<=SOMMA(B2:B12)/CONTA.NUMERI(A2:A12)		
17					

Figura 4.12 Calcolo dello ssm in Excel eseguendo direttamente i calcoli.

Per il calcolo dello scarto semplice medio Excel non possiede alcuna funzione: i calcoli andranno eseguiti in modo esplicito seguendo la traccia della Tabella 4.10.

Il valore assoluto degli scarti si calcola utilizzando la funzione `ASS()`, ponendo per argomento lo scarto come mostrato in Figura 4.12. Nella cella `B16` si è posta la formula (4.25) che calcola il rapporto tra somma degli scarti e numerosità. Invece di scrivere direttamente il valore 11 si è utilizzata la funzione[8] `CONTA.NUMERI()` che esegue automaticamente il conteggio dei numeri trovati nell'intervallo `A2:A12`.

Esempio 4.17

Riprendendo i dati dell'Esempio 4.15 relativo alla durata dei ricoveri presso un Policlinico, calcoliamo lo scarto quadratico medio estraendo la radice quadrata della varianza:

$$\sigma = \sqrt{\frac{10.860,29}{378}} = 5,36$$

[8] In questo esempio può sembrare eccessivo utilizzare una funzione per contare 11 righe, sarebbe più semplice contarle sul video. Si utilizza questa formula sotto la spinta di due considerazioni: la prima è quella che gli esempi in un testo devono essere elementari ma le applicazioni vere possono avere anche centinaia di righe ed allora è indispensabile saper utilizzare le funzioni; la seconda riguarda il fatto che Excel è uno strumento per il calcolo automatico: scrivendo 11 nella formula la si rende rigida, mentre utilizzando la funzione si potranno aggiungere o togliere righe dalla tabella e la formula si adeguerà automaticamente.

Si può affermare che per un dato reparto di chirurgia nella settimana selezionata la durata media del ricovero è $\mu = 14{,}14$ giorni con uno sqm $\sigma = 5{,}36$ giorni. Ripetendo il ragionamento svolto nell'Esempio 4.16 potremo dire che la media dei giorni di ricovero si aggira tra 14,1 – 5,4 = 8,7 giorni e 14,1 + 5,4 = 19,5 giorni.

Per calcolare lo scarto semplice medio secondo la formula (4.26) sarà necessario calcolare gli scarti in valore assoluto e moltiplicarli per le frequenze n_i come mostrato nella Tabella 4.11:

Tabella 4.11 Calcolo dello scarto semplice medio.

giorni ricovero x_i	numero pazienti n_i	$x_i \times n_i$	$x_i - \mu$	$(x_i - \mu) \times n_i$	$\lvert x_i - \mu \rvert \times n_i$
2	7	14	-12,1	-85,0	85,0
4	17	68	-10,1	-172,4	172,4
5	7	35	-9,1	-64,0	64,0
6	1	6	-8,1	-8,1	8,1
7	5	35	-7,1	-35,7	35,7
8	74	592	-6,1	-454,6	454,6
12	2	24	-2,1	-4,3	4,3
13	16	208	-1,1	-18,3	18,3
16	163	2608	1,9	302,7	302,7
19	51	969	4,9	247,7	247,7
22	33	726	7,9	259,3	259,3
28	1	28	13,9	13,9	13,9
33	1	33	18,9	18,9	18,9
Totale	*378*	*5.346*		–	*1.684,9*

Dopo aver calcolato i valori nelle colonne si calcola il rapporto:

$$\delta = \frac{1.684,9}{378} = 4,5$$

Proposta di soluzione in Excel

Per calcolare lo sqm riprendiamo la tabella dell'Esempio 4.15 (vedi Figura 4.10): si tratterà di aggiungere nella cella C17 alla formula della varianza l'estrazione della radice quadrata con la funzione `=RADQ()`. Avremo quindi `=RADQ(H15/B15)` come mostrato in Figura 4.13. □

Esempio 4.18

La Tabella 4.12 contiene il reddito in euro di 60 famiglie utenti di un servizio di una ASL piemontese. I dati sono stati raggruppati in quattro classi di frequenza. Calcolare il reddito familiare medio, lo scarto quadratico medio e lo scarto semplice medio.

	A	B	C	D	E	F	G	H	I
1	giorni ricovero x_i	numero pazienti n_i	$xi \times ni$	$x_i - \mu$		(xi - μ)2		(xi - μ)2 × ni	
2	2	7	14	-12,14	<=A2-C16	147,45	<=D2^2	1.032,14	<=F2*B2
3	4	17	68	-10,14	<=A3-C16	102,88	<=D3^2	1.748,92	<=F3*B3
4	5	7	35	-9,14	<=A4-C16	83,59	<=D4^2	585,14	<=F4*B4
5	6	1	6	-8,14		66,31		66,31	
6	7	5	35	-7,14		51,02		255,10	
7	8	74	592	-6,14		37,73		2.792,37	
8	12	2	24	-2,14		4,59		9,18	
9	13	16	208	-1,14		1,31		20,90	
10	16	163	2608	1,86		3,45		562,18	
11	19	51	969	4,86		23,59		1.203,18	
12	22	33	726	7,86		61,73		2.037,24	
13	28	1	28	13,86		192,02		192,02	
14	33	1	33	18,86		355,59		355,59	
15	**175**	**378**	**5346**					**10.860,29**	
16		**μ =**	**14,14**	**<=C15/B15**					
17		**σ =**	**5,36**	**<=RADQ(H15/B15)**					

Figura 4.13 Tabella per il calcolo dello sqm per dati ponderati.

Tabella 4.12 Calcolo dello scarto quadratico medio per dati raggruppati in classi.

redditi x_i	**numero famiglie n_i**	**valore centrale x_m**	$x_m \times n_i$	$x_i - \mu$	$(x_i - \mu)^2$	$(x_i - \mu)^2 \times n_i$
0 ⊢ 10	10	5	50	-15	225	2250
20 ⊢ 30	40	15	600	-5	25	1000
20 ⊢ 30	40	25	1000	5	25	1000
30 ⊢ 40	10	35	350	15	225	2250
Totale	*100*		*2000*	*0*		*6500*
			$\mu = 20$			$sqm = 8,06$

I dati sono espressi in migliaia di euro

Il reddito costituisce la variabile statistica x_i, il numero di famiglie rappresenta la frequenza n_i con cui si presenta il reddito, la somma del numero di famiglie rappresenta la numerosità totale $N = \sum_{i=1}^{k} n_i$ Per calcolare la frequenza per classe (quarta colonna della tabella) si dovrà utilizzare il valor medio della classe x_m ottenuto dividendo per due la somma degli estremi della classe:

$$x_m = \frac{10+20}{2} = 15$$

Osservando la Tabella 4.12, i passi da compiere per calcolare lo scarto quadratico medio saranno nell'ordine:

1. Sommare le frequenze n_i
2. Calcolare i valori centrali delle classi x_m
3. Moltiplicare la frequenza con il valore centrale della classe: $x_m \times n_i$
4. Sommare i valori ottenuti
5. Calcolare la media aritmetica $\mu = \frac{2000}{100} = 20$ ricordandosi che il reddito è espresso in migliaia di euro e quindi il valore medio corrisponderà a 20.000 euro
6. Calcolare gli scarti dalla media ed elevarli al quadrato
7. Moltiplicare gli scarti al quadrato per le frequenze n_i
8. Sommare i valori ottenuti
9. Dividere quest'ultima somma con la numerosità N ed estrarne la radice quadrata.

Si osservi che i valori in euro presenti sulla tabella andranno tutti moltiplicati per 1000. In conclusione, si può affermare che il reddito medio delle 100 famiglie e pari a 20.000 euro con uno sqm di 8.060 euro. Sottraendo al reddito medio lo sqm si avrà: 20.000 – 8.060 = 11.940; in modo analogo si aggiunga lo sqm alla media: 20.000 + 8.060 = 28.060. Avremo quindi con una certa approssimazione che nell'intervallo compreso tra 11.940 e 28.940 euro si collocheranno la maggior parte delle famiglie.

In modo analogo si può procedere per il calcolo dello scarto semplice medio come mostrato nella Tabella 4.13.

Tabella 4.13 Calcolo dello scarto semplice medio per dati raggruppati in classi.

redditi x_i	**numero famiglie n_i**	**valore centrale x_m**	$x_m \times n_i$	$x_i - \mu$	$\lvert x_i - \mu \rvert$	$\lvert x_i - \mu \rvert \times n_i$
0 ⊢ 10	10	5	50	-15	15	150
10 ⊢ 20	40	15	600	-5	5	200
20 ⊢ 30	40	25	1000	5	5	200
30 ⊢ 40	10	35	350	15	15	150
Totale	*100*	*80*	*2000*	*0*	*40*	*700*
		$\mu =$	*20*		*ssm =*	*7,00*

I dati sono espressi in migliaia di euro

Riprendendo l'elenco dei passi da percorrere mostrato prima si osservi che i primi cinque passi non mutano in questo nuovo calcolo, mentre i successivi saranno:

6. Calcolare gli scarti dalla media e ricavarne il valore assoluto
7. Moltiplicare gli scarti assoluti per le frequenze n_i
8. Sommare i valori ottenuti
9. Dividere quest'ultima somma con la numerosità N

Si otterrà il valore di 7,00 per lo scarto semplice medio.

	A	B	C	D	E	F	G	H	I
1	redditi x_i	numero famiglie n_i	valore centrale x_m	$x_m \times n_i$	$x_i - \mu$	$(x_i - \mu)^2$		$(x_i - \mu)^2 \times n_i$	
2	0 ⊢ 10	10	5	50	-15	225	<=E2^2	2250	<=F2*B2
3	10 ⊢ 20	40	15	600	-5	25	<=E3^2	1000	<=F2*B2
4	20 ⊢ 30	40	25	1000	5	25		1000	
5	30 ⊢ 40	10	35	350	15	225		2250	
6	*Totale:*	100		2000	0			6500	<=SOMMA(H2:H5)
7			$\mu =$	**20**		*sqm =*		**8,06**	<=RADQ(H6/B6)
8									

Figura 4.14 Lo schema e le formule Excel per il calcolo dello scarto quadratico medio per dati raggruppati in classi.

	A	B	C	D	E	F	G	H	I	J
1	redditi x_i	numero famiglie n_i	valore centrale x_m	$x_m \times n_i$	$x_i - \mu$	$\lvert x_i - \mu \rvert$		$\lvert x_i - \mu \rvert \times n_i$		
2	0 ⊢ 10	10	5	50	-15	15	<=ASS(E2)	150	<=F2*B2	
3	10 ⊢ 20	40	15	600	-5	5	<=ASS(E3)	200	<=F3*B3	
4	20 ⊢ 30	40	25	1000	5	5		200		
5	30 ⊢ 40	10	35	350	15	15		150		
6	*Totale:*	100		2000	0			700	<=SOMMA(H2:H5)	
7			$\mu =$	**20**		*ssm =*		**7,00**	<=H6/B6	

Figura 4.15 Lo schema e le formule Excel per il calcolo dello scarto semplice medio per dati raggruppati in classi.

Proposta di soluzione in Excel

Ripercorrendo i passi illustrati avremo le tabelle mostrate nella Figura 4.14 e Figura 4.15. Le formule impiegate non presentano particolari difficoltà. Per calcolare il valore assoluto si ricorre alla funzione Excel `ASS()` nelle formule della colonna `F`. □

4.5.4 Misure di disuguaglianza

Come sostenuto all'inizio del paragrafo, le misure di variabilità assoluta si dividono in misure di dispersione e *misure di disuguaglianza*. Le misure di disuguaglianza consistono nel calcolare la media delle differenze tra tutte le coppie di numeri; esse vengono utilizzate per indicare di quanto differiscono tra di loro le quantità rilevate.

La *differenza media assoluta* di Gini è definita dalla somma dei valori assoluti delle differenze tra tutte le coppie che si formano tra i valori x_i della variabile statistica. In formula sarà:

$$\Delta = \frac{\sum_{i=1}^{N}\sum_{h=1}^{N}|x_i - x_h|}{N(N-1)} \qquad (4.27)$$

Si considerano i valori assoluti delle differenze per evitare che ogni confronto tra due termini si annulli con il confronto di segno opposto. Nel caso di una distribuzione raggruppata in classi di frequenza la differenza media si calcola come:

$$\Delta = \frac{\sum_{i=1}^{s}\sum_{h=1}^{s}|x_i - x_h| \times n_i \times n_h}{N(N-1)} \tag{4.28}$$

Tale differenza media è definita *senza ripetizione* in quanto al denominatore non si considerano le differenze tra ciascun termine e se stesso. La differenza media *con ripetizione* sarà, invece, data da:

$$\Delta = \frac{\sum_{i=1}^{N}\sum_{h=1}^{N}|x_i - x_h|}{N^2} \tag{4.29}$$

Il calcolo della differenza media può effettuarsi anche attraverso la disposizione dei dati in forma di tabella (come nell'esempio successivo) o attraverso formule alternative utilizzate nel caso in cui N sia elevato[9].

Esempio 4.19

Consideriamo i seguenti redditi di quattro famiglie, espressi in migliaia di euro: 15, 20, 35, 70. Si calcoli la differenza media.

Per il calcolo della differenza media di Gini bisognerà individuare tutte le possibili coppie di valori della variabile x. Fissato il primo valore, 15, avremo tre coppie possibili:

15 - 20 15 – 35 15 - 70

Analogamente con il secondo valore 20 si avrà:

20 - 15 20 – 35 20 - 70

continuando così anche per gli altri due valori. Alla fine prendendo il valore assoluto di queste differenze potremo scrivere:

$$\frac{|15-20|+|15-35|+|15-70|+|20-15|+|20-35|+|20-70|+|35-15|+|35-20|+|35-70|+|70-15|+|70-20|+|70-35|}{4(4-1)}$$

che darà come risultato:

$$\frac{360}{12} = 30$$

Ricordiamo che 30 è espresso in migliaia di euro (30.000 euro). Avremo che 30.000 euro costituiscono la differenza media tra i valori del reddito delle quattro famiglie.

[9] Cfr. De Finetti, Paciello, ecc.

FFFF. Gli eventi sono tra loro *indipendenti* (ogni figlio che nasce conserva la stessa probabilità degli altri) e quindi si applicherà la formula:

$$P(FFFF) = P(F) \times P(F) \times P(F) \times P(F) = 0,49^4 = 0,0576 \cong \frac{1}{17}$$

ovvero con una probabiltà del 6% circa una famiglia con 4 figli avrà tutte femmine. Il caso 2) si risolve osservando un maschio e tre femmine possono venire in secondo le seguenti combinazioni: *MFFF FMFF FFMF FFFM*. Ognuna di queste combinazioni ha probabilità:

$$P(M) \times P(F) \times P(F) \times P(F) = (0,51) \times (0,49)^3 = 0,060$$

quindi la probabilità finale è la somma delle combinazioni:

$$0,060 + 0,060 + 0,060 + 0,060 = 0,24$$ □

6.3 Generazione di numeri casuali in Excel

Come tutti gli applicativi informatici che si rispettino Excel dispone di un generatore di numeri casuali[1] che permette di ottenere in modo semplice ottimi risultati. Scrivendo in una cella la formula `=CASUALE()` comparirà un numero compreso tra 0 e 1. Premendo tante volte il tasto `F9` (il comando "calcola" di Excel) quel numero cambierà automaticamente ogni volta.

Più interessante è la funzione `=CASUALE.TRA(<minimo>; <massimo>)` che accetta come parametri il limite inferiore e quello superiore entro cui far cadere il numero. Ad esempio scrivendo `=CASUALE.TRA(1; 2)` otteniamo come risultato solo il valore 1 o 2 che possiamo considerare come il lancio di una moneta. In modo analogo scrivendo `=CASUALE.TRA(1; 6)` otterremo la simulazione del lancio di un dado.

Esempio 6.6

Creare in un foglio Excel una simulazione di 10 lanci di una moneta mostrando il conteggio di quante Teste e quante Croci sono uscite e la frequenza relativa rispetto ai 10 lanci.
In un foglio Excel vuoto, nella cella `A1`, si scriva "Dieci lanci di una moneta". Nella cella `A2` la formula `=CASUALE.TRA(1; 2)`. Si selezioni la cella e si trascini in bas-

[1] Occorre molta prudenza con le parole: i computer non possono (almeno fino a oggi) generare numeri a caso perché agiscono nello spazio degli eventi deterministici (sono programmati in modo deterministico). Esistono tuttavia programmi che con trucchi anche raffinati generano dei numeri che appaiono casuali: in realtà appartengono alla categoria dei numeri *pseudo-casuali* molto utili nelle attività più comuni. Il termine "pseudo" si riferisce al fatto che se si conosce il "seme" da cui il computer inizia la generazione dei numeri (i pc usano il clock interno) è sempre possibile (almeno in via di principio) prevedere la sequenza che verrà generata.

	A	B	C	D	E	F	G
1	Dieci lanci di una moneta						
2	1	<=CASUALE.TRA(1; 2)					
3	1						
4	2	Numero teste uscite:		6	<=CONTA.SE(A2:A11; "=1")		
5	2	Numero croci uscite:		4	<=CONTA.SE(A2:A11; "=2")		
6	1						
7	2	Frequenza teste:		0,6	<=D4/10		
8	1	Frequenza croci:		0,4	<=D5/10		
9	2						
10	1						
11	1						
12							

Figura 6.1 Le funzioni impiegate per simulare dieci lanci di una moneta.

so la selezione fino alla cella `A11`. Compariranno 10 numeri: premendo il tasto[2] `F9` i numeri cambieranno in modo casuale. Nella cella `C4` scrivere "Numero teste uscite:", in `D4` scrivere la formula `=CONTA.SE(A2:A11; "=1")` che conterà il numero di volte che comparirà 1 nella dieci celle dell'intervallo. Ripetere le operazioni nelle celle `C5` e `D5` per inserire il conteggio sul 2. Il risultato finale dovrà apparire simile a quello della Figura 6.1.

Le funzioni `CASUALE()` e `CASUALE.TRA()` generano numeri pseudo casuali equiprobabili; per verificarlo è sufficiente copiare la formula dell'esempio in migliaia di celle e contare le frequenze relative: all'aumentare del numero delle celle la frequenza relativa si avvicinerà al valore atteso di 0,5. Dato che la distribuzione dei numeri è equiprobabile o *uniforme* è adatta a simulare eventi quali il lancio di monete o di dadi o l'estrazione di carte. Alla fine del capitolo vedremo altre funzionalità del generatore di numeri casuali di Excel. □

Esempio 6.7

Creare in un foglio Excel una simulazione del lancio di due dadi a sei facce.

Inserire in una cella la funzione `=CASUALE.TRA(1; 6)`, ripetere la stessa operazione nella casella accanto alla prima. Il risultato potrebbe essere simile alla Figura 6.2. Verificare che tutto funzioni generando nuovi numeri premendo il tasto `F9`. Adesso si possono chiamare gli amici e organizzare un tavolo di scommesse come in un qualunque Casinò di Las Vegas…

[2] Il tasto funzione F9 è il comando per effettuare il calcolo di tutte le formule presenti sul foglio. Di norma le impostazioni di Excel prevedono il ricalcolo automatico: ogni volta che viene effettuata una qualsivoglia operazione che comporta premere il tasto Invio, Excel ricalcola sempre tutte le formule. In generale questo fatto non viene notato perché i numeri sul foglio rimangono sempre gli stessi. L'unica eccezione è data dalla funzione CASUALE() che generando un numero continuamente diverso fa notare immediatamente la differenza. Per evitare la continua generazione di nuovi numeri, è opportuno, dopo averli creati, copiarli in altre celle con l'opzione Incolla valori. Se si vuole fermare il ricalcolo automatico per l'intero foglio si deve andare sulla finestra Opzioni di Excel, selezionare la scheda Formule e deselezionare la casella Automatico.

	A	B	C	D	E	F
1			Lancio di due dadi			
2	=CASUALE.TRA(1; 6) >		4	3	< =CASUALE.TRA(1; 6)	
3						

Figura 6.2 Simulazione del lancio di due dadi.

Le due funzioni `CASUALE()` e `CASUALE.TRA()` trovano utile impiego quando occorre selezionare dei nomi da un elenco. Se, per esempio, si dovessero scegliere 10 persone da un elenco di cento, sarebbe sufficiente scrivere la funzione `=CASUALE.TRA(1; 100)` e premere 10 volte il tasto `F9` per scegliere le dieci persone la cui riga corrisponde al numero uscito. □

6.4. Distribuzioni di probabilità

Il termine variabile casuale indica una funzione definita sullo spazio campionario Ω che associa ad ogni evento un unico numero reale creando così una corrispondenza tra il dominio Ω degli eventi e il codominio R dell'insieme dei numeri reali. Le variabili casuali (v.c. in sigla) prendono anche il nome di variabili aleatorie.

Molti fenomeni tipici delle scienze economiche e sociali possono essere descritti da modelli basati su variabili aleatorie (o casuali): questi modelli prendono il nome di distribuzioni di probabilità. Una variabile casuale può essere:

- *discreta*: se può assumere un insieme finito e numerabile di valori reali X_i in corrispondenza del presentarsi di un evento E_i, avente probabilità P_i.
- *continua*: se può assumere tutti i valori compresi in un intervallo reale (infinità non numerabile di valori).

6.4.1 Variabili casuali discrete

Nel caso di variabili discrete la probabilità P_i è una funzione reale della variabile X denominata *funzione di probabilità* e data da $P(X = x_i)$ ovvero dalla probabilità che la v.c. X assuma il valore x_i. Condizione affinché una v.c. discreta X sia ben definita è che è $P(x_i) \geq 0$ *e* $\sum_{i=1}^{\infty} P(x_i) = 1$. Si può rappresentare graficamente ponendo x_i sulle ascisse e $P(x_i)$ sulle ordinate. Altra importante funzione è la *funzione di ripartizione* (o funzione cumulata delle frequenze) che fornisce la probabilità che la v.c. X assuma valori minori o uguali di un dato valore x_i. Tale funzione è data da $F(x_i) = P(X \leq x) = \sum_{x \leq x_i} P(x_i)$. È una funzione monotona non decrescente tale che $0 \leq F(x_i) \leq 1$.

6.4.2 Variabili casuali continue

Nel caso di variabili continue la probabilità che la v.c. X assuma un valore corrispondente a un certo intervallo sarà denominata *funzione di densità* della probabilità e sarà data da $P(x \leq X \leq x + dx) = f(x)dx$.

La sua *funzione di ripartizione* $F(X)$ è di tipo continuo e la probabilità sarà espressa in termini di area. La relazione tra funzione di ripartizione e funzione di densità è la seguente:

$$F(x) = \int_{-\infty}^{x} f(x)dx .$$

Condizione affinché una v.c. continua sia ben definita è che

$$f(x) \geq 0 \text{ e } \int_{-\infty}^{+\infty} f(x)dx = 1.$$

Nell'ambito delle v.c. discrete e continue è opportuno analizzare alcune tra le variabili casuali più frequentemente utilizzate.

6.5 Variabili casuali discrete

Le variabili casuali *discrete* maggiormente utilizzate per l'analisi statistica sono:

- v.c. Binomiale
- v.c. di Poisson

6.5.1 Variabile casuale binomiale

La distribuzione binomiale è una variabile casuale ottenuta ripetendo per n volte e nelle medesime condizioni la variabile casuale di Bernoulli, ovvero una variabile di tipo dicotomico che ammette solo valori di $x = 0, 1$ a seconda che l'evento E si verifichi ($x = 0$) o non si verifichi ($x = 1$).

La v.c. binomiale viene definita anche come *v.c. delle prove ripetute* ed equivale allo schema di estrazione con ripetizione; essa rappresenta il numero di successi che si verificano in una sequenza di n prove "indipendenti" nelle quali è costante la probabilità di successo p.

La *funzione di probabilità* che in n prove si verifichi x volte l'evento dipende da due parametri (n, p) ed è data da:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \qquad (6.10)$$

Dove:

$P(X)$	probabilità di ottenere X successi dati n e p
n	ampiezza del campione

7
Campionamento statistico

7.1 Descrizione del campione

Le indagini statistiche possono essere totali o parziali (indagini campionarie). La probabilità di ricorrere ad una analisi fondata solamente su una parte dell'intera popolazione, invece che sul totale, risale ai primi decenni del secolo scorso, quando alcuni studiosi (Kiaer, Bowley e Jensen, Hansen e Hurwitz, Cochran) gettarono le basi per la teoria dei campioni distinguendo i campioni a scelta ragionata dai campioni probabilistici, in cui la selezione delle unità da inserire è esclusivamente affidata al caso.

Un *campione statistico* è costituito da un numero ridotto di unità della popolazione oggetto di studio, estratte con criteri tali da rappresentare in maniera fedele le caratteristiche del collettivo totale per consentirne così la generalizzazione dei risultati ottenuti all'intera popolazione.

Si studiano i campioni al posto delle popolazioni al fine di ridurre i costi e i tempi di rilevazione e per la minore complessità organizzativa. Le rilevazioni campionarie si distinguono in:

- *rappresentative* quando le unità del campione vengono selezionate dalla popolazione con un metodo statistico tale da presentare il fenomeno collettivo in scala ridotta;
- *non rappresentative*, se le unità del campione vengono individuate in modo non dipendente dal caso ma soltanto in base a criteri che dipendono dal problema oggetto di studio o dalla convenienza economica.

Un piano di campionamento è la definizione di una procedura di selezione delle n unità statistiche che comporranno il campione, mediante l'estrazione da una popolazione finita costituita da N unità. I principali piani di campionamento variano in funzione del tipo di scelta che si effettua per la determinazione del campione:

- *campionamento probabilistico*, se le unità del campione vengono estratte in maniera casuale e ciascuna di esse possiede la stessa probabilità di essere inclusa nel campione

- *campionamento non probabilistico*, se le unità prescelte vengono selezionate con un procedimento fondato sulla conoscenza della popolazione da parte dell'operatore.

A seconda del tipo di campionamento scelto vengono definiti diversi tipi di piani di campionamento.

7.2 Campionamento probabilistico

I piani di campionamento probabilistico vengono anche definiti come campioni a scelta casuale.

7.2.1 Campionamento casuale semplice con ripetizione

Il campionamento casuale semplice è il metodo più elementare e chiaro nel quale tutti i soggetti della popolazione hanno la stessa probabilità di essere inclusi nel campione. Esso consiste nella estrazione di un campione di n unità da una popolazione di N unità tutte identiche per forma, dimensione e peso. Trattandosi di procedura con ripetizione si avrà che l'unità selezionata verrà di volta in volta reinserita nella popolazione di partenza e, quindi, potrà essere nuovamente estratta, lasciando inoltre la probabilità di estrazione di un elemento sempre costante e pari a $P_i = \frac{1}{N^n}$. Tale metodo consente numerosi vantaggi ed è anche molto utilizzato a patto di poter disporre della lista di tutte le unità della popolazione, che nel caso di popolazioni umane vuol dire un elenco anagrafico di tutti i soggetti. Ovviamente le basi di dati devono essere sia precise che complete, senza alcuna eccezione.

L'estrazione del campione viene fatta in genere numerando tutte le unità presenti sulla lista ed estraendo, come nel gioco della tombola, dei numeri da un'urna. Si può ricorrere alle tavole dei numeri casuali leggendo a caso un elemento posto all'incrocio tra una riga e una colonna e proseguendo così fino a esaurire tutti gli elementi richiesti. In Excel si può utilizzare il generatore di numeri casuali come nell'esempio seguente.

Esempio 7.1

Utilizzando le tavole dei numeri casuali si estraggano 50 soggetti da un elenco che ne contiene 815. Dopo aver numerato progressivamente l'elenco dei soggetti da 1 a 815, basterà creare o scegliere 50 numeri casuali.

Proposta di soluzione in Excel

Dopo aver numerato progressivamente l'elenco dei soggetti da 1 a 815, in un nuovo foglio di lavoro scrivere nella cella A1 la formula `=CASUALE.TRA(1; 815)` che genera un numero intero casuale compreso tra i limiti passati come parametro. Trascinare questa formula nelle restanti 49 celle in basso per ottenere 50 numeri casuali. Ogni volta che si preme il tasto F9 l'intera sequenza di numeri verrà ricalcolata.

Un altro modo consiste nell'utilizzare il comando `Dati/Analisi dati` e scegliere nella finestra che si aprirà la riga `Generazione di un numero casuale`. Comparirà la finestra di Figura 7.1 dove sono stati poste nella prima e seconda riga i valori 5 e 10 per indicare la necessità di ottenere 5 colonne di valori (`numero di variabili`) per 10 valori ciascuna. Si sceglie il tipo di distribuzione `Uniforme` (equiprobabile), si assegnano i limiti entro i quali saranno generati i numeri e infine si assegna la prima riga di caselle dove andrà posto l'output. Il risultato apparirà come in Figura 7.2.

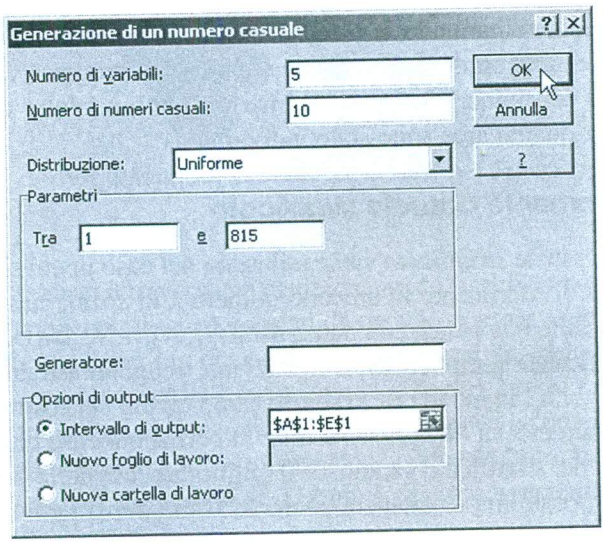

Figura 7.1 La finestra del comando `Dati/Analisi dati/Generazione di un numero casuale`.

	A	B	C	D	E
1	357	285	435	623	496
2	242	472	134	174	694
3	395	578	698	727	29
4	253	426	674	340	522
5	792	426	72	549	623
6	297	169	408	351	704
7	345	69	668	598	250
8	96	142	661	148	509
9	755	615	540	387	653
10	743	554	378	155	224
11					

Figura 7.2 Numeri casuali compresi nell'intervallo 1 – 815. □

7.2.2 Campionamento casuale semplice senza ripetizione

Il campionamento casuale semplice *senza ripetizione* consiste nella estrazione in blocco di un campione di n unità, tutte differenti tra di loro, da una popolazione di N unità. Trattandosi di procedura *senza ripetizione* si avrà che l'unità selezionata non verrà reinserita nella popolazione di partenza; l'universo campionario sarà pari a $\binom{N}{n}$ e la probabilità di estrazione varierà a seconda dell'ordine di estrazione e sarà pari a $P_i = \frac{1}{\binom{N}{n}}$.

Un esempio di campionamento casuale senza ripetizione è il campionamento casuale *per area* ottenuto ripartendo le aree territoriali in tante piccole aree attraverso le carte topografiche e numerando le diverse aree in maniera progressiva. Basterà, poi, scegliere casualmente dei numeri (senza reinserirli) da un bussolotto contenente tanti foglietti numerati per quante aree sono state individuate.

7.2.3 Campionamento casuale stratificato

Il campionamento casuale *stratificato* viene utilizzato nel caso in cui sia possibile suddividere la popolazione di riferimento in un certo numero k di *strati*, ciascuno composto con elementi il più possibile omogenei, con numerosità $N_1, N_2,\ldots N_k$. Da ogni strato, attraverso il campionamento casuale semplice senza ripetizione, vengono estratti i relativi campioni di numerosità $n_1, n_2,\ldots n_k$, in modo che lo spazio campionario sia costituito dalla somma di tutti i campioni di numerosità $n_1, n_2,\ldots n_k$, estratti in corrispondenza dei k strati.

Il campionamento stratificato garantisce il miglioramento delle stime (se gli strati sono ben scelti) e la possibilità di ottenere anche le stime dei singoli sottogruppi o strati. Scegliendo sottocampioni omogenei per ogni strato si aumenta l'efficienza giacché ogni singolo sottocampione diminuirà la propria variabilità. Il principale difetto che possiede è quello di essere molto oneroso in termini di tempo impiegato e di risorse necessarie.

Volendo realizzare una indagine sul costo medio sostenuto da una famiglia per il sostentamento dei propri figli in età scolare sarà opportuno individuare il numero di famiglie con figli e ripartirli in base al numero di figli. Ovvero, avremo un primo strato costituito dalle famiglie con un figlio, un secondo strato dalle famiglie con due figli, un terzo strato dalle famiglie con tre figli, ecc. A tal punto sceglieremo un campione da ogni famiglia in maniera da garantire la presenza nel campione di ogni tipologia familiare.

Esempio 7.2

In una indagine statistica mirata a valutare la percezione dei problemi ambientali da parte della comunità locale residente nel Parco Nazionale del Gargano si decide di effettuare un campionamento della popolazione. Il territorio si estende secondo Figura 7.3, rappresentativa della densità abitativa dei diversi comuni compresi nell'area territoriale del Parco, sulla base dei dati del 14° Censimento della Popolazione e delle Abitazioni del 2001:

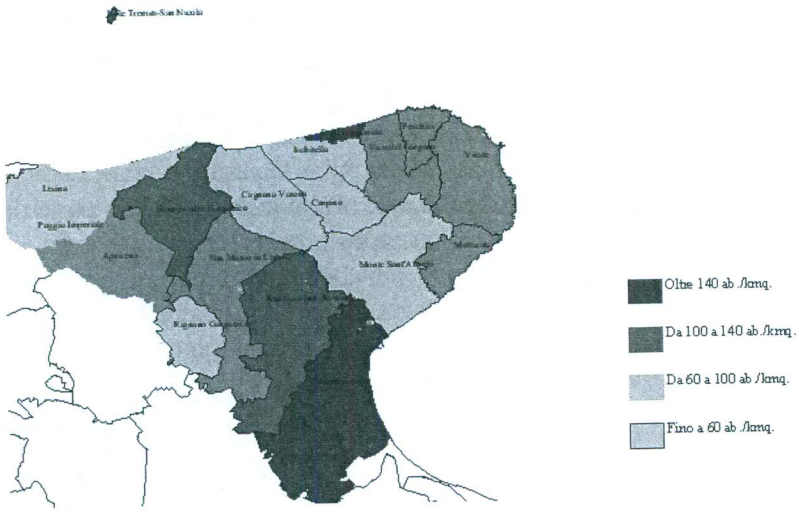

Figura 7.3 Densità demografica dei comuni compresi nel Parco del Gargano, Censimento 2001.

L'indagine è stata effettuata attraverso un campionamento stratificato dei comuni del parco su 7 strati, ritenuti omogenei, di cui 5 formati da un solo comune autorappresentativo per dimensioni demografiche o per ragioni dettate dalla vulnerabilità del sistema ecologico di questo arcipelago, come nel caso delle Isole Tremiti. Gli altri 2 strati raggruppano i comuni sulla base delle caratteristiche geografiche: comuni litoranei e comuni collinari.

Inoltre, è stata effettuata una sub stratificazione per sesso (maschi, femmine). Il campione individuato è, quindi, composto da 317 unità ripartite a seconda del sesso e dei diversi strati come nella Tabella 7.1:

Tabella 7.1 Distribuzione della popolazione per strati.

Strati	*Comuni*	*N. intervistati*		
		F	*M*	*Tot.*
1	Manfredonia	44	44	84
2	San Giovanni Rotondo	19	19	38
3	San Marco in Lamis	13	12	25
4	Sannicandro Garganico	14	14	28
5	Isole Tremiti – San Nicola	2	2	4
6	Lesina, Rodi Garganico, Peschici, Vieste, Mattinata	25	25	50
7	Cagnano Varano, Vico del Gargano, Carpino, Ischitella, Monte Sant'Angelo, Serra Capriola, Rignano Garganico, Apricena	44	44	88

□

7.2.4 Campionamento casuale a grappolo

Il campionamento casuale *a grappolo* viene utilizzato nel caso in cui la popolazione sia suddivisa in sottoinsiemi di unità naturali o artificiali (reparti ospedalieri, classi di studenti, abitanti di un quartiere, ecc.) che possiedono caratteristiche molteplici. Consiste nell'estrazione, con campionamento casuale semplice senza ripetizione, di n grappoli tra gli N possibili che costituiscono la popolazione. Tutte le unità facenti parte del grappolo prescelto fanno parte del campione. Presupposto necessario è che vi sia la massima omogeneità tra i diversi grappoli ed eterogeneità all'interno di ogni grappolo. Rispetto al campionamento stratificato, nel quale vengono incluse solo alcune unità statistiche appartenenti a tutti gli strati (omogenei al loro interno), nel campionamento a grappolo vengono incluse tutte le unità statistiche appartenenti ai soli grappoli prescelti (eterogenei al loro interno). L'oggetto del campionamento non è, quindi, la singola unità statistica, ma il grappolo.

Per esempio, in una indagine sulla soddisfazione del servizio effettuato da una azienda municipalizzata di un comune del Centro Italia si decide di effettuare un campionamento a grappolo suddividendo la popolazione in quartieri (se si è certi che tali quartieri siano omogenei tra loro) e scegliendo tutti gli abitanti residenti solo in alcuni dei quartieri. In tal modo avremo dei grappoli omogenei tra di loro ed eterogenei al loro interno.

7.2.5 Campionamento casuale a due stadi

Il campionamento casuale *a due stadi* risulta essere molto simile al campionamento a grappolo, in quanto sostituisce la fase finale della rilevazione totale delle unità all'interno dei grappoli con una ulteriore estrazione casuale, creando così un secondo stadio di campionamento.

In sintesi il campionamento si distingue in due fasi. Nella prima si opera suddividendo la popolazione su più livelli gerarchici (grappoli di una popolazione) ed effettuando una selezione di un certo numero di grappoli con estrazione casuale senza ripetizione. Successivamente da ciascuno dei grappoli selezionati si estrae un certo numero di unità campionarie secondo un piano di campionamento prescelto.

Per esempio, se si vuole condurre una indagine sui pazienti di un certo tipo di ambulatori, partendo dal livello regionale, si considerano tutte le ASL e le aziende ospedaliere dove si effettua una prima estrazione per includerne solo alcune all'interno del campione. Successivamente su ciascuna delle ASL scelte si effettua una seconda estrazione sulla base dei servizi presenti e, infine, si selezionano i pazienti del singolo servizio.

7.2.6 Campionamento sistematico

Molta importanza assume il metodo di campionamento sistematico che è una procedura del tutto equivalente a quella del campionamento casuale semplice. Consiste nell'estrarre da un popolazione n elementi in modo sistematico semplicemente scorrendo la

lista dei soggetti e selezionandone uno ogni intervallo $K = \frac{1}{f}$. L'interesse per tale metodo non risiede nel fatto di evitare il ricorso a numeri casuali o a estrazione di bigliettini da un'urna quanto al fatto che si può non conoscere l'elenco dei soggetti da intervistare o la numerosità n del campione. Se si devono intervistare i clienti di un centro commerciale basterà porsi davanti agli ingressi e intervistare, poniamo, 1 ogni 20 clienti che escono. È il metodo impiegato negli *Exit polls* quando vengono realizzate le interviste durante lo svolgimento delle elezioni. Questa procedura per generare un campione probabilistico deve rispettare due vincoli: il primo è che tutta la popolazione deve possedere la stessa probabilità di essere estratta e quindi nel caso delle interviste fuori da centro commerciale si deve operare lungo tutto il periodo di apertura, senza eccezioni, e in secondo luogo viene intervistata una persona ogni 20, senza alcuna eccezione altrimenti si rischia di perdere i benefici della casualità.

7.3 Campionamento non probabilistico

I piani di campionamento non probabilistico vengono definiti sovente come campioni *a scelta ragionata.* Tra di essi possiamo citarne i più comuni.

7.3.1 Campionamento a scelta ragionata o di giudizio

Il campionamento *a scelta ragionata* viene utilizzato nel caso in cui si intenda svolgere un'indagine localizzata su poche unità territoriali rappresentative e si voglia evitare che l'oscillazione casuale delle risposte possa allontanarsi troppo dalla popolazione. In tal caso le unità statistiche vengono individuate mediante criteri razionali di rappresentatività e autorevolezza, basati sulla conoscenza specialistica, da parte dell'operatore, del fenomeno nel suo complesso. Per esempio in un'indagine su determinate caratteristiche della popolazione di un comune si possono considerare solo alcuni quartieri (uno della periferia e uno del centro) e la scelta viene effettuata dal ricercatore sulla base delle proprie conoscenze della struttura sociale. Un'altra possibilità si verifica quando vengono intervistate poche persone, purché testimoni del fenomeno da considerare, che per la propria conoscenza e competenza possono esprimere un giudizio di merito approfondito e motivato.

Per esempio, una grande impresa vuole realizzare un nuovo stabilimento produttivo ed ha la possibilità di scegliere dove localizzarlo a seconda di tre possibili province. Può essere svolta un'indagine sulle caratteristiche socio economiche del territorio intervistando i sindaci dei comuni interessati, il presidente della Provincia e il capo dell'opposizione nel consiglio provinciale, il presidente della Camera di Commercio, dirigenti delle banche che operano in quei territori, imprenditori ed esponenti di istituzioni, ecc. Si parla in questi casi di campioni per testimoni privilegiati.

7.3.2 Campionamento per quote

Il campionamento *per quote* viene utilizzato quando i soggetti vengono estratti in modo che il campione rispecchi delle proporzioni predefinite rispetto alla popolazione: la popolazione si stratifica in base ad alcune caratteristiche (zona geografica, sesso, età, reddito, titolo di studio, ecc.) indicando le quote relative a ogni raggruppamento. Finora il metodo di campionamento è identico a quanto visto per il campionamento stratificato: in questo caso i soggetti vengono però scelti direttamente e liberamente dal ricercatore e non selezionati a caso. È il caso più diffuso di campionamento nelle ricerche di mercato e nei sondaggi di opinione perché, utilizzando i soggetti provenienti da cerchie di conoscenti o in gruppi sociali omogenei si ottiene un notevole risparmio dei tempi di reperimento dei soggetti, grande risparmio nei costi e soprattutto non è necessario trovare gli elenchi di componenti della popolazione.

Illustriamo quanto detto con un esempio. Dai dati del censimento si ricava la composizione percentuale della popolazione di un comune in base al genere (maschi 49% e femmine 51%), all'età (giovani 24%, adulti 35% ed anziani 41%) e ad altre caratteristiche utili all'indagine quali per esempio il tipo di attività lavorativa dipendente (impiegati, operai). Per quanto riguarda il tipo di lavoro dipendente la composizione percentuale sarà la seguente: per le donne (45% impiegate e 55% operaie) e per gli uomini (65% impiegati e 35% operai). Incrociando i dati si ottiene la Tabella 7.2 distinguendo il numero di maschi e di femmine e il tipo di lavoro dipendente. Per semplificare il calcolo poniamo che il campione abbia una dimensione di 100 unità; esso dovrà essere composto da 49 maschi (di cui 32 impiegati e 17 operai) e 51 donne (23 impiegate e 28 operaie). La ricerca dei soggetti da inserire nel campione si ottiene andando a cercare le persone in modo libero o in base alle conoscenze personali, con l'unico vincolo di rispettare *esattamente* il numero previsto per ogni quota.

Tabella 7.2 Distribuzione della popolazione per quote.

Maschi × 100			
	Giovani	**Adulti**	**Anziani**
Impiegati	9	10	13
Operai	3	6	8
Femmine × 100			
Impiegati	4	9	10
Operai	8	10	10

7.3.3 Campionamento a valanga

Il campionamento *a valanga* viene utilizzato nelle interviste telefoniche quando vengono prescelte poche unità con una determinata caratteristica, che a loro volta individuano "a catena" altre unità aventi la stessa caratteristica.

Molte popolazioni oggetto di interesse nelle indagini statistiche sono *nascoste* o clandestine. Si pensi per esempio agli evasori delle imposte (in Italia si stima che il 20-

lista dei soggetti e selezionandone uno ogni intervallo $K = \frac{1}{f}$. L'interesse per tale metodo non risiede nel fatto di evitare il ricorso a numeri casuali o a estrazione di bigliettini da un'urna quanto al fatto che si può non conoscere l'elenco dei soggetti da intervistare o la numerosità n del campione. Se si devono intervistare i clienti di un centro commerciale basterà porsi davanti agli ingressi e intervistare, poniamo, 1 ogni 20 clienti che escono. È il metodo impiegato negli *Exit polls* quando vengono realizzate le interviste durante lo svolgimento delle elezioni. Questa procedura per generare un campione probabilistico deve rispettare due vincoli: il primo è che tutta la popolazione deve possedere la stessa probabilità di essere estratta e quindi nel caso delle interviste fuori da centro commerciale si deve operare lungo tutto il periodo di apertura, senza eccezioni, e in secondo luogo viene intervistata una persona ogni 20, senza alcuna eccezione altrimenti si rischia di perdere i benefici della casualità.

7.3 Campionamento non probabilistico

I piani di campionamento non probabilistico vengono definiti sovente come campioni *a scelta ragionata.* Tra di essi possiamo citarne i più comuni.

7.3.1 Campionamento a scelta ragionata o di giudizio

Il campionamento *a scelta ragionata* viene utilizzato nel caso in cui si intenda svolgere un'indagine localizzata su poche unità territoriali rappresentative e si voglia evitare che l'oscillazione casuale delle risposte possa allontanarsi troppo dalla popolazione. In tal caso le unità statistiche vengono individuate mediante criteri razionali di rappresentatività e autorevolezza, basati sulla conoscenza specialistica, da parte dell'operatore, del fenomeno nel suo complesso. Per esempio in un'indagine su determinate caratteristiche della popolazione di un comune si possono considerare solo alcuni quartieri (uno della periferia e uno del centro) e la scelta viene effettuata dal ricercatore sulla base delle proprie conoscenze della struttura sociale. Un'altra possibilità si verifica quando vengono intervistate poche persone, purché testimoni del fenomeno da considerare, che per la propria conoscenza e competenza possono esprimere un giudizio di merito approfondito e motivato.

Per esempio, una grande impresa vuole realizzare un nuovo stabilimento produttivo ed ha la possibilità di scegliere dove localizzarlo a seconda di tre possibili province. Può essere svolta un'indagine sulle caratteristiche socio economiche del territorio intervistando i sindaci dei comuni interessati, il presidente della Provincia e il capo dell'opposizione nel consiglio provinciale, il presidente della Camera di Commercio, dirigenti delle banche che operano in quei territori, imprenditori ed esponenti di istituzioni, ecc. Si parla in questi casi di campioni per testimoni privilegiati.

di rappresentatività legati a profili psicografici chiamati stili di vita. Si tratta di criteri che consentono una classificazione legata a fattori psicologici e socio-culturali, che integrano i classici criteri geografici o demografici.

Appare, quindi, evidente come il piano di campionamento prescelto sia di tipo non probabilistico (ragionato o per quota) in quanto le famiglie vengono prescelte in modo che il campione complessivo rispecchi le proporzioni predefinite.

I dati forniti dall'Auditel riguardano, nello specifico:

- *Share,* ovvero il rapporto percentuale tra gli ascoltatori di una certa emittente e il totale degli ascoltatori che stanno guardando la televisione sulle diverse reti
- *Audience*, ovvero il numero medio dei telespettatori di un certo programma. dato dal rapporto tra la sommatoria dei telespettatori presenti in ciascun minuto di un dato intervallo di tempo e la durata in minuti dell'intervallo stesso.
- *Percentuale di penetrazione,* ovvero il rapporto percentuale tra gli ascoltatori (di una data categoria) e il loro universo statistico di riferimento (ISTAT). Cioè: quanti vedono quel tale programma rispetto al totale della popolazione che comprende anche quelli che non vedono la TV.
- *Percentuale di permanenza*, è un indicatore della "fedeltà" di visione. Cioè il rapporto percentuale tra il numero di minuti visti mediamente dagli ascoltatori di un certo programma e la durata dello stesso.

Esempio 7.3

Si riporta nella Figura 7.4 un esempio di rilevazione dei dati campionari effettuata nell'arco di un mese, in merito alle diverse componenti di interesse per la misurazione dell'auditel di reti nazionali o locali.

	A	B	C	D	E	F	G	H	I	J	K
1	AUDITEL										
2	*Giorno tipo medio mensile -- Fasce standard*										
3	Febbraio 2007 (dal 04/02/07 al 03/03/07) Totale individui (56276)(.000)										
4	*Fasce orarie*	Dalle:	02.00	07.00	09.00	12.00	15.00	18.00	20.30	22.30	
5		Alle:	25.59	09.00	12.00	15.00	18.00	20.30	22.30	25.59	
6	TV 1	am	2453	1427	1175	2960	2080	5157	6837	2812	
7		%sh	24.15	29.83	24.07	20.06	19.20	28.36	25.66	25.64	
8		%pe	4.36	2.54	2.09	5.26	3.70	9.16	12.15	5.00	
10	TV 2	am	1019	597	690	2023	1352	1155	2356	875	
11		%sh	10.03	12.48	14.13	13.71	12.48	6.35	8.84	7.98	
12		%pe	1.81	1.06	1.23	3.59	2.40	2.05	4.19	1.55	
14	TV 3	am	913	224	424	1291	796	1947	2784	912	
15		%sh	8.99	4.68	8.69	8.75	7.35	10.71	10.45	8.32	
16		%pe	1.62	0.40	0.75	2.29	1.41	3.46	4.95	1.62	
18	TV 4	am	4386	2248	2288	6274	4227	8259	11977	4599	
19		%sh	43.19	47.00	46.87	42.52	39.03	45.42	44.95	41.93	
20		%pe	7.79	3.99	4.07	11.15	7.51	14.68	21.28	8.17	
21	Legenda: am = ascolto medio, %sh = % share, %pe = % penetrazione										
22	L'ora 25,59 indica le due di notte meno 1 minuto.										
23	Rilevazioni ed elaborazioni AGB Nielsen Media Research										
24											

Figura 7.4 Esempio di rilevazione con dati auditel. □

7.4.2 Sondaggi di opinione

I sondaggi di opinione vengono effettuati frequentemente per la rilevazione delle opinioni di una collettività su argomenti e questioni diversificate, quali per esempio l'attualità, il supporto alle decisioni da parte delle istituzioni, la verifica della qualità di un servizio offerto, ecc.

Tre i principali campi di attività vi sono:

- inchieste sociali sull'opinione pubblica, cioè studi e analisi di fatti ed eventi volti a sondare gli atteggiamenti dell'opinione pubblica;
- eventi di attualità e di costume, per i quali occorre operare in modo veloce e tempestivo per registrare le posizioni dell'opinione pubblica;
- rilevazione delle abitudini di lettura di quotidiani e riviste al fine di mettere a punto un prodotto qualitativamente competitivo.

Essi vengono effettuati seguendo differenti criteri di campionamento che devono obbligatoriamente essere resi noti al pubblico al fine di consentire una valutazione oggettiva dei risultati. Con riferimento alle tecniche di campionamento, quella maggiormente adottata risulta essere il campionamento non probabilistico "per quota". I sondaggi d'opinione vengono generalmente svolti con metodologia telefonica C.A.T.I. oppure attraverso il panel di famiglie.

A scopo esemplificativo si illustrano degli esempi tratti dal sito dell'autorità Garante delle Comunicazioni (www.agcom.it) o sul sito del Centro Studi Investimenti Sociali - Censis (www.censis.it) .

Esempio 7.4

SONDAGGIO Censis "reagire a un mondo difficile. Giovani e giovani con sclerosi multipla: due punti di vista sul futuro"

DOCUMENTO INFORMATIVO COMPLETO

(in ottemperanza al regolamento dell'Autorità per le Garanzie nelle Comunicazioni in materia di pubblicazione e diffusione dei sondaggi sui mezzi di comunicazione di massa: delibera 153/02/CSP, allegato A, art. 3, pubblicato su G.U. 185 del 8/8/2002)

Autore: Censis – Codres

Committente e Acquirente: AISM

Tipo di rilevazione: sondaggio di opinione

Oggetto del sondaggio: Opinioni, comportamenti e aspettative sul futuro dei giovani malati di sclerosi multipla

Universo di riferimento: popolazione italiana giovanile (tra 18 e 40 anni)

Tipo di campione: Campionamento casuale a due stadi con stratificazione proporzionale. Le unità del primo stadio sono costituite dai comuni, scelti in funzione dell'area geo-

grafica e dell'ampiezza demografica; quelle del secondo stadio dai giovani stratificati per sesso e classe di età.

Estensione territoriale: nazionale

Data di realizzazione sondaggio: 20/1/2005 - 11/2/2005

Metodologia di raccolta delle informazioni: interviste telefoniche mediante metodologia CATI (*Computer Assisted Personal Interviewed*)

Verifica della coerenza delle risposte alle diverse domande: l'utilizzo del sistema CATI garantisce l'affidabilità dei risultati e la rapidità dei tempi di elaborazione, grazie al salvataggio automatico delle risposte su supporto informatico e alla possibilità di verifiche automatiche

Numerosita' campionaria: 992 casi

Rappresentatività dei risultati: il margine di errore relativo ai risultati del sondaggio sul totale dei casi, al livello di significatività del 95%, è compreso fra ± 3,1%

Numero di contatti: interviste complete n=992 (18% sul totale contatti, pari a 5.450), cadute per fuori quota/rifiuti/non eleggibili 4.464

La documentazione completa è disponile sul sito www.agcom.it □

7.4.3 Proiezioni elettorali

Le proiezioni elettorali vengono utilizzate per le rilevazioni sui temi politici e istituzionali. In particolare esse trovano larga applicazione in occasione di elezioni politiche o referendum per la rilevazione di due tipologie di indagini:

- *exit-poll*, ovvero previsioni di voto rilevate sulla base di interviste realizzate su un campione di elettori all'uscita dalle sezioni elettorali;
- *proiezioni elettorali*, ovvero previsioni di voto rilevate sulla base dello spoglio delle schede effettuato su un campione rappresentativo di seggi elettorali. Tale previsione varia a seconda del numero di seggi rilevati e tende al valore reale via via che i dati relativi ai seggi aumentano.

L'*Exit Poll* è, quindi, un sondaggio realizzato all'uscita dei seggi elettorali (scuole e/o uffici preposti). L'intervistatore chiede agli elettori che hanno appena votato di replicare, in segreto, il voto appena espresso all'interno di un fac-simile della scheda elettorale da riporre in un'apposita urna. Gli elettori che hanno appena votato sono contattati seguendo un semplice ma importantissimo principio: quello dell'intervallo prefissato (una persona ogni cinque), che garantisce la casualità del campione. I dati riportati dagli intervistati nei fac-simile di scheda sono comunicati dagli intervistatori tramite telefono alla sede operativa di Nexus secondo intervalli prestabiliti (generalmente ogni due ore). Giunti in sede i dati sono opportunamente elaborati per la consegna dei risultati finali. I risultati finali degli exit poll vengono comunicati istantaneamente al momento della chiusura dei seggi. I dati vengono tradizionalmente forniti all'interno di "forchette", o

intervalli, che definiscono il campo più probabile di esistenza del dato (per es. 52-54%). L'ampiezza delle forchette varia a seconda del grado di incertezza che grava sul dato, ma non è mai superiore ai + 2 punti percentuali.

Le *proiezioni*, a differenza degli exit poll, non sono un sondaggio. Esse si basano su una estrapolazione di risultati reali, tratti da sezioni campionarie, riportati all'universo delle sezioni di riferimento. Perciò nella fase pre-elettorale e negli Exit poll si parla di "voto di paglia" mentre per le proiezioni elettorali si parla di "voto di pietra". I rilevatori si presentano alla sezione loro assegnata circa mezz'ora prima della chiusura dei seggi in modo da poter rilevare il numero degli iscritti e degli elettori che hanno votato fino a quel momento, con l'apposito modulo di rilevazione dello spoglio. Al termine dello spoglio il rilevatore comunica alla sede operativa tre tipi di dati:

- il numero degli elettori iscritti e dei votanti effettivi;
- il numero delle schede bianche e nulle;
- il numero dei voti validi ottenuti da ciascuna lista e/o candidato.

I dati, giunti in sede, vengono trattati secondo consolidate tecniche statistiche, per fornire i risultati finali, tradizionalmente espressi in valori percentuali a un decimale (per es. 53,6%). I dati delle proiezioni elettorali possono essere forniti dal momento della disponibilità dei dati relativi al 5% delle sezioni scrutinate in quanto la soglia di errore associata non è superiore al + 5%. A partire dal 15% delle sezioni scrutinate i dati assumono maggiore stabilità avendo associato un intervallo di confidenza mediamente inferiore al + 3%. La soglia massima di errore tollerata con il 100% delle sezioni scrutinate è pari a + 1,5%.

Per analizzare le caratteristiche di un sondaggio politico è possibile sfogliare una qualunque rivista o quotidiano oppure consultare siti Internet. Tra i più accreditati vi sono:
www.sondaggipoliticoelettorali.it/
www.agcom.it/sondaggi/sondaggi_index.htm

Esempio 7.5

Consideriamo un esempio di Sondaggio Politico-Elettorale dal titolo: *Le attese degli Italiani dopo la Crisi di Governo*. Tale sondaggio è stato pubblicato in data 2/3/2007 con le seguenti caratteristiche di fondo:

Autore: IPR Marketing

Committente/ Acquirente: Repubblica.it

Criteri seguiti per la formazione del campione:

I questionari sono stati somministrati telefonicamente con l'ausilio del sistema Cati. L'elaborazione dei dati è avvenuta con il programma SPSS.

Il campione è stato disaggregato per sesso, età ed area di residenza in maniera proporzionale rispetto ai dati ufficiali della popolazione. Fonte Istat

Metodo di raccolta delle informazioni:

Metodo casuale con numeri telefonici estratti dall'elenco telefonico.

1.000 elettori, disaggregati per sesso, età ed area di residenza in maniera tale da essere rappresentativi di tutta la popolazione adulta residente

Data in cui è stato realizzato il sondaggio: 01/03/2007

QUESTIONARIO

Domanda:

Indipendentemente dalle sue idee politiche, per poter avere il suo apprezzamento, il Governo Prodi nei prossimi mesi dovrebbe:

(seguono risposte a scelta multipla)

Il sondaggio ha portato ai seguenti risultati riportati nella Figura 7.5. □

LE PRIORITA' DEL GOVERNO PRODI

TAV 1 INDIPENDENTEMENTE DALLE SUE IDEE POLITICHE, PER POTER AVERE IL SUO APPREZZAMENTO IL GOVERNO PRODI NEI PROSSIMI MESI DOVREBBE:

	TOTALE CAMPIONE 01-03-2007
RIFORMARE LA LEGGE ELETTORALE	
SI	81
NO	9
NON SA	10
TOTALE	100
CONFERMARE LA REALIZZAZIONE DELLA TAV TORINO -LIONE	
SI	62
NO	23
NON SA	15
TOTALE	100
CONTINUARE CON LE LIBERLIZZAZIONI AVVIATE DA BERSANI	
SI	55
NO	27
NON SA	18
TOTALE	100
APPROVAZIONE DICO IN PARLAMENTO	
SI	55
NO	34
NON SA	11
TOTALE	100
RIFINANZIARE LA MISSIONE DEI SOLDATI IN AFGHANISTAN	
SI	37
NO	52
NON SA	11
TOTALE	100
RIFORMARE LE PENSIONI PREVEDENDO L'AUMENTO DELL'ETÀ PENSIONABILE	
SI	15
NO	79
NON SA	6
TOTALE	100

FONTE: IPR Marketing

Figura 7.5 Esempio di Sondaggio Politico-Elettorale. □