FACOLTA' DI INGEGNERIA

CORSO DI LAUREA IN INGEGNERIA PER L'AMBIENTE ED IL TERRITORIO

CORSO DI STATISTICA E CALCOLO DELLE PROBABILITA'

PROF. PASQUALE VERSACE

SCHEDA DIDATTICA N°5

ARGOMENTO:

VERIFICA DEL MODELLO

A.A. 2008-09

PROBLEMA GENERALE

L'analisi statistica si articola nelle seguenti fasi :

- identificazione della variabile casuale di interesse;
- raccolta di dati campionari;
- scelta del modello probabilistico (ipotesi statistica);
- stima dei parametri del modello;
- verifica del modello (o dell'ipotesi statistica).

Il modello probabilistico in genere è caratterizzato da una distribuzione selezionata tra quelle note, per esempio Normale, Poisson, etc., e da valori dei parametri ottenuti in modo empirico in funzione dei dati campionari.

Per verificare la validità del modello adottato si opera un confronto tra popolazione (ipotizzata) e campione (reale).

Esistono diversi modi di confrontare modello e dati osservati, quelli che verranno considerati nelle pagine che seguono sono:

- Confronto grafico;
- Test di significatività sui parametri (distribuzione nota e distribuzione ricavata mediante metodo Montecarlo);
- Test di bontà dell'adattamento.

CONFRONTO GRAFICO

Una prima valutazione sull'attitudine di una funzione di distribuzione ad interpretare i risultati di una serie di osservazioni si può ottenere attraverso un confronto grafico tra modello e dati. Ad esempio si possono considerare l'istogramma di frequenza derivato dal campione e la funzione densità di probabilità del modello (PDF- Probability Density Function) rappresentata sotto forma di istogramma e discretizzata negli stessi intervalli considerati per le frequenze (Fig. 1).

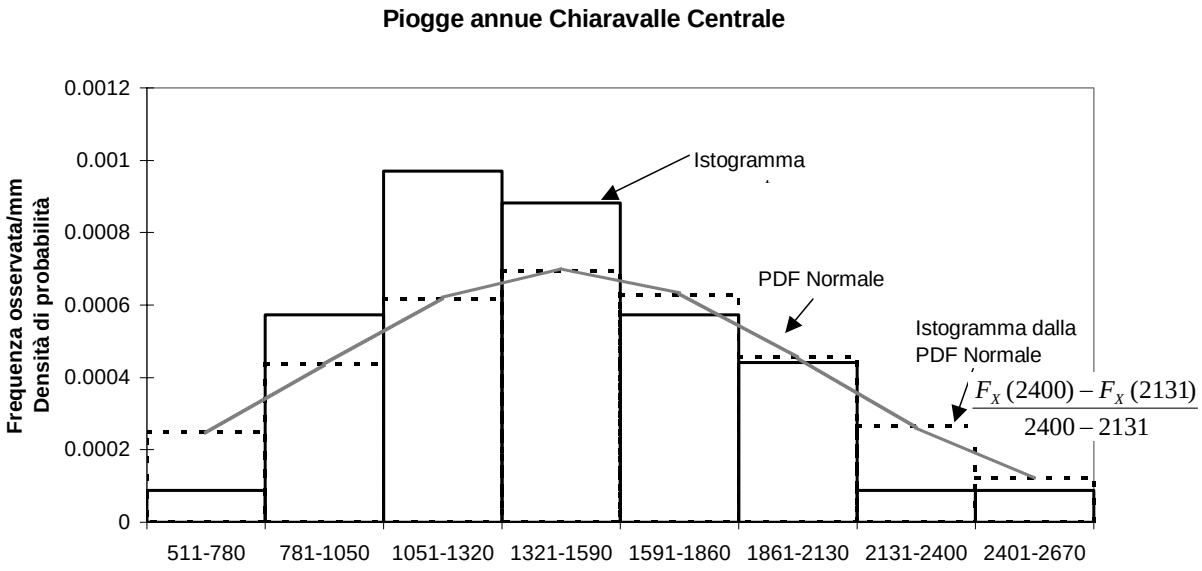


Fig.1- Confronto tra PDF normale ed istogramma dei dati osservati

Utili indicazioni si possono ricavare, anche, riportando su un cartogramma di coordinate *x*, $F_X(x)$ la funzione di probabilità cumulata (CDF- Cumulative Distribution Function) e le frequenze cumulate relative ai dati sperimentali (Fig. 2).

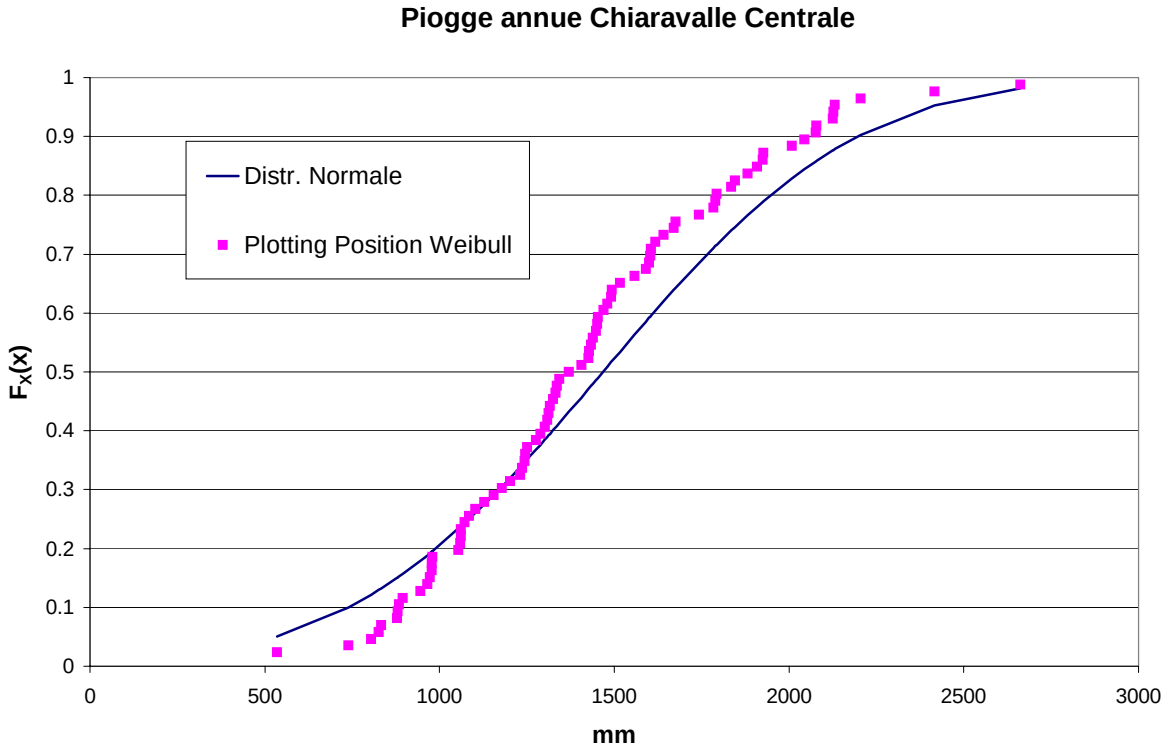


Figura 2- Confronto tra curva di frequenza cumulata teorica e plotting position.

L’informazione che si ottiene è in qualche modo più accurata rispetto al confronto istogrammi-PDF, per il quale è necessario raggruppare i dati osservati ed il modello matematico in intervalli predefiniti, con una perdita di informazioni riguardo l’esatto valore assunto dalle osservazioni.

Con la rappresentazione utilizzata, tuttavia, non si riesce ad avere una chiara visione dell’adattamento tra curva teorica e dati sperimentali in prossimità dei valori estremi (massimi e minimi), ovvero nelle “code” della distribuzione.

E’ più utile eseguire tale confronto servendosi dei cosiddetti *cartogrammi probabilistici,* ovvero diagrammi con scale deformate, a seconda del tipo di distribuzione considerata, in modo tale che la funzione di probabilità cumulata sia rappresentata da una retta.

In generale si riportano in un diagramma cartesiano come ascisse i valori di X, variabile casuale in esame, e come ordinate i valori di Y, variabile ridotta, scelta in modo che:

- ci sia tra X e Y un legame di tipo lineare $Y=a+bX$
- la funzione di probabilità cumulata della Y sia priva di parametri e dipenda quindi solo dai valori y assunti dalla Y. Sia cioè $F_Y(y) = \phi(y)$.

Tenendo conto del fatto che se tra i valori di x e y esiste un legame lineare vale la relazione:

$$F_X(x)=F_Y(y) \qquad (1)$$

si può procedere come segue.

Su un asse parallelo a quello delle y, si riporta in corrispondenza di ciascun valore di y quello assunto da $F_Y(y)$, che sarà ovviamente in una scala non lineare. Si utilizza successivamente come asse delle ordinate l’asse delle $F_Y(y)$, che per la (1) coincide con l’asse delle $F_X(x)$.

In definitiva se su un diagramma cartesiano tralasciando di riportare la y, si riportano come ascisse le x e come ordinate direttamente le $F_X(x)$ nella scala deformata, la funzione di probabilità cumulata risulterà rappresentata come una retta.

Tali diagrammi sono usati per esaminare distribuzioni riconducibili a quella normale (carta probabilistica normale) o quella lognormale (carta probabilistica logaritmico-normale) o alla distribuzione di Gumbel (carta probabilistica doppio esponenziale).

Un esempio applicativo può rendere più chiara la procedura.

Esempio 1: Carta probabilistica Normale

$$F_X(x) = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{x} e^{-\frac{(x-\mu)}{2\sigma^2}} dx$$

$$Y = \frac{X - \mu}{\sigma} \qquad Y = a + bX \qquad \text{ponendo } a = -\frac{\mu}{\sigma} \text{ e } b = \frac{1}{\sigma}$$

$$F_Y(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{y^2}{2}} dy \qquad F_X(x) = F_Y(y)$$

Si procede riportando sulle ordinate i valori delle Y (variabile ridotta) a scala lineare e si determinano i valori assunti da tale variabile in corrispondenza di assegnati valori di $F_Y(y)$:

$F_Y(y)$=0,1; 0,2, 0,3; 0,4; 0,5……. 0,998; 0,999

$F_Y(y)$	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	0,95	0.98
y	-1,282	-0,842	-0,524	-0,253	0,000	0,253	0,524	0,842	1,282	1,645	2,053

$F_Y(y)$	0.99	0.998	0.999
y	2,326	2,878	3,090

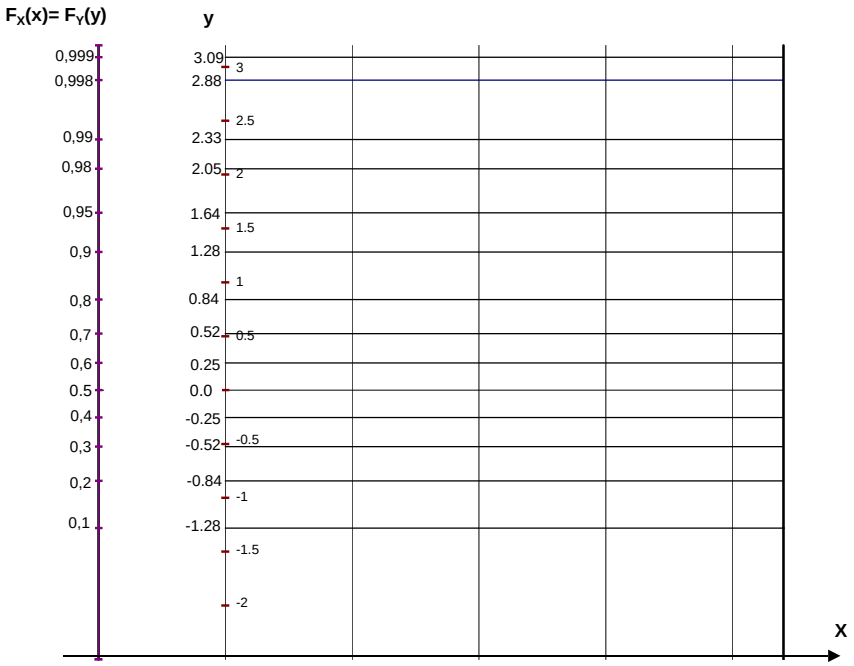


Fig. 3 – Cartogramma probabilistico normale.

Per le piogge annue registrate dalla stazione di Chiaravalle Centrale si ottiene, ad esempio, il diagramma di seguito riportato.

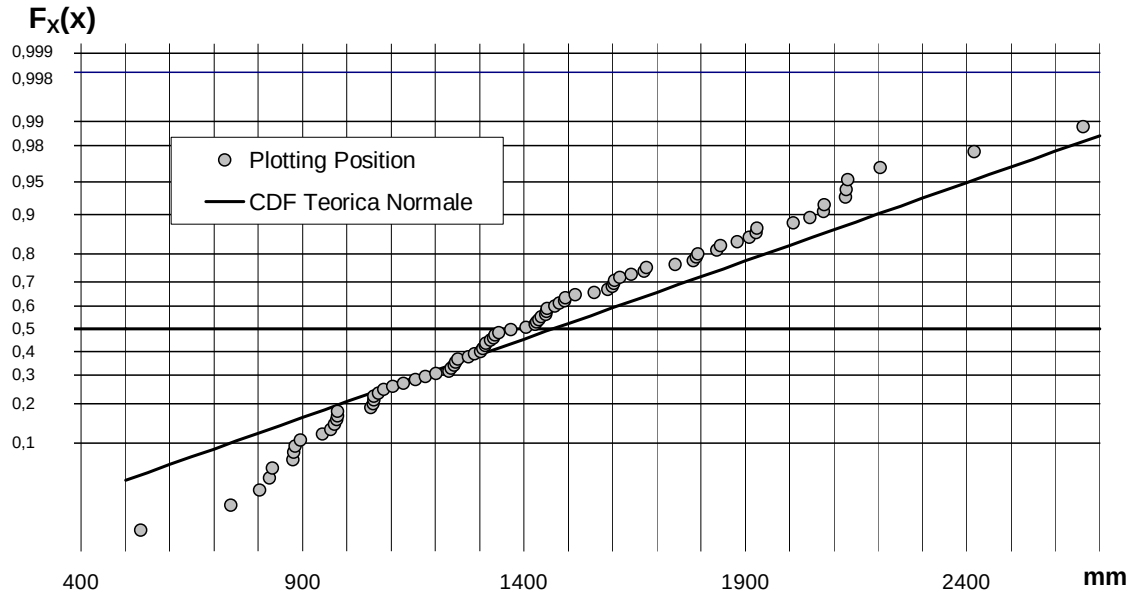


Fig. 4 – Confronto su carta probabilistica normale di CDF teorica e Plotting Position di Weibull.

Esempio 2 : Carta probabilistica DoppioEsponenziale (Gumbel)

$F_X(x) = e^{-e^{-\alpha(x-\varepsilon)}}$

$Y = \alpha(X - \varepsilon)$

$F_Y(y) = e^{-e^{-y}}$ $\qquad\qquad$ $F_X(x) = F_Y(y)$

In maniera analoga alla carta probabilistica normale si procede riportando sulle ordinate i valori delle Y (variabile ridotta) a scala lineare e si determinano i valori assunti da tale variabile in corrispondenza di assegnati valori di $F_Y(y)$:

$F_Y(y)$=0,1; 0,2, 0,3; 0,4; 0,5……. 0,998; 0,999

$F_Y(y)$	0,1	0,2	0,3	0,367	0,4	0,5	0,6	0,7	0,8	0,9
$y=-\ln(-\ln(F_Y(y))$	-0,834	-0,476	-0,186	0,000	0,087	0,367	0,672	1,031	1,500	2,250

$F_Y(y)$	0,95	0.98	0.99	0.998	0.999
$y=-\ln(-\ln(F_Y(y))$	2,970	3,902	4,600	6,214	6,907

Rappresentati gli elementi di un campione su una carta probabilistica, se il tipo di distribuzione ipotizzato è adatto a rappresentare le osservazioni, queste devono disporsi più o meno intorno ad una retta. Si tratta comunque di confronti basati su valutazioni soggettive e, quindi, non molto affidabili.

TEORIA CAMPIONARIA E VERIFICA DEL MODELLO

Un metodo più efficace di verifica del modello è il test di significatività che è tipico della teoria campionaria, cioè della sezione del calcolo delle probabilità che consente di valutare la probabilità che un particolare campione possa provenire da una determinata popolazione.

La verifica del modello è, in realtà, dal punto di vista concettuale un problema diverso. Si deve, infatti, decidere se l'ipotesi statistica fatta circa la popolazione di provenienza del campione sia o meno giusta. Si assume cioè vero il campione e si verifica se è attendibile l'ipotesi circa la provenienza del campione. Nella teoria campionaria, invece, la popolazione è nota e si vuole controllare se il campione proviene o meno da quella popolazione.

Anche se la differenza concettuale è evidente, dal punto di vista operativo la procedura nei due casi è identica: si formula un'ipotesi H_0, detta *ipotesi nulla*, sulla v.c. X e, attraverso i risultati campionari, si decide se accettare o rifiutare H_0.

Per *ipotesi statistica* si intende un'affermazione sulla distribuzione di probabilità di una v.c. e può tanto riguardare il modello adottato che il valore stimato per i parametri. Ad esempio è possibile sia verificare l'adattamento di una distribuzione normale, sia verificare l'ipotesi che la popolazione abbia una data media e/o varianza.

In molti casi, si formula un'ipotesi statistica per il solo scopo di rigettarla. Per esempio, se si vuole determinare se una certa moneta sia truccata, si formula l'ipotesi che la moneta sia "buona" e cioè che $p = 0.5$, dove p è la probabilità che si presenti testa. Se, basandoci sulla supposizione che una certa ipotesi sia vera, troviamo che il risultato osservato su un campione casuale differisce notevolmente da quello atteso, dovremmo dire che la differenza osservata è significativa e quindi rifiutare l'ipotesi. Per esempio se in 20 lanci di una moneta viene testa 19 volte, dovremmo rigettare l'ipotesi che la moneta sia buona, anche se è possibile che lanciando 20 volte una moneta buona esca 19 volte testa. Ma la probabilità di un tale evento è così piccola che appare preferibile ritenere che la moneta sia truccata.

E' necessario sottolineare che in un test di questo tipo la decisione sulla validità di H_0 viene adottata in base al campione specifico di cui si dispone e da cui si deducono le informazioni

riguardanti l'intera popolazione. Nella pratica, inoltre, spesso la decisione di accettare o rifiutare un'ipotesi viene dedotta da un indice che sintetizza le informazioni del campione.

Nota la popolazione *P*, e disponendo di uno (o più) campioni *C*, si formula l'**ipotesi nulla H_0**: ad esempio ***C* proviene da *P***.
Si possono avere quattro possibili situazioni riassunte nella tab. 1

REALTÀ	**RISPOSTA**	
SI	SI	OK
SI	NO	Errore del I° tipo
NO	SI	Errore del II° tipo
NO	NO	OK

Tabella. 1

Rifiutando l'ipotesi H_0 quando questa è vera si commette un errore del I° tipo, mentre accettando H_0 quando è falsa si commette un errore del II° tipo.
Si definisce *livello di significatività* la probabilità α di commettere un errore del I° tipo; sia β, invece, la probabilità di commettere un errore del II° tipo.
Affinché un test delle ipotesi sia efficace deve essere configurato in modo da minimizzare la probabilità di commettere errori di decisione. In genere, per una data ampiezza del campione ogni tentativo di diminuire un tipo di errore è accompagnato da un incremento dell'altro; il solo modo di ridurre sia α che β è di aumentare la numerosità del campione.

Nelle applicazioni si decide in primo luogo su quale grandezza effettuare il test (media, varianza, coefficiente di asimmetria, quantili, etc.).
Si fissa poi il livello di significatività da attribuire la test. In genere si pone: $\alpha = 0.05$ (5%), $\alpha = 0.01$ (1%), $\alpha = 0.001$ (1‰).
Successivamente, lo spazio parametrico dell'indice caratterizzante la decisione da prendere, ossia l'insieme dei possibili valori che esso può assumere, si ripartisce in una regione di accettazione ed una di rifiuto sulla base del livello di significatività adottato (ad esempio fig.5).
In questo modo α rappresenta la probabilità che il campione non ricada nella regione di accettazione quando, invece, H_0 è vera.

Il valore dell'indice dedotto sulla base del campione disponibile, infine, viene confrontato con un valore critico che definisce la regione di accettazione. Se dal confronto il valore calcolato non ricade in tale regione, l'ipotesi H_0 dovrà essere rifiutata.
Per identificare la regione di accettazione è necessario conoscere la distribuzione di probabilità della *statistica* utilizzata per il test. A seconda dei casi tale distribuzione è nota oppure deve essere ricostruita con tecniche di simulazione.

Test su *statistiche* con distribuzione nota

Per campioni di grandi dimensioni, la distribuzione campionaria di alcune statistiche è una distribuzione normale. Si consideri una v.c. X Normale con media μ e scarto quadratico σ noti, e di tale variabile si abbia un campione casuale (x_1, x_2,…, x_N) di N dati con media campionaria pari a $\bar{x}$ e varianza s^2.
$\overline{X}$ è anch'essa una variabile casuale che dipende dal campione estratto ed è distribuita secondo la legge normale con parametri:

$$E[\overline{X}]=\mu \qquad\qquad Var[\overline{X}]=\frac{\sigma^2}{N}$$

La media si distribuisce in maniera normale anche quando, qualunque sia la distribuzione della popolazione, il campione estratto abbia numerosità N>30.

Per ciò che riguarda la distribuzione di s^2 si può verificare che $\frac{(N-1)s^2}{\sigma^2}$ è una variabile casuale Chi-quadrato con (N-1) gradi di libertà.

Esempio

Si consideri una popolazione caratterizzata da:

$$\mu = 100 \qquad\qquad \sigma = 20$$

ed un campione di $N = 30$ dati e media campionaria

$$\bar{x} = \sum_{i=1}^{N} \frac{x_i}{N} = 90$$

Sotto l'ipotesi fondamentale che la popolazione di partenza sia distribuita in maniera Normale, la distribuzione dello stimatore $\overline{X}$ della media μ è ancora Normale.

$$\overline{X} \quad \Rightarrow \quad N(\mu, \frac{\sigma}{\sqrt{N}}) = N(100, \frac{20}{\sqrt{30}})$$

Ipotesi nulla H_0: il campione con media 90 e dimensione campionaria 30 proviene dalla popolazione con media 100 e scarto quadratico medio 20.

Fissiamo un livello di significatività pari al 5%.

Si consideri la variabile standardizzata $U=\frac{\overline{X}-\mu}{(\sigma/\sqrt{N})}$ per la quale si ha la PDF riportata in figura 5.

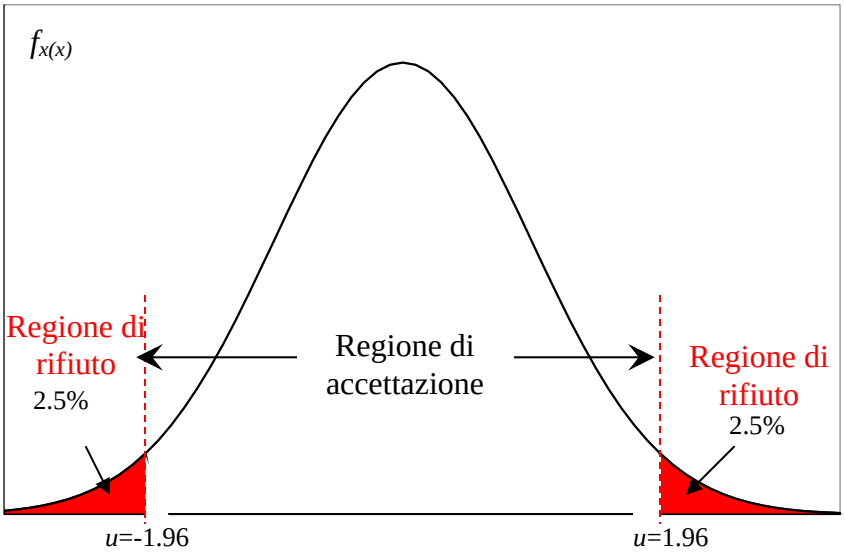


Fig. 5 – Curva normale standardizzata con evidenziazione della regione di accettazione e della regione di rifiuto.

Come indicato nella figura possiamo essere fiduciosi al 95% che se l'ipotesi è vera il valore u della media campionaria **standardizzata** dovrà cadere tra -1.96 e 1.96 (infatti l'area della curva normale standardizzata compresa tra questi due valori è pari a 0.95).

$u=\frac{\overline{x}-\mu}{(\sigma/\sqrt{N})}$=-1.96 $\qquad \overline{x}=\mu - 1.96\frac{\sigma}{\sqrt{N}}$ =100- 1.96·3.65=92.85

$u=\frac{\overline{x}-\mu}{(\sigma/\sqrt{N})}$=1.96 $\qquad \overline{x}=\mu + 1.96\frac{\sigma}{\sqrt{N}}$ =100+1.96·3.65=107.15

Si possono avere, quindi, i seguenti casi:

$\overline{x} < 92.85$ respingo H_0

$92.85 < \overline{x} < 107.15$ accetto H_0

$\overline{x} > 107.15$ respingo H_0

Nel caso in esame l'ipotesi viene respinta. La giustificazione intuitiva per il rifiuto è che la media campionaria è significativamente differente dal valore di μ, e quindi l'ipotesi formulata è verosimilmente sbagliata. Bisogna sottolineare, però, che l'ipotesi fatta potrebbe essere giusta qualora l'osservazione disponibile fosse, invece, un evento "raro". □

Nel test precedente abbiamo posto l'attenzione sui valori estremi della statistica considerata o del corrispondente valore *u* standardizzato rispetto alla media e allo scarto quadratico, cioè abbiamo posto l'attenzione su entrambe le code della distribuzione. Per tale ragione questi tipi di test sono detti *test a due code*. Spesso tuttavia possiamo essere interessati ai valori estremi di un solo lato, cioè poniamo l'attenzione su una sola coda della distribuzione. Tali test sono detti *test ad una coda* e la regione critica in questi casi è posta da un solo lato della distribuzione, con area pari al livello di significatività prescelto (fig.6).

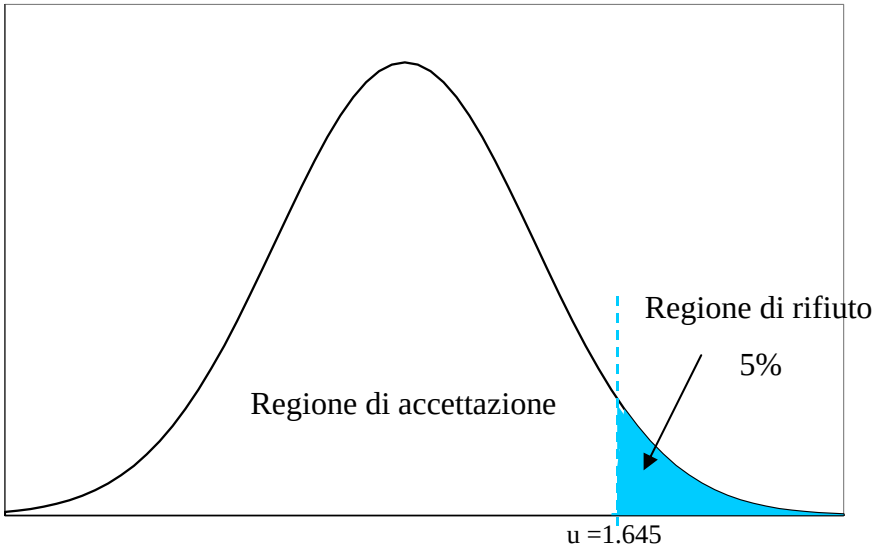


Fig. 6 – Curva normale standardizzata con evidenziazione della regione di rifiuto nel caso di test ad una coda.

La tab.2 riassume i valori critici di *u* per i test ad una coda e per i test a due code, relativi a diversi livelli di significatività.

Livello di significatività α	0.1	0.05	0.01	0.005
Valori critici di *u* per test a una coda	-1.28 o 1.28	-1.645 o 1.645	-2.33 o 2.33	-2.88 o 2.88
Valori critici di *u* per test a due code	-1.645 e 1.645	-1.96 e 1.96	-2.58 e 2.58	-3.08 e 3.08

Tabella. 2

Test su *statistiche* di cui non si conosce la distribuzione

Lo stesso tipo di analisi condotta nell'esempio precedente sulla media può essere effettuata su:

1. varianza σ^2
2. coefficiente di asimmetria γ_1
3. frattile X_F

Per poter applicare il test è però necessario conoscere la distribuzione di probabilità della variabile considerata. In alcuni casi, si è detto, è nota la distribuzione campionaria in particolari ipotesi (distribuzione normale, parametri noti, etc.), negli altri casi è necessario fare ricorso a tecniche di tipo Montecarlo che costituiscono un metodo approssimato per la soluzione di problemi di derivazione di distribuzioni di variabili casuali.

Con il metodo Montecarlo si approssima la distribuzione di probabilità che si intende cercare attraverso i risultati di molti esperimenti ripetuti artificialmente.

Si consideri l'esempio della media campionaria, di cui in questo caso si ipotizza di non conoscere quale sia la distribuzione e si determini in modo approssimato la funzione densità di probabilità della variabile $\overline{X}$ attraverso una serie di esperimenti. Ciascun esperimento consisterà nella creazione di un certo numero di dati campionari sintetici di cui è possibile determinare il valore medio. Un numero molto elevato di tali esperimenti permette di tracciare un istogramma che approssima molto bene la forma della funzione densità di probabilità $f_X(\bar{x})$.

La costruzione di dati sperimentali sintetici avviene in due fasi:

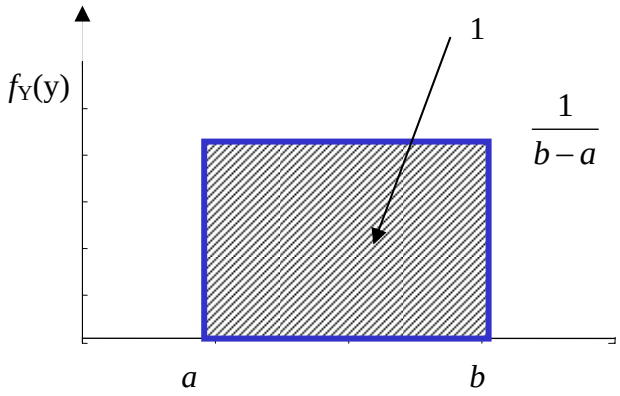
- si genera una serie di numeri casuali in maniera tale che ogni valore sia ugualmente probabile,
- si associa a ciascun numero casuale generato un particolare valore della variabile di interesse.

Per capire meglio la procedura è necessario introdurre un nuovo modello probabilistico: la distribuzione uniforme.

Distribuzione uniforme

Una variabile casuale Y distribuita con legge uniforme è caratterizzata dalla seguente funzione di densità e funzione di probabilità cumulata:

$$f_Y(y) = \begin{cases} \dfrac{1}{b-a} & a \le y \le b \\ 0 & \text{altrimenti} \end{cases} \qquad F_Y(y) = \begin{cases} 0 & y < a \\ \dfrac{y-a}{b-a} & a \le y \le b \\ 1 & y > b \end{cases}$$



Considerando $a = 0$ e $b = 1$ si ha un particolare caso di distribuzione uniforme per cui la frequenza cumulata è così definita.

$$F_Y(y) = \begin{cases} 0 & y < 0 \\ y & 0 \le y \le 1 \\ 1 & y > 1 \end{cases}$$

La media e la varianza di una variabile casuale distribuita uniformemente tra 0 e 1 sono[*]:

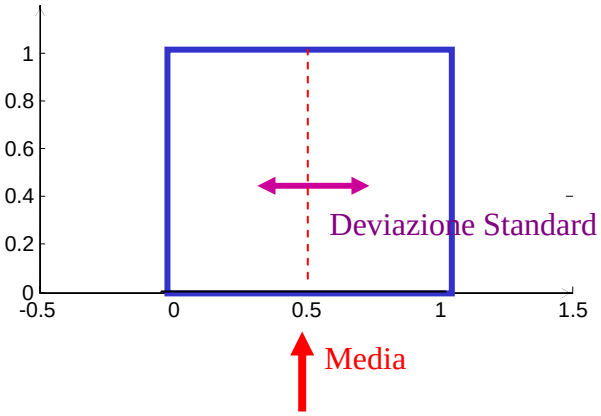
$$\bar{y} = \frac{1}{2} \qquad \sigma_y^2 = \frac{1}{12}$$

---

[*]

$$\bar{y} = \int_{-\infty}^{\infty} y\, p_y(y)dy = \int_0^1 y\,(1)\,dy = \frac{1}{2}$$

$$\sigma_y^2 = \int_{-\infty}^{\infty} y^2\, p_y(y)dy - \bar{y}^2 = \int_0^1 y^2\,(1)\,dy - \bar{y}^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$



Si genera un numero casuale *y* compreso tra 0 ed 1 e si assume che esso provenga da una distribuzione uniforme tra 0 ed 1 e, quindi, che la probabilità $F_Y(y)$ abbia lo stesso valore di y (fig. 7). In questo modo si genera un valore casuale della funzione di probabilità cumulata.

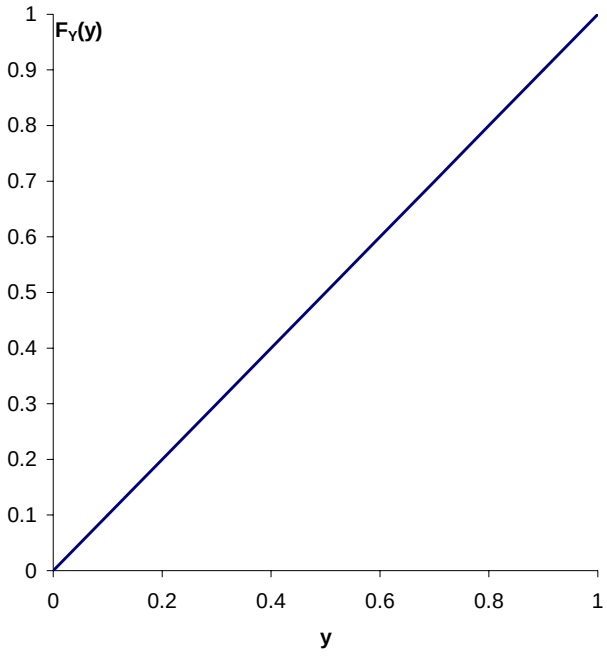


Fig. 7 – Andamento di $F_Y(y)$ al variare di $y \in [0,1]$ per la distribuzione uniforme nel caso a=0 e b=1

Un modo banale di generare numeri casuali equiprobabili compresi tra 0 e 1 è quello di predisporre 10000 bigliettini su ognuno dei quali si riportano valori che differiscono di 1/10000:

0.0001

0.0002

0.0003

…

0.9998

0.9999

1.0000

Estraendo a caso *k* bigliettini si ottengono *k* valori casuali $y \in [0,1]$ e, quindi, altrettanti valori $F_Y(y) \in [0,1]$.

Tale procedura, ovviamente, viene simulata su calcolatori mediante l'applicazione di algoritmi matematici e statistici che generano numeri casuali provenienti da una distribuzione uniforme con limiti 0 e 1. Un esempio di tali generatori è la funzione matematica "**casuale**" disponibile su EXCEL.

Esempio. Coefficiente di asimmetria campionario. Distribuzione Gumbel.

Consideriamo una popolazione distribuita secondo la legge probabilistica di Gumbel di cui sono noti i parametri α ed ε.

$$F_X(x) = e^{-e^{-\alpha(x-\varepsilon)}} \qquad (2)$$

In questo caso non è nota la distribuzione del coefficiente di asimmetria campionario g_1 e si determina, quindi, la funzione densità di probabilità di tale variabile attraverso una serie di esperimenti. In particolare:

- si generano *N* valori di $F_Y(y)$ come descritto nel paragrafo precedente,
- si pone $F_Y(y)=F_X(x)$
- si associa ad ogni valore di $F_X(x)$ la *x* corrispondente attraverso il legame definito dalla (2), ottenendo così un campione di *N* valori di una variabile casuale distribuita con legge di Gumbel.
- si calcola il coefficiente di asimmetria campionario g_1
- si ripete l'operazione *k* volte (*k* molto grande 10.000÷100.000) ottenendo *k* valori di g_1 (in pratica si ricostruisce la distribuzione di probabilità di g_1).

Ordinando i valori di g_1 in maniera crescente è possibile identificare $g_{1\,0.025}$, $g_{1\,0.975}$ (rispettivamente il 250° ed il 9750° valore su 10.000), che diventano i limiti della regione di accettazione per un eventuale test a due code con livello di significatività al 5%.□

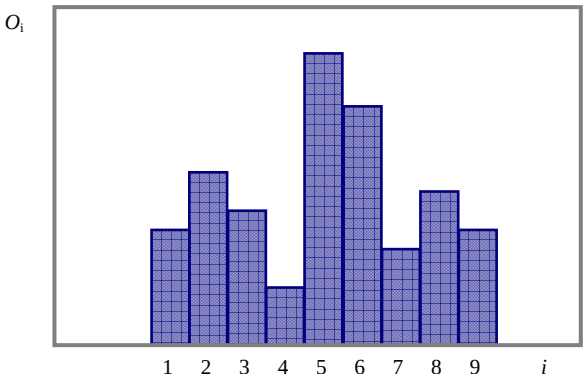
TEST SULL'ADATTAMENTO DI UNA DISTRIBUZIONE

In questo tipo di test, invece di considerare l'andamento della media o di un'altra statistica che descrive in modo sintetico il campione, si opera sull'intero campione.

Il test χ^2 (chi-quadrato) rappresenta uno dei test statistici più utilizzati ed è usato per verificare l'ipotesi H_0 che un certo campione di dati provenga da una specifica distribuzione nota (nella forma e nei parametri).

In pratica si procede nel seguente modo:

- si suddivide lo spazio campionario in *k* intervalli;
- si calcola O_i, frequenza assoluta osservata per il generico intervallo *i*-esimo, ossia il numero di osservazioni che ricadono in tale intervallo.



- sulla base del modello considerato nell'ipotesi H_0, si calcolano le probabilità p_i relative a ciascuna delle classi considerate e le frequenze teoriche attese, $E_i = Np_i$, pari al valor medio del numero di eventi che dovrebbero rientrare nella classe i-esima.

Nel dividere lo spazio campionario in *k* classi, si deve rispettare la condizione che Np_i risulti maggiore uguale di 5.

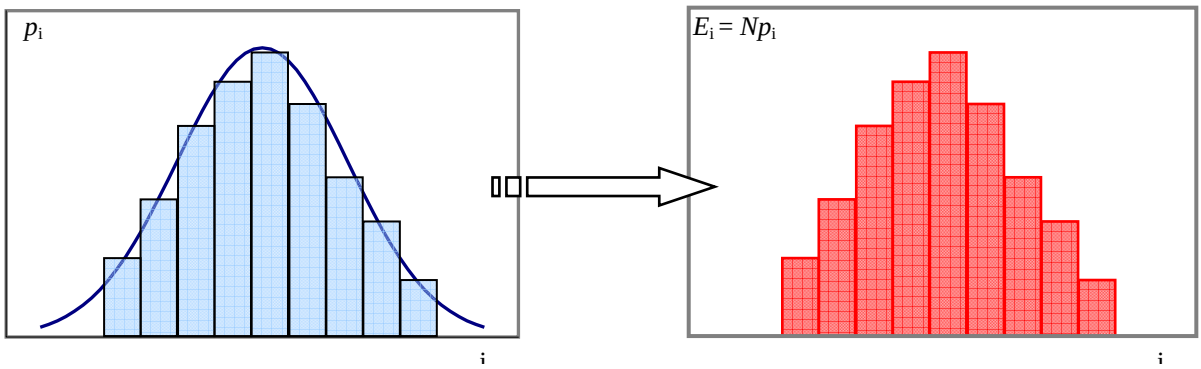
La frequenza attesa si calcola come:

$$E_i = Np_i$$

con

$$p_i = F(y_u) - F(y_l)$$

in cui *F* è la funzione di distribuzione cumulata del modello statistico da testare, y_u è il limite superiore per la classe *i*, e y_l è il suo limite inferiore.



Nell'esecuzione del test, in molti casi, si usa l'accortezza di operare una ripartizione in classi basata su un criterio di equiprobabilità in modo tale che risulti $p_i = 1/k$ uguale per tutte le classi e che *k* sia pari al massimo numero intero che soddisfa la relazione $\frac{N}{k} \geq 5$.

In tale caso i limiti superiore ed inferiore di ciascuna classe, quindi, vengono fissati come i frattili corrispondenti alle diverse frequenze cumulate.

Il confronto tra frequenze osservate e frequenze teoriche attese, definisce una grandezza, valida per il campione in esame,:

$$D = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \chi^2_{clc.}$$

D è distribuita secondo il modello probabilistico $\chi^2(\theta)$:

$$f_X(x) = \frac{1}{2^{\theta/2}\Gamma(\theta/2)} e^{-\frac{x}{2}} x^{\frac{\theta}{2}-1}$$

in cui:

θ = gradi di libertà = k - r - 1

r = numero di parametri della distribuzione indagata.

Per un fissato α livello di significatività si può individuare da apposite tabelle il valore limite $\chi^2_{(\alpha,\theta)}$ tale che:

$P[\chi^2 > \chi^2_{(\alpha,\theta)}] = \alpha$

Il criterio di accettazione dell'ipotesi statistica è del tipo

$\chi^2_{calc} > \chi^2_{(\alpha,\theta)}$ rifiuto H_0

$\chi^2_{calc} \leq \chi^2_{(\alpha,\theta)}$ accetto H_0

TABELLA 3 - VALORI CRITICI PER LA DISTRIBUZIONE DEL χ^2

θ	α = 0,05	α = 0,01	θ	α = 0,05	α = 0,01
1	3,841	6,635	**22**	33,924	40,289
2	5,991	9,210	**23**	35,172	41,638
3	7,815	11,345	**24**	36,415	42,980
4	9,488	13,277	**25**	37,652	44,314
5	11,070	15,086	**26**	38,885	45,642
6	12,592	16,812	**27**	40,113	46,963
7	14,067	18,475	**28**	41,337	48,278
8	15,507	20,090	**29**	42,557	49,588
9	16,916	21,666	**30**	43,773	50,892
10	18,307	23,209	**40**	55,758	63,691
11	19,675	24,725	**45**	61,656	69,957
12	21,026	26,217	**50**	67,505	76,154
13	22,362	27,688	**55**	73,311	82,292
14	23,685	29,141	**60**	79,082	88,379
15	24,996	30,578	**65**	84,821	94,422
16	26,296	32,000	**70**	90,531	100,430
17	27,587	33,409	**75**	96,217	106,390
18	28,869	34,805	**80**	101,880	112,330
19	30,144	36,191	**85**	107,520	118,240
20	31,410	37,566	**90**	113,150	124,120
21	32,671	38,932	**95**	118,750	129,970

Esempio 1

In 200 lanci di una moneta sono state osservate 115 teste e 85 croci. Provare l'ipotesi che la moneta è buona ($p_T = p_C = 0.5$) usando un livello di significatività dello 0.05 e dello 0.01.

Le frequenze osservate di teste e croci sono rispettivamente $O_T = 115$ e $O_C = 85$. Le frequenze attese di teste e di croci sono rispettivamente $E_T = Np_T = 100$ e $E_C = Np_C = 100$ se la moneta è buona. Allora

$$\chi^2_{calc} = \frac{(115-100)^2}{100} + \frac{(85-100)^2}{100} = 4.5$$

Poiché k=2 ed r=0, si ottiene θ=1. Il valore critico $\chi^2_{(0.05,1)}$ è 3.84. Poiché 4.5 > 3.84 rifiutiamo l'ipotesi che la moneta sia buona al livello di significatività dello 0.05. Il valore critico

$\chi^2_{(0.01,1)}$ è 6.63. Poiché 4.5 < 6.63 non possiamo rifiutare l'ipotesi che la moneta sia buona al livello di significatività dello 0.01. □

Esempio 2

La tabella 4 mostra le frequenze osservate e attese in 120 lanci di un dado. Provare l'ipotesi che il dado sia "equilibrato" usando un livello di significatività dello 0.05.

La condizione da verificare si traduce nella ipotesi statistica H_0: p_i =1/6.

Faccia						
Frequenze osservate	25	17	15	23	24	16
Frequenze attese	20	20	20	20	20	20

Tabella 4

Le frequenze attese sono state calcolate come:

$Np_i = 120 \cdot 1/6 = 20$

ed il valore del χ^2 calcolato dal campione a disposizione risulta pertanto:

$$\chi^2_{calc} = \frac{(25-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(15-20)^2}{20} + \frac{(23-20)^2}{20} + \frac{(24-20)^2}{20} + \frac{(16-20)^2}{20} = 5$$

Poiché il numero di classi è k = 6 ed r=0 si ha:

θ = gradi di libertà = k - r – 1 = 6 – 1 = 5

Il valore critico per 5 gradi di libertà e α = 0.05 è 11.1. Allora, poiché 5 < 11.1 non possiamo rifiutare l'ipotesi che il dado sia buono per il livello di significatività fissato. □

Esempio 3

Data la serie di valori, riportati in tabella 5, cui è stato adattato il modello probabilistico Normale, effettuare il test del χ^2 (livello di significatività pari a 0.05).

Tabella 5

451	813	411	807	797	696
377	615	650	524	655	658
761	421	618	916	1067	358
760	754	533	800	367	246
246	734	386	710	569	612

La media μ e la deviazione standard σ, stimate con il metodo dei momenti, sono pari rispettivamente a 610.4 e 200.51.

Per stabilire il numero di classi in cui suddividere il campione si è considerato il criterio dell'equiprobabilità. Partendo dalla condizione $Np_i \geq 5$, è stata fissata una frequenza teorica di occorrenza in ogni classe pari a $Np_i = 5$. Da cui si ricava:

$$p_i = \frac{5}{N} = \frac{5}{30} = 0.1666 .$$

Il numero k di classi da considerare affinché la probabilità p_i che un dato ricada in ciascuna classe risulti pari 0.1666 deve essere, quindi, pari a $k = (1/0.1666) = 6$.

Restano ancora da stabilire gli estremi di variazione di ciascuna classe. Tali valori sono stabiliti considerando i frattili relativi alle frequenze cumulate teoriche $F_x(x)$=0.1666, $F_x(x)$=0.3333, etc. definite dalla probabilità di occorrenza in ciascuna classe p_i. In particolare, per il caso in esame, nel quale è stata considerata la distribuzione Normale, tali valori possono essere desunti mediante l'applicazione della funzione "INV. NORM." di EXCEL o tramite la tabella relativa alle aree sottese dalla curva normale standardizzata.

Ad esempio, con l'ausilio della tabella (si lasciano allo studente i dettagli dell'applicazione), si può verificare che la v.c. normale standardizzata corrispondente alla probabilità di non superamento $F_x(x)$=0.1666 è circa pari a -0.968. Per risalire, quindi, al frattile della v.c. X corrispondente alla distribuzione teorica basta calcolare:

$$x = z \cdot \sigma + \mu = -0.968 \cdot 200.51 + 610.4 = 416.3$$

In maniera del tutto analoga si procede per la determinazione degli estremi corrispondenti alle probabilità di non superamento teoriche successive.

Infine, si valutano le frequenze osservate O_i, considerando il numero di dati del campione che ricadono in ciascuna classe.

I risultati ottenuti sono sintetizzati nella tabella riassuntiva che segue.

k	p_i	$F_x(x)$	x	O_i	$E_i=Np_i$	O_i-E_i	$\frac{(O_i - E_i)^2}{E_i}$
		0	0				
1	0.1666			7	5	2	0.8
		0.1666	416.4				
2	0.1666			3	5	-2	0.8
		0.3333	524.0				
3	0.1666			2	5	-3	1.8
		0.5	610.4				
4	0.1666			7	5	2	0.8
		0.6666	696.7				
5	0.1666			7	5	2	0.8
		0.8333	804.3				
6	0.1666			4	5	–1	0.2
		1	∞				
							$\chi^2_{calc} = 5.2$

I gradi di libertà, considerato che la funzione in esame è caratterizzata da due parametri, sono:

$\theta = k - r - 1 = 6 - 2 - 1 = 3$

Per il livello di significatività fissato e per il numero di gradi di libertà si ha dalla tabella 2

$$\chi^2 = 7.815$$

Poiché risulta $\chi^2 > \chi^2_{calc}$ il test risulta soddisfatto.□