STATISTICA

Corso di Laurea in Biotecnologie

- **Introduzione alla Statistica**
- **Parte I - Statistica Descrittiva**
- **Parte II - Calcolo delle Probabilità**
- **Parte III - Inferenza Statistica**

Paolo Vidoni
Dipartimento di Scienze Statistiche
Università di Udine

INTRODUZIONE ALLA STATISTICA

(Pace e Salvan, *Introduzione alla Statistica I - Statistica Descrittiva*, CEDAM Padova, 1996)

Lo studio di un fenomeno di interesse richiede spesso l'analisi di **informazioni espresse in forma quantitativa (i dati)**.

La Statistica fornisce concetti e strumenti per evidenziare gli aspetti rilevanti racchiusi nei dati e per quantificare la forza delle conclusioni che si possono dedurre da tale analisi.

La Statistica è una Matematica Applicata, che pur avendo come riferimento concreto i dati o il particolare fenomeno di interesse, interviene con principi e metodologie proprie.

La Statistica è di supporto a varie discipline, quali l'Economia, la Finanza, la Sociologia, la Biologia, le Scienze Naturali, ecc.

I dati si ottengono sia tramite *osservazione* sia tramite *sperimentazione*.

- Nella **sperimentazione** i dati sono *creati in circostanze controllate*. L'esperimento può essere *replicato un numero di volte potenzialmente infinito*, mantenendo fede ad un determinato protocollo sperimentale.

  Esempio. Sono esempi di sperimentazioni: la pesata di una modesta quantità di reagente con una bilancia di precisione; il lancio di un dado; la valutazione della qualità di un prodotto industriale; l'estrazione di un campione di individui da una popolazione nota. ◇

- Nell'**osservazione** il *fenomeno di interesse è precostituito* e i dati esistono in natura. I dati sono *finiti* e vengono rilevati direttamente per come si presentano. Sono tipicamente osservazioni di caratteristiche antropometiche o demografiche rilevate con *indagini censuarie*.

  Esempio. Il sesso, l'età, la statura e il gruppo sanguigno dei residenti nel Comune di Udine al 31 dicembre 2003. ◇

Sia nei dati ottenuti tramite sperimentazione che tramite osservazione si rileva usualmente la presenza di una certa **variabilità**.

I dati rappresentano l'informazione disponibile su certe caratteristiche di una popolazione. Al variare dell'**unità statistica** u, entro l'aggregato $\mathcal{U}$ di tutte le unità, detto **popolazione**, variano certe caratteristiche misurate su u.

Esempio. Al variare dell'individuo residente nel Comune di Udine al 31 dicembre 2003, cambia il sesso, l'età, la statura e il gruppo sanguigno. Se si ripetono le pesate della medesima quantità di reagente, si ottengono valori di misura diversi. $\diamond$

È necessario individuare in modo non equivoco la popolazione di interesse.

È essenziale distinguere tra *popolazioni reali* e *popolazioni virtuali*.

- **Popolazioni reali**: sono costituite da unità che hanno un'*esistenza fisica* simultanea al momento della rilevazione; sono popolazioni effettive e quindi *finite*. Può essere esaminata in modo completo (**censimento**) o parziale (**campionamento**).

- **Popolazioni virtuali**: hanno un'*esistenza concettuale* e sono evocate dalla potenziale replicabilità a piacere della sperimentazione. Sono (potenzialmente) *infinite* e quindi esaminabili solo in modo parziale (**campionamento**), considerando il numero finito di volte con cui la sperimentazione viene ripetuta.

Esempio. Un esempio di popolazione reale è l'insieme dei residenti nel Comune di Udine al 31 dicembre 2003. Un esempio di popolazione virtuale è l'insieme di tutte le possibili replicazioni (potenzialmente infinite) della pesata di una quantità di reagente. ◇

Quando si esaminano, con riferimento a determinate caratteristiche di interesse, *tutte* le unità di una popolazione reale, si effettua un **censimento** (indagine di tipo censuario).

La **Statistica descrittiva** fornisce strumenti e metodi per descrivere le caratteristiche della popolazione, sulla base dei dati disponibili. Le finalità sono principalmente di tipo descrittivo, poiché si sintetizzano le informazioni disponibili, che riguardano la totalità della popolazione.

Un **campione** è un aggregato di unità statistiche, appartenenti ad una popolazione reale o virtuale, selezionate mediante l'esperimento di campionamento.

L'**esperimento di campionamento** è un particolare esperimento (il cui ruolo è centrale in Statistica), assimilabile all'estrazione casuale di alcuni elementi da un'urna.

Per l'inerente replicabilità dell'estrazione del campione, i dati campionari vanno interpretati come sperimentali, anche se provengono dall'osservazione di caratteristiche di alcune unità di una popolazione reale, finita. Ad esempio, le caratteristiche antropometriche di un campione di 1000 residenti nel Comune di Udine al 31 dicembre 2003.

Le popolazioni reali possono essere studiate per via campionaria o censuaria, mentre per le popolazioni virtuali la strategia campionaria è la sola possibile.

Anche quando si conduce un'indagine di tipo campionario, l'obiettivo non muta: si desidera acquisire informazione sull'intera popolazione (reale o virtuale), con riferimento a particolari caratteristiche di interesse.

Affinchè il campione porti informazioni sull'intera popolazione, la sua *estrazione* deve essere *casuale*.

La **Statistica inferenziale** fornisce strumenti e metodi per ricavare dai dati campionari informazioni sulla popolazione di riferimento e per quantificare la fiducia da accordare a tali informazioni.

L'esperimento di campionamento è un **esperimento casuale** (**aleatorio**), dal momento che risultano possibili una pluralità di esiti (campioni osservati) e prima di effettuare il campionamento non è possibile individuare con certezza quale potenziale campione verrà selezionato.

Il **Calcolo delle Probabilità** fornisce gli strumenti matematici per lo studio di esperimenti casuali, e in particolare degli esperimenti di campionamento.

In questa sede non si considererano le problematiche riconducibili alle diverse strategie di campionamento casuale, rilevanti nel caso di campionamento da popolazioni finite.

Nell'ambito della parte dedicata alla Statistica inferenziale si considereranno principalmente campioni casuali tratti da una popolazione di interesse, (virtualmente) infinita.

Quando si effettua un'indagine campionaria che coinvolge più aspetti sumultaneamente, è utile distinguere tra **esperimento programmato** e **esperimento osservativo**.

Esempio. Un esempio di esperimento programmato. Per valutare l'efficacia di un nuovo fertilizzante, si suddivide in lotti omogenei un appezzamento di terreno. Metà dei lotti, scelti casualmente, sono trattati con il nuovo prodotto e metà con quello tradizionale. Alla fine si vuole confrontare la quantità di prodotto. ◇

Esempio. Un esempio di esperimento osservativo. Per valutare i danni del fumo sull'apparato respiratorio, si seleziona un campione di individui omogenei e, classificandoli in *fumatori* e *non fumatori*, si riporta il numero di malattie riscontrate nell'ultimo anno. ◇

In entrambi i casi le unità statistiche possono subire un **trattamento** (fertilizzante, fumo) e forniscono una **risposta** (prodotto, malattie).

La differenza è che solo nel primo caso lo sperimentatore può decidere come assegnare il trattamento alle singole unità.

STATISTICA DESCRITTIVA

(Pace e Salvan, *Introduzione alla Statistica I - Statistica Descrittiva*, CEDAM Padova, 1996)

Alcune nozioni di base saranno utili anche per la Statistica inferenziale.

Come premessa ad una analisi inferenziale, è possibile effettuare uno studio descrittivo con riferimento al particolare campione osservato.

Si suppone che i dati siano già stati acquisiti e che siano disponibili nella forma di **matrice dei dati**, di cui la tabella sottostante è un esempio. Questi sono i cosidetti **dati grezzi**.

Unità u	SESSO	ETÀ (a.c.)	LIVISTR (*)	DIST (km)
Andrea	M	28	2	5.0
Claudio	M	17	4	7.5
Lucia	F	20	4	12.0
Giuseppe	M	32	2	3.2
Mara	F	16	1	(**)
Luca	M	34	2	12.3
Aldo	M	18	1	25.0
Arianna	F	25	2	7.7

(*) con codificazione numerica: 1 per *Licenza Elem.*, 2 per *Licenza Media*, 3 per *Diploma Sec.*, 4 per *Laurea*; (**) dato mancante.

(*Da Pace e Salvan, 1996*)

- Ogni *riga* corrisponde ad una unità statistica e contiene i valori su essa rilevati delle variabili di interesse (sesso, età in anni compiuti, livello di istruzione, distanza dal luogo di lavoro).

- Ogni *colonna* corrisponde ad una variabile e contiene i valori di tale variabile rilevati sulle varie unità.

Si forniscono alcune definizioni utili anche in ambito inferenziale.

Definizione. Una **variabile** è una caratteristica delle unità statistiche che, al variare dell'unità, può assumere una pluralità di valori.

Definizione. Le **modalità** di una variabile sono i valori che essa può assumere (e si presumono noti preliminarmente). Sono, in genere, aggettivi, valori numerici, espressioni verbali.

Le variabili si indicano con le lettere maiuscole, ad esempio Y, mentre una generica modalità si indica con la corrispondente lettera minuscola, y. L'insieme $\mathcal{Y}$ è l'insieme di tutte le possibili modalità di Y.

Esempio. Y="SESSO", con $\mathcal{Y} = \{M, F\}$; Y="LIVELLO DI ISTRUZIONE", con $\mathcal{Y} = \{1, 2, 3, 4\}$, avendo scelto la codifica della tabella precedente; Y="ETÀ (a.c.)", con $\mathcal{Y} = \{0, 1, 2, \ldots\}$; Y="REDDITO", con $\mathcal{Y} = \mathbf{R}^+$, anche se si può pensare che il reddito vari su scaglioni prefissati. $\diamond$

Le variabili si possono classificare nel seguente modo.

- Variabili **qualitative**, se le modalità sono espresse in forma verbale.

  In particolare, variabili **qualitative sconnesse o nominali**, per le quali non è possibile individuare un ordinamento naturale delle modalità (ad esempio, "RELIGIONE PROFESSATA", "COLORE DEGLI OCCHI") e variabili **qualitative ordinali**, per le quali è invece possibile individuare un ordinamento naturale delle modalità (ad esempio, "LIVELLO DI ISTRUZIONE").

- Variabili **quantitative**, se le modalità sono espresse in forma numerica (da non confondere con le codifiche numeriche).

  In particolare, variabili **quantitative discrete**, se $\mathcal{Y}$ è un insieme finito o al più numerabile (ad esempio, "ETÀ (a.c.)") e variabili **quantitative continue**, se $\mathcal{Y}$ è un insieme continuo (ad es. "DISTANZA DAL LUOGO DI LAVORO", "ALTEZZA", "REDDITO"). Si noti che la continuità va intesa come *potenziale continuità* o come opportuno *riferimento semplificativo*.

Si consideri una popolazione finita $\mathcal{U}$, oggetto di studio, costituita da N unità statistiche, in simboli $|\mathcal{U}| = N$. La popolazione viene esaminata completamente (indagine censuaria) con riferimanto alle k variabili di interesse $Y_1, \ldots, Y_k$.

Si considerano **analisi statistiche univariate**, che prendono in esame una sola variabile, indicata con Y.

La variabile Y viene rilevata su $\mathcal{U}$ e si ottiene la seguente successione di valori (modalità) $(y_1, \ldots, y_i, \ldots, y_N)$, dove y_i, $i = 1, \ldots, N$, è il valore (modalità) assunto da Y con riferimento all'unità $u_i \in \mathcal{U}$.

Per distinguere tra variabile e risultato della sua rilevazione sulla popolazione $\mathcal{U}$ si introduce la seguente definizione.

Definizione. Una **variabile statistica** è una *corrispondenza empirica* tra le unità statistiche e le modalità ad esse associate, con riferimento alla variabile di interesse Y. In pratica, una variabile statistica corrisponde alla rilevazione $(y_1, \ldots, y_i, \ldots, y_N)$, che può essere vista come una colonna della matrice dei dati.

La stessa variabile rilevata su popolazioni diverse dà luogo, in genere, a variabili statistiche differenti. Spesso si usa il simbolo Y anche per indicare la variabile statistica.

Esempio. Con riferimento alla Tabella di pagina 8, alla variabile $Y =$"ETÀ (a.c.)" corrisponde la variabile statistica $(28, 17, 20, 32, 16, 34, 18, 25)$, con $N = 8$. $\diamond$

Poiché non tutte le modalità potenzialmente assumibili dalla variabile Y possono venire effettivamente rilevate in una popolazione, può essere utile la seguente definizione.

Definizione. Si dice **supporto** della variabile statistica Y, e si indica con S_Y, l'insieme delle modalità di Y effettivamente osservate nella popolazione $\mathcal{U}$; $S_Y = \{y_1, \ldots, y_j, \ldots, y_J\}$.

Si noti che $J \leq N$.

Le modalità osservate, che concorrono a costituire S_Y, sono tra loro distinte, cioé vanno prese una volta sola anche se ripetute.

Nel caso di variabili qualitative ordinali e quantitative si suppone che le modalità appartenenti al supporto vengano ordinate secondo un ordine crescente. Ad esempio, se Y è quantitativa, si considera $y_1 < y_2 < \ldots < y_J$.

Esempio. Con riferimento alla variabile statistica "ETÀ (a.c.)" riportata nella Tabella di pagina 8, il supporto è $S_Y = \{16, 17, 18, 20, 25, 28, 32, 34\}$, mentre $\mathcal{Y} = \{0, 1, 2, \ldots\}$. $\diamond$

Se l'ordine con cui le unità statistiche vengono rilevate non è importante, può essere utile passare dalla variabile statistica (dati in forma grezza) ad una **tabella di frequenza**.

Definizione. Se $y_j \in S_Y$, $j = 1, \ldots, J$, è una delle modalità osservate di Y, si dice **frequenza assoluta** di y_j il numero di volte che y_j risulta osservata. Si indica con f_j. Evidentemente, $f_j > 0$, $j = 1, \ldots, J$, e $\sum_{j=1}^{J} f_j = N$.

Definizione. Sia Y una variabile statistica con supporto $S_Y = \{y_1, \ldots, y_J\}$. La lista delle modalità osservate accompagnate dalle rispettive frequenze assolute è detta **distribuzione di frequenza assoluta**.

Si rappresenta mediante una **tabella di frequenza** del tipo

Modalità	y_1	$\cdots$	y_j	$\cdots$	y_J	Totale
Frequenza	f_1	$\cdots$	f_j	$\cdots$	f_J	$\sum_{j=1}^{J} f_j$

Esempi di tabelle di frequenza (assoluta) ricavabili dai dati grezzi di pag. 8.

SESSO	Frequenza
M	5
F	3
Totale	8

LIVISTR	Frequenza
Licenza Media	2
Diploma Sec.	4
Laurea	2
Totale	8

Una tabella di frequenza riferita ad una variabile statistica qualitativa è detta **serie statistica**.

Se la variabile statistica è quantitativa continua, si osservano, a meno di effetti di arrotondamento, tante modalità distinte quante sono le unità statistiche, ossia $J = N$. Quindi, S_Y corrisponde all'insieme dei dati grezzi e $f_j = 1$, $j = 1, \ldots, J$.

Questo può accadere, in alcuni casi, anche con variabili statistiche quantitative discrete.

È conveniente definire **classi di modalità** contigue e contare le unità che appartengono a ciascuna classe. Si ottiene la seguente tabella di frequenza (assoluta) con modalità raggruppate in classi

Classi	$y_0 \dashv y_1$	$\cdots$	$y_{j-1} \dashv y_j$	$\cdots$	$y_{J-1} \dashv y_J$	Totale
Freq.	f_1	$\cdots$	f_j	$\cdots$	f_J	$\sum_{j=1}^{J} f_j$

dove f_j è la frequenza assoluta associata alla classe $y_{j-1} \dashv y_j$, che corrisponde all'intervallo $(y_{j-1}, y_j]$. Analogamente, $y_{j-1} \vdash y_j$ corrisponde all'intervallo $[y_{j-1}, y_j)$ e y_J- indica $(y_J, +\infty)$.

Una tabella di frequenza così ottenuta è detta **seriazione statistica**.

Esempio di seriazione ottenuta dai dati grezzi di pag. 8.

DIST	Frequenza
$0 \dashv 5$	2
$5 \dashv 15$	4
$15-$	1
Totale	7

Le classi vanno definite di modo che

- non siano né troppe né troppo poche;
- siano disgiunte;
- comprendano tutte le modalità osservate.

Le classi non hanno necessariamente un'ampiezza costante.

Definizione. La **frequenza relativa** di una modalità y_j, o di una classe di modalità (ad esempio $y_{j-1} \dashv y_j$), è la frazione o proporzione p_j di unità statistiche rilevate portatrici di tale modalità o classe di modalità.

Se f_j è la associata frequenza assoluta, allora la frequenza relativa p_j è tale che

$$p_j = \frac{f_j}{\sum_{j=1}^{J} f_j} = \frac{f_j}{N}, \quad j = 1, \ldots, J.$$

Evidentemente, $p_j > 0$, $j = 1, \ldots, J$, e $\sum_{j=1}^{J} p_j = 1$.

Si possono definire anche le frequenze relative percentuali, definite come $p_j 100$, $j = 1, \ldots, J$.

Le frequenze relative sono utili per percepire il peso delle varie modalità e per operare confronti tra diverse popolazioni.

Si consideri, a proposito, la seguente tabella che fornisce la distribuzione per sesso della popolazione residente in Italia (confini attuali) in due censimenti; i dati sono espressi in migliaia.

		Freq. ass.	Freq. rel.	Freq. rel. %
1861	M	13399	0.5089	50.89
	F	12929	0.4911	49.11
	Totale	26328	1	100
1981	M	27506	0.4863	48.63
	F	29051	0.5137	51.37
	Totale	56557	1	100

Se $S_Y = \{y_1\}$, allora $J = 1$, $f_1 = N$, $p_1 = 1$ e la variabile statistica Y è detta **degenere**.

Quando si hanno variabili qualitative ordinali o quantitative, può essere utile considerare la seguente definizione.

Definizione. Sia Y una variabile statistica qualitativa ordinale o quantitativa con la associata tabella di frequenza assoluta o relativa. La **frequenza assoluta cumulata** F_j o, analogamente, la **frequenza relativa cumulata** P_j definiscono la frequenza assoluta o relativa di modalità o classi di modalità non superiori alla j-esima, $j = 1, \ldots, J$.

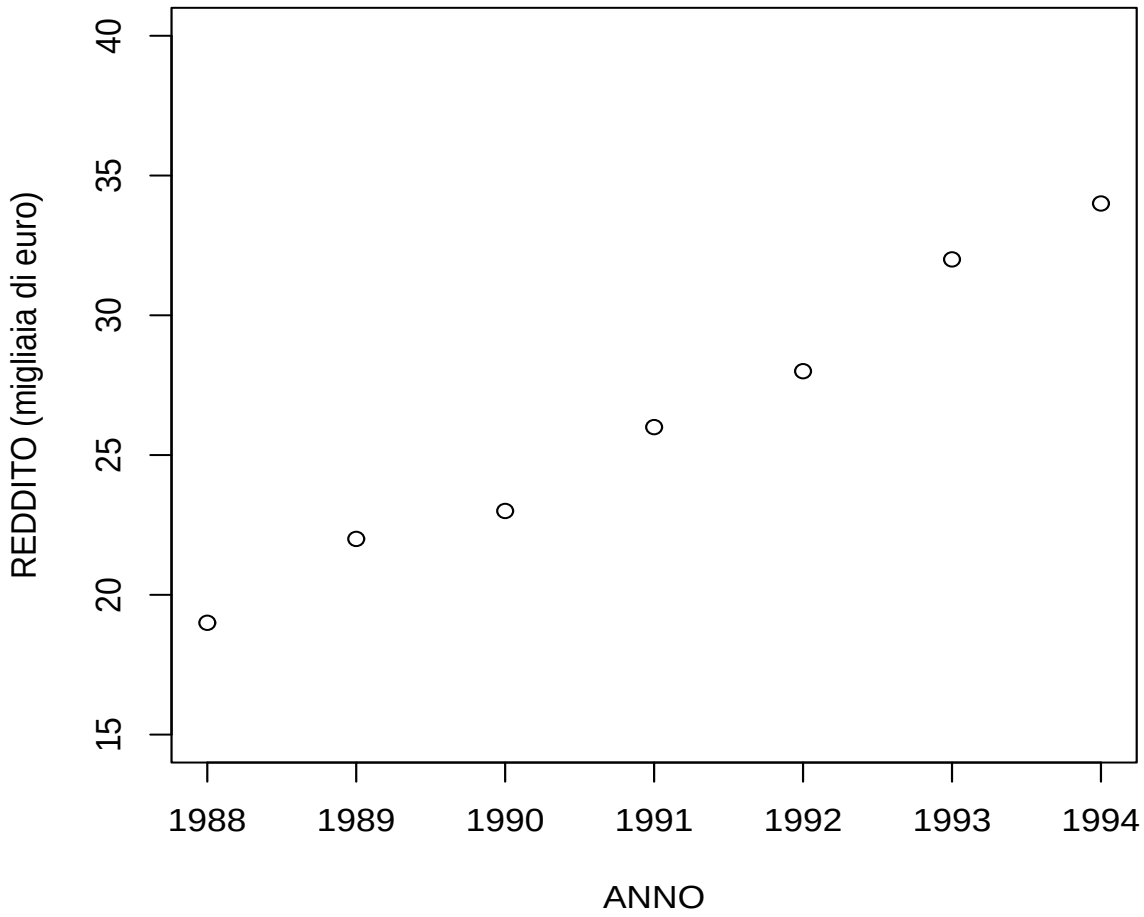
Più precisamente

$$F_j = \sum_{i=1}^{j} f_i, \quad P_j = \sum_{i=1}^{j} p_i, \quad j = 1, \ldots, J.$$

Evidentemente, $F_1 = f_1$, $F_J = N$, $P_1 = p_1$, $P_J = 1$.

Oltre alle tabelle di frequenza, risulta utile introdurre alcune rappresentazioni grafiche, dette **diagrammi statistici**.

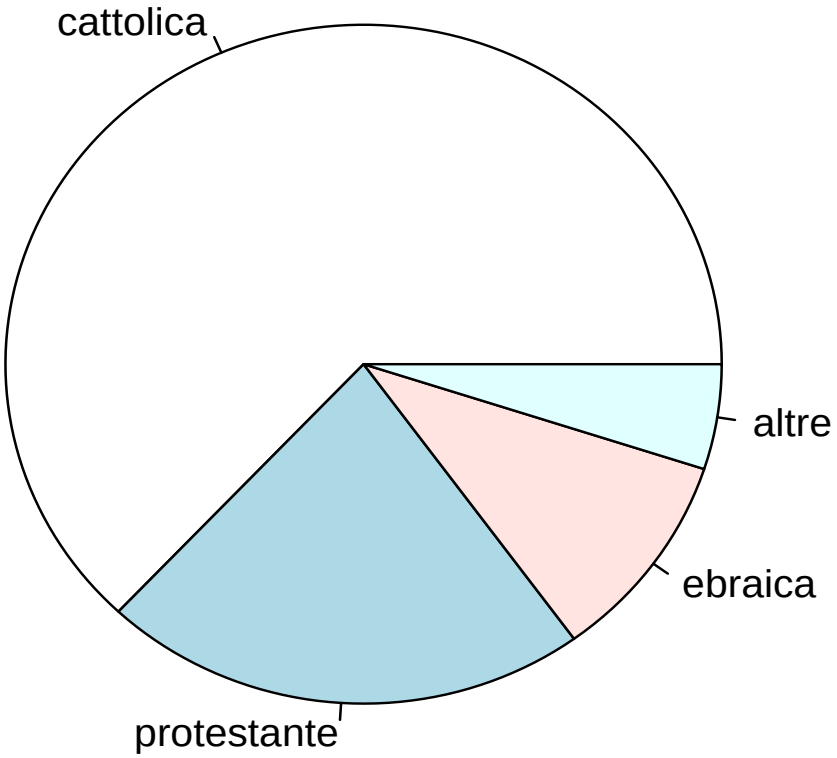
Per rappresentare dati quantitativi in forma grezza si usano semplici rappresentazioni sul piano cartesiano, dove ogni punto indica la modalità assunta dalla singola unità statistica.



Per rappresentare tabelle di frequenza relativa o assoluta ci sono varie rappresentazioni grafiche, utili per le varie tipologie di variabili in esame.

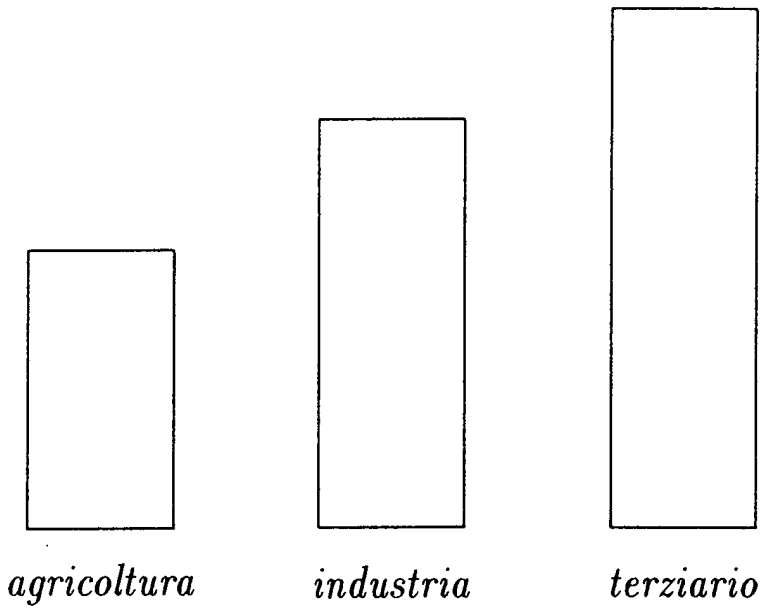
Per rappresentare serie statistiche sconnesse (riferite a variabili qualitative sconnesse) si possono utilizzare **diagrammi circolari**.

Criterio costruttivo: angolo al centro (area dei settori circolari) proporzionale alla frequenza della modalità.



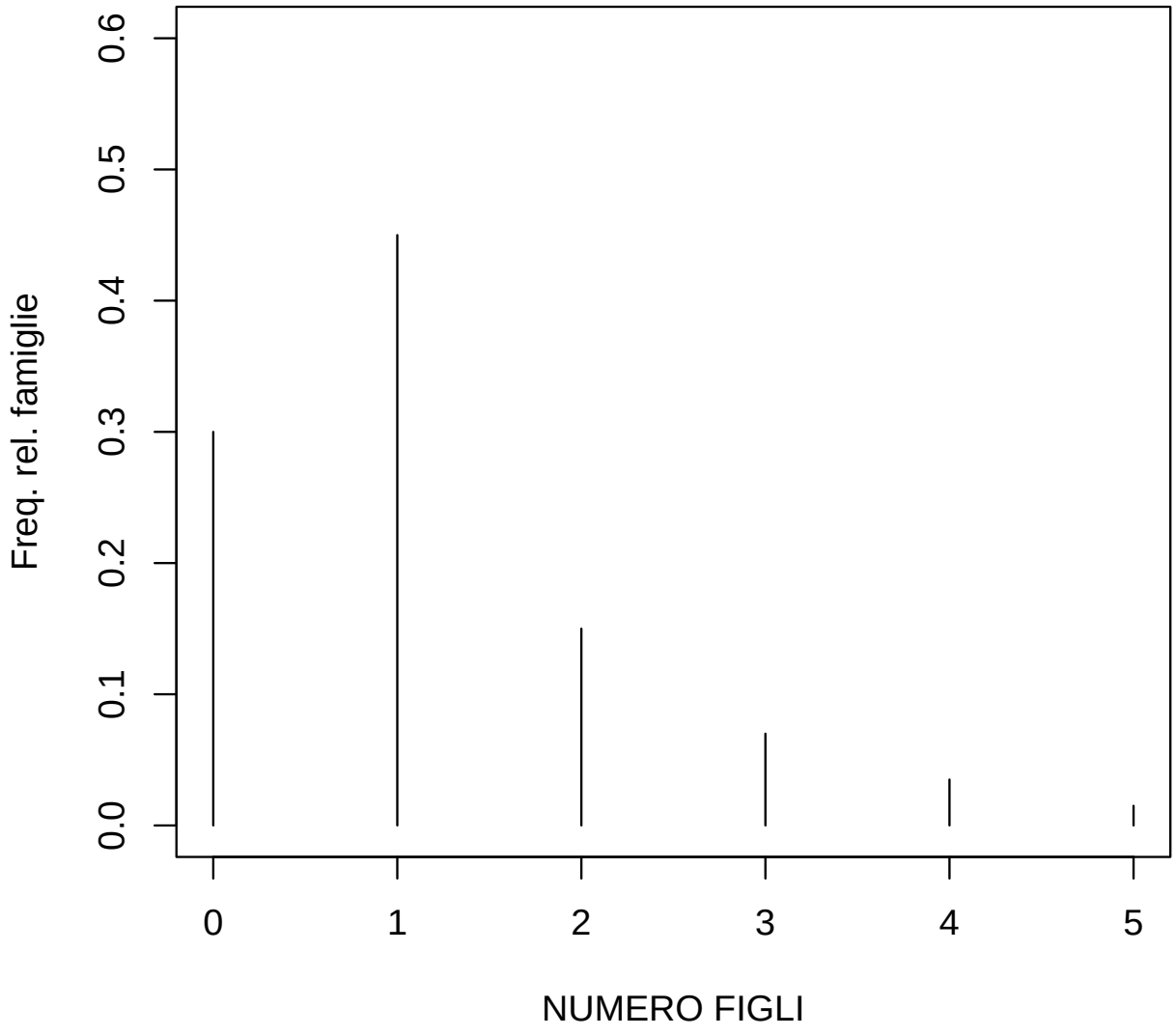
Per rappresentare serie statistiche (riferite a variabili qualitative sconnesse o ordinali) si possono utilizzare **diagrammi con rettangoli distanziati**.

Criterio costruttivo: altezza del rettangolo proporzionale alla frequenza della modalità.



Per rappresentare distribuzioni di frequenza assoluta o relativa (riferite a variabili quantitative discrete) si possono utilizzare **diagrammi con bastoncini**.

Criterio costruttivo: altezza del bastoncino proporzionale o pari alla frequenza (relativa o assoluta) della modalità.



Per rappresentare distribuzioni di frequenza assoluta o relativa con modalità raggruppate in classi (seriazioni) si possono utilizzare gli **istogrammi**.

L'istogramma è un insieme di rettangoli adiacenti, ognuno rappresentativo di una classe, posti su un piano cartesiano.

Il rettangolo corrispondente alla classe j-esima, ad esempio $y_{j-1} \dashv y_j$, $j = 1, \ldots, J$, ha come base l'intervallo $(y_{j-1}, y_j]$ e

- altezza (e quindi area) proporzionale a, oppure pari a, $f_j/(y_j - y_{j-1})$: **istogramma delle frequenze assolute**;

- altezza (e quindi area) proporzionale a, oppure pari a, $p_j/(y_j - y_{j-1})$: **istogramma delle frequenze relative**.

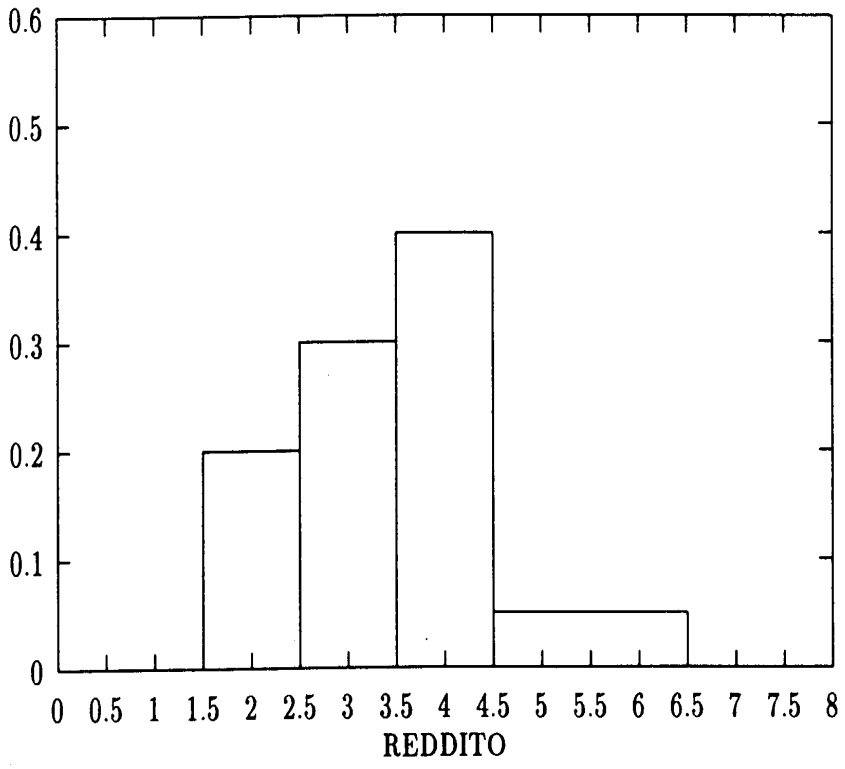
Se le classi estreme sono aperte, ad esempio $-y_1$ e $y_{J-1}-$, vanno chiuse scegliendo opportunamente gli estremi y_0 e y_J.

L'istogramma viene utilizzato usualmente con riferimento a variabili statistiche quantitative continue. In alcuni casi può essere utilizzato anche per descrivere distribuzioni di frequenza associate a variabili statistiche quantitative discrete.

Si consideri la seguente seriazione

Reddito	1.5 ⊣2.5	2.5⊣3.5	3.5⊣4.5	4.5⊣6.5	Tot.
freq. rel.	0.2	0.3	0.4	0.1	1

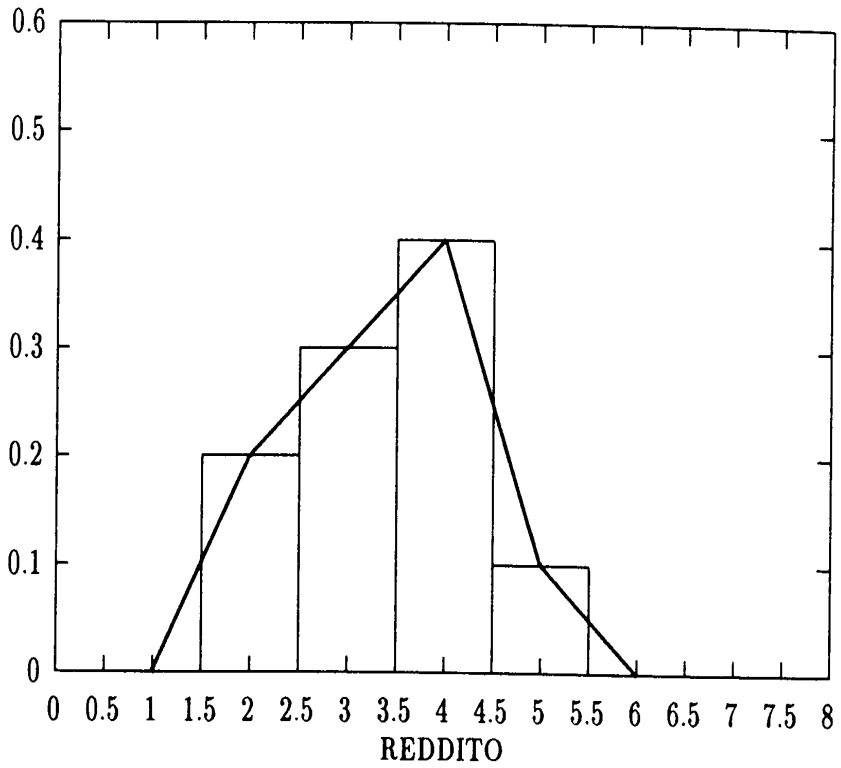
e l'associato istogramma



Un **poligono di frequenza** è uno smussamento locale dell'istogramma.

Per costruirlo si introducono due classi adiacenti alle classi esterne $y_0 \dashv y_1$ e $y_{J-1} \dashv y_J$, ognuna con ampiezza uguale alla classe vicina e frequenza assoluta pari a zero.

Il poligono di frequenza si ottiene individuando i punti medi dei lati superiori dei rettangoli dell'istogramma e tracciando la associata linea spezzata.



Per alcuni scopi di presentazione dei dati, riferiti ad una variabile statistica Y, una tabella di frequenza può non rappresentare una sintesi sufficientemente concisa.

In molti casi risulta interessante indagare i seguenti aspetti dei dati:

- la **posizione**: il *centro* dei dati, espresso nell'ordine di grandezza di Y;

- la **variabilità**: la *dispersione* dei dati.

Spesso interessano anche altri aspetti legati alla **forma** della distribuzione di frequenza, quali l'asimmetria e la pesantezza delle code.

Nel seguito si presenteranno alcuni indici sintetici che descrivono la posizione e la variabilità di una variabile statistica.

Si considerano i principali **indici di posizione**: la media aritmetica, la mediana e la moda.

La **media aritmetica** si può calcolare per una *variabile quantitativa* Y e si indica con $E(Y)$, con μ_Y o semplicemente con μ.

Esempio. Sia $Y = (27, 30, 30)$ la variabile statistica che descrive i voti riportati in tre esami da uno studente. La media aritmetica dei voti è $29 = (27 + 30 + 30)/3$. Si noti che 29 non corrisponde a nessuno dei voti ottenuti. Se $Y = (28, 30, 30)$, la media aritmetica dei voti è $28.3 = (28+30+30)/3$, che non corrisponde a nessuna potenziale modalità per Y. In entramni i casi la media sintetizza i valori osservati indicandone un centro. $\diamond$

Se si dispone dei *dati grezzi* $Y = (y_1, \ldots, y_N)$, allora la media aritmetica corriponde a

$$E(Y) = \frac{1}{N} \sum_{i=1}^{N} y_i.$$

Si noti che E(Y) corrisponde al valore di equiripartizione sulle unità statistiche del totale delle osservazioni.

Se, con riferimento ad una variabile statistica quantitativa discreta Y, si dispone della *tabella di frequenza assoluta o relativa*, allora

$$E(Y) = \frac{1}{N} \sum_{j=1}^{J} y_j f_j = \sum_{j=1}^{J} y_j p_j.$$

Se, con riferimento ad una variabile statistica quantitativa continua Y, si dispone della *tabella di frequenza assoluta o relativa con modalità raggruppate in classi* (ad esempio $y_{j-1} \dashv y_j$, $j = 1, \ldots, J$), è necessario calcolare il punto centrale $y_j^c = (y_{j-1} + y_j)/2$, $j = 1, \ldots, J$, delle singole classi.

In questo caso

$$E(Y) = \frac{1}{N} \sum_{j=1}^{J} y_j^c f_j = \sum_{j=1}^{J} y_j^c p_j.$$

Questa procedura per il calcolo di $E(Y)$ è equivalente a quella che si definisce quando si dispone dei dati grezzi se viene soddisfatta una delle seguenti ipotesi

- le osservazioni che cadono in una classe coincidono con il punto centrale della classe;

- le osservazioni sono distribuite in modo uniforme nella classe di appartenenza.

Non è detto che $E(Y)$ coincida con una delle modalità osservate o osservabili.

La media aritmetica risente della presenza di osservazioni anomale.

Esistono altre tipologie di medie, che non vengono considerate in questa sede.

Esempio. Si consideri la seguente tabella di frequenza

y_j	f_j
0	109
1	65
2	22
3	3
4	1
Totale	200

È immediato concludere che $E(Y) = 122/200 = 0.61$
$\diamond$

Esempio. Si consideri la seguente tabella di frequenza con modalità raggruppate in classi

Classe	$0 \dashv 10$	$10 \dashv 15$	$15 \dashv 20$	Totale
freq. rel.	0.30	0.52	0.18	1

I valori centrali delle classi sono, rispettivamente, $y_1^c = 5$, $y_2^c = 12.5$ e $y_3^c = 17.5$, da cui si conclude che

$$E(Y) = 5 \cdot 0.30 + 12.5 \cdot 0.52 + 17.5 \cdot 0.18 = 11.15.$$

$\diamond$

Se ci sono classi aperte, il punto centrale viene individuato dopo aver convenientemente "chiuso la classe".

La media aritmetica soddisfa le seguenti **proprietà**.

1) Proprietà di Cauchy: sia $S_Y = \{y_1, \ldots, y_J\}$, con $y_1 < \cdots < y_J$, allora

$$y_1 \leq E(Y) \leq y_J.$$

La media è compresa tra il più piccolo e il più grande valore osservato.

Infatti, per ogni $j = 1, \ldots, J$

$$y_1 \leq y_j \leq y_J \quad \Rightarrow \quad y_1 p_j \leq y_j p_j \leq y_J p_j \quad \Rightarrow$$

$$\sum_{j=1}^{J} y_1 p_j \leq \sum_{j=1}^{J} y_j p_j \leq \sum_{j=1}^{J} y_J p_j \quad \Rightarrow$$

$$y_1 \sum_{j=1}^{J} p_j \leq \sum_{j=1}^{J} y_j p_j \leq y_J \sum_{j=1}^{J} p_j,$$

da cui si ottiene la tesi, poiché $\sum_{j=1}^{J} p_j = 1$.

2) Proprietà di baricentro: sia $Y - E(Y)$ la variabile scarto di Y dalla sua media $E(Y)$, allora

$$E(Y - E(Y)) = 0.$$

Infatti, considerando i dati grezzi,

$$E(Y - E(Y)) = \frac{1}{N}\sum_{i=1}^{N}(y_i - E(Y)) = \frac{1}{N}\sum_{i=1}^{N} y_i - \frac{1}{N}\sum_{i=1}^{N} E(Y)$$

$$= E(Y) - \frac{1}{N}\, NE(Y) = 0.$$

3) Proprietà di linearità: sia $aY + b$, $a, b \in \mathbf{R}$, una trasformata lineare della variabile Y, allora

$$E(aY + b) = aE(Y) + b.$$

Infatti, considerando i dati grezzi,

$$E(aY + b) = \frac{1}{N}\sum_{i=1}^{N}(ay_i + b) = \frac{1}{N}\sum_{i=1}^{N} ay_i + \frac{1}{N}\sum_{i=1}^{N} b$$

$$= a\frac{1}{N}\sum_{i=1}^{N} y_i + \frac{1}{N}\, Nb = aE(Y) + b.$$

La **mediana** si può calcolare per una *variabile qualitativa ordinale o quantitativa* Y e si indica con $y_{0.5}$.

È quel valore di Y che, rispetto all'ordinamento non decrescente delle osservazioni (dati grezzi), risulta preceduto e seguito dalla stessa porzione di osservazioni (il 50%), a meno degli effetti di discretezza.

Definizione. La mediana di una variabile statistica Y corrisponde a ogni valore $y_{0.5}$ che soddisfa simultaneamente alle seguenti condizioni:

- almeno il 50% delle unità statistiche presenta modalità inferiori o pari a $y_{0.5}$;
- almeno il 50% delle unità statistiche presenta modalità superiori o pari a $y_{0.5}$.

Se si dispone dei *dati grezzi* $Y = (y_1, \ldots, y_N)$, ordinati secondo un ordinamento non decrescente, allora la mediana di $y_{0.5}$ corrisponde

- alla modalità che si trova nella posizione $(N+1)/2$, se N è **dispari**, cioè $y_{0.5} = y_{(N+1)/2}$;
- alle modalità che si trovano nelle posizioni $N/2$ e $(N/2)+1$, se N è **pari**, cioè $y_{0.5} = y_{N/2}$ e $y_{0.5} = y_{(N/2)+1}$.

Si noti che, se $y_{N/2}$ e $y_{(N/2)+1}$ non coincidono, la mediana può non essere unica.

Nel caso di variabili quantitative con N pari, si può avere anche un intervallo di valori $[y_{N/2}, y_{(N/2)+1}]$ che soddisfano alla definizione di mediana. In questo caso si può prendere il punto di mezzo come *mediana convenzionale*.

Esempio. Si consideri la variabile statistica qualitativa ordinale Y che descrive il voto di licenza media di $N = 5$ studenti, *opportunamente ordinati*,

$$Y = (sufficiente, sufficiente, buono, buono, ottimo).$$

Poiché N è dispari, $y_{0.5} = y_{(N+1)/2} = y_3 = buono$.

Si può anche verificare che $buono$ è l'unica modalità che verifica le condizioni di pag. 33.

Se invece

$$Y = (suff., suff., suff., buono, buono, ottimo),$$

N è pari, quindi $y_{0.5} = y_{N/2} = y_3 = suff.$ e $y_{0.5} = y_{(N/2)+1} = y_4 = buono$.

Si può anche verificare che sia $suff.$ sia $buono$ verificano le condizioni di pag. 33.

Infine, se

$$Y = (suff., suff., buono, buono, ottimo, ottimo),$$

N è pari, ma $y_{0.5} = y_{N/2} = y_3 = buono$ e $y_{0.5} = y_{(N/2)+1} = y_4 = buono$; quindi, $buono$ è l'unica mediana. $\diamond$

Esempio. Si consideri la variabile statistica quantitativa discreta Y che descrive il numero di puntate, di una serie televisiva, viste da 8 famiglie

$$Y = (0, 1, 3, 3, 5, 6, 6, 6);$$

i valori osservati sono stati ordinati opportunamente.

Poiché N è pari, $y_{0.5} = y_{N/2} = y_4 = 3$ e $y_{0.5} = y_{(N/2)+1} = y_5 = 4$. In questo caso, sia 3 che 4 sono valori mediani e, in generale, ogni punto dell'intervallo $[3, 4]$ è un valore mediano, dato che verifica le condizioni di pag. 33.

Se invece

$$Y = (0, 1, 3, 3, 5, 6, 6),$$

N è dispari e c'è un'unica mediana $y_{0.5} = y_{(N+1)/2} = y_4 = 3$. $\diamond$

Se non si dispone dei dati grezzi, ma soltanto della *distribuzione di frequenza relativa o assoluta* corrispondente, si può operare nel seguente modo.

Sia Y una variabile statistica qualitativa ordinale o quantitativa, con supporto $S_Y = \{y_1, \ldots, y_J\}$, dove le modalità si suppongono ordinate in senso crescente.

Se sono note le associate *frequenze assolute* f_j, $j = 1, \ldots, J$, e quindi la dimensione N della popolazione, la mediana corrisponde,

- se N è **dispari**, alla modalità y_j che presenta la frequenza assoluta cumulata F_j più piccola tale che $F_j \geq (N+1)/2$;

- se N è **pari**, alla modalità y_j che presenta la frequenza assoluta cumulata F_j più piccola tale che $F_j \geq N/2$ e alla modalità y_j che presenta la frequenza assoluta cumulata F_j piccola tale che $F_j \geq (N/2) + 1$.

Nel caso con N pari si possono avere due valori mediani distinti o più di due, se si considerano variabili quantitative.

Se sono note solo le associate *frequenze relative* p_j, $j = 1, \ldots, J$, e quindi la dimensione N della popolazione non risulta nota, allora la mediana corrisponde ad ogni valore $y_{0.5}$ che soddisfa simultaneamante le seguenti condizioni:

- la frequenza relativa di modalità inferiori o pari a $y_{0.5}$ è maggiore o uguale a 0.5;

- la frequenza relativa di modalità superiori o pari a $y_{0.5}$ è maggiore o uguale a 0.5.

La mediana è un indice di posizione *robusto* rispetto a valori anomali dei dati.

Se non si dispone dei dati grezzi, ma soltanto della *distribuzione di frequenza relativa o assoluta con modalità raggruppate in classi*, si può operare allo stesso modo.

Quindi, si individueranno *classi mediane*.

Esempio. Sia Y la variabile quantitativa discreta che descrive il numero di componenti delle famiglie residenti al censimento 1981.

Si consideri la tabella di frequenza riferita alla regione Liguria

No. componenti	f	F	p	P
1	197906	197906	0.272	0.272
2	203709	401615	0.281	0.553
3	168536	570151	0.232	0.785
4	117509	687660	0.162	0.947
5	29727	717387	0.041	0.988
6	6577	723964	0.009	0.997
7	1707	725671	0.002	0.999
8 o più	906	726577	0.001	1
Totale	726577		1	

Poiché $N = 726577$ è dispari, la mediana è unica e corrisponde alla modalità della famiglia che si trova nella posizione $(N+1)/2 = 363289$, dopo avere ordinato le famiglie secondo il numero crescente di componenti. Tale famiglia presenta modalità 2, quindi $y_{0.5} = 2$.

Si noti che a 2 corrisponde la frequenza assoluta cumulata più piccola che risulta maggiore o uguale a $(N+1)/2 = 363289$.

Se si considerano le frequenze relative, $y_{0.5} = 2$ è l'unico valore che verifica entrambe le condizioni di pag. 37.
$\diamond$

Si consideri la tabella di frequenza riferita alla regione Campania

No. componenti	f	F	p	P
1	225641	225641	0.144	0.144
2	304325	529966	0.194	0.338
3	278879	808845	0.178	0.516
4	355488	1164333	0.226	0.742
5	228494	1392827	0.146	0.888
6	98924	1491751	0.063	0.951
7	42894	1534645	0.027	0.978
8 o più	34999	1569644	0.022	1
Totale	1569644		1	

Poiché $N = 1569644$ è pari, la mediana corrisponde alla modalità della famiglia che si trova nella posizione $N/2 = 784822$ e alla modalità della famiglia che si trova nella posizione $(N/2) + 1 = 784823$, dopo avere ordinato le famiglie secondo il numero crescente di componenti. Tali famiglie presentano la stessa modalità 3, quindi $y_{0.5} = 3$.

Si noti che a 3 corrisponde la frequenza assoluta cumulata più piccola che risulta maggiore o uguale sia a $N/2 = 784822$ che a $(N/2) + 1 = 784823$.

Se si considerano le frequenze relative, $y_{0.5} = 3$ è l'unico valore che verifica entrambe le condizioni di pag. 37.
$\diamond$

La mediana può venire interpretata come una particolarizzazione della nozione generale di **quantile** di livello α, con $\alpha \in (0,1)$, indicato con la scrittura y_α.

Data una *variabile qualitativa ordinale o quantitativa* Y, y_α è quel valore che, rispetto all'ordinamento non decrescente delle osservazioni (dati grezzi), risulta preceduto da $\alpha 100\%$ osservazioni e seguito da $(1-\alpha)100\%$ osservazioni, a meno degli effetti di discretezza.

Definizione. Il quantile di livello α, con $\alpha \in (0,1)$, di una variabile statistica Y corrisponde a ogni valore y_α che soddisfa simultaneamente alle seguenti condizioni:

- almeno $\alpha 100\%$ unità statistiche presenta modalità inferiori o pari a y_α;
- almeno $(1-\alpha)100\%$ unità statistiche presenta modalità superiori o pari a y_α.

È evidente che, se $\alpha = 0.5$, si ottiene la definizione di mediana.

I quantili di livello $\alpha = 0.25, 0.5, 0.75$ vengono chiamati **quartili**.

I quantili di livello $\alpha = 0.10, 0.20, \ldots, 0.90$ vengono chiamati **decili**.

I quantili di livello $\alpha = 0.01, 0.02, \ldots, 0.99$ vengono chiamati **percentili**.

Se si dispone dei *dati grezzi* $Y = (y_1, \ldots, y_N)$, ordinati secondo un ordinamento non decrescente, allora y_α corrisponde

- alla modalità che si trova nella posizione $[N\alpha] + 1$, se $N\alpha$ è un numero **non intero**, cioè $y_\alpha = y_{[N\alpha]+1}$;

- alle modalità che si trovano nelle posizioni $N\alpha$ e $N\alpha + 1$, se $N\alpha$ è un numero **intero**, cioè $y_\alpha = y_{N\alpha}$ e $y_\alpha = y_{N\alpha+1}$.

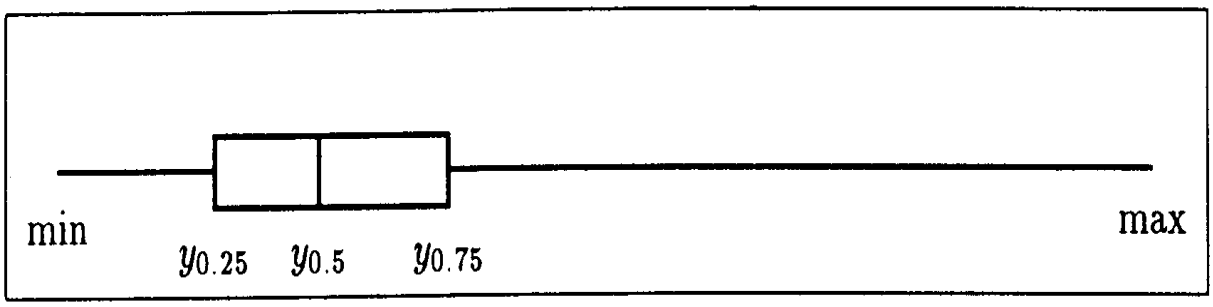
Si noti che, se $y_{N\alpha}$ e $y_{N\alpha+1}$ non coincidono, il quantile può non essere unico.

Nel caso di variabili quantitative con $N\alpha$ intero, si può avere anche un intervallo di valori $[y_{N\alpha}, y_{N\alpha+1}]$ che soddisfano alla definizione di quantile.

Con $[x]$, $x \in \mathbf{R}$ si indica la parte intera del numero x; ad esempio, $[12.274] = 12$.

Con riferimento alla nozione di quantile si possono fare considerazioni analoghe a quelle introdotte per la mediana.

Spesso si ricorre a una rappresentazione grafica, detta **diagramma a scatola e baffi** (*box and whiskers plot*) del tipo illustrato dalla figura sottostante.



La *scatola* contiene il 50% centrale della distribuzione di frequenza ed è delimitata dal primo quartile $y_{0.25}$ e dal terzo quartile $y_{0.75}$.

In corrispondenza della mediana $y_{0.5}$ viene tracciata una *linea verticale*.

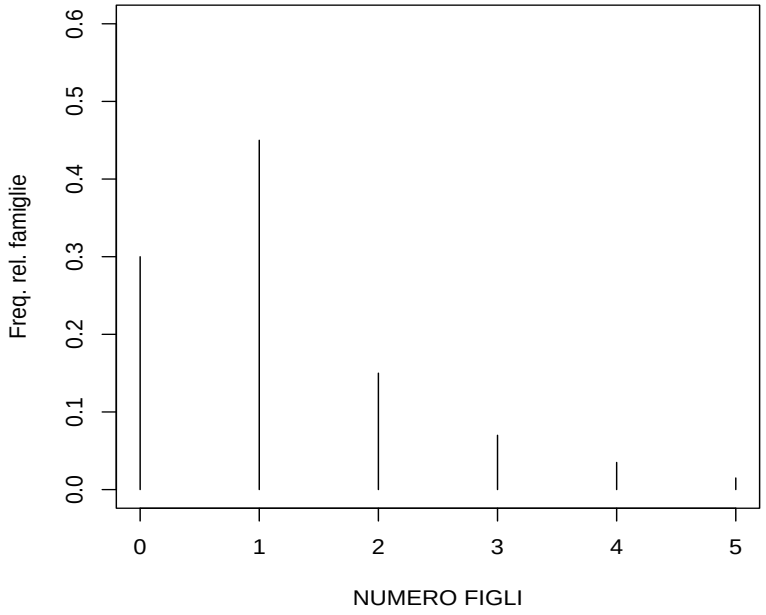
I *baffi* si prolungano fino al valore minimo e massimo osservati o fino ai percentili $y_{0.01}$ e $y_{0.99}$.

La **moda** si può calcolare per una *variabile qualitativa o quantitativa* Y e si indica con y_{mo}.

Definizione. La moda di una variabile statistica Y corrisponde al valore y_{mo} del supporto S_Y a cui è associata la frequenza, relativa o assoluta, più alta.

La moda è la modalità più comune e non è detto che sia unica.

Dal grafico sottostante si conclude che la moda $y_{mo} = 1$ ed è unica; in questo caso la distribuzione di frequenza è detta **unimodale**.



Ci possono essere anche distribuzioni **bimodali** o **multimodali**.

Nel caso in cui si abbia una tabella di frequenza con modalità raggruppate in classi, si può individuare la **classe modale**, se le *classi hanno tutte la stessa ampiezza*.

Si introducono alcuni **indici di variabilità**, utili per *variabili quantitative*.

Non si considerano, in questa sede, gli indici di variabilità per *variabili qualitative*, detti anche **indici di mutabilità**.

Con riferimento ad una variabile Y rilevata sulla popolazione $\mathcal{U}$, la variabilità si traduce nella diversificazione delle modalità osservate. Se Y è quantitativa, tale diversificazione si intende sia come *diversità* di valori osservati sia come *distanza* fra tali valori.

Esempio. Se Y è una variabile statistica degenere, $S_Y = \{y_1\}$ e la sua variabilità è nulla. Se $Y_1 = (1, 1, 1, 2, 2)$ e $Y_2 = (1, 1, 1, 10, 10)$, i due supporti corrispondenti contengono due modalità, ma la variabilità di Y_2 è più accentuata di quella di Y_1. $\diamond$

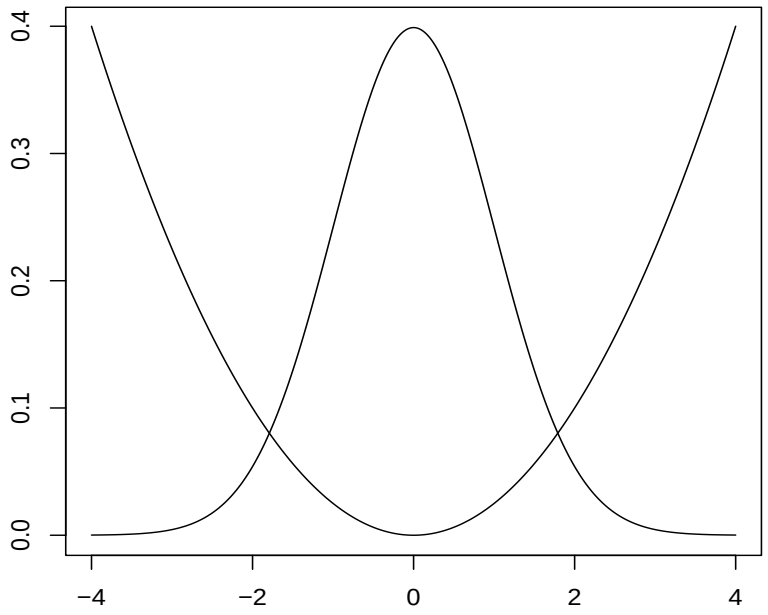
Due semplici indici di variabilità sono il campo di variazione e lo scarto interquartilico.

Definizione. Sia Y una variabile statistica quantitativa con supporto $S_Y = \{y_1, \ldots, y_J\}$, dove $y_1 < \cdots < y_J$. Il **campo di variazione** (*range*) corrisponde a

$$R_Y = y_J - y_1.$$

Se Y è degenere, $R_Y = 0$; altrimenti $R_Y > 0$.

R_Y è un indice piuttosto povero, come dimostra il seguente grafico dove le due curve esprimono lo stesso campo di variazione a fronte di una diversa variabilità.



Definizione. Sia Y una variabile statistica quantitativa, lo **scarto interquartilico** corrisponde a

$$SI_Y = y_{0.75} - y_{0.25}.$$

SI_Y esprime la lunghezza della scatola nel diagramma di pag. 42 ed è l'intevallo dove cade il 50% centrale della distribuzione di frequenza.

L'indice SI_Y può essere nullo anche per variabili non degeneri; ad esempio, si annulla per $Y = (1, 2, 2, 2, 2, 2, 5)$, poiché $y_{0.75} = y_{0.25} = 2$.

Il più importante indice di variabilità per variabili quantitative è la varianza.

Definizione. Sia Y una variabile statistica quantitativa con media aritmetica $E(Y)$. Si dice **varianza** di Y, e si indica con $V(Y)$, con σ_Y^2 o semplicemente con σ^2, la quantità

$$V(Y) = E\{(Y - E(Y))^2\}.$$

Si noti che $V(Y)$ è definita come la media aritmetica della variabile scarto $Y - E(Y)$ elevata al quadrato.

Quindi, per il calcolo di $V(Y)$, si può riprendere quanto detto con riferimento alla media aritmetica.

Se si dispone dei *dati grezzi* $Y = (y_1, \ldots, y_N)$ e si è preventivamente calcolata $E(Y)$, allora la varianza corriponde a

$$V(Y) = \frac{1}{N} \sum_{i=1}^{N} (y_i - E(Y))^2.$$

Se si dispone della *tabella di frequenza assoluta o relativa*, allora

$$V(Y) = \frac{1}{N} \sum_{j=1}^{J} (y_j - E(Y))^2 f_j = \sum_{j=1}^{J} (y_j - E(Y))^2 p_j.$$

Se si dispone della *tabella di frequenza assoluta o relativa con modalità raggruppate in classi* (ad esempio $y_{j-1} \dashv y_j$, $j = 1, \ldots, J$), è necessario calcolare il punto centrale $y_j^c = (y_{j-1} + y_j)/2$, $j = 1, \ldots, J$, delle singole classi.

In questo caso

$$V(Y) = \frac{1}{N} \sum_{j=1}^{J} (y_j^c - E(Y))^2 f_j = \sum_{j=1}^{J} (y_j^c - E(Y))^2 p_j.$$

Questa procedura per il calcolo di $V(Y)$ è equivalente a quella che si definisce quando si dispone dei dati grezzi se *le osservazioni che cadono in una classe coincidono con il punto centrale della classe*.

Esempio. Si consideri la tabella di frequenza di pag. 29, da cui si è ricavato che $E(Y) = 0.61$.

y_j	f_j
0	109
1	65
2	22
3	3
4	1
Totale	200

È immediato concludere che

$$V(Y) = \frac{1}{200}\big[(0-0.61)^2 \cdot 109 + (1-0.61)^2 \cdot 65$$
$$+(2-0.61)^2 \cdot 22 + (3-0.61)^2 \cdot 3$$
$$+(4-0.61)^2 \cdot 1\big] = 0.608.$$

$\diamond$

Esempio. Si consideri la tabella di frequenza di pag. 30, con modalità raggruppate in classi, da cui si è ricavato che $E(Y) = 11.15$.

Classe	$0 \dashv 10$	$10 \dashv 15$	$15 \dashv 20$	Totale
freq. rel.	0.30	0.52	0.18	1

Poiché i valori centrali delle classi sono, rispettivamente, $y_1^c = 5$, $y_2^c = 12.5$ e $y_3^c = 17.5$, si conclude che

$$V(Y) = (5 - 11.15)^2 \cdot 0.30 + (12.5 - 11.15)^2 \cdot 0.52$$
$$+(17.5 - 11.15)^2 \cdot 0.18 = 19.55.$$

$\diamond$

La varianza è espressa nel quadrato dell'unità di misura dei dati originari. Un indice dimensionalmente omogeneo con i dati è fornito dalla seguente definizione.

Definizione. Si dice **scarto quadratico medio** di Y, e si indica con σ_Y o con σ, la radice quadrata aritmetica (l'unica positiva) della varianza

$$\sigma_Y = \sqrt{V(Y)}.$$

La varianza soddisfa le seguenti **proprietà**.

1) Proprietà di non negatività: $V(Y) \geq 0$, con $V(Y) = 0$ se e solo se Y è degenere.

2) Formula per il calcolo:

$$V(Y) = E(Y^2) - (E(Y))^2.$$

Infatti, sfruttando la proprietà di linearità della media aritmetica,

$$V(Y) = E\{(Y - E(Y))^2\} = E\{Y^2 + (E(Y))^2 - 2YE(Y)\}$$
$$= E(Y^2) + (E(Y))^2 - 2E(Y)E(Y) = E(Y^2) - (E(Y))^2.$$

3) Proprietà di invarianza per traslazioni:

$$V(Y+b) = V(Y), \quad b \in \mathbf{R}.$$

Infatti, sfruttando la proprietà di linearità della media aritmetica,

$$V(Y+b) = E\{(Y+b-E(Y+b))^2\} = E\{(Y+b-E(Y)-b)^2\}$$
$$= E\{(Y-E(Y))^2\} = V(Y).$$

4) Proprietà di omogeneità di secondo grado:

$$V(aY) = a^2 V(Y), \quad a \in \mathbf{R}.$$

Infatti, sfruttando la proprietà di linearità della media aritmetica,

$$V(aY) = E\{(aY - E(aY))^2\} = E\{(aY - aE(Y))^2\}$$
$$= E\{a^2(Y-E(Y))^2\} = a^2 E\{(Y-E(Y))^2\} = a^2 V(Y).$$

Dalla **2)** discende che, se $E(Y) = 0$ allora $V(Y) = E(Y^2)$.

Dalla **3)** e dalla **4)** discende che $V(aY + b) = a^2 V(Y)$.

Una variabile Y tale che $E(Y) = 0$ è detta **centrata**.

Una variabile Y tale che $E(Y) = 0$ e $V(Y) = 1$ è detta **standardizzata**.

Come conseguenza delle proprietà della media aritmetica e della varianza si conclude che:

- data una variabile Y, allora $Z = (Y - E(Y))/\sqrt{V(Y)}$ è una variabile standardizzata;

- data una variabile Z standardizzata, allora la variabile $Y = \sigma Z + \mu$ è tale che $E(Y) = \mu$ e $V(Y) = \sigma^2$.

La proprietà **2)**, detta formula per il calcolo, fornisce effettivamente una procedura alternativa per il calcolo della varianza, come si può rilevare anche dal seguente esempio.

Esempio. Si consideri la tabella di frequenza di pag. 29, da cui si è ricavato che $E(Y) = 0.61$.

y_j	f_j
0	109
1	65
2	22
3	3
4	1
Totale	200

È immediato concludere che

$$E(Y^2) = \frac{1}{200}\left[0^2 \cdot 109 + 1^2 \cdot 65 + 2^2 \cdot 22 + 3^2 \cdot 3 + 4^2 \cdot 1\right]$$
$$= 0.98.$$

Da cui si ottiene che

$$V(Y) = E(Y^2) - (E(Y))^2 = 0.98 - 0.61^2 = 0.608,$$

che coincide con il valore ottenuto a pag. 48. $\diamond$

Con riferimento a variabili Y che assumono solo *valori positivi* si può introdurre un indice adimensionale di variabilità detto **coefficiente di variazione**

$$CV_Y = \frac{\sigma_Y}{\mu_Y}.$$

È un indice di variabilità relativa, nel senso che misura la variabilità dei dati tenendo conto dell'ordine di grandezza del fenomeno.

Inoltre, essendo un numero puro, permette il confronto tra popolazioni.

Esempio. Si consideri la seguente tabella di frequenza che riporta le merci e i passeggeri sbarcati, con riferimento agli scali portuali di alcune regioni italiane nel 1988.

Regione	Merci (migliaia di tonnellate)	Passeggeri (migliaia)
Friuli V.-G.	22806	42
Veneto	21849	248
Emilia-R.	12627	3
Marche	4937	266

Ci si chiede se è più variabile, tra le unità statistiche (le regioni), lo sbarco di merci (variabile Y_1) o lo sbarco di passeggeri (variabile Y_2).

Si ottiene che

$$E(Y_1) = 15554.75, \quad V(Y_1) = 53376613,$$
$$E(Y_2) = 139.75, \quad V(Y_2) = 13978.33,$$

ma

$$CV_{Y_1} = 0.47, \quad CV_{Y_2} = 0.85.$$

Quindi, nonostante la varianza di Y_1 sia più elevata, risulta maggiore, in termini relativi, la variabilità del numero di passeggeri. $\diamond$