Statistica di base mecatti

Sociologia (Università degli Studi di Milano-Bicocca)

STATISTICA: insieme di metodologie e tecniche per la trattazione quantitativa dei fenomeni osservabili nella realtà sociale, in natura, in laboratorio.
Per trattazione quantitativa intendiamo un percorso logico che prevede: l'*osservazione*(rilevazione), l'*analisi* (elaborazione), la *comprensione* (trasformazioni di dati in informazioni). Successivamente prendiamo decisioni. Questo processo è svolto per prendere qualsiasi tipo di decisione.
Definizioni base:
Popolazione statistica (**U**) -> collettivo delle unità statistiche su cui interessa un particolare fenomeno. Può essere chiamato target, è composto da un *insieme* di unità statistiche. È la popolazione su cui manifesta un fenomeno.
Fenomeno statistico (**X**) è il fenomeno d'interesse per la statistica, è la cornice di *caratterizzazione* o un *concetto.*
Manifestazione/modalità(**x**) -> le modalità in cui si manifesta il fenomeno (può essere svariate cose).
Numerosità di U (**N**) -> in genere è un numero intere, se è un numero tanto elevato può essere anche considerato come ∞. È il numero di unità statistiche che compongono la popolazione.
Classificazione dei fenomeni statistici
-> QUALITATIVI (fenomeni che si manifestano nella popolazione osservata attraverso attributi o categorie, qualità appunto). Possono essere ordinali (si manifestano con attributi e categorie che si possono ordinare secondo un qualche criterio oggettivo e convenzionalmente accettato) e categoriali (non c'è un criterio oggettivo per ordinare le categorie).
-> QUANTITATIVI (si manifestano nella popolazione osservata attraverso numeri, quantità appunto). Possono essere discreti(possiamo contare, enumerare) e continui (si possono misurare con una unità di misura o con un intervallo).
Scale di modalità: sono costituite dall'insieme di tutte le diverse manifestazioni di X su U. Devono rispettare due principi generali: esaustività (deve prevedere tutte le possibili manifestazioni di X che potenzialmente si possono osservare su U) e mutua esclusività (le modalità si devono escludere a vicenda).
Le scale di modalità sono:

- ➔ QUALITATIVE (se le modalità sono attributi o categorie) . Si dividono in scale qualitative ordinali (se sono ordinabili secondo un criterio oggettivo o convenzionalmente accettato) o sconnesse (se non possono essere ordinate secondo un criterio oggettivo).
- ➔ QUANTITATIVE (le modalità sono numeri).. Si dividono in scale quantitative rapporto (l'origine è 0 e ha un significato assoluto, cioè assenza del fenomeno) o non rapporto (l'origine non è assoluto ma convenzionale -es. gradi).

Si dice *scala dicotomica (o binaria)* una SdM con solo due modalità. Indicheremo con *k* il numero di diverse modalità della scala utilizzata. Per quanto riguarda gli intervalli indicheremo con x_l l'*estremo inferiore* e con x_L l'*estremo superiore*. Indichiamo con x_i la generica manifestazione; con x_i:x_l⊢x_L quando x_i è un intervallo; con i=1….k le diverse manifestazioni del fenomeno.
Alcuni fenomeni quantitativi possono essere rilevati su scala qualitativa, e viceversa. La natura di un fenomeno può essere discreta o quantitativa in base alle modalità che vogliamo attribuirgli noi: se un numero specifico o un intervallo.
La STATISTICA ha due funzioni: statistica descrittiva (funzione di descrivere il comportamento di X su U) e la statistica inferenziale (estendere i dati osservati, sull'intera popolazione).
La statistica descrittiva si compone di: -**statistica mono-variata** (ha per oggetto un unico fenomeno e come obbiettivo la descrizione sintetica del suo comportamento su U); -**statistica bi-variata** (due fenomeni, obbiettivo: individuazione e studio delle eventuali relazioni statistiche fra i due); -**multi-variata** (fenomeni sono più di due, obbiettivo: descriverne il comportamento congiunto e studiarne le relazioni, congiuntamente e per loro sottoinsiemi).
La statistica inferenziale è basata su dei campioni di tipo casuale scelti sulla totalità dei dati che esaurirebbero l'osservazione di U. Vi sono elementi di teoria delle probabilità.

DISTRIBUZIONI DI FREQUENZE, TABELLE E GRAFICI

Il risultato della rilevazione è una serie confusa di modalità x_i che si manifestano su N e sono dati in modo sparso. Questi vengono definiti dati grezzi i quali devono essere posti a sintesi successive, con l'obbiettivo di far emergere dati e informazioni utili a descrivere il comportamento di X su U. La prima fase di sintesi consiste nella creazione di tabelle e grafici che rendano i dati più leggibili. Queste tabelle prendono il nome di variabili statistiche.

La frequenza assoluta di ciascuna modalità osservata x_i è il numero di unità statistiche, tra le N osservate, manifesta quella modalità x_i di X. Indicheremo la frequenza assoluta con f_i.

L'insieme delle *k* frequenze (assolute) è detta distribuzione di frequenze assolute di X su U.

N è la somma delle frequenze assolute. Il complesso della tabella (p. 27) costituisce la v.s. (k coppie di tipo modalità, frequenza).

Schema p.27

x_i contiene le modalità, mentre f_i può contenere solo numeri interi ≥ 0 e con somma pari a N. Il complesso della tabella costituisce la variabile statistica, che è l'insieme di k coppie del tipo "modalità, frequenza".

Se l'obbiettivo è confrontare le distribuzioni di frequenze di X in due (o più) popolazioni dovremo depurare le frequenze assolute dall'influenza di N costruendo le frequenze relative. La frequenza relativa associata alla modalità x_i è il rapporto tra le frequenze assolute e la numerosità N. Indicheremo la frequenza relativa con p_i e in formule : $p_i = \frac{f_i}{N}$. Otteniamo il peso che ciascuna modalità ha sull'intera popolazione.

DIMOSTRAZIONE DELLA FORMULA:

$\sum_{i=1}^{k} p_i = 1$ *k*=numero qualunque intero.

N

$$\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} \frac{f_i}{N} = \frac{f_1 + f_2 + \cdots + f_k}{N} = \frac{1}{N}\sum_{i=1}^{k} f_i = \frac{1}{N} \times N = 1$$

La colonna delle frequenze relative costituisce la distribuzione di frequenze relative di X su U.

Quando il fenomeno è **almeno ordinale** (qualitativo ordinale o quantitativo) possiamo fare un'ulteriore analisi. Quando abbiamo questi tipi di analisi è consuetudine costruire la v.s. ponendo in ordine in senso crescente le x_i. Sommare, tecnicamente *cumulare*, le frequenze associate alle modalità inferiori di x_i, ci fa costruire le frequenze cumulate. Indicheremo le frequenze assolute cumulate con F_i e le frequenze relative cumulate con la "phi" maiuscola: Φ_i.

$F_i = f_1 + f_2 + \cdots + f_i = \sum_{j=1}^{i} f_j$ $\quad \Phi_i = p_1 + p_2 + \cdots + p_i = \sum_{j=1}^{i} p_j = \frac{F_i}{N}$

$F_i = Freq\,\{X \leq x_i\}$ -> fenomeno che si presenta in maniera minore o uguale alla i-esima.

$\Phi_i = Freq\ rel\,\{X \leq x_i\}$ -> freq.rel.con cui il fenomeno si presenta sulla popolazione in maniera ≤alla i-esima.

Proprietà delle frequenze cumulate:

- ➔ Fenomeni almeno ordinali
- ➔ $0 < F_i < N$; $F_1 = f_1$; $F_k = N$ $\qquad$ $0 < \Phi_i < 1$; $\Phi_1 = p_1$; $\Phi_k = 1$
- ➔ Tra le frequenze cumulate esiste una corrispondenza biunivoca e ricorsiva: se conosciamo le frequenze (assolute o relative) possiamo ottenere le cumulate e se conosciamo le cumulate possiamo ottenere le frequenze. In formule: $F_i - F_{i-1} = f_i \;\; e \;\; F_{i-1} + f_i = F_i$; $\Phi_1 - \Phi_{i-1} = p_i \;\; e \;\; \Phi_{i-1} + p_i = \Phi_i$

Discorso diverso si deve fare per quanto riguarda i fenomeni quantitativi continui che si rilevano con degli intervalli. La v.s. ci informa che in quell'intervallo ci sono f_i unità statistiche, ma non ci informa in che modo

sono distribuite all'interno dell'intervallo. Allora possiamo fare due ipotesi (entrambe corrette): l'ipotesi del valore centrale (l'obbiettivo è qui di assegnare a ciascuna delle unità statistiche che cadono nell'intervallo x_l:x_l⊦$_L$ un unico punto, interno all'intervallo stesso: $x_i^* = \frac{x_l + x_L}{2}$. Questo di solito fa riferimento a fenomeni con intervalli di ampiezza uniforme) e l'ipotesi di distribuzione uniforme (possiamo assumere l'ipotesi che le f_i siano distribuite in modo uniforme ed equidistante lungo tutto l'intervallo. Questo è utilizzato di più quando gli intervalli sono di ampiezza diversa).
L'ampiezza dell'intervallo è la differenza tra l'estremo superiore e l'estremo inferiore $x_L - x_l$.
Per riuscire a confrontare due intervalli tra di loro abbiamo bisogno della densità di frequenza, la quale è la frequenza dell'intervallo depurata dall'influenza dell'ampiezza; la indicheremo con la phi minuscola φ_i.

$$\varphi_i = \frac{f_i}{x_L - x_l}$$

Le densità di frequenza φ_i danno un'idea dell'addensamento delle frequenze all'interno degli intervalli e sono utili quando le diverse ampiezze degli intervalli rendono difficile l'interpretazione delle frequenze.
Il concetto di densità di frequenza può essere anche rappresentato utilizzando le frequenze relative invece che le assolute, a seconda del contesto:

$$\frac{p_i}{x_L - x_l} = \frac{f_i}{N(x_L - x_l)} = \frac{\varphi_i}{N}$$

Il metodo più opportuno per disegnare e per capire meglio una v.s. è attraverso dei grafici, soprattutto l'istogramma. Sulle ordinate (le x) scriviamo le modalità (o gli intervalli) e sulle ascisse (le y) le variabili.
Per quanto riguarda i fenomeni quantitativi continui (rilevati tramite gli intervalli) possiamo disegnarli in due modi: *istogramma a bastoncini* (ipotizziamo che i valori siano sul valore centrale dell'intervallo e eleviamo un bastoncino solo dal valore centrale dell'intervallo) oppure con *istogramma vero e proprio* (eleviamo un rettangolo per ogni area dell'intervallo, se ipotizziamo che i valori sono equamente distribuiti all'interno dell'intervallo). Sulle y andranno o φ_i oppure φ_i/N.
L'area dei rettangoli sarà base per altezza (dove l'altezza sarò rappresentata appunto dalla densità di frequenza, l'area dalla frequenza e la base l'estremo superiore meno l'estremo inferiore).
ESEMPI SUL QUADERNO E SUL LIBRO A P.39 E SS.

VALORI MEDI

I valori mdi sono dei valori sintetici che mettono in evidenza un particolare aspetto del comportamento di X su U. Quello che affronteremo sono moda, mediana e media aritmetica.
MODA -> la moda di una v.s. è la modalità a cui è associata la frequenza più elevata tra le k osservate. La moda si può calcolare per ogni tipo di fenomeno, l'informativa che ci dà è ridotta ma può essere associata la sua p_i in modo da avere più informazioni. Per indicare la moda useremo la notazione x_0 (x con zero). Chiamiamo intervallo modale quello a cui è associata la densità φ_i più elevata tra le k osservate. è più conveniente far coincidere la moda con il valore centrale dell'intervallo.
MEDIANA -> la mediana è un valore medio utilizzabile solo sui fenomeni almeno ordinali (quantitativi oppure qualitativi ordinali). La mediana di X è la modalità che, nell'ordinamento, occupa la posizione centrale. La indicheremo con $x_{0.5}$ (x con zero cinque). Il 50% di U manifesta modalità $x_i \leq x_{0.5}$, e il restante 50% $x_i \geq x_{0.5}$
Come si calcola? Non appena le frequenze relative cumulate ϕ_i superano il 50%, quella è la mediana.
Per quanto riguarda i fenomeni continui (da rilevare con intervalli), possiamo ipotizzare che la mediana sia il valore centrale x_i^* di quell'intervallo; oppure ipotizziamo che i valori siano equamente distribuiti all'interno dell'intervallo. In quest'ultimo caso (caso più accettabile) per calcolare la mediana ci sarà un calcolo più complicato: $x_{0.5} = x_i + \left(\frac{N}{2} - F_{i-1}\right)\frac{x_L - x_l}{f_i} = x_l + (0.5 - \phi_{i-1})\frac{x_L - x_l}{p_i}$.
PAGINA 54+QUADERNO: Prendiamo l'istogramma costruito con le frequenze assolute f_i e guardiamo il pezzo di istogramma che riguarda l'intervallo mediano. Il "pezzo" che ci interessa è un rettangolo di altezza $\varphi_i = f_i/x_L - x_l$, di base $x_L - x_l$e di area f_i. Per determinare la mediana bisogna aggiungere a x_l il pezzo che manca

per raggiungere $x_{0.5}$. Sull'istogramma le aree sono le frequenze, sappiamo che tutta l'area dell'istogramma vale N e che la mediana divide N in due parti (N/2). L'area a sinistra di x_l coincide con la frequenza di tutte le modalità $\leq x_l$, cioè la frequenza cumulata F_{i-1}. Ne segue che per differenza possiamo calcolare l'area del sotto-rettangolo che ci interessa: $area\ del\ sottorettangolo\ evidenziato = \frac{N}{2} - F_{i-1}$.

Siccome l'area del rettangolo è base per altezza, si ottiene la base dividendo l'area per l'altezza: $base\ del\ sottorettangolo\ evidenziato = \frac{\frac{N}{2} - F_{i-1}}{\varphi_i}$.

Infine cci ricordiamo la definizione di densità di frequenza: $\varphi_i = {f_i}/{x_L - x_l}$.

Mettiamo insieme tutti i pezzi: $x_{0.5} = x_l + \frac{\left(\frac{N}{2} - F_{i-1}\right)}{\varphi} = x_l + \left(\frac{N}{2} - F_{i-1}\right)\frac{x_L - x_l}{f_i}$.

La stessa formula si può anche calcolare utilizzato le frequenze relative e sarà: $x_{0.5} = x_l + (0.5 - \phi_{i-1})\frac{x_L - x_l}{p_i}$

MEDIA ARITMETICA-> La media aritmetica (che indicheremo con $\bar{x}$, "x medio") è calcolabile su fenomeni quantitativi (o qualitativi ordinali rilevati con scala quantitativa), è espressa con la stessa unità di misura con cui X si manifesta su U, ci dà un'informazione sintetica dell'ordine di grandezza di X su U ed è una sintesi dell'INTERA v.s.. è semplice da calcolare: bisogna moltiplicare le k modalità osservate per le f_i, sommare il tutto e infine dividere per il numero N di unità statistiche osservate. In formule $\bar{x} = \frac{1}{N}\sum_{i=1}^{k} x_i f_i = \sum_{i=1}^{k} x_i p_i$.

Se X è quantitativo continuo e le sue modalità sono degli intervalli, la media $\bar{x}$ è in genere calcolata sull'ipotesi del valore centrale x_i^*.

PROPRIETA' DELLA MEDIA ARITMETICA

Proprietà di internalità. Il valore della media aritmetica è sempre compreso tra la più piccola e la più grande delle modalità osservate di X: in formula $x_{min} \leq \bar{x} \leq x_{max}$.

Proprietà di omogeneità.Se X e Y sono due fenomeni diversi ma collegati tra lodo dalla formula $Y = aX$ dove a è un qualunque numero (costante) diverso da 0, si dice che Y è una trasformazione di scala di X: la media aritmetica di Y si ottiene dalla media aritmetica di X, con la stessa identica trasformazione: $\bar{y} = a\bar{x}$.

Proprietà associativa. Quando U è molto numero è una pratica sensata utilizzare dati aggregati anziché dati individuale. Formalmente si tratta di considerare U di numerosità N, suddivisa in un certo numero, diciamo h, di sottopopolazioni U_j ciascuna di numerosità N_j con j=1,…,h e $\sum_{j=1}^{h} N_j = N$. Quello che ci interessa è sempre sapere la media generale sull'intera U. Non disponiamo però dei dati individuali (le xi e le fi) ma solo dei dati aggregati, cioè le medie $\bar{x_j}$ nelle sottopopolazioni.

SCHEMA P.69

La proprietà che ci serve è quella associativa: la media (generale) di X (su U) è sempre raggiungibile dai dati aggregati (sulle sottopopolazioni U_j), basta calcolare la media delle medie delle sottopopolazioni. Si tratta di usare le medie parziali$\bar{x}_j$ al posto delle modalità x_i e le numerosità N_j al posto delle frequenze f_i. In formule: $\bar{x} = \frac{1}{N}\sum_{J=1}^{h} \bar{x}_j N_j$. -> formula non ufficiale…. $\bar{x} = \frac{\bar{x}_1 \times N1 + \bar{x}_2 \times N2 + \cdots + \bar{x}_k \times Nk}{N1 + N2 + \cdots + Nk}$.

Proprietà di annullamento degli scarti. La media aritmetica svolge il suo lavoro di sintesi della v.s. garantendo la compensazione delle differenze tra i valori x_i osservati e il valore medio di sintesi $\bar{x}$. In formule è più chiaro: Le differenze $(x_i - \bar{x})$ sono dette scarti o deviazioni della media aritmetica. Se poi si tiene conto del fatto che il valore x_i è presente su U con frequenza f_i, si ha lo scarto ponderato $(x_i - \bar{x})f_i$. Quando lo scarto è positivo si ha un valore sopra-media, se no sotto-media. Proprietà: I valori sopra e sotto-media si compensano, cioè se si sommano tutti i k scarti ponderati si ottiene (sempre) 0. È garantito solo per la media.

DIMOSTRAZIONE di $\sum_{i=1}^{k}(x_i - \bar{x}\ )f_i = 0$

$$\sum_{i=1}^{k}(x_i-\bar{x})f_i=\sum_{i=1}^{k}x_if_i-\sum_{i=1}^{k}\bar{x}f_i=N\bar{x}-\bar{x}\sum_{i=1}^{k}f_i=N\bar{x}-\bar{x}N=0$$

Proprietà di mantenimento e di equidistribuzione del totale. La somma di tutti i valori di su tutte le N unità osservate prende il nome di *totale di X*: in formule $\sum_{i=1}^{k}x_if_i=\boldsymbol{totale\ di\ X}\ (\boldsymbol{su\ U})$.
Questa formula è uguale a $N\bar{x}$ che è uguale a $\sum_{i=1}^{k}\bar{x}f_i$. Questa formula definisce un'altra proprietà esclusiva della media aritmetica: se ai valori x_i osservati sostituiamo la media aritmetica $\bar{x}$ che sintetizza tutti, il totale di X non cambia. Allora la media aritmetica mantiene inalterato il totale; inoltre, se il totale di X fosse diviso in parti uguali tra le N unità di U, a ciascuna unità toccherebbe una quota di totale pari a $\bar{x}$. Allora la media aritmetica equidistribuisce il totale di X su N unità di U.

VARIABILITA'

ESEMPIO DI TRILUSSA PAGINA 83

La variabilità (o dispersione di X) è l'attitudine di un fenomeno quantitativo a manifestarsi nelle N unità di U, con modalità tra loro diverse e distanti. È lo scopo della statistica: la variabilità è ciò che rende necessario il ricorso alla strumentazione statistica per l'analisi e la comprensione di un fenomeno su U.
La variabilità assume valore 0, in assenza di essa (ovvero quando le modalità sono costanti); assume valori positivi quando X si manifesta su U con molteplici e differenti modalità e assume valori sempre più elevati all'aumentare della variabilità. Una misura di variabilità che utilizza tutta la v.s. è la deviazione standard di X (chiamata anche scarto quadratico medio). Si tratta della misura di variabilità più nota e utilizzata e si identifica con la lettera sigma (σ). Questa confronta ciascuna delle k modalità osservate con un unico valore fisso scelto come polo di confronto.

Formula della deviazione standard. $\sigma=\sqrt{\frac{1}{N}\sum_{i=1}^{k}(x_i-\bar{x})^2f_i}$ Formula alternativa.$\sigma=\sqrt{\frac{1}{N}\sum_{i=1}^{k}x_i^2f_i-\overline{x^2}}$

Come salta fuori la formula? Ogni modalità è confrontata con la media aritmetica(la quale essendo una sintesi della v.s. è un ottimo punto di riferimento); la differenza (xi-$\bar{x}$) può risultare positiva o negativa, il segno qui è ininfluente e ci interessa la distanza dalla media, quindi si eleva al quadrato in modo da enfatizzare la distanza e facilitare i calcoli; gli scarti quadratici, poi, vengono ponderati con le frequenze; poiché gli scarti sono k, li sintetizziamo tutti in una media sommando e dividendo poi per N; infinte, si ristabilisce l'ordine di grandezza e dell'unità di misura inserendo la radice quadrata.
σ misura la variabilità di X considerando la dispersione dei valori intorno al loro valore medio. Ci dice che X si manifesta su U con valori che in media distano da $\bar{x}\ per\ \pm\sigma$.
A partire da sigma possiamo calcolare la varianza e la devianza. La varianza si compone elevando al quadrato tutto sigma, in modo però da avere alterato il risultato e non è una buona misura di variabilità; però ha vantaggi nel calcolo. La devianza deriva dalla varianza moltiplicata per N.$Varianza\rightarrow\ \sigma^2=\frac{1}{N}\sum_{i=1}^{k}(x_i-\bar{x})^2f_i\ \ e\ Devianza\rightarrow\ N\sigma^2=\sum_{i=1}^{k}(x_i-\bar{x})^2f_i$
Per confrontare la variabilità di un fenomeno rispetto alla variabilità di un altro fenomeno abbiamo bisogno di una formula relativa: quella più utilizzata è il coefficiente di variabilità di X che si costruisce ponendo la deviazione standard a rapporto con la media aritmetica. $cv=\frac{\sigma}{\bar{x}}$. Ricordiamoci che il risultato è relativo alla media non a N !!.

NUMERI INDICE

Quando analizziamo uno stesso fenomeno che viene ripetuto nel tempo si parla di dati longitudinali; per rilevare questi dati, non utilizziamo la v.s. ma una serie storica. Quando si rileva una serie storica l'obbiettivo è di descrivere e analizzare il comportamento di X nel tempo. Per analizzare le serie storiche dovremo creare degli indicatori sintetici per analizzare l'evoluzione nel tempo (numero indice). Utilizzeremo la t minuscola per indicare gli istanti temporali di osservazione; la T maiuscola per l'ultimo istante di rilevazione. Il numero indice è il rapporto tra due modalità x_t rilevate in due differenti istanti temporali. Il numero indice può essere costituito a base fissa o a base mobile.

A base fissa, si sceglie un istante temporale come base da porre al denominare e rimane invariata per l'intera serie di indici; in genere si sceglie l'istante iniziale di rilevazione.$Indice\ a\ base\ fissa\ (istante\ 1) = \frac{x_t}{x_1} \times 100$

A base mobile, si fa il rapporto di ciascuna modalità in base a quella precedente: $\frac{x_t}{x_{t-1}}$.

Utilizzeremo la notazione NI.

Se vogliamo sapere quanto ci manca (o quanto abbiamo in più) per arrivare a 100 usiamo la variazione temporale di X. $v = (NI\ a\ base\ fissa - 100) \quad v = (NI\ a\ base\ mobile - 100)$. La prima è la variazione percentuale rispetto all'anno base, la seconda è la variazione percentuale annua.

Le serie di NI e le variazioni percentuali possono essere viste come v.s. quindi possono essere sintetizzate con un valore medio che prende il nome di tasso di variazione medio annuo di X. È la percentuale con cui X è mediamente variato di anno in anno lungo tutto il periodo della serie. Lo indicheremo con $\bar{v}$ (vi medio).

$$\bar{v} = \left(\sqrt[T-1]{\frac{x_T}{x_1}} - 1 \right) \times 100 = \left[\left(\frac{x_T}{x_1} \right)^{\frac{1}{T-1}} \right] \times 100$$

Se $\bar{v}$ è minore di 0 l'evoluzione è in diminuzione; se no in aumento. $\bar{v} = 0$ non ci sono evoluzioni.

TABELLE A DOPPIA ENTRATA

L'obbiettivo, con la statistica descrittiva bi-variata, diventa la descrizione del comportamento congiunto di X e Y su U e la loro relazione statistica. I fenomeni sono osservati congiuntamente su ciascuna delle N unità, quindi il risultato della rilevazioni sarà un insieme di N coppie (di tipo *x,y*). Per organizzare i dati grezzi utilizzeremo le tabelle a doppia entrata. Tale tabella è composta da righe e colonne e useremo l'indice i con riferimento al fenomeno X, che avrà k modalità; useremo l'indice j con riferimento al fenomeno Y, che avrà h modalità. Le modalità di X saranno x_i ; quelle di Y saranno y_j. L'interno della tabella si avrà contando le manifestazioni della medesima coppia. Ai margini si pongono le somme di colonna e di riga.
SCHEMA P. 108

Sulla tabella a doppia entrata si avranno sia informazioni di tipo bivariato (X e Y condizionati), sia informazioni di tipo monovariato (X e Y considerati singolarmente).
All'interno della tabella si trova la frequenza con cui si manifesta ciascuna coppia di modalità, all'incrocio tra la i-esima riga e la j-esima colonna. Queste frequenze (riguardanti entrambi i fenomeni) prendono il nome di frequenze congiunte (indicate con f_{ij}). L'interno della tabella costituisce la variabile statistica doppiache sta alla base della stat. Descrittiva bi-variata. $\sum_{i=1}^{k}\sum_{j=1}^{h} f_{ij} = \sum_{j=1}^{h}\sum_{i=1}^{k} f_{ij} = N$.
Per quanto riguarda i margini delle tabelle, troviamo frequenze che riguardano i fenomeni presi singolarmente -> frequenze marginali (informazione di tipo mono-variato). Le frequenze marginali si ottengono sommando le frequenze congiunte che stanno sulla stessa riga ($f_{i.}$) o sulla stessa colonna ($f_{.j}$). $\sum_{j=1}^{h} f_{ij} = f_{i.}$; $\sum_{i=1}^{k} f_{ij} = f_{.j}$.
SCHEMA P.110

Fissando l'attenzione sulle singole righe o colonne separatamente si costruiscono le v.s. condizionate $Y|x_i$ (Y dato x_i) e $X|y_j$ (X condizionato y_j).
Considerare le righe separatamente significa ridurre l'attenzione dell'intera U di N unità, alla sottopopolazione di f_i unità che manifestano la modalità x_i di X e, in questa sottopopolazione, si guarda il comportamento di Y. La v.s. condizionata $Y|x_i$ descrive il comportamento di Y sulle sole f_i, unità statistiche che sono omogenee rispetto a X perché manifestano la medesima modalità x_i, che chiameremo modalità condizionante. Stesso discorso va fatto per le colonne.
Dalle v.s. condizionate possiamo arrivare alle frequenze condizionate che vengono chiamate percentuali di riga e percentuali di colonna.

$$Frequenze\ condizionate\ di\ Y|x_i = \frac{f_{ij}}{f_{i.}}\ (\times 100\ danno\ le\ perc.\ di\ riga)$$

$$Frequenze\ condizionate\ di\ X|y_i = \frac{f_{ij}}{f_{.j}} (\times 100\ danno\ le\ perc.\ di\ colonna)$$

Il fenomeno condizionante è anche chiamato variabile esplicativa, il fenomeno condizionato variabile rispost.

INDIPENDENZA, CONNESSIONE E ASSOCIAZIONE

I fenomeni quantitativa hanno una strumentazione statistica più ampia di quelli qualitativi. Gli strumenti di questo capitolo si possono applicare ad entrambi i fenomeni ma sono più consigliati per quelli qualitativi.
Se tra X e Y non esiste alcuna relazione statistica, parleremo di indipendenza statistica. Il metodo per stabilire se X e Y sono indipendenti consiste nel confrontare le frequenze condizionate con le frequenze marginali. Il

confronto è possibile solo tra frequenze relativa. Le f. condizionate sono già relativa, mentre quelle marginali si ottengono dividendo quelle assolute per N ($f_{i.}/_N$ per X, e $f_{.j}/_N$ per Y). Se tutte le k serie di frequenze condizionate sono uguali tra loro e uguali alle marginali (relative): X e Y sono indipendenti e quindi non esiste indipendenza statistica (i.s.) . $Condizione\ di\ i.s.: \ f_{ij}/_{fi.} = f_{.j}/_N \ per\ tutti\ gli\ indici\ i = 1..k\ e\ j = 1..h.$
Facendo un semplice passaggio algebrico sulla condizione di i.s. (moltiplicare entrambi i membri per $f_{i.}$) si ottengono le f. congiunte che realizzano la condizione di i.s. e le chiameremo <u>frequenze teoriche (o attese) di i.s.</u> e le indicheremo con un *. $f_{ij}^* = f_{i.}f_{.j}/_N$. Queste renderebbero vera l'i.s.. Quando le tabelle (osservata e teorica) coincidono si avrà indipendenza statistica (metodo alternativo per verificarla). Il concetto di indipendenza statistica è simmetrico: tra X e Y esiste i.s.: X è indipendente, Y è indipendente.
Se non è verificata l'i.s. allora ci sarà <u>connessione</u> tra i due fenomeni. Il passo successivo sarà capire se la connessione (relazione) tra X e Y è forte o debole. L'intensità è tanto più elevata, quanto la tabella osservata è lontana dalla tabella teorica. Il metodo più utilizzato per guardare questa lontananza consiste nella differenza tra valore osservato e valore teorico: $f_{ij} - f_{ij}^*$. Quando non sono nulle (c'è connessione) possono essere vicine e lontane dallo 0. Le differenze possono essere positive e negative, per consentirci di misurare la connessione dobbiamo togliere il segno elevando al quadrato e capiremo quanto sono grandi le differenze.
La misura di connessione sarà la *chi* greca χ. $Indice\ di\ connessione\ \chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{h} (f_{ij} - f_{ij}^*)^2/f_{ij}^*$.

Se ci fosse i.s. $\chi^2 = 0$. $Formula\ alternativa \rightarrow \ \chi^2 = N\left(\sum_{i=1}^{k}\sum_{j=1}^{h} f_{ij}^2/f_{i.}f_{.j} - 1\right)$.

<u>Normalizzazione</u>-> procedimento con cui si trasforma un indicatore statistico assoluto in una percentuale. Chiamiamo I una generica misura statistica. Se di I conosciamo il suo valore *minimo* (che chiameremo I_{min}, che sarà il valore che assumerebbe la misura in assenza di ciò che stiamo misurando di X) e il valore *massimo* (I_{max}; cioè il valore che assumerebbe nel caso che X presenti al livello massimo ciò che stiamo misurando) possiamo trasformare l'indicatore assoluto in percentuale, normalizzandolo con la formula:

$$Normalizzazione\ di\ un\ indice\ I = \frac{I - I_{min}}{I_{max} - I_{min}}.$$

Visto che il valore assoluto dell'indice di connessione χ^2 non consente una valutazione, dobbiamo normalizzarlo. Il valore minimo (di χ^2è lo 0) si normalizza rapportando al suo valore massimo. Il valore massimo del χ^2 è il valore che l'indice assumerebbe in caso di massima connessione tra i due fenomeni, cioè in caso di una relazione statistica perfetta.

$$Indice\ di\ connessione\ normalizzato\ \frac{\chi^2}{N \times min\{k-1, h-1\}}$$

Valore massimo del χ^2 : è il valore pari a N moltiplicato per il più piccolo tra il numero delle righe (k) e il numero delle colonne (h), meno 1. In formule → $N \times min\{k-1, h-1\}$.
<u>DIMOSTRAZIONE DI QUESTA FORMULA:</u>

DIPENDENZA E CORRELAZIONE

Quando almeno uno dei due fenomeni è quantitativo possiamo aumentare il livello di analisi introducendo relazioni e strumenti statistici più raffinati. Utilizzando sia le frequenze sia le modalità dei fenomeni è possibile dare un verso alla relazione (stabilire se e quanto X influenza Y, o viceversa). Quando poi entrambi i fenomeni sono quantitativi è possibile esplorare ancora di più la natura e la tipologia della relazione (con strumenti grafici e indicatori sintetici).

Quando almeno uno dei due fenomeni è quantitativo possiamo sintetizzare la tabella a doppia entrata con le medie e le varianze. Per Y (o X) continuo utilizziamo il valore centrale dell'intervallo.

Media marginale di Y. E' *la media della v.s. marginale di* Y.

$$\bar{y} = \frac{1}{N}\sum_{j=1}^{h} y_j f_{.j}$$

Varianza marginale di Y. E' *la varianza della v.s. marginale di* Y.

$$\sigma_Y^2 = \frac{1}{N}\sum_{j=1}^{h}(y_j - \bar{y})^2 f_{.j} = \frac{1}{N}\sum_{j=1}^{h} y_j^2 f_{.j} - \bar{y}^2$$

Medie e varianze marginali sono ponderate con le frequenze marginali.

Media condizionata di Y dato x_i. E' *la media della v.s. condizionata* $Y|x_i$ *che si legge sulla* $i-$ *esima riga della tabella.*

$$\bar{y}|x_i = \frac{1}{f_{i.}}\sum_{j=1}^{h} y_i f_{ij} \quad (l'indice\ i \text{ è } fisso)$$

Varianza condizionata di Y dato x_i. E' *la varianza della v.s. condizionata* $Y|x_i$ *che si legge sulla* $i-$ *esima riga della tabella.*

$$\sigma^2_{Y|x_i} = \frac{1}{f_{i.}}\sum_{j=1}^{h}(y_j - \bar{y}_{x_i})^2 f_{ij} \quad (l'indice\ i \text{ è } fisso).$$

La proprietà più importante che riguarda la media e la varianza è la proprietà associativa. Di medie condizionate ne abbiamo k e ciascuna si riferisce a una sotto-popolazione di numerosità $f_{i.}$. Si possono sintetizzare a loro volta in una media. La media (aritmetica) delle medie condizionate, ponderata con le numerosità delle sotto-popolazioni coincide con la media marginale. $\frac{1}{N}\sum_{i=1}^{k}\bar{y}|x_i f_{i.} = \bar{y}$

Se entrambi i fenomeni sono quantitativi è possibile trattare matematicamente l'intera v.s. bivariata utilizzando le coppie modalità (x_i,y_j) di entrambi i fenomeni oltre alle frequenze congiunte f_{ij}. Sulla v.s. bivariata si definiscono una sorta di media bivariata chiamata momento misto (mu: "μ") e una sorta di misura di variabilità congiunta: la covarianza (sigma: "σ"). $\mu_{XY} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{h} x_i y_j f_{ij}$

$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{h}(x_i - \bar{x})(y_j - \bar{y}) f_{ij}$ La covarianza può risultare positiva, negativa o nulla. Quando entrambi i fenomeni sono quantitativi è possibile fare un'ulteriore passo: possiamo studiare la natura della relazione statistica, dandole una formulazione matematica e rappresentandola in un grafico.

Il diagramma a dispersione è lo strumento utile per visualizzare il tipo di relazione esistente tra due fenomeni X e Y quantitativi. È un diagramma cartesiano con gli assi intestati alle modalità dei due fenomeni. E ci fa vedere se tra X e Y esiste una relazione statistica (se la nuvola di punti si presenta strutturata) oppure no. La struttura con cui i punti si presentano ci da indicazioni circa il tipo di relazione statistica esistente (cioè la sua formulazione matematica).

SCHEMI PAGINA 150

Parliamo di serie doppia quando le xi e le fi sono tutte diverse e non abbiamo le medie condizionate. È più facile da disegnare. Esempio p. 149

Interpretazione geometrica della covarianza

Cominciamo a rappresentare sul diagramma anche le medie marginali che appaiono nella formula di σ_{XY}(divido il diagramma in 4 parti in base alle medie); la covarianza è basata sugli scarti presi con il loro segno. A seconda che le modalità siano sopra o sotto la media, questi scarti sono positivi o negativi e cioè corrispondono ad una particolare dispersione sul piano cartesiano; σ_{XY} è basata sui prodotti, allora le quattro zone evidenziate sul diagramma contribuiscono al calcolo di sigma.

SCHEMA PAGINA 152

La relazione statistica di tipo lineare (tra X e Y quantitativi) è chiamata correlazione lineare o semplicemente correlazione. Quando la covarianza è positiva allora X e Y sono positivamente correlati, se è negativa sono negativamente correlati, se è nulla (=0) allora X e Y sono incorrelati (non esiste una relazione di tipo lineare). Una volta capito che X e Y sono correlati dobbiamo capire il grado di correlazione tramite il coefficiente di correlazione lineare (indicato con la lettera greca rho "ρ") $\rho_{XY} = {}^{\sigma_{XY}}/_{\sigma_X \sigma_Y}$.

Può assumere valori che vanno da -1 a +1; quando è uguale a zero, X e Y sono incorrelati. $\rho = \pm 1$ sono perfettamente e negativamente/positivamente correlati.

GRAFICO PERFETTA CORRELAZIONE POSITIVA E NEGATIVA

REGRESSIONE

Dopo aver imparato a stabilire l'esistenza di una relazione statistica in una coppia di fenomeni, a misurare l'intensità, analizzare il verso e studiarne la natura, ora ci occuperemo di modellarla.

Per modello si intende una rappresentazione semplificata (e approssimata della realtà). Per modello statistico una interpretazione matematica della relazione tra X e Y nella tabella osservata, quindi consiste in una formulazione matematica che ne coglie l'andamento di fondo, semplificandolo.

Il più noto strumento statistico per la costruzione di un modello è la regressione. Un modello di regressione interpreta la dipendenza di Y da X: è una formula da applicare a X per approssimare Y.$\hat{Y} = f(X)$. Dove f denota una qualunque funzione di X e il simbolo sopra la Y indica che stiamo approssimando la realtà osservata con una curva matematica semplice e regolare. Avremo un modello per prevedere e simulare un certo fenomeno, si tratterà di un modello statistico basato su dati osservati presso le N unità che compongono la U di riferimento, lo chiameremo modello di regressione(curva teorica, tramite equazione; utilizzeremo i dati osservati per costruire il modello teorico per sostituire la spezzata di regressione: i punti

del diagramma osservati, uniti con dei segmenti, prendono il nome di spezzata di regressione che è una curva empirica, cioè è basata sui dati osservati ed è quindi irregolare e spigolosa, con il modello di regressione dovremmo trovare una curva liscia e regolare che approssima questa spezzata di regressione).
Il fenomeno condizionato Y ha il ruolo di variabile dipendente (ed è anche chiamato variabile risposta), il fenomeno X ha il ruolo di variabile indipendente (chiamato anche variabile esplicativa o regressore).
Il modello di regressione adatto per interpretare la correlazione (cioè la relazione lineare tra X e Y) è la retta di regressione (o modello di regressione lineare).
$\hat{Y} = a + bX$; a e b sono detti parametri della retta. La a è l'intercetta (cioè il punto in cui la retta interseca l'asse verticale delle ordinate); la b è il coefficiente angolare (determina l'inclinazione della retta e la sua pendenza: più è elevato più è ripida; meno è elevato, più la retta sarà piatta; negativo -> retta decrescente).
Fare la regressione lineare significa utilizzare dati per assegnare un *valore* ai parametri *a* e *b* della retta. Il metodo più utilizzato, che approssima al meglio la spezzata di regressione, è quello dei minimi quadrati (mq). Esso consiste nell'assegnare ai parametri dei valori che randano minima la distanza tra dati osservati e la retta di regressione.

Condizione dei minimi quadrati → $\sum_{i=1}^{k}\sum_{j=1}^{h}(y_j - \hat{y}_i)^2 f_{ij} = \min_{a,b}$. Dove $\hat{y} = a + bx_i$

1.Valori reali osservati → y_j
2.Modello → retta di regressione $\hat{Y} = a + bX$
5.Distanza totale tra i dati reali e valori teorici

3.Valori teorici approssimati mediante il modello → $y_i = \widehat{a + bx_i}$

4.Distanza tra i dati reali e i valori teorici → è la differenza tra yj e y cappuccio; va elevata al quadrato per eliminare l'influenza del segno e ponderata con f_{ij}

Soluzione dei minimi quadrati → $b = \frac{\sigma_{XY}}{\sigma_X^2}$ e $a = \bar{y} - b\bar{x}$.

Se *b* maggiore di 0: retta dei m.q. crescente; viceversa: decrescente. Il valore *a* ci dice quanto vale $\hat{Y}$ quando X=0, mentre il valore di *b* ci dice di quanto varia $\hat{Y}$ quando X aumenta di 1. In altre parole, se prendiamo due valori per X che distano di 1, i corrispondenti valori di $\hat{Y}$, secondo il modello, differiscono di *b*.
Una volta trovata l'equazione della retta basta trovare due punti sulla retta (genericamente sostituendo x con 0 e y con 0: $(0, a)$ e $(-{}^{a}/_{b}, 0)$).
Ora dobbiamo capire quanto è affidabile questo modello (cioè quanto si adatta alla realtà). Allora dobbiamo valutare la bontà della regressione misurando l'adattamento (accostamento) della retta dei m.q. ai dati osservati reali. Dobbiamo accertarci che la distanza sia piccola o magari nulla attraverso l'analisi dei residui.
La distanza totale tra valori reali e la retta ci dà il residuo totale della retta, chiamato anche devianza residua. $DR = \sum_{i=1}^{k}\sum_{j=1}^{h}(y_j - \hat{y}_i)^2 f_{ij}$. Il residuo della retta dei m.q. è nullo (DR=0) quando sono tutte nulle le distanze tra valori osservati e quelli teorici, quindi la retta si adatta perfettamente ai dati reali.
Per capire se la distanza sia tanta o poca (modello buono o cattivo) dobbiamo misurare la bontà di adattamento dei m.q. normalizzando il residuo.
Devianza totale → $DT = N\sigma_Y^2 = \sum_{j=1}^{k}(y_j - \hat{y}_i)f_{.j}$. La DT si scompone di due parti: *devianza residua* e *devianza spiegata (DT=DR+DS)* . *Devianza spiegata* → $\sum_{i=1}^{h}(\bar{y}_j - \hat{y}_i)^2 f_{i.}$. DS parte catturata dalla retta dei m.q.; DR parte residua, non catturata.
FORMULE ALTERNATIVE PIU' SEMPLICI → → → $DR = DT\,(1 - \rho_{XY}^2)$ e $DS = DT \times \rho_{XY}^2$.
Bontà di adattamento : $\frac{DS}{DT} = \frac{DT \times \rho_{XY}^2}{DT} = \rho_{XY}^2$.
Se $\rho_{XY}^2 = 0$: se cioè DS=0 e DR=DT la retta lascia tutto residuo e non spiega niente (X e Y sono incorrelati).

Se $\rho_{XY}^2 = 1$: se cioè DR=0 e DS=DT la retta non lascia alcun residuo e spiega perfettamente la variabilità di Y (sono perfettamente correlati).
I valori di ρ_{XY}^2 intermedi tra 0 e 1 sono interpretabili come percentuali di variabilità di Y spiegata della retta dei m.q.
Se la bontà di adattamento risulta bassa o comunque non la riteniamo sufficiente ai fini della nostra ricerca possiamo proseguire in due modi:

1. *Cambiare il modello*: la retta è modello di grado 1 (lineare). Se non ha un buon adattamento si può provare con un modello di grado 2 (parabola) o di grado superiore. Si parla in questo caso di regressione polinomiale. Oppure si può provare a cambiare totalmente il modello passando dalla retta ad un modello complicato → regressione non lineare.
2. *Aumentare i dati osservati da utilizzare nella regressione*: si tratta di osservare congiuntamente 2 o più fenomeni e di porre Y in funzione di tutti i fenomeni osservati che vengono chiamati *variabili esplicative o covariate.* Con questa seconda strada passiamo all'*analisi statistica multivariata* e alla *regressione multipla*.

DALLA DESCRIZIONE ALL'INFERENZA

Quando abbiamo dei dati parziali (cioè relativi ad un sotto insieme, che chiameremo campione di numerosità *n*) e vogliamo estendere l'analisi del comportamento di X all'intera popolazione di U, parliamo di inferire dal campione all'intera popolazione.
L'osservazione esaustiva della popolazione U (con tutti i dati) prende il nome di *censimento*; se abbiamo dati solo relativi ad un campione avremo una *rilevazione campionario*. Ragioni perché è più frequente la rilevazione campionaria: ragioni di budget (richiede risorsi ridotte rispetto a un censimento) e ragioni di precisione (consente maggiore cura, precisione e profondità dell'indagine perché non c'è un numero elevato).
SCHEMA P. 191.

Il termine inferenza indica il generico passaggio dalla premessa alla conclusione. Un caso speciale è l'inferenza induttiva che procede dal particolare al generale. L'inferenza statistica è una inferenza induttiva che procede dal campione alla popolazione. Quindi il campione ha il carattere della rappresentatività e della causalità (è un campione scelto in modo casuale).

CASO, PROBABILITA' E VARIABILI CASUALI

Lo strumento formale per fare inferenza statistica è la variabile casuale (v.c.).
Cominciamo con il considerare la dicotomia tra situazione deterministica e situazione casuale. Rappresentiamo con il tradizionale disegno l'insieme di circostanza che sono causa (determinano) un certo risultato. Chiamo il risultato evento e lo indichiamo con E. Ci si trova in situazione deterministica quando è noto l'intero insieme di circostanze che determinano E. E è quindi prevedibile a priori con certezza. Situazione casuale viceversa: l'insieme è noto parzialmente.
SCHEMA P. 195

La parte di circostanze ignote che impediscono di prevedere a priori con certezza il risultato E definisce il caso
Esperimento casule: esperimento condotto sotto l'effetto del caso (nota solo una parte delle circostanze).
Evento elementare: ciascuno dei possibili esiti di un esperimento casuale.
Spazio campionario: insieme di tutti gli esiti di un esperimento casuale (omega: Ω).
Evento casuale: sottoinsieme di Ω. Di solito è quello che ci attendiamo o che vogliamo individuare (indicato con E. Gli elementi di E sono eventi elementari; un evento elementare che è contenuto in omega può appartenere o non appartenere ad E; ma non viceversa.

La probabilità di un evento casuale E è un numero associato a E che ne quantifica a priori il grado di incertezza ovvero la possibilità di realizzazione. Ci limitiamo a dare due definizioni di probabilità:

1. Definizione classica: P(E) è il rapporto tra il numero di *casi favorevoli* a E e il numero di *tutti i casi possibili*, posto che tutti siano ugualmente possibili. $\frac{casi\ favorevoli}{casi\ possibili}$ solo con equiprobabilità.
2. Definizione frequentistica o statistica: Questa definizione si basa sulla cosiddetta *legge empirica del caso*, cioè una regola che non si può dimostrare matematicamente ma che si osserva sistematicamente nella pratica. L'evento E di cui si vuole calcolare la probabilità P(E) è pensato come il risultato di un esperimento casuale ripetibile un gran numero N di volte sempre nelle stesse condizioni. Al termine di tali N prove, E si sarà verificato *f* volte (e non si sarà verificato le rimanente N-*f* volte). La legge empirica del caso dice che la frequenza relativa *f*/N del verificarsi di E tende a stabilizzarsi intorno a un certo valore man mano che aumenta il numero N di ripetizioni

dell'esperimento (sempre nelle stesse condizioni). La definizione frequentistica di probabilità si basa su questa legge empirica e stabilisce che la probabilità di E è proprio quel valore, intorno al quale tede a stabilizzarsi la frequenza relativa dopo un numero sufficientemente grande di prove. In formule: $P(E) = \lim_{N\to\infty} \frac{f}{N}$.

La definizione frequentistica ci permette di considerare spazi campionari virtualmente infiniti e di calcolare la probabilità di eventi che non sono tutti ugualmente possibili; però, la ripetibilità delle prove deve effettuarsi tutta nelle stesse condizioni.

Possiamo pensare la variabile casuale come lo strumento matematico che permette di concentrarsi sulle sole caratteristiche dell'esperimento che interessano e che trasforma gli eventi casuali in numeri reali, conservandone comunque la probabilità. Variabile casuale: è una funzione con dominio nello spazio campionario Ω e codominio nell'insieme dei numeri reali, a cui rimangono associate le probabilità degli eventi di Ω. $X: \Omega \to R$.
SCHEMA DEL QUADERNO

La somma delle probabilità di tutti i valori x della v.c. X è pari ad 1, in perfetta analogia con la somma delle frequenze relative per una v.s. La probabilità associate costituiscono la *funzione di probabilità*.
V.c. discreta X. V.c. che assume un numero finito (o infinito numerabile) di valori x che di solito sono numero interi.
Funzione di probabilità di X. È associata a una v.c. discreta, ne descrive completamente le probabilità e ha sempre somma 1. In formule: $P(X = x)\ con\ \sum_x P(X = x) = 1$
Media o valore atteso. È definita e calcolata come per la v.s. ma usando le probabilità al posto delle frequenze. Il simbolo per indicare la media di una v.c. X è standard e fa riferimento all'inglese *Expectation*. Formula: $E(X) = \sum_x x \times P(X = x)$. E(X) si legge "E di X" ed è la media della v.c. X.
Varianza. È definita e calcolata come per la v.s. ma usando la probabilità al posto delle frequenze. È una misura della variabilità di X, cioè della dispersione dei suoi valori intorno al suo valore atteso, ponderata con le probabilità. In formule: $V(X) = \sum_x [x - E(X)] \times P(X = x)$. Si legge "V di X" → "varianza di X.
Deviazione standard. La chiameremo *standard deviation* per non confonderla con quella della v.s. La varianza è elevata al quadrato; quando serve ripristinare l'ordine di grandezza basta mettere una radice quadrata e si ottiene la deviazione standard, useremo il simbolo SD. $SD(X) = \sqrt{V(X)}$.
Per fare inferenza statistica si usano alcune v.c. speciali. Una di queste è la v.c. binomiale.

La variabile casuale binomiale è una particolare v.c. discreta. Serve per modellare situazioni casuali che hanno 3 caratteristiche: - l'esperimento casuale consiste nell'esecuzione di n prove indipendenti (l'esito di ciascuna prova non influisce sull'esito della successiva); -ciascuna prova può avere come esito uno (e soltanto uno) di due eventi tra loro contrari ed esaustivi (che chiameremo *successo* e *insuccesso*, in base a quello che vogliamo osservare); -in ciascuna prova, la probabilità del successo, che denoteremo con *p*, è nota ed è costante, Poiché *p* è una probabilità, è un numero compreso tra 0 e 1 e conseguentemente è nota anche la probabilità dell'insuccesso: $P(successo) = p \quad 0 < p < 1 \quad P(insuccesso) = 1 - p$
Per indicare la v.c. binomiale useremo la notazione $X \sim Bin(n, p)$ che si legge "X è una v.c. binomiale con parametri n e p". Il numero di prove *n* e la probabilità di successo *p* sono infatti dei parametri.
Ora possiamo solo immaginare una generica struttura dei nostri eventi elementari. Ciascuna prova può avere un Successo o Insuccesso e di prove ne facciamo *n*. Allora il generico risultato della serie di *n* prove (cioè il generico evento elementare) è una n-upla.

SCHEMA P. 205

La variabile casuale binomiale concentra l'attenzione sul numero di successi delle *n* prove indipendenti. I suoi possibili valori allora sono numeri interi da 0 a *n*.
In formule: $X \sim Bin(n,p) \quad con \quad n > 0 \; intero, 0 < p < 1 \; e \; x = 0, 1, 2, \dots, n$
Consideriamo una n-upla che contiene x successi e (n-x) insuccessi nell'ordine più semplice: prima tutti S e dopo tutti I. Ciascun S si verifica con probabilità *p*, gli I con probabilità (1-*p*). SCHEMA P. 205

La probabilità di questa n-upla è quindi: $p^x(1-p)^{n-x}$. Per contare il numero di possibili combinazioni di x successi e (n-x) insuccessi in ordine diverso si usa il coefficiente binomiale $\binom{n}{x}$ che si legge "*n* su *x*". $\binom{n}{x} = \frac{n!}{x!(n-x)!}$. Non ci resta che mettere insieme tutti i pezzi e così abbiamo la funzione di probabilità.
Funzione di probabilità: $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad con \; x = 0, 1, \dots, n$.
La media informa sul numero *atteso* di successi nelle *n* prove. La varianza e la deviazione standard misurano la *dispersione* del numero di successi *intorno* al valore medio atteso. In particolare la deviazione standard ci dice di quanto, in media su tutte le possibili n-uple, il numero di successi si discosta dal numero medio atteso.

Media di X~Bin(n, p): $E(X) = n \times p$
Varianza di X~Bin(n, p): $V(X) = n \times p(1-p)$
Deviazione standard di X~Bin(n, p): $SD(X) = \sqrt{V(X)} = \sqrt{np(1-p)}$

Per fare inferenza statistica su fenomeni statistici continui servono le v.c. continue.
Le v.c. continue assumono infiniti valori. Nel continuo occorre fare riferimento a insiemi di valori, cioè intervalli. I singoli punti perdono significato e la probabilità è calcolabile solo per gli intervalli.
Le v.c. continue non hanno la funzione di probabilità P(X=x). Hanno però la funzione di densità indicata con la lettera greca fi φ. La funzione di densità serve per calcolare la probabilità di intervalli di valori di una v.c. X continua. Nel continuo le probabilità sono aree. L'area sottesa al grafico della funzione di densità φ(x) (si legge di x) in un intervallo è la probabilità che X assuma valori in quell'intervallo.
Per calcolare l'area sotto una φ(x) curvilinea e calcolare la probabilità di intervalli diventa una particolare operazione che richiederebbe gli integrali ma noi usiamo dei casi più semplici.

Una speciale v.c. continua è la **v.c. normale.** È la più nota tra le v.c. continue e la più usata. *Normalmente* si presta bene a interpretare molti fenomeni continui, inoltre, quando si fa inferenza statistica e qualunque sia il campo di applicazione, *normalmente* è lì che va a parare. Useremo la notazione $X \sim N(\mu, \sigma^2)$ che si legge "X è una v.c. normale di parametri mu e sigma quadro": mu rappresenta la media e siga quadro la varianza della normale. La Normale e la sua funzione di densità hanno 10 proprietà:

1. Assume tutti i possibili valori $-\infty < x < +\infty$.
2. Essendo continua ha la funzione di densità (non la funzione di probabilità) ed ha una forma campanulare (una campana centrata su mu)
3. L'area sottesa alla φ(x) fornisce l'intera probabilità = 1
4. Il parametro μ è la media di $X \sim N(\mu, \sigma^2)$. In formule E(X)=μ
5. Il parametro σ^2 è la varianza di $X \sim N(\mu, \sigma^2)$. In formule V(X)= σ^2

6. La curva campanulare ha la proprietà di essere simmetrica rispetto a mu: $P(X) \geq \mu \ e \ P(X) \leq \mu$
7. Mu rappresenta anche la media di X e la sua moda. $x_{0.5} = x_0 = \bar{x} = \mu$.
8. Si dice che la campana è formata da una pancia e due code con flessi (punti sull'asse delle ascisse dove la curva cambia concavità) pari a $\mu \pm \sigma$.
9. Mu stabilisce la posizione della curva (e se si sposta, la curva si sposta, ma non cambia la forma). Sigma stabilisce la forma della curva (se è più piccolo la curva sarà stretta e alta, se è più grande sarà larga e bassa).
10. Se prendiamo una parte(a,b) della curva, la probabilità che X assume un valore tra a e b coincide con l'aria sottesa a quell'intervallo.

DISEGNO CHE C'E' SUL QUADERNO

La Normale tende a manifestarsi con un valore sistematico prevalente (μ); i valori più probabili saranno vicini a tale valore; i valori lontani da mu sono rari e poco probabili.

Standardizzare una quantità statistica significa operare una trasformazione con lo scopo di depurarla da unità di misura e grandezza, rendendola confrontabile con altri dati standardizzati perché tutti riferibili a un'unica situazione standard.

$Xstandard = \frac{X-E(X)}{\sqrt{V(X)}}$: togliamo la sua media e dividiamo per la sua deviazione standard.

In questo modo avremo media nulla (=0) e varianza pari a 1 (e quindi anche deviazione standard). Cioè diventa riferibile a un'unica situazione. E(X standard)=0 e V(X standard)=SD(X standard)=1

Standardizzando una v.c. normale $X \sim N(\mu, \sigma^2)$ s, con la sua media μ e la sua deviazione standard $\sqrt{\sigma^2} = \sigma$ si ottiene la **v.c. normale standardizzata**, che indicheremo con Z. Useremo la notazione $Z \sim N(0,1)$.

$Z = \frac{X-\mu}{\sqrt{\sigma^2}} = \frac{X-\mu}{\sigma}$. E(Z)=0 e V(Z)=SD(Z)=1

Dal valore ottenuto controlliamo le tavole (le quali però segnano solo la parte a sinistra di quel numero P(Z≤z) , quindi servirà fare delle operazioni aritmetiche semplici).

CAMPIONAMENTO ED ERRORE CAMPIONARIO

Ora abbiamo gli strumenti per introdurre, comprendere e usare gli strumenti della statistica inferenziale. Il primo passo consiste nel procurarci i dati. In ambito inferenziale questo significa procurarci il campione che è un sottoinsieme dell'intera popolazione U su cui ci interessa studiare il fenomeno. L'inferenza statistica si basa su *campioni casuali* (l'operazione di scelta casuale si chiama campionamento). Il numero *n* è la numerosità o ampiezza campionaria e di solito è prefissato e molto più piccolo di N.

L'insieme dei metodi di campionamento prende il nome di teoria dei campioni. Gli elementi che vedremo noi sono basati sul metodo più semplice: campione bernoulliano. Un campione bernoulliano è il risultato di n estrazioni casuali da U condotte tutte nelle stesse condizioni, cioè indipendenti tra loro. Si tratta di estrazioni con reinserimento e equiprobabili.

Solitamente, però avvengono senza reinserimento (per evitare di estrarre la stessa persona) e si parla in questo caso di campione casuale semplice o anche SRS(simple random sample.

Se n è sufficientemente grande e allo stesso tempo n è piccolo rispetto a N, il che è in genere ciò che accade, le due tecniche con o senza risultato portano i risultati equivalenti. Frazione di campionamento n/N sufficientemente piccola.
Ciascuno dei differenti campioni estraibili da U può darci un'immagine più o meno fede di U perché fornisce un'informazione parziale e differente circa il comportamento su U che ci interessa. Questo è il concetto di variabilità campionaria. Il processo di inferenza statistica avviene sotto effetto della variabilità campionaria, la conseguenza è quella che comporta necessariamente incertezza e rischio di errore (chiameremo questo concetto errore di campionamento).
Quando si dispone solo di dati campionati (parziali e casuali) la distribuzione del fenomeno di interessi su U e i reali valori delle sue sintesi statistiche sono ignoti e li chiameremo parametri. I parametri ignoti sono l'oggetto dell'inferenza statistica. Le sintesi statistiche di X rappresentano i corrispondenti parametri ignoti di U. In particolare E(X)=media del fenomeno in U, la indicheremo con μ e la V(X)=varianza del fenomeno in U con σ^2.
Ciascuna *osservazione campionaria* X_i è il risultato di un esperimento casuale; è pertanto un evento casuale e può coincide con uno dei possibili valori della v.c. X. Allora, anche il risultato di ogni estrazione campionario è interpretato da una v.c. X_i che chiameremo v.c. estrazione campionaria. Poiché nel campione bernoulliano le estrazioni sono indipendente, allora le v.c. estrazioni campionarie X_i sono tra loro indipendenti. Infine, poiché X_i può coincidere con qualunque dei possibili valori del fenomeno, a sua volta interpretato dalla v.c. Z, si ha anche che ciascuna estrazione campionaria X_i è identica a X, e in quanto identica, ha stessa media e stessa varianza.
La statistica inferenziale offre metodologie per risolvere due grandi classi di problemi di inferenza:

1. la *stima* dei parametri, con l'obbiettivo di usare i dati campionari per inferire il valore dei parametri ignoti;
2. la *verifica di ipotesi statistiche*, con l'obbiettivo di usare i dati campionari per inferire se è accettabile o meno un valore che si ipotizza per i parametri ignoti.

STIME E STIMATORI

Ora impariamo a stimare i parametri ignoti. Per farlo esistono due classi di metodi: stima *puntuale* (con un unico valore) e stima *intervallare* (con un intervallo di valori). Qui l'errore campionario assume l'aspetto di *errore di stima* (quanto è più piccolo, più precisa e affidabile è la stima).

La stima puntuale è la metodologia statistica che utilizza le informazioni campionarie per: calcolare un (unico) valore puntuale per sostituirlo all'ignoto parametro; controllare in termini di probabilità se e quanto la sostituzione è affidabile e accurata.
Iniziamo a stimare 3 parametri: la *media* del fenomeno in U (che corrisponde alla media μ di X); la *varianza* del fenomeno in U (che corrisponde alla varianza σ^2 di X); una *percentuale* di valori di X di interesse, che indicheremo con *p* e che vedremo a cosa corrisponde. Li calcoleremo per *analogia*.
Stabilire se una stima è affidabile e sufficientemente precisa significa controllare e misurare l'errore campionario in termini di probabilità. La stima di un parametro è il risultato di un calcolo, un'elaborazione eseguita sugli n dati x1…xi…xn, per ottenere un unico numero da sostituire all'intero parametro in U (che è e rimane ignoto). Per controllare l'errore di stima dobbiamo tenere conto di tutti i possibili risultati ottenibili da tutti i possibili campioni. Per fare questo affianchiamo al concetto di stima il concetto di stimatore. Lo stimatore è la stessa funzione (formula) che definisce la stima, ma applicata alla v.c. estrazioni campionarie X1…Xi…Xn.
Lo stimatore è quindi una v.c. che interpreta tutti i possibili valori della stima su tutti i possibili campioni estraibili. Quindi, la stima è un numero, ottenuto sul campione effettivamente estratto e l'unico a disposizione; lo stimatore è una v.c. che tiene conto di tutte le possibili stime ottenibili su tutti i possibili

campioni estraibili. Questo serve per interpretare la variabilità campionaria e per controllare l'errore campionario.

Per stimare l'ignota media μ di U usiamo la media (aritmetica) degli n dati campionari. Chiameremo questa stima *media campionaria* e la indicheremo con $\bar{x}$.

MEDIA CAMPIONARIA (STIMA): $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Il corrispondente stimatore è la seguente v.c. STIMATORE: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Studiando le caratteristiche dello stimatore (cioè media, varianza e funzione di probabilità o densità) si definiscono le proprietà statistiche di uno stimatore. Le vedremo considerando lo stimatore media campionaria.

La più nota proprietà è la *non distorsione*: uno stimatore è *non distorto* se il suo valore atteso coincide con il parametro oggetto di stima (qualunque esso sia). Se questo non succede, lo stimatore è distorto.
Tra tutti i possibili campioni ve ne sono alcuni che forniscono sotto-stime del parametro, altri sovra-stime del parametro e altri valori vicini o identici al parametro oggetto di stima. Richiedere che uno stimatore non sia distorto significa garantire che sovra-stime e sotto-stime si compensino sul totale dei campioni estraibili e che in media lo stimatore coincide con ciò che si vuole stimare.Quando è verificata questa proprietà si parlerà di stima non distorta a garanzia dell'affidabilità dell'inferenza.

La media campionaria è stima per l'ignota media μ in U. Il corrispondente stimatore è *non distorto* per mu perché il suo valore atteso è proprio uguale a mu.

NON DISTORSIONE MEDIA CAMP. $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Dimostrazione

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n \times \mu = \mu$$

Se lo stimatore è distorto il suo valore atteso non coincide con il parametro da stimare. Questo succede quando sotto e sovra stime non si compensano. Una stima non distorta è affidabile perché è uno dei possibili valori di uno stimatore che in media coincide con ciò che si vuole stimare.
Uno stimatore può essere non distorto ma può anche essere sempre lontano da ciò che vogliamo stimare e dunque non è un buon stimatore. Per capire se è un buon stimatore o meno abbiamo bisogno dell'*errore quadratico medio* (misura quanto è preciso e quanto è vicino lo stimatore al parametro ignoto).

Quello che vogliamo fare ora è esprimere in formule l'errore quadratico medio. Un punto di partenza è la differenza tra $(\bar{x} - \mu)$; visto che parliamo dello stimatore consideriamo X e lo eleviamo al quadrato per eliminare l'influenza del segno e consideriamo l'errore medio di stima, mediano a tutti i campioni estraibili. $MSE = E(\bar{X} - \mu)^2$.
Quindi l'errore quadratico medio di uno stimatore è il valore atteso della differenza al quadrato tra lo stimatore e il parametro che si vuole stimare. Misura la dispersione dei valori dello stimatore. Più piccola, più preciso. L'MSE corrisponde alla varianza più la distorsione al quadrato. MSE=V+Dist^2
Visto che la media campionaria non è distorta, l'errore quadratico medio della media campionaria corrisponde alla sua varianza. Vediamo allora:

$MSE(\bar{X}) = E(\bar{X} - \mu)^2 = V(\bar{X})$ dove $V(\bar{X}) = \frac{\sigma^2}{n}$

$DIMOSTRAZIONE$:

$$V(\bar{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} V(X_i) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \sigma^2 = \left(\frac{1}{n}\right)^2 n\sigma^2 = \frac{\sigma^2}{n}$$

L'errore quadratico medio è in relazione diretta con sigma quadro e in relazione inversa con n. L'errore della media campionaria, quindi, è tanto minore quanto più grande è il campione.

Un'altra proprietà auspicabile per uno stimatore è la consistenza che riguarda la precisione (o accuratezza). A un buon stimatore si richiede che sia sempre più preciso, riducendo l'errore di stima, all'aumentare dell'ampiezza campionaria n, quando cioè aumentano i dati introdotti nel processo di stima.
Se lo stimatore non è distorto come la media campionaria, per essere consistente basta che la sua varianza diventi sempre più piccola al crescere dell'ampiezza campionaria n.

Anche la proprietà di efficienza relativa riguarda la precisione di uno stimatore. È un criterio di scelta quando si dispone di due (o più) stimatori per lo stesso parametro ignoto.
Se si tratta di stimatori non distorti, MSE coincide con la varianza e dunque il confronto avviene tra le varianze; lo stimatore non distorto con varianza inferiore è il più efficiente tra quelli a disposizione.
Un teorema (difficile da dimostrare) stabilisce che lo stimatore media campionaria sia il più efficienti tra tutti i possibili stimatori non distorti per mu.
Abbiamo analizzato a media campionaria, adesso impariamo altri parametri.

Il parametro ignoto da studiare ora è la varianza del fenomeno nella popolazione. La stima più naturale per la varianza di U è la varianza del campione $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Questa volta però non si può usare questa formula perché lo stimatore è distorto per sigma quadro e tenderà a sotto-stimare. Per ottenere uno stimatore *non distorto* allora dobbiamo dividere per n-1: varianza campionaria corretta (s)
$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. La quantità (n-1) è chiamata *gradi di libertà*. La varianza non è distorta, è consistente (perché l'errore s2 diminuisce al crescere dell'ampiezza campionaria). Non avremo la deviazione standard perché sarebbe distorta per sigma.
L'MSE è quadratico (cioè misura l'errore di stima prendendo le differenze tra stimatore e parametro elevate al quadrato) e questo produce effetti collaterali e per ristabilire l'ordine di grandezza allora dobbiamo prendere la $\sqrt{MSE}$ che è una misura teorica dell'errore medio di stima. La stima dell'errore medio di stima, calcolata con gli stessi dati campionari è detta standard error (SE).

$$\text{SE dello stimatore} = \sqrt{MSE} \text{ e se lo stimatore non è distorto: } \sqrt{V}$$

SE della media campionaria. Poiché è uno stimatore non distorto, si tratta di stimare la radice della varianza della media campionario (quindi la radice di sigma quadro fratto l'ampiezza campionaria), stimando con sigma quadro la varianza campionaria corretta. Quindi : $SE(\bar{X}) = \sqrt{\frac{s^2}{n}}$.
SE è un numero calcolato sul campione che stima l'errore medio che si commette sostituendo all'ignoto parametro la stima calcolata sul medesimo campione.

Nella ricerca sociale interessano i fenomeni *dicotomici*. L'oggetto della stima, in questo caso, è la *percentuale* di unità statistiche o casi, che tra tutte quelle che compongono la U, è classificabile in una determinata categoria. Facciamo allora riferimento ai fenomeni categoriali.
Scelta l'ampiezza campionaria *n*, si estrae da U un campione bernoulliano, il risultato sarà l'insieme di unità classificabili o non nella categoria che ci interessa. La stima più naturale per l'ignota frequenza relativa *p* di soggetti classificabili nella categoria di interesse, è corrispondente alla frequenza relativa nel campione, cioè la frequenza relativa campionaria che indicheremo con $\hat{p}$ (pi cappuccio).
L'affidabilità di questa stima risiede nelle proprietà statistiche del corrispondente stimatore $\hat{P}$. Questo assume valore 1 in corrispondenza dei soggetti classificabili nella categoria che ci interessa, 0 in quelli non classificabili. Allora il campione sarà un insieme di 0 e 1 .

La somma dei dati campionari ci dà il numero di soggetti campionati, che tra gli *n* estratti, sono classificabili nella categoria che ci interessa: $Stima\ della\ percentuale\ \hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Visto le caratteristiche (campione bernoulliano, prove indipendenti, S e I) avremo una variabile causale binomiale con parametri n e p per lo stimatore: $Stimatore$: $\hat{P} = \frac{Bin\ (n,p)}{n}$

Allora si determinano velocemente il valore atteso, varianza e standard error dello stimatore P cappuccio.

Non distorsione della freq. relativa (percentuale)campionaria:

$$E(\hat{P}) = E\left(\frac{Bin\ (n,p)}{n}\right) = \frac{1}{n}E[Bin\ (n,p)] = \frac{np}{n} = p$$

Allora il suo MSE coincide con la varianza

$$MSE(\hat{P}) = V(\hat{P}) = V\left[\frac{Bin\ (n,p)}{n}\right] = \left(\frac{1}{n}\right)^2 V[in(n,p)] = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

Standard Error

$$SE(\hat{P}) = \sqrt{V(\hat{P})} = \sqrt{\frac{p(1-p)}{n}}$$

INTERVALLI DI CONFIDENZA

La stima puntuale è un metodo sempre applicabile (è sempre calcolabile a partire dai soli dati campionari), è semplice (perché si procede per analogia), però: è difficile avvicinarsi ed azzeccare il parametro ignoto l'affidabilità della stima puntuale risiede nella garanzia probabilistica offerta dalle proprietà teorico-formali del corrispondente stimatore.

La stima intervallare, a garanzia della sua affidabilità, offre un numero associabile che misura la proprietà con cui il corrispondente stimatore contiene effettivamente l'ignoto parametro. L'errore di campionamento lo possiamo fissare noi.

Intervallo di confidenza → per un ignoto parametro. È un intervallo di valori calcolato sui dati campionari ,per il quale si può confidare, a un prescelto livello probabilistico, che contenga l'ignoto valore del parametro.

A *favore* della stima intervallare (IC): meno rischioso (è più facile, attraverso un intervallo, avvicinarsi al parametro ignoto); è più informativo (ma meno preciso, offre informazione più ampia di un unico valore); è più affidabile(quantificabile con una probabilità scelta a priori).

Contro: ha un elevato livello di complessità e servono elle informazioni ausiliari a priori .

Un IC non è sempre calcolabile sulla base dei soli dati campionari, ma è calcolabile solo se ci si trova o in una o nell'altra delle seguenti situazioni:

1. è noto (o ipotizzabile con un elevato grado di sicurezza) che il fenomeno X in U è ben interpretato da una v.c. Normale. Questa situazione la chiameremo popolazione normale
2. La numerosità del campione *n* è sufficientemente grande perché valgono opportuni teoremi di teoria delle probabilità Chiameremo questa situazione grandi campioni.

POPOLAZIONE NORMALE

- SIGMA QUADRO NOTO

Ipotizziamo di avere un fenomeno, ben interpretato con una v.c. Normale con media μ ignota ma varianza σ^2 nota. In formule: $X \sim N(\mu, \sigma^2\ \text{nota})$. Un teorema della probabilità ci garantisce che se X è normale anche lo stimatore media campionaria $\bar{X}$ lo è. Quindi se la varianza è nota avremo: $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\ \text{nota}\right)$.

Questa è una informazione ausiliaria.

La metodologia di costruzione di un IC prevede 5 passi:

1. Si estrae un campione bernoulliano di ampiezza n e ci si procurano i dati campionari.
2. Si calcola la stima puntuale per mu, cioè la media del campione: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

3. Si sceglie la probabilità di sbagliare, cioè di costruire una IC che non contiene mu. Indicheremo questa probabilità con alfa α. Allora la probabilità di fare bene sarà (1-α). L'alfa più utilizzano corrisponde a 0.05 oppure 0.1 oppure 0.01; (1-α) sarà: 95%, 90% o 99% .
4. Siccome abbiamo l'informazione ausiliare a priori: $\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} = Z \sim N(0,1)$ e noi sappiamo come calcolare la probabilità quando abbiamo un intervallo della Normale standardizzata, allora usiamo "al contrario" e con l'α scelto al punto precedente, poniamo:

   $P\left(a \leq \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \leq b = P(a \leq Z \leq b\right) = 1-\alpha$. Sappiamo che per la Normale le probabilità sono aree; all'interno dell'intervallo (a, b) c'è una probabilità pari a (1-α) mentre all'esterno una probabilità α che dividiamo in α/2 a sinistra e uguale a destra. Li indicheremo con $-z_{\alpha/2}$ e $z_{\alpha/2}$. Questi li chiameremo Z-score. Troviamo lo Z-score positivo sulle tavole della normale standardizzata Z-N(0,1).
   DISEGNO P. 258

   Infine si inverte questa relazione probabilistica in modo da avere un intervallo centrato su mu che si vuole stimare:

$$P\left(\bar{X} - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right) = 1-\alpha$$

   Nel caso di popolazione normale questa probabilità è vera. Sostituendo i dati campionari si otterrà un intervallo che è l'IC che cerchiamo. A tale intervallo rimane associato il numero (1-α) a garanzia probabilistica dell'affidabilità dell'IC costruito. Per questo 1-α è chiamato anche livello di confidenza (misura di quanto possiamo fidarci che l'IC contenga il valore ignoto del parametro).
5. Si sostituisce allo stimatore media campionaria all'interno delle parentesi tonde della probabilità scritta sopra, il valore della stima xmedio calcolato sull'unico campione estratto.

IC per μ a l.c. $(1-\alpha)$ con popolazione normale e sigma quadro nota

$$\left[\bar{x} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} \;,\quad \bar{x} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}\right]$$

ESEMPIO SUL QUADERNO….

- SIGMA QUADRO E VALORE ATTESO (MEDIA) IGNOTI

Siamo ancora nella condizione iniziale di na v.c. normale. Ma in questo caso non abbiamo nemmeno la varianza. Come si procede?

La stima per σ^2 è la varianza campionaria corretta con i gradi di libertà: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

In questo casi visto che non possiamo standardizzare (assenza di sigma quadro), studentizziamo:

$$Studentizzazione\ di\ \bar{X} \rightarrow \frac{\bar{X}-\mu}{\sqrt{s^2/n}}$$

Se utilizziamo le lettere maiuscole, ovviamente indicheremo gli indicatori. Un altro teorema della probabilità ci assicura che lo stimatore media campionaria non è più una $Z \sim (0,1)$ ma è un'altra v.c. chiamata T di student

La *T di Student* ha sempre forma campanulare ed è centrata sullo zero, ma ha varianza più grande rispetto ad 1. Ha le cose più pesanti (cioè sono più lontane dall'asse delle ascisse). La v.c. di Student ha un solo

parametro, detto *grado di libertà* (n-1). $\frac{\bar{X}-\mu}{\sqrt{s^2/n}} = T_{n-1}$. Le tavole della T di student identificano la parte che sta a destra (a differenza delle tavole della normale standardizzata).
DISEGNO T DI STUDENT

➔ Ora vediamo davvero come si agisce quando abbiamo sia media che sigma quadro ignoto
Anche in questo caso ci sono 5 passi da fare:

1. Si estrare un campione bernoulliano di ampiezza n e ci si procura i dati campionari
2. Si calcolano le stime puntuali; in questo caso per ENTRAMBI i parametri (media campionaria e varianza campionaria corretta; mu e s^2)
3. Si sceglie il livello di confidenza (1-α) da cui si ottiene la probabilità di sbagliare α e la probabilità α/2 delle code
4. Si studentizza lo stimatore media campionaria e si ottiene la v.c. T di Student con (n-) gradi di libertà. Quello che cambia è che salteranno fuori dei T-score. Quindi avremo:

$$P\left(-t_{\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\sqrt{\frac{s^2}{n}}} \leq t_{\frac{\alpha}{2}}\right) = 1-\alpha$$

Il T-score positivo si trova sulle tavole della T di student, il suo simmetrico si ottiene cambiando semplicemente il segno. Ora invertiamo la disuguaglianza all'interno delle parendesi e riscriviamo la probabilità centrata su mu:

$$P\left(\bar{X} - t_{\alpha/2}\sqrt{\frac{s^2}{n}} \leq \mu \leq \bar{X} + t_{\alpha}\sqrt{\frac{s^2}{n}}\right) = 1-\alpha$$

5. Sostituendo i dati campionari si ottiene finalmente l'IC che cerchiamo:

IC per μ a l.c. $(1-\alpha)$ con popolazione normale e sigma quadro ignota

$$\left[\bar{x} - t_{\frac{\alpha}{2}}\sqrt{\frac{s^2}{n}} \;,\quad \bar{x} + t_{\frac{\alpha}{2}}\sqrt{\frac{s^2}{n}}\right]$$

Ora ci mettiamo nel caso in cui non si sa nulla circa il fenomeno in U: non si hanno informazioni a priori, e comunque si sa che la popolazione *non è normale*.
Siamo certamente in un caso della popolazione non normale quando vogliamo stimare una percentuale e siamo in presenza di un fenomeno dicotomico. In questo caso sappiamo che è adeguata la v.c. binomiale.
Se non abbiamo informazioni ausiliare a priori di X, dobbiamo allora avere molti dati, cioè essere nel caso di *grandi campioni*. Solo se il campione è sufficiente grande possiamo appellarci ad un teorema delle probabilità fondamentale nell'inferenza statistica → il Teorema Centrale del Limite (TCL). Qualunque sia la distribuzione del fenomeno X in U, se l'ampiezza campionaria *n* tende all'infinito allora gli stimatori standardizzati della media campionaria $\bar{X}$ (stimatore per μ) e frequenza campionaria $\hat{P}$ (stimatore per p) sono normali. Questo è il risultato teorico del TCL.
Quindi quando *n* è sufficientemente grande gli stimatori sono approssimativamente *Normali*. In formule:

$$\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \approx N(0,1) \quad \text{e} \quad \frac{\hat{P}-p}{\sqrt{p(1-p)/n}} \approx N(0,1)$$

Qui la normalità è approssimata e conseguentemente si tratterà di IC approssimati per grandi campioni con un effettivo l.c. approssimativamente pari all'(1-α) scelto.

IC approssimato per grandi campioni per μ a l. c. approssimativamente pari a $(1-\alpha)$

$$\left[\bar{x} - z_{\frac{\alpha}{2}}\sqrt{\frac{s^2}{n}};\ \bar{x} + z_{\frac{\alpha}{2}}\sqrt{\frac{s^2}{n}}\right]$$

IC approssimato per grandi campioni per p a l. c. approssimativamente pari a $(1-\alpha)$

$$\left[\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

GRANDI CAMPIONI

Un IC è tanto più preciso più è stretto, cioè meno è ampio. L'ampiezza di un IC ne definisce, quindi, la precisione. Un IC ha struttura generale di questo tipo: *stima(puntuale)±score*SE(stima)*

Allora l'ampiezza di un IC ha la seguente forma generale:

$$(stima + score{\times}SE) - (stima - score{\times}SE) = 2score{\times}SE$$

Ne deduciamo che la precisione della stima intervallare dipende dalla precisione della stima puntuale, a sua volta stimata mediante lo standard error SE: più piccolo è SE, meno ampio e più preciso è l'IC.

Minore è SE($\bar{x}$)=s^2/n, minore è l'ampiezza dell'IC e dunque maggiore la precisione e accuratezza.

Osserviamo, infine, che l'ampiezza dell'IC dipende dal livello di confidenza e dall'ampiezza campionaria.

Allora il l.c. e la precisione di un IC sono tra loro collegati e a loro volta sono collegati all'ampiezza campionaria.

La precisione di un IC è in relazione inversa con il livello di confidenza e in relazione diretta con la numerosità campionaria. Cioè: a parità di ampiezza campionaria, un aumento del livello di confidenza provoca una diminuzione della precisione e viceversa; a parità di livello di confidenza, un aumento della numerosità campionaria provoca un aumento della precisione e viceversa.

Ma quanto deve essere grande n?

La *pianificazione di* n è strategica per l'inferenza statistica. Introduciamo l'*errore assoluto di stima*: è il modo più semplice per misurare l'errore di stima senza alterare l'ordine di grandezza e unità di misura.

Partiamo dalla differenza tra stima e parametro ignoto presa tra valore assoluto per eliminare il segno.

Per esempio per la media μ e la media campionaria $\bar{x}$, l'errore assoluto è: $|\bar{X} - \mu|$

Quindi l'errore assoluto di stima sarà: $|stimatore - parametro|$

È possibile scegliere a priori (prima di estrarre il campione) sia l'errore massimo che tollereremo (Err) sia il livello di probabilità con cui vogliamo che questo accada.

Cominciamo con il caso della media. Ora vogliamo decidere quando deve essere grande il campione affinchè, usando la media del campione per stimare la media dell'intera popolazione, commettiamo un errore assoluto non superiore a un certo margine tollerato. Siamo in condizioni di incertezza a causa della parzialità e casualità dei dati campionari e cerchiamo allora di fare una buona stima con una buona prob.

Scegliamo allora: la probabilità (1-α) di fare bene (90, 95 o 99%) e scegliamo il nostro margine di errore massimo tollerato, che lo indicheremo con *Err*.

Ora, poniamo la probabilità di fare bene, cioè di commettere un errore assoluto di stima *non più grande* del livello *err* che siamo disposti a tollerare, pari a livello (1-α) prescelto. $(1-\alpha) = P(|\bar{X} - \mu| \leq Err)$

Questa probabilità si può riscrivere così: $1 - \alpha = P(-Err \leq \bar{X} - \mu \leq +Err)$

Poiché stiamo cercando *n* sufficientemente grande, possiamo standardizzare e usare la Z~N(0,1):

$$1-\alpha=\frac{-Err}{\sqrt{\sigma^2/n}}\leq\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}\leq\frac{+Err}{\sqrt{\sigma^2/n}}=\frac{-Err}{\sqrt{\sigma^2/n}}\leq Z\leq\frac{+Err}{\sqrt{\sigma^2/n}}$$

Abbiamo così ritrovato il solito intervallo di calori Z~N(0,1) di probabilità (1-α) con -Err/… e +Err/… che corrispondono ai soliti Z-score. Vale cioè l'uguaglianza: $\frac{Err}{\sqrt{\sigma^2/n}}=Z_{\alpha/2}$

Risolvendo questa uguaglianza nella nostra incognita finalmente otteniamo:

$$n=\frac{Z_{\alpha/2}^2\sigma}{Err^2}$$

Che è l'ampiezza campionaria che con probabilità 1-alfa garantisce un errore assoluto di stima non superiore al nostro margine di errore *Err.*

LA FORMULA SI PUO' UTILIZZARE SOLO SE SI DISPONE INFORMAZIONI AUSILIARIE A PRIORI SULLA VARIABILITA' DEL FENOMENO X NELLA POPOLAZIONE U DI INTERESSE.

Un caso speciale si ha quando il fenomeno di interesse è qualitativo dicotomico e il parametro di stima è la frequenza relativa *p* di soggetti che appartengono a una data categoria che la corrispondente frequenza del campione (p cappuccio) è una buona stima. Ora vogliamo decidere quanto deve essere grande il campione affinchè usando p cappuccio per stimare p commettiamo un errore assoluto non superiore a un certo livello massimo di errore tollerato *Err*.

In questo caso ci metteremo nel caso della situazione peggiore per stimare p (50% succ, 50% insucc).

Sostituendo nella formula

$$n=\frac{Z_{\alpha/2}^2\sigma}{Err^2}=\frac{Z_{\frac{\alpha}{2}}^2(0.5*0.5)}{Err^2}=\frac{Z_{\alpha/2}^2*0.25}{Err^2}=\frac{Z_{\alpha/2}^2}{4Err^2}$$

Questa è l'ampiezza campionaria che garantisce la massima tutela dell'errore di stima, perché assume la situazione peggiore.

TEST STATISTICI

Immaginiamo ora di lavorare in un contesto applicativo che ci permette di formulare un'*ipotesi* circa il valore dell'ignoto parametro in U o circa qualche aspetto statistico del fenomeno nella popolazione. I dati campionari sono allora impiegati per stabilire se tale ipotesi è ragionevolmente *accettabile* o *rifiutabile*.

Il più classico test statistico è il *test di significatività*.

Quando non si dispone di una osservazione completa su una U di riferimento , ma solo di dati parziali parliamo di *ipotesi statistica*: è una congettura riguardante una qualche caratteristica statistica del fenomeno in U. Tale congettura è formulata a priori, cioè prima di estrarre il campione. Proviene dall'esterno e dipende dal contesto applicativo e dagli obbiettivi di ricerca, non dai dati campionari.

L'*ipotesi nulla* è la formalizzazione, cioè traduzione in simboli e formule, dell'ipotesi statistica che abbiamo emesso e che vogliamo sottoporre a verifica con un test statistico. Indicheremo l'ipotesi nulla con la notazione standard H_0.

La verifica di ipotesi è la metodologia inferenziale che porta a decidere se accettare o rifiutare l'ipotesi nulla. Il *test statistico* è la regola pratica che porta a questa decisione e con la nostra strumentazione base ci concentriamo solo sul *test di significatività*.

Un test statistico (cioè la regola che porta ad accettare o rifiutare H0) è basata su dati campionari e quindi su un'osservazione parziale dell'intera U di riferimento. È dunque condotto in condizioni incerte.

Possiamo fare due errori: *errore di I specie*(l'errato rifiuto di H0 vera) e *errore di II specie*(errata accettazione di H0 falsa). Noi ci occuperemo solo di quello di I specie.

Scegliamo a priori la probabilità di commettere l'errore e questa la chiameremo α.

$\alpha = P(rifiutare\ H_0|H_0)$ ovvero la probabilità di rifiutare h dato h0. E 1-α allora sarà la probabilità di non sbagliare accettando h0 e questo si chiama *livello di significatività(l.s.)*. Anche in questo caso abbiamo bisogno di informazioni ausiliarie del tipo *popolazione normale* oppure *grandi campioni*.

Mettiamoci nel caso di popolazione normale: sappiamo che il fenomeno di interesse è ben interpretato da una v.c. Normale con media μ ignota, ma con l'informazione della varianza: $X \sim N(\mu, \sigma^2 nota)$ quindi sappiamo che lo stimatore media campionaria sarà normale anch'esso: $\bar{X} \sim N(\mu, \frac{\sigma^2}{n} nota)$. Adesso immaginiamo che le condizioni in cui stiamo lavorando ci consentano di emettere l'ipotesi statistica che il valore dell'ignoto parametro μ sia un certo numero. Il simbolo per indicarlo sarà μ_0 "mu con zero". Avremo quindi: $H_0: \mu = \mu_0$. La verifica del test statistico consiste in 6 passi.

1. Si estrae il campione bernoulliano di ampiezza *n* e ci si procurano i dati campionari
2. Si calcola la stima di ciò che è ignoto: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
3. SI sceglie la probabilità di sbagliare α, cioè di commettere l'errore di rifiuto di H_0. Allora la probabilità di fare bene (ovvero il l.s.) sarà (1-α).
4. Si standardizza assumendo che H_0 sia vera e quindi utilizziamo al posto dell'ignota μ, μ_0 ipotizzando H_0. Otteniamo in questo modo la *statistica test*

$$STATISTICA\ TEST \rightarrow \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} = Z \sim N(0,1)$$

Poiché la media campionaria è una buona stima per mu, osserviamo che:

- Se $H_0: \mu = \mu_0$ è *vera*, allora la differenza tra $\bar{X} - \mu_0$ tende a risultare più piccola (vicina a 0)
- Se $H_0: \mu = \mu_0$ è *falsa*, allora la differenza tende a risultare grande (lontana in più o meno)

Allora i valori della statistica test Z: intorno allo 0 depongono a favore dell'accettazione, più lontani il rifiuto. Quindi Z è divisa in due zone: zona di accettazione (zona di valori a favore dall'accettazione di H0 pari a 1-α) e regione critica (zona di valori che depongono per il rifiuto di H0 e sono i valori corrispondenti alle due code e quindi sono pari a α/2).
SCHEMA DEL DISEGNO A CAMPANA P.285

Valore critico del test. È il punto sull'asse delle ascisse che identifica la soglia tra zona di accettazione zona critica. È lo Z-score $Z_{\alpha/2}$ che ci garantisce la probabilità di sbagliare che abbiamo scelto.

$$P(rifiutare\ H_0|H_0) = P\left(\frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} \leq -Z_{\frac{\alpha}{2}}\ \ oppure\ \ \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} \geq +Z_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha$$

$$P(accettare\ H_0|H_0) = P\left(-Z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} \leq Z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

5. Sostituendo nella statistica test Z i valori noti a priori, cioè μ_0,σ^2 e *n*, e la stima X medio calcolata sui dati campionari, si ottiene un numero che chiamiamo *valore sperimentale*: $\frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}}$

6. Si rifiuta $H_0: \mu = \mu_0$ a livello $(1-\alpha)0$ se ill valore sperimentale cade in una delle regioni critiche.

$$\frac{\bar{X}-\mu_0}{\sqrt{\sigma^2/n}} \leq -Z_{\frac{\alpha}{2}} \ oppure \ \frac{\bar{X}-\mu_0}{\sqrt{\sigma^2/n}} \geq +Z_{\frac{\alpha}{2}}$$

E SE ABBIAMO ANCHE σ^2 IGNOTO?
I passi sono simili.
Dopo aver raccolto i dati del campione bernoulliano, nel secondo passo oltre che stimare la media del campione dobbiamo anche stimare la varianza del campione: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
Dopo scegliamo il livello di significatività e nel quarto passo invece studendizziamo.
La nostra *statistica test* sarà così: $\frac{\bar{X}-\mu_0}{\sqrt{S^2/n}} = T_{n-1}$. Quindi per ottenere il valore critico otterremo un T-score.
Il valore critico sarà così:

$$P(rifiutare\ H_0|H_0) = P\left(\frac{\bar{X}-\mu_0}{\sqrt{S^2/n}} \leq -t_{\frac{\alpha}{2}} \ oppure \ \frac{\bar{X}-\mu_0}{\sqrt{S^2/n}} \geq +t_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha$$

Nel quinto passo si avrà il nostro valore sperimentale: $\frac{\bar{X}-\mu_0}{\sqrt{S^2/n}}$ e finalmente (al sesto passo) faremo il test.

Fino a qui abbiamo verificato solo ipotesi del tipo $H_0: \mu = \mu_0$. Se il test porta all'accettazione di H_0 si conclude che μ è uguale al valore ipotizzato a livello di significatività 1-α. Se viceversa il test porta al rifiuto di H_0 si conclude che μ è diverso da μ_0 con probabilità di sbagliare pari a α. Chiameremo allora questo tipo di ipotesi *bilaterali* (dove avremo due zone sotto due code e quindi li chiamiamo anche *test a due code*).
Nella pratica però sono utili anche ipotesi *unilaterali*, cioè l'ipotesi del tipo $H_0: \mu \geq \mu_0$

Per verificare ipotesi nulle unilaterali si pone tutta la regione critica solo sotto un'unica coda della statistica test (quella più lontana all'ipotesi nulla) e quindi avremo un *test a una coda.*
Vediamo $H_0: \mu \leq \mu_0$
Oltre alle differenze $\bar{x} - \mu_0$ vicine allo zero, anche tutte le differenze negative depongono dell'*accettazione* di $H_0: \mu \leq \mu_0$. Tali differenze però positive e troppo grandi depongono invece il *rifiuto* di $H_0: \mu \leq \mu_0$.
Per verificare $H_0: \mu \leq \mu_0$ avremo un T-test a una coda con regione critica sotto la coda di destra.
Avremo una sola coda di probabilità α.
Vediamo $H_0: \mu \geq \mu_0$
Il ragionamento si ribalta. Le differenze positive vicino allo zero depongono l'accettazione; le differenze negative e troppo grandi depongono il rifiuto. In questo caso la regione critica sarà tutta a sinistra.

Pensiamo ora alle situazioni in cui non si dispongono di informazioni ausiliarie a priori. Tutte le situazioni non previste dal caso di popolazione Normale. È necessario compensare questa mancanza con campioni sufficientemente grandi (*grandi campioni*)
Solo se il campione è sufficientemente grande possiamo applicare il TCL e recuperare la normalità degli stimatori (per la media e per la frequenza relativa percentuale):

$$\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \approx N(0,1) \quad \text{e} \quad \frac{\hat{P}-p}{\sqrt{p(1-p)/n}} \approx N(0,1)$$

Siccome ci basiamo su risultati approssimati possiamo costruire il test in tutti i casi in cui non si ha normalità della popolazione, ma si tratterà di *approssimati per grandi campioni*.

Quando il fenomeno d'interesse è categoriale (qualitativo e dicotomico) il parametro dell'ignoto oggetto di inferenza è la frequenza relativa *p* (o percentuale *p*100*) di soggetti che in U sono classificabili nella categoria che ci preme e che chiamiamo convenzionalmente *successo*. L'ipotesi nulla sarà del tipo:

$$H_0: p = p_0 (bilaterale) \quad e \quad H_0: p \geq p_0 \; o \; H_0: p \leq p_0 \; (unilaterale)$$

Se *n* è sufficientemente grande, per verificare questo tipo di ipotesi si utilizzano i Z-test a due o una coda. Ripercorriamo i sei passi:

1. Ricaviamo i dati campionari
2. Si calcola la stima puntuale per *p*. Siccome p è frequenza relativa di unità statistiche che nella popolazione sono classificate nella categoria successo, allora la sua stima sarà la corrispondente frequenza $\hat{p}$ del campione
3. Si sceglie il l.s. del test (1-α) da cui si deriva la probabilità di rifiuto α (o α/2)
4. La statistica test si ottiene tramite la standardizzazione:

$$\frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \cong Z \sim N(0,1)$$

Il valore critico sarà allora uno Z-score da cercare sulle tavole della Z

VALORE CRITICO

***Per il test a una coda*:**

$$P(rifiutare\ H_0 | H_{o)} = P\left(\frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \geq Z_\alpha\right) = \alpha$$

***Per il test a due code*:**

$$= P\left(\frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \geq Z_{\alpha/2}\right) + P\left(\frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leq -Z_{\alpha/2}\right) = \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha$$

5. Il valore sperimentale si calcola sostituendo nella statistica test i valori noti e le stime campionarie
6. Per costruire il testi come regola di rifiuto ci ricordiamo che stiamo lavorando con un test approssimato per grandi campioni quindi con probabilità di sbagliare approssimativamente pari al prescelto α, se il valore sperimentale cade nella regione critica.

TEST

A UNA CODA SI RIFIUTA $H_0: p \leq p_0$ se:

$$\frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \geq Z_\alpha$$

A DUE CODE SI RIFIUTA $H_0: p = p_0$ se:

$$\frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \geq Z_{\frac{\alpha}{2}} \; oppure \; se \; \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leq -Z_{\alpha/2}$$

Di solito le analisi statistiche si fanno al computer e nella pratica i test statistici non si eseguono aa mano usando le tavole. Il computer esegue test producendo un unico numero con il quale possiamo decidere se accettare o rifiutare H0 qualunque sia il livello di significatività che voglia fissare. Tale valore si chiama *p-value*. Il p-value dunque è una probabilità, un numero compreso tra 0 e 1. Esso è il minimo livello α per rifiutare H_0 (data H_0 vera).

Se il p-value è più piccolo del livello prescelto α (per un test a una coda) o di α/2 (per un test a due code) allora si rifiuta H0.
Il computer fornisce il p-value in sostituzione del valore critico. Il valore critico dipende sempre dall'α prescelto ed è diverso per i diversi l.s. Quando si esegue il test al computer si decide se accettare o rifiutare H0 confrontando due probabilità: il p-value (fornito dal pc) e il livello α o α/2. Le due procedure sono equivalenti perché portano allo stesso risultato.