

Lezioni di Statistica di base

(a cura di Giancarlo Ferrari)

Definizione di statistica

- Scienza che ha per oggetto lo studio di fenomeni collettivi suscettibili di misura e di descrizione quantitativa, specie quando il numero di individui interessato è talmente elevato da escludere la possibilità o la convenienza di seguire le vicende di ogni singolo individuo (Vocabolario delle Lingua Italiana Treccani ed.1995)
- Tecnica che ha come scopo la conoscenza quantitativa dei fenomeni collettivi (G. Leti- Istituzioni di statistica, 1985)

Sai che d'è la statistica? E' 'na cosa
Che serve pe' fa' un conto in generale
De la gente che nasce, che sta male,
che more, che va in carcere e che sposa.

Ma pe' me la statistica curiosa
È dove c'entra la percentuale,
pe' via che, lì, la media è sempre uguale
puro co' le persone bisognose.

Me spiego: da li conti che se fanno
Secondo le statistica d'adesso
Risurta che te tocca un pollo all'anno
e se nun entra nelle spese tue,
t'entra ne la statistica lo stesso
perché c'è n'antro che se ne magna due

(La statistica, Trilussa, 1914)

Introduzione

Lo studio dei fenomeni collettivi oggetto delle indagini statistiche può avere due obiettivi fondamentali: (i) DESCRIVERE un fenomeno e sintetizzarlo in funzione delle caratteristiche che se ne intendono studiare (in questo caso non vengono fatte particolari ipotesi sulla popolazione oggetto di studio); (ii) sottoporre a confronto gruppi (collettivi) di individui per studiarne le differenze oppure verificare delle ipotesi su una popolazione. Tale branca va sotto il nome di statistica INFERENZIALE.

Oggetto della indagine di tipo statistico può essere rappresentato da:

[] Popolazione: se il collettivo comprende tutte le unità omogenee rispetto ad una caratteristica comune;

[] Campione: se il collettivo in esame costituisce un sottoinsieme della popolazione di riferimento (l'utilizzo di un campione permette di ridurre i costi di una indagine statistica).

Nel primo caso l'indagine viene detta di tipo CENSUARIO mentre nel secondo caso si tratta di una indagine CAMPIONARIA.

Nota sul censimento: Il CENSIMENTO è la raccolta di dati su tutta la popolazione. In Italia viene effettuato ogni 10 anni (negli anni che terminano con 1 e quindi il prossimo sarà nel 2021). Viene definito CENSIMENTO DELLE POPOLAZIONI E DELLE ABITAZIONI e svolto con metodo classico mediante somministrazione di questionari auto-compilati. In particolare vengono raccolti dati su:

- ABITAZIONE e sue caratteristiche
- PERSONE DELLA FAMIGLIA
- PERSONE CHE NON ABITANO ABITUALMENTE nell'alloggio (ospiti occasionali, temporanei, presenti ma residenti altrove).

E' importante fare una distinzione tra la UNITA' di RILEVAZIONE (famiglia) e l'UNITA' STATISTICA (gli individui che compongono il nucleo familiare).

L'unità di rilevazione del censimento abitativo sono le abitazioni (abitate e non).

Nel corso del 15mo censimento del 2011 è stato possibile per la prima volta compilare il questionario via web.

1.1 Fasi di una indagine/ricerca statistica

In linea generale le fasi di una ricerca statistica è articolate nel seguente modo:

- (a) Definizione del problema che si intende affrontare;
- (b) Individuazione delle unità statistiche e di rilevazione;
- (c) Raccolta dati su ciascuna unità statistica;
- (d) Elaborazione e rappresentazione dei dati;
- (e) Conclusioni.

La statistica descrittiva si occupa di organizzare e sintetizzare le osservazioni in modo da riassumerne le caratteristiche generali. Ciò può essere fatto in vari modi mediante tabelle, grafici di varia natura e misure di sintesi, è necessario però definire prima che tipo di dati si hanno a disposizione. Inoltre i dati possono risultare da una raccolta diretta (DATI PRIMARI) oppure possono essere stati raccolti da altri. In questo caso si parla di DATI SECONDARI.

Come possono essere raccolti i dati?

- Predisposizione di questionari e loro utilizzo tramite interviste;
- Trasmissione di questionari per posta;
- Interviste telefoniche;
- Indagini di campo (ad esempio indagini campionarie con raccolta di campioni biologici).

Quelle appena descritte sono in genere indagini di tipo campionario nel corso delle quali vengono raccolti dati solo su una frazione della popolazione oggetto di indagine.

2.1 Variabili

Le caratteristiche che vengono studiate e che sono oggetto di raccolta dati sono dette VARIABILI. Le variabili sono espressione di caratteristiche degli individui.

L'altezza, la classe sociale, il numero di figli per famiglia, etc., sono tutte variabili che possono assumere valori o modalità diverse da individuo ad individuo.

Le variabili possono essere suddivise in due categorie ulteriormente scomponibili a seconda delle modalità che possono assumere.

2.1.1 Variabili di tipo qualitativo

Le variabili di tipo qualitativo sono anche chiamate categoriche poiché rappresentano una classificazione di caratteristiche. A loro volta possono essere suddivise in:

NOMINALI SCONNESSE: variabili le cui modalità non hanno un ordine (ad esempio gruppo sanguigno, colore dei capelli, sesso, etc..). Le variabili nominali che assumono uno di due distinti valori (ad esempio il sesso) sono denominate dicotomiche o binarie.

NOMINALI ORDINALI: le modalità sono ordinabili (ad esempio le persone possono essere classificate in base all'attività fisica: poco attive, attive, molto attive, etc..).

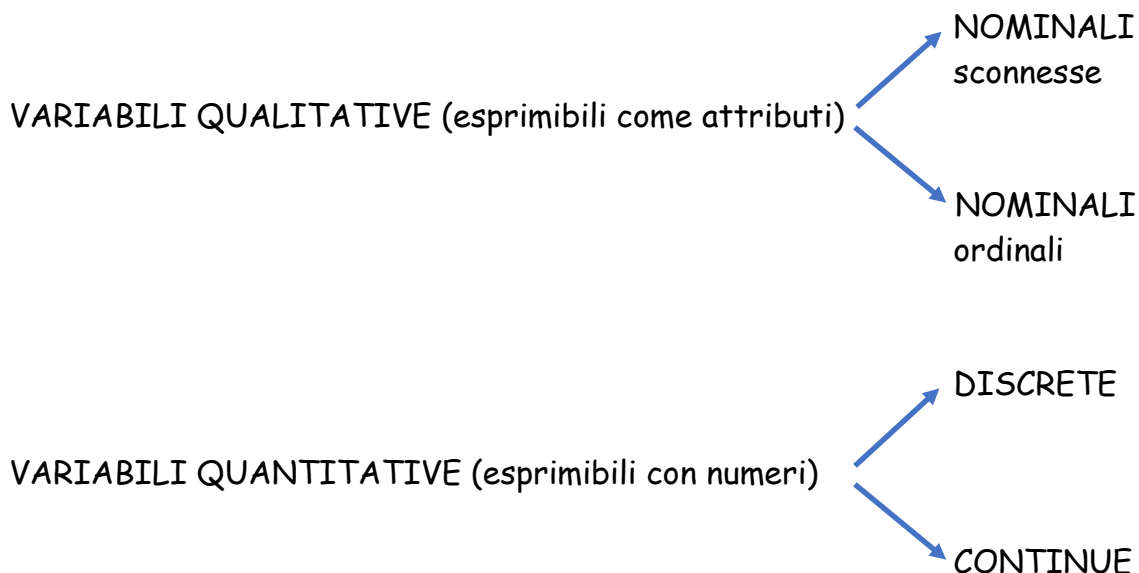
2.1.2 Variabili di tipo quantitativo

Le variabile quantitative assumono valori numerici (sono cioè misurabili) ed a seconda della modalità che il valore può assumere si possono classificare in:

DISCRETE: assumono soltanto valori interi (ad esempio numero di figli per famiglia, numero di stanze per abitazione, posti letto in una struttura ospedaliera, ecc...). Il risultato di una operazione aritmetica operato su dati discreti non necessariamente corrisponde ad un valore discreto (lo si vedrà quando verranno effettuate operazioni matematiche sui dati).

CONTINUE: possono assumere infinite modalità e la precisione con la quale una variabile quantitativa viene registrata dipende soltanto dallo strumento di misura.

Riassumendo:



IMPORTANTE: le categorie elencate hanno un ordine gerarchico di importanza crescente dalla nominale alla continua in quanto è sempre possibile passare da una all'altra di grado inferiore ma non è possibile fare il contrario.

Le tabelle statistiche e le rappresentazioni grafiche

2. Tabelle statistiche e grafici

2.1 Tabelle

I dati raccolti debbono in qualche modo essere organizzati, ed il primo passo in genere è la loro raccolta in una matrice di dati grezza che prende la seguente forma (si supponga che la variabile X possa assumere 3 modalità $[a, b, c]$ e che ogni individuo U non possa che essere classificato secondo una sola di tali modalità):

Tab. 2.1

X_i	U_j
X_a	U_1
X_a	U_2
X_c	U_3
X_b	U_4
X_a	U_5
X_b	U_6
X_b	U_7
X_a	U_8
X_c	U_9
X_c	U_{10}
X_c	U_{11}
X_a	U_{12}
X_a	U_{13}
X_b	U_{14}
X_b	U_{15}
X_c	U_{16}
X_a	U_{17}
X_a	U_{18}
X_a	U_{19}
X_b	U_{20}
X_c	U_{21}
X_a	U_{22}

Una prima organizzazione dei dati è costituita dalla loro organizzazione in una distribuzione di frequenza.

I dati della tabella 2.1 possono essere organizzati nel seguente semplice modo:

<i>Modalità</i>	<i>Frequenza</i>
X_a	10
X_b	6
X_c	6

La distribuzione di frequenza è una tabella che mostra i valori (le modalità) che possono essere assunti da una variabile e la frequenza con la quale ogni valore è stato osservato.

Nell'esempio che segue è mostrata una tabella che mostra la distribuzione di frequenza per una variabile qualitativa sconnessa.

Tabella 2.1 Numero di casi di AIDS osservati in un determinato periodo negli USA e nel Regno Unito classificati secondo classi di rischio:

Gruppi a rischio	Stati Uniti	Regno Unito
	Casi	Casi
Omo e/o bisessuali	28324	901
Tossicodipendenti	7109	15
Omosessuali tossicodipendenti	3224	17
Emofiliaci	430	60
Emotrasfusi	906	23
Rapporti eterosessuali	1618	37
Figli di donne HIV	599	12
Altri/non conosciuti	1323	2
Totale	43533	1067

(tratta da J.F. Osborn "Manuale di statistica medica", Soc. Ed. Universo - Roma)

La tabella successiva mostra invece una distribuzione di frequenza per una variabile di tipo qualitativo ordinale dove un collettivo di 150 persone ($n = 150$) viene classificato sulla base della attitudine al fumo che viene categorizzato secondo quattro modalità:

Tabella 2.2 (distribuzione di frequenza sulla base del livello di reddito)

Attitudine al fumo	Frequenza
NON FUMATORI	60
FUMATORI OCCASIONALI	15
FUMATORI MODERATI	45
FORTI FUMATORI	30

Nella tabella che segue viene invece rappresentata la distribuzione di frequenza di una variabile quantitativa discreta (numero di maschi in famiglie con otto bambini).

Tabella 2.3 Distribuzione di frequenza del numero di maschi in famiglie con 8 bambini

Numero di maschi	frequenza (n. di famiglie)
0	161
1	1152
2	3951
3	7603
4	10263
5	8498
6	4948
7	1655
8	264
Totale	38495

(tratta da J.F. Osborn "Manuale di statistica medica", Soc. Ed. Universo - Roma)

Nella tabella di frequenza precedente sono rappresentate 9 modalità (da 0 ad 8) e ciascuna modalità è rappresentata da un numero singolo corrispondente al numero esatto di figli per famiglia di cui se ne vuole studiare la distribuzione di frequenza.

In alcuni casi quando si costruisce una distribuzione di frequenza è necessario raggruppare le modalità in classi di intervallo.

Ciò è in realtà indispensabile per variabili quantitative di tipo continuo poiché sarebbe impossibile poter rappresentare la gamma di infiniti valori possibili che la variabile può assumere.

L'esempio che segue mostra una distribuzione di frequenza per una variabile quantitativa discreta le cui modalità sono state raggruppate in classi.

Tabella 2.4 Distribuzione di frequenza del numero di lesioni causate dal virus del vaiolo in membrane di uova

Numero delle lesioni	Frequenza (n. membrane)
0-	1
10-	6
20-	14
30-	14
40-	17
50-	8
60-	9
70-	3
80-	6
90-119	2
Totale	80

(tratta da J.F. Osborn "Manuale di statistica medica", Soc. Ed. Universo - Roma)

Il segno "-" a fianco di ciascuna classe ha un valore ben preciso e sta ad indicare "fino a ma non incluso il valore successivo". Pertanto l'intervallo 0- comprende tutte le osservazioni 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 ma non 10 che fa parte dell'intervallo successivo.

Come si vede dalla tabella ogni classe ha ampiezza uguale a 10 ad eccezione dell'ultima classe che ha ampiezza 30 (da 90 a 119).

In generale la simbologia adottata per definire qual'è la gamma di valori inclusi in intervalli di classe è la seguente:

Notazione	Estremo inferiore	Estremo superiore	Ampiezza
$x_i - x_{i+1}$	incluso	Escluso	$x_{i+1} - x_i$
$x_i - x_{i+1}$	escluso	Incluso	$x_{i+1} - x_i$
$x_i - x_{i+1}$	incluso	Incluso	$(x_{i+1} - x_i) + 1$

Il numero di classi di intervallo che si possono creare in una distribuzione di frequenza dipende dalla numerosità del campione e dalla convenienza. Se il numero di intervalli è esiguo non fornisce dettagli sufficienti.

In linea generale sono sufficienti da 7 a 20 intervalli.

Esiste una regola generale che può essere applicata chiamata Regola di Sturges dove il numero di classi K è dato dalla seguente formula:

$$K = 1 + (3,3 \times \log_{10} n)$$

Dove n rappresenta il totale delle frequenze.

Nel caso dei dati della tabella 2.3 si avrebbe:

$$K = 1 + (3,3 \times \log_{10} 80) = 1 + (3,3 \times 1,9) = 7,3$$

Vale a dire che nel caso della tabella 2.4 sarebbero stati sufficienti 7 classi senza perdita eccessiva di dettagli

E' utile talvolta conoscere la proporzione di valori che rientra in un determinato intervallo in una distribuzione di frequenza e non soltanto il numero assoluto.

La frequenza relativa di un intervallo è la percentuale del numero di osservazioni che appare nell'intervallo e si calcola dividendo il numero di osservazioni all'interno di un intervallo per il numero totale di osservazioni della tabella. Ad esempio (vedi tabella 2.5) la frequenza relativa dell'intervallo 0 - 9 si è ottenuta da $1/80 \times 100 = 1,25\%$.

Tabella 2.5 - Frequenze relative e cumulative delle lesioni prodotte dal virus del vaiolo su uova embrionate.

Classe (x)	Frequenza (f)	Frequenza relativa %	Frequenza cumulativa %
0 - 9	1	1,25%	1,25%
10 - 19	6	7,50%	8,75%
20 - 29	14	17,50%	26,25%
30 - 39	14	17,50%	43,75%
40 - 49	17	21,25%	65,00%
50 - 59	8	10,00%	75,00%
60 - 69	9	11,25%	86,25%
70 - 79	3	3,75%	90,00%
80 - 89	6	7,50%	97,50%
90 - 99	1	1,25%	98,75%
100 - 109	0	0,00%	98,75%
110 - 119	1	1,25%	100,00%

La frequenza relativa cumulativa di un intervallo (tabella 2.5) è la percentuale del numero totale di osservazioni che hanno un valore inferiore o uguale al limite superiore dell'intervallo stesso e si calcola sommando alle frequenze relative dell'intervallo le frequenze relative di tutti gli intervalli precedenti.

Ad esempio la frequenza cumulativa nell'intervallo 30-39 pari al 43,75% è data dalla somma della frequenza relativa dell'intervallo 0-9, 10-19, 20-29 e 30-39. Più semplicemente la frequenza cumulativa nell'intervallo 30-39 è data dalla somma tra la

sua frequenza relativa (17,50%) e la frequenza cumulativa dell'intervallo precedente (20-29 che è uguale a 26,25%).

La tabella così organizzata permette di iniziare a compiere una prima serie di osservazioni. La frequenza maggiore compare in corrispondenza dell'intervallo di classe 40-49 (21,25%) ed inoltre in corrispondenza di tale classe sono già rappresentate il 65% delle frequenze totali.

La rappresentazione per intervalli di classe comporta comunque una perdita di informazione data dal fatto che non siamo in grado di sapere la distribuzione dei valori all'interno di ciascuna classe. In teoria la classe 40-49 potrebbe essere rappresentata da un unico valore che si ripete 17 volte (tante quante le frequenze assolute nell'intervallo). L'assunzione che viene fatta quando si utilizzano gli intervalli di classe è che mediamente i valori sono pari al valore centrale della classe stessa. Nel caso della classe 40-49 significa che si assume che la media dei 17 valori sia pari a 45 (valore centrale della classe).

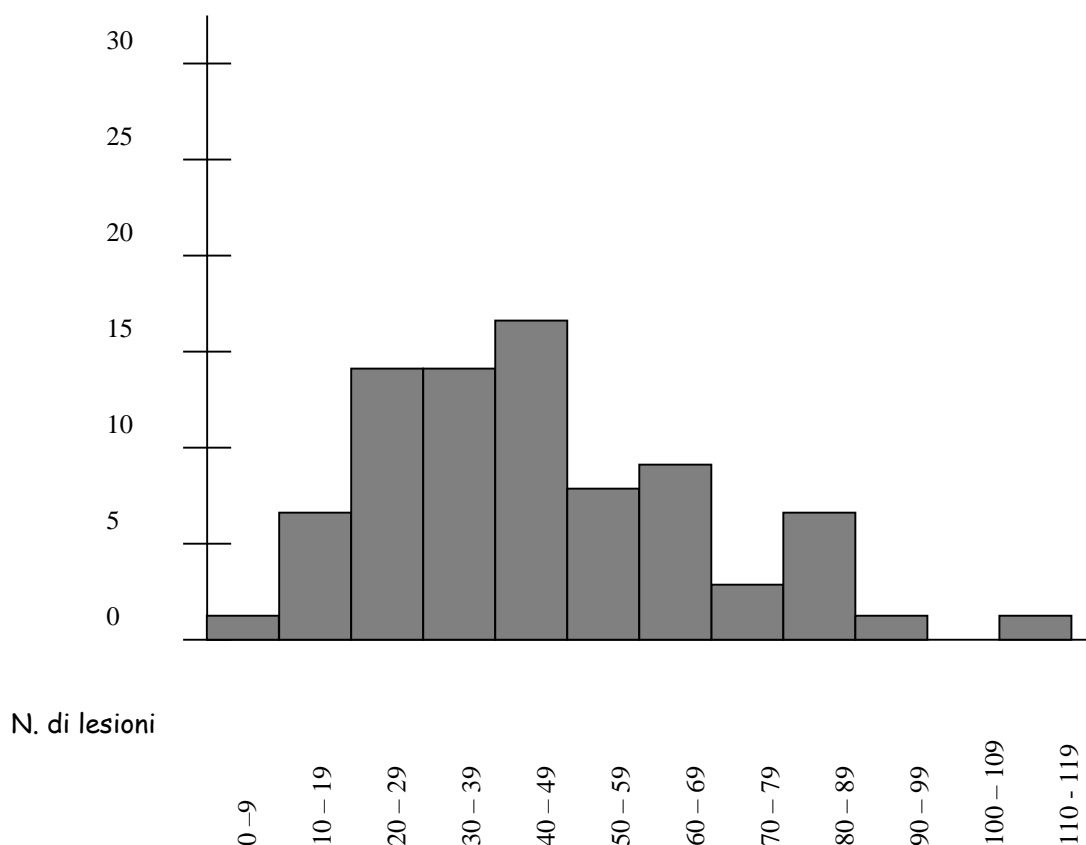
2.2 Grafici

Le distribuzioni di frequenza possono essere rappresentate attraverso la costruzione di istogrammi, diagrammi circolari, etc., la rappresentazione grafica che se ne può ottenere è in funzione del tipo di dati di cui si dispone.

2.2.1 Istogrammi

Nell'esempio che segue è rappresentato l'istogramma della distribuzione di frequenza del numero di lesioni causate dal virus del vaiolo su membrane di uova della tabella 2.4.

Fig. 2.1



La costruzione degli istogrammi nel caso gli intervalli siano tutti uguali (come nell'esempio precedente) è semplice e le ordinate corrispondono alle frequenze percentuali o assolute e la somma (l'area complessiva) è pari al 100% dei valori della distribuzione.

Nel caso in cui l'ampiezza delle classi non è uguale e si rappresentasse la distribuzione assumendo le altezze pari alle frequenze relative si otterrebbe un istogramma fuorviante.

Si considerino i dati della tabella 2.5 e si immagini di aggregare le classi 60-89 (creando un nuovo intervallo di ampiezza 30).

Tabella 2.6 - Frequenze relative e cumulative delle lesioni prodotte dal virus del vaiolo su uova embrionate con intervalli di classe diversi.

Classe (x)	Frequenza (f)	Frequenza relativa %	Frequenza cumulativa %
0 - 9	1	1,25%	1,25%
10 - 19	6	7,50%	8,75%
20 - 29	14	17,50%	26,25%
30 - 39	14	17,50%	43,75%
40 - 49	17	21,25%	65,00%

50 - 59	8	10,00%	75,00%
60 - 89	18	22,50%	97,50%
90 - 99	1	1,25%	98,75%
100 - 109	0	0,00%	98,75%
110 - 119	1	1,25%	100,00%

In questo caso se l'altezza della ordinata fosse posta uguale alle frequenze relative non considerando che c'è un intervallo di classe di ampiezza diversa ne risulterebbe un istogramma distorto.

Per ovviare a tale problema si può rappresentare l'altezza dei rettangoli in funzione di una quantità chiamata DENSITA' DI FREQUENZA. che può essere calcolata in vario modo.

In funzione delle frequenze assolute $H_i = \frac{n_i}{a_i}$ dove n_i è la frequenza assoluta nella i esima classe di ampiezza a .

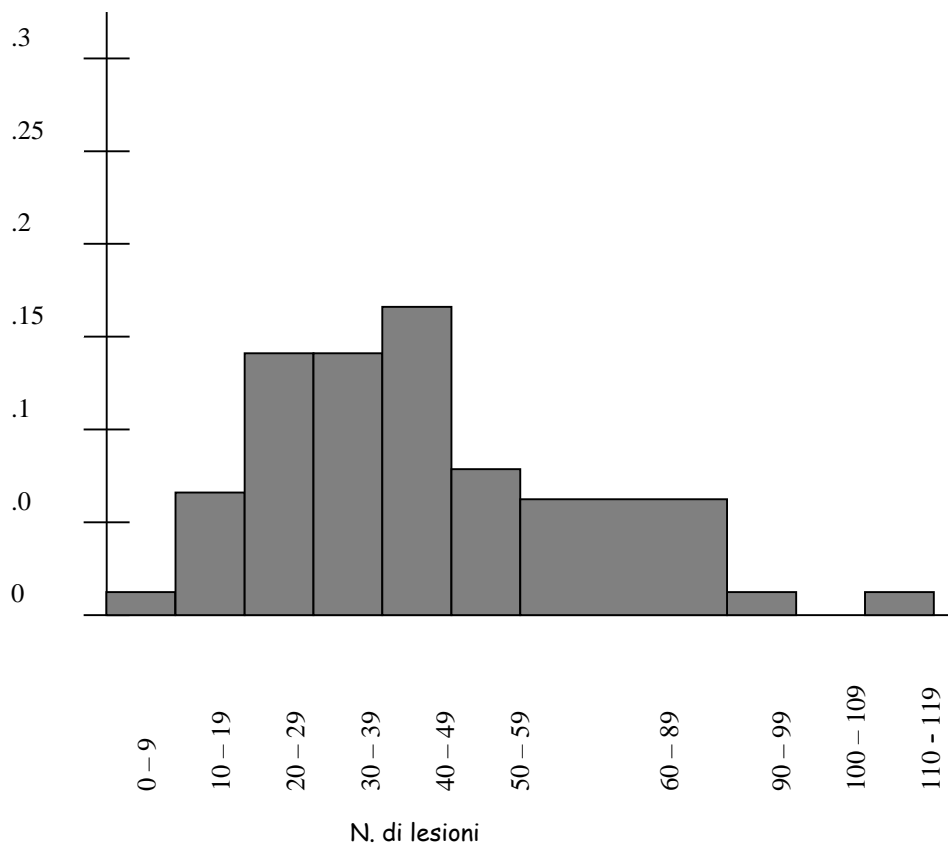
In funzione delle frequenze relative $f_i = \frac{f_i}{a_i}$ dove f_i è la frequenza relativa nella i esima classe di ampiezza a .

Nell'esempio successivo è rappresentata la costruzione dell'istogramma sulla base delle densità di frequenze illustrate nella tabella 2.7 calcolate sulle frequenze assolute.

Tab. 2.7

Classe (x)	Frequenza (f)	Intervallo di classe (d)	Altezza dell'ordinata (h)
0 - 9	1	10	1/10 = 0,1
10 - 19	6	10	6/10 = 0,6
20 - 29	14	10	14/10 = 1,4
30 - 39	14	10	14/10 = 1,4
40 - 49	17	10	17/10 = 1,7
50 - 59	8	10	8/10 = 0,8
60 - 89	18	30	18/30 = 0,6
90 - 99	1	10	1/10 = 0,1
100 - 109	0	10	0/10 = 0,0
110-119	1	10	1/10 = 0,1

Fig. 2.2

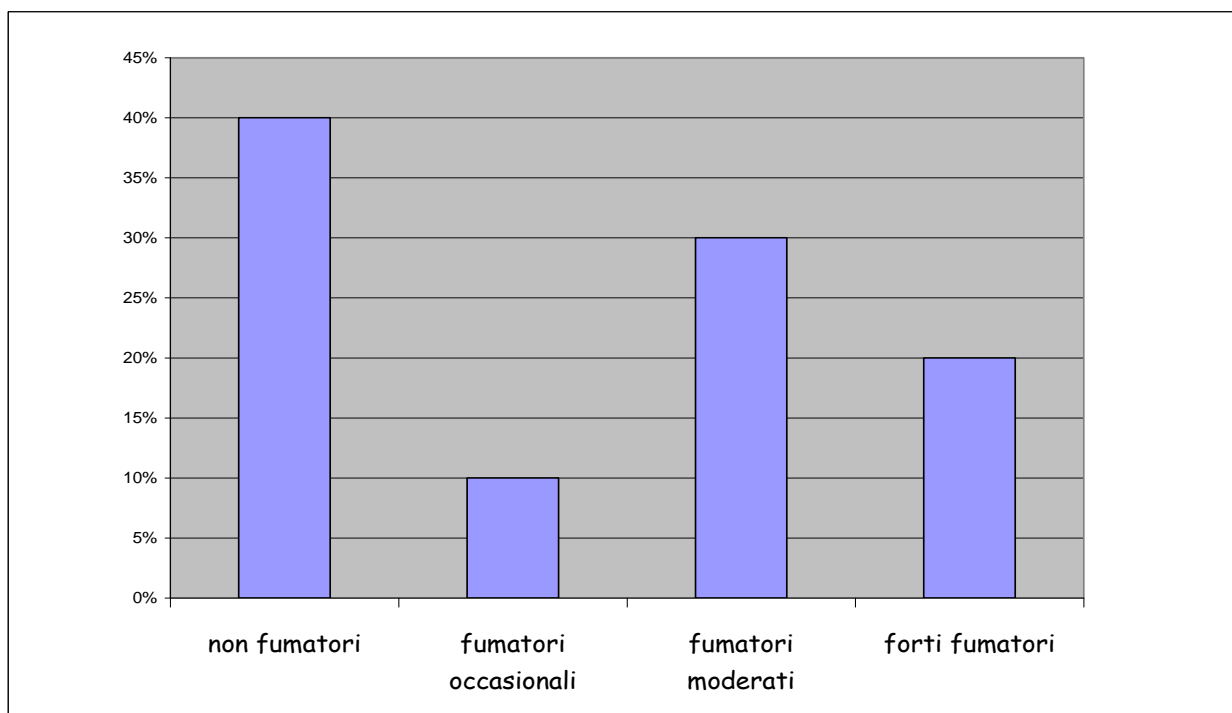


2.3 Diagrammi a barre

Il diagramma a barre viene utilizzato per la rappresentazione grafica delle distribuzioni di frequenza di variabili qualitative nominali o ordinali. A differenza dell'istogramma le barre verticali debbono essere separate poiché non implicano alcuna continuità tra un intervallo e l'altro. L'altezza delle barre rappresenta la frequenza assoluta o relativa del carattere che si sta studiando. L'ampiezza delle barre deve essere uguale per tutte le modalità del carattere.

Nella figura 2.4 è illustrata una distribuzione di una variabile nominale ordinale (attitudine al fumo di sigaretta).

Fig. 2.4



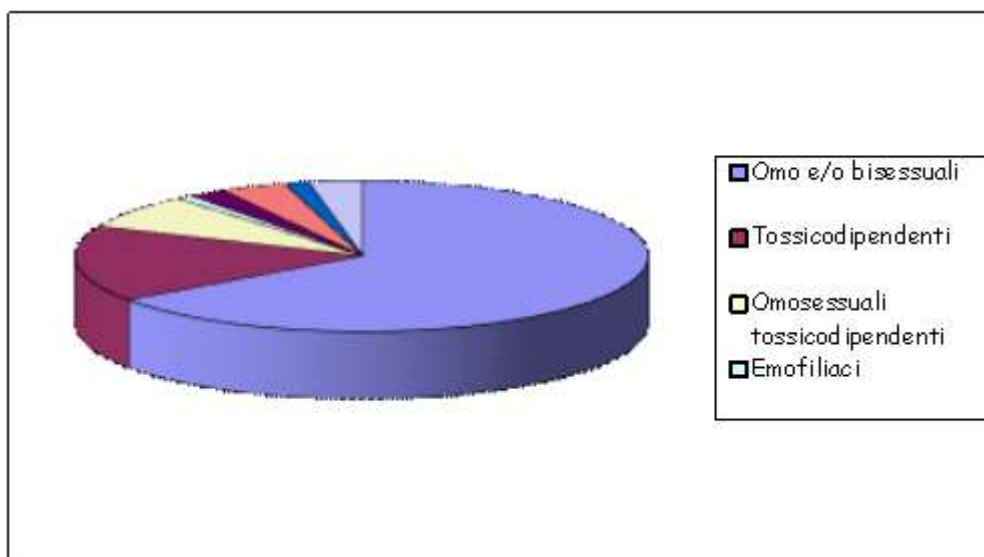
2.4 Diagrammi a settori circolari

Un altro modo di rappresentare graficamente i dati è mediante l'utilizzo di diagrammi a settori circolari (cosiddette torte) dove la frequenza o intensità totale del fenomeno è rappresentata dall'area di un cerchio. La frequenza o intensità di ciascuna modalità è rappresentata da ogni singolo settore circolare il cui angolo al centro è proporzionale alla frequenza o intensità della modalità rappresentata.

Le ampiezze α_i degli angoli dei settori vengono calcolate mediante la seguente proporzione:

$$\frac{\alpha_i}{360^0} = \frac{f_i}{f_{tot}}$$

Fig. 2.5 (dati dell'esempio riportato in tabella 2.1 - USA):



Misure di sintesi e di variabilità

3. Misure di posizione (o di intensità o di localizzazione)

Per poter rappresentare in maniera adeguata i dati di cui si dispone è necessario sintetizzarli ed è pertanto necessario calcolare alcune misure. Il primo passo è sicuramente quello di calcolare le misure di posizione (o valori medi).

Le più importanti misure di posizione sono le medie algebriche, la moda e la mediana.

In questa lezione tratteremo nell'ordine la stima delle medie per variabili di tipo quantitativo, per poi trattare il calcolo delle medie per variabili di tipo qualitativo (nominale) di tipo ordinale ed infine per variabili di tipo nominale sconnesso.

In statistica si considerano due tipi di medie:

Medie di calcolo (o ferme): sono quelle che soddisfano ad una condizione d'invarianza e che si calcolano tenendo conto di tutti i valori della distribuzione.

Medie di posizione (o lasche): sono quelle che si calcolano tenendo conto solo di alcuni valori.

La scelta del tipo di media da utilizzare dipende dal tipo di dati che si sta trattando.

3.1 Media aritmetica

Fa parte della famiglia delle medie algebriche. E' calcolabile per caratteri quantitativi ed è la somma dei valori delle singole osservazioni divisa per il loro numero.

Per indicare un calcolo da eseguire su ogni elemento è utile far riferimento ad un membro generico del gruppo di osservazioni attraverso un indice di comodo (i o j). Se ad esempio x è la variabile quantitativa "altezza di un gruppo di scolari" e x_1, x_2, \dots, x_n sono n valori di x , un generico valore può essere indicato con x_i .

In generale si può chiamare **media di una distribuzione** x_1, x_2, \dots, x_n rispetto a una funzione $f(x_1, \dots, x_n)$ quella quantità M che sostituita alle x_i nella funzione ne lascia invariato il risultato.

Se si definisce la media con il simbolo \bar{x} si ha¹:

$$x_1 + x_2 + x_3 + \dots + x_n = \bar{x} + \bar{x} + \bar{x} + \dots + \bar{x} = n\bar{x}$$

¹ I caratteri latini vengono utilizzati quando si trattano dati campionari. Quando ci si riferisce a dati non campionari ma che riguardano tutta la popolazione si utilizzano i caratteri greci e le lettere maiuscole dei caratteri in latino che rappresentano i valori individuali della variabile in studio. Ad esempio la media aritmetica di una intera popolazione verrebbe indicata con la lettera μ ed i generici valori come X_i

da cui si ricava la **media aritmetica** di più numeri:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

che viene definita *media aritmetica semplice* e si indica con il simbolo \bar{x} (x soprasegnato).

La somma al numeratore può essere indicata con il simbolo greco di sommatoria Σ (somma di) come indicato nella precedente formula. Gli indici che compaiono nel simbolo di sommatoria stanno a significare la somma dei generici valori x_i che vanno da 1 ad n.

Se i valori di x_i hanno frequenze diverse, ossia compaiono più volte nelle osservazioni (ad esempio il valore x_1 potrebbe comparire con frequenza assoluta y_1 , il valore x_2 con frequenza y_2 ecc.. la media diventa:

$$\bar{x} = \frac{x_1 y_1 + x_2 y_2 + \dots + x_n y_n}{y_1 + y_2 + \dots + y_n}$$

che prende il nome di *media aritmetica ponderata* perché le frequenze con le quali i valori compaiono vengono dette *pesi*.

Alternativamente la *media ponderata* può essere espressa come:

$$\bar{x} = \sum_{i=1}^n x_i f_i$$

dove

$$f_i = \frac{y_i}{\sum_{i=1}^n y_i}$$

sono le frequenze relative con cui ciascun valore compare nella distribuzione.

Qualora nella distribuzione di frequenza la variabile sia espressa in classi per effettuare il calcolo della media aritmetica invece che le singole determinazioni x_i sarà necessario utilizzare i valori corrispondenti ai valori centrali delle classi.

Il valore centrale di ciascuna classe equivale alla semi-somma dei due valori estremi dell'intervallo di classe considerato.

Assumendo il valore centrale si commette un errore di approssimazione che in genere non è grave (se l'ampiezza di ciascuna classe è piccola). Nel caso di classi ampie l'errore può essere rilevante.

La tabella 3.1 che segue mostra i dati delle altezze di un campione di 50 studenti e i cui valori individuali sono stati raggruppati in classi:

Tabella 3.1

Classi di altezza	Frequenze assolute	Valori centrali
151 - 156	4	$(156+151)/2 = 153,5$
156 - 161	9	$(161+156)/2 = 158,5$
161 - 166	15	$(166+161)/2 = 163,5$
166 - 171	7	$(171+166)/2 = 168,5$
171 - 176	8	$(176+171)/2 = 173,5$
176 - 181	3	$(181+176)/2 = 178,5$
181 - 186	3	$(186+181)/2 = 183,5$
186 - 191	1	$(191+186)/2 = 188,5$

Prima di procedere ad ulteriori calcoli si considerino le regole di approssimazione.

Si era già visto nella prima lezione a proposito della rappresentazione dei dati la simbologia da utilizzare. Nell'esempio riportato sopra il simbolo | - tra i due estremi dell'intervallo di classe sta ad indicare che il valore dell'estremo inferiore è incluso nell'intervallo mentre il valore dell'estremo superiore è incluso nell'intervallo successivo.

L'ampiezza di una classe è definita come la differenza tra il confine superiore e quello inferiore (l'ampiezza della classe 161|-166 è pertanto $166 - 161 = 5$).

E' importante notare che ogniqualevolta un insieme di dati viene suddiviso in classi, oltre agli arrotondamenti dei numeri, viene effettuata l'approssimazione di sostituire di fatto il valore centrale della classe ad ogni valore appartenente a quella classe assumendo come valore centrale la media aritmetica dei singoli valori della classe.

Una assunzione che viene inoltre fatta per dati raggruppati in classi è che i singoli valori all'interno della classe siano distribuiti in maniera omogenea (vale a dire equidistante). Ad esempio i 15 valori contenuti nell'intervallo 161 |-166 vanno immaginati come disposti ad una distanza l'uno dall'altro pari a $5/15 = 0,33$ (dove 5 è l'ampiezza della classe).

A questo punto possiamo procedere al calcolo del valore medio per dati raggruppati in classi.

Tabella 3.2

Classi di altezza	Frequenze assolute	Valori centrali	Frequenze relative
151 - 156	4	(156+151)/2 = 153,5	4/50 = 0,08 (8%)
156 - 161	9	(161+156)/2 = 158,5	9/50 = 0,18 (18%)
161 - 166	15	(166+161)/2 = 163,5	15/50 = 0,3 (30%)
166 - 171	7	(171+166)/2 = 168,5	7/50 = 0,14 (14%)
171 - 176	8	(176+171)/2 = 173,5	8/50 = 0,16 (16%)
176 - 181	3	(181+176)/2 = 178,5	3/50 = 0,06 (6%)
181 - 186	3	(186+181)/2 = 183,5	3/50 = 0,06 (6%)
186 - 191	1	(191+186)/2 = 188,5	1/50 = 0,02 (2%)
TOTALE	50		1 (100%)

La media aritmetica si può calcolare nel seguente modo:

$$\bar{x} = \frac{x_1y_1 + x_2y_2 + \dots + x_ny_n}{y_1 + y_2 + \dots + y_n} = \frac{\sum_{i=1}^k x_iy_i}{\sum_{i=1}^k y_i}$$

dove:

x_i è il valore centrale della generica classe i

y_i è la frequenza assoluta delle osservazioni rilevata in ciascuna classe generica i

Alternativamente ma in maniera del tutto analoga il valore medio può essere calcolato come segue:

$$\bar{x} = \sum_{i=1}^n x_i f_i$$

dove:

x_i è il valore centrale della generica classe i

f_i è la frequenza relativa delle osservazioni che ricadono in ciascuna classe generica i

Per i dati illustrati nella tabella delle altezze aggregate in classi si ha:

$$\bar{x} = \frac{(153,5 \cdot 4) + (158,5 \cdot 9) + (163,5 \cdot 15) + (168,5 \cdot 7) + (173,5 \cdot 8) + (178,5 \cdot 3) + (183,5 \cdot 3) + (188,5 \cdot 1)}{50} = 166,7$$

Oppure

$$\bar{x} = (153,5 \cdot 0,08) + (158,5 \cdot 0,18) + (163,5 \cdot 0,3) + (168,5 \cdot 0,14) + (173,5 \cdot 0,16) + (178,5 \cdot 0,06) + (183,5 \cdot 0,06) + (188,5 \cdot 0,02) = 166,7$$

La media aritmetica (semplice o ponderata) può essere utilizzata per variabili di tipo quantitativo di tipo continuo o discreto e non è adatta per variabili di tipo nominale o ordinale.

Una eccezione a questa regola è rappresentata da variabili di tipo qualitativo dove le modalità sono dicotomiche.

In questo caso per esprimere il valore medio si usano le proporzioni.

Si abbia ad esempio un campione di $n = 10$ pazienti di cui $r = 4$ sono donne e $n-r = 6$ sono maschi. La proporzione p dei pazienti che sono donne è pari a $r/n = 4/10 = 0,4$.

Se viene attribuito il valore numerico 1 alla modalità "donna" ed il valore 0 alla modalità "uomo" la somma di tutte le modalità "donna" è pari a 4 (che corrisponde al valore di r). Se si divide questa quantità per il numero totale di osservazioni $n = 10$ si ottiene il valore 0,4.

$$p = \frac{1 + 1 + 1 + 1 + 0 + 0 + 0 + 0 + 0 + 0}{10} = 0,4$$

La proporzione è quindi abbastanza simile ad una media e ciò è molto importante per le possibilità di analisi che saranno viste successivamente.

La media aritmetica gode delle seguenti proprietà:

Proprietà 1: *la somma algebrica degli scarti di ogni singolo valore dalla media è uguale a zero.*

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \dots + (x_n - \bar{x}) = 0$$

Lo stesso vale naturalmente per la *media aritmetica ponderata*.

Proprietà 2: *la somma dei quadrati degli scarti di ogni singolo valore dalla media è il minimo valore possibile.*

$$(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 = \min$$

Proprietà 3: aggiungendo (o sottraendo) a tutti i valori x_i la stessa quantità k , la media aritmetica è incrementata (o ridotta) di tale quantità (proprietà traslativa).

$$\frac{(x_1 \pm k) + (x_2 \pm k) + \dots + (x_n \pm k)}{n} = \bar{x} \pm k$$

Proprietà 4: moltiplicando (o dividendo) per la stessa quantità h (diversa da zero) la media aritmetica risulta moltiplicata (o divisa) per tale quantità.

$$\frac{(x_1 h) + (x_2 h) + \dots + (x_n h)}{n} = h \bar{x}$$

4.2 Media Geometrica

Un altro tipo di media algebrica che si incontra frequentemente è la media geometrica.

Tale tipo di media è calcolabile se i valori sono tutti positivi e non nulli.

Si definisce *media geometrica* dei valori x_1, x_2, \dots, x_n quel numero G che sostituito ai valori x_i ne lascia invariato il loro prodotto:

$$x_1 \cdot x_2 \cdot \dots \cdot x_n = G \cdot G \cdot \dots \cdot G = G^n$$

dalla quale si ricava che:

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

che è la media geometrica semplice.

Nel caso di valori x_i con frequenze o pesi y_i si ha:

$$x_1^{y_1} \cdot x_2^{y_2} \cdot \dots \cdot x_n^{y_n} = G^{y_1} \cdot G^{y_2} \cdot \dots \cdot G^{y_n} = G^{y_1 + y_2 + \dots + y_n}$$

e quindi

$$G = \sqrt[n]{x_1^{y_1} \cdot x_2^{y_2} \cdot \dots \cdot x_n^{y_n}}$$

dove

$$n = \sum_{i=1}^n y_i$$

che è la *media geometrica ponderata*.

E' chiaro che non si può calcolare la media geometrica se uno dei valori della serie è pari a zero poiché il prodotto sarebbe nullo.

Ad esempio se si deve calcolare la media della seguente successione di titoli sierologici:

2, 2, 2, 4, 4, 4, 4, 8, 8, 8, 64, 256

$$G = \sqrt[12]{2 \cdot 2 \cdot 2 \cdot 4 \cdot 4 \cdot 4 \cdot 4 \cdot 8 \cdot 8 \cdot 8 \cdot 64 \cdot 256} = 7,13$$

In maniera equivalente:

$$G = \sqrt[12]{2^3 \cdot 4^4 \cdot 8^3 \cdot 64^1 \cdot 256^1} = 7,13$$

che è la *media geometrica ponderata* e dove ciascun valore è stato elevato alla rispettiva frequenza assoluta osservata nella serie di dati.

Un modo più semplice per calcolare la media geometrica è di trasformare i valori originali della variabile nel loro logaritmo (in una qualunque base) e calcolare la media aritmetica dei logaritmi.

$$\log_{10} G = \frac{\log_{10} x_1 + \log_{10} x_2 + \dots + \log_{10} x_n}{n}$$

che nella serie illustrata nell'esempio precedente diventa (usando il log in base 10):

$$\log_{10} G = \frac{0,301 \cdot 0,301 \cdot 0,301 \cdot 0,602 \cdot 0,602 \cdot 0,602 \cdot 0,602 \cdot 0,903 \cdot 0,903 \cdot 0,903 \cdot 1,806 \cdot 2,408}{12} = 0,853$$

che è l'esponente al quale elevare la base utilizzata:

$$10^{0,853} = 7,13$$

Per quanto riguarda la *media geometrica ponderata* si ha:

$$\log_{10} G = \frac{(3 \cdot 0,301) + (4 \cdot 0,602) + (3 \cdot 0,903) + (1 \cdot 1,806) + (1 \cdot 2,408)}{3 + 4 + 3 + 1 + 1} = 0,853$$

3.3 Mediana

La mediana corrisponde al valore centrale quando i valori sono ordinati in ordine crescente, ovvero bipartisce la distribuzione lasciando dalle due parti della distribuzione un ugual numero di casi o di frequenze.

Siano $x_1, x_2, x_3, \dots, x_n$ i valori ordinati in senso crescente, si dice *Mediana (Me)* il valore che bipartisce le frequenze con le quali i valori vengono osservati in due metà uguali, 50% da una parte e 50% dall'altra.

La mediana è calcolabile solo per valori che possono essere ordinati e pertanto è calcolabile sia per variabili di tipo quantitativo (sia continue che discrete) che di tipo nominale ordinale.

In generale si ordinano gli n dati in ordine crescente e se il numero di osservazioni è dispari la mediana corrisponde al valore centrale, vale a dire che occupa la posizione $\frac{n+1}{2}$.

Se il numero n di osservazioni è pari la mediana è stimata utilizzando i due valori in corrispondenza delle posizioni $\left(\frac{n}{2}\right)$ ed $\left(\frac{n}{2} + 1\right)$ calcolandone la media aritmetica.

In questo caso se la variabile è di **tipo quantitativo** la mediana è stimata dalla media aritmetica dei due valori in corrispondenza di $\left(\frac{n}{2}\right)$ ed $\left[\left(\frac{n}{2}\right) + 1\right]$.

Ad esempio si abbia le serie di dati che rappresentano i giorni di degenza ospedaliera di 8 pazienti:

5, 5, 5, 7, 10, 20, 29, 104

La media aritmetica semplice sarebbe $184/8 = 23,1$ giorni che è un risultato molto atipico perché in realtà tutte le degenze ad eccezione di due sono al di sotto del valore medio. In questo caso il valore estremo pari a 104 influenza notevolmente il valore della media aritmetica semplice.

La mediana in questo caso fornisce un quadro più realistico, essendoci 8 osservazioni la mediana risulterà uguale alla media tra i valori compresi tra $\frac{n}{2}$ e $\left(\frac{n}{2} + 1\right)$, vale a dire tra i valori in quarta e quinta posizione:

$$Mediana = \frac{7 + 10}{2} = 8,5$$

Nell'esempio appena visto il calcolo della mediana è relativamente semplice poiché si tratta di dati ordinati individualmente.

Peraltro essendo la variabile di tipo quantitativo è stato possibile calcolare la media aritmetica dei due valori estremi dell'intervallo $(n/2)$ ed $[(n/2)+1]$.

Nel caso in cui le modalità con le quali la variabile quantitativa in studio è raggruppata in intervalli di classe le cose sono un pochino più complesse.

Si considererà il caso di una variabile di tipo quantitativo continuo raggruppata in intervalli di classe.

Si consideri la distribuzione in classi (di altezze in cm) della tabella 3.3

Tabella 3.3

Classi di altezza	Frequenze assolute	Valori centrali	Frequenze relative	Frequenza cumulativa
151 - 156	4	$(156+151)/2 = 153,5$	$4/50 = 0,08$ (8%)	0,08 (8%)
156 - 161	9	$(161+156)/2 = 158,5$	$9/50 = 0,18$ (18%)	0,26 (26%)
161 - 166	15	$(166+161)/2 = 163,5$	$15/50 = 0,3$ (30%)	0,56 (56%)
166 - 171	7	$(171+166)/2 = 168,5$	$7/50 = 0,14$ (14%)	0,7 (70%)
171 - 176	8	$(176+171)/2 = 173,5$	$8/50 = 0,16$ (16%)	0,86 (86%)
176 - 181	3	$(181+176)/2 = 178,5$	$3/50 = 0,06$ (6%)	0,92 (92%)
181 - 186	3	$(186+181)/2 = 183,5$	$3/50 = 0,06$ (6%)	0,98 (98%)
186 - 191	1	$(191+186)/2 = 188,5$	$1/50 = 0,02$ (2%)	1 (100%)
TOTALE	50		1 (100%)	

Per individuare la classe mediana è necessario identificare l'intervallo di classe all'interno del quale cade il 50% della distribuzione.

Dalla tabella si vede che tale intervallo (mediano) corrisponde alla classe 161|-166 poiché è l'intervallo di classe nel quale cade il valore 50% della frequenza cumulativa.

Tale intervallo di classe è inoltre compreso tra l'intervallo 156|-161 (con frequenza cumulativa pari al 26%) e 166|-171 (con frequenza cumulativa pari al 70%).

Inoltre la frequenza relativa dell'intervallo di classe 161|-166 è pari al 30%.

A questo punto o ci si può limitare ad indicare che la mediana è contenuta nell'intervallo di classe 161|-166 oppure, se volessimo esprimere la mediana come valore unico (piuttosto che come intervallo di classe), si ricorre alla seguente equazione (partendo dall'assunto che i singoli valori all'interno della classe siano uniformemente distribuiti).

$$M_e = c_{i-1} + \frac{c_i - c_{i-1}}{f_i} \cdot (0,5 - F_{i-1})$$

Dove:

c_{i-1} è il valore estremo inferiore della classe mediana

c_i è il valore estremo superiore della classe mediana

f_i è la frequenza relativa della classe mediana

F_{i-1} è la frequenza cumulativa nell'intervallo immediatamente precedente la classe mediana

Nel caso dell'esempio della distribuzione in classi delle altezze si ha:

$$M_e = 161 + \frac{166 - 161}{0,30} \cdot (0,5 - 0,26) = 165$$

Nel caso di variabili di tipo qualitativo ordinali non è possibile effettuare interpolazioni e pertanto la mediana sarà rappresentata dalla modalità nella quale cade il 50% delle frequenze.

Si abbia il seguente esempio (tabella 3.4): in un sondaggio effettuato all'interno di una facoltà composta da 250 studenti si vuole rilevare il carattere "Gradimento dei Professori" secondo cinque modalità: "molto deluso", "insoddisfatto", "parzialmente soddisfatto", "soddisfatto", "entusiasta".

Tabella 3.4

Modalità	Frequenze assolute	Frequenze relative	Frequenze cumulative
Molto deluso	36	0,144 (14,4%)	0,144 (14,4%)
Insoddisfatto	90	0,36 (36%)	0,504 (50,4%)
Parzialmente soddisfatto	63	0,252 (25,2%)	0,756 (75,6%)
Soddisfatto	51	0,204 (20,4%)	0,960 (96%)
Entusiasta	10	0,04 (4%)	1 (100%)
TOTALI	250	1 (100%)	

In questo caso la mediana (o meglio la classe mediana) è rappresentata dalla modalità "Insoddisfatto".

La mediana gode di una importante proprietà (per variabili di tipo quantitativo) che viene così definita: *la mediana corrisponde a quel valore per cui la somma degli scarti assoluti di ciascun valore dalla mediana è minima.*

Tale proprietà viene rappresentata nel seguente modo: $\sum_{i=1}^n |x_i - \bar{x}| = \min$

3.4 Moda

E' la più semplice delle medie e si definisce come la modalità che si presenta con la massima frequenza. E' calcolabile per qualsiasi tipo di dati.

Se si dispone di una serie di dati con valori discreti la *moda* è il valore con la massima frequenza.

Se i dati sono raggruppati in classi il calcolo della *moda* non presenta particolari difficoltà se le classi hanno tutte la stessa ampiezza, viceversa può presentare qualche difficoltà in più se le classi hanno ampiezza diversa.

In questo caso la classe modale è quella che ha la maggiore Densità di Frequenza che è ottenuta dividendo ciascuna frequenza per l'ampiezza della classe.

Per dati di tipo quantitativo continuo raggruppati in classi il valore della moda corrisponderà all'intervallo di classe che si presenta con maggiore frequenza.

Valore modale per dati quantitativi raggruppati in classi di frequenza della tabella 3.3:

Classi di altezza	Frequenze assolute	Valori centrali	Frequenze relative	Frequenza cumulativa
151 - 156	4	$(156+151)/2 = 153,5$	$4/50 = 0,08$ (8%)	0,08 (8%)
156 - 161	9	$(161+156)/2 = 158,5$	$9/50 = 0,18$ (18%)	0,26 (26%)
161 - 166	15	$(166+161)/2 = 163,5$	$15/50 = 0,3$ (30%)	0,56 (56%)
166 - 171	7	$(171+166)/2 = 168,5$	$7/50 = 0,14$ (14%)	0,7 (70%)
171 - 176	8	$(176+171)/2 = 173,5$	$8/50 = 0,16$ (16%)	0,86 (86%)
176 - 181	3	$(181+176)/2 = 178,5$	$3/50 = 0,06$ (6%)	0,92 (92%)
181 - 186	3	$(186+181)/2 = 183,5$	$3/50 = 0,06$ (6%)	0,98 (98%)
186 - 191	1	$(191+186)/2 = 188,5$	$1/50 = 0,02$ (2%)	1 (100%)
TOTALE	50		1 (100%)	

La Moda è rappresentata dalla frequenza più elevata che corrisponde alla classe 161|-166 che ha una frequenza relativa pari al 30%.

Da notare che la Moda e la Mediana sono entrambe nella stessa classe.

Nel caso di dati Nominali della tabella 3.4 la moda corrisponde alla modalità "Insoddisfatto" (la stessa modalità della Mediana) che ha una frequenza relativa pari al 36%:

Modalità	Frequenze assolute	Frequenze relative	Frequenze cumulative
Molto deluso	36	0,144 (14,4%)	0,144 (14,4%)
Insoddisfatto	90	0,36 (36%)	0,504 (50,4%)
Parzialmente soddisfatto	63	0,252 (25,2%)	0,756 (75,6%)
Soddisfatto	51	0,204 (20,4%)	0,960 (96%)
Entusiasta	10	0,04 (4%)	1 (100%)
TOTALI	250	1 (100%)	

Lo svantaggio della Moda è che può non essere unica, in tali casi si parla di distribuzioni bimodali, trimodali, etc...

Nella tabella 3.5 che segue è riportato quali sono gli indici di tendenza centrale calcolabili sui diversi tipi di variabili:

Tabella 3.5

	Tipi di variabili		
	Nominale sconnessa	Nominale ordinale	Quantitativa (discreta o continua)
Indice di tendenza centrale	MODA (ad eccezione di variabili di tipo dicotomico)	MODA, MEDIANA	MODA, MEDIANA, MEDIA ARITMETICA, MEDIA GEOMETRICA

4. Misure di variabilità

Calcolare il valore di tendenza centrale di una serie di dati non è una operazione sufficiente per avere una descrizione più completa delle osservazioni effettuate. Ad esempio una domanda che ci si può porre è quanto i singoli valori di una distribuzione tendono ad essere simili tra di loro oppure a distribuirsi in un intervallo più o meno ampio.

Una idea della distribuzione dei valori viene già dalla costruzione di tabelle o grafici, per poter però quantificare la variabilità dei dati è necessario misurarne l'entità. Anche in questo caso le misure di variabilità calcolabili sono diverse a seconda della tipologia di variabili con le quali si ha a che fare.

4.1 Campo di variazione (range)

Si definisce campo di variazione la differenza tra l'osservazione più grande e quella più piccola. Ad esempio considerando la seguente distribuzione quantitativa discreta:

5 15 8 9 20 4 14 10 6

il campo di variazione (o range) risulta essere $20 - 4 = 16$

Il campo di variazione:

- non prende in considerazione la variabilità delle osservazioni compresa tra i due estremi;
- tende ad aumentare con l'aumentare delle osservazioni (ed è scomodo avere una misura di variabilità che dipenda dal numero di osservazioni);
- in qualche modo ha gli stessi svantaggi della media aritmetica poiché risente dei valori estremi.

Il campo di variazione è calcolabile soltanto su dati di tipo numerico sui quali è possibile effettuare l'operazione di sottrazione (variabili quantitative).

4.2 Campo di variazione interquartile

Una misura di variabilità che non è influenzata dai valori estremi è il campo di variazione interquartile (o Range Interquartile). Il Range Interquartile è calcolabile solo per valori

di tipo quantitativo dove (analogamente al Campo di Variazione) sono eseguibili operazioni di sottrazione.

Si è visto in precedenza che una distribuzione di n osservazioni può essere suddivisa in due metà uguali e la Mediana corrisponde al valore (o la modalità) in corrispondenza del 50% delle n osservazioni.

Se su ciascuna metà delle osservazioni ne viene calcolata nuovamente la Mediana tali valori cadranno in corrispondenza rispettivamente del 25% e 75% delle osservazioni.

Il valore in corrispondenza del 25% delle osservazioni prende il nome di primo Quartile (Q_1), il secondo Quartile (Q_2) è il valore della Mediana mentre il terzo Quartile (Q_3) è il valore in corrispondenza del 75% delle n osservazioni.

La differenza tra i valori corrispondenti al 75% e quelli al 25% della distribuzione prende il nome di Campo di Variazione Interquartile (o Range Interquartile).

Ad esempio si abbia le serie di dati che rappresentano i giorni di degenza ospedaliera di 8 pazienti già visti in precedenza e rispetto ai quali ne era stato calcolato il valore Mediano ($Q_2 = 8,5$).

Essendo n un numero pari il valore mediano cade tra la quarta e la quinta osservazione. La Mediana a sinistra delle prime quattro osservazioni corrisponderà a Q_1 . Essendo anche in questo caso il numero di osservazioni pari la Mediana sarà data dalla media aritmetica dei valori in corrispondenza di $\frac{n}{2} = 2$ ed $\left(\frac{n}{2} + 1\right) = 3$ delle prime quattro osservazioni:

5, 5, 5, 7, 10, 20, 29, 104

$$\text{Pertanto } Q_1 = \frac{5+5}{2} = 5$$

Mentre Q_3 sarà pari alla Mediana delle osservazioni che cadono a destra del valore Mediano della intera distribuzione (dalla quinta alla ottava osservazione).

5, 5, 5, 7, 10, 20, 29, 104

$$\text{Pertanto } Q_3 = \frac{20+29}{2} = 24,5$$

Il Range Interquartile sarà quindi uguale a $24,5 - 5 = 19,5$

Come già visto nel caso del calcolo della *mediana* quando le osservazioni sono raggruppate in intervalli di classe si può utilizzare lo stesso approccio per calcolare i *quartili* effettuando le opportune sostituzioni.

Ad esempio se volessimo trovare i valori interquartili della tabella sottostante raggruppati in intervalli di classe della Tabella 3.3:

Classi di altezza	Frequenze assolute	Valori centrali	Frequenze relative	Frequenza cumulativa
151 - 156	4	$(156+151)/2 = 153,5$	$4/50 = 0,08$ (8%)	0,08 (8%)
156 - 161	9	$(161+156)/2 = 158,5$	$9/50 = 0,18$ (18%)	0,26 (26%)
161 - 166	15	$(166+161)/2 = 163,5$	$15/50 = 0,3$ (30%)	0,56 (56%)
166 - 171	7	$(171+166)/2 = 168,5$	$7/50 = 0,14$ (14%)	0,7 (70%)
171 - 176	8	$(176+171)/2 = 173,5$	$8/50 = 0,16$ (16%)	0,86 (86%)
176 - 181	3	$(181+176)/2 = 178,5$	$3/50 = 0,06$ (6%)	0,92 (92%)
181 - 186	3	$(186+181)/2 = 183,5$	$3/50 = 0,06$ (6%)	0,98 (98%)
186 - 191	1	$(191+186)/2 = 188,5$	$1/50 = 0,02$ (2%)	1 (100%)
TOTALE	50		1 (100%)	

Per individuare il *primo quartile* è necessario individuare l'intervallo di classe all'interno del quale è compreso il 25% della distribuzione.

Dalla tabella si vede che tale intervallo corrisponde alla classe 156|-161 poiché è l'intervallo di classe con frequenza cumulativa del 26% e che pertanto contiene il valore 25%.

Tale intervallo di classe è inoltre compreso tra l'intervallo di classe 151|-156 (con frequenza cumulativa pari all'8%) e 161|-166 (con frequenza cumulativa pari al 56%).

Inoltre la frequenza relativa dell'intervallo di classe 156|-161 è pari al 18%.

Il valore corrispondente al *primo quartile* lo si può calcolare a partire dalla stessa equazione utilizzata per il calcolo della mediana per valori raggruppati in classi:

$$M_e = c_{i-1} + \frac{c_i - c_{i-1}}{f_i} \cdot (0,5 - F_{i-1})$$

Dove al valore 0,5 dovrà essere sostituito il valore 0,25:

$$Q_{0,25} = c_{i-1} + \frac{c_i - c_{i-1}}{f_i} \cdot (0,25 - F_{i-1})$$

e dove pertanto il significato dei simboli diventa:

c_{i-1} è il valore estremo inferiore della classe del primo quartile

c_i è il valore estremo superiore della classe del primo quartile

f_i è la frequenza relativa della classe corrispondente al primo quartile

F_{i-1} è la frequenza cumulativa nell'intervallo immediatamente precedente

Quindi:

$$Q_{0,25} = 156 + \left(\frac{161 - 156}{0,18} \right) \cdot (0,25 - 0,08) = 160,72$$

In maniera analoga si deve procedere per il calcolo del valore corrispondente al terzo Quartile ($Q_{0,75}$).

Dalla tabella si vede che tale intervallo corrisponde alla classe 171 | - 176 poiché è l'intervallo di classe con frequenza cumulativa del 86% e che pertanto contiene il valore 75%.

Tale intervallo di classe è inoltre compreso tra l'intervallo di classe 166 | - 171 (con frequenza cumulativa pari al 70%) e 176 | - 181 (con frequenza cumulativa pari al 92%).

Inoltre la frequenza relativa dell'intervallo di classe 171 | - 176 è pari al 16%.

Inseriti tali valori nella equazione si ha:

$$Q_{0,75} = 171 + \left(\frac{176 - 171}{0,16} \right) \cdot (0,75 - 0,7) = 172,56$$

Il Range Interquartile è pertanto:

$$Range_{0,75-0,25} = 172,56 - 160,72 = 11,84$$

Tale risultato indica che il 50% delle osservazioni centrali sono contenute all'interno di tale intervallo.

A questo punto abbiamo acquisito i seguenti dati sulla distribuzione delle altezze illustrate nella tabella 4.2

$$\bar{x} = 166,7 \qquad \text{Mediana} = 165 \qquad \text{Range interquartile}_{0,75-0,25} = 11,84$$

Moda 161 | - 166 poiché è la classe con la maggiore frequenza (pari al 30%)

Vi sono indicazioni di una distribuzione abbastanza simmetrica. Il valore della media aritmetica e della mediana sono molto vicini ed inoltre la classe con maggiore frequenza contiene entrambi i valori della media aritmetica e della mediana.

4.2 Varianza e deviazione standard

La varianza è la misura di variabilità che si basa sulla differenza tra ogni singola osservazione e la media delle osservazioni stesse.

E' possibile calcolarla per variabili di tipo quantitativo e per variabili di tipo nominale che possono assumere solo due valori (variabili dicotomiche o binomiali).

Si supponga di rilevare in una popolazione di N individui i valori della variabile continua X_i . Se la media nella popolazione è pari a μ le differenze tra i valori di X e la media sono $(X_1 - \mu)$, $(X_2 - \mu)$, ..., $(X_n - \mu)$. E' ovvio che tanto più sono grandi le differenze tra i singoli valori ed il valor medio tanto più ciò implica che la variabilità delle osservazioni attorno alla media è grande.

Allo scopo di dare una misura della variabilità si potrebbe pensare di sommare tutte le deviazioni dei valori osservati dal valore medio e calcolare pertanto:

$$\sum (X_i - \mu)$$

questa quantità è però pari a 0 (è una delle proprietà della media).

Per superare il problema delle deviazioni negative che annullano le positive si elevano al quadrato i valori delle singole deviazioni:

$$\sum (X_i - \mu)^2$$

Tale quantità prende il nome di Devianza e rappresenta appunto la sommatoria delle singole deviazioni tra ogni singolo valore dalla media (ogni deviazione è elevata al quadrato in modo tale che ogni valore è positivo).

L'inconveniente maggiore della Devianza è che tende ad aumentare con il numero di osservazioni.

Un modo per ovviare a tale inconveniente è di calcolarne un valore medio.

$$\sum \frac{(X_i - \mu)^2}{N}$$

dove N rappresenta il totale delle osservazioni.

Tale quantità prende il nome di Varianza che calcolata sulla intera popolazione viene indicata con il simbolo σ^2 (sigma al quadrato).

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

N maiuscolo sta ad indicare che il valore della Varianza è calcolato su tutti gli individui che compongono la popolazione.

Nella realtà la varianza della popolazione non è quasi mai disponibile perché non sono disponibili i dati individuali della intera popolazione.

Ciò che viene utilizzata più spesso è la stima campionaria della Varianza che viene indicata come segue:

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

Si noti che per i dati campionari la Varianza viene indicata con il carattere latino s^2 , che la media viene indicata con la lettera x soprastegnata e che il numero di osservazioni viene indicato con n minuscola (indicando che si tratta di un campione estratto dalla popolazione composta da N individui).

Inoltre si osservi che per il calcolo della Varianza campionaria la somma delle deviazioni al quadrato viene divisa per $n-1$.

Il motivo è che in un campione di n osservazioni la somma $(x_i - \bar{x})^2$ sarà più piccola del valore $(x_i - \mu)^2$ poiché i valori estremi della distribuzione (che in genere sono poco frequenti) avranno meno probabilità di essere rappresentati nel campione (mentre lo saranno con certezza nella distribuzione della intera popolazione) e si è già visto che i dati estremi possono avere un peso rilevante nella stima dei valori medi.

La distorsione (bias) che viene pertanto introdotta è che i dati campionari sottostimano il valore della popolazione. Dividendo per $n-1$ il valore della varianza campionaria tende ad essere leggermente più grande e più vicino al valore della varianza della popolazione.

Il denominatore $(n-1)$ della stima della varianza campionaria è chiamato GRADI DI LIBERTA' ed è in genere indicato con lettera greca ν (ni) o con df (nei testi inglesi - degrees of freedom) o con gl (nei testi italiani).

Il concetto di 'gradi di libertà' è legato al fatto che una stima campionaria della varianza (dato un certo risultato) può essere pensato come generato da diverse serie di valori indipendenti ad eccezione dell'ultimo valore che è vincolato per poter restituire il valore stimato.

L'esempio che segue dovrebbe chiarire.

Si abbia la seguente serie di dati campionari di altezze degli individui: 167, 168, 171, 172, 173, 180, 185 il cui valore medio è pari a 173,71

La stima della varianza campionaria si ottiene nel seguente modo:

$$s^2 = \frac{(167 - 173,71)^2 + (168 - 173,71)^2 + (171 - 173,71)^2 + (172 - 173,71)^2 + (173 - 173,71)^2 + (180 - 173,71)^2 + (185 - 173,71)^2}{(7 - 1)} = \frac{255,428}{6} = 42,57$$

Se si estraesse un altro campione di 7 individui soltanto 6 dei 7 valori possono essere estratti in maniera casuale poiché il settimo è vincolato al risultato dato.

Si dice allora che i Gradi di Libertà della distribuzione è pari a 6 e l'ultimo valore sarà vincolato al fatto che il numeratore dovrà essere necessariamente pari a 255,428 (la somma del numeratore) per restituire un valore della Varianza campionaria pari a 42,57

Ad esempio se i primi 6 valori fossero uguali a 164, 166, 171, 172, 175, 176 l'ultimo valore non può che essere uguale a 182,9 che è l'unico valore per cui la somma dei quadrati del numeratore è uguale a 255,428

L'unico inconveniente della varianza è che utilizza come unità di misura il quadrato della unità di misura originaria e per ovviare a questo problema si utilizza la radice quadrata della Varianza.

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$$

La simbologia utilizzata si riferisce ad ipotetici dati della intera popolazione.

$$\text{Nel caso di dati campionari si ha: } s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Tale quantità è fondamentale in statistica e prende il nome di **Deviazione Standard**. La deviazione standard è una specie di deviazione media delle osservazioni dalla media ed ha il vantaggio di utilizzare le stesse unità di misura originarie.

Nel caso di dati raggruppati si utilizzano gli stessi criteri per il calcolo della media aritmetica dove i singoli valori vengono sostituiti dai valori centrali della classe:

$$s = \sqrt{\frac{\sum_{i=1}^k (m_i - \bar{x})^2 \cdot y_i}{(\sum_{i=1}^k y_i) - 1}}$$

Dove m_i corrisponde al valore centrale di ciascuna classe k_i e y_i è la frequenza delle osservazioni in ciascuna classe.

Si prenda come esempio i dati della tabella 3.1 (per i quali il valore medio è pari a 166,7)

Classi di altezza	Frequenze assolute	Valori centrali	$(m_i - \bar{x})^2$	$(m_i - \bar{x})^2 \cdot y_i$
151 - 156	4	$(156+151)/2 = 153,5$	$(153,5 - 166,7)^2$	$(153,5 - 166,7)^2 \cdot 4$
156 - 161	9	$(161+156)/2 = 158,5$	$(158,5 - 166,7)^2$	$(158,5 - 166,7)^2 \cdot 9$
161 - 166	15	$(166+161)/2 = 163,5$	$(163,5 - 166,7)^2$	$(163,5 - 166,7)^2 \cdot 15$
166 - 171	7	$(171+166)/2 = 168,5$	$(168,5 - 166,7)^2$	$(168,5 - 166,7)^2 \cdot 7$
171 - 176	8	$(176+171)/2 = 173,5$	$(173,5 - 166,7)^2$	$(173,5 - 166,7)^2 \cdot 8$
176 - 181	3	$(181+176)/2 = 178,5$	$(178,5 - 166,7)^2$	$(178,5 - 166,7)^2 \cdot 3$
181 - 186	3	$(186+181)/2 = 183,5$	$(183,5 - 166,7)^2$	$(183,5 - 166,7)^2 \cdot 3$
186 - 191	1	$(191+186)/2 = 188,5$	$(188,5 - 166,7)^2$	$(188,5 - 166,7)^2 \cdot 1$
Totale	50		Sommatoria	3588

$$s^2 = \frac{3588}{(50 - 1)} = 73,22$$

La Deviazione Standard è pari quindi a:

$$s = \sqrt{73,22} = 8,56$$

Nel caso di distribuzioni di frequenza simmetriche l'intervallo $\bar{x} \pm 2s$ contiene circa il 95% dei valori della distribuzione.

L'intervallo $\bar{x} \pm 2s$ viene anche chiamato Intervallo (o Range) di Normalità. Nel caso specifico il Range di Normalità è compreso tra:

$$166,7 - 2(8,56) \text{ e } 166,7 + 2(8,56)$$

Cioè tra i valori **149,58 e 183,81**

Una ulteriore misura di variabilità spesso usata è il Coefficiente di Variazione espresso dal rapporto tra la Deviazione Standard e la Media ed è solitamente espresso in percentuale.

$$CV = \frac{\sigma}{\mu} \cdot 100 \text{ che per dati campionari è stimato come } CV = \frac{s}{\bar{x}} \cdot 100$$

Il Coefficiente di Variazione è una misura relativa della variabilità rispetto al valore medio.

Nel caso dell'esempio della tabella 4.1 il Coefficiente di Variazione è il seguente:

$$CV = \frac{8,56}{166,7} \cdot 100 = 5,13\%$$

Il Coefficiente di Variazione suggerisce che non vi è molta dispersione dei dati attorno al valore medio.

Come accennato in precedenza il calcolo delle misure di variabilità come descritto nel paragrafo precedente è applicabile solamente a dati quantitativi.

Per variabili di tipo nominale (qualitative) si usano le proporzioni e un caso un po' speciale è rappresentato dalle variabili qualitative che possono assumere solo due valori, come ad esempio la variabile sesso.

Abbiamo già visto che a tali variabili può essere assegnato il valore 1 ad una modalità e 0 alla modalità alternativa.

Si abbia ad esempio un campione di $n = 10$ pazienti di cui $r = 4$ sono donne e $n-r = 6$ sono maschi. La proporzione p dei pazienti che sono donne è pari a $r/n = 4/10 = 0,4$.

Se viene attribuito il valore numerico 1 alla modalità "donna" ed il valore 0 alla modalità "uomo" la somma di tutte le modalità "donna" è pari a 4 (che corrisponde al valore di r). Se si divide questa quantità per il numero totale di osservazioni $n = 10$ si ottiene il valore 0,4.

$$p = \frac{1 + 1 + 1 + 1 + 0 + 0 + 0 + 0 + 0 + 0}{10} = 0,4$$

In genere si indica con 0 i fallimenti (nell'esempio la modalità uomo = 0) e 1 i successi (nell'esempio la modalità donna = 1).

In generale se p è la proporzione del campione classificata come 'successo' $(1-p) = q$ è classificata come proporzione di fallimenti.

Laddove, come in questo caso, una variabile qualitativa può assumere uno dei due possibili risultati si ha a che fare con una variabile di tipo BINOMIALE.

Una variabile binomiale si ottiene quando si osservano i risultati in n prove indipendenti ($n = 10$ nell'esempio precedente) e dove r è il numero di successi ed $(n-r)$ il numero di fallimenti.

Se volessimo calcolare la Varianza della variabile Sesso (indicata genericamente come x_i che può assumere solo due possibili valori 0 o 1) indichiamo la media con π si ha:

$$\sigma^2 = \frac{\sum (X_i - \pi)^2}{N}$$

Espandendo il numeratore ed effettuando una serie di semplificazioni si ottiene:

$$\sigma^2 = \pi (1 - \pi)$$

La deviazione standard sarà pertanto:

$$\sigma = \sqrt{\pi (1 - \pi)}$$

Per i dati campionari vale la stessa regola di utilizzare i caratteri latini e pertanto:

$$s = \sqrt{p (1 - p)}$$

E' un po' difficile attribuire un significato preciso a questi valori poiché i valori della variabile binomiale sono fissi (1 o 0).

Ciò che si può osservare è che il valore di s^2 (e di s) tende ad un massimo quando $p = 0,5$ che è la massima variabilità possibile in una popolazione dove la variabile in studio può assumere solo due modalità (il che è ragionevole poiché in una popolazione binomiale la massima variabilità coincide con valori di $\pi = 0,5$).

L'importanza di tale risultato lo si vedrà in seguito quando verranno trattati i dati campionari dei valori medi.

4.3 Indici di variabilità per dati qualitativi

Per quanto riguarda variabili di tipo qualitativo (per le quali non sono possibili operazioni di tipo algebrico) un modo per stimare la variabilità è l'Indice di eterogeneità di Gini attraverso il quale si può misurare la eterogeneità di una distribuzione a partire dai valori delle frequenze relative associate a ciascuna modalità.

Si prenda come esempio i dati riportati nella tabella 4.4 di una variabile qualitativa ordinale con 5 modalità ($k = 5$).

Tabella 4.4

Modalità	Frequenze assolute	Frequenze relative	Frequenze cumulative
Molto deluso	36	0,144 (14,4%)	0,144 (14,4%)
Insoddisfatto	90	0,36 (36%)	0,504 (50,4%)
Parzialmente soddisfatto	63	0,252 (25,2%)	0,756 (75,6%)
Soddisfatto	51	0,204 (20,4%)	0,960 (96%)
Entusiasta	10	0,04 (4%)	1 (100%)
TOTALI	250	1 (100%)	

Se i dati fossero distribuiti in maniera eterogenea ciò sta a significare che le frequenze relative di ciascuna modalità sono simili e che le k modalità sono tutte rappresentate. Viceversa la massima omogeneità la si ottiene quando i dati sono tutti concentrati su una sola delle k modalità considerate.

L'Indice di Gini si calcola come segue:

$$I = 1 - \sum_{i=1}^k f_i^2$$

dove f_i sono le frequenze relative delle k modalità della variabile in studio.

Nell'esempio dei dati riportati nella tabella 4.4 si ha:

$$I = 1 - (0,144^2 + 0,36^2 + 0,252^2 + 0,204^2 + 0,04^2) = 0,74$$

Nel caso in cui tutte le 250 frequenze si fossero concentrate su una unica modalità questa avrebbe preso valore 1 (100%) e 0 le altre. Il valore dell'Indice di Gini sarebbe stato pertanto (poniamo che tutti i 250 studenti avessero optato per la modalità "Parzialmente soddisfatto"):

$$I_{min} = 1 - (0^2 + 0^2 + 1^2 + 0^2 + 0^2) = 0$$

Tale risultato sta ad indicare appunto che non vi è eterogeneità nella distribuzione poiché delle 5 modalità considerate le frequenze sono concentrate soltanto su una.

Per poter effettuare confronti l'Indice di Gini deve essere normalizzato, vale a dire relativizzato rispetto al valore massimo che l'indice stesso può assumere nella distribuzione in studio.

Nell'esempio della tabella 4.4 il valore massimo lo si sarebbe potuto ottenere se tutte le frequenze relative fossero state uguali (massima eterogeneità), cioè se ogni modalità avesse avuto frequenza relativa pari a $1/k$:

$$I_{max} = 1 - \sum_{i=1}^k \left(\frac{1}{k}\right)^2 = 1 - \frac{k}{k^2} = 1 - \frac{1}{k} = \frac{k-1}{k} = \frac{5-1}{5} = 0,8$$

Pertanto l'Indice Normalizzato diventa:

$$I_{norm} = \frac{0,74}{0,8} = 0,93$$

L'Indice può anche essere espresso come percentuale stando ad indicare che la eterogeneità osservata è pari al 93% della massima eterogeneità osservabile sui dati.

Riassumendo:

Tabella 4.5

	Tipi di variabili		
	Nominale sconnessa	Nominale ordinale	Quantitativa (discreta o continua)
Indice di tendenza centrale ammissibili	MODA (ad eccezione di variabili di tipo dicotomico)	MODA, MEDIANA	MODA, MEDIANA, MEDIA ARITMETICA, MEDIA GEOMETRICA
Indici di variabilità ammissibili	INDICE DI ETEROGENEITÀ e (per le sole variabili di tipo dicotomico) VARIANZA E DEVIAZIONE STANDARD	INDICE DI ETEROGENEITÀ, DIFFERENZA INTERQUARTILE	DIFFERENZA INTERQUARTILE, VARIANZA, DEVIAZIONE STANDARD