Statistica

Capitolo 1
La rilevazione dei fenomeni statistici

Caratteri unità statistiche e collettivo
La statistica è la scienza che studia in termini qualitativi i **fenomeni collettivi**, ossia i fenomeni il cui studio richiede l'osservazione di un insieme di statistiche individuali.

Nome	Età	Sesso	Titolo di studio	Attività	Peso	Punteggio esercizi
Rossi	32	M	Laurea	Occupato	72	65
Bianchi	39	F	Laurea	Occupato	55	55
Nicoletti	46	M	Diploma	Disoccupato	79	53
Marcelli	28	M	Diploma	Studente	63	78
Petrone	51	F	Diploma	Casalinga	64	21

Come si evince dalla tabella, nome età sesso titolo di studio attività peso punteggio, sono le **caratteristiche** di un individuo. Questi caratteri assumono ad ogni individuo delle **modalità** (es. carattere=peso modalità=72). Le modalità possono essere numeriche che non numeriche. Nella tabella l'individuo è **l'unità elementare (unità statistica)** su cui sono osservati i caratteri. Un insieme di unità statistiche omogenee rispetto a uno o più caratteristiche costituisce un **collettivo statico** o una **popolazione**. I collettivi possono essere **di stato** (si individuano solo se si fissa un istante di tempo) o **di movimento** (se si parla di un intervallo di tempo). Le popolazioni possono anche essere **empiriche** (se osservabili) o **teoriche,** e possono racchiudere insiemi finiti o infiniti.

Classificazione dei caratteri statistici
Un carattere può assumere modalità differenti in corrispondenza delle diverse unità statistiche del collettivo. Le modalità devono essere **esaustive** (devono rappresentare tutti i possibili modi di manifestarsi di un carattere) e non **sovrapposte** (se ad ogni unità si può associare una sola modalità).
I caratteri si distinguono in

1. **Quantitativo (variabile)**
   - 1.1 **caratteri con scala a intervalli** non esiste uno zero assoluto o arbitrale (temperatura in gradi centigradi);
   - 1.2 **caratteri con scala a rapporti** esiste uno zero assoluto (peso);
   - 1.3 **caratteri discreti** l'insieme delle modalità può essere messo in corrispondenza biunivoca con un sottoinsieme di numeri interi (numero dei figli);
   - 1.4 **caratteri continui** l'insieme di modalità può essere messo in corrispondenza biunivoca con un insieme di numeri reali (altezza);
   - 1.5 **caratteri trasferibili** si dice trasferibile se considerata un'unità statistica possa cedere tutto o in parte il suo carattere a un'altra unità statistica;
   - 1.6 **caratteri non trasferibili;**
2. **Qualitativo (mutabile)**
   - 2.1 **carattere sconnesso** se date due sue modalità è solo possibile affermare se queste sono uguali o diverse (sesso);
   - 2.2 **carattere ordinato** se date due sue modalità è possibile solo dare un ordine specificando che una precede l'altra (titolo di studio);
     - 2.2.1 **caratteri ordinati rettilinei** che hanno una modalità di inizio e di fine;
       - 2.2.1.1 **Caratteri ordinati ciclici** che possono essere ordinati ma non hanno un inizio o una fine (es mese di nascita).

Suddivisione in classi di un carattere quantitativo

Se il carattere che si vuole analizzare presenta moltissime modalità distinte, si possono avere notevoli difficoltà nella comprensione dei dati osservati; in questi casi può essere conveniente fare un accorpamento delle modalità.

Se il carattere è quantitativo si procede alla **suddivisione in classi** che consiste nel suddividere l'insieme dei possibili valori in intervalli tra loro disgiunti. Cosi il carattere quantitativo passa da un livello di misura su scala di intervalli a un livello ordinale. Le classi sono caratterizzati da un'**ampiezza** che si intende la differenza tra estremo superiore ed estremo inferiore.

È opportuno definire le classi in modo tale che:

a. il numero sia abbastanza piccolo da fornire una sintesi adeguata ma sufficientemente grande da mantenere l'informazione con un livello accettabile;
b. siano tra loro disgiunte;
c. comprendano tutte le possibili modalità del carattere;
d. abbiano se possibile la stessa ampiezza.

Capitolo 2

Distribuzione di un carattere e sua rappresentazione

Dalle distribuzioni unitarie alle distribuzioni di frequenza

Dopo le fasi di acquisizione e di registrazione dei dati si passa alla loro elaborazione. La **distribuzione unitaria semplice** di un carattere è l'elencazione delle modalità osservate, unità per unità, nel collettivo preso in esame. Invece la **distribuzione unitaria multipla** si riferisce a più di un carattere. Per ottenere una maggiore sintesi delle modalità è utile scrivere una **frequenze assolute** (le modalità di un carattere vengono indicate con il numero di volte che si presentano). Tramite le frequenze possiamo ottenere una rappresentazione molto più sintetica una **distribuzione di frequenze**. Le distribuzioni di frequenza si dividono in:

1. **semplici** se è riferita ad un unico carattere
2. **doppia** sé riferita a due caratteri congiuntamente
3. **Multipla** se è riferita a più di due caratteri.

Frequenze relative, percentuali e cumulate

Dalle frequenze assolute si passa ad altre frequenze come quelle **relative** e **percentuali;** la prima è la frequenza assoluta diviso il numero di modalità osservate la seconda è la frequenza relativa per 100. Le frequenze relative e percentuali diventano significative se si vogliono confrontare due o più collettivi rispetto ad un carattere.

Data n unità statistiche dove n_j presentano la j-esima modalità, si definisce la frequenza relativa j-esima $f_j=\frac{n_j}{n}$, e frequenza percentuale $p_j=f_j \cdot 100$. La somma delle relative è pari a 1 invece delle percentuali è 100.

Nel caso in cui le modalità di un carattere in esame sono ordinate si può usare la **distribuzione di frequenza cumulata** che consiste nella somma di tutte le frequenze relative delle modalità precedenti. Dato un carattere X con K modalità ordinate in senso crescente, si indica con $N_j= n_1+n_2+ \dots +n_j$ la frequenza assoluta cumulata, con $F_j=f_1+f_2+ \dots +f_j$ la frequenza relativa cumulata e con $P_j=p_1+p_2+ \dots +p_j$ la frequenza percentuale cumulata.

Capitolo 3
Sintesi delle distruzione di un carattere – le medie

La media aritmetica
Nel caso in cui il carattere è quantitativo, la media più usata è la **media aritmetica** indicata con $\bar{x}$. La media aritmetica di un insieme di n valori osservati di un carattere quantitativo X è pari alla somma dei valori osservati divisa per il loro numero $\bar{x} = \frac{1}{n}(x_1+x_2+\ldots+x_m)$. Se il carattere è suddiviso in classi al posto della media aritmetica si può trovare il valore centrale, ossia il valore che si ottiene dalla semisomma degli estremi: $\bar{x} = \frac{1}{n}\sum_{j=1}^{k} c_j\, n_j$ dove c_j è il valore centrale della classe e n_j è la corrispondente frequenza assoluta. In alcuni casi si può dare importanza diversa alle diverse osservazioni attribuendogli un peso specifico in questo caso si usa la media ponderata $\bar{x}_p = \frac{x_1p_1+x_2p_2+\ldots+x_np_n}{p_1+p_2+\ldots+p_n}$

ci sono alcune proprietà della media aritmetica molto importanti:

1. la somma dei valori assunti da un insieme di unità statistiche è uguale al valore medio moltiplicato per le unità
2. la somma delle differenze tra i valori delle x_i e la loro media aritmetica $\bar{x}$ è pari a zero
3. la somma degli scarti al quadrato dei valori x_i da una costante c'è minima quando c'è uguale alla media aritmetica.

La media geometrica
La media geometrica è una media analitica utilizzata soprattutto nel caso in cui l'insieme dei dati è costruito da valori positivi generati da rapporti.
La media geometrica di un insieme di n valori positivi $x_1;x_2;x_n$ di un carattere quantitativo X è pari alla radice n-esima del prodotto dei singoli valori: $X_g = \sqrt[n]{(x_1 \cdot x_2 \cdot \ldots \cdot x_n)}$
Le proprietà della media geometrica sono

1. il prodotto dei valori assunti da un insieme di unità statistiche è uguale alla potenza n-esima della media geometrica
2. il logaritmo della media geometrica è uguale alla media aritmetica dei logaritmi

La mediana
La mediana è una media aritmetica un po' più "robusta" poiché non è condizionata dai valori più estremi.
La mediana (M_e) di un insieme di unità ordinate è la modalità presentata dall'unità centrale, dove per unità centrale si intende quell'unità che divide il collettivo in due parti di uguale numerosità.
Per calcolare la mediana è necessario procedere nel seguente modo:

1. ordinare le n unità in senso crescente rispetto alle modalità del carattere;
2. individuare la posizione in graduatoria dell'unità centrale: se n è dispari, la posizione è $\frac{n+1}{2}$ se n è pari si hanno due unità centrali con posizione $\frac{n}{2}$ e $(\frac{n}{2})+1$;
3. osservare la modalità presentata dall'unità centrale se né pari la mediana è $M_e=x_{(n+1)}$; se n è pari abbiamo due modalità $x_{\frac{n}{2}}$ e $x_{(\frac{n}{2}+1)}$. In ogni caso se il collettivo è quantitativo possiamo considerare come mediana la semisomma dei valori delle due unità centrali.

Se l'unità statistica è visualizzata dalle frequenze la mediana si trova cosi:

$M_e \cong I_m + \frac{(0{,}5-F_{m-1})}{(F_m - F_{m-1})}\Delta_m$

Dove

- I_m è l'estremo inferiore della classe mediana
- F_{m-1} è la frequenza relativa cumulata fino alla classe precedente a quella mediana

- F_m è la frequenza relativa cumulata fino alla classe mediana
- Δ_m è l'ampiezza della classe mediana

La moda

La moda è una media di posizione che può essere calcolata per qualsiasi tipo di carattere, in particolare anche per i caratteri qualitativi sconnessi. La moda è la modalità più frequente nel collettivo osservato.
Anche se la moda ci dice quale è il carattere più frequente nell'unità non ci dice però niente sugli altri caratteri. Per un'accuratezza migliore si può usare la **classe modale** che è definita come la classe alla quale corrisponde la frequenza più alta. Una distribuzione si dice **unimodale** se presenta un solo picco e **bimodale** se presenta due picchi di medesima altezza.

I percentili (quartili)

La mediana divide la distribuzione in due parti uguali, ognuna contenente il 50% delle unità. Si può dividere la distribuzione anche in cento parti, chiamando i valori suddivisi *percentili*. Si definiscono percentili quei valori che dividono la distribuzione in cento parti di uguale numerosità. In questa definizione si può considerare la mediana come cinquantesimo percentile.
I percentili di uso più frequente sono il 25esimo e il 75esimo, detti **primo quartile** (Q_1) e **terzo quartile** (Q_3).
Se la distribuzione di frequenza è suddivisa in classi, non è possibile trovare l'esatto valore del quartile, ma possiamo avvalerci di una sua approssimazione, con la seguente formula:

$$Q_1 \approx I_{Q_1} + \frac{(0{,}25 - F_{Q_{-1}})}{(F_Q - F_{Q_{-1}})} \Delta_{Q_1}$$

Capitolo 4

Sintesi della distribuzione di un carattere - la variabilità

La variabilità di una distribuzione

Come si sa la media è un indice che sintetizza la distribuzione del carattere, ma questo avviene soltanto se le unità presentano modalità vicini alla media. Bisogna quindi vedere il concetto **di variabilità** di un fenomeno.
La **variabilità** di una distribuzione esprime la tendenza delle unità del collettivo ad assumere diverse modalità del carattere. Per misurare la variabilità di una distribuzione è possibile utilizzare degli **indici di variabilità**.
Questi indici devono soddisfare almeno due requisiti:

1. deve assumere il suo valore minimo se e solo se tutte le unità della distribuzione presentano uguale modalità del carattere;
2. deve aumentare all'aumentare della diversità tra le due modalità assunte.

Indici basati sullo scostamento dalla media aritmetica

Tra gli indici di variabilità sono molto usai quelli che considerano le diversità dalla media aritmetica; la più nota è la **varianza.**

La **varianza** di un insieme di n valori osservati $x_1, x_2 \dots x_n$ di una variabile X con media aritmetica $\bar{x}$ è data da

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

DEVIANZA

quindi la varianza aritmetica è la media dei quadrati degli scarti della media aritmetica il numero della varianza è detto **devianza.**

La deviazione standard è la radice quadrata della varianza

$$\sigma = \sqrt{\sigma^2}$$

con questa operazione ci si riconduce a un indice di variabilità espresso nella stessa unità di misura della variabile considerata.

Il coefficiente di varianza CV della distribuzione di un carattere X di media $\bar{x}>0$ e deviazione standard σ è dato dal rapporto tra la deviazione e la media moltiplicato per 100

$$CV= \frac{\sigma}{\bar{x}} \cdot 100$$

Altre misure di variabilità sono gli **scostamenti semplici medi** che si ottengono come media aritmetica delle differenze in valore assoluto tra i valori osservati e una media

Si definisce **scostamento semplice medio dalla media aritmetica** la quantità:

$$S_{\bar{x}}= \frac{1}{n} \sum_{i=1}^{n} | x_i-\bar{x} |$$

Si definisce **scostamento semplice medio dalla mediana** la quantità:

$$S_{Me}= \frac{1}{n} \sum_{i=1}^{n}$$

Anche in questo caso ci sono i due rispettivi indici percentuali:

$$\frac{S_{\bar{x}}}{\bar{x}} \cdot 100 \qquad \frac{S_{Me}}{\bar{x}} \cdot 100$$

C'è una relazione che lega lo scostamento semplice medio dalla media aritmetica e la deviazione standard infatti si dimostra che: $S_{\bar{x}} \leq \sigma$

Il teorema di Chebyshev e la standardizzazione

Ci si può chiedere quali informazione una media e un indice di variabilità ci forniscano congiuntamente su una distribuzione incognita. Se come media e indice di variabilità usiamo la media aritmetica x_m e la deviazione standard σ, possiamo utilizzare il **teorema di Chebyshev.** Questo noto teorema afferma che, dato un carattere di cui si conoscono solamente la media aritmetica e la dev. Sta., la frequenza relativa delle unità che presentano valori esterni a un intervallo simmetrico rispetto alla media non può essere superiore a una certa quantità. Data una distribuzione di valori x_i dei quali si conoscono solo la media e la deviazione standard e dato un valore reale positivo k possiamo affermare che:

$$f(| x_i-\bar{x} |) \geq k\sigma \leq \frac{1}{k^2}$$

dove con f si intende la frequenza relativa dei valori del carattere X che soddisfano la diseguaglianza all'interno della parentesi.

Al teorema di Chebyshev è collegato un altro teorema quello di **Markov.**

Data una variabile X che assume solo valori non negativi x_i dei quali è nota la media $\bar{x}$ dato un qualsiasi valore a>0 possiamo affermare che:

$$f(X \geq a) \leq \frac{X}{a}$$

La concentrazione

Un carattere quantitativo trasferibile X con n valori osservati $x_1, x_2, \ldots, x_n$ si dice equidistribuito se ognuna delle n unità possiede uno /n dell'ammontare complessivo del carattere $A = \sum_{i=1}^{n} x_i$ ossia per ogni i sia che $x_i = \frac{A}{n} = \bar{x}$

Se non si verifica l'equidistribuzione, sussiste un certo grado di concentrazione del carattere, che non può essere misurato tramite opportuni indici.

La concentrazione del carattere evidenzia in modo più efficace e più interpretabile la variabilità dei caratteri trasferibili. In effetti, tanto più un carattere è concentrato, tanto più è elevata la variabilità del carattere.

La situazione di **massima concentrazione** si ha quando l'intero ammontare del carattere A è posseduto da una sola unità del collettivo, e cioè: $x_1 = x_2 = \ldots = X_{n-1} = 0$ e $x_n = A$

Si consideri un carattere quantitativo trasferibile X osservato su un quantitativo di n unità, ordinate in senso non decrescente secondo l'ammontare di carattere posseduto, ossia

$$X_1 \leq x_2 \leq \ldots \leq x_n$$

Indichiamo con:

- $A_i = x_1 + x_2 + \dots + x_i$ l'ammontare di carattere posseduto dalle i unità più povere.
- $Q_i = \frac{A_i}{A_n}$ la corrispondente frazione di ammontare
- $F_i = \frac{i}{n}$ la frequenta relativa cumulata delle prime i unità

Q_i e F_i hanno una stretta relazione evidenziata dalle seguenti proprietà:

- $F_i \geq Q_i$ per ogni i
- $F_i = Q_i$ per i = n, oppure per ogni i se $x_1 = x_2 = \dots = x_n$

Possiamo sintetizzare tali differenze attraverso il seguente indice:

$$C = \sum_{i=1}^{n-1} (F_i - Q_i)$$

Notiamo che la sommatoria arriva fino al termine (n-1)esimo poiché l'ennesima differenza è sempre uguale a zero, essendo $Q_n = F_n = 1$. Questo indice assume valore minimo quando tutte le differenze sono uguali a zero, cioè nel caso di equidistribuzione, e il suo valore di massimo nel caso di massima concentrazione, cioè quando $Q_i = 0$.

Per trasformare l'indice C in un indice di concentrazione negativo (variabile tra 0 e 1), basterà dividerlo per il suo massimo valore (**rapporto di concentrazione di Gini**).

Date le distribuzioni delle F_i e delle Qi relative alla distribuzione di un carattere quantitativo trasferibile X, osservato su n unità, con valori ordinati $X_1, X_2, \dots, X_n$ (con $X_i \leq X_{i+1}$), si definisce **rapporto di concentrazione di Gini** l'indice:

$$R = 1 - \frac{\sum_{i=1}^{n-1} Q_i}{\sum_{i=1}^{n-1} F_i}$$

Mediante le coppie di valore Q_i e F_i possiamo realizzare un grafico. Si consideri un piano cartesiano, in cui l'asse delle ascisse rappresenti i valori F_i, e l'asse delle ordinate i valori Q_i. Ogni valore è rappresentato da un punto sul piano; i punti limitrofi possono essere poi congiunti da una curva detta **spezzata di concentrazione** o **curva di Lorenz** (vedi grafico pag. 92).

Capitolo 5

Numeri indici, serie storiche e rapporti statistici

Misura del mutamento in una serie storica

L'osservazione sistematica nel tempo di un fenomeno permette di costruire una **serie storica.** Le osservazioni non hanno una temporalità specifica ma devono avere sempre la stessa temporalità (es. mese per mese anno per anno). Quello che in genere interessa studiare è l'entità delle variazioni avvenute tra due periodi di tempo contigui. La variazione di un periodo all'altro può essere misurata rapportando il valore della serie a un certo periodo t+1 con quello relativo al precedente ossia t, $\frac{y_{t+1}}{y_t}$. tale rapporto moltiplicato per 100 viene chiamato **tasso divariazione percentuale**. In generale, quando si è interessati a misurare l'entità dei mutamenti in una serie storica si possono effettuare dei rapporti tra due o più valori della serie. I valori cosi ottenuti vengono chiamati **numeri indici semplici**.

Numeri indici semplici

La serie di numeri indici può essere costruita in due diversi modi: a **base fissa** o a **base mobile**.

Una serie di numeri indici a base fissa esprime l'intensità di un fenomeno in ogni periodo di tempo come una quota dell'intensità di un periodo di riferimento chiamato base. Quindi un numero indice al tempo t di base s di ottiene dall'espressione: $I_{ts} = \frac{y_t}{y_s}$.

Un altro modo di analizzare l'andamento di un fenomeno è quello di confrontare l'intensità o la frequenza di un certo periodo con l'intensità o la frequenza del periodo precedente, cosi si usano numeri indici a base mobile.

Ponendo pari a 1 il numero indice per il periodo di tempo scelto come base e indicando con I_{ts} il generico numero indice relativo al tempo t e riferito al tempo base s si possono illustrare tre basilari proprietà:

a) se si confronta una situazione temporale con se stessa in numero indice vale 1
b) il numero indice I_{ts} è l'inverso del numero I_{st}
c) dati tre periodi t, s, r si ha $I_{ts} \cdot I_{sr} \cdot I_{rt} = 1$. Quest'ultima proprietà ci indica un principio di coerenza senza il quale non avrebbero senso operazioni quali il cambio di base.

Capitolo 6
Analisi dell'associazione tra due caratteri

Distribuzioni doppie di frequenze

Le determinazione di due caratteri su di un collettivo, possono essere organizzate sotto forma di **distribuzione unitaria doppia di frequenze**. Tale organizzazione è però inadatta a esplicitare le caratteristiche di un fenomeno ed è in gennere necessario sintetizzare le determinazioni dei caratteri tramite una **tabella di frequenze a doppia entrata** (distribuzione doppia di frequenze).

Dati due caratteri definiamo distribuzione doppia di frequenze l'insieme delle **frequenze congiunte** n_{ij} ovvero le frequenze assolute delle unità che presentano congiuntamente la modalità i-esima del primo carattere e la j-esima del secondo carattere. Es di tabella a doppia entrata con carattere X e Y e modalità H e K.

		Y					
		y_1	…	y_j	…	y_K	Totale
X	x_1	n_{11}	…	n_{1j}	…	n_{1K}	$n_{1\cdot}$
	…	…	…	…	…	…	…
	x_i	n_{i1}	…	n_{ij}	…	n_{ik}	$n_{i\cdot}$
	…	…	…	…	…	…	…
	x_H	n_{H1}	…	n_{Hj}	…	n_{HK}	$n_{H\cdot}$
	Totale	$n_{\cdot 1}$	…	$n_{\cdot j}$	…	$n_{\cdot K}$	n

L'ultima colonna e l'ultima riga sono le **distribuzioni marginali**. Esse corrispondono esattamente alle distribuzioni di frequenze semplici relative ai due caratteri esaminati. Le righe e le colonne interne identificano le cosiddette **distribuzioni condizionate**.

Nella tabella di frequenze a doppia entrata valgono le seguenti proprietà:

- $n_{i\cdot} = \sum_{j=1}^{K} n_{ij}$ per i = 1, …, H;
- $n_{\cdot j} = \sum_{i=1}^{H} n_{ij}$ per j = 1, …, K;
- $n = \sum_{i=1}^{H}\sum_{j=1}^{K} n_{ij} = \sum_{i=1}^{H} n_{i\cdot} = \sum_{j=1}^{K} n_{\cdot j}$

per ogni distribuzione condizionata di un carattere quantitativo si può calcolare **la media aritmetica condizionata**:

$$\bar{y}_{X=x_i} = \frac{1}{n_{i\cdot}} \sum_{j=1}^{K} y_j n_{ij}$$

oltre alla media da ogni distribuzione condizionata è possibile ricavare anche la **varianza condizionata:**

$$\bar{y}_{Y/X=x} = \frac{1}{n_{.j}} \sum_{i=1}^{K} (yj-\bar{y}_{X=x})2 \cdot n_{ij}$$

Se entrambi i caratteri sono di tipo ordinato è possibile definire le **frequenze cumulate** per la distribuzione doppia

Frequenza assoluta: $N_{ij} = \sum_{h=1}^{i} \sum_{k=1}^{K} n_{hk}$

Frequenza relativa: $F_{ij} = \sum_{h=1}^{i} \sum_{k=1}^{K} f_{hk} = \frac{1}{n} \sum_{h=1}^{i} \sum_{k=1}^{K} n_{hk}$

Dati due caratteri X e Y entrambi quantitativi, possiamo sintetizzare la distribuzione doppia mediante il punto di coordinate: $(\bar{x};\bar{y})$

Chiamato **punto medio o baricentro** della distribuzione

Analisi dell'associazione tra due caratteri: dipendenza, indipendenza, interdipendenza

La ricerca scientifica non si limita alla descrizione dei singoli fenomeni ognuno considerato indipendentemente dagli altri. Ad essa interessa soprattutto l'analisi della relazione che ognuno di questi può avere con gli altri. Tali ipotesi possono derivare da conoscenze scientifico-culturali oppure da ragionamenti logico-deduttivi.

Si parla di **dipendenza logica** tra due o più caratteri quando tra questi sono note a priori relazioni di causa-effetto.

Si parla invece di **indipendenza logica** fra due o più caratteri quando si suppone a priori che tra questi non possa sussistere alcuna relazione di causa-effetto. Per studiare la dipendenza tra due o più caratteri si usano due approcci: **o l'analisi della dipendenza** dove si studia come le modalità di un carattere "dipendono" da quelle di un altro carattere o **l'analisi dell'interdipendenza** in cui si assume che i caratteri abbiano tutti lo stesso ruolo e che la dipendenza fra loro siano bidirezionali.

Si parla invece di **indipendenza statistica** quando la conoscenza della modalità di uno dei due caratteri non migliora la "previsione" della modalità dell'altro.

Studio dell'associazione tra due caratteri in una tabella a doppia frequenza

La tabella doppia di frequenze è lo strumento più idoneo per indagare sulle relazioni esistenti tra le modalità di due caratteri qualitativi o quantitativi suddivisi in classi.

Molto importante è la dipendenza tra due variabili X e Y della tabella: si dirà che x è **indipendente** da Y se, qualunque sia la modalità con cui si manifesta il carattere Y, la distribuzione relativa condizionata di X non cambia. Se X è indipendente da y si può dimostrare che anche Y è indipendente da x.

Due caratteri x e y **si diranno indipendenti se** le distribuzioni relative condizionate rispetto alle modalità dell'altro sono uguali. Se due caratteri sono indipendenti, la **generica frequenza assoluta** corrisponde alla i-esima modalità di X e alla j-esima modalità di Y e dev'essere uguale a:

$$n_{ij} = \frac{n_{i.}\, n_{.j}}{n}$$

Le frequenze assolute di una tabella a doppia frequenza, ottenuta tramite la formula, saranno chiamate **frequenze teoriche di indipendenza** e indicate con n'_{ij}.

Se due caratteri non sono statisticamente indipendenti, ci si attende che fra essi sussista un qualche tipo di relazione. Un tipo di associazione è detta **spuria**. L'associazione spuria è un legame statistico empirico che si verifica tra due caratteri logicamente indipendenti.

Se il legame associativo tra i due caratteri non è spurio, se ne può affrontare lo studio secondo due ottiche: la **dipendenza** e l'**interdipendenza**.

Un carattere Y è *dipendente perfettamente* da X quando a ogni modalità di X è associata una sola modalità di Y, cioè quando, in una tabella doppia, per ogni i c'è solo un i per il quale $n_{ij} \neq 0$.

Tra due caratteri sussiste *interdipendenza perfetta* se a ogni modalità di uno dei due caratteri corrisponde una e una sola modalità dell'altro e viceversa (tabella quadrata). Vedi esempi pag. 138-139 (per indipendenza e interdipendenza).

Ogni situazione intermedia tra indipendenza e associazione perfetta esprime un certo grado di dipendenza.

Misura dell'associazione caratteri qualitativi sconnessi

La misura dell'associazione tra due caratteri qualitativi sconnessi avviene analizzando la distribuzione congiunta delle frequenze dei due caratteri. Per valutare l'interdipendenza si usano indici basati su un approccio simmetrico. Per la dipendenza invece si utilizza un approccio asimmetrico. Gli indici generali di associazione si basano sulle differenze tra le frequenze osservate n_{ij} e quelle teoriche di indipendenza n'_{ij}, che corrispondono alle frequenze che avremmo dovuto avere se, date le distribuzioni semplici, i caratteri fossero stati indipendenti.

Le differenze tra le frequenze osservate e quelle teoriche vengono dette **contingenze** (c_{ij}), con cui si costruisce la tabella delle contingenze (la somma delle contingenze dev'essere uguale a zero.

L'indice di associazione **Chi-quadrato** del Pearson è definito in modo seguente:

$$\chi^2 = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{c^2_{ij}}{n'_{ij}} \qquad \left(\chi^2 = \sum_i \sum_j \frac{n^2_{ij}}{n'_{ij}} - n\right)$$

Se i due caratteri sono perfettamente indipendenti, le contingenze devono essere tutte nulle, e dunque l'indice Chi-quadrato assumerà un valore nullo. Se, al contrario, i due caratteri sono associati, l'indice assumerà valore positivo, assumendo valori tanto più grandi, quanto più le contingenze sono più grandi (cioè i valori osservati sono distanti da quelli teorici).

In genere si preferisce utilizzare gli indici normalizzati che non dipendono dalle frequenze delle distribuzioni marginali; per fare questo si usa l'**indice di contingenza quadratica media**, cioè:

$$\phi^2 = \frac{\chi^2}{n}$$

Proprietà di ϕ^2:

Per una tabella frequenza doppia, con H righe e K colonne, il valore massimo di ϕ^2 è:

- Se H < K H-1 : il carattere X, a cui corrispondono le righe, dipende perfettamente da Y
- Se H > K K-1 : Y dipende perfettamente da X
- Se H = K H-1 = K-1 vi è perfetta interdipendenza tra i due caratteri.

Dalla proprietà appena esposta, si deduce che il valore massimo che può assumere ϕ^2, su una tabella formata da H righe e K colonne, è data da:

$$\max \phi^2 = \min [(H-1)\ (K-1)]$$

INDICE V DI CRAMER:

Cramer ha proposto di normalizzare l'indice ϕ^2 rapportandolo al suo valore massimo:

$$V = \sqrt{\frac{\phi^2}{\min [(H-1)\ (K-1)]}}$$

INDICE LAMBDA:

Un indice asimmetrico utile per analizzare la dipendenza è l'indice **lambda**. Questo indice si basa sull'analisi del miglioramento della previsione del carattere Y, data la modalità del carattere X. $E_2 = \sum_{i=1}^{h} (n_{i.} - n_{im_i})$

n_{im_i} è la frequenza della moda della riga i-esima. L'indice lambda è dato dal rapporto $\lambda = \frac{E_1 - E_2}{E_1} = \frac{\sum_{i=1}^{h} n_{im_i} - n_{.m}}{n - n_{.m}}$

Proprietà dell'indice lambda:

- $0 \le \lambda \le 1$
- $\lambda = 0$ se e solo se la conoscenza di una modalità di riga non è di nessun aiuto nel prevedere la modalità del carattere di colonna.

- $\lambda = 1$ se e solo se la conoscenza di una modalità di riga specifica completamente la modalità di colonna, in altre parole se ogni riga della tabella contiene al massimo una sola frequenza $\neq 0$.
- Se X e Y sono indipendenti, allora $\lambda = 0$, ma non è vero il contrario.

Capitolo 8
Probabilità: concetti base

Concetti primitivi
I concetti primitivi rappresentano le nozioni originarie e intuitive su cui viene costruita successivamente tutta la teoria. Nella geometria, per esempio, il punto e la retta sono concetti primitivi. Se si considera l'esempio del lancio del dado in questa situazione compaiono tre entità di concetti fondamentali:

- **la prova** o esperimento aleatorio è un esperimento che ha due o più possibili risultati e in cui c'è un certo grado di incertezza su quale di questi risultati si presenterà.
- **l'evento** si distingue in due tipi:
- **evento elementare** si intende uno dei risultati possibili delle prove
- **evento non-elementare**
- **la probabilità**

Il legame logico-formale tra queste entità è schematizzato nella seguente proposizione.
In una data **prova**, **l'evento** E si verifica con la **probabilità** $P(E)$.
Es nel lancio di un dato la faccia contrassegnata dal numero 5 ($E = 5$) si presenta con probabilità $P(E = 5) = 1/6$.
La probabilità è un numero compreso tra 0 e 1 che misura il grado di incertezza sul verificarsi di un evento.

Eventi e algebra degli eventi
Con il verificarsi di una prova non si possono considerare tutti gli eventi elementi ω_i l'insieme di tutti gli eventi è indicato con E. È conveniente introdurre una collezione di eventi $E = \{E_1, E_2, \ldots, E_n\}$ tutti sottoinsiemi di Ω la cui struttura matematica è quella di un'**algebra di Boole**.
POSTULATO 1 gli eventi formano un'**algebra di Boole**.
In questa struttura matematica sono definite le operazioni fondamentali:

1. La **negazione** di un evento A, ossia $\bar{A}$ (l'evento A non si verifica)
2. L'**intersezione** tra due eventi A e B, ossia $A \cap B$ (si verifica sia l'evento A che l'evento B)
3. L'**unione** tra due eventi A e B, ossia $A \cup B$ (si verifica o l'evento A o l'evento B)

È utile definire due eventi importanti:

- L'evento **impossibile** è definito come $A \cap \bar{A} = \emptyset$
- L'evento **certo** che si verifica sempre perché contiene tutti i possibili risultati dell'esperimento $\emptyset = \Omega$.

Due eventi A e B si dicono **incompatibili** se $A \cap B = \emptyset$

I postulati
Ad una generica prova è associato uno spazio campionario Ω e ad esso una collezione di eventi la cui struttura matematica è quella dell'algebra di Boole.

La **probabilità** è una funzione di insieme che associa a ogni evento $E_i \in E$ un numero reale, la probablità sarà indicata con $P(E_i)$.

POSTULATO 2: $P(A) \geq 0$

POSTULATO 3: $P(\Omega) = 1$

POSTULATO 4: $[A \cap B = \emptyset] \Rightarrow [P(A \cup B) = P(A) + P(B)]$

Misura della probabilità approccio classico

A partire dai postulati definiti è possibile definire in maniera intuitiva una misura della probabilità che si adatta in cui gli eventi elementari sono perfettamente noti, in numero finito e equipossibili

La **probabilità** è data dal rapporto tra il numero dei casi favorevoli all'evento e il numero di casi possibili purché essi siano tutti ugualmente possibili $P(E) = \frac{\text{n. di casi favorevoli}}{\text{n. di casi possibili}}$

Probabilità condizionate e indipendenza

In alcune situazioni si vuole valutare la probabilità di un evento sapendo che si è già verificato un altro evento a esso collegato (vedi esempio pag. 204). Per un caso di probabilità condizionata la formula è:

$$P(A \mid B) = \frac{\text{n. di casi favorevoli ad } (A \cap B)}{\text{n. di casi possibili a B}}$$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad \text{con } P(B) > 0$$

$$P(A \cap B) = P(B) \cdot P(A \mid B)$$

Se $P(A \mid B) = P(A)$ e $P(B \mid A) = P(B)$ allora si ricava che $P(B \mid A) = \frac{P(B \cap A)}{P(A)} = P(B) \Rightarrow \qquad P(B \cap A) = P(B) \cdot P(A)$

Capitolo 9

Variabili casuali e distribuzione di probabilità

Variabili casuali (aleatorie)

La variabile casuale (**aleatoria, stocastica**) X è una funzione definita sullo spazio campionario Ω che associa a ogni risultato elementare ω_i in un unico numero reale. Si devono distinguere due tipi di variabili casuali:

- **discreta** può assumere un insieme discreto (numerabili) di numeri reali
- **continua** può assumere tutti i valori compresi in un intervallo reale.

Variabili casuali discrete

Con le variabili casuali discrete si trova facilmente la distribuzione di probabilità (vedi es. pag. 219). Cosi in generale si può indicare con $P(X = x_i)$ la probabilità che la v. c. X assuma il valore x_i. Cosi si definisce la **funzione di probabilità** di una variabile casuale discreta X associa a ognuno dei possibili valori x_i la corrispondente probabilità $P(X = x_i)$. Dalla definizione discendono due evidenti proprietà:

$$\sum_{i=1}^{n} P(x_i) = 1 \qquad\qquad P(x_i) \geq 0$$

Data una v. c. discreta X, la funzione che fa corrispondere ai valori x le probabilità cumulate $P(X \leq x)$viene detta funzione di ripartizione ed è indicata con: $F(x) = P(X \leq x) = \sum_{w \leq x} P(X = w)$ (vedi es. pag. 220)

Variabili casuali continue

Si suppone per esempio che la v.c. X possa assumere tutti i valori dell'intervallo reale [0;1] e che assuma ciascun valore con la stessa probabilità (diversa da 0): comunque si fissi tale probabilità si ha che la somma delle probabilità è infinita. Piuttosto che assegnare una misura di probabilità ai singoli valori, possiamo assegnare una misura di probabilità a tutti i possibili intervalli sull'asse reale. A tale scopo so introduce la **funzione di densità** (è la funzione matematica f(x) per cui l'area sottesa alla funzione, corrispondente a un certo intervallo, è uguale alla probabilità che X assuma un valore in quell'intervallo $P(A \leq X \leq b) = \int_a^b f(x)\,dx$

Proprietà delle funzioni di densità

1. una funzione di densità non può mai assumere valori negativi, ossia $f(x) \geq 0$; ciò assicura che la probabilità che X cada in un qualsiasi intervallo sia non negativa
2. l'area totale sottesa alla funzione è uguale a 1 ossia $\int_{-\infty}^{+\infty} f(x)\,dx = 1$
3. la probabilità che la v.c. X assuma un particolare valore dell'intervallo è 0. Ciò è dovuto al fatto che un singolo valore corrisponde a un in intervallo di ampiezza 0, quindi la corrispondente area è anch'essa 0. Questo per esempio implica che non ha influenza l'inclusione, nel calcolo della probabilità degli estremi dell'intervallo, ossia: $P(a \leq X \leq b) = P(a < X < b)$

Data una v.c. continua X, la funzione che fa corrispondere ai valori x le probabilità cumulate $P(X \leq x)$ viene detta **funzione di ripartizione** e indicata con:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(w)\,dw$$

Valore atteso e varianza di una variabile casuale

Spesso si è interessati a conoscere il **valore medio** che un variabile casuale può assumere in un gran numero di prove. Tale valore viene chiamato **valore atteso o speranza matematica:**

$E(X) = \sum_i x_i P(x_i)$ se la v.c. è discreta

$E(X) = \int_{-\infty}^{+\infty} x\, f(x)\,dx$ se la v.c. è continua

Però la media fornisce solo la dimensione del fenomeno descritto dalla v.c. ma non fornisce nessuna informazione sulla sua variabilità. Per la variabilità si usa la **varianza**, per spiegare la varianza di una v.c. X dobbiamo innanzitutto spiegare come calcolare il valore atteso della funzione della v.c. X, Y = g(X), che per definizione è dato da $E(Y) = \sum_i y_i P(y_i)$ nel caso discreto e in quello continuo è data da $E(Y) = \int_{-\infty}^{+\infty} y\, f(y)\,dy$. Queste formule presumono la previa conoscenza di $P(y_i)$ e di f(y), tuttavia questo diventa superfluo perché per il calcolo del valore atteso si può utilizzare direttamente la funzione di probabilità o di densità infatti si dimostra che se la v.c. X è discreta $E(Y) = \sum_i y_i P(y_i) = \sum_i g(x_i) P(x_i)$ mentre se è continua $E(Y) = \int_{-\infty}^{+\infty} y\, f(y)\,dy = \int_{-\infty}^{+\infty} g(x)\, f(x)\,dx$ (vedi es pag 226)

La **varianza** V(X) di una v.c. X è definita da

$V(X) = \sum_i (x_i - E(X))^2\, p(x_i)$

$V(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 f(x)dx$

La radice quadrata della varianza di una v.c. X viene chiamata **deviazione standard** di X ed è indicata con:

$SD(X) = \sqrt{V(X)}$

Variabili casuali standardizzate e teorema di Chebyshev

I valori standardizzati esprimono la distanza tra le osservazioni e la media in termini di deviazione standard. Se X è una variabile casuale con valore atteso E(X) e deviazione standard SD(X), allora:

$$Y = \frac{X - E(X)}{SD(X)} \qquad \text{con } E(Y) = 0 \text{ e } V(Y) = 1$$

È una **variabile casuale standardizzata**.

Il teorema (o diseguaglianza) di Cebyshev (vedi cap 4)

Siano X una v.c. e k un valore reale positivo, allora vale la seguente diseguaglianza:

$$P(|X - E(X)| \geq k \cdot SD(X)) \leq \frac{1}{k^2}$$

Questo teorema previene la probabilità che che X assuma valori distanti dalla media più di k deviazioni standard e al più $1/k^2$.

Distribuzione di Bernoulli e binomiale

Si consideri una prova nella quale ha interesse solo verificare se un certo evento si è meno verificato. Tale variabile casuale viene detta **v.c. di Bernoulli**. Una v.c. di Bernoulli, indicata con X~, può assumere il valore 1 con probabilità π e il valore 0 con probabilità $1 - \pi$; la sua funzione di probabilità può essere espressa come

$$P(X = x) = \pi^x(1 - \pi)^{1-x} \qquad \text{per } x = 0,1$$

La media e la varianza di tale distribuzione sono date da:

$$E(X) = \pi \qquad V(X) = \pi\,(1 - \pi)$$

Distribuzione uniforme continua e normale

Una v.c. uniforme continua X, indicata con $X \approx U(a;b)$ è una v.c. che assume valori reali in un intervallo limitato [a;b] con a e b numeri reali. La **funzione di densità uniforme** è definita come:

$$f(x) \begin{cases} \frac{1}{b-a} & \text{se } a \leq x \leq b \\ 0 & \text{altrove} \end{cases}$$

La media e la varianza di tale distribuzione sono date da

$$E(X) = \frac{(a+b)}{2} \qquad V(X) = \frac{(a-b)^2}{12}$$

La **v.c. normale X**, indicata con $X \approx N(\mu;\sigma^2)$ è una v.c. continua che può assumere valori su tutto l'asse reale, con funzione di densità:

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \text{ con i parametri } -\infty \leq \mu \leq +\infty \text{ e } \sigma^2 > 0$$

La media e la varianza della v.c. normale sono date da:

$$E(X) = \mu \qquad V(X) = \sigma^2$$

Variabile casuale doppia

Fino adesso si è parlato di distribuzione della probabilità di una sola variabile casuale, derivata dall'associazione di un unico numero reale a ogni evento dello spazio campionario. Tale operazione si può estendere al caso in cui a ogni evento si possa associare un n-pla di numeri reali. Possiamo definire **una variabile casuale doppia** nel modo seguente:

si dice variabile casuale doppia una funzione (X,Y) definita sullo spazio campionario Ω, che associa a ogni risultato elementare detto ω_i, una coppia di numeri reali x,y.

Anche in questo caso si possono distinguere le **variabili casuali discrete** e le **variabili casuali continue**.

Quando una v.c. doppia può assumere solo un insieme finito e numerato di valori, parleremo di una v.c. doppia discreta a cui è associata una funzione di probabilità congiunta P(x,y).

$$P(x,y) \geq 0 \qquad \text{e} \qquad \sum_x \sum_y P(x,y) = 1$$

Quando invece la v.c. può assumere un insieme non numerabile di valori, parleremo di una v.c. doppia continua, a cui è associata una funzione di densità congiunta f(x,y).

$$F(x,y) \geq 0 \qquad \text{e} \qquad \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\,dx\,dy = 1$$

La **funzione di ripartizione congiunta** per la v.c. (X,Y) si definisce in modo simile a quanto visto per il caso univariato ed è data da:

v.c. doppia discreta $\qquad F(x,y) = P(X \leq x, Y \leq y) = \sum_{u \leq x}\sum_{v \leq y} P(u,v)$

v.c. doppia continua $\qquad F(x,y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)du\,dv$

Sommando o integrando la distribuzione di probabilità congiunta rispetto a tutti i valori della Y, si ottiene la distribuzione di probabilità della X, detta **distribuzione di probabilità marginale**.

v.c. doppia discreta $\qquad P(x) = \sum_{x} P(x,y) \qquad$ e uguale per P(y)

v.c. doppia continua $\qquad f(x) = \int_{-\infty}^{+\infty} f(x,y)\,dy \qquad$ e uguale per f(y)

La relazione di dipendenza di una variabile rispetto all'altra può essere studiata attraverso la **distribuzione di probabilità condizionata**:
data la variabile casuale doppia (X,Y), la **distribuzione di probabilità condizionata** di Y, data la X = x, è

caso discreto $\qquad P(y \mid x) = \frac{P(x;y)}{P(x)}$

caso continuo $\qquad f(y \mid x) = \frac{f(x;y)}{f(x)}$

vedi gli esempi a pag. 250-254

Considerata una v.c. doppia (X,Y), tra le due variabili casuali, c'è **indipendenza** se e solo se la distribuzione di probabilità congiunta può essere espressa dal prodotto delle distribuzioni marginali.

caso discreto $\qquad P(x,y) = P(x) \cdot P(y)$

caso continuo $\qquad f(x,y) = f(x) \cdot f(y)$

Il **valore atteso** di una combinazione lineare di 2 variabili casuali (variabile casuale doppia: $X = a_1X_1 + a_2X_2$) è dato da:

$$E(X + Y) = E(X) + E(Y)$$
$$E(X - Y) = E(X) - E(Y)$$

Consideriamo ora la funzione prodotto degli scarti dai valori attesi, g(X,Y) = [X – E(X)] [Y- E(Y)]. Il suo valore atteso è noto con il nome di **covarianza**, che si calcola nel seguente modo:

v.c. doppia discreta $\qquad \sigma_{xy} = \sum_{x}\sum_{y}(x \cdot y)P(x,y) - \sum_{x}xP(x)\ \sum_{y}yP(y)$

v.c. doppia continua $\qquad \sigma_{xy} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xy\,f(x,y)\,dx\,dy - \int_{-\infty}^{+\infty} x\,f(x)dx \int_{-\infty}^{+\infty} y\,f(y)dy$

Teorema del limite centrale

Una successione di variabili casuali X_1, X_2, X_3 …, con funzione di ripartizione $F_1(x)$, $F_2(x)$, $F_3(x)$ …, **converge in distribuzione** a una variabile casuale x se, per tutti i punti in cui F(x) è continua si ha

$$\lim_{n\to+\infty} F_n(x) = F(x)$$

La convergenza di distribuzione è alla base del **Teorema del limite centrale**, che dice che:

siano X_1, X_2, X_3, variabili casuali indipendenti e identicamente distribuite (iid), con media μ e varianza σ^2 finite

posto che $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

si ha che la v.c. è $Z_n = \frac{(\overline{X_n} - \mu)\sqrt{n}}{\sigma}$

Capitolo 10

Campionamento e distribuzioni campionarie

Popolazione e parametri della popolazione

Si deve distinguere tra **popolazione finita** e **popolazione infinita**. La popolazione finita è un insieme costituito da N unità; dato un carattere X osservato su tutta la popolazione si possono calcolare i **parametri della popolazione**, ossia le costanti che descrivono aspetti caratteristici della distribuzione del carattere della popolazione (vedi es pag. 272). La popolazione infinita è composta da un numero elevato di unità, ossia da tutte le unità potenzialmente osservabili e non necessariamente già esistenti fisicamente. Nelle popolazioni infinite il carattere d'interesse può essere rappresentato da una variabile casuale X con una certa distribuzione di probabilità. In questo caso è quindi consuetudine indicare con **popolazione X** la variabile casuale X

Media della popolazione (o valore atteso) di una popolazione infinita:

- discreta $\mu = E(X) = \sum_i$
- continua $\mu = E(X) = \int_{-\infty}^{+\infty} x\, f(x)dx$

Varianza della popolazione:

- discreta $\sigma^2 = Var(X) = \sum_{i=1}^{k} (x_i - \mu)^2\, p(x_i) = \sum_{i=1}^{k} x_i^2\, p(x_i) - \mu^2$
- continua $\sigma^2 = Var(X) = \int_{-\infty}^{+\infty} (x - \mu)^2\, f(x)\, dx = \int_{-\infty}^{+\infty} x^2\, f(x)dx - \mu^2$

Il campionamento da popolazioni infinite

Quando la popolazione è infinita non è possibile svolgere indagini totali e bisogna necessariamente ricorrere al campionamento. La caratteristica di interesse nelle popolazioni infinite può essere rappresentata da una variabile casuale X che possiede una certa distribuzione di probabilità.

Dalla popolazione X viene quindi estratto un sottoinsieme di unità statistiche, e tale procedura genera una n-pla di v.c., la cui determinazione numerica corrisponde a un n-pla osservazioni x, che costituisce il **campione osservato**.

Le variabili casuali x_i sono indipendenti, e quindi la n-pla di v.c. è una collezione di v.c. indipendenti e identicamente distribuite (iid). In tal caso si parla di **campione casuale**.

Una collezione di n variabili casuali X_1, X_2, …, X_n, ottenuta con un procedimento di estrazione dalla popolazione X, forma un **campione casuale** di dimensioni n della popolazione X se:

- X_n variabili sono casuali indipendenti
- Ogni v.c. X_i possiede la stessa distribuzione di probabilità della popolazione X.

Statistiche, campionari e distribuzioni campionarie

Sia $X_1, X_2, \ldots, X_n$ un campione casuale di n osservazioni appartenenti a una popolazione finita o infinita, si indica come **statistica campionaria** la funzione dei valori reali delle osservazioni campionarie $X_1, X_2, \ldots X_n$.

Le statistiche campionarie di uso comune sono le seguenti:

- media campionaria: $\bar{X} = \frac{1}{n} \sum_{i=1}^{K} X_i$
- varianza campionaria: $\sigma^2 = \frac{1}{n} \sum_{i=1}^{K} (X_i - \bar{X})^2$
- deviazione standard campionaria: $\sigma = \sqrt{\sigma^2}$
- massimo campionario: $X_n = \max (X_1, X_2, \ldots, X_n)$
- minimo campionario: $X_n = \min (X_1, X_2, \ldots, X_n)$
- intervallo di variazione campionario: $R = X_{(n)} - X_{(1)}$

Le statistiche non devono essere confuse con i parametri della popolazione, poiché questi ultimi si riferiscono all'intera popolazione, mentre le statistiche dipendono solamente dalle osservazioni campionarie.

La distribuzione della media campionaria nelle popolazioni infinite

Tra le statistiche più frequentemente utilizzate, la media campionaria ricopre un ruolo particolare dovuto alle sue proprietà campionarie. Il valore atteso e la varianza della media campionaria possono essere facilmente calcolati per le popolazioni infinite. Sia X la variabile casuale d'interesse e siano μ e σ^2 rispettivamente la media e la varianza della popolazione; in questo caso, $\mu = E(X)$ e $\sigma^2 = Var(X)$. Sia inoltre X_n un campione casuale di dimensioni n estratto dalla popolazione X; esso è formato da n variabili casuali indipendenti e identicamente distribuite (iid) con

$E(X_i)\ \mu$ e $Var(X,Y) = \sigma^2$ per ogni i appartenente a R

Sotto tali condizioni si ha:

- Il valore della media campionaria è uguale alla media della popolazione, ossia $E(\bar{X}) = \mu$
- La varianza della media campionaria è uguale alla varianza della popolazione divisa per la dimensione campionaria, ossia $Var(\bar{X}) = \frac{\sigma^2}{n}$

Se la popolazione ha una **distribuzione normale** $X \approx N(\mu; \sigma^2)$, allora la distribuzione campionaria sarà ancora una normale, cioè $\bar{X} \approx N(\mu; \frac{\sigma^2}{n})$

Se la popolazione X possiede una **distribuzione di Bernoulli** con parametro π, allora la distribuzione della media campionaria $\bar{X}$ sarà data da $P(\bar{X}=x) = \binom{n}{nx} \pi^{nx} (1-\pi)^{n-nx}$ con media π e varianza $\frac{\pi(1-\pi)}{n}$

Capitolo 11
Stima puntuale

Stima puntuale e stimatori

Quando non è possibile osservare tutte le unità della popolazione, alcune caratteristiche della popolazione (come media, varianza ecc.) restano incognite. In tal caso, si potranno ottenere informazioni circa il valore del parametro analizzando i dati provenienti da un campione. Bisogna quindi trovare una funzione dei dati campionari che fornisca una buona approssimazione del parametro ignoto; tale problema è noto come **stima puntuale** e il parametro che viene trovato è chiamato **stima puntuale del parametro**.

Sia X una variabile casuale che rappresenta il carattere osservato sulla popolazione d'interesse. Se la variabile casuale X è discreta, la sua funzione di probabilità sarà indicata da $p(X,\theta)$, se invece X è continua, la sua variabile di densità verrà indicata da $f(X,\theta)$, dove θ è il parametro che si vuole stimare.

Uno **stimatore** è una variabile casuale utilizzata per stimare una determinata caratteristica θ della popolazione e sarà indicato con $T = T(X_1 \ldots X_n)$

Il valore assunto da uno stimatore verrà detto **stima** e indicato con $t = t(x_1 \ldots x_n)$

Lo stimatore è una variabile casuale e quindi ha una sua distribuzione campionaria, la cui conoscenza permetterà di capire se lo stimatore scelto produrrà con elevata probabilità stime vicine al valore vero del parametro.

Stimatori corretti

La proprietà più intuitiva di uno stimatore è la **correttezza**. T è uno stimatore corretto di θ se il suo valore atteso è uguale al valore vero del parametro, quindi: $E(T) = \theta$ per ogni θ

Se $E(T) \neq \theta$ per qualche valore di θ, allora **T sarà distorto**.

La **distorsione di uno stimatore** è uguale a $B(T) = E(T) - \theta$

Stimatori efficienti e minimo errore quadratico medio

Per dire che T si avvicini ai valori di θ possiamo affermare che $|T - \theta|$ abbia valori piccoli e, di conseguenza, anche $(T - \theta)^2$ dovrà avere un piccolo valore. Questa quantità è chiamata **errore quadratico medio**, che ha come formula:

$$MSE(T) = E[(T - \theta)^2]$$

A questo punto di introduce l'**efficienza** di uno stimatore: T_1 è più efficiente di T_2 se possiede un MSE più piccolo.

L'errore quadratico medio di uno stimatore T è uguale alla somma della varianza dello stimatore e della sua distorsione al quadrato, quindi

$$MSE(T) = Var(T) + B(T) \qquad \text{dove } Var(T) = E\{[T - E(T)]^2\}$$

Se lo stimatore T è corretto, allora

$$MSE(T) = Var(T) \qquad \text{per tutti i possibili valori di } \theta$$

Stimatori consistenti e asintoticamente corretti

È importante valutare anche le proprietà asintotiche degli stimatori. Una delle proprietà asintotiche è la **consistenza**. Uno stimatore T_n di un parametro θ è **consistente** se il suo MSE tende a zero al tendere a $+\infty$ di n. Lo stimatore T_n è **consistente in media quadratica** se

$$\lim_{n \to +\infty} MSE(T_n) = \lim_{n \to +\infty} E(T_n - \theta)^2 = \lim_{n \to +\infty} Var(T_n) = 0$$

Uno stimatore T_n di parametro θ è **asintoticamente corretto** se

$$\lim_{n \to +\infty} E(T_n) = \theta \qquad \text{per ogni possibile valore di } \theta$$

Stima puntuale della media di una popolazione

Tra i parametri di una popolazione, un ruolo centrale è svolto dalla media. Da questo si evince che:

- La **media campionaria** è uno **stimatore corretto** della media della popolazione; questo vale per qualunque tipo di distribuzione della popolazione. Quindi si ha che l'errore quadratico medio coincide con la varianza, cioè

  $$MSE(\bar{X}) = Var(\bar{X}) = \frac{\sigma^2}{n}$$

- La **media campionaria** è uno **stimatore consistente** della media della popolazione, cioè

  $$\lim_{n\to+\infty} MSE(\bar{X}) = \lim_{n\to+\infty} Var(\bar{X}) = \lim_{n\to+\infty} \frac{\sigma^2}{n} = 0$$

Stima puntuale della proporzione in una popolazione

Una stima della proporzione di unità di una popolazione che presenta un certo attributo A può essere ottenuta in modo simile a quanto mostrato per la media della popolazione.

Data una popolazione X, distribuita con distribuzione Bernoulli, con parametro π, la media campionaria $\bar{X}$ è uno **stimatore corretto** di π, ossia $E(X) = \pi$ per ogni π tra 0 e 1

Stima puntuale della varianza di una popolazione

Un altro parametro di particolare interesse è la varianza di popolazione, da ciò si evince che $\sigma^2 = E[(X - \mu)^2]$

Dato un campione casuale di dimensioni n estratto da una popolazione X si definisce **varianza campionaria corretta** lo stimatore:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Dato un campione casuale estratto da una popolazione con media μ e varianza σ^2 ignote, allora S^2 è uno **stimatore corretto**, ossia

$$E(S^2) = \sigma^2 \qquad \text{per ogni } \sigma^2 > 0$$

Stima puntuale mediante il metodo della massima verosimiglianza

Il più importante metodo per la costruzione di stimatori puntuali si basa sulla **funzione di verosimiglianza**. Si consideri una variabile X discreta, la cui distribuzione dipenda solo dal parametro incognito θ e sia dato un certo campione osservato, ci possiamo chiedere qual è la possibilità di osservare quel dato campione per ogni θ.

La **funzione di verosimiglianza** $L(\theta)$ indica la **probabilità di osservare un campione fissato** al variare del parametro θ, ossia:

se X è discreta $L(\theta) = P(\text{dati osservati}; \theta) = \prod_{i=1}^{n} p(x_i; \theta)$ (vedi esempio pag. 309)

se X è continua $L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$

Se θ_1 θ_2 sono due distinti valori del parametro ignoto, e se $L(\theta_1) > L(\theta_2)$, diremmo che il valore θ_1 è **più verosimile** di θ_2.

La **stima di massima verosimiglianza** del parametro θ è il valore $\hat{\theta}$, che massimizza la funzione di verosimiglianza, ovvero:

$$L(\hat{\theta}) = \max L(\theta) \qquad \text{e} \qquad \log L(\hat{\theta}) = \max \log L(\theta)$$

Il modo più semplice di trovare il punto di massimo è quello basato sulla **derivata prima** della funzione log L (θ).

Gli **stimatori di massima verosimiglianza** per la media μ e la varianza σ^2 di una popolazione normale sono rispettivamente:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \text{e} \qquad \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})$$

Capitolo 12
Stima per intervallo

Stima per intervallo

In un **intervallo casuale** $[L_1(X_1, \ldots, X_n), L_2(X_1, \ldots, X_n)]$, si definisce **intervallo di confidenza** di livello $1 - \alpha$ per un parametro θ se contiene, con probabilità $1 - \alpha$, il parametro ignoto θ per una popolazione. Quindi

$$P[L_1(X_1, \ldots, X_n) \leq \theta \leq L_2(X_1, \ldots, X_n)] = 1 - \alpha$$ vedi esempio pag. 323-324

L'intervallo numerico $[l_1, l_2] = [L_1(x_1, \ldots, x_n), L_2(x_1, \ldots, x_n)]$ è una realizzazione dell'intervallo casuale $[L_1; L_2]$, ottenuta in corrispondenza del campione osservato, e viene quindi chiamato **intervallo di confidenza stimato**.

Analogie tra stima puntuale e stima intervallare

	Stima puntuale	**Stima intervallare**
Campione casuale	$X_1, \ldots, X_n$	$X_1, \ldots, X_n$
Obiettivo	Stima puntuale del parametro θ	Stima per intervallo del parametro θ
Strumento	Stimatore puntuale: $T = T(X_1, \ldots, X_n)$	Stimatore intervallo di confidenza: $[L_1, L_2] = [L_1(X_1, \ldots, X_n), L_2(X_1, \ldots, X_n)]$
Accuratezza	Errore quadratico medio: $MSE(T) = E[(T - \theta)^2]$	Livello di confidenza: $P(L_1 \leq \theta \leq L_2) = 1 - \alpha$
Dati campionari	$x_1, \ldots, x_n$	$x_1, \ldots, x_n$
Risultato	Stima puntuale: $t = T(x_1, \ldots, x_n)$	Intervallo di confidenza stimato: $[l_1, l_2] = [L_1(x_1, \ldots, x_n), L_2(x_1, \ldots, x_n)]$

Intervallo di confidenza per la media (σ^2 noto)

Nei problemi reali quando si vuole costruire un intervallo di confidenza per la media di una popolazione normale, raramente si conosce la varianza della popolazione. Il procedimento per ottenere un intervallo di confidenza per μ a partire da un campione casuale di dimensioni n, è tutto analogo a quello descritto nel paragrafo precedente. Tuttavia quando σ^2 è ignoto è necessario sostituirlo con una sua stima la **varianza campionaria corretta:**

$$S^2 = \frac{1}{n - 1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

se nella standardizzazione di $\bar{X}$, σ viene sostituito da $S = \sqrt{S^2}$ si ottiene la variabile casuale:

$$T = \frac{- \mu}{S/\sqrt{S^2}}$$

Naturalmente questa sostituzione non è priva di conseguenze.

La v.c. T è funzione delle due variabili casuali $\bar{X}$ e S, si distribuisce secondo una distribuzione t-student con $n - 1$ gradi di libertà.

Funzione pag 327

Intervallo di confidenza per una proporzione

Se si vuole studiare la presenza/assenza di un certo attributo A nella popolazione di interesse. La distribuzione del carattere può essere perciò rappresentata tramite una variabile casuale bernoulliana X il cui parametro d'interesse è π. Appellandosi al teorema del limite centrale, sappiamo che al crescere della dimensione campionaria la distribuzione della $\bar{X}$ può essere approssimata con quella di una normale con media π e varianza $\pi(1 - \pi)/n$ di conseguenza al crescere di n la variabile standardizzata:

$$\frac{\overline{X} - \pi}{\sqrt{\pi(1 - \pi)/n}}$$ tende a distribuirsi secondo una normale standardizzata

Osserviamo che in questo caso la varianza σ^2 gli estremi dell'intervallo dipendono dal parametro incognito π infatti:

$$1-\alpha \approx P\left(-Z_{\frac{\alpha}{2}} \le \frac{\overline{X}-\pi}{\sqrt{\pi(1-\pi)/n}} \le Z_{\frac{\alpha}{2}}\right) = P\left(\bar{X} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\pi(1-\pi)}{n}} \le \pi \le \bar{X} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

tuttavia all'aumentare della dimensione campionaria, poiché $\bar{X}$ è uno stimatore consistente di π anche lo stimatore $\bar{X}(1 - \bar{X})$ tenderà alla quantità $\pi(1 - \pi)$. Pertanto, la distribuzione della variabile casuale:

$$\frac{\overline{X}-\pi}{\sqrt{\overline{X}(1-\overline{X})/n}}$$

per una dimensione campionaria abbastanza elevata, uno stimatore sufficientemente accurato dell'**intervallo di confidenza per la proporzione** π al livello $1 - \alpha$ è dato da:

$$\left[\bar{X} - Z_{\frac{\alpha}{2}}\frac{\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{n}},\quad \bar{X} + Z_{\frac{\alpha}{2}}\frac{\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{n}}\right]$$

Capitolo 13
Teoria dei test statistici

Formulazione delle ipotesi

Il primo importante passo nella costruzione di un test statistico consiste nella definizione delle due possibili ipotesi, tra cui dobbiamo scegliere in base al risultato campionario (vedere es. pag. 343).

Per **ipotesi statistica** si intende una congettura riguardante un parametro θ della popolazione. Nell'approccio di Neyman-Pearson si distinguono due ipotesi contrapposte:

- **Ipotesi nulla**, indicata con H_0: si intende l'ipotesi preesistente all'osservazione dei dati campionari, ossia quella ritenuta vera fino a prova contraria.
- **Ipotesi alternativa**, indicata con H_1: si contrappone a quella nulla e potrebbe essere considerata più verosimile in base al risultato campionario.

Indichiamo con Θ lo **spazio parametrico**, ossia l'insieme di tutti i possibili valori che può assumere θ.
In termini generali, possiamo indicare le due ipotesi tramite un sistema del tipo:

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1 \end{cases}$$

SISTEMI DI IPOTESI

Le ipotesi possono essere semplici o composte (vedi es. pag. 345).

Un'ipotesi è detta **semplice** quando specifica completamente la popolazione; altrimenti è detta **composta**.

Se l'ipotesi composta riguardante il parametro θ individua un intervallo di valori, come per esempio $\theta \ge \theta_0$, questa si dirà **unidirezionale**, altrimenti, se è del tipo $\theta \ne \theta_0$, si dirà **bidirezionale**.

Regione di accettazione e regione di rifiuto

TEST STATISTICO

Il rifiuto o l'accettazione dell'ipotesi nulla dipende ovviamente dal campione osservato. Se l'informazione che si ricava dal campione contrasta in maniera evidente con l'ipotesi nulla, si rifiuta tale ipotesi; in caso contrario, accetteremo l'ipotesi nulla. Tale procedura è chiamata **test statistico**.

Un **test statistico** (o **test di ipotesi**) è una regola che permette di discriminare i campioni che portano all'accettazione dell'ipotesi nulla da quelli che portano al suo rifiuto.

STATISTICA TEST

La media campionaria è la statistica utilizzata per decidere se un determinato campione porta all'accettazione o al rifiuto dell'ipotesi nulla viene chiamata **statistica test**.

L'insieme dei valori della statistica test che portano all'accettazione dell'ipotesi nulla H_0 è chiamata **regione di accettazione**. L'insieme dei valori della statistica test che portano al rifiuto dell'ipotesi nulla H_0 è chiamata **regione di rifiuto** (vedere fig. pag. 347).

Si può osservare che la definizione di quali siano i valori appartenenti alla regione di accettazione dipende essenzialmente dal valore α scelto, detto anche **livello di significatività del test**: maggiore è il suo valore, più ampia sarà la regione di rifiuto.

La statistica test viene calcolata con la seguente formula:

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

Test con ipotesi nulla semplice

In questo sistema di ipotesi (**ipotesi alternativa bidirezionale o bilaterale**)

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_1 \end{cases}$$

Le regioni di rifiuto corrispondono alle due code della distribuzione, ciascuna pari a $\frac{\alpha}{2}$

Nel caso di ipotesi nulla o semplice, abbiamo altre due situazioni molto comuni:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta > \theta_1 \end{cases}$$

(**ipotesi alternative unidirezionali o unilaterali**)

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta < \theta_1 \end{cases}$$

Il p-value

Abbiamo visto che la conclusione alla quale ci conduce un test è quella di rifiutare o meno H_0 per un certo libello di significatività α prefissato. Tuttavia, poiché la scelta del valore α è arbitraria, la conclusione potrebbe dipendere da tale scelta. Un altro modo per evidenziare il risultato del test è quella di riportare il **p-value**.

Il **p-value** è dato dalla probabilità di osservare un valore della statistica test uguale o più estremo del valore ottenuto dal campione sotto l'ipotesi nulla.

Pertanto il p-value non è una quantità fissata come il livello di significatività, ma al contrario è una quantità che misura l'evidenza fornita dai dati contro l'ipotesi H_0: minore è il valore p-value, più è forte l'evidenza contro H_0.

Viene anche chiamato **livello di significatività osservato** (vedi es. pag. 351 con figura).

Errori del primo e del secondo tipo

Si commette un **errore del primo tipo** quando si rifiuta l'ipotesi nulla mentre questa è vera.

Si commette un **errore del secondo tipo** quando si accetta l'ipotesi nulla mentre questa è falsa (vedi tab. pag. 352).

Capitolo 14

Test per medie, proporzioni e varianze

Test per la media

In questo capitolo si affronteranno alcuni tipici problemi di verifica delle ipotesi, sulla base della teoria generale. Si considererà innanzitutto test per la verifica di ipotesi riguardanti i parametri di una singola popolazione e quelli relativi a ipotesi sulla media.

Si assuma che la popolazione di interesse segua una certa distribuzione di forma nota e che si voglia verificare un'ipotesi nulla H_0 riguardante la sua media sulla base di un campione di dimensioni n.

Per una popolazione normale, di varianza nota, il test sulla media può essere costruito facilmente:

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

μ_0 = valore della media ipotizzato

Z = differenza tra la media campionaria e il suo valore atteso

$\sigma / \sqrt{n}$ = deviazione standard

Sotto l'ipotesi nulla H_0 la statistica test Z si distribuisce come una normale standardizzata, perciò, fissato il livello di significatività, si hanno le seguenti regioni di rifiuto:

Ipotesi alternativa	Regione di rifiuto
$H_1 : \mu > \mu_0$	$Z \geq z_\alpha$
$H_1 : \mu < \mu_0$	$Z \leq -z_\alpha$
$H_1 : \mu \neq \mu_0$	$\lvert Z \rvert \geq z_{\alpha/2}$

(vedi es. pag. 367)

Test per la media di una popolazione normale con varianza incognita

In molte situazioni, la varianza della popolazione non è nota, quindi è necessario stimarla. Si usa in questo caso la varianza campionaria corretta S^2, ottenendo la seguente formula della statistica test:

$$T = \frac{\overline{X} - \mu_0}{S / \sqrt{n}}$$

Tale sostituzione implica che, quando il vero valore di μ è μ_0, la statistica T si distribuisca secondo una T-student, con n–1 gradi di libertà. Pertanto, si hanno le seguenti regioni di rifiuto:

Ipotesi alternativa	Regione di rifiuto
$H_1 : \mu > \mu_0$	$T \geq t_\alpha$
$H_1 : \mu < \mu_0$	$T \leq -t_\alpha$
$H_1 : \mu \neq \mu_0$	$\lvert T \rvert \geq t_{\alpha/2}$

(vedi es. pag. 368)

Test per la media di una popolazione non normale

Se la popolazione non segue la funzione di normalità, la statistica test dipende dalle dimensione campionaria del test. In particolare, sotto alcune condizioni di regolarità, si ha che sotto H_0 vale il seguente risultato:

Al tendere di n a infinito la statistica test è: $\dfrac{\overline{X} - \mu_0}{S / \sqrt{n}}$

La funzione tende a distribuirsi come una normale standardizzata (teorema del limite centrale – pag.15 dei riassunti).

(vedi es. pag. 369)

Come stabilire la dimensione campionaria

Finora n è stata considerata sempre come una quantità fissata nota. Tuttavia può essere interessante fissare n in modo tale che il test raggiunga una certa potenza sotto una specifica ipotesi alternativa.

Si tratta di determinare il valore di n in corrispondenza di un valore obiettivo degli errori α e β. La procedura per la determinazione della numerosità prevede i seguenti passi:

1. specificare la probabilità dell'errore di prima specie α;
2. specificare il valore μ_1 e il corrispondente valore β;
3. selezionare una stima iniziale di σ;
4. calcolare la numerosità campionaria.

(vedi es. pag. 371)

Test per una proporzione

Si consideri la situazione in cui X è una variabile casuale dicotomica e che segue una distribuzione bernoulliana con parametro π ($0 < \pi < 1$). In questo caso possiamo essere interessati a uno dei seguenti sistemi di ipotesi:

1. $H_0 : \pi = \pi_0$ contro $H_1 : \pi > \pi_0$
2. $H_0 : \pi = \pi_0$ contro $H_1 : \pi < \pi_0$
3. $H_0 : \pi = \pi_0$ contro $H_1 : \pi \neq \pi_0$

Per verificare queste ipotesi si può utilizzare la seguente statistica test:

$$Z = \frac{\overline{X} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

Si hanno le seguenti regioni di rifiuto:

Ipotesi alternativa	Regione di rifiuto
$H_1 : \pi > \pi_0$	$Z \geq z_\alpha$
$H_1 : \pi < \pi_0$	$Z \leq -z_\alpha$
$H_1 : \pi \neq \pi_0$	$\lvert Z \rvert \geq z_{\alpha/2}$

(vedi es. pag. 372)

Capitolo 16

Il modello di regressione lineare semplice

Relazione funzionale e relazione statistica tra due variabili

Quando si analizzano due o più caratteri quantitativi si può cercare di individuare una funzione che descriva in modo dettagliato la relazione che emerge dai dati. Se una delle variabili è considerata dipendente dall'altra, si utilizzerà un **modello di regressione**.

Si considerino due variabili quantitative Y X e supponiamo di essere interessati a comprendere come la variabile Y, considerata una variabile **dipendente o risposta**, sia influenzata dalla X, che assumiamo essere una variabile **esplicativa o indipendente**. Una variabile Y è una funzione di X se a ogni valore di X corrisponde uno e un solo valore di Y (**relazione funzionale**).

Una **relazione funzionale lineare** si può scrivere come:

$$Y = \beta_0 + \beta_1 X$$

Negli studi empirici, la relazione che può essere osservata tra le variabili Y e X non è mai una relazione matematica esatta; infatti, a un determinato valore di X possono corrispondere più valori di Y.

Per descrivere e analizzare fenomeni empirici è opportuno introdurre una relazione più complessa di quella funzionale, che prende il nome di **relazione statistica**. Una relazione statistica tra una variabile indipendente X e una variabile dipendente Y può essere descritta dall'equazione:

$$Y = f(X) + \varepsilon$$

ε = contributo di tutti gli altri fattori

In una relazione statistica vi è una componente deterministica rappresentata da f(x) e una componente stocastica rappresentata dalla variabile casuale ε.

Specificazione del modello di regressione lineare semplice

Il più semplice modello di regressione è il modello di **regressione lineare semplice** $f(X) = \beta_0 + \beta_1 X$, dove β_0 e β_1 vengono chiamati **coefficienti di regressione**.

La forza del modello lineare semplice dipende dal fatto che una funzione di regressione che non è lineare può essere spesso approssimata tramite una retta. Le assunzioni del modello di regressione lineare semplice si riferiscono al processo che genera le n coppie di dati disponibili. Queste assunzioni sono:

Assunzione 1: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ per ogni osservazione di i = 1, …, n

Assunzione 2: Le ε_i sono v.c. dipendenti con valore atteso $E(\varepsilon_i) = 0$ e varianza costante $V(\varepsilon_i) = \sigma^2$

Assunzione 3: I valori x_i della variabile esplicativa X sono noti senza errore.

La prima assunzione implica che tre le possibili funzioni f(X), che possono descrivere il legame tra la variabile indipendente e la variabile esplicativa, si è scelta la funzione lineare. Ogni ε_i è una v.c. poiché rappresenta gli **scostamenti di Y_i**. Si assume allora che le v.c. ε_i siano tra loro **statisticamente indipendenti**.

Poiché σ^2 è la media delle deviazioni al quadrato di tutti i possibili valori di ε_i e poiché questi valori hanno media nulla, si ottiene che **σ^2 è una misura della grandezza di ε_i**.

Poiché ε_i è una variabile casuale, anche la variabile dipendente Y, somma di una componente deterministica e di una stocastica, è **variabile casuale**.

Dalle assunzioni 2 e 3 si ricava che il valore atteso di Y, condizionato da $X = x_i$ è

$$E(Y_i \mid X = x_i) = \beta_0 + \beta_1 x_i$$

La varianza di Y_i è $V(Y_i) = V(\varepsilon_i) = \sigma^2$

Stima puntuale dei coefficienti di regressione

La vera retta di regressione deve passare attraverso una nuvola di punti, avvicinandosi il più possibile a essi. Chiameremo $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ il valore y fornito dalla retta stimata in corrispondenza di x_i. Il problema consiste nell'individuare i coefficienti di regressione (quindi β_0 e β_1) in modo tale che i valori stimati siano il più possibile vicini ai valori osservati. Si dirà che una retta ha un **migliore adattamento** ai dati osservati se, fissati i valori dei coefficienti di regressione, complessivamente gli scarti sono più piccoli.

Per valutare gli scarti si usa **il metodo di stima dei minimi quadrati**:

$$G(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Si chiamerà **residuo i-esimo** ($\hat{e}_i$) la differenza tra il valore osservato e il valore fornito dalla retta di regressione e si calcolerà così:

$$\hat{e}_i = y_i - \hat{y}_i$$

Per individuare i valori di β_0 e β_1 occorre calcolare le derivate parziali di $G(\beta_0, \beta_1)$ rispetto a β_0 e β_1 e porle uguali a zero. Dopo alcuni passaggi di semplificazione si ottengono le **stime dei coefficienti di regressione**. **le stime dei minimi quadrati dei coefficienti di regressione** sono date da:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$\bar{x}$ e $\bar{y}$ sono rispettivamente le medie campionarie di X e di Y

Poiché il numeratore di $\hat{\beta}_1$ è n volte la covarianza campionaria, mentre il denominatore è n volte la varianza campionaria non corretta della X, la stima di $\hat{\beta}_1$ può anche essere espressa come:

$$\hat{\beta}_1 = \frac{\sigma_{xy}}{\sigma_x^2} \qquad \text{oppure} \qquad \hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

La decomposizione della varianza totale e il coefficiente di determinazione

Le stime dei minimi quadrati dei coefficienti di regressione possiedono un'importante proprietà che consente di valutare le capacità previsive del modello stimato. Si può dimostrare che i valori stimati soddisfano la seguente relazione:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \hat{e}_i^2$$

La suddetta relazione è chiamata **decomposizione della varianza totale**.

Il termine a sinistra del segno di uguaglianza è la **devianza** della variabile dipendente Y, detta **somma totale dei quadrati** (SQT), quindi è:

$$SQT = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Il primo termine a destra dell'uguaglianza è detto **somma dei quadrati della regressione** (SQR), quindi è:

$$SQR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

Il secondo termine a destra dell'uguaglianza è detto **somma dei quadrati degli errori** (SQE), quindi è:

$$SQE = \sum_{i=1}^{n} \hat{e}_i^2$$

Pertanto: $SQT = SQR + SQE$

Da questa relazione si può definire un **indice statistico**, che misura la **bontà di adattamento** della retta di regressione ai dati. Dividendo SQR per il suo valore massimo SQT otteniamo il **coefficiente di determinazione**:

$$R_{xy}^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}$$

Il coefficiente di determinazione varia tra zero e uno. Vale zero **in assenza di relazione lineare statistica** tra le osservazioni, vale uno **in presenza di perfetta relazione lineare**.

Si può dimostrare che R_{xy}^2 corrisponde al quadrato del **coefficiente di correlazione lineare** tra x e y:

$$R_{xy}^2 = (p_{xy})^2 = \left(\frac{\sigma_{xy}}{\sigma_x \sigma_y}\right)^2$$

Inoltre, poiché $\hat{\beta}_1 = p_{xy}\frac{\sigma_y}{\sigma_x}$, la pendenza della retta di regressione corrisponde al segno del coefficiente di correlazione.

Proprietà degli stimatori dei coefficienti e della risposta media

Ovviamente, le stime dei coefficienti del modello di regressione lineare dipendono dal campione osservato, e al variare di questo generano le variabili casuali **stimatori dei coefficienti di regressione**, che indicheremo con B_0 e B_1.
È importante a questo punto considerare quali proprietà posseggono tali stimatori.

PROPRIETÀ DEGLI STIMATORI DEI MINIMI QUADRATI

1. B_0 e B_1 sono stimatori corretti di β_0 e β_1;
2. Nella classe degli stimatori corretti β_0 e β_1, che sono funzioni lineari di Y_i, gli stimatori dei minimi quadrati sono i più efficienti, cioè sono quelli che hanno **minima varianza** per qualsiasi valore dei parametri;
3. La varianza e la covarianza degli stimatori dei minimi quadrati sono dati da:

$$V(B_0) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad V(B_1) = \sigma^2 \left[\frac{1}{n} + \textbf{Errore.} \right]$$

$$Cov(B_0, B_1) = -\sigma^2 \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = -\bar{x} V(B_1)$$

Una misura della variabilità degli stimatori dei coefficienti di regressione è data dagli **errori standard**, ossia dalle radici quadrate di $V(B_0)$ e $V(B_1)$, indicate da

$$\sigma(B_0) = \sqrt{V(B_0)} \qquad \text{e} \qquad \sigma(B_1) = \sqrt{V(B_1)}$$

Lo **stimatore corretto della varianza dei residui** è dato da:

$$s^2 = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-2}$$