Domande E Risposte Esame Statistica

Statistica (Università degli Studi di Milano-Bicocca)

DESCRITTIVA MONOVARIATA

1- Definire e discutere i 4 elementi base: popolazione, fenomeno, modalità e numerosità

- POPOLAZIONE STATISTICA [U]. È il collettivo di unità statistiche su cui ci interessa studiare il fenomeno statistico. Insieme delle unità statistiche sulle quali interessa studiare il fenomeno.

- FENOMENO STATISTICO [X]. Sono molti e molto variegati. Sono di diversa natura: ci sono diversi strumenti statistici a seconda della diversa natura del fenomeno. Sono i fenomeni di interesse per la statistica. Si presentano con una molteplicità di manifestazioni, non con un'unica modalità. Nelle scienze sociali sono di particolare interesse i fenomeni che riguardano le popolazioni umane e lo società (il genere delle perone in un collettivo di interesse, reddito mensile, numero di esami, ma anche il peso delle buste di salame prodotte da una azienda in un certo tempo, etc)

- MODALITA' [x]. Può essere molte cose a seconda del fenomeno che stiamo studiando: attributo o categoria (talvolta ordinabile), numero o insieme di numeri (interi o reali a seconda che si conti o si misuri). Singola manifestazione del fenomeno indicato con la corrispondente lettera maiuscola.

- NUMEROSITA' DI U [N]. Numero intero positivo. Numero di unità statistiche che compongono la popolazione statistica. Su U di numerosità N sono presenti le manifestazioni x del fenomeno X. È possibile pensare fenomeni statistici su popolazioni infinite 8teroicamente), composte cioè da un numero virtualmente infinite di unità statistiche. Molte volte la dimensione della popolazione presa in esame è talmente elevata che ai fini dell'analisi statistica è conveniente pensarla infinita, anche se concretamente non lo è.

2- Dare una classificazione dei fenomeni statistici e un esempio per ciascun tipo

FENOMENI QUALITATIVI: si manifestano nella popolazione osservata attraverso attributi o categorie (qualità). Fenomeno che manifesta modalità che sono attributi o categorie (es: il genere,titolo di studio).

- Ordinali: pur essendo qualitativi si manifestano con attributi e categorie che si possono ordinare secondo un qualche criterio oggettivo e convenzionalmente accettato. Le sue modalità sono categorie che tutti ordiniamo allo stesso modo (es titolo di studio). Ordinamento oggettivo.
- Categoriali: fenomeni qualitativi per i quali non abbiamo un criterio oggettivo ma solo personale per ordinare le categorie con cui si manifesta. (es città di residenza). Ordinamento soggettivo.

FENOMENI QUANTITATIVI: si manifestano nella popolazione osservata attraverso numeri (quantità). Possono essere sempre ordinati perché fra i numeri esiste una relazione d'ordine naturale.

- Discreti: fenomeni quantitativi che possiamo contare (es esami registrati sul libretto). Fenomeno quantitativo dove x è numero intero
- Continui: fenomeni quantitativi che si possono misurare con un'opportuna unità di misura e con un corretto strumento di misurazione (es peso). Le sue manifestazioni sono intervalli. x dipende dall'unità di misura e dallo strumento.

3- Definire le scale di modalità (o di rilevazione) e darne una classificazione precisando il livello di analisi statistica consentito

Le scale di modalità sono gli strumenti tipici della rilevazione. La scala delle modalità con cui si rileva X è l'insieme di tutte le diverse manifestazioni di X osservabili su U, possono anche essere viste come il menù delle possibili risposte. La scelta delle scale è parte integrante del processo di costruzione di un buon questionari. Le scale devono rispettare i principi di esaustività, cioè devono prevedere tutte le possibili manifestazioni di X che potenzialmente si possono osservare su U e mutua esclusività, ovvero la scala con cui si effettua la rivelazione deve prevedere modalità che si escludono a vicenda senza possibilità di confusione e ambiguità

SCALE QUALITATIVE: le modalità sono attributi o categorie (qualità)

- Ordinali: i suoi attributi o categorie possono essere ordinati secondo un qualche criterio oggettivo o convenzionalmente accettato. Ordinamento oggettivo: ha un ordine naturale. (<,>) il livello di analisi è più profondo,perché fra le modalità oltre alla relazione uguale o diverso è istituibile anche la relazione minore , maggiore etc.
- Sconnesse: i suoi attributi o categorie non ammettono un ordinamento oggettivo ma solo casuale o personale (es scala dicotomica, 2 sole modalità, come fenomeno qualitativo genere M o F). Si possono istituire soltanto relazioni di uguaglianza o diversità

SCALE QUANTITATIVE: le modalità sono numeri (quantità). Si distinguono rispetto all'origine.

- Rapporto: l'origine è il numero 0 con significato assoluto, indica l'assenza del fenomeno. Scala numerica: posso utilizzare tutte le operazioni numeriche. (numero accessi a un sito internet, quantità grammi cioccolato presa). Permette tutte e 4 le operazioni elementari.
- Non rapporto: l'origine non è assoluta ma convenzionale, scelta secondo qualche criterio (es temperatura). Lo 0 non ha significato di origine. Non consentono l'operazione di divisione

Il tipo di scala determina le possibili relazioni istituibili tra le sue modalità.

I fenomeni quantitativi possono essere rilevati con scale qualitative e viceversa:

- Ad esempio il fenomeno reddito mensile è quantitativo, ma può essere rilevato con una scala qualitativa del tipo: alto, medio, basso. In questo modo otteniamo dati più attendibili, ma abbassiamo il livello di analisi statistica, passando da una scala quantitativa a una qualitativa ordinale. Il fenomeno deve essere ordinale

- Il fenomeno esito dell'esame è qualitativo perché dipende dal giudizio del professore ma si rileva con una scala quantitativa: da 18 a 30 trentesimi. Non va mai dimenticata la natura qualitativa del fenomeno, in quanto non è detto che uno studente che ha preso 30 è più preparato di uno studente che ha preso 25. La scala quantitativa è non rapporto.

- Infine vi sono fenomeni discreti che paiono continui, ad esempio il reddito: conto il denaro, ma ho bisogno di un'unità di misura ($,euro etc). Il consiglio è: quando il fenomeno quantitativo si presenta con un numero di modalità k molto elevato conviene trattarlo come se fosse continuo raggruppando le modalità in classi.

4- Definire la variabile statistica (vs) e le diverse distribuzioni di frequenza dandone l'interpretazione descrittiva

Una variabile statistica è l'insieme di k coppie del tipo "modalità, frequenza" tale che la somma delle frequenze (assolute) riproduce la numerosità di U. Insieme di coppie di elementi: il primo elemento è la manifestazione del fenomeno, il secondo è la frequenza con la quale si presenta nella popolazione. La **frequenza assoluta** è il numero di unità statistiche che fra le N osservate manifesta quella modalità xi di X e la indichiamo con fi . L'insieme delle k frequenze (assolute) è detta distribuzione di frequenze assolute di X su U. Le frequenze assolute sono sempre numeri interi e con somma pari a N. Non sono confrontabili fra popolazioni di numerosità diversa: sono direttamente influenzate da N e non sono nemmeno valutabili. Per fare dei confronti servono indicatori relativi, per fare delle valutazioni servono degli indicatori normalizzati. Quindi per confrontare popolazioni con numerosità diversa occorre costruire le **frequenze relative** che sono il rapporto fra la frequenza assoluta di xi e la numerosità N di U (pi), solo quantità adimensionali e perciò confrontabili. La numerosità della popolazione è la grandezza che disturba e impedisce il confronto, quindi va messa al denominatore. Sono comprese tra 0 e 1 e la loro sommo è pari a 1 (vedi dimostrazione quaderno)
Frequenze percentuali: sono le frequenze relative moltiplicate per 100. Sono sempre comprese tra 1 e 100 e la loro somma è pari a 100.
A livello di analisi statistica è preferibile lavorare con le pi mentre a livello di interpretazione e comunicazione dei risultati è conveniente passare alla fr percentuali, più comprensibili.

5- Definire le frequenze cumulate (assolute e relative); commentarne l'interpretazione e discuterne la relazione biunivoca con le frequenze assolute e relative

Solo per X almeno ordinale. Si possono cumulare le frequenze associate alle modalità inferiori di xi costruendo le frequenze cumulate (sia relative che assolute). La prima frequenza cumulata coincide con la frequenza della modalità più piccola, l'ultima frequenza cumulata assoluta coincide con la numerosità di N di U

Le **frequenze cumulate assolute** sono numeri compresi tra 0 e N. l'utima frequenza cumulata coincide con la numerosità N di U. le indichiamo con Fi. F1 = f1. Fk=N
Le **frequenze cumulate relative** permettono il confronto tra popolazioni di numerosità diversa. L'ultima frequenza cumulata relativa coincide con 1. Le indichiamo con phi maiuscolo. PHI1=p1. PHIk=1

Fra le frequenze assolute e relative e le corrispondenti frequenze cumulate esiste una corrispondenza biunivoca: data una distribuzione è possibile passare all'altra e viceversa.
Se conosciamo le frequenze (assolute o relative) possiamo ottenere le cumulate (sommando) e viceversa (sottraendo).

6- Con riferimento ad una variabile statistica con modalità intervallari, discutere comparativamente l'assunto iniziale di "valore centrale" e di "distribuzione uniforme" delle frequenze all'interno degli intervalli.

Fenomeni quantitativi continui: le modalità sono intervalli.
La v.s. ci informa che al generico intervallo xi: xl I- xL appartengono fi unità statistiche, non sappiamo in

quale degli infiniti punti che appartengono all'intervallo si posiziona ciascuna delle fi unità statistiche che cadono in xi: etc.
La distribuzione di frequenze all'interno degli intervalli è ignota: sappiamo che appartengono all'intervallo ma non sappiamo esattamente dove si trovano.
Per superare questo ostacolo si ricorre all'emissione di ipotesi, che siano ragionevoli sostenibili e convincenti.
Due ipotesi:

- **Ipotesi del valore centrale**: si assegna a ciascuna delle fi unità statistiche che cadono nell'intervallo xi: xl I- xL un unico punto interno all'intervallo stesso. Il metodo consiste nell'associare tutte le fi al valore centrale dell'intervallo, il quale è la semisomma dei due estremi. Lo indicheremo con x*i: xl+xL/2. In questo modo si supera il problema dell'ignota distribuzione di frequenze all'interno degli intervalli , ma si perde la natura continua. Questa ipotesi è sempre accettabile e apprezzabile, aumenta la semplicità. Utilizzo un diagramma a bastoncini.

- **ipotesi di uniforme distribuzione**: meno drastica della precedente, conserva e rappresenta la continuità. Distribuisce in maniera uniforme ed equidistante i soggetti all'interno dell'intervallo, in questo modo si da la stessa importanza a tutti i punti. Utilizzo un istogramma

7- Con riferimento ad una variabile statistica con modalità intervallari, definire e discutere la densità di frequenza e l'istogramma

La densità di frequenza di un intervallo è la frequenza dell'intervallo depurata dall'influenza dell'ampiezza, è una distribuzione per fenomeni continui (vedi formula pag 34 e 35).
Tanto più un intervallo è ampio tanto più è facile che contenga più casi di un intervallo meno ampio.
in ogni caso le densità di frequenza (le indichiamo con phi minuscolo) sono numeri reali e sempre positive , non hanno limite superiore. Danno un'idea di addensamento delle frequenze all'interno degli intervalli; a parità di frequenze, un intervallo ampio è meno denso di un intervallo più stretto.
Ci servono per individuare intervallo con densità maggiore.
In questo caso quindi abbiamo fenomeni quantitativi continui, e la **migliore rappresentazione grafica è l'istogramma**.
La distribuzione di frequenza all'interno degli intervalli è ignota; adottiamo quindi l'ipotesi di uniforme distribuzione la quale mantiene il carattere continuo del fenomeno, associando la frequenza a tutti gli infiniti punti dell'intervallo in modo che sia uniformemente distribuita.
In un istogramma area dei singoli rettangoli che si creano rappresenta la frequenza (pi).
Sappiamo che area rettangoli è base x altezza: nel nostro caso l'altezza corrisponde alla densità di frequenza relativa, mentre la base è l'ampiezza dell'intervallo.
Per rappresentare la distribuzione di **frequenze assolute**, sotto l'ipotesi di una distribuzione uniforme, si pongono gli intervalli sulle ascisse e la densità di frequenza sulle ordinate.
Se si vogliono rappresentare le **frequenze relative** , si pongono le densità di frequenze relative sulle ordinate.
L'area totale sottesa all'istogramma è pari a N se si rappresentano le frequenze assolute, mentre è pari a 1 se si rappresentano le relative .

8- Discutere e interpretare la distribuzione di Frequenza Cumulate per un fenomeno quantitativo continuo anche in relazione all'istogramma.

Su un istogramma sono automaticamente rappresentate come aree anche le frequenze cumulate assolute o relative, a seconda che le aree dei rettangoli rappresentino fi o pi.
L'istogramma permette il calcolo delle frequenze cumulate per qualunque valore del fenomeno continuo X; il calcolo avviene sotto l'ipotesi di uniforme distribuzione.

Sulle ascisse mettiamo gli intervalli, mentre sulle ordinate mettiamo le densità di frequenza

9- Definire la Moda e discuterne il calcolo e l'informazione descrittiva nel caso di fenomeni qualitativi e quantitativi. Discutere gli adattamenti necessari per il caso di modalità intervallari (o in classi).

La moda di una variabile statistica è la modalità a cui è associata la frequenza più elevata fra le k osservate, cioè la modalità più osservata.
Modalità più frequente in U. la indichiamo con x_o la moda è tanto più informativa tanto più elevata la frequenza corrispondente.
Quindi la moda è un valore medio di sintesi calcolabile per qualunque X ed è immediatamente individuabile.
Quando la v.s è sottoforma di tabella basta individuare la frequenza più elevata, quando è rappresentata graficamente la frequenza più elevata si individua a occhio, in quanto è la barra più alta in un diagramma a barre , spicchio più grande in un diagramma a torta.
Se X è continuo bisogna fare egli accorgimenti.
Se gli intervalli sono di ampiezza differente la frequenza(assoluta o relativa) perde la sua carica informativa per individuazione moda.
È necessario usare la **densità di frequenza**: chiamiamo intervallo modale quello a cui è associata la densità più elevata fra le k osservate.
È convenzione diffusa far coincidere la moda con il valore centrale dell'intervallo. In generale la moda non è un buon valore medio.

10- Definire e interpretare la mediana, in particolare discuterne il calcolo nel caso di modalità intervallari

La mediana **($x_{0.5}$)** è calcolabile solo X è almeno ordinale (fenomeni qualitativi ordinali oppure quantitativi).
La mediana di X è la modalità che occupa la posizione centrale nell'ordinamento.
Divide U: il 50% manifesta modalità xi<=x0.5, l'altro 50% modalità xi>=x0.5.
La mediana divide U in due gruppi ugualmente numerosi: in un gruppo stanno le unità che manifestano le modalità non superiori e nell'altro gruppo quelle che manifestano le modalità non inferiori.
Con X quantitativo continuo le modalità sono intervalli.
Si parla quindi di intervallo mediano.
Per calcolare la mediana (o l'intervallo mediano) bisogna scorrere la colonna delle frequenze cumulate relative e appena si trova (o si supera) lo 0.5, lì troviamo la mediana (o intervallo mediano). Nel caso di X quantitativo continuo, la mediana si trova all'interno dell'intervallo mediano.
Il problema quando abbiamo a che fare con intervalli è che la distribuzione all'interno di essi è ignota. Si possono utilizzare ipotesi valore entrale o quella di distribuzione uniforme.
La seconda come sappiamo è quella meno drastica; sotto questa ipotesi la mediana si identifica attraverso una formula (vedi pag 53). xl è estremo inferiore dell'intervallo mediano, Fi-1 oppure PHIi-1 è la frequenza dell'intervallo precedente, xL- xl è ampiezza intervallo mediano, fi oppure pi è la frequenza dell'intervallo mediano. Per determinare la mediana bisogna aggiungere a xl il pezzetto che manca per arrivare alla mediana stessa, il quale coincide con la base di un sotto-rettangolo.
Nell'istogramma le aree sono le frequenze, tutta l'area sotto istogramma vale N; la mediana divide quest'area in due parti uguali N/2.
L'area a sinistra di xl coincide con la frequenza di tutte le modalità <= xl, cioè la frequenza cumulata Fi-1.
Quindi la'era del sotto rettangolo è N/2 – Fi-1.

Otteniamo la base dividendo ultima formula per la densità (vedi pag 54). Vale lo stesso con le densità di frequenza relativa.

11- Discutere criticamente la media aritmetica quale sintesi di una variabile statistica quantitativa evidenziandone pregi e difetti

La media si può calcolare per fenomeni quantitativi o qualitativi ordinali rilevati con scala quantitativa.
Questo valore medio di sintesi ci permetto di manipolare algebricamente l'intera v.s. la indichiamo con x segnato.
È espressa nella stessa unità di misura con cui X si manifesta su U e ci da un info sintetica dell'ordine di grandezza di X su U.
Per calcolarla si moltiplica ciascuna delle k modalità osservate xi per il numero di volte in cui sono state osservate in U (cioè la loro frequenza fi) sommare tutto e infine dividere per N .
se utilizziamo le fr relative non dividiamo per N, in quanto le fr relative stesse sono già divise per la numerosità della popolazione. si può chiamare media ponderata, in quanto le modalità xi sono ponderate con le frequenze ed è divisa per la somma dei pesi della ponderazione (N).
ciascuna manifestazione del fenomeno è pesata con un certo peso (fr assolute o relative).
Nel caso di X continuo la media è calcolata utilizzando ipotesi del valore centrale.
Bisogna puntualizzare che la media è molto sensibile agli outlier, cioè valori anomali, molto diversi dagli altri, creano uno sbilanciamento e la media ne risente molto. La media aritmetica è il valore medio di sintesi più utilizzato perché gode di molte proprietà utili anche se non sempre da sola è informativa. La media non coglie la variabilità: non tutte le risorse sono equidistribuite.
La media dà un unico valore medio che bilancia la distribuzione tramite la proprietà dell'annullamento degli scarti la quale conferisce alla media il ruolo di baricentro della v.s e rappresenta una sintesi della tendenza centrale..
La media coglie solo un aspetto parziale del fenomeno e non basta: va associata ad un indicatore che coglie la variabilità.
A suo favore gode si 2 proprietà, che solo lei ha, cioè: mantenimento del totale e annullamento degli scarti. La prima è molto importante per fenomeni come il reddito medio.

12- Definire moda, mediana e media aritmetica, discuterne comparativamente il potenziale informativo e la scelta

- **MODA:** la moda di una v.s. è la modalità a cui è associata la frequenza più elevata tra le k osservate, cioè la modalità più osservata. È un valore medio calcolabile per X qualunque. Quando la v.s. è data sotto forma di tabella, basta scorrere la colonna delle frequenze (assolute o relative) e individuare la più elevata. Se la v.s. è rappresentata graficamente, la frequenza più elevata può essere individuata "a occhio". Talvolta la v.s. è priva di moda o è difficile individuarla, quindi non è un buon valore medio.

- **MEDIANA:** la mediana è calcolabile solo se X è almeno ordinale (fenomeni qualitativi ordinali oppure quantitativi). La mediana di X è la modalità che occupa la posizione centrale nell'ordinamento. Divide U: il 50% manifesta modalità xi<=x0.5, l'altro 50% modalità xi>=x0.5. La mediana divide U in due gruppi ugualmente numerosi: in un gruppo stanno le unità che manifestano le modalità non superiori e nell'altro gruppo quelle che manifestano le modalità non inferiori.

- **MEDIA:** La media si può calcolare per fenomeni quantitativi o qualitativi ordinali rilevati con scala quantitativa. Questo valore medio di sintesi ci permetto di manipolare algebricamente l'intera v.s. la indichiamo con x segnato. È espressa nella stessa unità di misura con cui X si

manifesta su U e ci da un info sintetica dell'ordine di grandezza di X su U. Per calcolarla si moltiplica ciascuna delle k modalità osservate xi per il numero di volte in cui sono state osservate in U (cioè la loro frequenza fi) sommare tutto e infine dividere per N . se utilizziamo le fr relative non dividiamo per N, in quanto le fr relative stesse sono già divise per la numerosità della popolazione. si può chiamare media ponderata, in quanto le modalità xi sono ponderate con le frequenze ed è divisa per la somma dei pesi della ponderazione (N). ciascuna manifestazione del fenomeno è pesata con un certo peso (fr assolute o relative). È molto sensibile agli outlier, cioè valori che sbilanciano l'interpretazione del fenomeno, sono valori "assurdi", molto diversi dagli altri. La media aritmetica è il valore medio di sintesi più utilizzato perché gode di molte proprietà utili anche se non sempre da sola è informativa. La media non coglie la variabilità: non tutte le risorse sono equidistribuite. La media dà un unico valore medio che bilancia la distribuzione. La media coglie solo un aspetto parziale del fenomeno e non basta: va associata ad un indicatore che coglie la variabilità.

In generale: Su un fenomeno quantitativo continuo posso calcolare moda, mediana e media.
Su un fenomeno qualitativo rilevato con scala sconnessa posso calcolare solo la media. Su un fenomeno ordinale posso calcolare solo la moda e la mediana.
Quando la v.s. è varia e complessa non basta un solo valore medio di sintesi, ma conviene costruire più valori. I valori medi non sono in alternativa tra loro ma sono complementari.
Spesso media, mediana e moda coincidono o sono molto "vicine".

13- Enunciare (a parole) la proprietà associativa della media aritmetica e discuterne l'utilità nelle applicazioni pratiche di ricerca sociale

Quando u è molto numerosa si utilizzano dati aggregati, si tratta di considerare U di numerosità N suddivisa in un certo numero di sottopopolazioni Uj, ciascuna di numerosità Nj con j=1,…h. Ci interessa sempre la media generale di X sull'intera U, ma non disponiamo di dati individuali, ma solo di dati aggregati e quindi di medie x segnato j della sottopopolazione, per ragioni di privacy. Grazie alla proprietà associativa della media aritmetica possiamo calcolare comunque la media generale, in quanto la media generale di X su U è sempre raggiungibile dai dati aggregati , basta calcolare media delle medie delle sottopopolazioni. Si tratta di usare le medie parziali al posto della modalità xi e le numerosità parziali Nj al posto delle frequenze fi. (vedi formula pag 70) La proprietà associativa è molto utile per tutelare la privacy e quando si ha a che fare con una grande mole di dati

14- Enunciare (a parole) la proprietà di equidistribuzione e mantenimento del totale della media aritmetica e discuterne l'utilità nelle applicazioni pratiche di ricerca sociale

- TOTALE: la somma di tutti i valori di X su tutte le N unità osservate. Dividendo il totale di X per N si ottiene la media aritmetica di X.
- EQUIDISTRIBUZIONE: è la situazione in cui tutta la popolazione manifesta un unico valore

La formula è la sommatoria per i che va da 1 a k della media moltiplicata per le frequenze assolute.
Quando parliamo di mantenimento del totale intendiamo dire che se sostituiamo ai valori osservati la media aritmetica che li sintetizza tutti, il totale X non cambia. Inoltre la media aritmetica equidistribuisce il totale di X sulle N unità di U.
Esempio del pollo di Trilussa: la media dà un unico valore medio che bilancia la distribuzione. La media coglie solo un aspetto parziale del fenomeno e non basta: nella realtà, non tutte le risorse sono equidistribuite. Se una persona mangia 2 polli, e un'altra non ne mangia nessuno, la media dà 1 pollo a testa ma questo risultato non rispecchia la realtà, in quanto la media non coglie la variabilità del fenomeno preso in considerazione, non colgie la differente distribuzione del totale tra le modalità.

15- Discutere il concetto di variabilità di un fenomeno quantitativo; descriverne (a parole) la metodologia di misura basata sugli scarti quadratici dalla media aritmetica.

VARIABILITA': attitudine di un fenomeno quantitativo a manifestarsi, sulle N unità di U, con modalità fra loro diverse e distanti.
È un aspetto essenziale nella descrizione statistica del comportamento di X su U.
Una misura (assoluta) della variabilità di X (su U) è un indice sintetico calcolato sulla v.s. con le seguenti caratteristiche:

- Assume valore 0 in assenza di variabilità, cioè nella situazione limite in cui X si manifesta sulle N unità di U con un'unica modalità, generando una v.s. costante
- Assume valori positivi (>0) quando X si manifesta su U con modalità molteplici e differenti, cioè in caso di variabilità
- Assume valori positivi e sempre più grandi all'aumentare della variabilità

Un modo semplice e intuitivo per costruire un indice con queste proprietà è confrontare la più piccola e la più grande fra le modalità osservate. Chiamiamo questo misura di variabilità **range.** è una misura assoluta di variabilità:

- Vale 0 quando X si manifesta con un'unica modalità e perciò xmax = xmin.
- Assume valori positivi quando X si manifesta con più modalità diverse e perciò xmax>xmin.

Il Range è una misura di variabilità grossolana, è sensibile ai valori anomali, quando xmin è molto piccola o xmax è molto grande. Inoltre è basato solo su 2 fra le k modalità osservate, quelle estreme, mentre il resto della v.s è ignorato.
Una misura di variabilità più raffinata è la **deviazione standard** di X (o scarto quadratico medio), meno sensibile ai valori anomali e la indichiamo con la lettere sigma. Anziché confrontare tra loro le singole modalità di X si confrontano ciascuna delle k modalità osservate xi con un unico valore fisso. Ogni modalità xi è confrontata con la media aritmetica. La differenza (xi - media) si chiama scarto o deviazione e può essere o positiva o negativa. A noi interessa la distanza che xi ha dalla media. Per eliminare l'influenza del segno, e quindi avere tutte le differenze positive, si considerano gli **scarti quadratici**, cioè (xi - media)^2. Il quadrato è semplice da trattare matematicamente e si preferisce al valore assoluto. Questi scarti quadratici vengono poi ponderati con le frequenze, tenendo conto che la modalità xi si presenta in U fi volte. Se si sommassero tutti gli scarti ponderati non al quadrato ma con il loro segno positivo o negativo, si otterrebbe sempre 0, un'altra ragione per eliminare l'influenza del segno. Infine avendo k scarti quadratici li sommiamo tutti e poi li dividiamo per N, ristabilendo poi l'ordine di grandezza e unità di misura prendendo la radice quadrata.
Sigma quindi misura la variabilità di X considerando la dispersione dei suoi valori intorno al suo valore medio e ci dice che X si manifesta su U con valori che distano dalla media + o – sigma. Questo valore sigma però non è facilmente interpretabile e nemmeno confrontabile con sigma calcolata su una diversa v.

16- Definire Deviazione standard, Varianza, Devianza; discutere comparativamente l'impiego e il potenziale informativo.

- DEVIEAZIONE STANDARD: rappresenta tutta la v.s. ed è meno sensibile agli eventuali valori anomali.
  Anziché confrontare tra loro le singole modalità di X si confrontano ciascuna delle k modalità osservate xi con un unico valore fisso.
  Ogni modalità xi è confrontata con la media aritmetica.
  Sigma quindi misura la variabilità di X considerando la dispersione dei suoi valori intorno al suo

valore medio e ci dice che X si manifesta su U con valori che distano dalla media + o – sigma. È una misura (assoluta) della variabilità di X (su U):

- Assume valore 0 in assenza di variabilità, cioè nella situazione limite in cui X si manifesta sulle N unità di U con un'unica modalità, generando una v.s. costante
- Assume valori positivi (>0) quando X si manifesta su U con modalità molteplici e differenti, cioè in caso di variabilità
- Assume valori positivi e sempre più grandi all'aumentare della variabilità

- VARIANZA: è la deviazione standard elevata al quadrato (si elimina la radice quadrata). Vale 0 in caso di assenza di variabilità e assume valori positivi e crescenti all'aumentare della variabilità di X su U. la varianza però non è una buona misura di variabilità in quanto l'ordine di grandezza e l'unità di misura sono alterati dal quadrato (pensiamo per esempio a $^2). L'eliminazione della radice quadrata ha notevoli vantaggi analitici, gode di parecchie proprietà statistiche di cui non gode invece la deviazione standard e ha potenzialità descrittive maggiori

- DEVIANZA: è la varianza moltiplicata per N. Anche essa è una misura di variabilità: vale 0 in assenza di variabilità e assume valori positivi e crescenti al crescere della variabilità, ma non è buona. È una quantità al quadrato ed è un totale anziché una media (perché non è divisa per N). è più conveniente da usare in linea teorica, non per misurare la variabilità.

Tutte e tre sono misure assolute di variabilità, cioè sono influenzate dall'ordine di grandezza e dall'unità di misura.
Di conseguenza non sono né valutabili, né confrontabili. Serve una misura relativa

17- Definire la relazione fra Deviazione Standard, Varianza e Devianza di un fenomeno quantitativo commentandone l'informazione descritta

18- Discutere il problema del confronto della variabilità fra due diversi fenomeni osservati sulla medesima popolazione statistica ovvero del medesimo fenomeno osservato su due diverse popolazioni.

La deviazione standard e la varianza sono misure assolute di variabilità, cioè sono influenzate dall'ordine di grandezza e dall'unità di misura. Di conseguenza non sono né valutabili, né confrontabili. Sono indici che ci dicono solo che X presenta variabilità, ma non se tale variabilità sia tanta o poca. Per confrontare e valutare la variabilità di X occorre costruire una misura di variabilità relativa, per togliere gli elementi di disturbo (unità di misura e ordine di grandezza). Per costruire una misura di variabilità relativa si mette a rapporto la misura assoluta con un valore medio che sintetizzi l'ordine di grandezza di X e che sia espressa nella medesima unità di misura: questo è il coefficiente di variazione.

19- Definire (a parole) il Coefficiente di Variazione e commentarne l'uso per il confronto e per la valutazione della variabilità di un fenomeno quantitativo.

Il coefficiente di variazione è una misura relativa. Si costruisce ponendo la deviazione standard sigma (misura assoluta di variabilità) a rapporto con la media aritmetica, ovvero la sintesi dell'ordine di grandezza ed espressa nella stessa unità di misura con cui è rilevato X. Il cv è un indice puro, cioè senza unità di misura, è confrontabile fra fenomeni con diverso ordine di grandezza e diversa unità di misura e fra fenomeni rilevati su popolazioni diverse. E' valutabile come percentuale della media. Valutare la variabilità di un fenomeno serve anche a valutare la capacità di sintesi della media aritmetica: più è alta la variabilità del fenomeno, meno informativa è la media aritmetica. Il cv è sempre positivo, può occasionalmente presentarsi inferiore a 1. Il cv non è la percentuale della variabilità , ma solo della media

20- Enunciare (a parole) una metodologia di costruzione di un indice normalizzato e discuterne l'utilità statistica.

La normalizzazione è il procedimento di trasformazione di un indicatore statistico assoluto in una percentuale. Per normalizzare un indice: trasformarlo in percentuale. Se è vicino a 0 la connessione è poca, se è vicino a 100 è forte. Per poter valutare e confrontare. Normalizzare un indice significa trasformarlo in un numero compreso nell'intervallo (0,1) in modo che, moltiplicato per 100, diventi una percentuale e diventi facilmente interpretabile. Trovo un numero compreso tra 0 (minimo) e 1 (massimo). Siamo quindi in grado di dire se ciò che I misura di X è tanto o poco: è poco se è vicino a 0, è tanto se è vicino a 1.(vedi libro pag 95)

DESCRITTIVA BIVARIATA

1- Descrivere come si organizza il risultato della rilevazione congiunta di una coppia di fenomeni statistici e discutere le distribuzioni di frequenza leggibili sulla tabella a doppia entrata.

I due fenomeni X e Y sono osservati congiuntamente (insieme) su ciascuna delle N unità che compongono la popolazione di interesse U. Il risultato della rilevazione è un insieme di N coppie del tipo (x, y) che prende il nome di matrice dei dati (grezzi). Il risultato della rilevazione congiunta viene organizzato in una tabella a doppia entrata composta da righe e colonne. La tabella a doppia entrata struttura i dati (grezzi) bivariati, organizza i casi osservati e dà le prime indicazioni circa l'eventuale relazione tra i due fenomeni. L'obiettivo è la descrizione del comportamento congiunto di due fenomeni.

- All'interno della tabella si leggono informazioni bivariate. Frequenze **congiunte** che riguardano entrambi i fenomeni (fij). La somma generale di tutte le frequenze congiunte riproduce la numerosità N di ; è una somma doppia poiché riguarda entrambi gli indici i e j, cioè sia per riga sia per colonna. L'interno della tabella a doppia entrata costituisce la variabile statistica doppia, strumento base della statistica descrittiva bivariata
- Ai margini della tabella si trovano le frequenze che riguardano i fenomeni X e Y separatamente: le frequenze **marginali**. Contengono frequenze univariate che ricavo da informazioni bivariate. Danno informazioni solo su un fenomeno, l'altro si trascura. Ci dicono come si comportano due fenomeni indipendenti l'uno dall'altro. Si leggono sulla riga e sulla colonna marginali della tabella. Si ottengono sommando le frequenze congiunte sulla stessa riga o sulla stessa colonna

Avremmo quindi: fi.= frequenze marginali di X e f.j= frequenze marginali di Y. La somma delle frequenze congiunte sulla i-esima riga da le frequenze marginali di X; la somma delle frequenze congiunte sulla j-esima colonna da le

frequenze marginali di Y; la domma di tutte le frequenze congiunte oppure di tutte le frequenze marginali riproduce la numerosità di N.
Vi sono anche le fr marginali relative di X : fi./N, danno risultato totale uguale a 1(somma unitaria) e le fr margina li relative di Y; f.j/N anche queste con somma unitaria.

2- Enunciare (a parole) e interpretare le frequenze marginali e le frequenze condizionate descrivendone il ruolo nella definizione di indipendenza statistica.

Ai margini della tabella si trovano le frequenze che riguardano i fenomeni X e Y separatamente: le frequenze **marginali**. Contengono frequenze univariate che ricavo da informazioni bivariate. Danno informazioni solo su un fenomeno, l'altro si trascura. Ci dicono come si comportano due fenomeni indipendenti l'uno dall'altro. Si leggono sulla riga e sulla colonna marginali della tabella. Si ottengono sommando le frequenze congiunte sulla stessa riga o sulla stessa colonna. Fi.= fr marginali di X, f.j = fr marginali di Y. Le frequenze marginali si ottengono sommando le frequenze congiunte che stanno sulla stessa riga o sulla stessa colonna.

Frequenze **condizionate**: frequenze relative ottenute dal rapporto fra le frequenze congiunte e la frequenza marginale della modalità con cui si condiziona. Si leggono sulle righe o sulle colonne separatamente. Questo tipo di frequenze, si costruiscono sulle varabile statistica condizionate (Y condizionato da xi si scrive cosi: YIxi; X condizionato da yi si scrive cosi: XIyi.).Si avranno tante vs condizionate quante sono le possibili modalità condizionanti; si hanno k variabili condizionate di tipo YIxi e ha variabili condizionate dell'altro tipo. Fr condizionate di YIxi = fij/fi (percentuali di riga).
Fr condizionate di XIyi= fij/f.j (percentuale di colonna). Danno informazione sul comportamento di un fenomeno condizionatamente dall'altro.
Se fra X e Y non esiste alcuna relazione statistica, allora X e Y sono statisticamente indipendenti. Per saperlo, confrontiamo le frequenze condizionate con le frequenze marginali (relative). Unico accorgimento è tener conto che le fr marginali si riferiscono all'intera U di numerosità N mentre le fr condizionate si riferiscono a sotto popolazioni di numerosità fi. o f.j. Quindi per poter effettuare un confronto dobbiamo trasformare le frequenze marginali in frequenze relative (fi./N o f.j/N). Se tutte le k serie di frequenze condizionate fij/fi. sono uguali tra loro e uguali alla marginale relativa f.j/N, ne deduciamo che X e Y sono statisticamente indipendenti. La condizione deve valere per tutti gli indici.

CONDIZIONE DI INDIPENDENZA STATISTICA : fij/fi. = f.j/N , per tutti gli indici i=1…k e j=1…h. non vi è differenza tra comportamento condizionato e marginale. Distribuzioni identiche tra loro

3- Dopo aver esposto il concetto e di Indipendenza Statistica, illustrare (a parole) la metodologia di verifica dell'esistenza o meno in una tabella a doppia entrata.

Posso rispondere citando ancora condizione di indipendenza statistica + :

A ogni tabella di dati rilevati nella realtà (tabella osservata) si può accostare la corrispondente tabella teorica di indipendenza statistica. Si mantengono fisse le marginali e si sostituiscono le frequenze congiunte osservate con le frequenze teoriche di indipendenza statistica. Queste fr teoriche di i.s si calcolano cosi: fi. f.j/N
L'obiettivo è di verificare se esiste relazione tra X e Y. Se le due tabelle non coincidono, ovvero non tutte le fr congiunte sono uguali alle fr teoriche di indipendenza e dunque la tabella osservata differisce dalla tabella teorica di indipendenza, la condizione di indipendenza non è verificata e quindi X e Y non sono statisticamente indipendenti. Il concetto di indipendenza statistica è simmetrico: X è indipendente da Y e contemporaneamente Y è indipendente da X quindi quando vale fij/fi.= f.j/N vale anche fij/f.j= fi./N.

Questo tipo di analisi è possibile per fenomeni di ogni natura.

4- Dopo aver esposto il concetto di Connessione fra due fenomeni statistici, descrivere (a parole) e discutere la costruzione dell'indice di connessione $\chi 2$ e i suoi possibili valori.

Se si conclude che X e Y non sono statisticamente indipendenti (non tutte le fr congiunte sono uguali alle fr teoriche di indipendenza e dunque la tabella osservata differisce dalla tabella teorica di indipendenza), allora fra X e Y esiste una qualche relazione statistica. X e Y sono quindi connessi.

Connessione: generica relazione statisticamente rilevabile in una coppia di fenomeni osservati sulla U d'interesse.

Il passo successivo è stabilire se la relazione rilevata tra X e Y è forte o debole, cioè misurare il grado di connessione. L'intensità di connessione è tanto più elevata quanto più la tabella osservata è lontana dalla tabella teorica di indipendenza. Il metodo per misurare la connessione consiste nel considerare la differenza tra le frequenze congiunte e le frequenze teoriche di indipendenza statistica. (fij- fij*). Queste differenze possono essere tutte nulle se fra X e Y esiste indipendenza statistica, cioè quando tabella osservata e teorica di i.s coincidono. Se queste differenze sono vicine a 0 la connessione è bassa: esiste una relazione debole fra X e Y e quindi i due fenomeni sono connessi ma si influenzano poco l'uno con l'altro. All'aumentare del valore di tali differenze si ha connessione sempre più alta, cioè una relazione forte tra X e Y, indicativa del fatto che i due fenomeni si influenzano e hanno molto a che fare l'uno con l'altro. In una tabella a doppia entrata con k righe e h colonne sono calcolabili kxh differenze Queste differenze possono essere sia positive che negative e ai nostri fini, cioè misurare la connessione, non interessa il segno di queste differenze ma solo quanto sono grandi; dobbiamo quindi eliminare il segno, e sintetizzare in un unico **indice** tutte le kxh differenze. Per eliminare l'effetto del segno eleveremo al quadrato queste differenze. L'indice che utilizzeremo lo indichiamo con x^2 (chi quadro) e lo chiamiamo **indice di connessione** : (fij-fij*)^2/fij* oppure fij^2/fi.f.j -1 (poi si sommano tutti i valori):

- Se le differenze sono uguali a 0 X e Y sono statisticamente indipendenti, indice di connessione =0
- Se i valori aumentano (elevandole al quadrato diventano ancora più grandi) tanto più elevato sarà chi quadro e quindi ci sarà maggior connessione tra i due fenomeni

È un indice assoluto quindi non è consentita la valutazione, non è interpretabile, e nemmeno confrontabile. Deve essere normalizzato (pag 123 normalizzazione)

5- Definire e interpretare i concetti di medie e varianze marginali e condizionate.

- Media marginale di Y: è la media della v.s. marginale di Y= sommatoria di j che va da 1 a n di yj x f.j/N. (stesso ragionamento media marginale di X). Conseguenza proprietà associativa media aritmetica; media delle medie condizionate riproduce la marginale.

- Varianza marginale di Y: è la varianza della v.s. marginale di Y= sommatoria di j che va da 1 a n di (yj – media marginale y)^2 x f.j

Medie e varianze marginali sono ponderate con le frequenze marginali.

Media condizionata di Y dato xi è la media della v.s. condizionata Y| xi che si legge sulla i-esima riga della tabella (l'indice è fisso)= sommatoria di j che va da 1 a n di yj x fij/ fi.

Varianza condizionata di Y dato xi è la varianza della v.s. condizionata Y| xi che si legge sulla i-esima riga della tabella (l'indice è fisso)= sommatoria di j che va da 1 a n di (yj – y segnato I xi)^2 x fij

Medie e varianze condizionate sono ponderate con le frequenze condizionate. In generale si può dire che Y dipende da X se le medie condizionate variano al variare del condizionamento. Se tutte le medie condizionate sono uguali tra loro e uguali alla marginale si parla di indipendenza in media da X, ovvero X non influisce su Y. Se invece le medie condizionate sono molto diverse tra loro allora abbiamo molta dipendenza, si parla cosi di variabilità delle medie condizionate.

6- Enunciare (a parole) la proprietà di la scomposizione della varianza marginale in varianza "nei" e "fra" gruppi; discuterne l'informazione statistico-descrittiva

Per la varianza vale la proprietà della scomposizione. La varianza marginale di Y si scompone nella somma di due componenti che chiamiamo varianza NEI e varianza FRA. Entrambe queste componenti ci dicono qualcosa di importante riguardo la relazione fra X e Y.

Varianza NEI: media varianze condizionate (pag 137). Sintetizza tutte le varianze condizionate , cioè sintetizza la variabilità di Y all'interno di sottopopolazioni omogenee rispetto a X, mantenendo fisso X. Misura la variabilità di Y che non dipende da X

Varianza FRA: è la varianza delle medie condizionate (pag 138). Sintetizza la variabilità all'esterno delle sotto popolazioni , ovvero fra una sotto popolazioni e l'altra. Misura la variabilità di Y che dipende da X

Quindi la proprietà di scomposizione ci dice che la varianza marginale di Y si spezza in due parti: varianza NEI e varianza FRA.(pag 138). Insieme riproducono l'intera variabilità di Y. Questa proprietà vale solo per la varianza e non per la deviazione standard, ci aiuta a capire quanto Y dipende da X. La relazione statistica di dipendenza è tanto più forte quanto più cresce la variabilità FRA

7- Dopo aver esposto il concetto di indipendenza in media, enunciare (a parole) la condizione e discutere il significato statistico descrittivo.

- INDIPENDENZA IN MEDIA DI Y DA X: Y e X sono connessi (non statisticamente indipendenti). Y dipende da X se tale relazione di connessione si riflette sulle medie condizionate che risultano diverse fra loro al variare di X (cioè condizionatamente alle modalità xi di X) e diverse dalla media marginale (cioè indipendentemente da X). Diremo invece che Y è indipendente in media da X se è sufficiente sintetizzare le distribuzioni condizionate nelle medie condizionate perché la relazione di connessione scompaia e le medie condizionate appaiono tutte uguali fra loro al variare di X e uguali alla media marginale

- CONDIZIONE DI INDIPENDENZA IN MEDIA DI Y DA X: Se tutte le medie condizionate sono uguali tra loro e uguali alla marginale si parla di indipendenza in media da X, ovvero X non influisce su Y.

Ricordiamo che l'i.m è più debole rispetto all'i.s , quindi non è necessariamente vero che se Y è i.m da X allora X e Y sono i.s

8- Dopo aver esposto il concetto di Dipendenza di un fenomeno dall'altro, enunciare la costruzione e interpretare i possibili valori degli indici di dipendenza n 2

Y dipende da X se le medie condizionate sono diverse tra di loro, variano. La varianza FRA misura la parte di variabilità

di Y che dipende da X. Quando Y è i,m da X, allora le k differenze (medie condizionate – media marginale di y) sono tutte uguai a zero. Quando y è i.m da X la varianza FRA vale 0. All'aumentare dell'influenza di X su Y le medie condizionate sono sempre più diverse fra loro e diverse dalla media marginale; allora le differenze sono sempre più grandi all'aumentare del grado di dipendenza di Y da X. È proprio sulla varianza FRA che si basa la misura della dipendenza di Y da X.

INDICE DI DIPENDENZA: lo indicheremo con la lettera eta (n) elevata al quadrato per ricordarci che ha a che fare con le varianze e non con la deviazione standard. Si calcola mettendo a rapporto la varianza FRA con la varianza marginale di y, misura quanto è forte la dipendenza di Y da X. Assume valori compresi tra 0 e 1 , è cioè un indie normalizzato che moltiplicato per 100 è interpretabile come percentuale di dipendenza di Y da X. Vale 0 quando il numeratore è 0 , quindi solo se la varianza FRA è uguale a 0, cioè quando Y è indipendente in media da X. Vale 1 quando varianza FRA = varianza marginale di Y: in questo caso tutta la variabilità di Y dipende da X. Tutti i valori intermedi sono interpretabili come percentuale di dipendenza di Y da X

Tutto questo vale anche se si vuole capire se X dipende da Y.

9- Esporre la metodologia di costruzione di un Diagramma a Dispersione per una coppia di fenomeni quantitativi nel caso di Serie Doppia e nel caso di Tabella a Doppia Entrata.

Quando entrambi i fenomeni sono quantitativi, oltre a rilevare l'esistenza, misurare il grado di una generica relazione e fissarne il verso (dipendenza), possiamo anche studiare la natura della relazione statistica tra X e Y, rappresentandola su grafici

Il diagramma a dispersione è uno strumento grafico utile per visualizzare il tipo di relazione esistente fra due fenomeni X e Y quantitativi. In inglese si chiama scatter plot. È un diagramma cartesiano, con ad esempio X sulle ascisse e Y sulle ordinate. Le coppie di valori osservati (xi, yj) sono viste come coordinate di punti sul diagramma. La tabella è rappresentata sul diagramma come una nuvola di kxh punti. Ci fa vedere che tipo di relazione statistica c'è tra X e Y, sempre che ci sia. Se tra X e Y esiste una certa relazione statistica , la nuvola di punti si presenta strutturata, appaiono più concentrati in particolari zone del diagramma. Tutto questo nel caos in cui abbiamo una serie doppia. (sufficiente un unico indice)

Nel caso in cui abbiamo una tabella a doppia a doppia entrata dovremmo costruire un grafico a 3 dimensioni ma è troppo complicato, quindi costruiremo un diagramma a dispersione a bolle. Le bolle più grandi sono quelle con maggiore peso (fi.j maggiore). I punti sul grafico avranno coordinate (xi, yj). Si può anche costruire la versione ridotta con le medie condizionate, rappresentato i punti di coordinate (xi, y segnato I xi). A seconda dell'andamento della linea che si crea possiamo dire se X e Y sono statisticamente dipendenti o meno. La nuvola di bolle non evidenzia nessuna struttura se i punti sono tutti sparpagliati. X e Y sono quindi statisticamente indipendenti

10- Definire il Diagramma a Dispersione e discuterne l'utilità nell'analisi dell'eventuale relazione esistente fra i due fenomeni.

Quando entrambi i fenomeni sono quantitativi, oltre a rilevare l'esistenza, misurare il grado di una generica relazione e fissarne il verso (dipendenza), possiamo anche studiare la natura della relazione statistica tra X e Y, rappresentandola su grafici

Il diagramma a dispersione è uno strumento grafico utile per visualizzare il tipo di relazione esistente fra due fenomeni X e Y quantitativi. In inglese si chiama scatter plot. È un diagramma cartesiano, con ad esempio X sulle ascisse e Y sulle ordinate. Le coppie di valori osservati (xi, yj) sono viste come coordinate di punti sul diagramma. La tabella è rappresentata sul diagramma come una nuvola di kxh punti. Ci fa vedere che tipo di relazione statistica c'è

tra X e Y, sempre che ci sia. Se tra X e Y esiste una certa relazione statistica , la nuvola di punti si presenta strutturata, appaiono più concentrati in particolari zone del diagramma.

A seconda di come si dispongono i punti si possono individuare diversi tipi di relazione: lineare, quadratica, logaritmica, esponenziale e infine cubica. Quando X e Y sono statisticamente indipendenti i punti si presentano sparpagliati sul diagramma senza nessuna struttura evidente.

11- Esporre il concetto di Correlazione in una coppia di fenomeni quantitativi, discutere il ruolo della covarianza e definire (a parole) il coefficiente di correlazione lineare ρ interpretandone i valori.

Con il termine **correlazione** si intende relazione esistente tra due (o più) fenomeni. Non è una relazione causa effetto ma rappresenta la capacità di un fenomeno di variare in funzione dell'altro. Quindi la relazione statistica tra due fenomeni è chiamata correlazione lineare. La misura della correlazione è basata sulla covarianza

La **covarianza** è una sorta di misura di variabilità congiunta. La indichiamo con la lettera sigma. Si devono calcolare gli scarti ponderati congiunti, utilizzando quindi entrambi gli indici i e j, ponderati con le frequenze congiunte. Può risultare sia positiva che negativa o anche nulla. Il suo valore non è direttamente interpretabile ma è utile per l'analisi di un'altra relazione statistica. Essendo basata sugli scarti (xi – x segnato) e (yj – y segnato) non elevati al quadrato, la covarianza si prende con il suo segno; a seconda che le modalità siano sopra o sotto media, questi scarti sono positivi o negativi. Il segno ci dice se l'andamento congiunto tra i due fenomeni è crescente o decrescente . Oltre che essere basata sugli scarti è basata sui prodotti degli scarti, quindi possiamo avere 4 diversi "risultati" in corrispondenza delle 4 zone del grafico, in quanto dividiamo il quadrante di destra in 4 parti , le quali contribuiscono al calcolo della covarianza stessa:

- Se xi > x segnato, yj > y segnato abbiamo scarti ponderati positivi e quindi covarianza positiva. Gli scarti positivi prevalgono su quelli negativi. Quando la covarianza è positiva X e Y sono positivamente correlati, al crescere dell'uno cresce anche l'altro.
- Se xi < x segnato, yj> y segnato abbiamo scarti ponderati negativi (- x + = -) e quindi covarianza negativa (oppure il contrario). Gli scarti negativi prevalgono su quelli positivi. Se la covarianza è negativa, X e Y sono negativamente correlati, cioè al crescere dell'uno decresce l'altro
- Se sia xi che yj < delle rispettive medie abbiamo scarti ponderati positivi e quindi covarianza positiva
- La covarianza può anche essere uguale a 0, cioè i punti sono disposti in modo che gli scarti positivi e negativi si compensino. Questo accade quando i punti son sparpagliati sul diagramma a dispersione senza alcuna struttura, cioè in caso di indipendenza statistica. X e Y non sono correlati. Succede anche quando i punti non sono strutturati linearmente sul grafico, quindi hanno una struttura diversa, come ad esempio quella quadratica. Ricordiamo che covarianza =0 non implica i.s, ma è il contrario.

Dopo aver constatato che X e Y sono correlati dobbiamo misurare il grado di questa correlazione, stabilire quindi se è forte e debole, usando il **coefficiente di correlazione lineare** , lo indicheremo con la lettera rho. Si costruisce dividendo la covarianza per il suo valore massimo (covarianza xy/ covarianza x moltiplicata per covarianza y). Questo coefficiente assume valori tra -1 e +1 e ci a indicazioni sul verso e sull'intensità della correlazione:

- Se è uguale a -1 X e Y sono perfettamente e negativamente correlati. I punti sul diagramma sono allineati lungo una retta con pendenza negativa , ovvero decrescente
- Se è uguale a +1 X e Y sono perfettamente e positivamente correlati. I punti sul diagramma sono allineati

lungo una retta con pendenza positiva, ovvero crescente

- Se è uguale a 0, X e Y sono incorrelati.
- I valori intermedi sono interpretabili come percentuale di correlazione, in particolare indicano percentuale di correlazione negativa se sono compresi tra – 1 e 0, percentuale di correlazione positiva se i valori sono compresi tra 0 e +1

12- Discutere, i concetti di incorrelazione, perfetta correlazione, indipendenza statistica e massima connessione; per ciascun caso confrontare i valori degli indici r e $\chi 2$

- Incorrelazione: assenza di relazione lineare fra X e Y , da cui segue che coefficiente di correlazione lineare(pxy) è uguale a 0

- Indipendenza statistica: assenza di qualunque relazione fra X e Y, da cui segue che chi quadro (x^2) è uguale a 0

- Perfetta correlazione: esiste una perfetta relazione lineare tra X e Y, quindi indice di correlazione lineare è uguale a + o – 1. L'implicazione matematica è : perfetta correlazione – perfetta dipendenza di Y da X e di X da Y – massima connessione. In formule : pxy=+ o – 1 n^2x e n^2y= 1 x^2=1. Non è vero il contrario.

Massima connessione: si ha quando esiste un legame perfetto fra X e Y avente qualunque natura (lineare e non). Con legame perfetto si intende che un fenomeno determina statisticamente l'altro: fissata una modalità xi, in U esiste un 'unica modalità yj e/o viceversa:

- Se la tabella è quadrata ,cioè k=h ovvero stesso numero di righe e di colonne, la massima connessione è biunivoca, ed implica la perfetta dipendenza di un fenomeno dall'altro , quindi x^2(normalizzato) =n^2x= n^2y=1
- Se la tabella è rettangolare la massima connessione è univoca; se k>h, la massima connessione univoca di riga implica che un solo fenomeno dipende perfettamente dall'altro (il fenomeno sulle colonne dipende da quello sulle righe); se k<h, sarà il contrario.

13- Discutere, i concetti di indipendenza in media, indipendenza statistica e massima connessione; per ciascun caso confrontare i valori degli indici Eta quadro e Chi Quadro

Massima connessione: si ha quando esiste un legame perfetto fra X e Y avente qualunque natura (lineare e non). Con legame perfetto si intende che un fenomeno determina statisticamente l'altro: fissata una modalità xi, in U esiste un 'unica modalità yj e/o viceversa:

- Se la tabella è quadrata ,cioè k=h ovvero stesso numero di righe e di colonne, la massima connessione è biunivoca, ed implica la perfetta dipendenza di un fenomeno dall'altro , quindi x^2(normalizzato) =n^2x= n^2y=1
- Se la tabella è rettangolare la massima connessione è univoca; se k>h, la massima connessione univoca di riga implica che un solo fenomeno dipende perfettamente dall'altro (il fenomeno sulle colonne dipende da quello sulle righe); se k<h, sarà il contrario.

Indipendenza statistica: assenza di qualunque relazione fra X e Y, da cui segue che chi quadro (x^2) è uguale a 0

Indipendenza in media: diciamo che Y è indipendente in media d X se è sufficiente sintetizzare le distribuzioni condizionate nelle medie condizionate perché la relazione di connessione scompaia e le medie condizionate appaiono tutte uguali tra loro e uguali alla media marginale di y. Indice di dipendenza (eta quadro) è uguale a 0 quando la variabilità esterna alle righe o parte di variabilità di Y che dipende da X (FRA) è uguale a 0,ovvero quando c'è indipendenza in media.

14- Esporre il concetto di Modello Statistico e discuterne l'utilità, esemplificando con il modello di regressione lineare semplice.

Un modello statistico è una formula che interpreta matematicamente il comportamento congiunto di X e Y. Costruirlo significa utilizzare i dati della tab osservata per individuare la formula che esprime Y in funzione di X. È una curva matematica semplice in grado di ben approssimata la realtà osservata, cioè di cogliere l'andamento di fondo (o trend) del comportamento congiunto di X e Y. Il metodo più noto è chiamato **regressione,** il più semplice ma anche il più utilizzato è chiamato **regressione lineare semplice**. Il modello di regressione interpreta Y in funzione di X, è una formula per approssimare Y e può essere utilizzato per fare previsioni o simulare valori di Y non osservati, costruendo nuovi scenari. Indichiamo Y con un cappuccio per indicare che stiamo approssimando la realtà osservata con un curva matematica più semplice e regolare : Y^=f(X). Questo modello di regressione vuole cogliere l'essenza della relazione tra X e Y e semplificare al realtà. La funzione è la seguente : Y^= a + bX, dove X è la variabile indipendente, fenomeno condizionante (interpreta variabilità di Y) ed è anche una variabile esplicativa ovvero riusciamo a spiegare Y con X. Al contrario Y^ è la variabile dipendente, è il fenomeno condizionato da X ed è la variabile risposta, in quanto risponde con un valore al variare di X. Il parametro a è l'intercetta, cioè il punto in cui la retta interseca l'asse verticale delle ordinate. È il valore di Y se x=0; Il parametro b è il coefficiente angolare e determina l'inclinazione della retta e la sua pendenza: se è positivo la retta è crescente mentre se è negativo la retta è decrescente.

15- Esporre e interpretare i concetti di: nuvola di punti su un diagramma a dispersione, spezzata di regressione e modello di regressione.

Il diagramma a dispersione è uno strumento grafico utile per visualizzare il tipo di relazione esistente fra due fenomeni X e Y quantitativi. In inglese si chiama scatter plot. È un diagramma cartesiano, con ad esempio X sulle ascisse e Y sulle ordinate. Le coppie di valori osservati (xi, yj) sono viste come coordinate di punti sul diagramma. La tabella è rappresentata sul diagramma come una **nuvola di kxh punti**. Ci fa vedere che tipo di relazione statistica c'è tra X e Y, sempre che ci sia. Se tra X e Y esiste una certa relazione statistica , la nuvola di punti si presenta strutturata, appaiono più concentrati in particolari zone del diagramma.

SPEZZATA DI REGRESSIONE: è una curva empirica (reale) basata sui dati osservati e per questo in genere si presenta irregolare e spigolosa.

MODELLO DI REGRESSIONE: curva teorica, si presenta liscia e regolare. Interpreta la dipendenza di Y da X. Può essere utilizzato per prevedere e simulare valori di Y non osservati. Il fenomeno condizionato Y ha ruolo di variabile dipendente e risposta (risponde con un valore al variare di X, il fenomeno condizionante X ha ruolo di variabile indipendente e esplicativa (ci permette di spiegare Y con X) .Con la regressione si va a individuare la curva matematica liscia e regolare che meglio approssima la spezzata di regressione. Il modello teorico approssima la spezzata di regressione (curva empirica).

16- Esporre (a parole) e discutere il criterio dei Minimi Quadrati per la determinazione

della Retta di Regressione

Per scegliere la retta che meglio approssima la spezzata di regressione, utilizziamo il criterio dei minimi quadrati: esprime la distanza tra i dati osservati e la retta di regressione e assegna ai parametri del modello il valore che rende minima tale distanza. Si calcola lo scarto quadratico tra yj(valori osservati reali) e yi cappuccio (valori teorici approssimati mediante il modello); questa differenza al quadrato (per eliminare influenza del segno) va ponderata con le frequenze congiunte osservate. La condizione dei minimi quadrati è appunto che questa differenza sia la più piccola possibile: dobbiamo quindi minimizzare la funzione. Da questo problema di minimo troviamo i valori dei parametri a e b della retta di regressione: a= y segnato – b moltiplicato per x segnato; b= covarianza xy/ varianza x^2. Sostituendo la soluzione dei minimi quadrati nella retta di regressione si ottiene la retta dei minimi quadrati cioè al sola retta che, tra le infinite, rende minima la distanza totale fra i dati osservati e il modello. Il coefficiente angolare b può essere sia positivo che negativo, in quanto al numeratore la covarianza può essere sia negativa che positiva, quindi:

- Correlazione positiva: covarianza positiva, b>0. Retta dei minimi quadrati crescente
- Correlazione negativa: covarianza negativa, b<0. Retta dei minimi quadrati decrescente

In generale quindi il valore di a ci dice quanto vale Y^ quando X=0, mentre b ci dice di quanto varia Y^ quando X aumenta di 1. Infine per disegnarla basta individuare 2 punti perché per due punti passa una sola retta. Quelli più semplici da individuare sono quelli in cui la retta interseca gli assi: poniamo prima X=0 e poi Y=0, due punti di coordinate (0;a) e (-a/b;0).

17- Definire (a parole) i parametri della retta di regressione dei minimi quadrati, discuterne i valori e l'interpretazione statistico-descritti Definire (a parole) i concetti di "devianza spiegata" e "devianza residua" di un modello di regressione e discutere il loro ruolo nella misura della bontà del modello

I valori dei parametri a e b della retta di regressione: a= y segnato – b moltiplicato per x segnato; b= covarianza xy/ varianza x^2. Sostituendo la soluzione dei minimi quadrati nella retta di regressione si ottiene la retta dei minimi quadrati cioè al sola retta che, tra le infinite, rende minima la distanza totale fra i dati osservati e il modello. Il coefficiente angolare b può essere sia positivo che negativo, in quanto al numeratore la covarianza può essere sia negativa che positiva, quindi:

- Correlazione positiva: covarianza positiva, b>0. Retta dei minimi quadrati crescente
- Correlazione negativa: covarianza negativa, b<0. Retta dei minimi quadrati decrescente

In generale quindi il valore di a ci dice quanto vale Y^ quando X=0, mentre b ci dice di quanto varia Y^ quando X aumenta di 1. Infine per disegnarla basta individuare 2 punti perché per due punti passa una sola retta. Quelli più semplici da individuare sono quelli in cui la retta interseca gli assi: poniamo prima X=0 e poi Y=0, due punti di coordinate (0;a) e (-a/b;0). In ogni caso le estrapolazioni non devono allontanarsi troppo dai dati osservati

Dopo aver sostituito i parametri a e b con le soluzioni dei minimi quadrati, la distanza totale tra i valori reali osservati e la retta ci da il residuo totale della retta, chiamato **devianza residua.** Il residuo delle retta dei minimi quadrati è nullo (DR=0) quando sono nulle le distanze fra i valori osservati e i valori teorici del modello, cioè quando la retta si adatta perfettamente ai dati reali. Questo avviene quando X e Y sono perfettamente correlati e i punti sul diagramma a dispersione si presentano allineati lungo una retta crescente o decrescente. In tutti gli altri casi la regressione lascia

un qualche residuo, quindi DR>0. La DR è una misura assoluta (formula pag 173)e quindi non è ne valutabile ne confrontabile. Quando il residuo non è nullo non sappiamo dire se è tanto o poco e quindi stabilire se la retta dei minimi quadrati è un modello buono o cattivo. Dobbiamo quindi normalizzare il residuo, cioè trasformarlo in una percentuale che ne consenta la valutazione. La varianza marginale di Y moltiplicata per N è chiamata **devianza totale,** la quale si scompone nella somma di due parti: una componente è la devianza residua , l'altra è la **devianza spiegata** . queste due componenti ci dicono qualcosa di importante circa la bontà del modello di regressione.Il modello sarà buono se è in grado di di spiegare come varia il fenomeno. Più la DS è alta , più è bassa la DR, se la DR è alta abbiamo un cattivo modello e la DS cattura meno variabilità.
Dobbiamo quindi costruire un indice interpretabile come percentuale che misuri l'adattamento della retta dei minimi quadrati ai dati osservati. Innanzitutto scomponiamo la DT in DS e DR e poi normalizziamo: DS/DT = DT x pxy^2/DT = pxy^2.

18- Definire (a parole) i concetti di "devianza spiegata" e "devianza residua" di un modello di regressione e discuterne le semplificazioni nel caso della retta dei minimi quadrati.

Dopo aver sostituito i parametri a e b con le soluzioni dei minimi quadrati, la distanza totale tra i valori reali osservati e la retta ci da il residuo totale della retta, chiamato **devianza residua.** Il residuo delle retta dei minimi quadrati è nullo (DR=0) quando sono nulle le distanze fra i valori osservati e i valori teorici del modello, cioè quando la retta si adatta perfettamente ai dati reali. Questo avviene quando X e Y sono perfettamente correlati e i punti sul diagramma a dispersione si presentano allineati lungo una retta crescente o decrescente. In tutti gli altri casi la regressione lascia un qualche residuo, quindi DR>0. La DR è una misura assoluta (formula pag 173)e quindi non è ne valutabile ne confrontabile. Quando il residuo non è nullo non sappiamo dire se è tanto o poco e quindi stabilire se la retta dei minimi quadrati è un modello buono o cattivo. Dobbiamo quindi normalizzare il residuo, cioè trasformarlo in una percentuale che ne consenta la valutazione. La varianza marginale di Y moltiplicata per N è chiamata **devianza totale,** la quale si scompone nella somma di due parti: una componente è la devianza residua , l'altra è la **devianza spiegata** . queste due componenti ci dicono qualcosa di importante circa la bontà del modello di regressione in quanto:

- La DS è la parte di variabilità di Y spiegata o catturata dalle retta dei minimi quadrati. Si calcola cosi: DT pxy^2
- La DR misura la variabilità residua , cioè la parte di variabilità di Y che non è catturata dalla retta dei minimi quadrati. Si calcola cosi: DT(1- pxy^2). DR=0 se e solo se pxy^2 =+ o – 1, cioè soltanto in caso di perfetta correlazione tra X e Y. DR e pxy^2 sono in relazione inversa , residuo retta mq è tanto più piccolo quanto più è elevata la correzione.

Dobbiamo quindi costruire un indice interpretabile come percentuale che misuri l'adattamento della retta dei minimi quadrati ai dati osservati. Innanzitutto scomponiamo la DT in DS e DR e poi normalizziamo: DS/DT = DT x pxy^2/DT = pxy^2. Questo è vero solo per la retta dei minimi quadrati , in quanto si adatta meglio alla realtà osservata quanto più è elevata la correlazione tra X e Y. DS/DT è compreso tra 0 e 1

- Pxy^2=0 se DS =0, quindi DR=DT, la retta lascia tutto residuo, è lo scenario peggiore , X e Y sono incorrelati
- Pxy^2=1 se DS=DT, quindi DR=0. La DS cattura tutta la variabilità di Y , il modello è perfetto non lascia residuo e distanze, X e Y sono perfettamente correlati
- I valori intermedi sono interpretabili come percentuali di variabilità di Y spiegata dalla retta dei minimi quadrati o come % di bontà del modello

Pxy^2 misura la bontà della retta dei minimi quadrati , + alta è la correlazione , migliore sarà la retta.

ULTIME DUE DOMANDE MOLTO SIMILI, INTERSCAMBIABILI.

INFERENZA

1- Discutere concetti e obiettivi dell'Inferenza statistica e le specifiche problematiche rispetto alla Statistica descrittiva.

Più spesso non si dispone dell'intera popolazione ma solo di dati parziali relativi ad un sottoinsieme di U campione. L'obiettivo dell'inferenza statistica è estendere l'analisi del comportamento di X all'intera U, in quanto si tratta di inferire dal campione all'intera popolazione. La rilevazione completa di U si chiama censimento mentre quando l'osservazione di X avviene solo su una parte di U si effettua una rilevazione campionaria. Le rilevazioni campionarie sono più frequenti e preferibili rispetto al censimento per ragioni di budget, in quanto una rilevazione campionaria richiede risorse ridotte in termini di tempo e costi, per ragioni di precisione e In alcuni casi la rilevazione parziale si impone rispetto alla rilevazione esaustiva perché quest'ultima risulta impossibile o sconveniente.
U è ignoto: non abbiamo tutte le info di tutte le unità statistiche e disponiamo solo dei dati campionari. L'inferenza statistica è un'inferenza induttiva (dal particolare al generale) che procede dal campione (una parte) alla popolazione (il tutto) ed è a rischio di errore, ma nonostante questo controlla effetto del caso attraverso la probabilità. Il campione deve essere rappresentativo, cioè è un'immagine in scala ridotta ma fedele dell'intera U, deve essere casuale, cioè scelto a caso da U stessa. La casualità del campione è garanzia della sua rappresentatività.
Con inferenza ci troviamo in una situazione casuale in quanto vi è sempre una componente di incertezza.

2- Esporre i concetti di Esperimento casuale, Evento elementare, Spazio campionario, Evento causale

Esperimento casuale: è un esperimento condotto sotto l'effetto del caso, cioè quando è nota solo una parte delle circostanze che consentirebbero di prevederne il risultato con certezza a priori. Di un esperimento casuale è possibile solo elencare a priori l'insieme dei possibili esiti. Un esempio sono i giochi d'azzardo con monete, dadi, roulette) eseguiti regolarmente e senza barare.

Evento elementare: ciascuno dei possibili esiti di un esperimento casuale

Spazio campionario: è l'insieme di tutti i possibili esiti di un esperimento casuale, elencabili a priori, è quindi l'insieme di tutti gli eventi elementari. Usiamo la lettera omega maiuscola per denotare lo spazio campionario

Evento casuale: è un sottoinsieme dello spazio campionario ed è un concetto più generale rispetto a quello di evento elementare in quanto un evento elementare è un singolo evento di omega, mentre l'evento casuale è un sottoinsieme di omega cioè un insieme di eventi elementari. Infatti può contenerne uno, nessuno, molti ,pochi. La notazione che usiamo per indicarlo è la E,i cui elementi sono appunto eventi elementari i quali possono appartenere o non appartenere a E.

3- Esporre, e discutere comparativamente, le definizioni classica e frequentista di probabilità

Classica: è la più antica e semplice ed è applicabile a spazi campionari finiti. P(E) è il rapporto (cioè una frazione) fra il numero di casi favorevoli a E e il numero di tutti i casi possibili. Due limiti: i casi devono essere ugualmente possibili (equiprobabili) e bisogna contare sia il numero di casi favorevoli che il numero di casi possibili (e con eventi complessi non sempre è possibile). Questa definizione è insufficiente: è impraticabile o impossibile contare i casi possibili e i casi

favorevoli al verificarsi di eventi complessi. La definizione classica è limitata solo ai casi in cui E è bilanciato, quindi finito e simmetrico.

Frequentista: definizione basata sull'osservazione, legge empirica del caso, si osserva 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 restati N-f volte). La legge empirica del caso dice che 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. La probabilità di E è quel valore intorno al quale tende a stabilizzarsi la frequenza relativa dopo un numero sufficientemente grande di prove:
P(E)= lim per N che tende a infinito di f/N. La definizione frequentista di probabilità è più ampia di quella classica: permette di considerare spazi campionari infiniti e di calcolare la probabilità di eventi anche quando i casi possibili non sono tutti ugualmente possibili. Permette di probabilizzare eventi più complessi. Ha anche i suoi problemi:

- Le prove devono essere ripetute tutte nelle stesse condizioni
- Casistiche "sufficientemente grandi"

La probabilità di un qualunque evento casuale E è un numero compreso tra 0 e 1. Elimina la circolarità e la limitatezza della definizione classica. Gli eventi non sono ugualmente possibili . Supera la definizione classica ma è ancora limitata: l'oggetto cresce sempre di più con l'avanzamento tecnologico.

4- Esporre il concetto di variabile casuale, la sua utilità in relazione all'esperimento casuale, e l'analogia con la variabile statistica (descrittiva) inclusi i concetti di media e varianza

La variabile casuale è 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. 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. La variabile casuale prende gli elementi di Ω e suoi sottoinsiemi e li trasforma in numeri reali, cioè in valori della variabile casuale. Nasce un'osservazione parziale e casuale della realtà. Il procedimento è analogo a quello della statistica descrittiva; si parte da un inferenza statistica (nel caso della descrittiva si parte appunto dalla statistica descrittiva), si crea la variabile casuale formata dalle coppie (nel caso della descrittiva si crea l v. statistica) e infine la v casuale trasforma gli elementi dei sottoinsiemi dello spazio campionario in valori, cioè probabilità del corrispondente evento (nel caso della descrittiva la v statistica contiene modalità e frequenze relative).

Una varabile casuale è detta discreta quando assume un numero finito di valori x che di solito sono numeri interi e la somma di tutti questi valori è pari a 1 in perfetta analogia con la somma delle frequenze relative della variabile statistica. Quindi sfruttando l'analogia con la v.s è possibile trasferire sulla v.c concetti della statistica descrittiva, come ad esempio il concetto di media che quando è riferita a una v.c viene chiamata **valore atteso** o expectation. Un altro concetto è quello della varianza che nel caso della vc è definita e calcolata come per la v.s, ma usando le 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

5- Esporre (a parole) le caratteristiche e le proprietà della variabile casuale Normale

discutendone il ruolo centrale nell'inferenza statistica.

La variabile casuale continua si usa per fare inferenza statistica su fenomeni continui, quelli che non si possono contare ma si possono misurare. Le v.c normali assumono **infiniti** valori e quindi occorre far riferimento a insiemi di valori, ovvero **intervalli**; la probabilità è calcolabile solo per gli intervalli. Siccome i singoli valori non sono visibili, non si parla più di probabilità P(X=x) ma di **funzione di densità**, che indichiamo con la lettera greca fi e serve per calcolare le probabilità di intervalli di valori. Infine le probabilità non sono singoli valori ma aree: l'area sottesa al grafico della funzione di densità fi(x) in un intervallo è la probabilità che X assuma valori in quell'intervallo.

La **variabile casuale Normale** è la più nota tra le v.c. continue ed è la più utile nell'inferenza statistica. È continua, cioè utilizzata per fare inferenza su fenomeni continui.

Notazione: X-N(μ, varianza) che si legge "X è una v.c. normale di parametri μ e varianza ". Il parametro μ può essere un numero reale qualunque mentre il parametro "varianza" è un numero reale positivo. Questa v.c presenta 10 proprietà:

1. È v.c. continua e assume tutti i possibili valori reali:
2. Ha la funzione di densità fi(x) . L'area sottesa al grafico della fi(x) in un certo intervallo rappresenta la probabilità che la v.c. assuma valori in quell'intervallo. La sua rappresentazione grafica è la curva a campana centrata sul valore μ.
3. L'area sottesa all'intera curva fi(x) corrisponde alla probabilità dell'intero intervallo "- infinito + infinito " ed è pari a 1.
4. Il parametro μ è la media di X-N (μ, varianza) .In formule: E(X)= μ
5. Il seconda parametro è la varianza di X-N (μ, varianza). In formule: V(X)= varianza e dunque SD(X)= radice quadrata della varianza = deviazione standard
6. La curva a campana è simmetrica rispetto a μ, cioè l'area sottesa alla curva a destra e a sinistra di μ è uguale e dunque pari a 0,5. P(X<= μ)= p(X>= μ)=0.5X assume valori sotto media e sopra media con la stessa probabilità.
7. Il parametro μ rappresenta anche la mediana e la moda di X
8. I flessi (le due code) corrispondono ai punti μ-varianza, μ+varianza , cioè una deviazione standard dal valore medio. Tra questi due valori c'è la pancia dove è compresa la maggior probabilità.La curva cambia concavità in corrispondenza dei flessi; dove c'è meno probabilità l'area si riduce
9. I parametri μ e varianza della normale, oltre che rappresentare moda, media, mediana e varianza di X, determinano anche la posizione e la forma della campana.
10. La probabilità di un qualunque intervallo di valori X è l'area sottesa alla campana in quell'intervallo.

La normale tende a manifestarsi con un valore sistematico prevalente (μ). I valori più probabili sono vicini a tale valore prevalente, i valori lontani da μ sono rari e poco probabili. Inoltre il parametro μ determina la posizione della curva a campana e variando μ la curva trasla ma la forma rimane uguale. Mentre il parametro "varianza" determina la forma della curva a campana, appiattita o appuntita, alterando i flessi; se il parametro varianza diminuisce i flessi si avvicinano e la curva si impenna, mentre se aumenta avviene il contrario. Infine possiamo dire che la v.c Normale è un buon interprete per fenomeni quantitativi, continui e misurabili. STANDADIZZAZIONE.

6- Esporre (a parole) le caratteristiche della variabile casuale Binomiale e discuterne l'utilità per l'inferenza su fenomeni dicotomici.

E' una particolare v.c discreta. L'esperimento casuale consiste nell'esecuzione di n prove indipendenti , cioè l'esito di ciascuna prova non influenza l'esito della prova successiva. Un esempio potrebbe essere un certo numero di estrazioni a caso condotte tutte nelle stesse condizioni, cioè con il reinserimento dell'unità estratta. Ogni prova può

avere come esito uno e soltanto uno di due eventi fra loro contrari ed esaustivi e li chiamiamo successo e insuccesso. In questo modo possiamo modellare i fenomeni dicotomici, cioè i fenomeni statistici che si manifestano solo con 2 modalità. Infine in ciascuna prova la probabilità del successo è nota e costante e denotata con al lettera p. p è compreso fra 0 e 1 e quindi la probabilità dell'insuccesso sarà 1-p.
Per indicare la v.c Binomiale usiamo la notazione: X-Bin (n,p). n e p sono chiamati parametri della v.c. Ciascuna prova può avere come esito o un successo o un insuccesso e di prove ne facciamo n; allora il generico risultato delle n prove è una n-upla di successi e insuccessi. Ogni n-upla può contenere 0 successi (quindi tutti successi) oppure 1 successo e n-1 insuccessi e cosi via fino a n successi. I possibili valori della v.c Binomiale quindi sono numeri interi da 0(tutti insuccessi) a n(tutti successi). La funzione di probabilità P(X=x) della v.c binomiale da la probabilità di ottenere x successi sulle n prove con x= 0,1,2…n. Scritta in maniera compatta la probabilità della n-upla è: p^x (1-p)^n-x. Il problema è che la n-upla contiene x successi e n-x insuccessi e quindi può presentarsi in molti ordini diversi; per contare il numero di possibili combinazioni di x successi e n-x insuccessi in ordine diverso si calcola il **coefficiente binomiale** che si legge n su x e si calcola cosi: n!/x!(n-x)!, dove n! si legge n fattoriale. Unendo il coefficiente binomiale alla probabilità della n-upla abbiamo **la funzione di probabilità** della v.c binomiale.
La v.c binomiale ha anche la media, la varianza e la deviazione standard:

- La media informa sul numero atteso di successi delle n prove: E(X)= np
- La varianza e la deviazione standard misurano la dispersione del numero di successi introno al valore medio atteso. In particolare la deviaizone standard ci dice di quanto il numero di successi si discosta dal numero medio atteso. V(X)=np (1-p). SD(X)= radice quadrata varianza

7- Esporre le caratteristiche, le proprietà e l'utilità della variabile casuale Normale = domanda 5.

8- Esporre e discutere comparativamente vantaggi e svantaggi di una rilevazione campionaria rispetto ad una rilevazione censuaria.

Più spesso non si dispone dell'intera popolazione ma solo di dati parziali relativi ad un sottoinsieme di U campione. L'obiettivo dell'inferenza statistica è estendere l'analisi del comportamento di X all'intera U, in quanto si tratta di inferire dal campione all'intera popolazione. La rilevazione completa di U si chiama censimento mentre quando l'osservazione di X avviene solo su una parte di U si effettua una rilevazione campionaria. Le rilevazioni campionarie sono più frequenti e preferibili rispetto al censimento per ragioni di budget, in quanto una rilevazione campionaria richiede risorse ridotte in termini di tempo e costi, per ragioni di precisione e In alcuni casi la rilevazione parziale si impone rispetto alla rilevazione esaustiva perché quest'ultima risulta impossibile o sconveniente. Il campione deve essere rappresentativo, cioè è un'immagine in scala ridotta ma fedele dell'intera U, deve essere casuale, cioè scelto a caso da U stessa. La casualità del campione è garanzia della sua rappresentatività. Quando scegliamo la rilevazione campionaria dobbiamo sapere che andiamo incontro all'errore campionario, il quale però può essere controllato e misurato tramite la probabilità

9- Esporre il concetto di Inferenza Statistica e discutere il ruolo della casualità del campione

Più spesso non si dispone dell'intera popolazione ma solo di dati parziali relativi ad un sottoinsieme di U campione. L'obiettivo dell'inferenza statistica è estendere l'analisi del comportamento di X all'intera U, in quanto si tratta di inferire dal campione all'intera popolazione. La rilevazione completa di U si chiama censimento mentre quando l'osservazione di X avviene solo su una parte di U si effettua una rilevazione campionaria. Le rilevazioni campionarie

sono più frequenti e preferibili rispetto al censimento per ragioni di budget, in quanto una rilevazione campionaria richiede risorse ridotte in termini di tempo e costi, per ragioni di precisione e In alcuni casi la rilevazione parziale si impone rispetto alla rilevazione esaustiva perché quest'ultima risulta impossibile o sconveniente.
U è ignoto: non abbiamo tutte le info di tutte le unità statistiche e disponiamo solo dei dati campionari. L'inferenza statistica è un'inferenza induttiva (dal particolare al generale) che procede dal campione (una parte) alla popolazione (il tutto) ed è a rischio di errore, ma nonostante questo controlla effetto del caso attraverso la probabilità. Il campione deve essere rappresentativo, cioè è un'immagine in scala ridotta ma fedele dell'intera U, deve essere casuale, cioè scelto a caso da U stessa. La casualità del campione è garanzia della sua rappresentatività.
Con inferenza ci troviamo in una situazione casuale in quanto vi è sempre una componente di incertezza. Ma possiamo gestirla in quanto disponiamo della teoria delle probabilità. Inoltre si parla di "caso" in quanto E non è imprevedibile a priori con certezza, è nota solo una parte delle circostanze che determinano E (un esempio sono i giochi di azzardo,se lanciamo moneta non sappiamo con certezza che cosa uscirà).

10- Esporre e discutere i concetti di campionamento, variabilità campionaria ed errore campionario

- CAMPIONAMENTO: è l'operazione di scelta casuale del campione di n unità statistiche fra le N che compongono l'intera U. Il numero n è detto ampiezza campionaria: di solito è scelto a priori ed è molto più piccolo di N (n<N). Il campionamento è allora un esperimento casuale in U: estrazione causale di unità statistiche. Ci sono molti modi per effettuare un campionamento e ne loro insieme formano una branchia della statistica: la teoria dei campioni. Noi abbiamo visto il tipo più semplice di campione casuale che chiamiamo bernoulliano.

- VARIABILITA' CAMPIONARIA: il campione è una parte della popolazione scelta casualmente e dalla stessa U sono estraibili molti diversi campioni. La casualità del campione è una garanzia della sua rappresentatività ma introduce anche incertezza. Ciascuno dei differenti campioni estraibili da U può darci un'immagine più o meno fedele di U perché fornisce un'informazione parziale e potenzialmente differente circa il comportamento su U del fenomeno.

- ERRORE CAMPIONARIO: l'inferenza statistica comporta necessariamente incertezza e rischio di errore, in quanto si utilizzano solo i dati noti di un campione estratto tra i tanti possibili. L'errore campionario è controllato e misurato con le probabilità.

11- Esporre e discutere il concetto di Campione Bernoulliano e di estrazione con e senza reinserimento; enunciare (a parole) le condizioni applicative in cui possono considerarsi equivalenti.

Il campione bernoulliano è il risultato di n estrazioni casuali da U condotte tutte nelle stesse condizioni, cioè tra loro indipendenti. Si tratta di effettuare n estrazioni con reinserimento fra le N unità di U tra loro equiprobabili. Se il campione è estratto senza reinserimento si chiama campione casuale semplice. Un campione bernoulliano è diverso da un campione casuale semplice perché può contenere duplicazioni. Se n è "sufficientemente grande" e allo stesso tempo è piccolo rispetto a N, le due tecniche con o senza reinserimento portano a risultati equivalenti: tutti gli strumenti di inferenza statistica che richiedono un campione bernoulliano si possono applicare anche a campioni senza reinserimento perché tendono a produrre risultati equivalenti. Infatti quando estraiamo una unità da una popolazione molto grande, se la reinseriamo in U prima di effettuare un'altra estrazione, la possibilità di ri estrarla è molto piccola, quasi 0. Allo stesso modo, se non la reinseriamo la possibilità di estrarre una qualunque delle rimanenti rimane praticamente invariata.

12- Esporre (a parole) la formalizzazione della variabilità campionaria col sistema delle n variabili casuali "osservazione campionaria" indipendenti e identiche al fenomeno X nella popolazione.

Quando si dispone di dati campionari la distribuzione del fenomeno di interesse U e i reali valori delle sue sintesi statistiche sono ignoti e li chiamiamo parametri, i quali sono l'oggetto dell'inferenza statistica. (media e varianza) L'esperimento casuale di campionamento fornisce solo n osservazioni del fenomeno e dunque solo n possibili valori della v.c X. Il campione bernoulliano viene indicato con la n-upla di valori: x1…xi…xn. Ciascuna **osservazione campionaria** xi è il risultato di un esperimento casuale; è quindi un evento casuale e può coincidere con uno qualunque dei possibili valori della v.c X. Allora anche il risultato di ogni estrazione campionaria è interpretato da una v.c Xi, che chiamiamo **v.c estrazione campionaria** di cui l'osservazione campionaria xi rappresenta uno dei possibili valori. Poiché nel campione bernoulliano le estrazioni sono indipendenti allora le v.c estrazioni campionarie Xi sono tra loro indipendenti. Infine ciascuna v.c estrazione campionaria Xi è identica a X, poiché Xi può coincidere con uno qualunque dei possibili valori del fenomeno. Essendo identica ha la stesa media e la stessa varianza.

13- Esporre (a parole) i concetti di Stima e Stimatore (puntuali) precisandone il ruolo e l'utilità per l'inferenza statistica.

STIMA PUNTUALE: calcola un unico valore puntuale per sostituirlo all'ignoto parametro. Controlla in termini di probabilità se e quanto la sostituzione è affidabile e accurata. La stima puntuale di un ignoto parametro è una qualche funzione dei dati campionari x1…xi…xn . La stima di un parametro è quindi il risultato di un calcolo per ottenere un unico numero da sostituire al parametro in U.Noi abbiamo imparato a stimare i 3 parametri ignoti più semplici, ovvero la media del fenomeno in U(che corrisponde alle media μ in X) , la varianza del fenomeno in U(""), una percentuale di valori di X di interesse che indichiamo con p. L'errore campionario assume l'aspetto di errore di stima: quanto più piccolo è l'errore di stima, tanto più precisa, accurata e affidabile è la stima. Per controllare l'errore di stima dobbiamo tener conto di tutti i possibili risultati ottenibili da tutti i possibili campioni.

STIMATORE: è la stessa funzione che definisce la stima ma applicata alle v.c. estrazioni campionarie X1…XI…Xn. Oggetto teorico interprete della variabilità campionaria. Lo stimatore è quindi una v.c. che interpreta tutti i possibili valori della stima su tutti i possibili campioni estraibili. La stima calcolata sul campione estratto è uno dei possibili valori dello stimatore

La stima è un numero ottenuto sul campione effettivamente estratto, fa riferimento ad un solo campione; lo stimatore è una variabile casuale che tiene conto di tutte le possibili stime ottenibili su tutti i possibili campioni estraibili. Serve per interpretare la variabilità campionaria e per controllare l'errore campionario. Fa riferimento all'intero spazio campionario.

14- Definire (a parole) la proprietà di non distorsione per uno stimatore puntale. Discuterne l'utilità per l'inferenza statistica. Quali stimatori non distorti conoscete?

La più nota e semplice proprietà richiesta a uno stimatore è detta non distorsione, la quale riguarda il valore atteso dello stimatore. Uno stimatore è non distorto se il suo valore atteso coincide con il parametro oggetto di stima; se questo non succede lo stimatore è distorto. Fra tutti i possibili campioni ve ne sono alcuni che forniscono una sotto-stima del parametro, altri una sovra-stima e altri ancora forniscono valori identici o vicini al parametro oggetto di stima. Richiedere che uno stimatore sia non distorto significa garantire che sovra e sotto stime si compensino su totale dei campioni estraibili e che in media lo stimatore coincida con ciò che si vuole stimare. La non distorsione si può accettare solo teoricamente, poiché lo stimatore è un oggetto teorico. Lo stimatore media campionaria è uno

stimatore non distorto per μ perché il suo valore atteso è proprio uguale a μ. Anche lo stimatore frequenze relativa(o percentuale) è non distorto

15- Esporre il problema della stima non distorta della varianza, descrivere (a parole) come si ottiene la varianza campionaria corretta e discutere il concetto di Gradi di Libertà.

Sulla base dei soli dati disponibili, cioè il campione bernoulliano, la stima più naturale per la varianza di U è la varianza del campione: sommatoria per i che va da 1 a n di (xi-x segnato)^2/n. In questo caso però il corrispondente stimatore è distorto per la varianza , cioè ha valore atteso che non coincide con ciò che si vuole stimare e ha tendenza a sotto stimare. Per ottenere uno stimatore non distorto basta dividere per n-1 anzichè per n nel calcolo della varianza del campione. Questa stima è chiamata varianza campionaria corretta e la indichiamo con s^2. La quantità n-1 è chiamata gradi di libertà. Utilizzando i gradi di libertà si ottiene una stima non distorta e consistente, cioè l'errore di stima che si commette stimando la varianza con s^2 diminuisce al crescere dell'ampiezza campionaria, ma questa riduzione è lenta e per ottenere stime sufficientemente precise occorrono campioni più grandi. Se si vuole stimare la deviazione standard anziché la varianza del fenomeno U bisogna ricordare che radice quadrata di s^2 in generale è distorta per la deviazione standard e unico modo per correggerla è aumentare ampiezza del campione n

16- Esporre il problema della stima puntuale della % (frequenza relativa) di una categoria di interesse (successo) per un fenomeno dicotomico, descrivere lo stimatore Frequenza relativa Campionaria e discuterne le proprietà inferenziali.

L'oggetto della stima è la percentuale di unità statistiche o casi che ,fra tutte quelle che compongono la U di riferimento, è classificabile in una data categoria. Scelta l'ampiezza campionaria n, si estrae da U un campione bernoulliano. Il risultato sarà un insieme di unità statistiche classificabili o non classificabili nella categoria che ci interessa. La stima più naturale per l'ignota frequenza relativa p è la corrispondente frequenza relativa del campione, ovvero la frequenza relativa campionaria che si indica con p^. X può assumere due valori 0 e 1: assume il valore 1 in corrispondenza di soggetti classificabili nella categoria di interesse, mentre assume valore 0 in caso contrario. La somma dei dati campionari ci da il numero di soggetti che fra gli n estratti sono classificabili nella categoria che ci interessa. Dividendo questa somma per l'ampiezza del campione si ottiene la stima. La stima p^ ha la stessa formula della media campionari. Per ciascun soggetto estratto possiamo chiamare successo il fatto di essere classificabile nella categoria di interesse e insuccesso il fatto di non essere classificabile. La somma dei dati campionari xi è quindi il numero di successi su n prove indipendenti, cioè uno dei possibili valori di una v.c binomiale di parametri n e p. il corrispondente stimatore lo indichiamo con P^ : Bin(n,p)/n. La v.c binomiale ha valore atteso pari a np e varianza pari a np(1-p). Possiamo quindi dire che la stima p^ è non distorta per p, perché il corrispondente stimatore ha valore atteso uguale a p , cioè P^ è stimatore non distorto per p. Il suo MSE coincide con la varianza. Lo stimatore frequenza relativa campionaria P^ è consistente per p ed è anche il più efficiente fra tutti gli stimatori non distorti per p.

17- Esporre (a parole) e discutere l'utilità della relazione fra lo stimatore Percentuale campionaria e la variabile casuale Binomiale

Per ciascun soggetto estratto possiamo chiamare successo il fatto di essere classificabile nella categoria di interesse e insuccesso il fatto di non essere classificabile. La somma dei dati campionari xi è quindi il numero di successi su n prove indipendenti, cioè uno dei possibili valori di una v.c binomiale di parametri n e p. il corrispondente stimatore lo indichiamo con P^ : Bin(n,p)/n. La v.c binomiale ha valore atteso pari a np e varianza pari a np(1-p). Possiamo quindi dire che la stima p^ è non distorta per p, perché il corrispondente stimatore ha valore atteso uguale a p , cioè P^ è stimatore non distorto per p. Il suo MSE coincide con la varianza. Lo stimatore frequenza relativa campionaria P^ è consistente per p ed è anche il più efficiente fra tutti gli stimatori non distorti per p.

18- Definire (a parole) l'errore quadratico medio e discuterne l'utilità esemplificando con il caso dello stimatore Media campionaria

L'errore quadratico medio è un modo per esprimere in formule l'errore campionario associato all'inferenza nel processo di stima , cioè l'errore di stima. Misura quanto lo stimatore è preciso, quanto è vicino all'ignoto parametro che si vuole stimare. Il punto di partenza è la differenza : (x segnato – μ). Tale errore può risultare positivo su alcuni campioni (sovra stime) o negativo su altri (sotto stime). Eleviamo al quadrato la differenza per eliminare l'influenza del segno e infine consideriamo l'errore medio di stima mediando su tutti i possibili campioni estraibili: E(x segnato – μ). Indicheremo questa quantità con MSE, errore quadratico medio. Quest'ultimo quindi è il valore atteso della differenza al quadrato tra lo stimatore e il parametro che si vuole stimare; è una quantità teorica che misura la **dispersione** dei valori dello stimatore intorno all'oggetto di stima. Quanto è più piccola tale dispersione tanto più è preciso e accurato lo stimatore. L'MSE è formato sia dalla sua varianza sia dalla sua distorsione al quadrato, quindi se la distorsione è =0 l'MSE coincide con la varianza.
Nel caso della media campionaria lo stimatore è non distorto e quindi MSE coincide con la varianza. Poi calcoliamo al varianza della media campionaria: varianza fenomeno in U/n, quindi MSE=V. Tanto più la V è grande , maggiore sarà la differenza nella stima che ciascun campione produce e non è una buona cosa. Se il fenomeno è molto variabile è difficile stimare qualcosa infatti:

- Maggiore è la variabilità , più siamo a rischio di errore
- Più il campione è grande più riduco l'errore di stima (basta ragionare sulla formula V/n)
- Errore di stima dipende sia da n che dalla varianza quindi vi deve essere un bilanciamento tra queste due cose.

19- Definire (a parole) lo Standard Error di uno stimatore, discuterne l'utilità ed esemplificare con il caso della Media campionaria

L' MSE è una misura teorica , è quadratico cioè misura l'errore di stima prendendo le differenze fra stimatore e parametro elevate al quadrato: questo produce effetti collaterali. Per rimediare prediamo la radice quadrata dell'MSE che è una misura teorica dell'errore medio di stima con la stessa unità di misura e lo stesso ordine di grandezza. La stima dell'errore medio di stima è detta **standard error** dello stimatore e lo indichiamo con SE = radice quadrata della V.
Nel caso della media campionaria , poiche X è stimatore non distorto si tratta di stimare la radice quadrata di V = radice quadrata della varianza/n, stimando il parametro(varianza) ignoto con la varianza campionaria corretta, quindi la formula è: radice quadrata di s^2/n.
SE è u numero calcolato sul campione che stima l'errore medio che si commette sostituendo all'ignoto parametro la stima calcolata sul medesimo campione.

20- Definire lo Standard Error di uno stimatore, discuterne l'utilità ed esemplificare con il caso della Frequenza relativa campionaria

21- Discutere comparativamente i vantaggi e gli svantaggi di una stima intervallare rispetto ad una stima puntuale.

La stima puntuale consiste nell'elaborazione di dati campionari per produrre un unico valore da sostituire al parametro ignoto.

Vantaggi:

- È generale e sempre calcolabile per ogni tipo di fenomeno (solo con dati campionari) non richiede info ausiliarie
- È semplice perché si procede per analogia in quanto: la media mu si stima con la media del campione x segnato, la varianza si stima con la varianza del campione s^2 e la fr relativa si stima con al corrispondente fr relativa campionaria p^

Svantaggi:

- è difficile avvicinarsi all'ignoto valore del parametro con un unico valore puntuale
- affidabilità della stima puntuale risiede nella garanzia probabilistica offerta dalle proprietà del corrispondente stimatore. L'errore medio di stima a livello pratico lo si può solo stimare con lo standard error e utilizzando gli stessi dati campionari

La stima intervallare è un insieme di valori (intervallo) calcolato dai dati campionar,i con una pre scelta % di affidabilità ,che contenga l'ignoto valore del parametro (intervallo di confidenza).

Vantaggi:

- ci da un informazione + ampia e meno precisa
- + facile da azzeccare il parametro ignoto, c'è un intero intervallo
- Affidabilità della stima intervallare è quantificata a priori con una probabilità al livello che più ci piace e ci conviene

Svantaggi:

- Aumento della complessità della procedura di stima servono più informazioni oltre ai dati campionari. Servono infatti info ausiliarie a priori
- Le info a priori non sempre sono facili da reperire , quindi sono ipotizzabili con il rischio di basarci su un 'ipotesi azzardata e lontana dalla realtà

22- Esporre e discutere le Informazioni ausiliarie (a priori) per la costruzione di Intervalli di Confidenza e l'impatto sul livello di confidenza.

L' intervallo di confidenza 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. In questo caso però sono necessarie informazioni ausiliarie a priori che a volte sono facili da reperire mentre altre volte sono solo ipotizzabili con il rischio di basarci su un'ipotesi azzardata e molto lontana realtà. Un intervallo di confidenza non è sempre producibile sulla base dei soli dati campionari, ma è calcolabile soltanto se abbiamo info ausiliarie.
Il livello di confidenza è una misura di quanto possiamo fidarci che l'intervallo di confidenza contenga l'ignoto valore del parametro. Quindi si dovrà scegliere la probabilità di sbagliare che si indica con alfa, mente la probabilità di fare bene (che è appunto il livello di confidenza) si indica con 1-alfa. A seconda dell'info ausiliaria che scelgo avrò diversi valori di alfa e quindi un livello di confidenza diverso.

23- Esporre (a parole) la metodologia per la costruzione di un Intervallo di Confidenza per la media di Popolazione Normale, definendo e discutendo il livello di confidenza

Il fenomeno di interesse in U è ben interpretato da una v.c Normale con media mu ignota ma varianza nota. X-N(mu,

varianza nota). In questo caso siamo in una popolazione normale con l'informazione circa il valore della varianza, in quanto appunto è nota (scarsa realisticità del caso). Sotto queste condizioni vogliamo costruire una stima intervallare per l'ignoto parametro mu. Un teorema di teoria delle probabilità ci garantisce che se X è normale anche lo stimatore media campionaria X segnato lo è , con media mu e varianza varianza/n. Quest'ultima è un info ausiliaria a priori, in quanto abbiamo detto che la varianza è nota. Standardizzando si ottiene allora la v.c Z normale standard, della quale calcoliamo la probabilità di qualunque intervallo utilizzando le tavole della Z : Z-N(0,1). La metodologia di costruzione prevede 5 passi:

1. si estrae un campione bernoulliano di ampiezza n e ci si procura i dati campionari.

2. si calcola la stima puntuale per mu, cioè la media del campione

3. si sceglie la probabilità di sbagliare ,cioè di costruire un intervallo di confidenza che non contiene mu. Indichiamo questa probabilità con la lettera alfa, e quindi la probabilità di fare bene con 1-alfa, cioè la probabilità di costruire un intervallo di confidenza che contiene l'ignoto parametro mu. Scegliere la probabilità di sbagliare=0 non è una buona idea: il rischio di errore campionario esiste sempre ed è ineliminabile. Nella pratica alfa è fissato a un livello standard di 0.05 oppure di 0.1 o 0.01. cosi la probabilità di fare bene può essere il 95%, 90% oppure il 99%

4.abbiamo l'info a priori della varianza nota e quindi: P(a<= X segnato – mu/radice quadrata varianza nota/n <=b = P(a<= Z<=b))= 1-alfa. Sappiamo che per la normale le probabilità sono aree. All'interno dell'intervallo (a,b) c'è una probabilità pari a 1-alfa mentre all'esterno c'è la probabilità alfa che dividiamo in alfa/2 a sinistra e alfa/2 a destra. Gli estremi di tale intervallo sono due valori della Z-N(0,1) simmetrici rispetto allo 0, li indichiamo con –zalfa/2 e zalfa/2 e li chiamiamo Z-score. Troviamo quello positivo dalle tavole, mentre i negativo si ottiene cambiando il segno. Possiamo riscrivere la probabilità cosi :

P(a<= Z <=b)=P(-Zscore<= Xsegnato – mu/radice quadrata varianza nota/n <= Zscore)= 1-alfa.

Infine si inverte questa relazione probabilistica in modo da ottenere un intervallo centrato sul parametro mu che si vuole stimare. Quindi mu sarà compreso tra lo stimatore meno lo z score moltiplicato per la radice quadrata della varianza nota /n e lo stimatore + "". Nel caso di popolazione normale questa probabilità è vera. A tale intervallo viene associato il numero 1-alfa ovvero il livello di confidenza che è una misura di quanto possiamo fidarci che l'intervallo di confidenza contenga l'ignoto valore del parametro

5. sostituisco i dati del mio campione, quindi il valore della stima calcolata sul campione estratto
LEGGO ANCHE NEL CASO DI POP NORMALE CON VARIANZA IGNTA (STUDENTIZZO CON LA T DI STUDENT E USO VARIANZA CAMPIONARIA CORRETTA CON GRADI DI LIBERTà,)

24- Esporre (a parole) la metodologia di costruzione di un Intervallo di Confidenza approssimato per Grandi Campioni, precisando quali informazioni ausiliarie sono necessarie e quali sono le conseguenze sulla validità dell'inferenza

Costruire un Intervallo d confidenza richiede delle informazioni in più. Se non abbiamo info ausiliarie a priori su X cioè se non siamo nel caso di popolazione normale, dobbiamo allora avere molti dati cioè essere nel caso di grandi campioni. Solo se il campione è sufficientemente grande possiamo appellarci a un teorema di teroai delle probabilità fondamentale nell'inferenza statistica. Questo teorema si chiama teorema centrale del limite, TCL. Se l'ampiezza campionari n tende all'infinto allora gli stimatori standardizzati media campionaria Xsegnato e frequenza relativa campionaria P^ sono normali. Questo è il risultato teorico. Nella pratica i campioni possono essere grandi ma non infiniti; quando n è sufficientemente grande gli stimatori media e percentuale standardizzati sono

approssimativamente normali. Si parlerà quindi di intervalli di confidenza **approssimati** per grandi campioni con un livello di confidenza a**pprossimativamente** pari all'1-alfa scelto. Gli IC approssimati per grandi campioni si possono usare per la media mu quando non si può assumere la normalità della popolazione (pag 265 formule). Si userà s^2 varianza campionaria corretta con i gradi di liberta. Uso lo z score.

25- Definire (a parole) la "precisione" di un intervallo di confidenza e discutere come può essere controllata attraverso la relazione con livello di confidenza e numerosità campionaria

Un IC è tanto più preciso quanto più è stretto, cioè quanto meno è ampio. L'ampiezza di un IC ne definisce la precisione e viene calcolata come la differenza fra estremo superiore e inferiore. Un IC ha una struttura generale di questo tipo: stima(puntuale) + o – score x SE(stima) dove il valore di score sarà uno Z-score o un T-score. La precisione di un IC è in relazione inversa con il livello di confidenza e in relazione diretta con la numerosità campionaria:

- A parità di ampiezza campionaria n, un aumento del livello di confidenza 1-alfa provoca una diminuzione di precisione cioè un aumento dell'ampiezza dell'IC e viceversa
- A parità di livello di confidenza au aumento della numerosità campionaria n provoca un aumento della precisione cioè una diminuzione dell'ampiezza dell'IC e viceversa. Inoltre aumentando n diminuisce SE

26- Esporre e discutere i concetti di Ipotesi Statistica, Ipotesi Nulla e Test statistico, a 2 e a 1 coda.

- IPOTESI STATISTICA: è una congettura riguardante una qualche caratteristica del fenomeno in U. tale congettura è formata a priori cioè prima di estrarre il campione. Proviene dall'esterno dipende dal contesto applicativo e dagli obiettivi di ricerca e non dai dati campionari. Può riguardare il valore di un parametro di U , ad esempio la media mu ,oppure una fr relativa etc.

- IPOTESI NULLA: è la formalizzazione ,cioè la traduzione in simboli e formule dell'ipotesi statistica che abbiamo emesso e che vogliamo sottoporre a verifica con un test statistico. La indicheremo con la notazione standard H0.

La verifica delle ipotesi è la metodologia inferenziale che partire dai dati campionari porta a decidere se accettare o rifiutare l'ipotesi nulla controllando probabilisticamente l'errore campionario. Il TEST STATISTICO è la regola pratica che porta a questa decisione, in particolare abbiamo visto il test di significatività

Tra i test statistici vi sono Test a 1 o a 2 code.
Test a 1 coda: test t
Statistico per verificare un' ipotesi unilaterale, cioè un'ipotesi nulla del tipo H0 : mu<mu0 (oppure >); per verificare ipotesi nulla unilaterale si pone la ragione critica (sotto la/le code) sotto un'unica coda della statistica test, quella più lontana dall'ipotesi nulla e si esegue appunto un test a una coda utilizzando un T-test studendizzando la statistica test. Quindi oltre alla differenze xsegnato – mu0 vicine allo 0 ,anche tutte le differenze negative depongono a favore dell'accettazione di H0: mu<muo, mentre le differenze positive e troppo grandi si trovano nella regione critica. Quindi per verificare l'ipotesi unilaterale H0 : mu<=mu0 si usa un T-test a 1 coda ponendo la regione critica tutta sotto la coda di destra mentre le coda di sinistra farà part della zona di accettazione. Non è necessario dividere la probabilità di sbagliare alfa per due , ovvero in alfa/2 sotto una coda e alfa/2 sotto un'altra, in quanto la regione critica è composta da una sola coda di probabilità alfa. Nel caso in cui l'ipotesi unilaterale è del tipo H0 : mu>= mu0 si ribalta il ragionamento.

Test a 2 code: test statistico per verificare ipotesi bilaterale. Ha la ragione critica formata dalle due zone sotto el due code della statistica test, ciascuna di probabilità alfa /2. L'ipotesi nulla è del tipo H0 : mu= mu0 (mentre nel caso di ipotesi unilaterale avevamo >= oppure <=). Se il test porta all'accettazione di H0 si conclude che mu è uguale al valore

ipotizzato mu0 a livello di significatività 1-alfa. Se invece il test porta al rifiuto di H0 si conclude che mu è diversa da mu0 con probabilità di sbagliare pari ad alfa.

27- Esporre e Discutere il concetto di misura & controllo dell'errore campionario per un Test statistico.

L'ipotesi nulla H0 può essere vera per X su U, oppure falsa. Il test statistico che utilizziamo e che ci porterà a stabilire se H0 è vera o falsa e quindi se accettarla o no è basato su dati campionari, cioè su un'osservazione parziale dell'intera U di riferimento. È dunque condotto in condizioni di incertezza. Accettare o rifiutare H0 sulla base dei dati campionari comporta inevitabilmente il rischio di commettere un errore. Si possono commettere due tipi di errori, l'errore di prima specie, che consiste nell'errato rifiuto di H0, e quello di seconda specie che è l'errata accettazione di H0. Noi ci siamo occupati solo del primo. Con un test di significatività si sceglie a priori la probabilità di commettere l'errore di prima specie, basta che questa probabilità non la scegliamo =0 perché il rischio di errore esiste sempre e non può essere nullo; la indichiamo sempre con alfa: alfa= P(rifiutare H0 I H0). 1- alfa sarà la probabilità di fare bene, cioè di non sbagliare accettando H0 perché H0 è vera, e questa probabilità è chiamata livello di significatività di un test statistico: alfa= P(accettare H0 I Ho)

28- Definire e discutere i concetti di accettazione/rifiuto dell'ipotesi nulla, errore di I specie e livello di significatività di un test statistico.

Accettare o rifiutare H0 sulla base dei dati campionari comporta inevitabilmente il rischio di commettere un errore. Si possono commettere due tipi di errori, l'errore di prima specie, che consiste nell'errato rifiuto di H0, e quello di seconda specie che è l'errata accettazione di H0. Noi ci siamo occupati solo del primo. Con un test di significatività si sceglie a priori la probabilità di commettere l'errore di prima specie, basta che questa probabilità non la scegliamo =0 perché il rischio di errore esiste sempre e non può essere nullo; la indichiamo sempre con alfa: alfa= P(rifiutare H0 I H0). 1- alfa sarà la probabilità di fare bene, cioè di non sbagliare accettando H0 perché H0 è vera, e questa probabilità è chiamata livello di significatività di un test statistico: alfa= P(accettare H0 I Ho). La probabilità di errore di prima specie alfa è in genere fissata a uno dei livelli standard 0.05, 0.1, 0.01; conseguentemente il livello di significatività sarà 95%, 90% o 99%.

29- Esporre la metodologia per la costruzione di un test statistico per ipotesi bilaterale sulla media di una Popolazione Normale

La condizione di popolazione normale non è difficile da riscontrare nella realtà. E' meno probabile invece avere un caso in cui la varianza è nota: se non abbiamo info su mu è molto probabile che manchino anche le info sulla varianza e quindi dovrà essere stimato anche questo parametro. Il fenomeno X è interpretabile con la v.c normale con entrambi i parametri mu e varianza ignoti. L'obiettivo è verificare l'ipotesi nulla H0: mu=mu0. La metodologia di costruzione si articola in 6 passi:

1. si estrae il campione bernoulliano di ampiezza n e si ottengono dati campionari.

2.si calcola la stima puntuale per tutto ciò che è ignoto, dunque entrambi i parametri. La media del campione xsegnato per mu e la varianza del campione corretta con i gradi di libertà per garantirci la non distorsione (quindi s^2, dividendo (xi – x segnato)^2 per n-1.

3. si sceglie il livello di significatività 1-alfa da cui si ricava la probabilità di sbagliare alfa e la probabilità delle due code

alfa/2.

4.siamo nel caso di pop normale quindi abbiamo la normalità della media campionaria , ma ci manca il valore della varianza per standardizzare. Usiamo quindi la stima non distorta s^2 e studentizzare. Effettuiamo la studentizzazione sempre sotto H0 cioè usando il valore ipotizzato mu0 al posto dell'ignoto valore vero mu. La statistica test che otteniamo è una T di student con n-1 gradi di libertà : Xsegnato – mu0/ radice quadrata di S^2/n = Tn-1. (lettere maiuscole perchè usiamo gli stimatori). Per ottenere il valore critico si tratterà di ricavare il T-score talfa/2 cercandolo sulle tavole della T. Con il valore critico e con il suo simmetrico, che si ottiene semplicemente cambiando segno ,possiamo individuare sotto la curva a campana della T di student la zona di accettazione e la regione critica del test

5. si calcola il valore sperimentale sostituendo nella statistica test i valori coti cioè mu0 e n e le due stime media del campione e s^2. Si ottiene cosi un numero: xsegnato – mu0/ rq di s^2/n

6.il test. Si rifiuta H0: mu=mu0 a livello 1-alfa se il valore sperimentale cade nella regione critica quindi se è <o = -talfa/2 oppure se è >=talfa/2. Nel caso in cui i gradi di libertà sono tanti (più di 30) si può usare lo Z-test anche se la varianza è ignota.

30- Esporre la metodologia di costruzione di un test approssimato per Grandi Campioni precisando quali informazioni ausiliarie sono necessarie e quali sono le conseguenze sulla validità dell'inferenza

Siamo in una situazione in cui non si dispone di info ausiliare a priori , non si sa nulla circa il fenomeno in U oppure non si ritiene realistica l'ipotesi che la popolazione sia normale oppure si sa che non lo è. In mancanza di info ausiliare a priori è necessario compensare con una quantità di dati campionari sufficientemente grande e si parla teoricamente di grandi campioni. Solo se il campione è sufficientemente grande possiamo applicare il TCL e recuperare la normalità degli stimatori per la media mu e la frequenza relativa p. Quando usiamo il TCL per grandi campioni stiamo usnado risultati approssimati. Non si ha quindi la normalità della popolazione ma, ma si tratterà di test approssimati per grandi campioni. La conseguenza è che l'effettivo livello di significatività è approssimativamente il valore 1-alfa scelto, e più aumenta l'ampiezza campionaria n più ne saremo vicino. Il test approssimato per grandi campioni è sempre Z-test anche quando il parametro varianza è ignoto. (pag 295 formule).
Nel caso in cui il fenomeno di interesseè qualitativo, cioè categoriale, dicotomico o ordinale il parametro ignoto di inferenza è la fr relativa p di soggetti che in U sono classificabili nella categoria che ci interessa e che chiamiamo successo. L'ipotesi nulla sarà: H0: p=p0 (bilaterale), H0: p<= oppure >= p0 (unilaterale). La procedura di costruzione del test si articola in 6 passi:

1.dati campionari di un campione bernoulliano di ampiezza n

2. si calcola la stima puntuale per p e la sua stima è la corrispondente frequenza p^ del campione

3. si sceglie il l.s del test 1-alfa da cui si ricava la probabilità di errato rifiuto alfa (e per test a 2 code la probabilità delle code alfa/2)

4.la statistica test si ottiene con la standardizzazione sotto H0. Essendo nel caso di grandi campioni la statistica test è approssimativamente una Z-N(0,1). Il valore critico sarà uno Z-score da cercare sulle tavole della Z (zalfa/2 per test a due code, zalfa per test a una coda) pag 296 formule

5.il valore sperimentale si calcola sostituendo nella statistica test i valori noti e le stime campionarie

6. per costruire il test come regola di rifiuto ci ricordiamo che stiamo lavorando con un test approssimato per grandi campioni con probabilità di sbagliare (cioè di rifiutare ipotesi che invece è vera) approssimativamente pari a alfa, se il valore sperimentale cade nella regione critica. Quindi si rifiuta se il vs <= -zalfa/2 o >=zalfa/2 nel caso el test a 2 code e se vs>= zalfa nel caso del test a 1 coda.

31- Esporre il concetto, l'utilità e la corretta interpretazione di p-value per l'esecuzione di test statistici a 1 o a 2 code.

Il p-value è un numero prodotto dal computer con il quale possiamo decidere se accettare o rifiutare H0 qualunque sia il livello di significatività che vogliamo fissare. Il p-value è una probabilità , un numero compreso tra 0 e 1, quindi un 'area sotto la curva della statistica test. È il minimo livello α per rifiutare H0. Se il p-value risulta più piccolo del livello prescelto α (per un test ad una coda) o di α/2 (per un test a due code) allora si rifiuta H0. Si stanno confrontando delle probabilità

Il p- value dipende solo dal valore sperimentale del test, cioè sui dati campionari e dunque rimane sempre lo stesso a qualunque livello di significatività, mentre il valore critico dipende sempre dall' alfa scelto ed è diverso per diversi livelli di significatività . Quando il p-value è molto piccolo, si rifiuta H0 praticamente a qualunque livello di significatività e si parla di test non significativo. Quando si rifiuta H0 il test è non significativo si conclude che fra i due fenomeni esiste relazione statistica.

Quando si esegue il test "a mano" si decide se accettare o rifiutare H0 confrontando due valori: il valore sperimentale e il valore critico. Nel grafico questi due valori stanno sull'asse delle ascisse e il valore critico si recupera dalle tavole una volta scelto il l.s. se si esegue il test al computer si decide se accettare o rifiutare H0 confrontando due probabilità: il p-value e il livello alfa o alfa mezzi. Le due procedure sono equivalenti , infatti quando p-value<alfa significa che il valore sperimentale cade nella regione critica.