Le variabili casuali

L'esigenza di semplificare la trattazione degli esperimenti casuali porta a tradurne gli eventi in numeri reali

ESEMPIO ILLUSTRATIVO

Tre monete ben equilibrate sono lanciate insieme. L'universo degli eventi è

$$\left\{\begin{matrix} E_1 = TTT, E_2 = TTC, E_3 = TCT, E_4 = CTT \\ E_5 = TCC, E_6 = CTC, E_7 = CCT, E_8 = CCC \end{matrix}\right\}$$

Indichiamo con "X" il numero totale di teste.

Per l'evento $E_4 = C \cap T \cap T$ **si ha X=2.**

Inoltre E_4 **ha probabilità 1/8 dato che gli 8 eventi dell'universo degli eventi sono equiprobabili.**

Da notare che anche gli eventi E_2 **ed** E_3 **producono X=2 per cui questo è un evento composto con probabilità 3/8.**

Le variabili casuali/2

L'evento "X=2" si verifica ogni volta che si verifica l'evento $E_2 \cup E_3 \cup E_4$

I valore assunto dalla "X" non è noto a priori perché dipende dall'esito casuale dell'esperimento

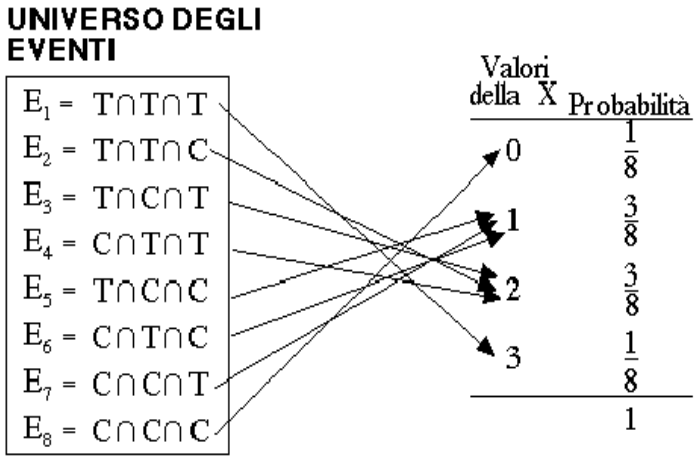
Per questo la "X" è detta variabile casuale *E' un prodotto dell'esperimento*

La tabella dei valori della X e delle probabilità con cui sono assunti si chiama:

DISTRIBUZIONE DI PROBABILITA' DELLA VARIABILE CASUALE

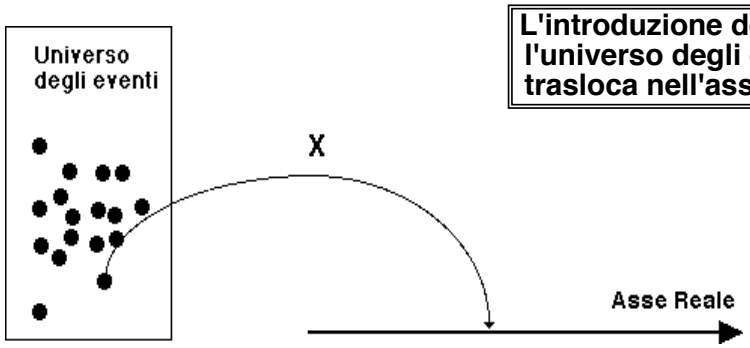
con essa si calcolano le probabilità degli eventi che interessano la "X". Ad esempio

$$P(X>1) = P(X=2) + P(X=3) = \frac{3}{8} + \frac{1}{8} = \frac{4}{8} = \frac{1}{2}$$



Le variabili casuali/3

La variabile casuale è una funzione che associa ad ogni evento dell'universo degli eventi uno ed un solo numero reale.



L'introduzione della "X" semplifica l'universo degli eventi perché lo trasloca nell'asse reale

La corrispondenza tra eventi e valori della X è univoca, ma non necessariament biunivoca: un dato valore della X può derivare da eventi diversi

Esempio

Si supponga che che le seguenti coppie di lettere siano equiprobabili

ab	aj	bg	eg	gj
ae	ak	bj	ej	gk
ag	be	bk	ek	jk

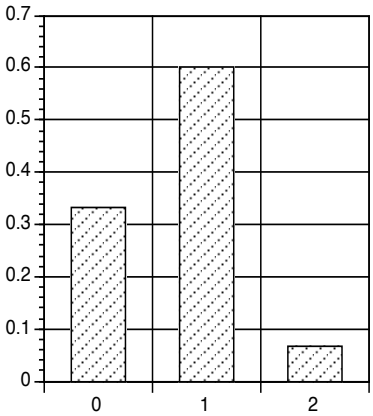
Sia "X" la variabile casuale: X=numero di vocali nella coppia

a) Costruire la distribuzione di probabilità

b) Rappresentarla graficamente

$X = x_i \quad è \quad un \quad evento$

X	P(X = x)
0	$6/15$
1	$8/15$
2	$1/15$
	1



Esercizio

Un esperimento saggia la praticabilità del green di una buca da golf con il tiro di 4 palline.

L'esito è incerto, ma riteniamo che la probabilità di mandare la pallina in buca sia del 70%.

Indichiamo con Y= "numero di palline in buca".

A) Quali sono gli eventi di interesse?

B) Quali probabilità sono associate ai valori della Y?

$$S = \left\{ \begin{array}{l} (0,0,0,0);\ (0,0,0,1);(0,0,1,0);\ (0,0,1,1);(0,1,0,0);\ (0,1,0,1);(0,1,1,0); \\ (0,1,1,1)(1,0,0,0);\ (1,0,0,1);(1,0,1,0);\ (1,0,1,1);(1,1,0,0);\ (1,1,0,1); \\ (1,1,1,0);\ (1,1,1,1); \end{array} \right\}$$

$$P\left[(x,x,x,x)\right] = 0.7^{a}0.3^{4-a};\ dove\ \ a = numero\ \ di\ \ "1"\ \ nella\ \ quaterna$$

$$Y \in \{0,1,2,3,4\};\ \ P(Y=y) = \{0.0081,\ \ 0.0756, 0.2656, 0.4116, 0.2406\}$$

Variabili casuali discrete e continue

La natura quantitativa dell'universo degli eventi porta a considerare

VARIABILI CASUALI DISCRETE

L'insieme dei valori possibili è formato da punti isolati che possono essere "contati" cioè posti in corrispondenza biunivoca con l'insieme dei naturali anche enumerabilmente infinito

Esempi

Scelta casuale di una famiglia da intervistare: X= Numero di componenti

Ispezione di un lotto di prodotti: X= numero di pezzi difettosi

Numero di molecole in un composto

VARIABILI CASUALI CONTINUE

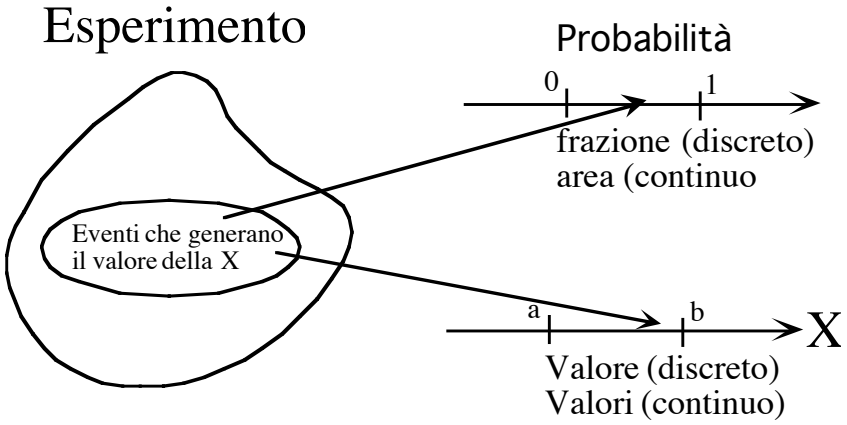
L'insieme dei valori possibili forma un intervallo di numeri reali

Esempio

Vita media di un componente di computer: X=tempo di funzionamento

Riflessione

Se l'obiettivo dell'esperimento è lo studio della variabile casuale perché non concentrarsi direttamente sulla sua funzione di distribuzione (X, p)?



I fenomeni soggetti alla sorte non nascono con il loro bravo modello, ma occorre decifrarlo, se possibile, in un opportuno esperimento.

Dopo che ciò è avvenuto e per tutti i casi riconosciuti analoghi si può adottare direttamente il modello di distribuzione.

Studio delle V.C. discrete

Una v.c. discreta è nota attraverso la sua distribuzione di probabilità formata dai valori possibili e dalle probabilità loro associate

Valori di X	x_1	x_2	...	x_i	...	totale
Probabilità	p_1	p_2	...	p_i	...	1

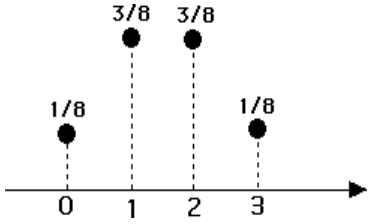
Il minuscolo indica i valori possibili cioè $P(X = x_i) = p_i$

Perchè le $\{p_i\}$ **costituiscano delle probabilità è necessario che:** $1)\ p_i \geq 0 \quad \forall i$ $2)\sum_i p_i = 1$

DISTRIBUZIONE DI PROBABILITA'

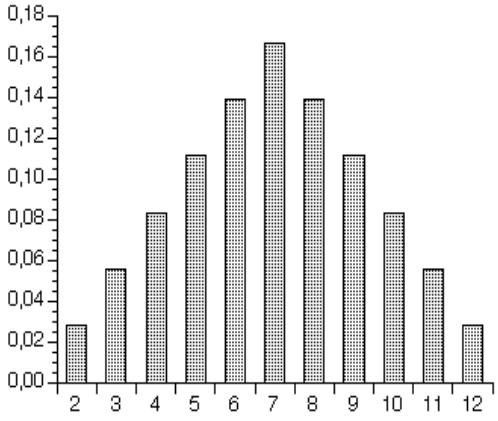
Esempio

Numero di teste in un lancio di tre monete equilibrate



Esempio

Lancio di due dadi. Studiare la variabile casuale X: SOMMA DEI DUE NUMERI CHE COMPAIONO NELLE DUE FACCE SUPERIORI



UNIVERSO DEGLI EVENTI

(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)
(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)

Poichè ogni coppia ha probabilità 1/36
la V.C. "X" avrà distribuzione di probabilità:

X	2	3	4	5	6	7	8	9	10	11	12	
P(X = x)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36	1

In questo caso le probabilità sono dettate dalle simmetrie dell'esperimento

Esercizio

In un'impresa che ha il 70% del personale di sesso maschile si scelgono a caso (con riposizione) due persone per una commissione. Il numero di donne "X" nella commissione è una variabile casuale

Possibili risultati:
(f,f) (f,m) (m,f) (m,m)

Corrispondenti valori della v.c.

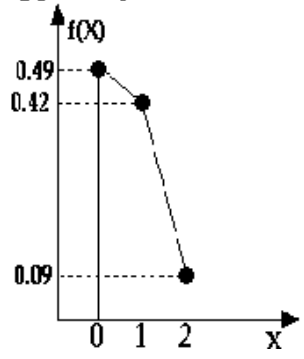
$$X = \begin{cases} 0 & \text{se } (m,m) \\ 1 & \text{se } (f,m) \text{ oppure } (m,f) \\ 2 & \text{se } (f,f) \end{cases}$$

Poiché il 1° componente è indipendente dal 2° si applica la legge del prodotto delle probabilità per cui

Evento	X	P(Evento)
(m,m)	0	0.7*0.7 = 0.49
(f,m)	1	0.3*0.7 = 0.21
(m,f)	1	0.7*0.3 = 0.21
(f,f)	2	0.3*0.3 = 0.09
		1.00

Distribuzione di probabilità

x	P(X = x)
0	0.49
1	0.42
2	0.09
	1.00



Sintesi delle variabili casuali

Si possono ricondurre le v.c. a pochi parametri descrittivi delle caratteristiche principali:

CENTRALITA' - VARIABILITA' - ASIMMETRIA

già discusse per le rilevazioni empiriche.

L'idea è che il verificarsi più o meno probabile di certi eventi risulta legato ad aspetti comprensibili e noti della variabile casuale.

Simbologia e definizioni non cambiano: si sostituiscono le frequenze con le probabilità

In particolare

Media aritmetica ponderata (Valore atteso)

$$E(X) = \sum_{i=1}^{n} x_i p_i$$

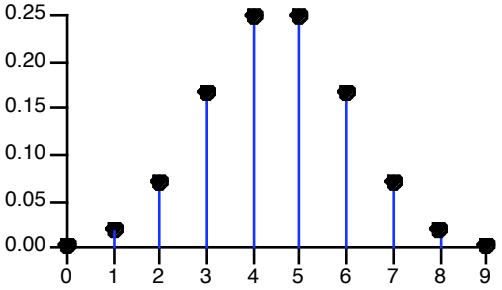
dove "E" sta per "Expectation" cioè aspettativa, valore atteso

Varianza

$$\sigma^2(X) = \sum_{i=1}^{n} x_i^2 p_i - \mu^2$$

Esempio

Un modello assegna le probabilità ai valori secondo la formula di seguito riportata e di cui è dato un esempio in figura per n=9.



$$p(i) = \frac{\binom{n}{i}}{2^n} \quad per \quad i = 0,1,2,\ldots,n;$$

Calcolare il valore atteso

$$\mu = \sum_{i=0}^{n} i \frac{\binom{n}{i}}{2^n} = \frac{n2^{n-1}}{2^n} = \frac{n}{2}$$

Esercizio

La autocarrozzeria RUG.PAL effettua 5 diverse categorie di interventi ed ogni settimana ripete lo stesso numero di interventi per ogni categoria.

Gli incassi sono i seguenti:

Servizio	Incasso	P(X)	F(X)
1	120	1/5	1/5
2	60	1/5	2/5
3	20	1/5	3/5
4	80	1/5	4/5
5	70	1/5	5/5

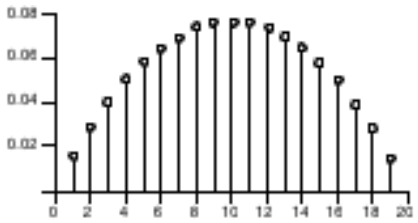
a) Qual'è l'incasso atteso per una settimana qualsiasi?

$$\mu = \frac{120+60+20+80+70}{5} = 70$$

b) Qual'è la varianza dell'incasso atteso?

$$\sigma^2 = \frac{120^2+60^2+20^2+80^2+70^2}{5} - 70^2 = 1040$$

Esercizio



Alcuni fenomeni si manifestano con probabilità speculari rispetto al centro. Un modello che risponde a tale requisito è:

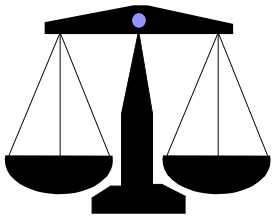
$$p_i = \frac{i*(n-i)}{\left(n/6\right)\left(n^2-1\right)}; \qquad i = 0,1,2,\ldots,n$$

Calcolare lo scarto quadratico medio se n=20

μ=10; σ=4.45

Qui si può adoperare il foglio elettronico

Equità dei giochi



Un gioco d'azzardo si dice EQUO se le poste dei giocatori sono proporzionali alle rispettive probabilità di vincita

ESEMPIO:
Nel lancio del dado l'uscita singola è data 5:1. Per una puntata di 5€, in caso di vincita dovrei incassare 30€ (i miei 5 più 25€ di vincita.

Se perdo, il banco dovrebbe trattenere solo 1€ restituire 4€ per compensare le sue maggiori probabilità:

$$5\left(\frac{1}{6}\right) - 1\left(\frac{5}{6}\right) = 1 - 1 = 0$$

La speranza matematica E(G) è l'importo certo che si è disposti a pagare per ricevere in cambio un importo aleatorio maggiore

Equità dei giochi/2

Se "p" è la probabilità di vincere una scommessa G in cui vi sia la promessa di vincere una cifra "x" con probabilità "p" e di perdere "y" con probabilità (1-p) l'esito atteso è:

$$E(G) = xp - y(1-p) = xp - y + yp = (x+y)p - y$$

Perché il gioco sia equo si deve avere

$$E(G) = 0 \Rightarrow (x+y)p - y = 0 \Rightarrow p = \frac{y}{x+y},\ 1-p = \frac{x}{x+y}$$

Ovvero: $x = \left(\frac{1-p}{p}\right)y$

Se perdo 1€ con probabilità1/30, quando vinco debbo incassare

$$X = \frac{\frac{30}{30} - \frac{1}{30}}{\frac{1}{30}} = \frac{29}{1} = 29$$

Esempio

Nella roulette americana decidete di giocare $ Y sul nero con P(N)=18/38 e P(N^c)=20/38 (in questo tipo di roulette ci sono lo "0" ed il "00" di colore verde).
Se esce il nero ricevete 2Y (inclusivi della vostra puntata).

E' un gioco equo? $X = \frac{\frac{20}{38}}{\frac{18}{38}} = \frac{20}{18} = 1.11$

Se il gioco fosse equo puntando 27$ se ne dovrebbero incassare 27+1.11*27 cioè la giocata più l'equa vincita.

Invece di (2.11*27)=56.97 di vincita il casinò dà 54. Questo si spiega per le spese organizzative, di manutenzione e gestione, ma lo scarto dell'11% è alto.

La tassa sulla stupidità

Se Ciccillo si gioca 5€ per un ambo sulla ruota di Cagliari (perché comincia con la sua lettera) la sua aspettativa di guadagno è

$$x = \left(\frac{\frac{7090}{8010}}{\frac{20}{8010}} \right) 5 = \left(\frac{7090}{20} \right) 5 = 1772.5$$

Il gestore pada invece 250 volte la posta cioè 1250€.

La differenza è in parte da attribuire alle spese di organizzazione, ma possono incidere con una decurtazione del 30% ?

Il lotto ed altri giochi gestiti dallo stato sono iniqui e sarebbe stupido giocarci qualora ci fossero alternative più convenienti

Tuttavia, la tassa sulla dabbenaggine dei giocatori trova parziale giustificazione nell'interesse pubblico con cui si impiegano i fondi così ottenuti.

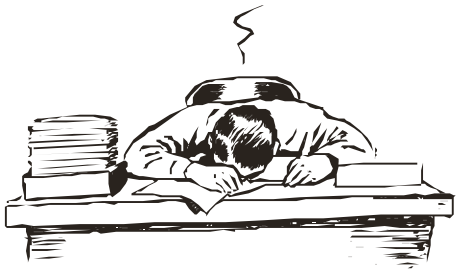
Dal discreto al continuo

Le distribuzioni di probabilità già viste servono a rappresentare delle caratteristiche discrete

L'insieme dei valori possibili è formato da punti isolati che possono essere "contati" cioè posti in corrispondenza biunivoca con l'insieme dei numeri naturali.

Sono inadatte per descrivere il comportamento di quantità i cui valori ricadono in un intervallo di numeri reali: Distanze,pesi, altezze, etc.

Per affrontare questi aspetti è necessario ampliare il nostro bagaglio di strumenti d'analisi



La funzione di densità

La funzione di densità di una v.c. continua "X" è

1. $f(X) \geq 0 \quad se \quad a \leq X \leq b$
2. $f(X) = 0 \quad se \quad (X < a) \cup (X > b)$
3. $\int_{-\infty}^{\infty} f(X)dx = 1$

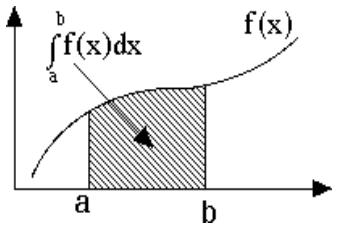
Le 1^ deriva dal fatto che la probabilità non può essere negativa.

La 2^ estende a tutto l'asse reale il campo di variazione della X.

L'intervallo (a,b) dove f(x)>0 è detto SUPPORTO della v.c.

La 3^ deriva dal fatto che l'evento certo deve avere probabilità 1

il simbolo " $\int_a^b$ **" indica l'integrale della f(X) in (a,b) cioè la misura dell'area sottesa alla curva della funzione f(x) nell'intervallo (a,b)**



N.B. La f(X) non dà la probabilità di X, ma è proporzionale alla probabilità che X ricada in un intervallo infinitesimo centrato su X.

Nota sulla distribuzione di probabilità

In genere, la distribuzione diuna variabile non può essere ricostruita a partire dalla situazione sperimentale.

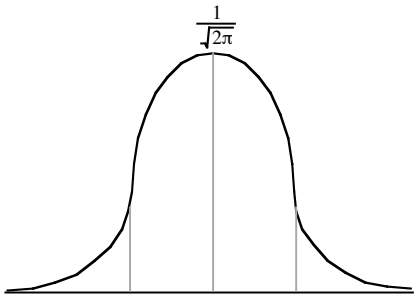
Talvolta lo schema probabilistico fornisce distribuzioni del tutto intrattabili.

Compito della Statistica è di definire dei "modelli" di distribuzione e di aiutare a scegliere quello più adatto alla particolare situazione di studio.

Tra i tanti modelli proposti e sviluppati dalla statistica studieremo in dettaglio solo il modello gaussiano

La curva normale (o gaussiana)

E' il modello di probabilità più noto e più usato in statistica



La tilde si legge "distribuito come"

$$X \sim N(\mu,\sigma) \rightarrow h(x) = \frac{e^{-\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}; \quad -\infty < \mu < \infty; \ \sigma > 0$$

Media aritmetica μ

Varianza σ^2

La curva è simmetrica intorno a *μ*

L'andamento campanulare e simmetrico della curva Normale sta ad indicare che:

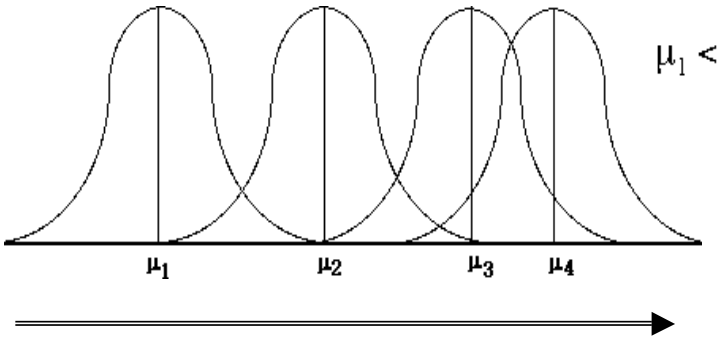
1) Gli scostamenti negativi dal centro sono altrettanto probabili di quelli positivi;
2) I valori sono addensati intorno al centro;
3) Gli scostamenti si verificano con probabilità decrescente man mano che diventano grandi in valore assoluto.

Significato del parametro "*μ*"

La densità è unimodale e l'ordinata massima si raggiunge per X=*μ* (la moda)

Quindi, il parametro *μ* rappresenta il valore più probabili nonché il valore atteso e quello che bipartisce il supporto dei valori.

Cambiando *μ* si modifica la collocazione del grafico



$\mu_1 < \mu_2 < \mu_3 < \mu_4$

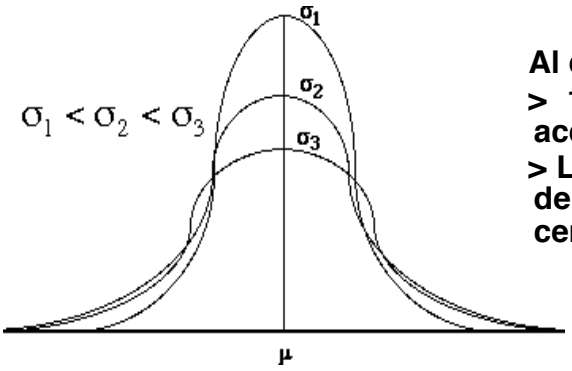
Al variare di *μ* il grafico resta inalterato nella sua forma. Si modifica solo la sua localizzazione: più a destra se *μ* aumenta; più a sinistra se *μ* diminuisce

Significato del parametro "σ"

Il σ corrisponde allo scarto quadratico medio.

La curvatura del grafico della distribuzione normale cambia due volte inflessione in corrispondenza dei punti x=*μ* ± σ. Inoltre

Quindi $f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$



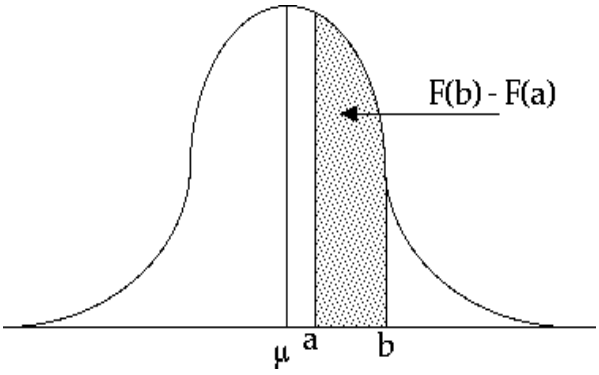
Al diminuire di σ :
> 1 due punti di flesso tendono ad accentrarsi;
> L'ordinata massima aumenta a causa del maggiore addensamento intorno al centro della distribuzione.

La funzione di ripartizione

La formula della Normale è complicata e l'integrale

$$F(x) = P(X \le x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$

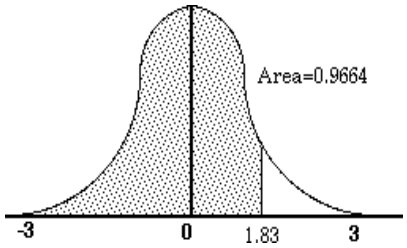
è calcolato con metodi di approssimazione numerica



Esprime la probabilità di osservare un valore della normale tra "a" e "b"

Calcolo delle aree sottese alla curva

Esempio_1: calcolare l'area compresa fra $-\infty$ e 1.83, in simboli: $\phi(1.83)$

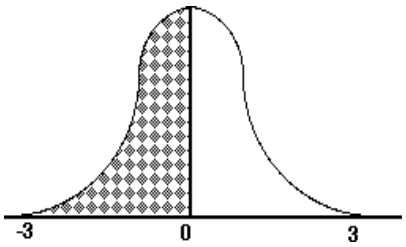


$\phi(1.83) = 0.9664$

$P(Z \leq 1.83)$

=DISTRIB.NORM.ST()

Esempio_2: calcolare l'area compresa fra $-\infty$ e 0, in simboli: $\phi(0)$



$2\phi(0)=1$ e quindi $\phi(0)=\frac{1}{2}$

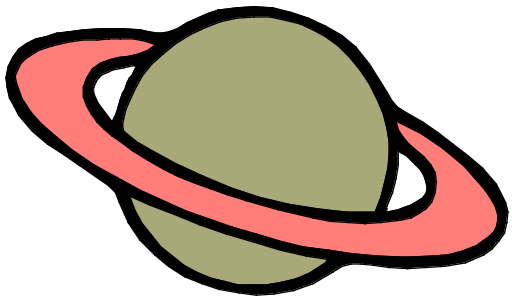
$P(Z \leq 0)$

Importanza della normale

Risiede nel fatto che moltissimi fenomeni possono esservi rappresentati.

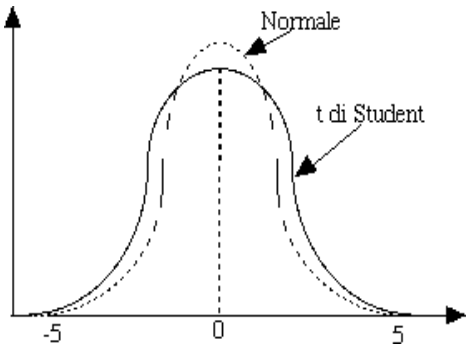
Infatti, la normale serve da efficace approssimazione di molte altre variabili casuali continue e discrete.

Già nel 17° secolo, Galileo discusse il comportamento delle misurazioni delle distanze astronomiche avendo in mente il modello normale della compensazione tra errori di segno opposto.



La t di Student

Questo modello è molto simile a quello gaussiano, ma ha code più "spesse" (ordinate estreme più alte)



$$-\infty < t < \infty;\ n > 2$$

$$E(t_n) = 0; \quad Var(t_n) = \frac{n}{n-2}$$

La varianza è superiore all'unità, ma si avvicina ad uno all'aumentare di "n"

L'elemento caratterizzante della t di Student sono "i gradi di libertà" cioè l'ampiezza campionaria ridotta di una unità: n-1. ("n-1" è il parametro della t di Student)

Per ogni grado di libertà esiste una t di Student, sebbene queste diventino poco distinguibili per n≥60.

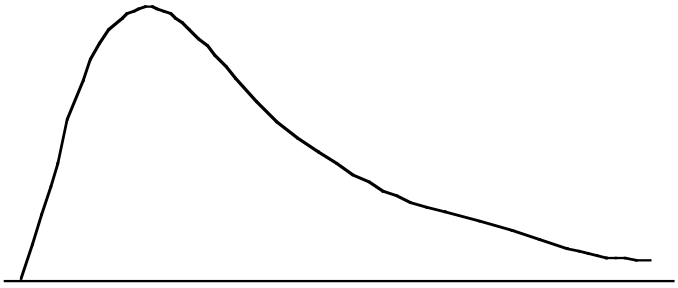
Questa v.c. è stata analizzata da W.S. Gosset, nel 1906, che firmò l'articolo con lo pseudonimo "STUDENT" ed è da allora nota come "La t di Student"

χ^2 (chi-quadrato)

Questo modello è definito per i soli non negativi e presenta una marcata asimmetria positiva

Anche in questo caso l'elemento caratterizzante sono "i gradi di libertà" cioè "g"

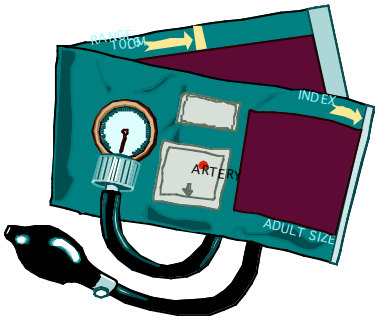
$$E\left(\chi^2\right) = g, \qquad Var\left(\chi^2\right) = 2g$$



Per "g" superiore a 30 la distribuzione del χ^2 si avvicina a quella gaussiana

La distribuzione del χ^2 si incontra nello studio campionario della variabilità.

Questa è importante nelle analisi cliniche e nel controllo della qualità



Statistica descrittiva ed inferenziale

L'escussione delle unità del campione rispetto alle variabili produce il data set

$$C = \{X_1,\ X_2,\ \ldots,\ X_n\}$$

cioè "m" osservazioni su di "n" unità. Nel nostro corso m=1 o 2.

Si parla di STATISTICA DESCRITTIVA se il data set è analizzato per quello che è senza uno sfondo su cui proiettare i dati

Emittenti	Ascolti	Emittenti	Ascolti	Emittenti	Ascolti	Emittenti	Ascolti
Radiouno	7616	Radioverderai	791	R.D.S.	2671	Lattemiele	1145
Radiodue	6137	Isoradio	594	Rete 105	2607	Radio cuore	1135
Radiotre	1458	Radio deejay	3687	RTL 102.5	2112	Radio Maria	1105
Stereorai	1282	Radio italia SMI	3178	Radio Radicale	1541	Italia Network	1056
CNR	1468	Radio Montecarlo	1460	Radio Kiss Kiss	1393	Kiss Kiss Italia	972
105 Classic	786						

Valore massimo per Radiouno; minimo per Isoradio; c'è un gruppo che si addensa intorno a 1000-1500 ascolti.

Le reti pubbliche sono più diffuse di quelle commerciali

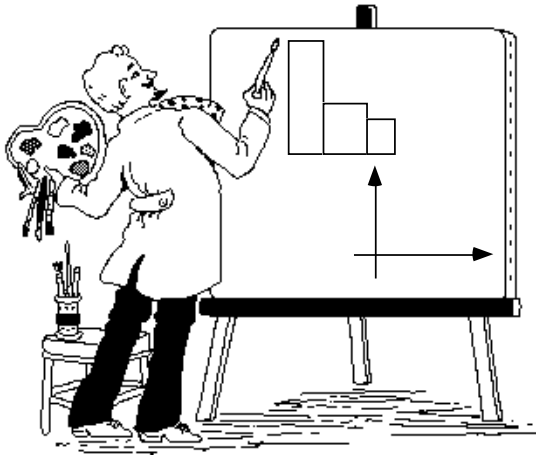
Statistica descrittiva

La STATISTICA DESCRITTIVA mira alla organizzazione, all'analisi tabellare e grafica nonché al calcolo di grandezze sintetiche di ciò che si è rinvenuto nella rilevazione

E' anche nota come analisi esplorativa *(Exploratory Data Analisys)* proposta soprattutto dall'americano J.W. Tukey nel 1977

In breve, si configura come una trattazione preliminare indispensabile per affrontare uno studio complesso.

Utilizza tecniche elementari, soprattutto grafiche.



Statistica inferenziale

Inizia laddove il data set è visto come la punta di un iceberg.

I dati sono solo una delle possibili realizzazioni e riguardano anche gli ascolti che potevano esserci, ma non ci sono stati nonché gli ascolti che ci saranno o potranno esserci in futuro.

In che modo ed in che misura possiamo estendere agli ascolti potenziali le quantità calcolate sui valori osservati?

Questa è STATISTICA INFERENZIALE

Logica della Inferenza statistica

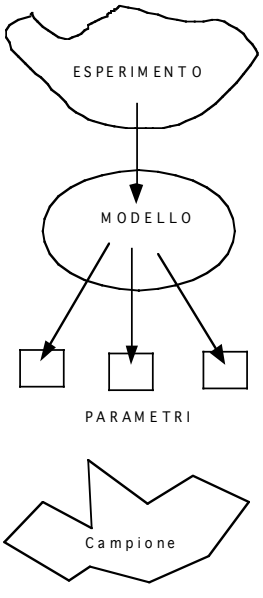
Le situazioni in cui la statistica si è più affermata sono gli esperimenti replicabili all'infinito

Le esigenze conoscitive si limitano spesso a poche caratteristiche dell'esperimento: valore atteso e varianza di una o più variabili.

Tali caratteristiche sono spesso i parametri del modello che descrive il comportamento delle variabili casuali.

Il modello di casualità è approssimabile dalla variabile casuale normale

Come sfruttare al meglio le informazioni del campione per determinare il valore dei parametri?



Le procedure inferenziali

Ciò che interessa il nostro corso è:

LA STIMA DEI PARAMETRI

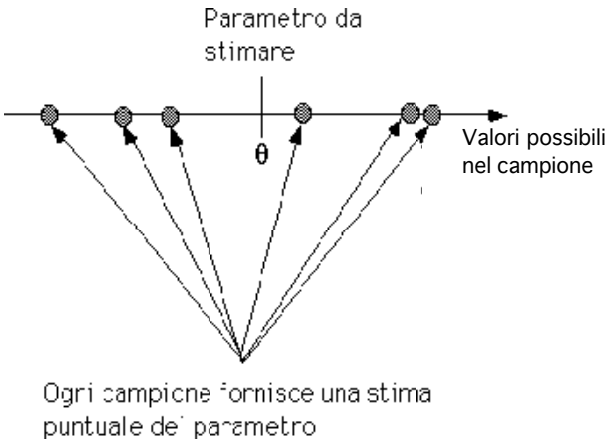
- **PUNTUALE: quando si propone un singolo valore come stima di un parametro dell variabile casuale.**
- **INTERVALLARE: quando si propone un ventaglio di valori ragionevoli come stime.**

LA VERIFICA DI IPOTESI

Da esperienze precedenti o dalla logica delle indagini si può supporre che i parametri abbiano determinati valori. Sono compatibili con le risultanze campionarie?

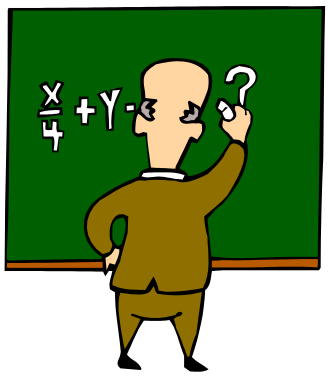
La stima puntuale

E' la procedura più semplice: in base alle osservazioni campionarie si ottiene il valore da sostituire al parametro da stimare



C'è da aspettarsi un certo scarto tra la stima puntuale ed il parametro incognito, ma in genere non conosciamo né l'entità né il segno dell'errore

Gli stimatori



Lo stimatore è una funzione NOTA dei valori inclusi in un campione casuale. Il suo valore è la STIMA

E' caratteristica quantitativa della popolazione dalla quale il campione è stato estratto.

Esempi di stimatori:

$$\text{Totale: } Q = \sum_{i=1}^{n} X_i; \quad \text{Campo di variazione: } R = X_{max} - X_{min}$$

$$\text{Media : } \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}; \quad \text{Varianza: } s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}; \quad \text{frazione di successi: } \pi = \frac{S_n}{n}$$

Uno stimatore è detto naturale se ciò che si calcola nel campione è in stretta analogia con ciò che si deve stimare nella popolazione

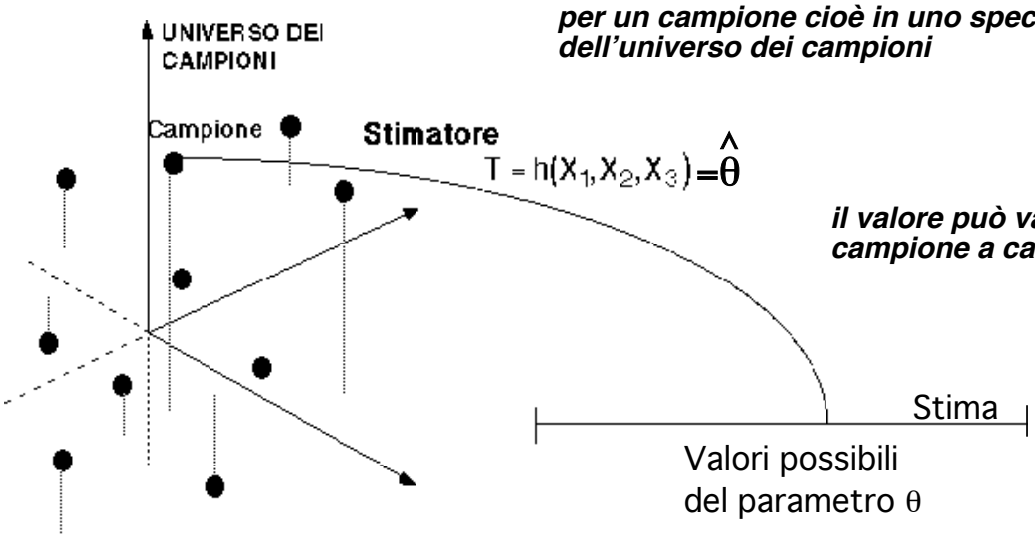
Stimatore e stima

ESEMPIO: Quale stipendio si può aspettare la manager di una USL?

Si sceglie un campione casuale diciamo di n=3 manager già in servizio e si calcola il valore atteso della loro retribuzione. Supponiamo che sia $\bar{x} = 65$

il valore "65 mila euro" è una STIMA del salario ipotetico, la media campionaria è uno STIMATORE del salario.

La stima è il valore assunto dallo stimatore per un campione cioè in uno specifico punto dell'universo dei campioni



il valore può variare da campione a campione

Esempio

L'estrazione del campione produce la n-tupla $(x_1, x_2, \ldots, x_n)$ i cui elementi sono le osservazioni campionarie

Ogni n-tupla, a sua volta, produce un valore dello stimatore

Esempio:

Si esamina un campione casuale di 10 imprese e si rileva X il numero di dipendenti regolari.

Il valore della X è casuale perché non è certa quale azienda finirà nel campione

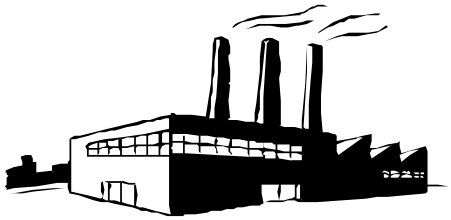
Osservazioni campionarie

5	0
3	2
1	4
3	2
2	3

Calcoliamo alcuni stimatori

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{5+0+3+2+1+4+3+2+2+3}{10} = \frac{25}{10} = 2.5$$

$$s^2 = \frac{\sum_{i=1}^{10} (x_i - 10)^2}{10-1} = \frac{2.5^2 + 2.5^2 + 0.5^2 + 0.5^2 + 1.5^2 + \\ +1.5^2 + 0.5^2 + 0.5^2 + 0.5^2 + 0.5^2}{9} = 2.06$$



La distribuzione degli stimatori

Lo stimatore è una variabile casuale connessa all'esperimento: estrazione casuale di un campione.

Conoscere la sua distribuzione ci serve per descrivere l'andamento dei risultati che si possono osservare replicando il piano di campionamento.

Dobbiamo ricordare che…

- **Stimare qualcosa significa dare un valore a quel qualcosa**
- **La stima ottenuta da un campione può essere diversa da quella ottenuta con un altro campione**
- **La stima tende differire dal parametro da stimare, ma se conosciamo la distribuzione campionaria dello stimatore possiamo quantificare probabilisticamente l'errore**

Legge forte dei grandi numeri

Di solito si ignora la variabile casuale che può descrivere in modo soddisfacente un dato aspetto della popolazione.

Di conseguenza non è possibile costruire la distribuzione di uno stimatore.

Inoltre, uno stesso stimatore ha una distribuzione campionaria diversa in dipendenza del tipo di variabile casuale che descrive la popolazione.

C'è una via d'uscita? La legge forte dei grandi numeri!

Se la distribuzione non è nota, ma il campione casuale è abbastanza numeroso e le estrazioni sono indipendenti (o virtualmente tali) è possibile approssimare la distribuzione degli stimatori con il MODELLO NORMALE

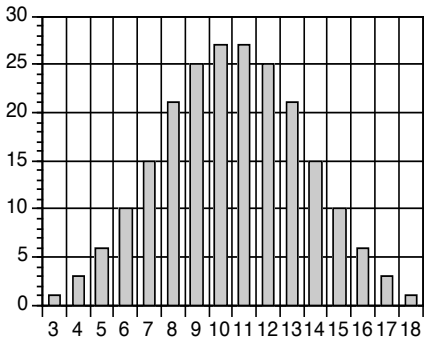
Esempi

Somma del lancio di "n" dadi (sorte benigna)

$$S = X_1 + X_2 + \ldots + X_n$$

$$E(S) = n * 3.5;\ \sigma^2(S) = n\frac{35}{12};\ Z = \frac{S - n * 3.5}{\sqrt{n} * 2.9167}$$

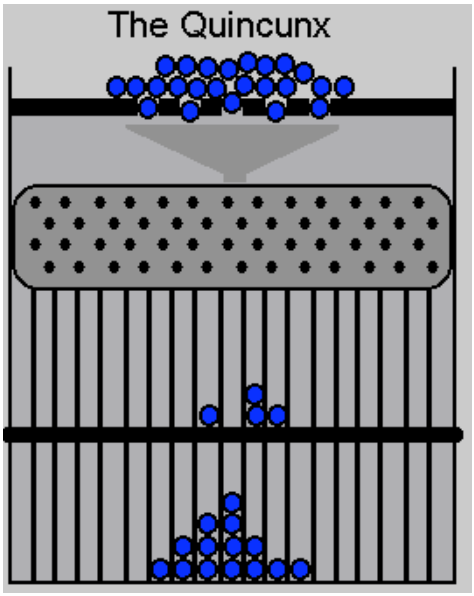
per n=3



Oscillazioni di un titolo di borsa (sorte selvaggia)

Alcune di queste possono essere catastrofiche o superpositive e non c'è convergenza alla normale perché la varianza è infinita

Quincux



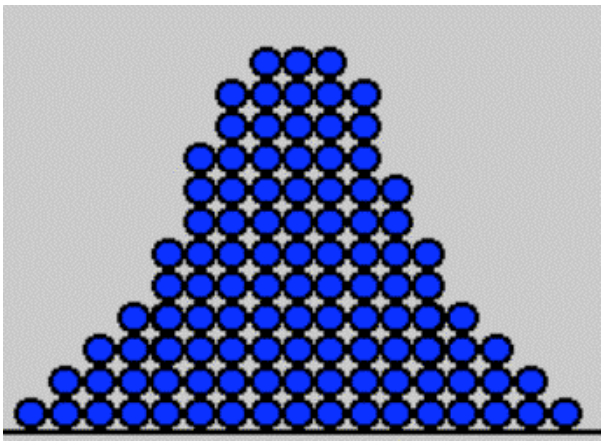
1) Le biglie entrano nell'imbuto dai vari fori

2) Le biglie escono dall'imbuto una alla volta

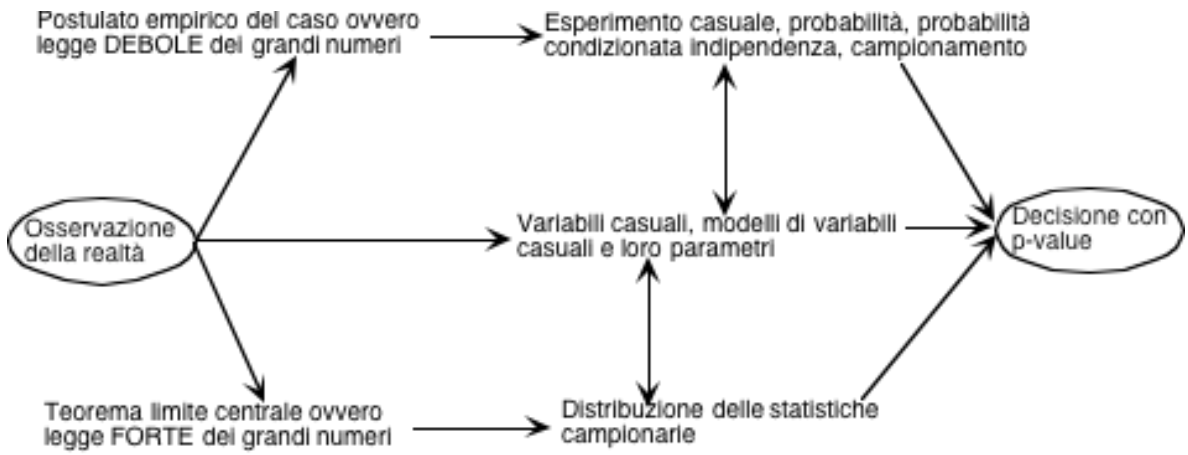
3) Le biglie rimbalzano a caso tra i vari pioli

4) Ogni biglia imbocca unasola scanalatura

Risultato finale



Schema dell'inferenza statistica



Test delle ipotesi

Non si è più alla ricerca di un valore (o un intervallo di valori) da sostituire al parametro incognito, ma si deve stabilire quale, tra due ipotesi, è più probabilmente VERA

Se la decisione si potesse basare sulla conoscenza totale si avrebbe una conclusione definitiva: l'ipotesi è VERA o FALSA.

Come in molte scienze sperimentali non potremo dimostrare vera o falsa una affermazione.

Potremo solo affermare: è più coerente o meno coerente con i nostri dati campionari.

Formalismo dei Test

L'IPOTESI STATISTICA H_0 è una asserzione verificabile su di una variabile casuale. In genere riguarda i suoi parametri.

Supponiamo di conoscere la funzione di densità della v.c.: f(X;θ) di cui ignoriamo il valore del parametro θ.

In genere, la H_0 ipotizza un certo valore del parametro e ne valuta la sua conformità con i dati campionari

$$H_0 : \ \theta - \theta_0 = 0$$

È detta "nulla" perché, di solito, niente cambia se viene accettata

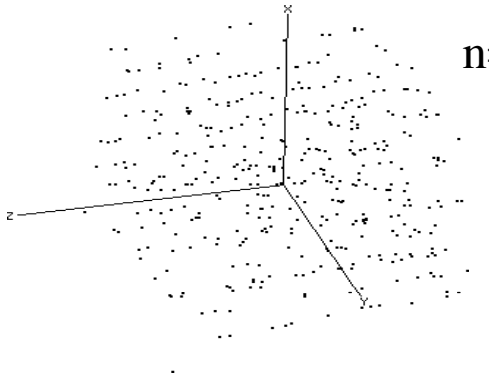
il valore θ_0 è scelto in base a varie considerazioni:

- **Dalla logica dell'esperimento**
- **Da esperienze precedenti o indagini pilota**
- **E' un livello critico *(baseline)***
- **E' un livello desiderato**

Formalismo del Test/2

Consideriamo L'UNIVERSO DEI CAMPIONI di ampiezza "n" cioè l'insieme di tutte le possibili realizzazioni della n-tupla …

$$(X_1, X_2, \ldots, X_n)$$



In tale spazio, il campione estratto è solo un punto.

L'idea del test è di assegnare ad ogni campione una probabilità detta

LIVELLO OSSERVATO DI SIGNIFICATIVITA' (p-Value)

e in base a questo decidere sull'ipotesi H_0.

La statistica test

E' l'anello di congiunzione tra universo dei campioni e valori del parametro

La distribuzione della statistica test T(X;θ) è funzione:

- Del campione casuale $(X_1, X_2, \ldots, X_n)$
- Del parametro da stimare: θ

Le statistiche test più in uso sono del tipo: $T(X;\theta) = \dfrac{T - \theta}{\sigma(T)}$

Spesso T(X;θ) è lo stimatore naturale del parametro a cui sono riferite le ipotesi

Applicazione/1

ESEMPIO:

Una docente sa che -storicamente- i risultati dei suoi esami scritti hanno μ=25, σ=3.

Però, nell'ultima prova, i risultati dei primi 10 compiti sono molto scadenti. Che si tratti un corso ad alta densità di ciucci?

$$Ipotesi\ \ nulla\quad H_0:\ (\mu - 25) = 0$$

$$Ipotesi\ \ alternativa\quad H_1:\ \mu < 25$$

Come statistica test sembra ovvio scegliere la media campionaria.

La docente rifiuterà H_0 se il voto medio osservato nel campione sarà molto più piccolo di 25.

Il fatto che si rifiuti H_0 non implica necessariamente che accetti H_1. Potrebbe anche decidere di rinviare la decisione in attesa di avere più dati.

L'ipotesi alternativa è inserita soprattutto per stabilire la direzione del test

Applicazione/2

N.B. si è sostituito l'evento punto μ=23 con l'evento intervallo $\mu \leq 23$

nel campione si trova: $\hat{\mu}$ =23

Se l'ipotesi H0 fosse vera, quale sarebbe la probabilità di osservare una media campionaria inferiore o uguale a 23?

Supponendo che la distribuzione sia gaussiana , avremo

"|" significa: sotto H0"

$$P(\hat{\mu} \leq 23|H_0) = P\left(Z \leq \sqrt{10}\left(\frac{23-25}{3}\right)|H_0\right) = P(Z \leq -2.11|H_0) = 0.0174$$

sotto l'ipotesi nulla (la media è 25) è altamente improbabile (1.7%) osservare un voto medio di 23 o meno nel campione estratto da una popolazione con μ=25

Ne consegue che lo scarto 23-25 non è attribuibile alle fluttuazioni campionarie, ma fa invece pensare ad una classe particolarmente ciuccia.

Applicazione/3

Dopo aver corretto altri 25 compiti risulta che, su questi, si ha $\hat{\mu} = 27$

A questo punto sorge un altro dubbio: che l'assistente del corso, imbranato come una foca, abbia mischiato i compiti con quelli della classe "advanced"

$$Ipotesi\ \ nulla \quad H_0: (\mu - 25) = 0$$
$$Ipotesi\ \ alternativa \quad H_1: \mu > 25$$

La direzione del test è ora verso i valori alti poiché i valori inferiori a 25 non fanno sorgere questo tipo di dubbio

$$P(\hat{\mu} \geq 27 | \mu = 25) \Rightarrow P\left(\frac{\hat{\mu} - 25}{1.6667} \geq \frac{27 - 25}{1.6667}\right) = P(Z \geq 1.2) = 1 - 0.8849 = 0.1151$$

Lo scarto 25-27 è poco compatibile con l'ipotesi sebbene non ci siano evidenze fortissime contro.

Un valore superiore maggiore o uguale di 27 lo si può trovare nell'11% dei campioni da popolazioni aventi media 25.

L'assistente si salva.

Esempio

La dott.ssa Angelina Romano propone una procedura d'ufficio che riduce i tempi medi rispetto agli attuali μ=75 minuti, pur conservando lo stesso σ=9.

$$Ipotesi\ \ nulla \quad H_0: (\mu - 75) = 0$$
$$Ipotesi\ \ alternativa \quad H_1: \mu < 75$$

Si considera un campione di n=25 pratiche. Quindi la media del campione -sotto H_0- sarà approssimata dalla gaussiana con μ=75 e σ=9/√25=1.80

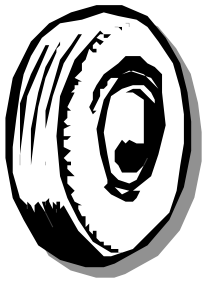
Nel campione si trova $\hat{\mu} = 69.8$

Qual 'è la probabilità di ottenere un valore della statistica test inferiore o uguale del valore osservato se ma media della popolazione è 75?

$$P(\hat{\mu} \leq 69.8 | \mu = 75) \Rightarrow P\left(\frac{\hat{\mu} - 75}{1.8} \leq \frac{69.8 - 75}{1.8}\right) = P\left(\frac{\hat{\mu} - 75}{1.8} \leq -2.89\right) = 0.0019$$

Esempio

L'amministratore delegato, Sig.ra Rosetta Gaudio, di una fabbrica di pneumatici sta valutando la modifica della trama del prodotto leader.



Lo studio di fattibilità segnala che la nuova trama è conveniente se la vita media dei prodotti supera le 20'000 miglia in condizioni standard.

$$Ipotesi\ \ nulla \quad H_0: \ (\mu - 20'\,000) = 0$$

$$Ipotesi\ \ alternativa \quad H_1: \ \mu > 20'\,000$$

Un campione di n=16 prototipi viene provato dando luogo a $\hat{\mu} = 20{,}758$

Si sa che σ=6'000. L'amministratore Gaudio che deve fare?

$$P\left(\hat{\mu} \geq 20758 \middle| \mu = 20000\right) \Rightarrow P\left(\frac{\hat{\mu} - 20000}{1500} \geq \frac{20758 - 20000}{1500}\right) =$$

$$P\left(\frac{\hat{\mu} - 20000}{1500} \geq 0.39\right) = 1 - P\left(\frac{\hat{\mu} - 20000}{1500} < 0.39\right) = 0.3483$$

P-value

Indica la probabilità che valori della statistica test -inferiori o uguali a quello osservato- siano sopravvenuti solo per effetto della sorte.

Quindi, il P-value misura la probabilità di sbagliare, nelle condizioni date, se si rifiuta l'ipotesi nulla

Nuova procedura amministrativa

$$Ipotesi\ \ nulla \quad H_0: \ (\mu - 75) = 0, \ \ P - value = 0.0019$$

La nuova procedura potrebbe essere non migliorativa rispetto alla vecchia solo in 2 casi su 1000 (circa). E' bene rifiutare H_0

Nuovo pneumatico

$$Ipotesi\ \ nulla \quad H_0: \ (\mu - 20'\,000) = 0, \ \ P - value = 0.3483$$

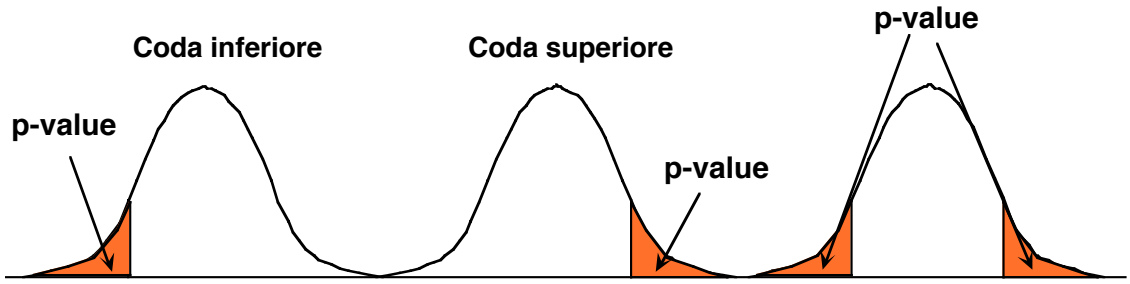
La nuova trama è migliorativa circa una volta su 3. Non è' consigliabile rifiutare H_0.

Precisazioni

Rispetto all'ipotesi che il parametro abbia un valore prefissato ci sono tre casi:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta < \theta_0 \end{cases}; \qquad \begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}; \qquad \begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_0 \end{cases}$$

Nei primi due il test è unidirezionale (o ad una coda), nel terzo è bidirezionale (o a due code).



Il giudizio sull'entità dello scarto tra valore campionario ed ipotesi è espresso in base alla distribuzione della statistica test

Guida

Se P -value ≤ 1%.	***Aldilà di ogni ragionevole dubbio si può rifiutare H_0***
Se 1% ≤ P -value ≤ 5%.	***Ci sono buone ragioni per rifiutare H_0***
Se 5% ≤ P -value ≤ 10%.	***Ci sono ragioni per rifiutare H_0, ma non sono del tutto convincenti***
Se P -value > 10%.	***E' consigliabile non rifiutare H_0***

I valori sono solo apparentemente bassi.

Le condizioni di applicabilità dei test (ad esempio la distribuzione normale) sono valide solo in parte).

Di conseguenza, solo una forte evidenza può convincere a rifiutare l'ipotesi nulla (angolatura conservativa)

Esempio

Una compagnia controlla la situazione di magazzino in base alle vendite dell'anno precedente. Si supponga che gli ordini dell'anno precedente abbiano dato luogo

$$\mu = 36 \text{ (miliardi)}; \quad \sigma = 2$$

Nel primo quadrimestre (n=225) si è riscontrato $\bar{x} = 35.82$

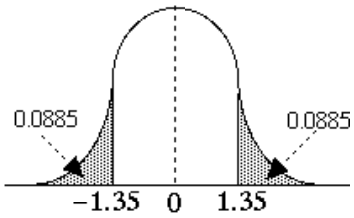
E' possibile che ci sia un diverso andamento delle vendite? $\begin{cases} H_0: \mu = 36 \\ H_1: \mu \neq 36 \end{cases}$

Supponendo σ noto avremo $Z_c = \sqrt{225}\left(\frac{35.82 - 36.00}{2}\right) = -1.35$

A quale livello è significativo?

$$\phi(-1.35) = 1 - \phi(1.35) = 0.0885$$

Poiché il test è a due code avremo $\alpha = 2 * (0.0885) = 0.177$



Cioè circa 18 volte su 100, il campione di n=225 estratto dalla popolazione N(36;2) disterà più di 1.35 dalla media ipotizzata. Il valore 35.82 non è allora troppo strano.

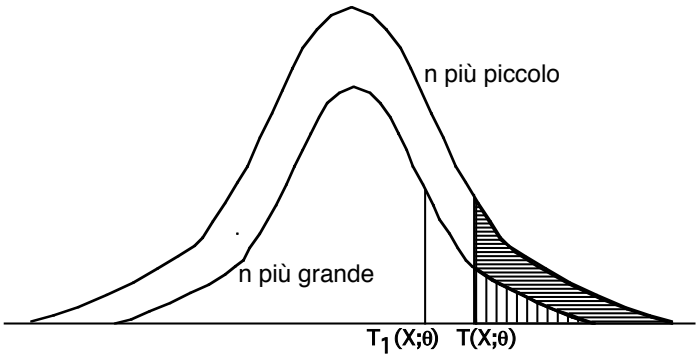
Non ci sono ragioni sufficienti per pensare ad un cambiamento

Ampiezza del campione e p-value

La statistica test è uno stimatore consistente del parametro sotto ipotesi.

Quindi, all'aumentare dell'ampiezza del campione, la sua variabilità si riduce.

Questo implica che le code della distribuzione della statistica test diventano più sottili.



A parità di p-value, la corrispondente statistica test è inferiore.

Ovvero, la stessa statistica test può avere un p-value più piccolo perché il campione è più grande.

Campioni molto grandi possono rendere significative dei valori della statistica test poco rilevanti dal punto di vista pratico.

Esercizio

La produzione media per ettaro è stata di 18 Kg con σ=7.12. L'uso di un nuovo fertilizzante su n=30 ettari sperimentali, ha dato luogo a

$$\hat{\mu} = 19.5$$

$$Ipotesi\ \ nulla \quad H_0 : \ (\mu - 18) = 0$$

$$P(\hat{\mu} \geq 19.5 | \mu = 18) \Rightarrow P\left(\frac{\hat{\mu} - 18}{7.12/\sqrt{30}} \geq \frac{19.5 - 18}{7.12/\sqrt{30}} \right) =$$

$$P(Z \geq 1.15) = 1 - P(Z < 1.15) = 0.1251$$

Il p-value è relativamente elevato. Non sembra ci sia sufficiente evidenza che il nuovo fertilizzante incida significativamente sulla produttività.

Excel1

Importo					
42.62		Ampiezza	34	=Conta.Numeri(A2:A35)	
46.68		Campionaria			
51.73					
62.48		Media	48.84	=Media(A2:A35)	
50.28		Campionaria			
43.86					
52.47		Varianza nota	25		
40.99		Popolazione			
37.97					
46.00		H_0: μ=	50		
45.14		H_1: μ<50			
44.62					
36.49		Statistica	-1.3501947	=(D5-D11)/radq(D8/D2)	
52.35		Test			
44.71					
46.21		p-Value	0.08847683	=Distr.Norm.St(D14)	
54.02					
68.47		Le ragioni per rifiutare H_0 non sono del tutto convincenti			
51.52					
55.42					
40.73					
50.75					
41.78					
61.31					
49.74					
48.24					
42.00					
49.19					
49.81					
45.34					
57.93					
54.32					
46.90					
48.56					

Importo medio speso in un super mercato da un campione di clienti

Excel2

Verifica di un possibile sbilanciamento di un indice aziendale

Indice				
19.36		Ampiezza	24	=Conta(A2:A25)
18.94		Campionaria		
25.42				
18.93		Media	17.80	=Media(A2:A25)
10.37		Campionaria		
11.12				
10.24		Varianza nota	36	
14.98		Popolazione		
14.89				
17.80		H_0: μ=	15.0	
20.26		H_1: μ > 0		
18.74				
19.42		Statistica	2.28487638	=(D5-D11)/radq(D8/D2)
16.15		Test		
11.02				
21.27	p-Value (coda superiore)		0.01116001	=1-Distr.Norm.St(D14)
21.11				
28.14				
29.43		Si può rifiutare l'ipotesi che l'indice aziendale sia in equilibrio		
16.42		Affermando che è sbilanciato verso l'alto si sbaglia l'1%		
13.54		delle volte		
10.03				
16.42				
23.17				

Excel3

Oscillazioni di un corso azionario. Verifica della stabilità.

Scarto				
-2.31	-0.36	Ampiezza Campionaria	42	=Conta(A2:A43)
-2.24	-4.56			
-3.35	-1.97	Media Campionaria	-0.79	=Media(A2:A43)
-0.84	-1.37			
-1.68	1.17	Varianza nota Popolazione	2.25	
-0.75	1.17			
1.36	0.25	H_0: μ=	0.0	
-3.72	-0.47	H_1: μ ≠ 0		
-1.88	-0.16			
1.31	-2.51	Statistica Test	-3.41908683	=(D5-D11)/radq(D8/D2)
0.37	-1.22			
-1.15	-1.22			
1.32	-0.83	p-Value (coda inferiore)	0.00031421	=Distr.Norm.St(D14)
-1.70	0.54	P-value (coda superiore)	0.99968579	
1.11	0.31	P-value (totale)	0.00062842	=2*Min(D17,D18)
-1.29	-1.71			
1.36	-1.43	Si può senzaltro rifiutare l'ipotesi che il corso azionario		
-1.81	0.76	si stia stabilizzando intorno allo zero		
-1.67	-1.47			
0.64	0.28			
-1.05	-0.49			

Se σ è incognito...

Il calcolo del p-value ha finora ipotizzato che σ fosse noto e questo è molto irrealistico.

In genere "σ" è stimato con

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \mu\right)^2}{n-1}}$$

La statistica test utilizzata in questo caso è basata sulla t di Student

$$t_c = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \sqrt{n}\left(\frac{\hat{\mu} - \mu}{s}\right)$$

Esempio

Una società che produce mobili ha sempre saputo che per costruire un arredo completo sono necessarie *μ*=20 ore/lavoro.

Di recente, si è anche assunta manodopera con contratto di formazione e si teme che questo faccia aumentare i tempi.

Dalla rilevazione dei tempi su di un campione casuale di n=9 arredi risulta:

25, 29, 23, 23, 31, 21, 27, 25, 33

$$\hat{\mu} = \frac{\sum_{i=1}^{9} x_i}{9} \quad 26.33; \quad \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{9}\left(x_i - \hat{\mu}\right)^2}{8}} = 4$$

$$t_c = \sqrt{9}\left(\frac{26.33 - 20}{4}\right) = 4.748$$

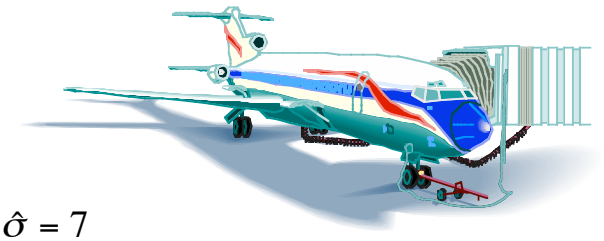
TDIST(4.748;8;1)

$$P\left(t_c \geq 4.748\right) = 1 - P\left(t_c \leq 4.748\right) = 0.0007$$

I tempi sono realmente aumentati, chi affermasse il contrario direbbe la verità solo 7 volte su 10'000.

Esempio

L'Antitrust ha disposto un controllo sulla compagnia aerea "Facta" che propone un volo AZ tra due noti scali a 110 minuti.



In un campione di n=49 si trova:

$$\hat{\mu} = 108, \quad \hat{\sigma} = 7$$

$$Ipotesi\ \ nulla \quad H_0: \ (\mu - 110) = 0$$

$$Ipotesi\ \ alternativa \quad H_1: \ \mu \neq 110$$

$$t_c = \sqrt{49}\left(\frac{108-110}{7}\right) = -2,$$

$$P(t_c \leq -2) = 1 - P(t_c \geq 2) = 1 - \left[1 - P(t_c \leq 2)\right] = P(t_c \leq 2) = 0.0256$$

$$p - value = 2(0.0256) = 0.0512$$

La dichiarazione della compagnia non sembra infondata, almeno alla luce del campione analizzato

Esercizio

La biologia di molti laghi cambia in peggio a causa delle piogge acide. Un lago è considerato NON Acido se Ph≥6

Ecco il livello di Ph rilevati da due studiosi italiani in n=15 laghi alpini

5.5	5.7	5.8	6.1	6.3
6.3	6.5	6.6	6.7	6.9
6.9	7.2	7.3	7.3	7.9

$$\hat{\mu} = 6.6$$

$$s = 0.6719$$

Cosa si può dire sui laghi della zona esaminata?

$$Ipotesi\ \ nulla \quad H_0: \ (\mu - 6) = 0$$

$$Ipotesi\ \ alternativa \quad H_1: \ (\mu > 6)$$

L'ipotesi può tranquillamente essere rifiutata con un errore probabile del 2 per mille

$$t_c = \sqrt{15}\left(\frac{6.6-6}{0.6719}\right) = 3.4586$$

$$P(t_c \geq 3.4586) = 1 - P(t_c \leq -3.4586) = 1 - \left[1 - P(t_c \leq 3.4586)\right] = P(t_c \leq 3.4586) = 0.0019$$

Esercizio

Offerte su base d'asta. La cifra di riferimento 28.7 milioni di euro costituisce la media delle proposte?

	A	B	C	D	E	F
1	offerte		TEST SULLA MEDIA -VARIANZA INCOGNITA			
2	26.56		Ampiezza	21	=Conta(A2:A22)	
3	27.25		Campionaria			
4	28.61					
5	25.73		Media	26.77	=Media(A2:A43)	
6	28.43		Campionaria			
7	26.82					
8	25.81		Varianza nota	2.2989		
9	26.63		Popolazione			
10	27.80					
11	20.29		H_0: μ=	28.7		
12	25.01		H_1: μ ≠ 28.7			
13	28.10					
14	29.98		Statistica	-3.840127	=Radq(D2)*(D5-D11)/D8	
15	29.42		Test			
16	29.23					
17	29.64		Gradi di libertà	20	=D2-1	
18	26.45					
19	26.56		P-value (totale)	0.00102216	=DISTRIB.T(-D14;D17;2)	
20	25.36					
21	23.04					
22	25.51		Si deve rifiutare l'ipotesi le offerte abbiano 28.7			
23			Come base di riferimento			

Test sulla proporzione

La percentuale di clienti di cui si finanzia il credito potrebbe sfuggire al controllo e deviare rispetto al 65% dello scorso anno.

L'ufficio fidi esegue un test al 10% e trova che, in un campione di n=315 sono stati affidati 214 clienti cioè H=0.6794

$$Ipotesi\ \ nulla\ \ \ H_0:\ (\pi - 0.65) = 0$$

$$Ipotesi\ \ alternativa\ \ \ H_1:\ \pi \neq 0.65$$

$$Z_c = \left(\frac{0.6794 - 0{,}65}{\sqrt{\frac{0.6794(1-0.6794)}{315}}} \right) = 1.118$$

$$P(Z_c \leq -1.118) = 0.1318$$

$$p - value = 2(0.1318) = 0.2636$$

Non si può rifiutare H_0

Esercizio

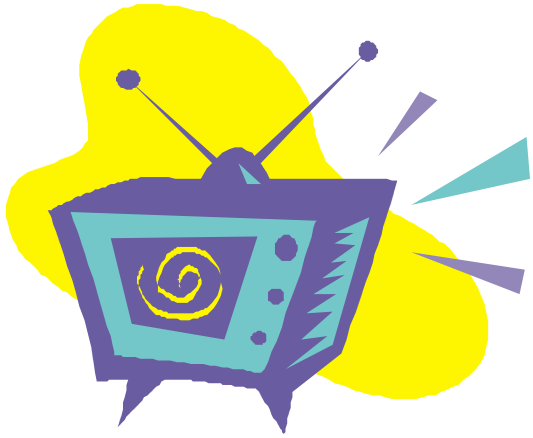
In un rapporto AUDITEL è riportato che le persone che guardano i film in televisione dopo mezzanotte sono equamente suddivise tra i due sessi.

In un campione casuale di 400 persone che guardano la televisione nelle ore piccole si sono trovati 180 donne.

a) Si effettui un test su H_0: π=50% contro H_1: π ≠ 50%;
b) Si effettui un test su H_0: π=50% contro H_1: π < 50%

a) $Z_c = 2, \;\; p-value = 4.6\%$

b) $Z_c = -2, \;\; p-value = 2.3\%$



Esercizio

Un'indagine sulle intenzioni di voto di un campione di universitari ha prodotto i seguenti risultati. Qual è la probabilità che il "non voto" resti sotto i 2/3?

	A	B	C	D	E	F	G
1	**Posizione**		TEST SULLE PROPORZIONI -Uso della Normale				
2	Voto	Non voto	Ampiezza	50	=Conta.Valori(A2:A51)		
3	Non voto	Non voto	Campionaria				
4	Non voto	Non voto					
5	Non voto	Non voto	Risposta	Non voto	=Media(A2:A43)		
6	Voto	Non voto	Di interessa				
7	Non voto	Voto					
8	Voto	Non voto	Conteggio	32	=CONTA.SE(D5)		
9	Non voto	Voto	risposte				
10	Non voto	Non voto					
11	Non voto	Non voto	Proporzione	0.6400	=D11/D2		
12	Voto	Voto	campionaria				
13	Voto	Non voto					
14	Non voto	Non voto	Approssimazione	0.0679	=RADQ(D11*(1-D11)/D2)		
15	Voto	Non voto	Dev.Standard				
16	Non voto	Voto					
17	Non voto	Non voto	H_0: π=0.66	0.6600			
18	Voto	Non voto	H_1: μ < 0.66				
19	Non voto	Voto					
20	Non voto	Voto	Statistica	-0.2946278	=(D11-D17)/D14		
21	Non Voto	Voto	Test				
22	Voto	Voto					
23	Non voto	Non voto	p-Value	0.38413917	=Distr.Norm.St(D20)		
24	Non voto	Non voto					
25	Non voto	Voto	Non c'evidenza campionaria che la percentuale del "non voto"				
26	Voto	Non voto	si mantenga al di sotto dei 2/3.				

Test sulla differenza tra medie
(Z-test cioè varianze note)

Indivhiamo con D la differenza ipotizzata tra le due medie (spesso è zero) abbiamo

$$Z_c = \frac{(\hat{\mu}_1 - \hat{\mu}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$H_1 : D < D_0 \Rightarrow p - value = \Phi(Z_c)$$
$$H_1 : D > D_0 \Rightarrow p - value = 1 - \Phi(Z_c)$$
$$H_1 : D \neq D_0 \Rightarrow p - value = 2\left[1 - \Phi(|Z_c|)\right]$$

La stessa procedura si applica se le ampiezze dei campioni sono entrambe molto grandi ed i campioni sono indipendenti

Basterà sostituire alle varianze incognite le loro stime campionarie

Esempio

La proprietà di un ristorante vuole sapere se la nuova campagna pubblicitaria ha aumentato gli incassi. Ecco i risultati per il periodo precedente la campagna e per il periodo seguente

Prima	Dopo
$n_1 = 50$	$n_2 = 30$
$\hat{\mu}_1 = 12.55$	$\hat{\mu}_2 = 13.30$
$\hat{\sigma}_1 = 215$	$\hat{\sigma}_2 = 238$

Per verificare l'ipotesi che la campagna sia stata efficace verifichiamo che $\mu_1 < \mu_2$

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 < \mu_2 \end{cases} \qquad Z_c = \frac{(12.55 - 13,30) - 0}{\sqrt{\frac{215^2}{50} + \frac{238^2}{30}}} = \frac{-75}{53.03} = -1.41$$

Poiché il p-value del test è del 7.9% non possiamo rifiutare l'ipotesi H_0.

A questo punto non c'è evidenza della efficacia della campagna

Esercizio

Un'indagine sulla spesa media di trasporti tra operati ed impiegati ha dato luogo ai seguenti risultati

La differenza nelle spese è significativa all'uno per mille

	A	B	C	D	E	F
1				TEST SULA DIFFERENZA TRA MEDIE - VARIANZE NOTE		
2	Operai	Impiegati		SRUMENTI-ANALISI DEI DATI-		
3	11.14	13.02		Test z: due campioni per medie		
4	16.51	19.77				
5	15.07	17.96		Test z: due campioni per medie		
6	18.41	22.15				
7	10.49	12.20			*Variabile 1*	*Variabile 2*
8	19.70	13.77		Media	15.9503	17.8072
9	14.56	17.32		Varianza nota	3.5	4.1
10	16.12	19.28		Osservazioni	35	19
11	17.52	21.03		Differenza ipotizzata per le medie	0	
12	21.60	22.67		z	-3.3044	
13	16.11	19.27		P(Z<=z) una coda	0.0005	
14	17.44	15.91		z critico una coda	1.6449	
15	16.88	20.23		P(Z<=z) due code	0.0010	
16	12.80	15.11		z critico due code	1.9600	
17	12.94	15.28				
18	16.98	20.36				
19	15.86	18.96				
20	15.95	14.04				
21	16.68	19.99				
22	11.14					
23	18.78					
24	12.01					
25	16.68					
26	15.41					
27	14.44					
28	16.00					
29	14.94					
30	19.10					
31	15.85					
32	18.41					
33	14.81					
34	14.84					
35	15.99					
36	19.98					
37	17.11					

Test sulla differenza tra medie
(t-test, varianze eguali)

$$t_c = \frac{(\hat{\mu}_1 - \hat{\mu}_2) - (\mu_1 - \mu_2)}{S_A \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

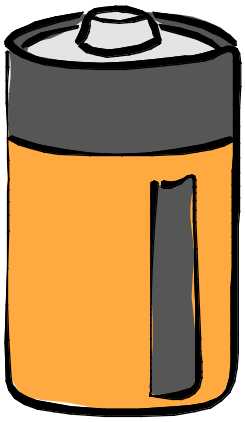
$$H_1 : D < D_0 \Rightarrow p - value = T_{n-1}(t_c)$$
$$H_1 : D > D_0 \Rightarrow p - value = 1 - T_{n-1}(t_c) \qquad n = n_1 + n_2 - 1$$
$$H_1 : D \neq D_0 \Rightarrow p - value = 2\left[1 - T_{n-1}(|t_c|)\right]$$

S_A è lo scarto quadratico medio aggregato

$$S_A^2 = \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}$$

Quando i campioni sono piccoli questo test è abbastanza valido perché non è molto sensibile anche a congrue differenze tra le varianze

Esempio



Una ditta produttrice di batterie vanta una durata media μ_1=75 ore con SQM di 7 ore (valori ricavati da un campione casuale semplice di n_1=20 batterie)

Un campione di n_2=15 batterie di una ditta concorrente ha dato luogo ad una media di μ_2=70 ore con SQM di 5 ore.

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 > \mu_2 \end{cases}$$

$$S_A = \sqrt{\frac{19*7^2 + 14*5^2}{20+15-2}} = \sqrt{\frac{1281}{33}} = 6.23 \qquad t_c = \frac{(75-70)-0}{6.23\sqrt{\frac{1}{20}+\frac{1}{15}}} = \frac{5}{2.13} = 2.35$$

$$p - value = 1 - T_{33}(2.35) = 1.25\%$$

La ditta potrebbe avere qualche ragione a sostenere che le sue batterie durano più a lungo rispetto a quelle dei concorrenti

Esercizio

Tempi medi di completamento di un compito distinti per sesso

	A	B	C	D	E	F	G
1	TEST SULLA DIFFERENZA TRA MEDIE - PICCOLI CAMPIONI - VARIANZE INCOGNITE MA UGUALI						
2							
3	Maschi	Femmine					
4	118.38	118.38		Test t: due campioni assumendo uguale varianza			
5	118.38	105.74					
6	105.74	109.35			*Variabile 1*	*Variabile 2*	
7	109.35	117.19		Media	119.3386	120.3630	
8	117.19	113.26		Varianza	66.8361	65.5155	
9	113.26	126.59		Osservazioni	15	19	
10	126.59	125.65		Varianza complessiva	66.0933		
11	125.65	107.41		Differenza ipotizzata pe	0.0000		
12	107.41	121.63		gdl	32.0000		
13	121.63	124.45		Stat t	-0.3648		
14	124.45	123.42		P(T<=t) una coda	0.3588		
15	123.42	121.52		t critico una coda	1.6939		
16	121.52	119.90		P(T<=t) due code	0.7176		
17	119.90	137.22		t critico due code	2.0369		
18	137.22	119.27					
19		120.52					
20		117.90					
21		135.22					
22		122.27					
23							

Non ci sono evidenze convincenti che il sesso di chi esegue il compito faccia la differenza nei tempi di completamento

Test sulla differenza tra medie
(t-test, varianze diverse)

Indivhiamo con D la differenza ipotizzata tra le due medie (spesso è zero) abbiamo

$$t_c = \frac{(\hat{\mu}_1 - \hat{\mu}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

$$H_1 : D < D_0 \Rightarrow p - value = T_{df}(t_c)$$
$$H_1 : D > D_0 \Rightarrow p - value = 1 - T_{df1}(t_c)$$
$$H_1 : D \neq D_0 \Rightarrow p - value = 2\left[1 - T_{df}(|t_c|)\right]$$

I gradi di libertà sono approssimati dalla formula

$$df = \left[\frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\frac{\left(\frac{\hat{\sigma}_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{\hat{\sigma}_2^2}{n_2}\right)^2}{n_2 - 1}} \right]$$

Esempio

Confronto di due processi produttivi;

1° tipo. Su di un campione di 13 pezzi, ha funzionato in media 202 ore con un SQM di 9 ore.

2° tipo. Su di un campione di 10 pezzi, è durato in media 195 ore con SQM di 6 ore.

Verifichiamo il p-value dell'ipotesi di uguaglianza tra le durate in un test bilaterale

$$Ipotesi\ \ nulla \quad H_0 : \ D = 0$$
$$Ipotesi\ \ alternativa : \quad H_1 \quad D \neq 0$$

$$t_c = \frac{(202 - 195) - 0}{\sqrt{\frac{81}{13} + \frac{36}{10}}} = 2.23$$

P-value: 3.7%

$$df = \left[\frac{\left(\frac{81}{13} + \frac{36}{10}\right)^2}{\frac{\left(\frac{81}{13}\right)^2}{12} + \frac{\left(\frac{36}{10}\right)^2}{9}} \right] = \left[\frac{96.44}{4.675}\right] = [20.629] = 20$$

Si può rifiutare H_0 , ma non sarebbe sbagliato cercare più dati

Esercizio

Confronto dei guadagni iniziali tra Laureati non laureati.

Lo scarto di 300 euro non risulta significativo

	A	B	C	D	E	F
1	TEST T- DUE CAMPIONI ASSUMENDO VARIANZE DIVERSE					
2						
3	Laureati	Non laureati				
4	2123.77	1885.65		Test t: due campioni assumendo varianze diverse		
5	1720.96	1281.44				
6	1924.34	1586.50			*Variabile 1*	*Variabile 2*
7	1670.92	1206.38		Media	1783.5170	1356.0815
8	1559.65	1039.48		Varianza	55948.8161	140122.4605
9	1386.49	779.73		Osservazioni	36	25
10	1598.77	1098.16		Differenza ipotizzata	300	
11	1969.72	1654.58		gdl	37	
12	1616.39	1124.58		Stat t	1.50613	
13	1516.71	975.07		P(T<=t) una coda	0.07026	
14	1758.07	1337.10		t critico una coda	1.68709	
15	1664.51	1196.76		P(T<=t) due code	0.14052	
16	1639.49	1159.24		t critico due code	2.02619	
17	1878.81	1518.21				
18	1798.33	1397.49				
19	2076.27	1814.40				
20	1695.70	1243.55				
21	1777.92	1366.88				
22	2443.37	2365.05				
23	2002.36	1703.54				
24	2155.88	1933.81				
25	1481.22	921.83				
26	1531.60	997.40				
27	1555.76	1033.63				
28	1721.02	1281.54				
29	2011.78					
30	2120.05					
31	1691.47					
32	1815.58					
33	1904.83					
34	1517.31					
35	1927.16					
36	1913.04					
37	1894.05					
38	1744.33					
39	1398.98					