La statistica inferenziale

INDICE

Prima parte

La statistica e le variabili casuali

CAPITOLO 1: LA STATISTICA E L'INDAGINE STATISTICA

1. LA STATISTICA

La *statistica* è l'arte e la scienza di raccogliere, analizzare, interpretare e presentare dati nell'intento di trasformarli in un'informazione utile.

Esistono due rami principali della statistica:

- *Statistica descrittiva;*
- *Statistica inferenziale.*

2. LA STATISTICA DESCRITTIVA

In *statistica descrittiva* si parte dall'assunto che i dati oggetto di studio siano frutto di un'*indagine censuaria*, anche detto censimento.

Il *censimento* è una tipologia di indagine molto lunga, costosa e complessa, in quanto parte dal presupposto di analizzare una popolazione intera di unità.

(ESEMPIO: Censimento che fa l'INPS, circa ogni 5/10 anni, sulla famiglia, della popolazione italiana)

In statistica descrittiva, che appunto, come dice la parola descrive un fenomeno, abbiamo:

- *Obiettivo:* descrivere il fenomeno;
- *Dati:* intera popolazione;
- *Strumenti:* tabelle di indagine e grafici;
- *Risultato:* sintesi statistica di un determinato carattere.

3. LA STATISTICA INFERENZIALE

In *statistica inferenziale* si parte dal presupposto che la popolazione, ossia i nostri *dati,* siano molto numerosi e, quindi, come dato oggetto di studio prendiamo solo una parte di essa, che prende il nome di campione, l'indagine, che quindi andremo a fare, prende il nome di *indagine campionaria,* anche detto campionamento.

(Il campionamento, a differenza del censimento, è un'indagine veloce, economica e semplice, in quanto parte dal presupposto di analizzare solo una parte delle unità che compongono la popolazione).

Nell'inferenza statistica, la quale è la branca di statistica che si occupa di analizzare i dati ottenuti da un campione al fine di stimare alcune caratteristiche di un fenomeno definito sull'intera popolazione di riferimento, abbiamo:

- *Obiettivo:* verificare o stimare un'ipotesi statistica;
- *Dati:* campione casuale;
- *Strumenti:* Variabili casuali, metodi di stima, test di ipotesi, modello di regressione;
- *Risultato:* stima di caratteristiche ignote di un carattere.

Distinguiamo quindi due concetti:

- ❖ Popolazione: Insieme di *unità omogenee* rispetto ad una o più caratteristiche di interesse;
- ❖ Campione: *Sottoinsieme rappresentativo* delle unità statistiche della popolazione di riferimento.

I risultati e le decisioni dell'inferenza statistica comportano un rischio, dovuto alla *limitatezza delle informazioni usate*.

Il rischio può essere *valutato e controllato* in termini probabilistici con gli strumenti dell'inferenza statistica.

La statistica inferenziale consente di ottenere un'informazione sul parametro θ (si legge teta) (informazione ignota della popolazione), attraverso la stima t_n (ossia la stima di teta), che sebbene sia affetta da un errore ε (si legge epsilon) (è *un errore accettabile, controllabile e misurabile* del ricercatore attraverso le scelte metodologiche inerenti alla procedura inferenziale adottata) è comunque utile poiché attenua l'ignoranza su un fenomeno oggetto di studio.

4. L'INDAGINE STATISTICA

L'*indagine statistica* è l'insieme di operazioni mediante le quali si acquisiscono informazioni su uno o più fenomeni attinenti ad una popolazione di riferimento.

Lo scopo dell'indagine statistica è quello di produrre delle *statistiche,* ovvero descrizioni riassuntive di carattere quantitativo riguardanti il collettivo di interesse.

La progettazione e l'esecuzione di un'indagine sono il frutto di un impegno multidisciplinare che coinvolge necessariamente un elevato numero di professionalità.

L'attività di progettazione deve procedere prendendo in considerazione tutti gli aspetti coinvolti, da quelli riguardanti i fenomeni di interesse a quelli di carattere più operativo.

Ma quali sono le fasi?

1. *Definizione del problema* (approccio esplorativo o confermativo);
2. *Raccolta dei dati* (con scelta delle unità e dei caratteri) (N.B se i dati saranno raccolti erroneamente, l'indagine sarà sbagliata);
3. *Scelta del metodo di analisi;*
4. *Interpretazione dell'indagine statistica* (significatività e rilevanza dei risultati);
5. *Conclusioni* (coerenza con il problema posto).

Tutto ciò avviene con un controllo qualità e prevede:

- *Progettazione;*
- *Realizzazione;*
- *Elaborazione;*
- *Presentazione.*

Quando l'indagine avviene osservando tutte le unità della popolazione, si parlerà di indagine *censuaria,* se invece viene fatta analizzando solo una parte di queste unità si parla di indagine *campionaria,* con la quale si userà un *disegno di campionamento*.

Con la prima, ossia quella censuaria, avremo una rilevazione totale del fenomeno, a differenza di quella campionaria con la quale avremo una rilevazione parziale, o meglio, si ottiene *una stima* del fenomeno analizzato.

Spesso e volentieri ci si chiede quale indagine sia la migliore, ma, in realtà, bisogna considerare diversi aspetti:

- *Tempi di rilevazione* (Censuaria=Lunga; Campionaria=Breve);
- *Costi di rilevazione* (Censuaria=Costosa; Campionaria=Economica);
- *Ricchezza dei dettagli della rilevazione* (Censuaria=Poche info; Campionaria=Più info e di qualità);
- *Errori associati alla rilevazione* (Censuaria=Più errori perché troppe unità; Campionaria=Meno errori perché meno unità).

Vedendola così ci sembra che l'indagine campionaria sia migliore, ma ricordiamoci anche che con la campionaria andiamo a realizzare una stima e prendendo in considerazione poche unità, la stima potrebbe essere vicina al valore della realtà come potrebbe non esserlo, ma quasi mai corrisponderà al valore della realtà, invece, con la censuaria avremo meno errori di stima, ciò significa che il nostro valore finale sarà il più possibile vicino o sarà proprio il valore corrispondente alla realtà del fenomeno analizzato.

Abbiamo diverse tecniche per acquisire le informazioni e bisogna considerare:

- *Estensione degli scopi perseguiti;*
- *Ampiezza del collettivo di riferimento;*
- *Costo e budget disponibile;*
- *Grado di precisione considerato.*

Ricorda che osservare è la via più diretta e immediata per studiare i comportamenti e che domandare è la via obbligata per esplorare motivazioni, aspettative, credenze, sentimenti, percezioni e atteggiamenti.

Tante volte si ricorre all'indagine campionaria perché è impossibile attuare un'indagine censuaria, quando, ad esempio, parliamo di popolazioni a circuito chiuso, o quando l'osservazione del fenomeno comporta la distribuzione delle unità che si osservano.

Gli aspetti positivi dell'indagine campionaria sono i *costi bassi, la rapidità nel raccogliere i dati, trattare i dati e pubblicare i risultati, e, infine, la profondità della ricerca.*

Oltre all'errore campionario, ossia gli *errori commessi durante la rilevazione dei dati*, troviamo l'errore non campionario, provocato dalle *imprecisioni e dalle inaccuratezze commesse o subite durante un'indagine* (Esempio: rifiuti a rispondere)

La dimensione ridotta del campione consente una maggiore accuratezza, grazie anche al *personale qualificato* che svolge l'indagine.

Si parla di precisione di stima e di accuratezza. Nel primo caso ricorda che *più è grande il campione più la stima sarà precisa rispetto alla realtà*. Nel secondo caso si basa sui *pochi errori commessi durante il passaggio dei dati su un adeguato supporto di elaborazione* (Esempio: Excel).

Abbiamo quattro casi di indagine campionaria:

- *Inaccurata e imprecisa;*
- *Accurata ma imprecisa;*
- *Inaccurata ma precisa;*
- *Accurata e precisa.*

Il concetto che racchiude in sé sia la precisione di stima che l'accuratezza è rappresentato dall'attendibilità di un'indagine,

ARGOMENTI, SOTTOFORMA DI DOMANDA, DEL CAPITOLO 1

- *Che cos'è la statistica?*
- *Quali tipologie di statistiche abbiamo e in cosa si differenziano?*
- *Cos'è la popolazione?*
- *Cos'è il campione?*
- *Perché la statistica inferenziale comporta un rischio?*
- *Cosa fa la statistica inferenziale nello specifico?*
- *Cos'è l'indagine statistica?*
- *Quali sono le fasi dell'indagine statistica?*
- *Cos'è il controllo qualità dell'indagine statistica e cosa prevede?*
- *Cos'è l'indagine censuaria?*
- *Cos'è l'indagine campionaria?*
- *Cosa si intende per rilevazione parziale e rilevazione totale?*
- *Cosa consideriamo per capire quale indagine sia la migliore?*
- *Personalmente a cosa dobbiamo tenere conto quando scegliamo il metodo di indagine statistica da applicare al fenomeno di ricerca?*
- *Perché è importante osservare e domandare?*
- *Quali sono gli aspetti positivi dell'immagine campionaria?*
- *Cosa si intende per errore campionario?*
- *Cosa si intende per errore non campionario?*
- *Cos'è la precisione di stima?*
- *Cos'è l'accuratezza?*
- *Quali sono i quattro casi di attendibilità di indagine?*

CAPITOLO 2: LE VARIABILI CASUALI

Con le variabili causali (v.c) si rappresentano i risultati possibili di un esperimento aleatorio che consiste nella realizzazione di azioni i cui esiti sono *incerti*, o anche detti *aleatori*.

In generale, una variabile casuale *è una regola che associa ad ogni evento un unico numero reale che assume determinati valori con determinate probabilità*. Quindi si assegna a ciascun valore la variabile che la probabilità può assumere. Ma ricordiamo sempre che la probabilità può assumere un numero relativo tra 0 e 1, quindi, una percentuale tra lo 0% e il 100%.

In generale la *variabile statistica* deriva dalla classificazione di dati rilevati, cioè viene definita empiricamente una volta conosciuti i dati ed averli classificati.

La variabile casuale, indicata con X, assume valori nello spazio dei numeri reali secondo una funziona di probabilità P(X). La funzione di probabilità sarà quindi definita nel seguente modo $f(x)=P(X=x)$.

Il concetto di variabile casuale è diverso dal concetto di *variabile definita su una popolazione* di cui io posso conoscere o meno il valore che questa assume sulle singole unità.

1. LE VARIABILI CASUALI DISCRETE E CONTINUE

Le variabili casuali possono essere discrete o continue:

- Una *variabile casuale discreta* assume un numero finito o un'infinità numerabile di valori;
- Una *variabile casuale continua* può assumere un qualsiasi valore reale, il quale è determinato in un intervallo.

1.1 La variabile casuale discreta

Con le variabili casuali discrete abbiamo visto che abbiamo un numero finito di valori reali, indicabili come $(x_1, x_2, \cdots, x_n)$ con probabilità $(\rho_1, \rho_2, \dots, \rho_n)$. La funzione di probabilità sarà la stessa che abbiamo già visto in precedenza, tenendo conto del fatto che la funzione sarà maggiore o uguale a 0 e che la sommatoria delle varie probabilità delle sequenze sarà pari a 1. Quando parliamo di sommatoria, che indichiamo con $\sum_i f(x_i) = 1$, possiamo anche parlare di

probabilità cumulata, descritta dalla *funzione di ripartizione* e indicata con $F(X_K)$, avremo quindi: $F(X_K) = f(x_1) + f(x_2), \dots, f(x_k)$.

Tra le variabili casuali discrete troviamo: la variabile casuale di Bernoulli e la variabile casuale binomiale.

1.2 La variabile casuale continua

Quando abbiamo parlato della variabile casuale continua abbiamo detto che è una v.c che può assumere un numero infinito di valori compresi in un intervallo, o in uno spazio, con *ampiezza finita o infinita*. A differenza di quella continua non possiamo calcolare una cumulata, si parla di paradosso della continuità e lo si risolve ricorrendo al concetto di area, ossia, *la probabilità sarà data da un intervallo che possiamo rappresentare su un istogramma e l'area di quella parte di istogramma sarà pari alla nostra probabilità*.

A questo concetto è legata la funzione di densità di probabilità, o *densità di probabilità*, la quale è una funzione matematica *f(x)* per cui l'area sottesa alla funzione, corrispondente ad un certo intervallo è uguale alla probabilità che *x* assuma un valore in quell'intervallo. La funzione verrà calcolata attraverso un integrale, che noi non andremo a calcolare.

Una variabile casuale continua è la variabile casuale normale.

2. LA VARIABILE CASUALE DI BERNOULLI

La variabile casuale di Bernoulli è una variabile che trae origine da una prova nella quale interessa verificare se l'evento *E* si è verificato o meno. Essa è legata a una tipologia di esperimento/prova detta dicotomica, o *dicotomizzabile*, i cui due possibili risultati vengono indicati con il termine "Successo", il quale vale *1* e indicheremo con *p*, ed "Insuccesso", il quale vale *0* e indicheremo con *1-p*, o con *q*.

La variabile casuale di Bernoulli presenta una sola proprietà, la quale è proprio la probabilità di Successo, quindi, indicheremo *X~Ber (p)*.

La sua distribuzione di probabilità sarà: $P(X=X)=p^x(1-p)^{1-x}$.
Questa presenta due condizioni, ossia che il suo *valore atteso, o media, sia uguale alla probabilità di successo (p)* e invece la sua *varianza sarà data dal prodotto tra la probabilità di successo (p) e la probabilità di insuccesso (1-p), (q)*.

- *E(x)=p;*

- *Var(x)=p(1-p).*

Il calcolo per una v.c di Bernoulli si risolverà calcolando la probabilità di successo di ciascuna sequenza, per poi calcolare la cumulata di ciascuna p.

3. LA VARIABILE CASUALE BINOMIALE

Alla v.c di Bernoulli è legata la variabile casuale binomiale, che prevede proprio un *esperimento binomiale,* dato da un *numero di prove Bernoulliane*, e ogni prova può avere solo due possibili risultati, tutte le prove sono *indipendenti* e ognuna della quali ha la *stessa probabilità di successo*. Avremo così le seguenti condizioni:

- Successo o insuccesso;
- Probabilità costante in tutte le prove;
- Estrazioni indipendenti e con ripetizione.

Avremo come v.c binomiale *x* il *numero di successi di n prove*, e indicheremo sempre il successo con *p* e l'insuccesso con *1-p*.

Questa volta, per poter calcolare questa probabilità useremo la *funzione di probabilità con il fattoriale*, anche chiamato *coefficiente binomiale,* la quale sarà una funzione che possiamo utilizzare anche per le v.c Bernoulliane quando abbiamo un numero di sequenze molto alto. La nostra funzione sarà, quindi:

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x}$$

In questo caso le condizioni, o proprietà della v.c binomiale saranno il numero di prove e la probabilità di successo, che indicheremo con *X~Bin(n;p)*. Anche in questo caso avremo che la media, o valore atteso, sarà uguale al *prodotto tra il numero di prove e la probabilità di successo*, invece, la varianza sarà *data dal prodotto tra il prodotto di n e p e la probabilità di insuccesso:*

- *E(x)=np;*
- *Var(x)=np(1-p).*

Possiamo, inoltre, calcolare la binomiale relativa, ossia la proporzione di successi in n prove, data proprio dal rapporto tra *x* e *n*: *x/n*. In questo caso la loro media e la loro varianza saranno:

- *E(x/n)=p;*
- *Var(x/n)=*$\frac{p(1-p)}{n}$

Inoltre, nel caso di un numero di prove, ossia *n*, molto *grande* avremo un'approssimazione della distribuzione binomiale. La quale consiste nel calcolo di z, useremo quindi la tavola z, che poi analizzeremo con la variabile casuale normale.

4. LA VARIABILE CASUALE NORMALE

La variabile casuale normale è anche chiamata curva degli errori accidentali, questo perché è legata agli *errori di una misurazione fisica*. Essa la si utilizza per gli esperimenti che *si distribuiscono in maniera normale* ed è rappresentata attraverso un istogramma, anche chiamata *curva Gaussiana* da Gauss.

Avremo questa tipologia di variabile casuale in tre casi:

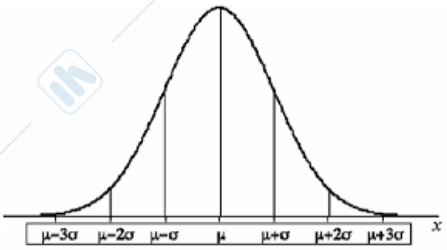
- *Misurazione di una grandezza fisica;*
- *Distribuzione di una caratteristica della popolazione;*
- *Dimensione effettiva degli oggetti prodotti in serie e che si cerca di produrre in modo identico.*

Quindi la prima cosa che ci si chiede è "Di quale casistica stiamo parlando?"

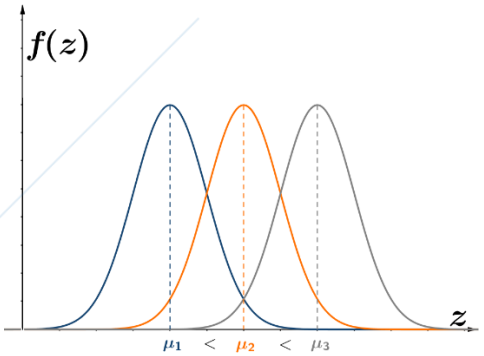
La distribuzione di una v.c. normale presenta inoltre due parametri, i quali sono la media e la varianza. Scriveremo quindi: $X \sim N(\mu;\sigma^2)$

La distribuzione normale presenta una serie di caratteristiche, le quali sono:

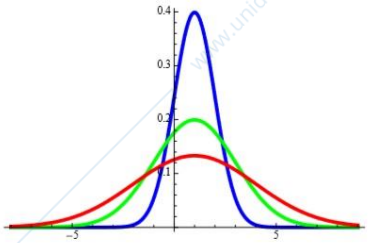
1. Forma campanulare e simmetrica, come possiamo anche osservare dal seguente grafico;
2. Media, mediana e moda sono coincidenti e dividono in due il grafico;
3. Ha due punti di flesso che indicheremo con $\mu \pm \sigma$ equidistanti dalla media (il punto di flesso è il punto in cui la curva da concava diventa convessa);
4. Circa il 68% dei valori della variabile è compreso nell'intervallo $\mu \pm \sigma$;
5. Circa il 95% dei valori della variabile è compreso nell'intervallo $\mu \pm 2\sigma$;
6. Circa il 98% dei valori della variabile è compreso nell'intervallo $\mu \pm 3\sigma$;



7. Un aumento (o una *diminuzione*) della media determina uno slittamento, a parità di forma, della curva sull'asse x verso destra (o verso *sinistra*);



8. Un aumento (o una *diminuzione*) della varianza determina una minore (o una *maggiore*) *concentrazione dei valori attorno alla media*; quindi, la curva si restringerà (platicurtica) (o si *amplierà* -leptocurtica). Ricorda che se cambia σ^2, cambierà anche s.q.m;



9. La somma di più (k) variabili casuali normali, darà ancora una variabile casuale normale;
10. Una trasformazione lineare di una variabile casuale normale sarà ancora una variabile casuale normale.

La variabile casuale normale è largamente utilizzata come *modello probabilistico* per i più svariati fenomeni. Essa sarà sempre definita *tra* $-\infty$ *e* $+\infty$, cioè su tutta la retta *dell'asse x*, quindi su tutto l'insieme *R* (numeri reali).

La sua *funzione di densità di probabilità* sarà calcolata mediante l'utilizzo del *numero di Nepero*, ma anche in questo caso si andrà a semplificare il calcolo. Andiamo, in realtà, a standardizzare il calcolo, ma perché?
Nel tempo, la media e la varianza tendono a modificarsi e a variare, ciò significa andare, ogni volta, a dover ricalcolare la probabilità di una variabile casuale normale. Tutto ciò può essere semplificato, attraverso la standardizzazione della nostra distribuzione, con questo andremo poi a calcolare l'area, ossia la probabilità, che ci interessa sul grafico standardizzato. Il quale presenterà una serie di caratteristiche:

- La μ sarà uguale a 0 e la σ^2 sarà uguale a 1;
- L'intera area dell'istogramma sarà pari a 1, così come il valore massimo che diamo alla probabilità è 1 (100%);

- Dividendo a metà il grafico con la media, avremo che la metà di destra e la metà di sinistra avranno rispettivamente un'area pari a 0,50; ossia una probabilità uguale a 0,50 (50%).

Qualsiasi distribuzione normale può essere ricondotta a una distribuzione standardizzata con media e una varianza unitaria, mediante la trasformazione che avviene con questa formula: $z = \frac{x-\mu}{\sigma}$. In questo modo i valori sotto la curva gaussiana possono essere tabulati. Questa *z*, infatti, sarà il valore da trovare all'interno *della tavola della distribuzione normale standardizzata z*, che ci darà l'area sottesa alla curva di densità normale standardizzata calcolata tra 0 e Z, ossia, ci darà il valore dell'area, incrociando il valore di riga e quello colonna.
Su questa tavola avremo:

- Riga: L'unità e il primo decimale di z;
- Colonna: Il secondo decimale di z (se è presente un terzo decimale, approssimiamo).

Parliamo di valori che vanno da 0 a Z perché seppure avessimo valori negativi, *l'area (probabilità) sarà sempre positiva*, vedremo solo che cambia la disposizione sul grafico.

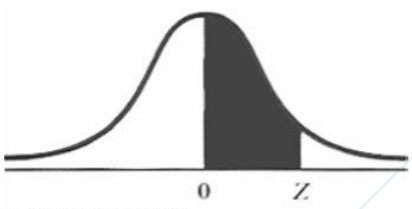


Tavola della distribuzione Normale Standardizzata

All'interno della tabella abbiamo la probabilità
Riga: primo decimale;
Colonna: secondo decimale.
Si incrociano riga e colonna, questo ci dà l'area che ci serve per l'istogramma

	Area sottesa alla curva di densità normale standardizzata calcolata tra **0** e **Z**									
Z	**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

Come già abbiamo visto nel paragrafo precedente, possiamo avere anche un'*approssimazione della distribuzione binomiale,* quando il numero di prove è molto grande. Avremo due casistiche, nel caso in cui parliamo di binomiale o binomiale relativa:

- Binomiale: $z = \frac{x-n\rho}{\sqrt{n \cdot \rho \cdot q}}$;
- Binomiale relativa: $z = \frac{\left(\frac{x}{n}\right)-\rho}{\sqrt{\frac{p \cdot q}{n}}}$.

Dalla variabile casuale normale è possibile derivare ulteriori variabili casuali che per il loro impiego nella statistica inferenziale assumono un notevole interesse. Esse sono:

- *Variabile casuale* χ^2.
- *Variabile casuale T di Student;*
- *Variabile casuale F di Fisher;*

5. LA VARIABILE CASUALE χ^2

La variabile casuale x^2, è un ulteriore variabile casuale continua, come quella di Fisher, di Student e normale. Sul supporto che prende i valori *da 0 a* $+\infty$.

Essa è definita mediante i gradi di libertà ed è una *distribuzione asimmetrica, continua e definita per valori reali non negativi.*

La *funzione di densità* dipende da un unico parametro chiamato grado di libertà, un intero positivo indicato con la lettera *v*. In formule la indicheremo nel seguente modo: $x \sim \chi^2_{v,\alpha} = \sum_{g=1}^{v} z_g^2$.

Per valori elevati, ossia $n \geq 80$ si approssimerà come una variabile casuale normale.

Questa ha un valore atteso uguale ai gradi di libertà e una varianza pari al doppio dei gradi di libertà.

La si applica sull'inferenza sulla varianza di una popolazione, per i *test di adattamento** e per il *calcolo dell'indipendenza statistica**.

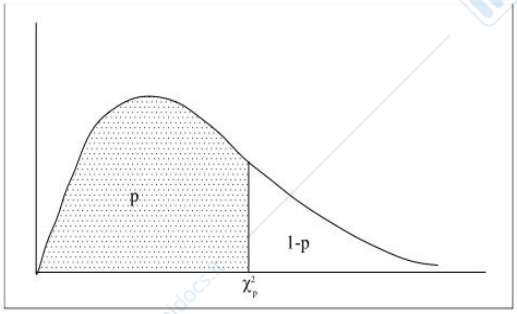
Useremo, quindi, la tavola del chi quadrato:

- Riga: cerchiamo il grado di libertà, dato da n-1;
- Colonna: cerchiamo il valore di α, che vedremo poi a cosa corrisponde.

VALORI PERCENTILI (χ^2_P) PER LA DISTRIBUZIONE CHI-QUADRATO CON ν GRADI DI LIBERTÀ

	α	α	α	α	α	α	α	α	α	α	α	α	α	α
	0.995	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01	0.005	0.001
ν	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P
1	0.0000	0.0002	0.0010	0.0039	0.0158	0.102	0.455	1.323	2.706	3.841	5.024	6.635	7.879	10.827
2	0.0100	0.0201	0.0506	0.1026	0.211	0.575	1.386	2.773	4.605	5.991	7.378	9.210	10.597	13.815
3	0.0717	0.1148	0.2158	0.352	0.584	1.213	2.366	4.108	6.251	7.815	9.348	11.345	12.838	16.266
4	0.207	0.297	0.484	0.711	1.064	1.923	3.357	5.385	7.779	9.488	11.143	13.277	14.860	18.466
5	0.412	0.554	0.831	1.145	1.610	2.675	4.351	6.626	9.236	11.070	12.832	15.086	16.750	20.515
6	0.676	0.872	1.237	1.635	2.204	3.455	5.348	7.841	10.645	12.592	14.449	16.812	18.548	22.457
7	0.989	1.239	1.690	2.167	2.833	4.255	6.346	9.037	12.017	14.067	16.013	18.475	20.278	24.321
8	1.344	1.647	2.180	2.733	3.490	5.071	7.344	10.219	13.362	15.507	17.535	20.090	21.955	26.124
9	1.735	2.088	2.700	3.325	4.168	5.899	8.343	11.389	14.684	16.919	19.023	21.666	23.589	27.877
10	2.156	2.558	3.247	3.940	4.865	6.737	9.342	12.549	15.987	18.307	20.483	23.209	25.188	29.588
11	2.603	3.053	3.816	4.575	5.578	7.584	10.341	13.701	17.275	19.675	21.920	24.725	26.757	31.264
12	3.074	3.571	4.404	5.226	6.304	8.438	11.340	14.845	18.549	21.026	23.337	26.217	28.300	32.909
13	3.565	4.107	5.009	5.892	7.041	9.299	12.340	15.984	19.812	22.362	24.736	27.688	29.819	34.527
14	4.075	4.660	5.629	6.571	7.790	10.165	13.339	17.117	21.064	23.685	26.119	29.141	31.319	36.124
15	4.601	5.229	6.262	7.261	8.547	11.037	14.339	18.245	22.307	24.996	27.488	30.578	32.801	37.698
16	5.142	5.812	6.908	7.962	9.312	11.912	15.338	19.369	23.542	26.296	28.845	32.000	34.267	39.252
17	5.697	6.408	7.564	8.672	10.085	12.792	16.338	20.489	24.769	27.587	30.191	33.409	35.718	40.791
18	6.265	7.015	8.231	9.390	10.865	13.675	17.338	21.605	25.989	28.869	31.526	34.805	37.156	42.312
19	6.844	7.633	8.907	10.117	11.651	14.562	18.338	22.718	27.204	30.144	32.852	36.191	38.582	43.819
20	7.434	8.260	9.591	10.851	12.443	15.452	19.337	23.828	28.412	31.410	34.170	37.566	39.997	45.314
21	8.034	8.897	10.283	11.591	13.240	16.344	20.337	24.935	29.615	32.671	35.479	38.932	41.401	46.796
22	8.643	9.542	10.982	12.338	14.041	17.240	21.337	26.039	30.813	33.924	36.781	40.289	42.796	48.268
23	9.260	10.196	11.689	13.091	14.848	18.137	22.337	27.141	32.007	35.172	38.076	41.638	44.181	49.728
24	9.886	10.856	12.401	13.848	15.659	19.037	23.337	28.241	33.196	36.415	39.364	42.980	45.558	51.179
25	10.520	11.524	13.120	14.611	16.473	19.939	24.337	29.339	34.382	37.652	40.646	44.314	46.928	52.619
26	11.160	12.198	13.844	15.379	17.292	20.843	25.336	30.435	35.563	38.885	41.923	45.642	48.290	54.051
27	11.808	12.878	14.573	16.151	18.114	21.749	26.336	31.528	36.741	40.113	43.195	46.963	49.645	55.475
28	12.461	13.565	15.308	16.928	18.939	22.657	27.336	32.620	37.916	41.337	44.461	48.278	50.994	56.892
29	13.121	14.256	16.047	17.708	19.768	23.567	28.336	33.711	39.087	42.557	45.722	49.588	52.335	58.301
30	13.787	14.953	16.791	18.493	20.599	24.478	29.336	34.800	40.256	43.773	46.979	50.892	53.672	59.702
40	20.707	22.164	24.433	26.509	29.051	33.660	39.335	45.616	51.805	55.758	59.342	63.691	66.766	73.403
50	27.991	29.707	32.357	34.764	37.689	42.942	49.335	56.334	63.167	67.505	71.420	76.154	79.490	86.660
60	35.534	37.485	40.482	43.188	46.459	52.294	59.335	66.981	74.397	79.082	83.298	88.379	91.952	99.608
70	43.275	45.442	48.758	51.739	55.329	61.698	69.334	77.577	85.527	90.531	95.023	100.425	104.215	112.317
80	51.172	53.540	57.153	60.391	64.278	71.145	79.334	88.130	96.578	101.879	106.629	112.329	116.321	124.839
90	59.196	61.754	65.647	69.126	73.291	80.625	89.334	98.650	107.565	113.145	118.136	124.116	128.299	137.208
100	67.328	70.065	74.222	77.929	82.358	90.133	99.334	109.141	118.498	124.342	129.561	135.807	140.170	149.449

N.B.: il percentile *p* rappresenta la probabilità di non superamento e quindi è uguale a (1 - il livello di significatività alfa)



6. LA VARIABILE CASUALE T DI STUDENT

La variabile casuale T di Student è una variabile *standardizzata*, usata nei casi in cui n<100, ma noi la useremo quando $n \leq 30$, secondo il Teorema del Limite Centrale che analizzeremo nel prossimo capitolo.

Questa è il rapporto tra una variabile casuale normale standardizzata e una radice della variabile casuale chi quadrato, con i cosiddetti *gradi di libertà*, che indicheremo con *v*, i quali sono indipendenti tra di loro.
Avremo quindi: $x \sim Tv,\alpha = \frac{z}{\sqrt{\frac{x^2}{v}}}$

Essa è una *variabile casuale continua* definita sull'intero asse reale, ha forma, anch'essa campanulare platicurtica ed è simmetrica rispetto alla media. Inoltre come parametro, o condizione, principale ha i gradi libertà, i quali corrispondono a *n-1*.
La sua media standardizzata è uguale a 0, invece la sua varianza sarà data dal seguente rapporto: $\frac{v}{v-2}$.

Essa si applica:

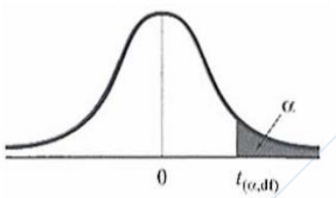
- *Sulla media della popolazione normale;*
- *Sulla differenza tra le media di due popolazioni normali;*
- *Sui parametri del modello di regressione.*

Per trovare il valore che cerchiamo sulla tavola t di student, andremo a cercare:

- Riga: Gradi di liberta;
- Colonna: Il valore di $\propto/2$.

I gradi di libertà che prendiamo in considerazione arrivano fino a 29 (n=30).

Tavola della distribuzione T di Student



Gradi di libertà	Area nella coda di destra								
	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	3.078	6.314	12.706	15.894	31.821	63.656	127.321	318.289	636.578
2	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.328	31.600
3	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.214	12.924
4	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.894	6.869
6	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922
19	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.689
28	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.660
30	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
31	1.309	1.696	2.040	2.144	2.453	2.744	3.022	3.375	3.633
32	1.309	1.694	2.037	2.141	2.449	2.738	3.015	3.365	3.622
33	1.308	1.692	2.035	2.138	2.445	2.733	3.008	3.356	3.611
34	1.307	1.691	2.032	2.136	2.441	2.728	3.002	3.348	3.601
35	1.306	1.690	2.030	2.133	2.438	2.724	2.996	3.340	3.591
36	1.306	1.688	2.028	2.131	2.434	2.719	2.990	3.333	3.582
37	1.305	1.687	2.026	2.129	2.431	2.715	2.985	3.326	3.574
38	1.304	1.686	2.024	2.127	2.429	2.712	2.980	3.319	3.566
39	1.304	1.685	2.023	2.125	2.426	2.708	2.976	3.313	3.558
40	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
41	1.303	1.683	2.020	2.121	2.421	2.701	2.967	3.301	3.544
42	1.302	1.682	2.018	2.120	2.418	2.698	2.963	3.296	3.538
43	1.302	1.681	2.017	2.118	2.416	2.695	2.959	3.291	3.532
44	1.301	1.680	2.015	2.116	2.414	2.692	2.956	3.286	3.526
45	1.301	1.679	2.014	2.115	2.412	2.690	2.952	3.281	3.520
46	1.300	1.679	2.013	2.114	2.410	2.687	2.949	3.277	3.515
47	1.300	1.678	2.012	2.112	2.408	2.685	2.946	3.273	3.510
48	1.299	1.677	2.011	2.111	2.407	2.682	2.943	3.269	3.505
49	1.299	1.677	2.010	2.110	2.405	2.680	2.940	3.265	3.500
50	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496

Tavola della distribuzione T di Student (continua)

Gradi di libertà	Area nella coda di destra								
	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
51	1.298	1.675	2.008	2.108	2.402	2.676	2.934	3.258	3.492
52	1.298	1.675	2.007	2.107	2.400	2.674	2.932	3.255	3.488
53	1.298	1.674	2.006	2.106	2.399	2.672	2.929	3.251	3.484
54	1.297	1.674	2.005	2.105	2.397	2.670	2.927	3.248	3.480
55	1.297	1.673	2.004	2.104	2.396	2.668	2.925	3.245	3.476
56	1.297	1.673	2.003	2.103	2.395	2.667	2.923	3.242	3.473
57	1.297	1.672	2.002	2.102	2.394	2.665	2.920	3.239	3.469
58	1.296	1.672	2.002	2.101	2.392	2.663	2.918	3.237	3.466
59	1.296	1.671	2.001	2.100	2.391	2.662	2.916	3.234	3.463
60	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
61	1.296	1.670	2.000	2.099	2.389	2.659	2.913	3.229	3.457
62	1.295	1.670	1.999	2.098	2.388	2.657	2.911	3.227	3.454
63	1.295	1.669	1.998	2.097	2.387	2.656	2.909	3.225	3.452
64	1.295	1.669	1.998	2.096	2.386	2.655	2.908	3.223	3.449
65	1.295	1.669	1.997	2.096	2.385	2.654	2.906	3.220	3.447
66	1.295	1.668	1.997	2.095	2.384	2.652	2.904	3.218	3.444
67	1.294	1.668	1.996	2.095	2.383	2.651	2.903	3.216	3.442
68	1.294	1.668	1.995	2.094	2.382	2.650	2.902	3.214	3.439
69	1.294	1.667	1.995	2.093	2.382	2.649	2.900	3.213	3.437
70	1.294	1.667	1.994	2.093	2.381	2.648	2.899	3.211	3.435
71	1.294	1.667	1.994	2.092	2.380	2.647	2.897	3.209	3.433
72	1.293	1.666	1.993	2.092	2.379	2.646	2.896	3.207	3.431
73	1.293	1.666	1.993	2.091	2.379	2.645	2.895	3.206	3.429
74	1.293	1.666	1.993	2.091	2.378	2.644	2.894	3.204	3.427
75	1.293	1.665	1.992	2.090	2.377	2.643	2.892	3.202	3.425
76	1.293	1.665	1.992	2.090	2.376	2.642	2.891	3.201	3.423
77	1.293	1.665	1.991	2.089	2.376	2.641	2.890	3.199	3.421
78	1.292	1.665	1.991	2.089	2.375	2.640	2.889	3.198	3.420
79	1.292	1.664	1.990	2.088	2.374	2.639	2.888	3.197	3.418
80	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
81	1.292	1.664	1.990	2.087	2.373	2.638	2.886	3.194	3.415
82	1.292	1.664	1.989	2.087	2.373	2.637	2.885	3.193	3.413
83	1.292	1.663	1.989	2.087	2.372	2.636	2.884	3.191	3.412
84	1.292	1.663	1.989	2.086	2.372	2.636	2.883	3.190	3.410
85	1.292	1.663	1.988	2.086	2.371	2.635	2.882	3.189	3.409
86	1.291	1.663	1.988	2.085	2.370	2.634	2.881	3.188	3.407
87	1.291	1.663	1.988	2.085	2.370	2.634	2.880	3.187	3.406
88	1.291	1.662	1.987	2.085	2.369	2.633	2.880	3.185	3.405
89	1.291	1.662	1.987	2.084	2.369	2.632	2.879	3.184	3.403
90	1.291	1.662	1.987	2.084	2.368	2.632	2.878	3.183	3.402
91	1.291	1.662	1.986	2.084	2.368	2.631	2.877	3.182	3.401
92	1.291	1.662	1.986	2.083	2.368	2.630	2.876	3.181	3.399
93	1.291	1.661	1.986	2.083	2.367	2.630	2.876	3.180	3.398
94	1.291	1.661	1.986	2.083	2.367	2.629	2.875	3.179	3.397
95	1.291	1.661	1.985	2.082	2.366	2.629	2.874	3.178	3.396
96	1.290	1.661	1.985	2.082	2.366	2.628	2.873	3.177	3.395
97	1.290	1.661	1.985	2.082	2.365	2.627	2.873	3.176	3.394
98	1.290	1.661	1.984	2.081	2.365	2.627	2.872	3.176	3.393
99	1.290	1.660	1.984	2.081	2.365	2.626	2.871	3.175	3.391
100	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
101	1.290	1.660	1.984	2.081	2.364	2.625	2.870	3.173	3.389
102	1.290	1.660	1.983	2.080	2.363	2.625	2.869	3.172	3.389
103	1.290	1.660	1.983	2.080	2.363	2.624	2.869	3.171	3.388
104	1.290	1.660	1.983	2.080	2.363	2.624	2.868	3.170	3.387
105	1.290	1.659	1.983	2.080	2.362	2.623	2.868	3.170	3.386
106	1.290	1.659	1.983	2.079	2.362	2.623	2.867	3.169	3.385
107	1.290	1.659	1.982	2.079	2.362	2.623	2.866	3.168	3.384
108	1.289	1.659	1.982	2.079	2.361	2.622	2.866	3.167	3.383
109	1.289	1.659	1.982	2.079	2.361	2.622	2.865	3.167	3.382
110	1.289	1.659	1.982	2.078	2.361	2.621	2.865	3.166	3.381
30000	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291

7. LA VARIABILE CASUALE F DI FISHER

La variabile casuale F di Fisher può essere definita come il rapporto tra due variabili casuali χ^2 indipendenti e rapportate ai *rispettivi gradi di libertà*.

Avremo quindi: $X \sim F_{v_1;v_2} = \frac{\frac{X^2_{v_1}}{V_1}}{\frac{X^2_{v_2}}{V_2}}$

È una distribuzione *asimmetrica, continua e definita* per valori reali non negativi, essa dipende da due parametri, ossia i *gradi di libertà* v_1 *e* v_2.

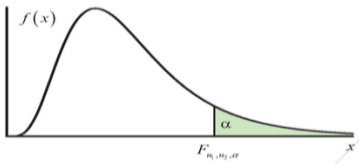
Essa si applica al rapporto tra varianze tra due popolazioni normali e al *modello di regressione**.

Abbiamo come media e varianza le seguenti formule:

- Media: $E(x) = \frac{v_1}{v_2} - 2$;
- Varianza: $\text{Var}(x) = \frac{2^2_{v_2}(v_1+v_2-2)}{v_1(v_2-2)^2(v_2-4)}$.

Anche in questo si userà la tavola ma della F di Fisher.

Percentili della variabile casuale F di Fisher



$\alpha = 0.05$

Denom. n_2	Numeratore n_1																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.	199.	215.	224.	230.	234.	236.	238.	240.	241.	243.	246.	248.	249.	250.	251.	252.	253.	254.
2	18.5	19.0	19.1	19.2	19.3	19.3	19.3	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

FORMULARIO VARIABILI CASUALI

Variabile casuale di Bernoulli	Formula
Sequenza	$f(x_n) = p(1-p)$
Frequenza cumulata	$F(X_K) = f(x_1) + f(x_2) + \cdots + f(x_k)$
Media	$\mu = p$
Varianza	$\sigma^2 = p(1-p)$

Variabile casuale normale	Formula
Valore z standardizzato	$P\ (Z=\frac{X-\mu}{\sigma})$
Formula inversa per la x	$X=\mu + (\sigma z)$

Variabile casuale binomiale	Formula
Probabilità	$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x}$
Media	$\mu = np$
Varianza	$\sigma^2 = np(1-p)$
Binomiale relativa	x/n
Media proporzionale	$\mu = p$
Varianza proporzionale	$\sigma^2 = \frac{p(1-p)}{n}$
Approssimazione binomiale	$z = \frac{x - n\rho}{\sqrt{n \cdot \rho \cdot q}}$
Approssimazione binomiale relativa	$z = \frac{\left(\frac{x}{n}\right) - \rho}{\sqrt{\frac{p \cdot q}{n}}}$

Spiegazione esercizi Step by Step sulle variabili casuali

Variabile casuale di Bernoulli

- Scriviamo i dati che abbiamo, per questo esercizio ci serviranno:
  - *Tipologia di esperimento;*
  - *Variabile casuale analizzata;*
  - *Numero di unità del campione;*
  - *Probabilità di successo;*
  - *Probabilità di insuccesso.*
- Calcoliamo la probabilità di ogni singola sequenza, *moltiplicandola probabilità di successo e si insuccesso in base all'alternanza della sequenza*;
- Calcoliamo la probabilità finale con la frequenza cumulata, ossia la *sommatoria delle singole sequenze*;
- Scriviamo il commento finale;
- *Se richiesto calcoliamo la media e la varianza.*

Variabile casuale binomiale

- Scriviamo i dati che abbiamo, per questo esercizio ci serviranno:
  - *Variabile casuale analizzata;*
  - *Numero di unità del campione;*
  - *Probabilità di successo;*
  - *Probabilità di insuccesso;*
  - *Proprietà della binomiale;*
- Calcoliamo la probabilità con la formula del fattoriale;
- *Se serve sommiamo le singole probabilità per quella finale*;
- Scriviamo il commento finale;
- *Se richiesto calcoliamo la media e la varianza;*
- *Se richiesto calcoliamo la proporzione del numero di successi e commentiamo;*
- *Se richiesto calcoliamo la media e la varianza proporzionale.*

Variabile casuale binomiale con approssimazione

- Scriviamo i dati che abbiamo, per questo esercizio ci serviranno:
  - *Variabile casuale analizzata;*
  - *Numero di unità del campione, che sarà molto alto;*
  - *Probabilità di successo;*
  - *Probabilità di insuccesso;*
  - *Proprietà della binomiale.*
- Calcoliamo la probabilità con la formula della z approssimata;
- Cerchiamo z;

- Scriviamo il commento finale;
- *Se richiesto calcoliamo la proporzione del numero di successi.*

Variabile casuale normale

- Scriviamo i dati che abbiamo, per questo esercizio ci serviranno:
  - *Numero di unità del campione;*
  - *Media;*
  - *Varianza;*
  - *Scarto quadratico medio;*
  - *Proprietà della normale;*
  - *Proprietà della normale approssimate.*
- Disegniamo il grafico con i valori che abbiamo;
- Calcoliamo il valore della z;
- Disegniamo il grafico con i valori di z;
- Cerchiamo le aree da 0 a z sulla tavola di z;
- Calcolo la probabilità in base al caso che ho, ricordando che:
  - *Se ho un intervallo che comprende lo zero tra due valori, sommerò le due aree;*
  - *Se ho un intervallo che va da un valore a infinito, sottraiamo 0,50 con l'area trovata sulla tavola;*
  - *Se ho un intervallo tra due valori che non comprende lo zero, sottraiamo l'area maggiore con la minore;*
  - *Se ho un intervallo che va da 0 a un valore, l'area sarà uguale all'area trovata in tavola corrispondente alla z di quel valore che non è zero;*
  - *Se ho un intervallo tra o e infinito, l'area sarà 0,50;*
  - *Se ho un intervallo che va da un numero inferiore (superiore) a 0 a più infinito (meno infinito) sommerò 0,50 con l'area trovata nella tavola.*
- Scriviamo il commento finale.

Calcolo inverso per trovare un valore della v.c normale

- Scriviamo i dati che abbiamo, per questo esercizio ci serviranno:
  - *Il valore della probabilità;*
  - *Media;*
  - *Scarto quadratico medio.*
- Sottraiamo 0,50 con la probabilità;
- Cerchiamo il valore ottenuto dalla sottrazione nella tavola;
- Uso la formula inversa per trovare x;
- Disegniamo il grafico con i valori della x e della media;
- Scriviamo il commento finale.

Esercitazione

1. **Esercizio:**

In un mazzo di carte colorate ne consideriamo a caso 4, ogni carta sarà di un solo colore, o rosso o blu:

- **a)** Calcolare la probabilità che nel campione di 4 carte ci siano 3 carte rosse;
- **b)** Calcolare la media e la varianza.

2. **Esercizio:**

Un recente studio ha riportato che solo il 40% degli elettori americani intende votare alle prossime elezioni presidenziali. Si immagini di selezionare un casualmente 10 elettori. Utilizzando la funzione di probabilità binomiale si chiede di determinare la probabilità che:

- **a)** Tutti e 10 elettori intendano votare alle prossime elezioni;
- **b)** Nessuno dei 10 elettori intendano votare alle prossime elezioni;
- **c)** Solo 3 elettori intendano votare alle prossime elezioni;
- **d)** Almeno 7 elettori intendano votare alle prossime elezioni;
- **e)** Al massimo 4 elettori intendano votare alle prossime elezioni;
- **f)** Da 4 a 6 elettori intendano votare alle prossime elezioni;
- **g)** Calcolare la media e la varianza;
- **h)** Calcolare la proporzione del numero di successi per le prove ottenute (binomiale relativa) del punto c;
- **i)** Calcolare la media e la varianza proporzionale.

3. **Esercizio:**

Su 400 monete:

- **a)** Calcolare la probabilità che esce croce 75 volte attraverso l'approssimazione binomiale;
- **b)** Calcolare la proporzione del numero di successi per le prove ottenute, attraverso l'approssimazione binomiale relativa.

4. **Esercizio:**

Il tempo necessario per sviluppare un nuovo prodotto, dal concetto iniziale al design finale, presso la Klobes Engineering segue una distribuzione normale, con media di 120 giorni e una varianza di 100 giorni. Calcolare quanto è probabile che il tempo di sviluppo cada nei seguenti intervalli:

- **a)** Tra 112 e 120 giorni;
- **b)** Tra 120 e 135 giorni;
- **c)** Tra 110 e 138 giorni;
- **d)** Tra 95 e 115 giorni;
- **e)** Tra 130 e 135 giorni;

f) Oltre 105 giorni;
g) Meno di 125 giorni;
h) Oltre 128 giorni;
i) Meno di 110 giorni;
j) Oltre 120 giorni;
k) Stabilire quei giorni g_0 per i quali la probabilità che il tempo di sviluppo sia maggiore di g_0 è pari a 0,30;
l) Stabilire quei giorni g_0 per i quali la probabilità che il tempo di sviluppo sia minore di g_0 è pari a 0,42.

5. Esercizio:

Su 100 carte di colore rosso e blu:

c) Calcolare la probabilità che escono 25 volte le carte rosse;
d) Calcolare approssimativamente la proporzione del numero di successi per le prove ottenute;
e) Commentare i risultati ottenuti.

6. Esercizio:

Un recente studio della National Highway Traffic Safety Administration, ha affermato che il tasso di utilizzo del casco per i motociclisti negli USA è del 55%. Se questa affermazione è vera, osservando 15 motociclisti selezionati casualmente, calcolare la probabilità e commentarla, riguardo i seguenti eventi:

a) Tutti e 15 motociclisti usano il casco;
b) Esattamente 12 di essi usano il casco;
c) Almeno 9 di essi usano il casco;
d) Da 9 a 11 di essi usano il casco;
e) Calcolare la media e la varianza;
f) Calcolare la proporzione del numero di successi per le prove ottenute (binomiale relativa) del punto b;
g) Calcolare la media e la varianza proporzionale.

7. Esercizio:

Calcolare la probabilità che su 5 lanci di moneta non truccata, esca testa 4 volte. Indicare, poi, la media e la varianza e commentare i risultati ottenuti.

8. Esercizio:

Un'azienda produce parti di motore di precisione che devono essere conformi a specifiche di diametro abbastanza rigide. I diametri seguono una distribuzione normale con una media di 125 mm e una deviazione standard di 4 mm. Calcolare la probabilità che il diametro sia:

a) Tra 120 e 130 mm;

b) Tra 131 e 133 mm;
c) Meno di 136 mm;
d) Più di 132 mm;
e) Meno 116 mm;
f) Stabilire quei mm mm_0 per i quali la probabilità che il diametro sia minore di mm_0 è pari a 0,25.

ARGOMENTI, SOTTOFORMA DI DOMANDA, DEL CAPITOLO 2

- *Che cosa sono le variabili casuali?*
- *Qual è la differenza con la variabile definita su una popolazione?*
- *Che cosa è una variabile casuale discreta? Quali sono?*
- *Che cosa è una variabile casuale continua? Quali sono?*
- *Che cosa è il paradosso della continuità?*
- *Che cosa si intende per variabile casuale di Bernoulli?*
- *Che cosa si intende per variabile casuale binomiale? Quali sono le condizioni?*
- *Che cosa si intende per variabile casuale normale? Perché si chiama curva degli errori accidentali?*
- *Quali sono i tre casi di una v.c. normale?*
- *Quali sono le caratteristiche di una v.c. normale?*
- *Che cosa si intende per normale standardizzata? Quali sono le sue caratteristiche?*
- *Che cosa si intende con approssimazione della distribuzione binomiale?*
- *Che cosa è una variabile casuale chi quadrato e quali sono le sue caratteristiche?*
- *Che cosa è una variabile casuale T di Student e quali sono le sue caratteristiche?*
- *Che cosa è una variabile casuale F di Fisher e quali sono le sue caratteristiche?*

Seconda parte

Il campionamento, le stime e i test statistici

CAPITOLO 3: IL CAMPIONAMENTO

Le prime indagini campionarie risalgono *tra la fine dell'800 e l'inizio degli anni '30 del '900*. Karla Marx fu il primo a fare *interviste non strutturate* sulle condizioni di vita dei lettori della rivista "Révue Socialiste", poi abbiamo avuto Weber che indagò sullo stile di vita degli operai. Infine, *Neyman* con campioni opportunamente scelti, il suo lavoro fu "On the two different aspecys of the rapresentative method: The method of stratified sampling and the method of purposive selection", pubblicato nel 1934, con questo Neyman dimostrò che: data una procedura di selezione casuale dei campioni di una data popolazione il carattere di rappresentatività sarà presente in ognuno dei possibili campioni selezionabili secondo tale procedura ciò implica la possibilità di utilizzare un qualunque di questi campioni per poter indurre conoscenza sull'intera popolazione.

Uno dei motivi per i quali usiamo le indagini campionarie è perché non possiamo esaminare ogni singola unità della popolazione; quindi, questa tipologia di indagine velocizza il processo.

Ma perché non possiamo esaminare tutte le unità?

1. *Limitate risorse disponibili;*
2. *L'intera popolazione non è fisicamente raggiungibile;*
3. *L'intera popolazione non è del tutto nota.*

Per superare questi problemi si ricorre, quindi, ad una rilevazione campionaria, ossia, una rilevazione che consente di studiare le caratteristiche di una popolazione attraverso l'osservazione di una popolazione stessa, ossia la popolazione del campione.

1. IL CAMPIONAMENTO

Il campionamento è il procedimento che usiamo per *ricavare parte delle unità della popolazione.*

L'indagine campionaria ci fornisce una *stima*, ossia un'indagine inferenziale, di un *valore approssimato*, ossia un parametro θ, *soggetto ad errori*.

2. IL DISEGNO DI CAMPIONAMENTO

L'insieme delle decisioni per formare un campione si definisce disegno di campionamento. (Esempio: Nelle ricerche ISTAT se andassimo a guardare le note metodologiche potremmo vedere che tipologia di disegno di campionamento è stato utilizzato).

Il disegno di campionamento si articola in questo modo:

1. *Definire la popolazione di riferimento;*
2. *Specificare la lista o le liste della popolazione;*
3. *Scegliere il metodo di campionamento* (esempio: probabilistico);
4. *Determinare la dimensione campionaria;*
5. *Estrarre il campione;*
6. *Segnare i dati estratti dal campione;*
7. *Controllare il tasso di risposta.*

Ricorda che quando si controlla il tasso di risposta bisogna vedere se ci sono state delle problematiche, ad esempio, se tra le unità dei soggetti estratti, ci sono unità che non hanno risposto a tutte le domande, in questo caso, possiamo estrarre dal resto della popolazione un'unità che ha caratteristiche molto simili all'altra e sostituire le risposte non date con l'una con quello dato dall'altro, o addirittura, sostituire totalmente l'unità. *L'importante è che abbia caratteristiche simili o uguali.*

2.1 Popolazione e campione statistico

Abbiamo già, precedentemente, discusso la differenza tra popolazione e campione statistico, ma c'è da dire altro.

La *popolazione*, come sappiamo, è l'*insieme di unità oggetto di rilevazione,* definiti dai contenuti, nello spazio e nel tempo.

La popolazione può essere:

- Teorica (o target), è l'insieme di tutti i casi che compongono la popolazione;
- Accessibile (o campionaria) che sono i casi effettivamente raggiungibili.

Si ha una lista della popolazione che può essere divisa in più o un solo elenco e la sua accuratezza è connessa alla casualità del processo di estrazione campionaria e alla rappresentatività dei campioni ottenuti.

Il *campione statistico,* come abbiamo visto, è l'insieme di *n* unità selezionate su un *N* popolazione. Esso è regolato *dalla legge del caso*, cioè dalla probabilità. Ricorda che esistono metodi e tecniche di selezione dei casi, infatti, le fonti devono essere note e disponibili.

2.2 Errore di campionamento

La stima di un parametro comporta un errore di campionamento, ossia, dovuto *dall'impossibilità di determinare con esattezza il suo valore*.

Al di fuori di un intervallo abbiamo, quindi, l'errore, rappresentato dalla *differenza tra i risultati ottenuti dal campionamento e la vera caratteristica della popolazione che vogliamo stimare,* ossia quanto si discosta la stima del valore dal valore reale. (Come quando in economia viene fatto un budget previsionale e un'analisi degli scostamenti, successivamente a ciò che è realmente accaduto, lo scostamento sarebbe l'errore di previsione, che ha lo stesso funzionamento dell'errore di campionamento)

Il parametro sarà quindi dato dalla *stima più o meno l'errore di campionamento*.

$$(V=v\pm\epsilon)$$

(V=Parametro; v=stima; ϵ (epsilon) =errore di campionamento)

L'errore di campionamento è direttamente proporzionale all'intervallo di fiducia, ossia quanto ci fidiamo di quella stima, che indichiamo con $\propto$ (alfa), e alla variabilità del fenomeno studiato e proporzionale all'ampiezza del campione.

Ricorda che, più il campione è grande più l'errore è piccolo, e viceversa.

Oltre all'errore di campionamento (o errore campionario), abbiamo poi altre tre tipologie di errori:

- *Errore o distorsione per dati mancanti* (parte dell'informazione campionaria non è stata rilevata e quindi l'informazione disponibile è meno completa di quanto previsto nel disegno di indagine)*;*
- *Errore di misurazione* (attiene alla fase di rilevazione dei dati e dipende soprattutto dalla abilità e destrezza dei rilevatori)*;*
- *Errore di copertura o distorsione nella selezione del campione* (riguarda tutte quelle casistiche in cui il campione selezionato dalla popolazione non è rappresentativo della stessa e quindi vi è una distorsione nella selezione).

Le prime due tipologie sono attribuibili alla fase di rilevazione, invece, l'errore di copertura e quello di campionamento riguardano la selezione del campione.

3. METODI DI CAMPIONAMENTO

Ciò che a noi interessa è l'estrazione di un campione casuale, ossia una *n-pla di variabili casuali X,* $C= (X_1, X_2, \ldots, X_n)$*, indipendenti e identicamente distribuite (I.I.D) con distribuzione di probabilità pari a quella della popolazione.*

Questa n-pla di v.c X viene indicata come una *n-pla di osservazioni* C_K, in quanto la sua particolare realizzazione è il campione osservato.

Il campione lo indicheremo con Ω.

Quindi, ciascuna variabile casuale di osservazione campionaria X, avrà la stessa distribuzione e gli stessi parametri della variabile X della popolazione, la quale ha unità indipendenti e identicamente distribuite (I.I.D).

Quindi, il campionamento è *l'attività di selezione del sottoinsieme della popolazione su cui rilevare i dati*, ossia il campione, che gioca un ruolo fondamentale per il collegamento tra le statistiche campionarie e i parametri reali della popolazione.

Le unità di una popolazione sono solitamente elencate in una o più liste dalle quali si effettua l'estrazione casuale per rilevare il campione.

Quando una popolazione viene da un'unica lista avremo il cosiddetto "*Campionamento casuale semplice*" (CCS).

In generale il *campionamento casuale semplice* è un metodo in grado di *garantire che ogni sottoinsieme di n unità della popolazione ha la medesima possibilità di essere selezionato*.

In pratica il CCS avviene selezionando in modo completamente casuale le *n* unità del campione, *una alla volta oppure tutte in blocco*.

Nel CCS tutte le unità hanno uguale probabilità di essere incluse nel campione. *Data una popolazione di N unità e un campione di dimensione n, la probabilità che l'i-nesimo individuo sia incluso nel campione sarà pari alla frazione di campionamento che indichiamo con* $ps_i = f = n/N$ (Rapporto tra le *n* unità del campionamento e le *N* unità della popolazione)

La selezione delle unità può avvenire in due modi, con ripetizione o senza ripetizione:

- Con *ripetizione* (o con remissione/ o con reinserimento), ricrea la situazione di partenza ad ogni estrazione in quanto mettiamo nuovamente l'unità *n* nella popolazione *N*. Si generano numeri casuali compresi tra 1 e N, con la teorica possibilità che uno stesso numero sia generato due o più volte e quindi che una stessa unità sia selezionata più di una volta nel campione. Questo, volendo indicarlo in formule, si scriverà: $C_N^n = N^n$;
- Senza *ripetizione* (o senza remissione/ o senza reinserimento), quando non inseriamo nuovamente l'unità *n*; quindi, avremo diverse situazioni (diverse

probabilità) ad ogni estrazione. Questo è il metodo più efficiente perché è equivalente a pescare *n* oggetti da un'urna che ne contiene *N*, senza reinserire l'oggetto estratto dopo ogni selezione. Volendolo indicare in formule scriveremo: $C_N^n = \binom{N}{n} = \frac{N!}{(N-n)!}$.

La *probabilità di inclusione* delle *n* unità sarà sempre costante.

La popolazione può, però, avere *infinite unità statistiche*, ossia da tutte quelle potenzialmente osservabili. Il carattere di interesse è la variabile casuale x, che potrà essere sintetizzata dal suo valore atteso e dalla varianza. *Qualunque metodo di estrazione si assimilerà ad un processo di estrazione con reintroduzione*.

Se, invece, avessimo più liste dalle quali selezionare le nostre unità, si parlerebbe di *"Campionamento casuale complesso"*.

Abbiamo diverse tipologie di campionamento casuale:

- *Campionamento casuale stratificato;*
- *Campionamento casuale a grappoli;*
- *Campionamento casuale a più stadi;*
- *Campionamento casuale sistematico.*

3.1 Il campionamento casuale stratificato

Il campionamento casuale stratificato va a *stratificare un collettivo in sottopopolazioni* (o sottogruppi che hanno tra di loro caratteristiche comuni), dette strati, il più possibile omogenee rispetto alla variabile (/parametro) oggetto di studio.

Gli strati sono formati sulla base di *un'informazione a priori già nota,* inoltre, per ogni strato viene effettuato un campionamento casuale semplice, questo garantisce di avere un campione più efficiente su cui basare le conclusioni relative alla popolazione. Questo è tanto più vero quanto più le differenze tra i gruppi (strati) sono rilevanti e le differenze tra unità all'interno dei gruppi sono continue.

Perché ci conviene stratificare?

- *Per evidenziare insiemi di unità significative;*
- *Per separare sottopopolazioni con caratteristiche speciali;*
- *Per utilizzare informazioni note, mantenendo la casualità dell'estrazione.*

Quando *N* popolazione è suddivisa in *H* strati, allora ogni strato sarà una N_h sottopopolazione, quindi volendolo scrivere in formule avremo: $\bigcup_{h=1}^{H} N_h = N$.

Inoltre, il campione C_h estratto dallo strato H è idoneo a rappresentare l'intera popolazione *N*, allora scriveremo: $\bigcup_{h=1}^{H} C_h = C$.

Come vediamo, le due formule sono simili, in quanto il campione è in grado di parlare a nome di tutta la popolazione.

(Esempio: Si vuole indagare sulla frequenza d'uso dei social media dei ragazzi universitari, ma faremo delle sottopopolazioni, dividendo i ragazzi per il Corso di Laurea che frequentano, questo permetterà di avere calcoli più precisi rispetto alla realtà)

3.2 Il campionamento casuale a grappoli

Con il campionamento causale a grappoli, anche denominato campionamento casuale a cluster, la popolazione viene *suddivisa in maniera naturale o artificiale in insiemi di unità legate da vincoli di continuità spaziale o di altra natura*. (Esempio: una classe, una casa, un bar etc.)

Questa tipologia di campionamento conviene specialmente quando si hanno grandi popolazioni.

Le unità elementari non sono scelte direttamente ma sono estratti alcuni *grappoli,* non le singole unità, ciò significa che tutte le unità del grappolo entrano a far parte del campione. (Esempio: classe, quindi tutti gli alunni presenti nella classe)

I cluster, o grappoli, sono sempre gruppi di unità della popolazione tendenzialmente simili tra loro e quindi, secondo lo stesso criterio, simili alla popolazione stessa.

Questa tipologia di campionamento pur preservando la componente di casualità necessaria a garantire la qualità dei risultati inferenziali, in certe situazioni, è più conveniente dal punto di vista pratico.

Abbiamo due tipologie di campionamento a grappoli:

- Campionamento a due stadi, che consiste nella doppia selezione casuale di cluster prima di unità poi, infatti, successivamente vedremo che prende il nome anche di *campionamento casuale a più stadi*;
- Campionamento ad uno stadio, quando vengono incluse nel campione tutte le unità dei cluster selezionati casualmente nella prima fase.

Il campionamento a grappoli è dato da una *mancata disponibilità della lista delle unità elementari* e da *un'opportunità di tipo organizzativo*.

3.4 Il campionamento casuale a più stadi

È possibile ricorrere al campionamento casuale a più stadi (anche detto *campionamento casuale a due stadi*, come abbiamo già visto prima) quando il collettivo di riferimento è strutturato gerarchicamente, ossia quando la popolazione di unità finali è contenuta in un insieme di unità di livello superiore.

Si utilizza quando non è disponibile una lista complessiva delle unità della popolazione.

Questo schema ricalca quello a grappoli, ma la differenza è che una volta selezionati i grappoli anche le unità elementari appartenenti a questi saranno selezionate attraverso un'estrazione casuale senza ripetizione.

3.5 Il campionamento casuale sistematico

Il campionamento casuale sistematico rappresenta una valida alternativa per motivi di praticità. Infatti, nella pratica, il CCS è poco utilizzato a causa dei costi e dei tempi elevati.

Con il campionamento casuale sistematico *si presuppone che si disponga di una lista di tutte le unità della popolazione dalla quale viene selezionata casualmente una*. Si procede, quindi, con una selezione ad intervalli fissi della unità da inserire nel campione scorrendo la lista della popolazione a partire dalla prima unità selezionata a caso.

Una volta formato il campione completo sarà possibile trattare i risultati ottenuti in modo analogo ai risultati provenienti dal CCS.

4. LA STATISTICA CAMPIONARIA E I SUOI PARAMETRI

Si definisce statistica campionaria una qualunque *funzione a valori reali delle osservazioni campionarie* X_n. Una statistica campionaria è rappresentata da una qualunque funzione calcolata su dati campionari $T_n = t(x_1, x_2, \dots, x_n)$.

Le statistiche campionarie non devono essere confuse con i parametri della popolazione, in quanto:

- Un parametro si riferisce all'intera popolazione e ne rappresenta una caratteristica;

- Una statistica campionaria dipende unicamente dalle osservazioni campionarie e, siccome queste sono variabili casuali, anch'essa sarà una variabile casuale con una propria distribuzione di probabilità.

In statistica andiamo poi a valutare la validità, la quale può essere:

- *Interna,* ossia misura quanto i risultati di uno stadio sono corretti per il campione di individui che sono stati studiati;
- *Esterna,* ossia il grado di generalizzabilità delle conclusioni tratte da uno studio campionario.

Il campione deve, quindi, essere eterogeneo (con varie caratteristiche), accurato (pochi errori), rappresentativo (caratteristiche che interessano più unità) e ampio.

5. DISTRIBUZIONE CAMPIONARIA, TLC E PARAMETRI CAMPIONARI

Una statistica T_n di variabili casuali (campione casuale), sarà anch'essa una *variabile casuale che assumerà valori diversi al variare delle realizzazioni campionarie*.

La distribuzione di probabilità di una statistica campionaria è detta distribuzione campionaria.

Quindi, ogni statistica è una sintesi delle variabili casuali campionari, con le quali troveremo legati alcuni parametri campionari:

1. *Media campionaria,* indicata con $\bar{x}$;
2. *Varianza campionaria,* indicata con S^2;
3. *Varianza campionaria corretta/deviazione standard campionaria corretta,* indicata con $\hat{s}_n^2$;
4. *Scarto quadratico medio campionario corretto,* indicato con $\hat{S}$;
5. *Grado di libertà,* indicato con ν, dato da *n meno 1*.

5.1 Il Teorema del Limite Centrale

Prima di analizzare i parametri campionari, è giusto parlare e analizzare il Teorema del Limite Centrale (TLC), il quale afferma che *la somma (o la media) di un elevato numero di variabili casuali è approssimativamente normale, indipendentemente dalla distribuzione soggiacente*.

Per convenzione avremo che se n>30 tratteremo la distribuzione come una distribuzione normale, se invece, dovessimo avere n<=30 allora useremo la distribuzione t di student, ma ricordiamoci che la t di student arriva fino a n=100. (Vedremo nei prossimi capitoli che si intende per distribuzione normale e t di student).

5.2 Media campionaria

Tra le statistiche di più frequente utilizzo, la media campionaria ha grande valore. In formule scriveremo: $\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_1$.

Ricordiamo però che:

- Il *valore atteso della media campionaria* è pari alla media della popolazione;
- La *varianza della media campionaria* è pari alla varianza della popolazione divisa per la dimensione del campione;
- L'*errore standard della media campionaria* è pari alla radice quadrata della varianza della media campionaria.

Ricorda che se la popolazione ha una *distribuzione normale* allora la distribuzione della media campionaria sarà ancora una *normale*. Se la tipologia di distribuzione non fosse nota, allora si può pervenire ad una buona conoscenza approssimata della distribuzione della media campionaria attraverso il Teorema del Limite Centrale, che dopo analizzeremo, ma in generale, diremo che per *n* abbastanza grande *(n>30, se è minore allora avremo t di student, ma la t di student potremmo anche utilizzarla per una n<=100)* la media campionaria si distribuisce approssimativamente come una variabile casuale normale.

Quando la dimensione del campione è abbastanza grande e la frazione di campionamento abbastanza piccola, allora per il TLC *la distribuzione della media campionaria può essere approssimata a quella normale*.

La distribuzione della media campionaria è meno variabile della distribuzione della popolazione e che, inoltre, la *riduzione della variabilità è tanto più forte quanto maggiore è la dimensione del campione*.

Ricorda che:

- n=1, *i risultati ottenuti con lo schema di campionamento con reintroduzione coincidono con quelli ottenuti nel campionamento senza reintroduzione*;
- n=N, *la varianza della media campionaria nello schema di campionamento senza reintroduzione è nulla e non si ha alcuna certezza legata al campionamento*;
- n<N, il fattore di correzione utilizzato nello schema è <1. Questo vuol dire che *la varianza campionaria della media campionaria nello schema senza reintroduzione è minore di quella che si ottiene nello schema con reintroduzione*;
- n<<N, il fattore di correzione per lo schema di campionamento senza reintroduzione è prossimo a 1. *La differenza tra i due schemi di campionamento sarà quindi trascurabile*.

Per analizzare tutto ciò verrà quindi fatta una media della media e la distribuzione campionaria potrà tenere conto o meno dell'ordine di distribuzione delle *n* unità.

7.3 Varianza campionaria

Sia una popolazione normale con media pari a μ e varianza pari a σ^2 e sia un campione casuale di dimensione *n*, avremo che:

- Si definisce varianza campionaria non corretta, quando *il suo valore atteso non è uguale alla varianza della popolazione*, ma differisce da esso della quantità σ^2 / μ ed avremo la seguente statistica: $\boldsymbol{S^2}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2$;
- Si definisce varianza campionaria corretta, quando il suo valore atteso della statistica $\hat{S}^2$ è proprio uguale alla varianza della popolazione x, ed avremo la seguente statistica: $\boldsymbol{\hat{s}_n^2}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{\chi}_n)^2$.

La differenza tra le due statistiche diventa trascurabile quando la dimensione campionaria è abbastanza grande.

5.4 Proporzione campionaria

Si definisce proporzione campionaria la seguente statistica: $\rho_n=\frac{1}{n}\sum_{i=1}^{n}x_i$.

Rappresenta *la media campionaria di un campione estratto da una popolazione bernoulliana*. Il suo valore medio equivale alla frequenza relativa o proporzione dei successi (rapporto tra successi e prove).

Quindi:

- Il *valore atteso della proporzione campionaria* è pari alla media della popolazione;
- La *varianza della proporzione campionaria* è pari alla varianza della popolazione divisa per la dimensione del campione;
- *L'errore standard della proporzione campionaria* è pari alla radice della varianza della popolazione campionaria.

Nonostante parliamo di una popolazione Bernoulliana (che nei prossimi capitoli analizzeremo meglio in che consiste), possiamo applicare il teorema del limite centrale, quindi per n abbastanza grande, la proporzione campionaria si distribuisce approssimativamente come una v.c. Normale.

ARGOMENTI, SOTTOFORMA DI DOMANDA, DEL CAPITOLO 3

- *Quando e chi ha iniziato a parlare del campionamento?*
- *Quali sono i limiti di un'indagine censuaria?*
- *In cosa consiste il campionamento o rilevazione campionaria?*
- *Cosa ci fornisce l'indagine campionaria?*
- *Cos'è il disegno di campionamento?*
- *Come si articolano le fasi di disegno di campionamento?*
- *Cosa si intende per controllo del tasso di risposta?*
- *Cos'è la popolazione? Qual è la differenza tra popolazione teorica e accessibile?*
- *Cos'è la lista della popolazione?*
- *Cos'è l'errore di campionamento?*
- *Quali altri errori conosciamo?*
- *Qual è la definizione di estrazione di un campione casuale?*
- *Cos'è il campionamento casuale semplice? Che si intende con ripetizione e senza ripetizione?*
- *Cos'è un campionamento casuale complesso?*
- *Cos'è un campionamento casuale stratificato?*
- *Cos'è un campionamento casuale a grappoli?*
- *Cos'è un campionamento casuale a più stadi?*
- *Cos'è un campionamento casuale sistematico?*
- *Cosa si intende per statistica campionaria e parametro?*
- *Cos'è una distribuzione campionaria?*
- *Cosa si intende per Teorema del Limite Centrale?*
- *Quali sono i parametri campionari?*

CAPITOLO 4: LA TEORIA DELLA STIMA

La teoria della stima ha come obiettivo principale la *stima di parametri, scalari o vettoriali, a partire da dati misurati e/o empirici, la cui distribuzione è influenzata dai valori effettivi assunti da tali parametri.*

In altre parole, essa include l'insieme degli strumenti statistici che consentono di derivare, a partire da un campione casuale, una conoscenza, affetta da errore, su alcune caratteristiche ignote della popolazione, ossia i parametri.

L'obiettivo è quello di trovare su un campione casuale x un valore, o un insieme di valori, che più si avvicinano al parametro θ o all'insieme di parametri ignoti di una popolazione definita su una variabile X.

Una funzione delle osservazioni campionaria è una variabile casuale che viene definita stimatore e il valore che assume lo stimatore nel campione è la stima, ossia la realizzazione campionaria della variabile casuale.

Lo stimatore è una *statistica campionaria* che viene utilizzata per stimare un parametro della popolazione e scegliere uno stimatore significa individuare quale statistica campionaria gode delle proprietà desiderabili in un processo di stima. Ciascun stimatore avrà una propria distribuzione campionaria che, in generale, ammetterà una media e una varianza dello stimatore dei parametri teta. Alcune statistiche campionaria assumono il ruolo di stimatore, perché godono di alcune proprietà che le rendono migliori rispetto ad altre per stimare un parametro della popolazione.

Le proprietà degli stimatori aiutano a capire qual è lo stimatore più desiderabile, per conoscere le proprietà degli stimatori è necessario, però, conoscere la loro struttura probabilistica, cioè la loro distribuzione di probabilità.

Le proprietà sono diverse per *piccoli campioni* e *grandi campioni*.

1. LA PROPRIETÀ DEI PICCOLI CAMPIONI

1.1 La correttezza

Uno stimatore è corretto e non distorto quando il suo valore atteso è pari al parametro da stimare e la sua distorsione (*BIAS*) è nulla.

Avremo che:

- *La media campionaria è uno stimatore corretto per la media della popolazione;*
- *La varianza campionaria è uno stimatore distorto per la varianza della popolazione;*
- *La varianza è uno stimatore corretti per la varianza della popolazione.*

1.2 L'efficienza

Dati due stimatori entrambi corretti, il più efficiente sarà quello con varianza minore, in quanto avrà stime più accurate e con meno errori.

L'errore quadratico medio misura proprio l'efficienza dello stimatore e consente di individuare lo stimatore migliore considerando anche quelli distorti.

Uno stimatore si dice efficiente in senso assoluto quando la sua varianza è minore a quella di qualsiasi altro stimatore corretto

Secondo il Teorema di Cramer-Rao si individua, sotto certe condizioni di regolarità, *un valore minimo della varianza di un qualunque stimatore corretto di un parametro teta*. Se uno stimatore T_n ha una varianza pari a questo limite allora è lo stimatore efficiente in senso assoluto, ossia il migliore tra i possibili stimatori.

È possibile dimostrare, inoltre, che sotto condizioni molto generali, ad ogni parametro teta è possibile associare un valore minimo della varianza dello stimatore corretto, al di sotto del quale *tale varianza non può scendere*, se lo stimatore raggiunge tale limite verrà definito stimatore a varianza minima. L'esistenza di una varianza minima per ogni parametro teta non implica necessariamente l'esistenza di uno stimatore a varianza minima. È possibile, comunque, ipotizzare che ad ogni parametro corrisponda almeno uno stimatore asintoticamente efficiente.

Abbiamo poi anche l'efficienza assoluta per la media campionaria, quando la *varianza sarà maggiore o uguale al rapporto tra la varianza e la popolazione e l'efficienza assoluta per la proporzione campionaria*, invece, avremo efficienza assoluta per la proporzione campionaria, quando *la varianza sarà maggiore o uguale al prodotto della probabilità di successo e di insuccesso rapportate alla popolazione*.

1.3 Sufficienza

Lo stimatore è sufficiente per teta se la distribuzione di probabilità condizionata dalla X non dipende dal parametro teta. Sarà sufficiente quando è in grado di

sintetizzare tutta l'informazione sulla popolazione X contenuta in un campione casuale estratto da essa.

2. LA PROPRIETÀ DEI GRANDI CAMPIONI

Le proprietà dei grandi campioni sono:

- correttezza asintotica, ossia quando il *valore dello stimatore è asintoticamente uguale al parametro da stimare quando la dimensione del campione tende verso infinito*;
- la seconda proprietà è la consistenza, ossia, *quando lo stimatore T per il parametro teta è consistente in media quadratica al crescere della dimensione campionaria se il suo errore quadratico medio tende a zero*, ciò significa che al crescere di n la varianza e la distorsione dello stimatore tenderanno contemporaneamente a zero. Avremo:
  - La consistenza in media quadratica della media campionaria, che è uno stimatore consistente in media quadratica;
  - La consistenza in media quadratica della proporzione campionaria.

3. STIMA PUNTUALE E STIMA INTERVALLARE

Nella teoria della stima distinguiamo:

- Stima puntuale, che assegna *un solo valore* al parametro teta;
- Stima intervallare, che assegna un *intervallo* che include il parametro teta, con una prefissata fiducia.

3.1 Stima puntuale

Con la stima puntuale lo stimatore *t* ottiene un particolare valore. L'accuratezza della stima puntuale dipende dall'errore standard della stima.

Lo stimatore, essendo una funzione delle osservazioni campionarie, è una variabile casuale. La conoscenza della sua distribuzione campionaria consente di capire se lo stimatore scelto produrrà, con elevata probabilità, *stime «vicine» al vero valore del parametro*. L'accuratezza delle stime, cioè la loro «vicinanza» attesa, considerando le proprietà dello stimatore sarà data dall' Errore Standard (SE).

Abbiamo diversi casi:

- Stima puntuale della media μ, la quale è la media calcolata nel particolare campione estratto, ossia, $\bar{\chi} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Con accuratezza della stima pari a $SE(\bar{\chi}) = \frac{\sigma}{\sqrt{n}}$;
- Stima puntuale della probabilità di successo per una distribuzione di Bernoulli, la quale è la media calcolata nel particolare campione estratto, avremo così, $\overline{p_n} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x}{n}$. Con accuratezza della stima pari a $SE(\rho_n) = \sqrt{\frac{\pi(1-\pi)}{n}}$.

Esistono dei limiti riguardo alla stima puntuale, in quanto l'approccio si caratterizza per un intrinseco punto di debolezza. La stima ottenuta sul campione potrebbe differire notevolmente dal parametro teta. Infatti anche se lo stimatore gode delle proprietà desiderabili, avremo che:

- La probabilità che la stima sia esattamente pari a θ è nulla;
- Per alcuni campioni rari si potrebbe ottenere una stima molto lontana dal parametro θ.

3.2 Stima intervallare

Nella procedura di stima per intervallo si fornisce un intervallo di stime plausibili al quale è associato un fissato *livello di fiducia*, o affidabilità. Con la stima per intervallo si determina un intervallo di valori intorno alla stima puntuale che ci si aspetta contenga, con un certo livello di fiducia, il valore del parametro incognito.

Avremo quindi uno *stimatore* e la sua *distribuzione* di probabilità nota, avremo un *livello di fiducia* indicato come 1-α, che è una probabilità che indica l'affidabilità della stima, e infine *abbiamo l'intervallo di confidenza*, ossia un *insieme di valori* di stima plausibili per il parametro teta.

L'intervallo di confidenza è un intervallo casuale che comprende il parametro incognito θ con una probabilità nota $P\ L1 \leq \theta \leq L2 = 1 - \alpha$. Inoltre, gli estremi dell'intervallo casuale, $L1$ ed $L2$, sono v.c., funzione dello stimatore T_n.

Dopo aver estratto il campione, si sostituiscono le osservazioni campionarie nella funzione dello stimatore ottenendo le realizzazioni delle v.c. estremi dell'intervallo, $l1$ ed $l2$. Una volta estratto il campione, non si potrà esprimere tale intervallo in termini probabilistici ed è per tal motivo che si parla di «intervalli di confidenza» (o di fiducia) $IC\ l1 \leq \theta \leq l2 = 1 - \alpha$.

4. IC PER LA MEDIA μ CON VARIANZA σ^2 NOTA

Data una popolazione di distribuzione normale con parametri media e varianza, di cui la media è ignota, e sapendo che $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. Avremo che la nostra variabile standardizzata $Z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ si distribuisce come una popolazione normale e sarà così possibile costruire *l'intervallo di confidenza per la media*:

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

5. IC PER LA MEDIA μ CON VARIANZA σ^2 NON NOTA

Data una popolazione di distribuzione normale con parametri media e varianza, ignoti, e sapendo che $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. Avremo che la nostra variabile standardizzata $T=\frac{\bar{x}-\mu}{S/\sqrt{n}}$ si distribuisce come una popolazione normale, ma con la quale calcoleremo la T di Student con n-1 gradi di libertà per una n<30. Essendo ignora la varianza nella procedura si usa il suo stimatore, ossia la varianza campionaria corretta, che facendo la radice, diventerà lo s.q.m corretto, e sarà così possibile costruire *l'intervallo di confidenza per la media*:

$$IC\left[\bar{x} - t_{\alpha/2,n-1}\frac{\hat{s}}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1}\frac{\hat{s}}{\sqrt{n}}\right] = 1 - \alpha$$

L'ampiezza dell'IC nel caso di varianza ignota è generalmente maggiore di quello in cui la varianza è nota, Da un punto di vista logico, l'utilizzo di una stima della varianza, in luogo della conoscenza del parametro σ 2 , introduce un'ulteriore fonte di incertezza nel processo inferenziale e quindi la stima sarà meno accurata.

Nel caso di grandi campioni, avremo che, secondo il TLC useremo la tavola standardizzata della z, quindi avremo come intervallo di confidenza per la media:

$$IC\left[\bar{x} - z_{\alpha/2}\frac{\hat{s}}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\hat{s}}{\sqrt{n}}\right] = 1 - \alpha$$

6. IC PER LA MEDIA μ CON VARIANZA σ^2 NOTA PER POPOLAZIONI NON NORMALI

La dipendenza dell'intervallo di confidenza dalla conoscenza e dal tipo di distribuzione della popolazione X perde importanza quando la numerosità campionaria è tale da ottenere una buona approssimazione asintotica dello stimatore, allora avremo che il valore standardizzato di Z sarà $Z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$. Con intervallo di confidenza per la media pari a:

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

7. IC PER LA MEDIA μ CON VARIANZA σ^2 NON NOTA PER POPOLAZIONI NON NORMALI

In questo caso avremo che la Z sarà $Z=\frac{\bar{x}-\mu}{S/\sqrt{n}}$ e che il nostro intervallo di confidenza sarà pari a

$$\left[\bar{x} - z_{\alpha/2}\frac{\hat{s}}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\hat{s}}{\sqrt{n}}\right]$$

8. IC PER LA PROPORZIONE π DI UNA POPOLAZIONE DI BERNOULLI

Data una popolazione di Bernoulli con proporzione ignota e sapendo che $Pn \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$. Per il TLC la variabile standardizzata Z si approssimerà asintoticamente a $Z = \frac{P_n - \pi}{\sqrt{\frac{P_n(1-P_n)}{n}}}$, *l'intervallo di confidenza per la proporzione* sarà:

$$IC\left[p_n - z_{\alpha/2}\sqrt{\frac{p_n(1-p_n)}{n}} \leq \pi \leq p_n + z_{\alpha/2}\sqrt{\frac{p_n(1-p_n)}{n}}\right] = 1-\alpha$$

9. ERRORE CAMPIONARIO E NUMEROSITÀ OTTIMALE

La semi lunghezza dell'intervallo di confidenza prende il nome di errore campionario e misura l'accuratezza della stima intervallare dipendendo da il livello di confidenza, la dimensione campionaria e la varianza. A parità di queste caratteristiche, la lunghezza dell'intervallo non cambierà al variare delle estrazioni campionarie.

Al crescere della popolazione campionaria, si riduce l'errore, l'ampiezza dell'IC e si accresce l'accuratezza dell'IC.

Partendo dalle relazioni tra dimensione campionaria ed errore campionario, è possibile determinare la dimensione ottimale per una prefissata precisione della stima.

- Media $\epsilon = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \rightarrow n^{ott} = \left(\frac{z_{\alpha/2}\,\sigma}{\epsilon}\right)^2$
- Proporzione $\epsilon = z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} \rightarrow n^{ott} = \left(\frac{z_{\alpha/2}\sqrt{\pi(1-\pi}}{\epsilon}\right)^2$

Per il calcolo della numerosità ottimale è necessario fissare il livello di confidenza e soprattutto conoscere la varianza della popolazione.

È possibile utilizzare stime delle varianze provenienti da indagini precedenti.

Inoltre, per popolazioni bernoulliane, in assenza di informazioni sul fenomeno, è possibile ipotizzare la massima incertezza $\pi\,(1-\pi) = 0.25$.

FORMULARIO INTERVALLO DI CONFIDENZA

Popolazione	Parametri noti	Parametri ignoti	Dimensione del campione	Intervallo di confidenza
Intervallo di confidenza per la media				
Normale	σ^2	μ	*Qualsiasi*	$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
Normale		μ, σ^2	*Piccolo*	$\bar{x} \pm t_{\alpha/2,n-1}\frac{\hat{s}}{\sqrt{n}}$
Qualsiasi		μ, σ^2	*Grande*	$\bar{x} \pm z_{\alpha/2}\frac{\hat{s}}{\sqrt{n}}$
Intervallo di confidenza per la proporzione				
Bernoulli		π	*Grande*	$p_n \pm z_{\alpha/2}\sqrt{\frac{p_n(1-p_n)}{n}}$

Caso	Errore campionario	Numerosità campionaria ottimale
Media	$\epsilon = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$n^{ott} = (\frac{Z_{\frac{\alpha}{2}}\sigma}{\epsilon})^2$
Proporzione	$\epsilon = Z_{\alpha/2}\sqrt{\frac{\pi\,(1-\pi\,)}{n}}$	$n^{ott} = (\frac{Z_{\frac{\alpha}{2}}\sqrt{\pi(1-\pi)}}{\epsilon})^2$

Quando non abbiamo il P-greco per il calcolo della numerosità ottimale, questo vale 0,25

Spiegazione esercizi Step by Step sugli intervalli di confidenza

Intervallo di confidenza per la media

- Scriviamo i dati che abbiamo, per questo esercizio ci serviranno:
  - *Varianza/Varianza campionaria;*
  - *Numero di unità del campione-fai attenzione a quanto sia numeroso!;*
  - *Media campionaria;*
  - *Livello di fiducia e valore di alfa;*;
- *Dividiamo alfa con 2 e cerchiamo il valore di Z nella tavola Z/Sottraiamo n con 1 e cerchiamo il valore della T nella tavola T;*
- Calcoliamo l'intervallo di confidenza con la formula più adatta;
- Scriviamo il commento finale.

Intervallo di confidenza per la proporzione

- Scriviamo i dati che abbiamo, per questo esercizio ci serviranno:
  - *Probabilità di successo;*
  - *Probabilità di insuccesso;*
  - *Numero di unità del campione-fai attenzione a quanto sia numeroso!;*
  - *Livello di fiducia e valore di alfa;*;
- *Dividiamo alfa con 2 e cerchiamo il valore di Z nella tavola Z;*
- Calcoliamo l'intervallo di confidenza;
- Scriviamo il commento finale.

Errore campionario e dimensione ottimale per la media

- Scriviamo i dati che abbiamo, per questo esercizio ci serviranno:
  - *Scarto quadratico medio;*
  - *Numero di unità del campione-fai attenzione a quanto sia numeroso!;*
  - *Livello di fiducia e valore di alfa;*;
- *Dividiamo alfa con 2 e cerchiamo il valore di Z nella tavola Z;*
- Calcoliamo l'errore campionario;
- Usiamo l'errore campionario come dato;
- Calcoliamo la numerosità ottimale.

Errore campionario e dimensione ottimale per la proporzione

- Scriviamo i dati che abbiamo, per questo esercizio ci serviranno:
  - *Proporzione;*
  - *Numero di unità del campione-fai attenzione a quanto sia numeroso!;*

    - *Livello di fiducia e valore di alfa;*;
- *Dividiamo alfa con 2 e cerchiamo il valore di Z nella tavola Z;*
- Calcoliamo l'errore campionario;
- Usiamo l'errore campionario come dato;
- Calcoliamo la numerosità ottimale.

Esercitazione

1. **Esercizio:**

Il prezzo in € di una camera alberghiera (doppia) in una nota località turistica è distribuito come una v.c. Normale con media ignota e varianza pari a 221,5.
Si vuole stimare il prezzo medio attraverso un intervallo di confidenza con un livello di confidenza $1 - \alpha = 0.95$.
A tal fine si estrae un CCS di dimensione 10:

(118.1, 133.3, 146.4, 166.7, 125.2, 106.3, 141.7, 139.2, 143.3, 141.1)

2. **Esercizio:**

Il prezzo in € di una camera alberghiera (doppia) in una nota località turistica è distribuito come una v.c. Normale con media e varianza ignota.
Si vuole stimare il prezzo medio attraverso un intervallo di confidenza con un livello di confidenza $1 - \alpha = 0.95$.
A tal fine si estrae un CCS di dimensione 10:

(118.1, 133.3, 146.4, 166.7, 125.2, 106.3, 141.7, 139.2, 143.3, 141.1)

3. **Esercizio:**

Il prezzo in € di una camera alberghiera (doppia) in una nota località turistica è distribuito come una v.c. Normale con media e varianza ignota
Si vuole stimare il prezzo medio attraverso un intervallo di confidenza con un livello di confidenza $1 - \alpha = 0.95$.
A tal fine si estrae un CCS di dimensione 38, con media campionaria corretta pari a 140 e varianza campionaria corretta pari a 260.

4. **Esercizio:**

Un tour operator internazionale vuole stimare la proporzione di clienti di età compresa tra i 18 e i 30 anni che si dichiarano soddisfatti dei servizi offerti. Si vuole quindi stimare la proporzione di soddisfatti attraverso un intervallo di confidenza con un livello di confidenza $1 - \alpha = 0.95$. A tal fine si estrae un CCS di dimensione 150. Tra questi 87 si dischiarano soddisfatti.

5. **Esercizio:**

Il National Center for Health Statistics riporta che il peso medio di un campione di uomini statunitensi di età compresa tra 20 e 74 anni è pari a 87kg. Nel 1960 la media era di 75kg. Supponendo che la dimensione campionaria sia 1200 e che la deviazione standard del peso della popolazione sia pari a 8,25kg, determinare:

a) l'errore campionario al 95% di confidenza per la stima intervallare del peso medio maschile nella popolazione;

b) determinare la numerosità campionaria per ottenere un intervallo di ampiezza non superiore a 5.

6. Esercizio:

Supponiamo che la dimensione campionaria dei clienti della Palmetto Solutions, sia 500 e che 120 dei clienti intervistati siano favorevoli alla modifica del prodotto al livello di confidenza del 90%.

a) Calcolare l'errore campionario;

b) Si supponga un errore non superiore a 0,02 e un livello di confidenza del 95%. Si calcoli la dimensione campionaria ottimale con stima di probabilità di successo pari a 0,38.

7. Esercizio:

Si desidera stimare gli anni di anzianità media dei dipendenti che lavorano alla Kaneko Ltd. Viene selezionato casualmente un campione di 20 lavoratori dell'azienda. L'anzianità media campionaria è di 13,6 anni. Supponiamo di sapere che la deviazione standard campionaria sia pari a 5,2 anni. Determinare un intervallo di confidenza con livello di fiducia del 80% e commentare il risultato.

8. Esercizio:

Si desidera stimare il punteggio medio SAT per tutti gli studenti che hanno frequentato il corso di preparazione al test SAT di Ethan-Davies negli ultimi due anni. Si sceglie un CCS di 100 studenti da un elenco completo di tutti gli studenti di Ethan-Davies che hanno frequentato il corso negli ultimi due anni, rilevando che il punteggio medio SAT per il campione è pari a 1940 punti. Supponiamo di conoscere che la deviazione standard della popolazione sia di 83 punti.

a) Produrre un intervallo di confidenza del 95% del punteggio medio al test SAT per la popolazione degli studenti Ethan-Davis e commentare il risultato.;

b) Calcolare l'errore campionario;

c) Determinare la numerosità ottimale quando l'intervallo abbia un'ampiezza dimezzata.

9. Esercizio:

Tra il 9 e il 16 febbraio Harris Interactive Polling ha realizzato un sondaggio telefonico intervistando un campione di 20 adulti selezionati in modo casuale. Una delle domande poste era: Pensi che le compagnie aree dovrebbero fornire dati personali di identificazione riguardanti i passeggeri alle agenzie federali per migliorare la sicurezza?. Il 73% ha risposto di si.

a) Costruire un intervallo di confidenza al 90% per la proporzione di adulti che la pensa in questo modo e commentare;
b) Calcolare l'errore campionario;
c) Individuare la numerosità ottimale con un errore non superiore a 0,05.

10. Esercizio:

L'ufficio di controllo qualità di Hershey Mechanical vuole stimare la vita media delle nuove guarnizioni O-ring che l'azienda ha appena introdotto nel mercato. Su un campione casuale di 40 O-ring si osserva una vita media di 1560 ore. La deviazione standard campionaria è di 68 ore. Costruire e commentare l'intervallo di confidenza con un livello di fiducia del 95%.

ARGOMENTI, SOTTOFORMA DI DOMANDA, DEL CAPITOLO 4

- *Che cosa è la teoria della stima?*
- *Che cosa è la stima e cosa è lo stimatore?*
- *Quali sono le proprietà dei piccoli campioni?*
- *Che cosa si intende per correttezza?*
- *Che cosa si intende per efficienza?*
- *Che cosa si intende per sufficienza?*
- *Quali sono le proprietà dei grandi campioni?*
- *Che cosa si intende per correttezza asintotica?*
- *Che cosa si intende per consistenza? Quali tipologie ne abbiamo?*
- *Che cosa si intende per stima puntuale?*
- *Cosa si intende e quali sono i limiti della stima puntuale?*
- *Che cosa si intende per stima intervallare?*
- *Quali sono le varie tipologie di intervalli di confidenza e che variabili casuali usiamo?*
- *Che cosa si intende per errore campionario?*
- *Che cosa si intende per numerosità ottimale?*

CAPITOLO 5: TEORIA DEI TEST STATISTICI

Nell'inferenza statistica una delle maggiori aree di interesse è la verifica delle ipotesi statiche, la quale trova ampia applicazione in *problemi reali e nelle aree più svariate*, come l'università, la medicina e la ricerca.

Si tratta di *verificare la plausibilità di una ipotesi attraverso l'osservazione dei dati* e eventualmente smentire o confermare l'ipotesi.

Si chiama test statistico ogni *procedura atta a verificare la conformità (o plausibilità) probabilistica tra un campione casuale e una popolazione* al fine di determinare l'accettazione o il rigetto di un'ipotesi riguardante la popolazione.
L'accettazione o il rigetto di un'ipotesi avviene, quindi, *verificando la ragionevolezza della stessa rispetto alla evidenza empirica*.

Una prima formalizzazione al problema dei test è stata data da R.A. Fisher con l'approccio dei test di significatività basati sul *concetto di discrepanza statistica* tra l'ipotesi formulata (ipotesi nulla) e i dati osservati. Successivamente, J. Neyman e E.S. Pearson formularono la moderna teoria dei test statistici esplicitando *l'esistenza di un'ipotesi alternativa a quella nulla, ossia lo status quo del fenomeno.* Quest'ultima è proprio quella che vogliamo analizzare.

Nel test di ipotesi, viene fatta un' affermazione, chiamata ipotesi, riguardante alcuni aspetti della popolazione e viene selezionato un campione nel tentativo di stabilire se l' affermazione è corretta o meno.

1. TEST E IPOTESI STATISTICHE

La definizione di test statistico si basa sui seguenti passi:

- *Definizione delle ipotesi statistiche;*
- *Individuazione delle regioni di accettazione e di rigetto;*
- *Scelta di una statistica test;*
- *Scelta del livello di significatività del test;*
- *Definizione della regola di decisione del test.*

1.1 Le ipotesi statistiche

Il primo passo nella costruzione di un test statistico consiste nella formulazione delle due possibili ipotesi. Si definisce ipotesi statistica una congettura riguardante un parametro θ sulla popolazione X.

Nell'approccio Neyman-Pearson, si distinguono due ipotesi contrapposte:

- L'ipotesi nulla, indicata con H_0, con la quale si intende *l'ipotesi preesistente* all'osservazione dei dati campionari, ossia quella ritenuta vera fino a prova contraria. Si dice nulla in quanto essa rappresenta lo «status quo»;
- L'ipotesi alternativa, indicata con H_1, con la quale si intende *l'ipotesi formulata sulla base di dubbi per i quali si cerca una conferma/smentita* sulla base di un'evidenza empirica L'aggettivo «alternativa» sta ad indicare la sua *contrapposizione allo «status quo»* definito nell'ipotesi nulla.

Abbiamo così uno spazio parametrico Θ, ossia, *l'insieme di tutti i possibili valori che può assumere* θ. La formulazione delle ipotesi nulla e alternativa definisce una *partizione dello spazio parametrico* in due parti.

Le ipotesi possono essere di due tipologie:

- Un'ipotesi semplice che individua un valore puntuale per θ e quindi specifica completamente la popolazione *X, l'ipotesi nulla è un ipotesi semplice;*
- Un'ipotesi composta che individua un intervallo di valori per θ e quindi non specifica un'unica popolazione ma un insieme di popolazioni, *l'ipotesi alternativa è un'ipotesi composta.* Questa può essere:
    - Unidirezionale, quando individua, per il parametro θ, un *unico intervallo di valori* $\theta \leq \theta_0$ oppure $\theta \geq \theta_0$;
    - Bidirezionale, quando individua, per il parametro θ, *due intervalli di valori* $\theta \neq \theta_0$ che equivale a $\theta \leq \theta_0 \cup \theta \geq \theta_0$.

1.2 Regioni di accettazione e di rifiuto

Il campione osservato ci consente di individuare la regione di accettazione o di rifiuto dell'ipotesi nulla. Se 'informazione ricavata dal campione è in *evidente contrasto con l'ipotesi nulla* allora si propenderà per il rifiuto della stessa. In caso contrario, una *«somiglianza» dell'informazione campionaria all'ipotesi nulla* porterà ad una decisione di accettazione della stessa. Per individuare la regione di rifiuto però dobbiamo comprendere la tipologia della distribuzione della popolazione e

calcolare la nostra *statistica test*. La regione di rifiuto è determinata dall'insieme dei campioni a cui sono associati statistiche test che assumono valori nelle code della distribuzione e quindi con probabilità bassa.
Nell'ambito grafico possiamo determinare i valori critici, che possiamo indicare con C_1 e C_2, queste determinano il limite tra le regioni di accettazione (A) e di rifiuto (R). La scelta delle soglie è legata alla probabilità α scelta ed il valore α è detto livello di significatività del test.

1.3 La statistica test

Una statistica test è la *statistica campionaria utilizzata per effettuare la verifica di conformità probabilistica del campione rispetto alle ipotesi formulate sulla popolazione*.

La scelta della statistica test è speculare a quella dello stimatore. Essa deve godere di *alcune proprietà desiderabili affinché la procedura di verifica delle ipotesi sia accurata*. In altre parole, una statistica campionaria (che gode delle proprietà desiderabili) assumerà il ruolo di stimatore quando è impiegata in un processo di stima mentre assumerà il ruolo di statistica test quando è invece impiegata in una procedura di test statistico.

La statistica test crea una corrispondenza tra lo spazio parametrico e lo spazio campionario.

1.4 Errori del primo e del secondo tipo e la potenza del test

Nell'eseguire un test statistico si possono distinguere quattro diverse situazioni in cui si evincono due diversi tipi di errore:

- L'errore di primo tipo, indicato con α che si commette quando *rifiutiamo l'ipotesi nulla quando questa è vera*, ed è l'errore più grave che possiamo commettere!
- L'errore di secondo tipo, indicato con β che si commette quando *accettiamo l'ipotesi nulla quando questa è falsa*.

		Decisione del test	
		$Accetto\ H_0$	$Rifiuto\ H_0$
Stato della realtà	H_0 è $vera$	Decisione corretta $1-\alpha$	Errore di I° tipo E_1 α
	H_0 è $falsa$	Errore di II° tipo E_2 β	Decisione corretta $1-\beta$

Nel caso di decisione corretta 1- β, *ossia rifiutare l'ipotesi nulla quando è falsa,* essa è detta anche potenza del test. Il valore della potenza del test dipende da α e dal valore assunto dall'ipotesi alternativa θ_1 e dalla dimensione campionaria n. Inoltre, la funzione di potenza del test $\boldsymbol{\pi\ \theta}$ è la funzione che *descrive la probabilità* $1-\beta$ *al variare del valore dell'ipotesi alternativa*. Abbiamo che:

- *La potenza raggiunge il suo minimo, pari ad* α *quando le due ipotesi coincidono;*
- *La potenza cresce al divergere delle due ipotesi;*
- *La pendenza della funzione* $\boldsymbol{\pi\ \theta}$ *dipende dalla numerosità campionaria, maggiore è* n *e maggiore sarà* $1-\beta$ *a parità dele altre scelte in merito al test.*

2. TEST UNIFORMEMENTE PIÙ POTENTI

Tra tutti i possibili test statistici si preferirà quello che *presenta il più piccolo valore sia di* α *sia di* β *per ogni possibile valore* θ del sottospazio parametrico.

Un test è detto uniformemente più potente a livello $\boldsymbol{\alpha}$ se per un dato $\alpha\ \pi*(\theta) \geq \pi(\theta)$ per ogni $\theta \in \Theta_1$ rispetto a qualunque altro test di pari significatività per θ.

Un test unidirezionale è uniformemente più potente rispetto ad un test bidirezionale effettuato nelle stesse condizioni.

3. INTERVALLI DI CONFIDENZA E TEST STATISTICI

Esiste un importante legame tra intervalli di confidenza e test di ipotesi. È possibile costruire test a partire da intervalli di confidenza appropriati.

In altre parole, effettuare un test bidirezionale equivale a *costruire un intervallo di confidenza e verificare poi se l'ipotesi nulla è compresa o meno nello stesso*.

Nel caso in cui l'ipotesi nulla è compresa nell'intervallo allora si accetterà H_0 ritenendola maggiormente conforme all'osservazione empirica. Viceversa, se il valore non è compreso nell'intervallo si propenderà per il rifiuto dell'ipotesi nulla.

4. P-VALUE

Il p-value (o valore p) di un test di verifica d'ipotesi è dato dalla *probabilità di osservare un valore della statistica test uguale o più estremo del valore ottenuto dal*

campione, sotto l'ipotesi nulla. Talvolta viene chiamato anche livello di significatività osservato, ossia essa è la nostra statistica test osservata (Z_{oss}/T_{oss}).

Il p-value non è una quantità fissa come il livello di significatività, ma al contrario *è un valore che misura l'evidenza fornita dai* dati contro l'ipotesi nulla.

Esso è il minimo livello di significatività per il quale l'ipotesi nulla verrebbe rifiutata, in base ai dati campionari.

Minore sarà il valore del p-value e più forte sarà l'evidenza verso il rifiuto di H_0.

Esso ci dà una regola di decisione, o meglio, rifiuto H_0 se Pvalue ≤ α. Questa regola non cambia al mutare della natura del test. Il valore del p-value e il suo confronto con α fornisce una immediata evidenza della forza con cui si rifiuta H_0.

α esprime la soglia di rischio (di commettere $E1$) massima che si intende assumere nello svolgimento di un test. Esso è quindi il livello di significatività scelto apriori e mi consente di poter osservate il mio valore teorico (Z_{teo}/T_{teo}).

Pvalue esprime invece il rischio di incorrere in $E1$ che, una volta estratto il campione, realmente si assume se si decide di rifiutare $H0$. Esso è quindi il livello di significatività osservato.

Diremo quindi che:

- Se il p-value è minore o uguale del livello di significatività alfa, l'ipotesi nulla deve essere rifiutata in favore dell'ipotesi alternativa;
- Se il p-value è maggiore del livello di significatività alfa, l'ipotesi nulla va accettata.

La regola decisionale basata sul P-value è sempre uguale, la differenza da un problema all'altro sta nel calcolo del P-value.

Ipotesi alternativa	Regione critica	P-value
$\boldsymbol{\mu < \mu_0}$	$z_{oss} \leq -z_\alpha$	P(Z≤ z_{oss})
$\boldsymbol{\mu > \mu_0}$	$z_{oss} \geq z_\alpha$	P(Z≥ z_{oss})

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

5. TEST PER LA μ MEDIA DI UNA POPOLAZIONE NORMALE CON VARIANZA NOTA

Se avessimo una distribuzione normale con media da sottoporre a verifica mediante ipotesi e la varianza che è nota, avremo che ci calcoleremo in questo modo il *valore osservato*, ossia la *statistica test*:

$$Z_{oss} = Z = \frac{\bar{x} - \mu_0}{\sigma \,/\, \sqrt{n}}$$

Il numeratore rappresenta lo scostamento della media campionaria dal valore ipotizzato nell'ipotesi nulla. Invece, il denominatore rappresenta l'errore standard di $\bar{x}$.

La statistica test osservata misura, quindi, *lo scostamento osservato nel campione rispetto allo scostamento medio dello stimatore*. La definizione di una regola di decisione basata su Z significa quindi valutare se lo scostamento rientra nella variabilità media (delle fluttuazioni campionarie) oppure è frutto di un significativo allontanamento dallo status quo.

Il *valore teorico* sarà invece uguale a Z_α.

Avremo un'ipotesi nulla $H_0: \mu = \mu_0$.

Tipo di test	Ipotesi alternativa	Regione di rifiuto	
		In forma standardizzata	*In forma esplicita*
Unidirezionale verso destra	$H_1: \mu > \mu_0$	$Z \geq z_\alpha$	$\bar{X} \geq \mu_0 + z_\alpha \sigma/\sqrt{n}$
Unidirezionale verso sinistra	$H_1: \mu < \mu_0$	$Z \leq -z_\alpha$	$\bar{X} \leq \mu_0 - z_\alpha \sigma/\sqrt{n}$
Bidirezionale	$H_1: \mu \neq \mu_0$	$\lvert Z \rvert \geq z_{\alpha/2}$	$\bar{X} \leq \mu_0 - z_{\alpha/2} \sigma/\sqrt{n}$ oppure $\bar{X} \geq \mu_0 + z_{\alpha/2} \sigma/\sqrt{n}$

6. TEST PER LA μ MEDIA DI UNA POPOLAZIONE NORMALE CON VARIANZA IGNOTA

Se avessimo una distribuzione normale con media da sottoporre a verifica mediante ipotesi e la varianza ignota, avremo che ci calcoleremo in questo modo *il valore osservato*, ossia la *statistica test*:

$$T_{oss} = T = \frac{\bar{x} - \mu_0}{S \,/\, \sqrt{n}}$$

Dove S è lo stimatore della varianza.

In questo caso la statistica test si distribuirà, analogamente al caso degli intervalli di confidenza, come una v.c. $t - Student$ con $n - 1$ gradi di libertà.

Il *valore teorico* sarà invece uguale a $t_{\alpha,n-1}$.

Avremo un'ipotesi nulla $H_0: \mu = \mu_0$.

Tipo di test	Ipotesi alternativa	Regione di rifiuto	
		In forma standardizzata	*In forma esplicita*
Unidirezionale verso destra	$H_1: \mu > \mu_0$	$T \geq t_{\alpha,n-1}$	$\bar{X} \geq \mu_0 + t_{\alpha,n-1} s/\sqrt{n}$
Unidirezionale verso sinistra	$H_1: \mu < \mu_0$	$T \leq -t_{\alpha,n-1}$	$\bar{X} \leq \mu_0 - t_{\alpha,n-1} s/\sqrt{n}$
Bidirezionale	$H_1: \mu \neq \mu_0$	$\lvert T \rvert \geq t_{\alpha/2,n-1}$	$\bar{X} \leq \mu_0 - t_{\alpha/2,n-1} s/\sqrt{n}$ oppure $\bar{X} \geq \mu_0 + t_{\alpha/2,n-1} s/\sqrt{n}$

7. TEST PER LA PROPORZIONE

Se avessimo una distribuzione di Bernoulli con parametro principale la probabilità di successo e sapendo che $\rho_n \sim N\left(\pi, \frac{\pi(1-\tilde{n})}{n}\right)$. Per il *teorema del limite centrale* calcoleremo il nostro *valore osservato*, ossia la nostra *statistica test*, mediante la Z e avremo:

$$Z_{oss} = Z = \frac{\rho_n - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

In questo caso quindi il test può essere effettuato solo quando la dimensione campionaria è sufficientemente grande da consentire l'applicazione del Teorema del Limite Centrale.

Il *valore teorico* sarà invece uguale a Z_α.

Avremo un'ipotesi nulla $H_0: \pi = \pi_0$.

Tipo di test	Ipotesi alternativa	Regione di rifiuto	
		In forma standardizzata	*In forma esplicita*
Unidirezionale verso destra	$H_1: \pi > \pi_0$	$Z \geq z_\alpha$	$P_n \geq \pi_0 + z_\alpha \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$
Unidirezionale verso sinistra	$H_1: \pi < \pi_0$	$Z \leq -z_\alpha$	$P_n \leq \pi_0 - z_\alpha \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$
Bidirezionale	$H_1: \pi \neq \pi_0$	$\lvert Z \rvert \geq z_{\alpha/2}$	$P_n \leq \pi_0 - z_{\alpha/2} \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$ oppure $P_n \geq \pi_0 + z_{\alpha/2} \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$

8. TEST PER LA DIFFERENZA TRA DUE μ MEDIE DI DUE POPOLAZIONE NORMALE CON VARIANZA NOTA

Se avessimo due popolazioni normali, ci calcoleremo il nostro *valore osservato*, ossia la *nostra statistica test*:

$$Z = \frac{(\bar{X}_A - \bar{X}_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

Il *valore teorico* sarà invece uguale a Z_α.

Avremo un'ipotesi nulla $H_0: \mu_A - \mu_B = 0$.

Tipo di test	Ipotesi alternativa	Regione di rifiuto	
		In forma standardizzata	*In forma esplicita*
Unidirezionale verso destra	$H_1: \mu_A - \mu_B > 0$	$Z \geq z_\alpha$	$\bar{X}_A - \bar{X}_B \geq z_\alpha \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$
Unidirezionale verso sinistra	$H_1: \mu_A - \mu_B < 0$	$Z \leq -z_\alpha$	$\bar{X}_A - \bar{X}_B \leq -z_\alpha \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$
Bidirezionale	$H_1: \mu_A - \mu_B \neq 0$	$\lvert Z \rvert \geq z_{\alpha/2}$	$\bar{X}_A - \bar{X}_B \geq z_{\alpha/2} \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$ oppure $\bar{X}_A - \bar{X}_B \leq -z_{\alpha/2} \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$

9. TEST PER LA DIFFERENZA TRA DUE μ MEDIE DI DUE POPOLAZIONE NORMALE CON VARIANZE IGNOTE MA UGUALI

Se avessimo due popolazioni normali con varianze uguali ma ignote, ci calcoleremo il nostro *valore osservato*, ossia la *nostra statistica test*:

$$T = \frac{(\bar{X}_A - \bar{X}_B)}{S_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

Dove *Sp* 2 è lo stimatore della varianza comune già illustrato in precedenza nelle slide relative alla costruzione degli intervalli di confidenza.

Il *valore teorico* sarà invece uguale a $T_{\alpha,(n_\alpha + n_B - 2)}$.

Avremo un'ipotesi nulla $H_0: \mu_A - \mu_B = 0$.

Tipo di test	Ipotesi alternativa	Regione di rifiuto	
		In forma standardizzata	*In forma esplicita*
Unidirezionale verso destra	$H_1: \mu_A - \mu_B > 0$	$T \geq t_{\alpha,(n_A+n_B-2)}$	$\bar{X}_A - \bar{X}_B \geq t_\alpha S_p \sqrt{\frac{1}{n_A}+\frac{1}{n_B}}$
Unidirezionale verso sinistra	$H_1: \mu_A - \mu_B < 0$	$T \leq -t_{\alpha,(n_A+n_B-2)}$	$\bar{X}_A - \bar{X}_B \leq -t_\alpha S_p \sqrt{\frac{1}{n_A}+\frac{1}{n_B}}$
Bidirezionale	$H_1: \mu_A - \mu_B \neq 0$	$\lvert T \rvert \geq t_{\alpha/2,(n_A+n_B-2)}$	$\bar{X}_A - \bar{X}_B \geq t_{\alpha/2} S_p \sqrt{\frac{1}{n_A}+\frac{1}{n_B}}$ oppure $\bar{X}_A - \bar{X}_B \leq -t_{\alpha/2} S_p \sqrt{\frac{1}{n_A}+\frac{1}{n_B}}$

10. TEST PER LA DIFFERENZA TRA DUE μ MEDIE DI DUE POPOLAZIONE NORMALE CON VARIANZE IGNOTE

Se avessimo due popolazioni normali con varianze ignote, ci calcoleremo il nostro *valore osservato*, ossia la *nostra statistica test*:

$$Z = \frac{(\bar{X}_A - \bar{X}_B)}{\sqrt{\frac{S_A^2}{n_A}+\frac{S_B^2}{n_B}}}$$

In questo caso quindi il test può essere effettuato solo quando entrambe le dimensioni campionarie sono sufficientemente grandi da consentire l'applicazione del Teorema del Limite Centrale.

Il *valore teorico* sarà invece uguale a Z_α.

Avremo un'ipotesi nulla $H_0: \mu_A - \mu_B = 0$.

Tipo di test	Ipotesi alternativa	Regione di rifiuto	
		In forma standardizzata	*In forma esplicita*
Unidirezionale verso destra	$H_1: \mu_A - \mu_B > 0$	$Z \geq z_\alpha$	$\bar{X}_A - \bar{X}_B \geq z_\alpha \sqrt{\frac{S_A^2}{n_A}+\frac{S_B^2}{n_B}}$
Unidirezionale verso sinistra	$H_1: \mu_A - \mu_B < 0$	$Z \leq -z_\alpha$	$\bar{X}_A - \bar{X}_B \leq -z_\alpha \sqrt{\frac{S_A^2}{n_A}+\frac{S_B^2}{n_B}}$
Bidirezionale	$H_1: \mu_A - \mu_B \neq 0$	$\lvert Z \rvert \geq z_{\alpha/2}$	$\bar{X}_A - \bar{X}_B \geq z_{\alpha/2} \sqrt{\frac{S_A^2}{n_A}+\frac{S_B^2}{n_B}}$ oppure $\bar{X}_A - \bar{X}_B \leq -z_{\alpha/2} \sqrt{\frac{S_A^2}{n_A}+\frac{S_B^2}{n_B}}$

11. TEST PER LA μ MEDIA DI UNA POPOLAZIONE NON NORMALE

Se avessimo una distribuzione non normale e ignota, ma con media da sottoporre a verifica, avremo un *valore osservato*, ossia una *statistica test:*

$$Z_{oss} = Z = \frac{\bar{x} - \mu_0}{S / \sqrt{n}}$$

In questo caso quindi il test può essere effettuato solo quando la dimensione campionaria è sufficientemente grande da consentire l'applicazione del Teorema del Limite Centrale.

Si deve, quindi, stabilire la dimensione campionaria. Solitamente, nella fase di definizione di un test statistico, una delle prime scelte da affrontare è quello di definire la dimensione ottimale del campione affinché il *test possa ritenersi sufficientemente accurato*.

Per calcolare la dimensione ottimale bisogna specificare:

- il livello di significatività del test;
- il valore della media dell'ipotesi alternativa e il corrispondente valore dell'errore di secondo tipo;
- una stima iniziale della varianza.

Il calcolo della dimensione campionaria con relativa ipotesi alternativa sarà:

Tipo di test	Ipotesi alternativa	Calcolo della numerosità ottimale
Unidirezionale verso destra	$H_1: \mu > \mu_0$	$n = \left[\frac{\sigma(z_\alpha + z_\beta)}{\mu_1 - \mu_0}\right]^2$
Unidirezionale verso sinistra	$H_1: \mu < \mu_0$	
Bidirezionale	$H_1: \mu \neq \mu_0$	$n = \left[\frac{\sigma(z_{\alpha/2} + z_\beta)}{\mu_1 - \mu_0}\right]^2$

12. TEST NON PARAMETRICI

I test descritti sino ad ora sono detti parametrici in quanto il loro utilizzo *necessita della assunzione apriori che la popolazione appartenga a una determinata famiglia di distribuzioni*. Si chiamano test non parametrici, quelle procedure di verifica delle ipotesi che *prescindono dalla conoscenza della distribuzione della popolazione*.

Il test di indipendenza è uno dei test parametrici più utilizzato. Esso permette di *verificare se tra due variabili sussiste o meno associazione*. Si applica quindi tra due

variabili la cui distribuzione congiunta può essere rappresentata in una tabella di contingenza. È anche detto «Test chi quadrato di Pearson».

Siano X e Y due variabili, rispettivamente con K e H modalità, *l'obiettivo del test è quello di formulare due ipotesi sulla indipendenza tra le due variabili.*

Avremo quindi due ipotesi:

- H_0: *X e Y sono indipendenti*
- H_1: *X e Y non sono indipendenti*

E di verificare la conformità probabilistica di tali ipotesi attraverso l'estrazione campionaria e la misurazione dell'associazione.

La *statistica test* utilizzata sarà:

$$x^2 = \sum_{i}^{k} \sum_{j}^{H} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

Dove nij sono le frequenze osservate nella tabella di contingenza campionaria, invece, $\hat{n}_{ij}$ sono le frequenze attese ottenute come $\hat{n}_{ij} = \frac{n_i \cdot n_j}{N}$.

Ipotesi alternativa	Regione di rifiuto
$H_1: \pi_{ij} \neq \pi_i \pi_j$	$\chi^2 \geq \chi^2_{\alpha,(K-1)(H-1)}$

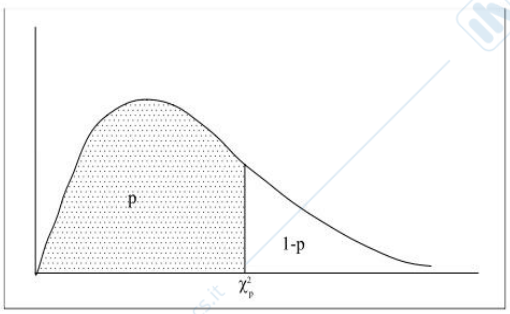
Ricordiamo la tabella del chi quadrato.

Nel grafico finale 1-p rappresenta la regione di rifiuto, invece p la regione di accettazione.

VALORI PERCENTILI (χ^2_P) PER LA DISTRIBUZIONE CHI-QUADRATO CON ν GRADI DI LIBERTÀ

	α	α	α	α	α	α	α	α	α	α	α	α	α	α
	0.995	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01	0.005	0.001
ν	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P
1	0.0000	0.0002	0.0010	0.0039	0.0158	0.102	0.455	1.323	2.706	3.841	5.024	6.635	7.879	10.827
2	0.0100	0.0201	0.0506	0.1026	0.211	0.575	1.386	2.773	4.605	5.991	7.378	9.210	10.597	13.815
3	0.0717	0.1148	0.2158	0.352	0.584	1.213	2.366	4.108	6.251	7.815	9.348	11.345	12.838	16.266
4	0.207	0.297	0.484	0.711	1.064	1.923	3.357	5.385	7.779	9.488	11.143	13.277	14.860	18.466
5	0.412	0.554	0.831	1.145	1.610	2.675	4.351	6.626	9.236	11.070	12.832	15.086	16.750	20.515
6	0.676	0.872	1.237	1.635	2.204	3.455	5.348	7.841	10.645	12.592	14.449	16.812	18.548	22.457
7	0.989	1.239	1.690	2.167	2.833	4.255	6.346	9.037	12.017	14.067	16.013	18.475	20.278	24.321
8	1.344	1.647	2.180	2.733	3.490	5.071	7.344	10.219	13.362	15.507	17.535	20.090	21.955	26.124
9	1.735	2.088	2.700	3.325	4.168	5.899	8.343	11.389	14.684	16.919	19.023	21.666	23.589	27.877
10	2.156	2.558	3.247	3.940	4.865	6.737	9.342	12.549	15.987	18.307	20.483	23.209	25.188	29.588
11	2.603	3.053	3.816	4.575	5.578	7.584	10.341	13.701	17.275	19.675	21.920	24.725	26.757	31.264
12	3.074	3.571	4.404	5.226	6.304	8.438	11.340	14.845	18.549	21.026	23.337	26.217	28.300	32.909
13	3.565	4.107	5.009	5.892	7.041	9.299	12.340	15.984	19.812	22.362	24.736	27.688	29.819	34.527
14	4.075	4.660	5.629	6.571	7.790	10.165	13.339	17.117	21.064	23.685	26.119	29.141	31.319	36.124
15	4.601	5.229	6.262	7.261	8.547	11.037	14.339	18.245	22.307	24.996	27.488	30.578	32.801	37.698
16	5.142	5.812	6.908	7.962	9.312	11.912	15.338	19.369	23.542	26.296	28.845	32.000	34.267	39.252
17	5.697	6.408	7.564	8.672	10.085	12.792	16.338	20.489	24.769	27.587	30.191	33.409	35.718	40.791
18	6.265	7.015	8.231	9.390	10.865	13.675	17.338	21.605	25.989	28.869	31.526	34.805	37.156	42.312
19	6.844	7.633	8.907	10.117	11.651	14.562	18.338	22.718	27.204	30.144	32.852	36.191	38.582	43.819
20	7.434	8.260	9.591	10.851	12.443	15.452	19.337	23.828	28.412	31.410	34.170	37.566	39.997	45.314
21	8.034	8.897	10.283	11.591	13.240	16.344	20.337	24.935	29.615	32.671	35.479	38.932	41.401	46.796
22	8.643	9.542	10.982	12.338	14.041	17.240	21.337	26.039	30.813	33.924	36.781	40.289	42.796	48.268
23	9.260	10.196	11.689	13.091	14.848	18.137	22.337	27.141	32.007	35.172	38.076	41.638	44.181	49.728
24	9.886	10.856	12.401	13.848	15.659	19.037	23.337	28.241	33.196	36.415	39.364	42.980	45.558	51.179
25	10.520	11.524	13.120	14.611	16.473	19.939	24.337	29.339	34.382	37.652	40.646	44.314	46.928	52.619
26	11.160	12.198	13.844	15.379	17.292	20.843	25.336	30.435	35.563	38.885	41.923	45.642	48.290	54.051
27	11.808	12.878	14.573	16.151	18.114	21.749	26.336	31.528	36.741	40.113	43.195	46.963	49.645	55.475
28	12.461	13.565	15.308	16.928	18.939	22.657	27.336	32.620	37.916	41.337	44.461	48.278	50.994	56.892
29	13.121	14.256	16.047	17.708	19.768	23.567	28.336	33.711	39.087	42.557	45.722	49.588	52.335	58.301
30	13.787	14.953	16.791	18.493	20.599	24.478	29.336	34.800	40.256	43.773	46.979	50.892	53.672	59.702
40	20.707	22.164	24.433	26.509	29.051	33.660	39.335	45.616	51.805	55.758	59.342	63.691	66.766	73.403
50	27.991	29.707	32.357	34.764	37.689	42.942	49.335	56.334	63.167	67.505	71.420	76.154	79.490	86.660
60	35.534	37.485	40.482	43.188	46.459	52.294	59.335	66.981	74.397	79.082	83.298	88.379	91.952	99.608
70	43.275	45.442	48.758	51.739	55.329	61.698	69.334	77.577	85.527	90.531	95.023	100.425	104.215	112.317
80	51.172	53.540	57.153	60.391	64.278	71.145	79.334	88.130	96.578	101.879	106.629	112.329	116.321	124.839
90	59.196	61.754	65.647	69.126	73.291	80.625	89.334	98.650	107.565	113.145	118.136	124.116	128.299	137.208
100	67.328	70.065	74.222	77.929	82.358	90.133	99.334	109.141	118.498	124.342	129.561	135.807	140.170	149.449

N.B.: il percentile *p* rappresenta la probabilità di non superamento e quindi è uguale a (1 - il livello di significatività alfa)



FORMULARIO TEST DI IPOTESI

TEST PER LA MEDIA DI UNA POPOLAZIONE NORMALE CON VARIANZA NOTA

TIPO DI TEST	SISTEMA DI IPOTESI	VALORE OSSERVATO	VALORE TEORICO	REGIONE DI RIFIUTO STANDARDIZZATA
Unidirezionale verso destra	$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \geq \mu_0 \end{cases}$	$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$	Z_α	$Z \geq Z_\alpha$
Unidirezionale verso sinistra	$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \leq \mu_0 \end{cases}$	$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$	Z_α	$Z \leq -Z_\alpha$
Bidirezionale	$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases}$	$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$	$Z_{\alpha/2}$	$\lvert Z \rvert \geq Z_{\alpha/2}$

TEST PER LA MEDIA DI UNA POPOLAZIONE NORMALE CON VARIANZA IGNOTA

TIPO DI TEST	SISTEMA DI IPOTESI	VALORE OSSERVATO	VALORE TEORICO	REGIONE DI RIFIUTO STANDARDIZZATA
Unidirezionale verso destra	$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \geq \mu_0 \end{cases}$	$T = \frac{\bar{x} - \mu_0}{S / \sqrt{n}}$	$T_{\alpha,n-1}$	$T \geq T_{\alpha,n-1}$
Unidirezionale verso sinistra	$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \leq \mu_0 \end{cases}$	$T = \frac{\bar{x} - \mu_0}{S / \sqrt{n}}$	$T_{\alpha,n-1}$	$T \leq -T_{\alpha,n-1}$
Bidirezionale	$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases}$	$T = \frac{\bar{x} - \mu_0}{S / \sqrt{n}}$	$T_{\alpha/2,n-1}$	$\lvert T \rvert \geq T_{\frac{\alpha}{2},n-1}$

TEST PER LA PROPORZIONE

TIPO DI TEST	SISTEMA DI IPOTESI	VALORE OSSERVATO	VALORE TEORICO	REGIONE DI RIFIUTO STANDARDIZZATA
Unidirezionale verso destra	$\begin{cases} H_0: \pi = \pi_0 \\ H_1: \pi \geq \pi_0 \end{cases}$	$Z = \frac{P_n - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$	Z_α	$Z \geq Z_\alpha$
Unidirezionale verso sinistra	$\begin{cases} H_0: \pi = \pi_0 \\ H_1: \pi \leq \pi_0 \end{cases}$	$Z = \frac{P_n - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$	Z_α	$Z \leq -Z_\alpha$
Bidirezionale	$\begin{cases} H_0: \pi = \pi_0 \\ H_1: \pi \neq \pi_0 \end{cases}$	$Z = \frac{P_n - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$	$Z_{\alpha/2}$	$\lvert Z \rvert \geq Z_{\alpha/2}$

TEST PER LA DIFFERENZA TRA DUE MEDIE DI DUE N CON VARIANZA NOTA

TIPO DI TEST	SISTEMA DI IPOTESI	VALORE OSSERVATO	VALORE TEORICO	REGIONE DI RIFIUTO STANDARDIZZATA
Unidirezionale verso destra	$\begin{cases} H_0: \mu_A - \mu_B = 0 \\ H_0: \mu_A - \mu_B \geq 0 \end{cases}$	$Z = \frac{(\overline{X_A} - \overline{X_B})}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$	Z_α	$Z \geq Z_\alpha$

Unidirezionale verso sinistra	$\begin{cases} H_0: \mu_A - \mu_B = 0 \\ H_0: \mu_A - \mu_B \leq 0 \end{cases}$	$Z = \frac{(\overline{X_A} - \overline{X_B})}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$	Z_α	$Z \leq -Z_\alpha$
Bidirezionale	$\begin{cases} H_0: \mu_A - \mu_B = 0 \\ H_0: \mu_A - \mu_B \neq 0 \end{cases}$	$Z = \frac{(\overline{X_A} - \overline{X_B})}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$	$Z_{\alpha/2}$	$\lvert Z \rvert \geq Z_{\alpha/2}$

TEST PER LA DIFFERENZA TRA DUE MEDIE DI DUE N CON VARIANZE IGNOTE MA UGUALI

TIPO DI TEST	SISTEMA DI IPOTESI	VALORE OSSERVATO	VALORE TEORICO	REGIONE DI RIFIUTO STANDARDIZZATA
Unidirezionale verso destra	$\begin{cases} H_0: \mu_A - \mu_B = 0 \\ H_0: \mu_A - \mu_B \geq 0 \end{cases}$	$T = \frac{(\overline{X_A} - \overline{X_B})}{S_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$	$T_{\alpha,(n_A+n_B-2)}$	$T \geq T_{\alpha,(n_A+n_B-2)}$
Unidirezionale verso sinistra	$\begin{cases} H_0: \mu_A - \mu_B = 0 \\ H_0: \mu_A - \mu_B \leq 0 \end{cases}$	$T = \frac{(\overline{X_A} - \overline{X_B})}{S_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$	$T_{\alpha,(n_A+n_B-2)}$	$T \leq -T_{\alpha,(n_A+n_B-2)}$
Bidirezionale	$\begin{cases} H_0: \mu_A - \mu_B = 0 \\ H_0: \mu_A - \mu_B \neq 0 \end{cases}$	$T = \frac{(\overline{X_A} - \overline{X_B})}{S_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$	$T_{\alpha/2,(n_A+n_B-2)}$	$\lvert T \rvert \geq T_{\alpha/2,(n_A+n_B-2)}$

TEST PER LA DIFFERENZA TRA DUE MEDIE DI DUE N CON VARIANZE IGNOTE

TIPO DI TEST	SISTEMA DI IPOTESI	VALORE OSSERVATO	VALORE TEORICO	REGIONE DI RIFIUTO STANDARDIZZATA
Unidirezionale verso destra	$\begin{cases} H_0: \mu_A - \mu_B = 0 \\ H_0: \mu_A - \mu_B \geq 0 \end{cases}$	$Z = \frac{(\overline{X_A} - \overline{X_B})}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}$	Z_α	$Z \geq Z_\alpha$
Unidirezionale verso sinistra	$\begin{cases} H_0: \mu_A - \mu_B = 0 \\ H_0: \mu_A - \mu_B \leq 0 \end{cases}$	$Z = \frac{(\overline{X_A} - \overline{X_B})}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}$	Z_α	$Z \leq -Z_\alpha$
Bidirezionale	$\begin{cases} H_0: \mu_A - \mu_B = 0 \\ H_0: \mu_A - \mu_B \neq 0 \end{cases}$	$Z = \frac{(\overline{X_A} - \overline{X_B})}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}$	$Z_{\alpha/2}$	$\lvert Z \rvert \geq Z_{\alpha/2}$

TEST PER LA MEDIA DI UNA POPOLAZIONE IGNOTA

TIPO DI TEST	SISTEMA DI IPOTESI	VALORE OSSERVATO	NUMEROSITA' OTTIMALE
Unidirezionale verso destra	$\begin{cases} H_0: \mu = \mu_0 \\ H_0: \mu \geq \mu_0 \end{cases}$	$Z = \frac{\bar{x} - \mu_0}{S / \sqrt{n}}$	$n = [\frac{\sigma(Z_\alpha + Z_\beta)}{\mu_1 - \mu_2}]^2$
Unidirezionale verso sinistra	$\begin{cases} H_0: \mu = \mu_0 \\ H_0: \mu \leq \mu_0 \end{cases}$	$Z = \frac{\bar{x} - \mu_0}{S / \sqrt{n}}$	$n = [\frac{\sigma(Z_\alpha + Z_\beta)}{\mu_1 - \mu_2}]^2$

Bidirezionale	$\begin{cases} H_0: \mu = \mu_0 \\ H_0: \mu \neq \mu_0 \end{cases}$	$Z = \frac{\bar{x} - \mu_0}{S / \sqrt{n}}$	$n = [\frac{\sigma(Z_{\alpha/2} + Z_{\beta})}{\mu_1 - \mu_2}]^2$

TEST DI INDIPENDENZA

SISTEMA DI IPOTESI	STATISTICA TEST	REGIONE DI RIFIUTO STANDARDIZZATA
$\begin{cases} \boldsymbol{H_0: X \text{ e } Y \text{ sono indipendenti}} \\ \boldsymbol{H_0: X \text{ e } Y \text{ non sono indipendenti}} \end{cases}$	$x^2 = \sum_i^k \sum_j^H \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$ *Ricorda:* $\hat{n}_{ij} = \frac{n_i \cdot n_j}{N}$	$X^2 \geq X^2_{\alpha,(K-1)(H-1)}$

VARIANZA CAMPIONARIA CORRETTA

$$\hat{s}^2 = \frac{\sum(x_1 - \bar{x})^2}{n - 1}$$

P-VALUE

Ipotesi alternativa	Regione critica	P-value
$\boldsymbol{\mu < \mu_0}$	$z_{oss} \leq -z_{\alpha}$	P(Z≤ z_{oss})
$\boldsymbol{\mu > \mu_0}$	$z_{oss} \geq z_{\alpha}$	P(Z≥ z_{oss})

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

P-value:

- Cerco la Z osservata calcolata nell' ipotesi nella tavola, incrociando riga e colonna;
- 0,5-Z (TROVATA)=P-value;
- Imposto la regola in base a che direzione abbiamo.
- Se avessi avuto la T, avrei calcolato la Z con la S.

Spiegazione esercizi Step by Step sui test statistici

Test statistici per la media

- Scriviamo i dati che abbiamo, per questo esercizio ci serviranno:
  - *Media;*
  - *Media campionaria;*
  - *Varianza/Varianza campionaria;*
  - *Scarto quadratico medio/Scarto quadratico medio campionario;*
  - *Livello di significatività;*
  - *Numero unità del campione.*
- Scriviamo il sistema di ipotesi;
- Calcoliamo il test statistico più adatto a caso;
- Cerchiamo il valore teorico più adatto al caso;
- Stabiliamo la regola di rifiuto standardizzata/p-value più adatta al caso;
- Scriviamo il commento finale.

Test statistici per la proporzione

- Scriviamo i dati che abbiamo, per questo esercizio ci serviranno:
  - *Probabilità di successo;*
  - *Probabilità di insuccesso;*
  - *Livello di significatività;*
  - *Numero unità del campione.*
- Scriviamo il sistema di ipotesi;
- Calcoliamo il test statistico più adatto a caso;
- Cerchiamo il valore teorico più adatto al caso;
- Stabiliamo la regola di rifiuto standardizzata/p-value più adatta al caso;
- Scriviamo il commento finale.

Analisi di dipendenza

- Scriviamo i dati che abbiamo, per questo esercizio ci serviranno:
  - *Tabella di contingenza;*
  - *Livello di significatività;*
  - *Numero unità del campione.*
- Scriviamo il sistema di ipotesi;
- Cerchiamo il chi quadrato osservato;
- Cerchiamo il valore teorico ;
- Stabiliamo la regola di rifiuto standardizzata;
- Scriviamo il commento finale.

Errore di I e di II tipo

- Scriviamo i dati che abbiamo, per questo esercizio ci serviranno:
  - *Media dello status quo;*
  - *Media dell'ipotesi alternativa;*
  - *Media che ci definisce la regola di rifiuto;*
  - *Livello di significatività;*
  - *Valore di Z;*
  - *Numero unità del campione.*
- Scriviamo il sistema di ipotesi;
- Calcoliamo la Z di alfa;
- Cerchiamo l'area di Z di alfa;
- Sottraiamo 0,50 dall'area di Z;
- Calcoliamo la Z di beta;
- Cerchiamo l'area di Z di beta;
- Sottraiamo 0,50 dall'area di Z;
- Disegniamo il grafico.

Potenza del test per la media

- Scriviamo i dati che abbiamo, per questo esercizio ci serviranno:
  - *Media dell'ipotesi alternativa;*
  - *Media campionaria;*
  - *Media dello status quo;*
  - *Scarto quadratico medio/Scarto quadratico medio campionario;*
  - *Livello di significatività;*
  - *Valore di Z/Valore di T;;*
  - *Numero unità del campione.*
- Impostiamo il rifiuto più adatto al caso;
- Calcoliamo la Z;
- Cerchiamo la Z in tavola, trovando beta;
- Sottraiamo 0,50 da quel valore;
- Calcoliamo la potenza del test.

Potenza del test per la proporzione

- Scriviamo i dati che abbiamo, per questo esercizio ci serviranno:
  - *Proporzione dell'ipotesi alternativa;*
  - *Proporzione campionaria;*
  - *Proporzione dello Status quo;*
  - *Probabilità di insuccesso;*
  - *Livello di significatività;*
  - *Valore di Z;*

    - *Numero unità del campione.*
- Impostiamo il rifiuto più adatto al caso;
- Calcoliamo la Z;
- Cerchiamo la Z in tavola, trovando beta;
- Sottraiamo 0,50 da quel valore;
- Calcoliamo la potenza del test.

P-Value

- Cerco la Z osservata calcolata nell' ipotesi nella tavola, incrociando riga e colonna;
- 0,5-Z (TROVATA)=P-value;
- Imposto la regola in base a che direzione abbiamo.
- Se avessi avuto la T, avrei calcolato la Z con la S.

Esercitazione

1. **Esercizio:**

Un'azienda inizia a produrre un nuovo tipo di lettore mp3 da viaggio con la durata media di una ricarica delle di 140 ore. È noto che la durata si distribuisce come una normale con varianza pari a 400. Cosa si può concludere a un livello di significatività al 5%, se un campione di 49 lettori selezionati casualmente nel momento in cui:

a) un cliente dice che la durata media sia stata di 150 ore? Indicare anche il P-value;
b) un cliente dice che la durata media sia stata di 130 ore? Indica anche il P-value;
c) un cliente dice che la durata sia stata delle volte di 135 ore e quindi diverso da 140?.

2. **Esercizio:**

Un'azienda inizia a produrre un nuovo tipo di lettore mp3 da viaggio con la durata media di una ricarica delle di 140 ore. È noto che la durata si distribuisce come una normale con varianza ignota. Cosa si può concludere a un livello di significatività al 5%, e con una varianza campionaria corretta pari a 360 se un campione di 49 lettori selezionati casualmente nel momento in cui:

a. un cliente dice che la durata media sia stata di 147 ore?
b. un cliente dice che la durata media sia stata di 132 ore?
c. un cliente dice che la durata sia stata delle volte di 142 ore e quindi diversa di 140?

3. **Esercizio:**

Una campagna promozionale di una nota compagnia aerea rivolta alla clientela business afferma che i 2/3 dei viaggiatori che viaggiano per lavoro utilizzano il computer portatile durante il volo. Effettuare un test con un livello di significatività pari al 5%.

a. Per verificare tale affermazione si estrae un campione casuale di 546 clienti business dei quali 355 affermano di utilizzare il computer durante il viaggio. Indicare anche il P-value;
b. Si stima però che lo 0,50 usi il computer durante il volo. Indicare anche il P-value;
c. Si stima però che lo 0,80 usi il computer durante il volo. Indicare anche il P-value.

4. Esercizio:

Gli inserzionisti televisivi scelgono gli spettatori che daranno loro il pubblico demograficamente desiderato come età, reddito e livello di istruzione più adatti al loro prodotto. In un campione casuale di 200 spettatori che guardano regolarmente il programma A, l'età media degli spettatori è di 31,2 anni. In un campione casuale di 300 spettatori che guardano regolarmente il programma B, l'età media è di 29,8 anni. Si assuma che la deviazione standard dell'età della popolazione che segue il programma A sia di 5,5 anni e quella della popolazione che segue il programma B è di 5,1 anni. Effettuare un test per verificare l'ipotesi che non ci sia differenza tra l'età media delle due popolazioni di spettatori, a livello di significatività del 5%. Indicare anche il P-value.

5. Esercizio:

Gli inserzionisti televisivi scelgono gli spettatori che daranno loro il pubblico demograficamente desiderato come età, reddito e livello di istruzione più adatti al loro prodotto. In un campione casuale di 200 spettatori che guardano regolarmente il programma A, l'età media degli spettatori è di 31,2 anni. In un campione casuale di 300 spettatori che guardano regolarmente il programma B, l'età media è di 29,8 anni. Si assuma che la deviazione standard è ignota ma diversa e varianze campionarie corrette pari 1 a 5,3 anni e la seconda a 5,7 anni. Effettuare un test per verificare l'ipotesi che non ci sia differenza tra l'età media delle due popolazioni di spettatori, a livello di significatività del 5%.

6. Esercizio:

Si vuole verificare l'indipendenza tra la fascia di prezzo di una camera B&B e la città in cui è venduta. A tal fine si estrae un campione di 150 esercizi B&B e si riportano i risultati nella tabella seguente. Verificare l'ipotesi di indipendenza con un livello di significatività del 5%.

	Fascia bassa	Fascia media
Napoli	20	30
Milano	40	60
Roma	15	35

7. Esercizio:

Il produttore di una nota compagnia telefonica sostiene che la durata del cellulare prodotto sia mediamente di 16 ore e scarto quadratico medio pari a 2 ore. L'acquirente, sulla base dell'utilizzo del cellulare, sospetta che pur distribuendosi normalmente con scarto quadratico medio pari a 2 ore, la durata media del cellulare sia di 13 ore. Si mettono allora in produzione 3 cellulari che verranno acquistati solo se la loro durata sia maggiore di 15 ore.

a) Si calcolino l'errore di primo tipo e del secondo tipo;
b) Si calcoli la potenza del test.

8. Esercizio:

Il supervisore delle operazioni presso un centro di assistenza clienti online è preoccupato per il tempo medio che intercorre tra la ricezione di una domanda o un reclamo di un cliente e l'invio di una risposta. La società va orgogliosa in un tempo di risposta medio non superiore alle 36 ore. Viene selezionato un campione casuale di 10 richieste recenti da parte di clienti, rilevando un tempo medio campionario di risposta di 41 ore e una deviazione standard campionaria di 8 ore. Determinare se questi dati portano a rifiutare l'ipotesi nulla, in quanto il tempo di risposta sarà diverso da 36 ore a un livello di significatività del 5%.

9. Esercizio:

Se gli addetti alla produzione svolgono correttamente il loro lavoro, le scatole di detersivo in polvere prodotte da una nota azienda che produce detersivi dovrebbero contenere, in media, 850 grammi di prodotto. L'azienda selezione periodicamente un campione di 36 scatole di detersivo e ne controlla il contenuto. Effettuare un test di ipotesi appropriato usando un livello di significatività del 5%. Assumere che il processo di riempimento abbia una deviazione standard nota di 8,5 grammi indipendentemente dalla media.

a) Supponiamo di osservare un campione e rilevare una media campionaria di 843 grammi per scatola. Quale sarebbe la conclusione?
b) Supponiamo di selezionare un campione che abbia una media di 853 grammi per scatola. Quale sarebbe la conclusione? Calcolare il P-value;
c) Si mettono in produzione le scatole e si nota un peso medio di 860. Sulla base del punto b calcolare l'errore di primo e di secondo tipo;
d) Calcolare la potenza del test sulla base del punto b.

10. Esercizio:

50 malati della stessa patologia sono divisi in due gruppi di 25 malati ciascuno. Il primo gruppo è trattato con un farmaco vecchio e l'altro con un nuovo. Si controlli l'esistenza di una dipendenza tra i caratteri con un livello di significatività dello 0,05.

	Guariti	Non guariti
Farmaco Vecchio	11	14
Farmaco Nuovo	19	6

11. Esercizio:

L'anno scorso, il 15% degli studenti senior della Lousiana Tech ha riferito di aver già avviato un'attività con l'intento di proseguirla dopo la laurea. Da un CCS di 150 studenti senior attuali dell'istituto, risulta che 18 studenti del campione hanno dichiarato di aver avviato un'attività. Alla luce dei dati, possiamo affermare che l'attuale percentuale di studenti hanno iniziato un'attività non è la stessa dell'anno scorso? Eseguire un test di ipotesi con un livello dell'1% e calcolare il P-value.

ARGOMENTI, SOTTOFORMA DI DOMANDA, DEL CAPITOLO 5

- *Che cosa è la verifica delle ipotesi statistiche?*
- *Che cosa è un test statistico?*
- *Cosa si intende per accettazione e rigetto?*
- *Che cosa è l'approccio dei test di significatività di Fisher?*
- *Che cosa è la teoria dei Test statistici di Neyman e Pearson?*
- *Quali sono le fasi di un test statistico?*
- *Cosa si intende per ipotesi nulla e alternativa?*
- *Cosa si intende per ipotesi semplice e composta?*
- *Cosa si intende per ipotesi composta unidirezionale e bidirezionale?*
- *Cosa è lo spazio parametrico?*
- *Cosa è la statistica test?*
- *Cosa è si intende per errore di primo e secondo tipo?*
- *Cosa è la potenza del test?*
- *Cosa si intende per test uniformemente più potenti?*
- *Qual è il rapporto con gli intervalli di confidenza?*
- *Cosa è P-value? In che senso parliamo di regola di decisione?*
- *Quali sono i vari test statistici che abbiamo?*
- *Cosa si intende per test non parametrici? Cosa è un test di indipendenza?*

Terza parte

Il modello di regressione lineare semplice

CAPITOLO 6: IL MODELLO DI REGRESSIONE LINEARE SEMPLICE

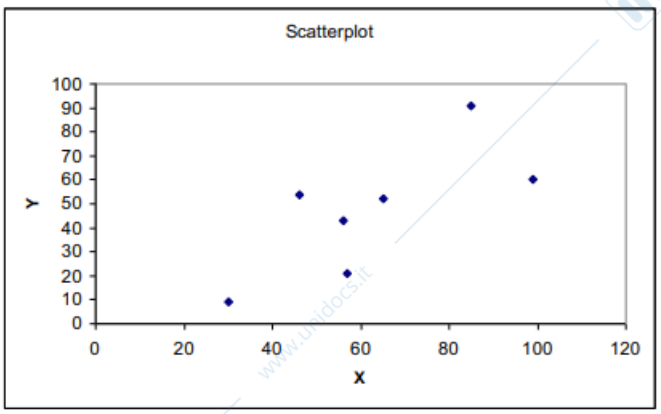
1. LA RELAZIONE STATISTICA

Per poter rappresentare una *relazione statistica* viene utilizzato il grafico di dispersione, o *diagramma a dispersione, o scatter plot*, il quale è un tipo di grafico in cui due variabili di un set di dati sono riportate su uno spazio cartesiano.

I dati sono visualizzati tramite una *collezione di punti* ciascuno con una posizione sull'asse orizzontale determinato da una variabile, la quale sarà sotto il controllo dello sperimentatore, ciò significa che avrà il *ruolo di variabile indipendente X* e, invece, sull'asse verticale determinato dall'altra variabile, ossia *la variabile dipendente Y*.

Il grafico a dispersione può essere utile per visualizzare il grado di correlazione, o meglio di dipendenza lineare, tra le due variabili e rappresentare la relazione casuale tra la variabile indipendente e quella dipendente.

Il grafico avrà una forma simile:



Una relazione statistica tra una variabile indipendente X e una variabile dipendente Y è descritta dall'equazione: $Y=f(X)+\in$. In questa equazione troviamo una componente deterministica, ossia, *f(X)* la quale è la funzione di regressione, essa definisce, infatti *il contributo della variabile indipendente, esplicitato da una funzione deterministica, ai valori assunti dalla variabile di risposta Y*. Troviamo, infine, una componente stocastica o aleatoria, ossia il valore di *epsilon* ci rappresenta l'errore, ossia, *il contributo di tutti gli altri fattori, non osservati, in grado di influenzare la Y e che sono quindi sintetizzati in una variabile casuale* (es. variabili omesse, errori di misurazione).

Quando è possibile osservare le variabili *X* e *Y* sull'intera popolazione allora si parla di interpolazione statistica, ossia una procedura che attraverso una funzione

matematica mira a *descrivere al meglio la nuvola dei punti osservata nel diagramma a dispersione*.

Diversamente, si parla di modello di regressione quando l'obiettivo è quello di *stimare la relazione tra due variabili statistiche attraverso un campione casuale e una procedura inferenziale*.

L'analisi di regressione consiste in un insieme di tecniche statistiche finalizzate a identificare una funzione matematica che pone in relazione due o più variabili. Tale funzione dovrebbe permettere di prevedere il valore di una o più variabili conoscendo il valore di altre variabili.

2. IL MODELLO DI REGRESSIONE LINEARE SEMPLICE

Lo scopo dell'analisi di regressione lineare è di formalizzare un modello che esprima *il legame lineare esistente in media tra la variabile Y ed una variabile X*. Il concetto può essere descritto mediante l'equazione $Y = \beta_1 + \beta_2 X + \epsilon$, dove troviamo la componente deterministica f(x) da forma lineare, ossia rappresentata da una retta, in cui β_1 rappresenta l'intercetta, invece β_2 è il coefficiente angolare.

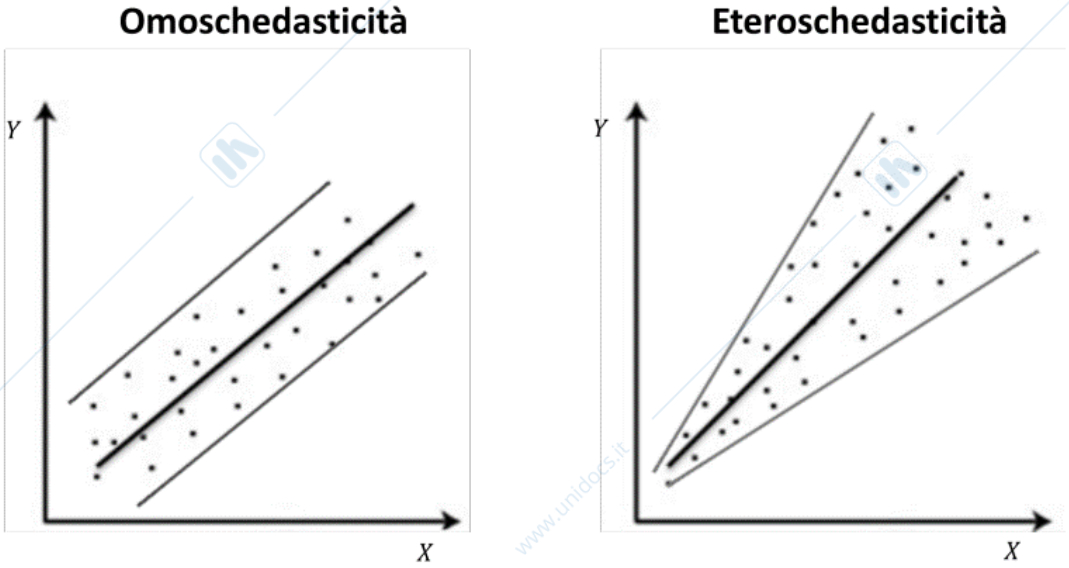
Ma da cosa è caratterizzato il modello? Il modello di regressione lineare è detto:

- Semplice, poiché considera la relazione tra due sole variabili;
- Classico, poiché le ipotesi su cui si basa per la stima dei parametri sono dette "*ipotesi classiche*";
- Di regressione, poiché con esso si intende *stimare o predire il valore medio della variabile dipendente sulla base di valori prefissati della variabile esplicativa,* per cui si dice che la variabile dipendente regredisce verso la media al variare dei valori della variabile esplicativa $E(Y|x_i) = \beta_1 + \beta_2 x_i$.
- Lineare, riferito ai parametri indipendentemente dalle variabili che possono essere opportunamente trasformate.

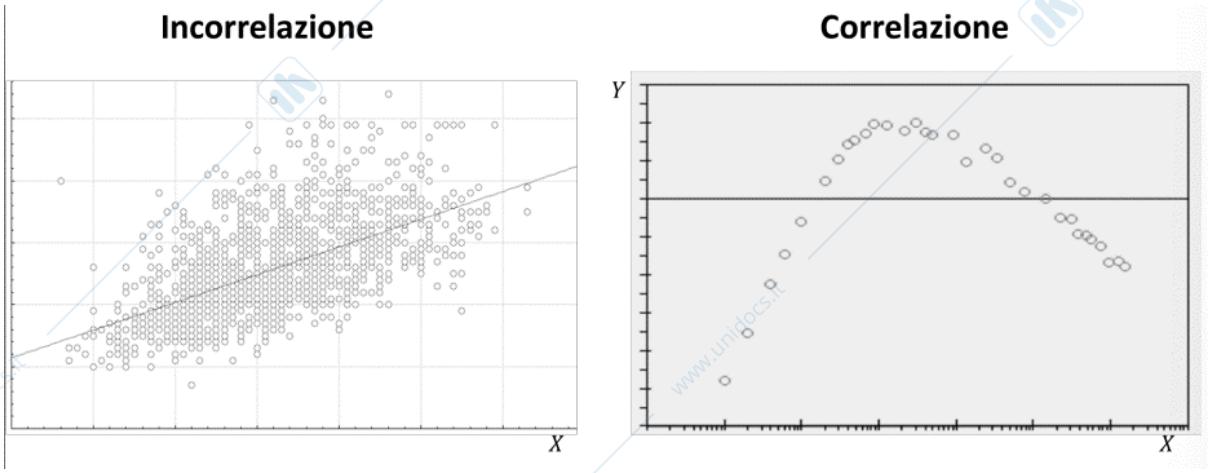
Possiamo poi definire le *ipotesi classiche del modello di regressione lineare:*

1. Non sistemaricità degli errori, ossia ciascun valore atteso (media) di ciascuna *variabile casuale errore* è uguale a zero;

2. Omoschedasticità degli errori, ossia la varianza dell'errore è costante e finita;



3. Incorrelazione degli errori, ossia la covarianza degli errori è uguale a zero, da cui deriva che gli errori sono incorrellati ma non necessariamenre indipendenti;



4. X non è una variabile casuale, ovvero non è correlata con l'errore;
5. Il modello è correttamente specificato, ovvero la relazione lineare tra le variabili è correttamente formalizzata nel modello;
6. La varianza di X non deve essere eccessivamente elevata, alyrimenti un'analisi lineare condurrebbe a soluzioni informative.

Con le prime tre ipotesi si può definire l'errore come una *variabile casuale white noise*.
Possiamo andare a definire il parametro e il suo stimatore con la seguente tabella:

Descrizione	Parametro	Stimatore	Stima puntuale
Intercetta	β_1	$\hat{\beta}_1$	$\hat{b}_1$
Coefficiente di regressione	β_2	$\hat{\beta}_2$	$\hat{b}_2$
Errori	ϵ	$\boldsymbol{e}$	e
Varianza degli errori	σ^2	$\hat{\boldsymbol{\sigma}}^{\boldsymbol{2}}$	$\hat{\sigma}^2$
Coefficiente di correlazione lineare	ρ	$\boldsymbol{r}$	r

NB. Le stime *e* sono dette residui

3. LA STIMA DEI PARAMETRI DEL MODELLO

Per poter realizzare una stima dei parametri del modello si usa il Metodo dei Minimi Quadrati, anche chiamato OLS, ossia *Ordinary Least Square*.

Il metodo dei minimi quadrati identifica come retta che meglio si adatta ai dati che minimizza la somma dei quadrati delle distanze verticali dei punti dalla retta, cioè la somma delle differenze tra i valori osservati della variabile dipendente e i valori previsti mediante l'equazione della retta.

La stima dei parametri del modello di regressione è ottenuta attraverso il metodo dei minimi quadrati, ossia *minimizzando la somma dei quadrati degli errori.*

Utilizziamo questa formula: $min\ Q\ (\beta_1, \beta_2) = \sum_i (y_i - \beta_1 - \beta_2 x_i)^2 = \sum_i \in_i^2$

Dalla soluzione di questo problema di minimo si ottengono gli stimatori dei due parametri del modello, ossia il coefficiente di regressione e lo stimatore dell'intercetta:

- *Stimatore del coefficiente di regressione*:

$$\hat{\beta}_2 = \frac{S_{X,Y}}{S_X^2} = \frac{\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum(x_i - \bar{x})^2} = \frac{\frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum x_i^2 - \bar{x}^2}$$

- *Stimatore dell'intercetta:* $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{\chi}$.

Ma quali sono le proprietà del metodo dei minimi quadrati?

1. La retta stimata passa per il punto medio di coordinate $(\bar{x}, \bar{y})$;
2. Il valore atteso delle stime $\hat{Y}$ è pari al valore atteso di Y siccome, il valore medio dell'errore è pari a 0, allora $E(Y) = E(\hat{Y}) + E(\in)$, quindi il valore medio di Y è uguale al valore medio di $\hat{Y}$;
3. $\sum(x_i e_i) = 0$.

Queste proprietà, secondo il Teorema di Gauss-Markov, confermano che gli stimatori dei minimi quadrati sono lineari, non distorti e con varianza minima *(BLUE-Best Linear Unbiased Estimators)*

4. LA BONTÀ DI ADATTAMENTO

La bontà di adattamento si valuta misurando quanta parte della variabilità della variabile dipendente è spiegata dalla parte deterministica del modello.

A tal fine si scompone la devianza di Y (devianza totale) nelle due fonti:

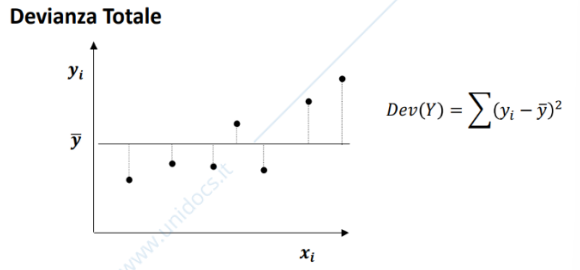
$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$$

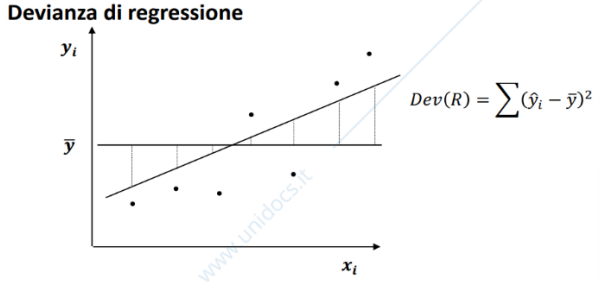
Dove la prima sommatoria è la devianza di regressione *Dev (R)*, invece la seconda sommatoria è la devianza dei residui *Dev (e)*.

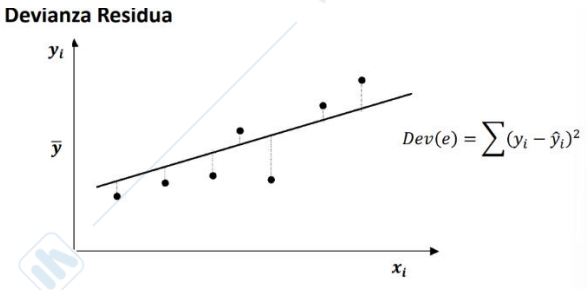
Dalla scomposizione della devianza si deriva l'indice R^2 che misura la bontà di adattamento del modello, dato dal rapporto tra la devianza di regressione e la devianza dei residui.

$$R^2 = \frac{Dev(R)}{Dev(Y)} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$ oppure $$R^2 = 1 - \frac{Dev(e)}{Dev(Y)} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

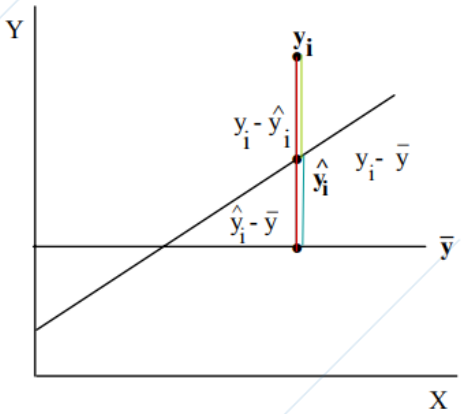
Esso è un indice che varia da zero ad uno, esprimendo un buon grado di adattamento lineare qualora il suo valore è prossimo ad uno. Nel caso della regressione lineare semplice vale la seguente relazione: $R^2 = \rho^2$.







Graficamente quindi...



Devianza Totale

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

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

Devianza Regressione | *Devianza Residua*

5. LA VERIFICA DEL MODELLO

Una volta stimato il modello attraverso un campione statistico si procede alla verifica dello stesso.

La verifica del modello consiste in una serie di procedure inferenziali mirate ad *una discussione del risultato ottenuto*. In particolare, tale discussione affronta i seguenti aspetti:

- *La significatività statistica dei parametri stimati;*
- *La coerenza interpretativa dei parametri stimati;*
- *La casualità delle v.c. errori;*
- *La capacità esplicativa del modello.*

La verifica avviene sulla base di quanto dice il Teorema di Gauss-Markov, che come abbiamo già visto, dice che sotto le ipotesi classiche del modello di regressione lineare semplice, gli stimatori dei minimi quadrati $(\hat{\beta}_1, \hat{\beta}_2)$ per i parametri (β_1, β_2) *sono lineari, non distorti e i più efficienti nella classe degli stimatori lineari e non distorti (BLUE, Best Linear Unbiased Estimators).*

Riprendiamo la tabella degli stimatori:

Descrizione	Parametro	Stimatore	Stima puntuale
Intercetta	β_1	$\hat{\beta}_1$	$\hat{b}_1$
Coefficiente di regressione	β_2	$\hat{\beta}_2$	$\hat{b}_2$
Errori	ϵ	$\boldsymbol{e}$	e
Varianza degli errori	σ^2	$\boldsymbol{\hat{\sigma}^2}$	$\hat{\sigma}^2$
Coefficiente di correlazione lineare	ρ	$\boldsymbol{r}$	r

5.1 Caratteristiche dello stimatore $\hat{\beta}_1$

Lo stimatore $\hat{\beta}_1$ è lo stimatore dell'intercetta di regressione.

Esso è dato da: $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$.

Secondo il Teorema di Gauss Markov è BLUE:

- È corretto $E(\hat{\beta}_1) = \beta_1$;
- È il più efficiente nella classe degli stimatori lineari, in quanto: $Var\ (\hat{\beta}_1) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{Dev\ (x)}\right)$;
- Se la varianza degli errori σ^2 non è nota allora si utilizzerà lo stimatore $\hat{\sigma}^2$, dove: $s^2(\beta_1) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{Dev\ (x)}\right)$.

5.2 Caratteristiche dello stimatore $\hat{\beta}_2$

Lo stimatore $\hat{\beta}_2$ è lo stimatore del coefficiente di regressione.

Esso è dato da: $\hat{\beta}_2 = \frac{s_{x,y}}{s_x^2}$.

Secondo il Teorema di Gauss Markov è BLUE:

- È corretto $E(\hat{\beta}_2) = \beta_2$;
- È il più efficiente nella classe degli stimatori lineari, in quanto: $Var\ (\hat{\beta}_2) = \frac{\sigma^2}{DEV\ (X)}$;
- Se la varianza degli errori σ^2 non è nota allora si utilizzerà lo stimatore $\hat{\sigma}^2$, dove: $s^2(\hat{\beta}_2) = \frac{\hat{\sigma}^2}{DEv(x)}$.

5.3 Caratteristiche dello stimatore $\hat{\sigma}^2$

Lo stimatore $\hat{\sigma}^2$ è lo stimatore della varianza degli errori.

Esso è dato da: $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$.

Esso è corretto quando: $E(\hat{\sigma}^2) = \sigma^2$

5.4 Verifica del modello stimato con test sul parametro β_1

Ipotesi nulla $H_0: \beta_1 = 0$

Statistica test con varianza nota: $Z = \frac{\hat{\beta}_1 - \beta_{1,H_0}}{\sqrt{Var(\hat{\beta}_1)}} \sim N(0,1)$

Statistica test con varianza non nota: $T = \frac{\hat{\beta}_1 - \beta_{1,H_0}}{\sqrt{S^2(\hat{\beta}_1)}} \sim t_{n-2}$

	Tipo di test	Ipotesi alternativa	Regione di rifiuto	
			In forma standardizzata	*In forma esplicita*
σ^2 noto	Unidirezionale verso destra	$H_1: \beta_1 > 0$	$Z \geq z_\alpha$	$\hat{\beta}_1 \geq z_\alpha \sqrt{Var(\hat{\beta}_1)}$
	Unidirezionale verso sinistra	$H_1: \beta_1 < 0$	$Z \leq -z_\alpha$	$\hat{\beta}_1 \leq -z_\alpha \sqrt{Var(\hat{\beta}_1)}$
	Bidirezionale	$H_1: \beta_1 \neq 0$	$\lvert Z \rvert \geq z_{\alpha/2}$	$\hat{\beta}_1 \geq z_{\alpha/2}\sqrt{Var(\hat{\beta}_1)}$ oppure $\hat{\beta}_1 \leq -z_{\alpha/2}\sqrt{Var(\hat{\beta}_1)}$
σ^2 non noto	Unidirezionale verso destra	$H_1: \beta_1 > 0$	$t \geq t_{\alpha,n-2}$	$\hat{\beta}_1 \geq t_{\alpha,n-2}\sqrt{S^2(\hat{\beta}_1)}$
	Unidirezionale verso sinistra	$H_1: \beta_1 < 0$	$t \leq -t_{\alpha,n-2}$	$\hat{\beta}_1 \leq -t_{\alpha,n-2}\sqrt{S^2(\hat{\beta}_1)}$
	Bidirezionale	$H_1: \beta_1 \neq 0$	$\lvert t \rvert \geq t_{\frac{\alpha}{2},n-2}$	$\hat{\beta}_1 \geq t_{\frac{\alpha}{2},n-2}\sqrt{S^2(\hat{\beta}_1)}$ oppure $\hat{\beta}_1 \leq -t_{\frac{\alpha}{2},n-2}\sqrt{S^2(\hat{\beta}_1)}$

- **Intervallo di confidenza per β_1** (ipotizzando σ^2 non noto)

$$IC\left[\hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\sqrt{S^2(\hat{\beta}_1)} \leq \beta_1 \leq \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\sqrt{S^2(\hat{\beta}_1)}\right] = 1 - \alpha$$

5.5 Verifica del modello stimato con test sul parametro β_2

Ipotesi nulla $H_0: \beta_2 = 0$

Statistica test con varianza nota: $Z = \frac{\hat{\beta}_2 - \beta_{2,H_0}}{\sqrt{Var(\hat{\beta}_2)}} \sim N(0,1)$

Statistica test con varianza non nota: $T = \frac{\hat{\beta}_2 - \beta_{2,H_0}}{\sqrt{S^2(\hat{\beta}_2)}} \sim t_{n-2}$

	Tipo di test	Ipotesi alternativa	Regione di rifiuto	
			In forma standardizzata	*In forma esplicita*
σ^2 nota	Unidirezionale verso destra	$H_1: \beta_2 > 0$	$Z \geq z_\alpha$	$\hat{\beta}_2 \geq z_\alpha \sqrt{Var(\hat{\beta}_2)}$
	Unidirezionale verso sinistra	$H_1: \beta_2 < 0$	$Z \leq -z_\alpha$	$\hat{\beta}_2 \leq -z_\alpha \sqrt{Var(\hat{\beta}_2)}$
	Bidirezionale	$H_1: \beta_2 \neq 0$	$\lvert Z \rvert \geq z_{\alpha/2}$	$\hat{\beta}_2 \geq z_{\alpha/2}\sqrt{Var(\hat{\beta}_2)}$ oppure $\hat{\beta}_2 \leq -z_{\alpha/2}\sqrt{Var(\hat{\beta}_2)}$
σ^2 non nota	Unidirezionale verso destra	$H_1: \beta_2 > 0$	$t \geq t_{\alpha,n-2}$	$\hat{\beta}_2 \geq t_{\alpha,n-2}\sqrt{S^2(\hat{\beta}_2)}$
	Unidirezionale verso sinistra	$H_1: \beta_2 < 0$	$t \leq -t_{\alpha,n-2}$	$\hat{\beta}_2 \leq -t_{\alpha,n-2}\sqrt{S^2(\hat{\beta}_2)}$
	Bidirezionale	$H_1: \beta_2 \neq 0$	$\lvert t \rvert \geq t_{\frac{\alpha}{2},n-2}$	$\hat{\beta}_2 \geq t_{\frac{\alpha}{2},n-2}\sqrt{S^2(\hat{\beta}_2)}$ oppure $\hat{\beta}_2 \leq -t_{\frac{\alpha}{2},n-2}\sqrt{S^2(\hat{\beta}_2)}$

- **Intervallo di confidenza per β_2** (ipotizzando σ^2 non noto)

$$IC\left[\hat{\beta}_2 + t_{\frac{\alpha}{2},n-2}\sqrt{S^2(\hat{\beta}_2)} \leq \beta_2 \leq \hat{\beta}_2 + t_{\frac{\alpha}{2},n-2}\sqrt{S^2(\hat{\beta}_2)}\right] = 1 - \alpha$$

6. L'IPOTESI DI NORMALITÀ DELLE V.C. ERRORI

La derivazione degli stimatori e delle loro proprietà è stata ottenuta esclusivamente sulla base delle ipotesi classiche, che non richiedono alcuna assunzione sulla distribuzione delle v.c. ϵ_i.

Per poter effettuare test di ipotesi e intervalli di confidenza sulle stime dei parametri è necessario derivare la distribuzione degli stimatori $(\hat{\beta}, \hat{\beta}_2)$.

A tal fine si aggiunge una ulteriore assunzione: la normalità delle v.c. ϵ_i.

$\epsilon_i \sim N(0, \sigma^2)$ *e per i=1,…n indipendenti*

Sotto tale ipotesi di normalità degli errori gli stimatori ottenuti con il metodo dei minimi quadrati coincidono con quelli che si otterrebbero con il metodo della massima verosimiglianza.

Ciò implica che:

- *Gli stimatori $(\hat{\beta}_1,\hat{\beta}_2)$sono anche sufficienti e efficienti in senso assoluto;*
- *Gli stimatori $(\hat{\beta}_1,\hat{\beta}_2)$ si distribuiscono normalmente qualunque sia la dimensione campionaria.*

Allora ne deriva che:

$$\hat{\beta}_1 \sim N\left(\beta_1, Var(\hat{\beta}_1)\right)$$

$$\hat{\beta}_2 \sim N\left(\beta_2, Var(\hat{\beta}_2)\right)$$

7. L'ANALISI DEI RESIDUI

Se le assunzioni classiche e la normalità degli errori possono ritenersi veritiere, gli stimatori dei minimi quadrati godono di proprietà desiderabili e possono essere utilizzati per inferire sulla popolazione.

Se una o più assunzioni risultano *violate* si dice che il modello è mal specificato. In questo caso le procedure inferenziali devono essere impiegate con molta prudenza.

La principale tecnica che consente di investigare le cause di un modello mal specificato è basata sull'analisi grafica dei residui.

Il grafico dei residui è un *grafico di dispersione* in cui l'asse delle ordinate è riferito ai residui e_i e l'asse delle ascisse ai valori stimati $\hat{y}_i$(o in alcuni casi ai valori χ_i).

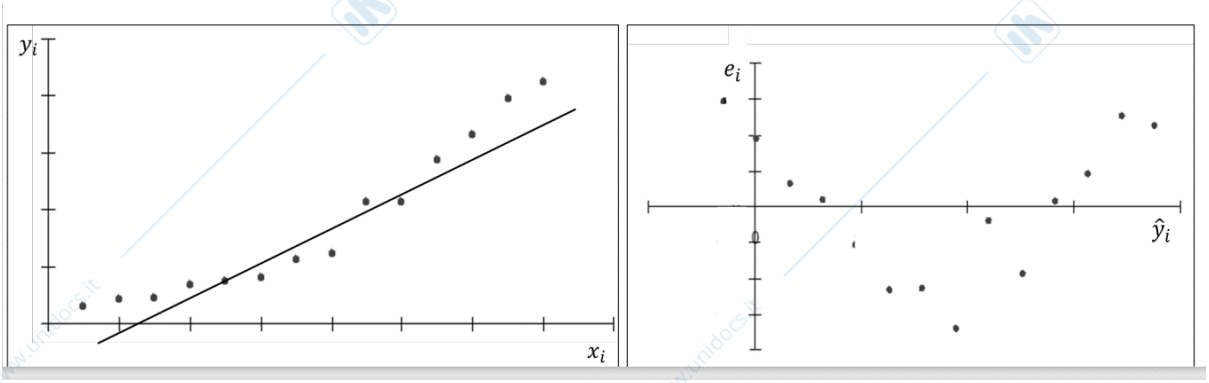
Se il modello è ben specificato, la distribuzione empirica dei residui nel piano cartesiano rifletterà le assunzioni specificate per gli errori $\in_i$:

- I residui tenderanno a distribuirsi casualmente attorno allo zero;
- Non presenteranno comportamenti sistematici o tendenze di fondo (white noise).

Possiamo avere diversi casi.

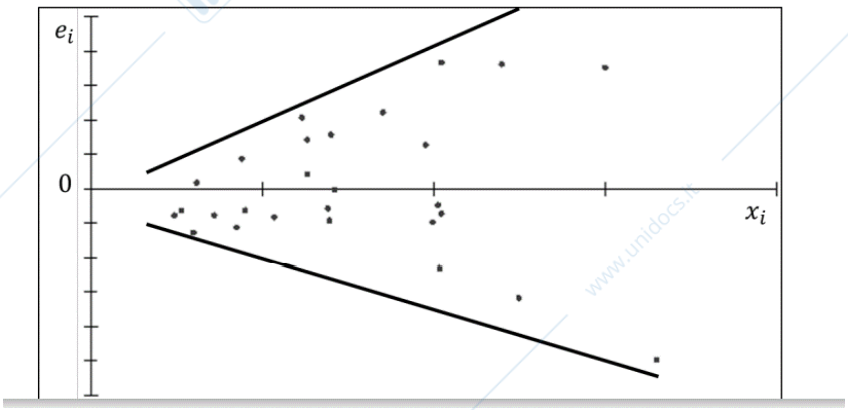
7.1 Linearità

Il diagramma di dispersione e il grafico dei residui indicano una violazione dell'assunzione di linearità. Infatti, è chiaramente visibile l'esistenza di una relazione non lineare (curvilinea) tra le due variabili *X* e Y.



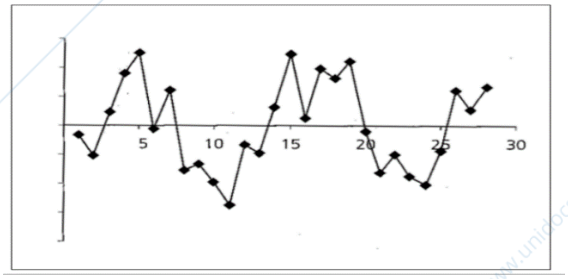
7.2 Omoschedasticità

In questo caso il grafico mostra una violazione dell'ipotesi di omoschedasticità degli errori. L'ampiezza della nuvola, e quindi la variabilità dei residui, tende ad aumentare al crescere dei valori della X.



7.3 Incorrelazione

Quando le osservazioni sono collezionate secondo una sequenza temporale può accadere che gli errori sia correlati.

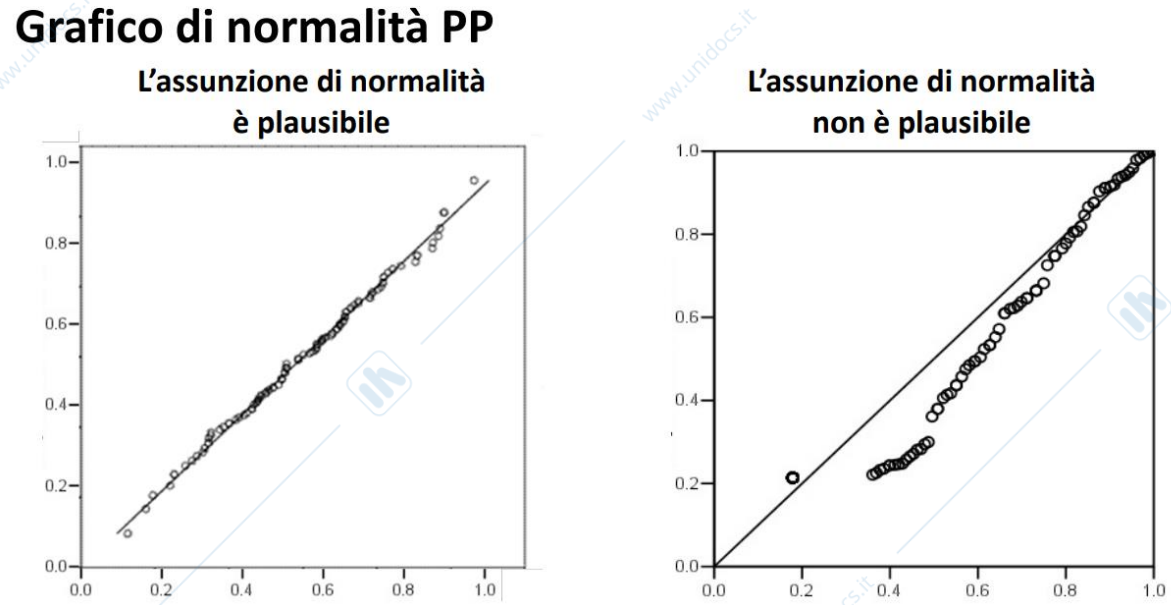


7.4 Normalità di PP

In questo grafico si confrontano le proporzioni cumulate di una variabile numerica (in questo caso i residui standardizzati) e le proporzioni cumulate attese in caso di distribuzione normale della stessa variabile.

Se l'ipotesi di normalità è veritiera allora ci si attenderà una corrispondenza tra le proporzioni osservate e quelle attese. In questo caso i punti si distribuiranno sulla bisettrice del piano cartesiano.

Maggiore sarà la divergenza dalla bisettrice e maggiore sarà l'allontanamento dall'ipotesi di normalità della variabile considerata.



FORMULARIO MODELLO DI REGRESSIONE

STIMATORE DELLA VARIANZA DEGLI ERRORI

STIMATORE	$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$
MEDIA	$E(\hat{\sigma}^2) = \sigma^2$

STIMATORE DEL COEFFICIENTE DI REGRESSIONE E VERIFICA DEL MODELLO STIMATO

STIMATORE	$\hat{\beta}_2 = \frac{S_{X,Y}}{S_X^2}$
MEDIA	$E(\hat{\beta}_2) = \beta_2$
VARIANZA	$Var(\hat{\beta}_2) = \frac{\sigma^2}{Dev(x)}$
VARIANZA DEGLI ERRORI NON NOTA	$S^2(\hat{\beta}_2) = \frac{\hat{\sigma}^2}{Dev(x)}$
IPOTESI NULLA	$H_0: \beta_2 = 0$
STATISTICA TEST VARIANZA NOTA	$z = \frac{\hat{\beta}_2 - \beta_{2,H_0}}{\sqrt{VAR(\hat{\beta}_2)}}$
STATISTICA TEST VARIANZA NON NOTA	$T = \frac{\hat{\beta}_2 - \beta_{2,H_0}}{\sqrt{S^2(\hat{\beta}_2)}}$

	Tipo di test	Ipotesi alternativa	Regione di rifiuto	
			In forma standardizzata	*In forma esplicita*
σ^2 nota	Unidirezionale verso destra	$H_1: \beta_2 > 0$	$Z \geq z_\alpha$	$\hat{\beta}_2 \geq z_\alpha \sqrt{Var(\hat{\beta}_2)}$
	Unidirezionale verso sinistra	$H_1: \beta_2 < 0$	$Z \leq -z_\alpha$	$\hat{\beta}_2 \leq -z_\alpha \sqrt{Var(\hat{\beta}_2)}$
	Bidirezionale	$H_1: \beta_2 \neq 0$	$\lvert Z \rvert \geq z_{\alpha/2}$	$\hat{\beta}_2 \geq z_{\alpha/2}\sqrt{Var(\hat{\beta}_2)}$ oppure $\hat{\beta}_2 \leq -z_{\alpha/2}\sqrt{Var(\hat{\beta}_2)}$
σ^2 non nota	Unidirezionale verso destra	$H_1: \beta_2 > 0$	$t \geq t_{\alpha,n-2}$	$\hat{\beta}_2 \geq t_{\alpha,n-2}\sqrt{S^2(\hat{\beta}_2)}$
	Unidirezionale verso sinistra	$H_1: \beta_2 < 0$	$t \leq -t_{\alpha,n-2}$	$\hat{\beta}_2 \leq -t_{\alpha,n-2}\sqrt{S^2(\hat{\beta}_2)}$
	Bidirezionale	$H_1: \beta_2 \neq 0$	$\lvert t \rvert \geq t_{\frac{\alpha}{2},n-2}$	$\hat{\beta}_2 \geq t_{\frac{\alpha}{2},n-2}\sqrt{S^2(\hat{\beta}_2)}$ oppure $\hat{\beta}_2 \leq -t_{\frac{\alpha}{2},n-2}\sqrt{S^2(\hat{\beta}_2)}$

- **Intervallo di confidenza per β_2** (ipotizzando σ^2 non noto)

$$IC\left[\hat{\beta}_2 + t_{\frac{\alpha}{2},n-2}\sqrt{S^2(\hat{\beta}_2)} \leq \beta_2 \leq \hat{\beta}_2 + t_{\frac{\alpha}{2},n-2}\sqrt{S^2(\hat{\beta}_2)}\right] = 1-\alpha$$

STIMATORE DELL'INTERCETTA DI REGRESSIONE E VERIFICA DEL MODELLO STIMATO

STIMATORE	$\hat{\beta}_1 = \bar{y} - \tilde{\beta}_2 \bar{x}$
MEDIA	$E(\hat{\beta}_1) = \beta_1$
VARIANZA	$Var(\hat{\beta}_1) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sigma}\right)$
VARIANZA DEGLI ERRORI NON NOTA	$S^2(\hat{\beta}_1) = Var(\hat{\beta}_1) = \bar{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sigma}\right)$
IPOTESI NULLA	$H_0: \beta_1 = 0$
STATISTICA TEST VARIANZA NOTA	$z = \frac{\hat{\beta}_1 - \beta_{1,H_0}}{\sqrt{VAR(\hat{\beta}_1)}}$
STATISTICA TEST VARIANZA NON NOTA	$T = \frac{\hat{\beta}_1 - \beta_{1,H_0}}{\sqrt{S^2(\hat{\beta}_1)}}$

	Tipo di test	Ipotesi alternativa	Regione di rifiuto	
			In forma standardizzata	*In forma esplicita*
σ^2 noto	Unidirezionale verso destra	$H_1: \beta_1 > 0$	$Z \geq z_\alpha$	$\hat{\beta}_1 \geq z_\alpha \sqrt{Var(\hat{\beta}_1)}$
	Unidirezionale verso sinistra	$H_1: \beta_1 < 0$	$Z \leq -z_\alpha$	$\hat{\beta}_1 \leq -z_\alpha \sqrt{Var(\hat{\beta}_1)}$
	Bidirezionale	$H_1: \beta_1 \neq 0$	$\lvert Z \rvert \geq z_{\alpha/2}$	$\hat{\beta}_1 \geq z_{\alpha/2} \sqrt{Var(\hat{\beta}_1)}$ oppure $\hat{\beta}_1 \leq -z_{\alpha/2} \sqrt{Var(\hat{\beta}_1)}$
σ^2 non noto	Unidirezionale verso destra	$H_1: \beta_1 > 0$	$t \geq t_{\alpha,n-2}$	$\hat{\beta}_1 \geq t_{\alpha,n-2} \sqrt{S^2(\hat{\beta}_1)}$
	Unidirezionale verso sinistra	$H_1: \beta_1 < 0$	$t \leq -t_{\alpha,n-2}$	$\hat{\beta}_1 \leq -t_{\alpha,n-2} \sqrt{S^2(\hat{\beta}_1)}$
	Bidirezionale	$H_1: \beta_1 \neq 0$	$\lvert t \rvert \geq t_{\frac{\alpha}{2},n-2}$	$\hat{\beta}_1 \geq t_{\frac{\alpha}{2},n-2} \sqrt{S^2(\hat{\beta}_1)}$ oppure $\hat{\beta}_1 \leq -t_{\frac{\alpha}{2},n-2} \sqrt{S^2(\hat{\beta}_1)}$

- **Intervallo di confidenza per β_1** (ipotizzando σ^2 non noto)

$$IC\left[\hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\sqrt{S^2(\hat{\beta}_1)} \leq \beta_1 \leq \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\sqrt{S^2(\hat{\beta}_1)}\right] = 1 - \alpha$$

Spiegazione esercizi Step by Step sul modello di regressione lineare semplice

Relazione tra variabili

- Scriviamo i dati che abbiamo, per questo esercizio ci serviranno:
  - *Variabile dipendente;*
  - *Variabile indipendente;*
  - *Numero delle unità nel campione;*
  - *Sommatoria dei prodotti tra le due variabili;*
  - *Sommatoria delle variabili indipendenti al quadrato;*
  - *Sommatoria delle variabili dipendenti al quadrato;*
  - *Media delle variabili indipendenti;*
  - *Media al quadrato delle variabili indipendenti;*
  - *Media delle variabili dipendenti;*
  - *Media al quadrato delle variabili dipendenti.*
- Scriviamo la funzione di relazione;
- Calcoliamo lo stimatore del coefficiente angolare;
- Analizziamo la covarianza;
- Calcoliamo lo stimatore dell'intercetta di regressione;
- Scriviamo il commento finale;
- Facciamo un esempio con la funzione.

Test statistici per i valori beta:

- Scriviamo i dati che abbiamo, per questo esercizio ci serviranno:
  - *Numero delle unità nel campione;*
  - *Sommatoria delle variabili dipendenti al quadrato;*
  - *Media al quadrato delle variabili indipendenti;*
  - *Media al quadrato delle variabili dipendenti;*
  - *Devianza di x;*
  - *Quadrato dello stimatore del coefficiente;*
  - *Stimatore del coefficiente;*
  - *Stimatore dell'intercetta;*
  - *Livello di significatività.*
- Scriviamo i sistemi di ipotesi di entrambi i parametri;
- Calcoliamo il test statistico più adatto a caso per il primo parametro;
- Cerchiamo il valore teorico più adatto al caso;
- Stabiliamo la regola di rifiuto standardizzata;
- Scriviamo il commento finale;
- *Ripetiamo per il secondo parametro.*

Intervalli di confidenza della regressione

- Scriviamo i dati che abbiamo, per questo esercizio ci serviranno:
  - *Stimatori;*
  - *Valori teorici;*
  - *Varianze degli errori;*
- Calcoliamo gli intervalli di confidenza;
- Scriviamo il commento finale.

Esercitazione

1. Esercizio:

Una grande società che possiede una catena di negozi situati in diverse città italiane vuole valutare l'impatto avuto da un investimento in pubblicità su quotidiani locali e attraverso volantini fatto nel mese di settembre. A tal fine selezione un campione casuale di 20 negozi, situati in città simili per dimensione demografica e per altre caratteristiche che possono influenzare il comportamento d'acquisto. Per ciascun negozio vengono rilevati: investimento in pubblicità (x) e volume delle vendite (y) fra il 20 settembre e il 30 novembre. Di seguito sono riportati i dati e consideriamo che lo scatter plot di questa analisi avrà una nuvola di regressione crescente, ossia, positiva.

$\sum_{i=1}^{n} x_i y_i = 1115{,}16$; $\sum_{i=1}^{n} x_i^2 = 267{,}29$; $\sum_{i=1}^{n} y_i^2 = 4814{,}08$; $\bar{x} = 3{,}41$; $\bar{Y} = 15{,}18$

a) Sulla base del campione osservato si vogliono stimare i parametri del modello che approssima il valore mediamente assunto dal volume delle vendite attraverso una funzione dei valori assunti dall'investimento in pubblicità;
b) Ipotizzando che gli errori abbiano una distribuzione normale, verificare se i coefficienti del modello sono significativamente diversi da zero, con un livello di significatività dell'1%;
c) Sulla base dei dati ottenuti calcolare l'intervallo di confidenza β_1 e di β_2.

2. Esercizio:

Un campione di 10 soggetti di età superiore ai 60 anni è stato sottoposto ad un test di abilità motorie. I punteggi al test (X) e le rispettive età (Y) dei soggetti sono riportati di seguito:

Età	60	65	72	80	67	75	77	79	81	77
Punteggio	40	25	16	18	35	14	10	15	12	18

a) Sulla base del campione osservato si vogliono stimare i parametri del modello che approssima il valore mediamente assunto dal punteggio attraverso una funzione dei valori assunti dall'età, facendo un esempio con punteggio pari a 19. Specificare, inoltre, attraverso il segno della covarianza, se la retta di regressione sarà positiva o negativa;

b) Valutare con un opportuno test se esiste un legame lineare tra il risultato del test e l'età, ossia se i nostri parametri siano diversi da 0, con un livello di significatività pari a 0,05;

c) Sulla base dei dati ottenuti calcolare l'intervallo di confidenza β1 e di β2.

3. Esercizio:

Nella seguente sono riportati per 10 autovetture di marche diverse, i dati relativi al numero di km percorribili in città con un litro di carburante (x) e alla cilindrata del motore in cc (y):

Km/lt	13,1	6,7	10,8	8,2	9,6	11	12,4	6,7	11,1	11,3
Cilindrata in cc	1590	2835	1590	2474	1950	1590	1393	2835	1590	1868

a) Si disegni e si commenti il diagramma di dispersione;

b) Sulla base del campione osservato si vogliono stimare i parametri del modello, ossia della retta di regressione, che lega i km percorribili per litro alla cilindrata delle autovetture e si interpretino i risultati ottenuti, inoltre, facendo un esempio con km/lt pari a 12 calcolare il valore della cilindrata;

c) Ipotizzando che gli errori abbiano una distribuzione normale, verificare se i coefficienti del modello sono significativamente diversi da zero, con un livello di significatività dell'5%;

d) Sulla base dei dati ottenuti calcolare gli intervalli di confidenza.

ARGOMENTI, SOTTOFORMA DI DOMANDA, DEL CAPITOLO 6

- *Che cosa è un diagramma a dispersione? Come può anche essere chiamato?*
- *Cosa abbiamo sull'asse x e sull'asse y?*
- *Cosa è la componente deterministica e cosa è la componente stocastica?*
- *Che cosa è l'interpolazione statistica?*
- *Cosa è il modello di regressione?*
- *Qual è il significato delle parole MODELLO CLASSICO DI REGRESSIONE LINEARE SEMPLICE?*
- *Quali sono le ipotesi classiche del modello di regressione lineare?*
- *Quali sono i parametri e gli stimatori?*
- *Cosa si intende per Metodo dei Minimi Quadrati?*
- *Quali sono le proprietà del metodo dei minimi quadrati?*
- *Cosa è il Teorema di Gauss-Markov?*
- *Cosa è la bontà di adattamento?*
- *Cosa si intende per verifica del modello?*
- *Quali sono le caratteristiche dello stimatore dell'intercetta di regressione?*
- *Quali sono le caratteristiche dello stimatore del coefficiente di regressione?*
- *Quali sono le caratteristiche dello stimatore della varianza degli errori?*
- *Cosa implica l'ipotesi di normalità delle variabili casuali errori?*
- *Cosa è l'analisi dei residui?*
- *Cosa si intende per linearità?*
- *Cosa si intende per omoschedasticità?*
- *Cosa si intende per incorrelazione?*
- *Cosa si intende per normalità?*