RICHIAMI DI STATISTICA DESCRITTIVA E DI INFERENZA:

LA VERIFICA DI IPOTESI: TEST BASATI SU UN CAMPIONE

La verifica di ipotesi

Finora abbiamo visto come si utilizza una statistica campionaria per stimare un parametro della popolazione

Un'altra fase dell'inferenza è quella che consente di verificare delle ipotesi sui parametri

Obiettivo primario: capire e ridurre l'incertezza per prendere decisioni
Obiettivo secondario: controllare il rischio del prendere decisioni sulla base delle statistiche campionarie

Esempio: in una azienda che produce scatole metalliche si intende valutare se il processo produttivo opera in modo tale da garantire che la lunghezza del lato maggiore sia pari a 368 mm. Viene estratto un campione di 25 scatole. Se la lunghezza delle scatole risultasse diversa sarebbe necessario un intervento correttivo, altrimenti no.

La verifica di ipotesi

Verifica di ipotesi: metodologia per fare inferenza sui parametri della popolazione alla luce dell'analisi delle differenze tra i risultati osservati (statistica campionaria) e quelli che ci aspetteremmo se una qualche ipotesi sulla popolazione fosse vera.

La verifica di ipotesi ha inizio con la formulazione del sistema di ipotesi sottoposto a verifica.

Il sistema di ipotesi fa sempre riferimento a qualche parametro della popolazione. Consiste sempre in due ipotesi contrapposte.

Riquadro 6.1 L'ipotesi nulla e l'ipotesi alternativa

✓ **1.** L'ipotesi nulla H_0 è l'ipotesi sottoposta a verifica.

✓ **2.** L'ipotesi alternativa H_1 è specificata come ipotesi opposta all'ipotesi nulla e rappresenta la conclusione raggiunta quando l'ipotesi nulla è rifiutata.

(continua)

La verifica di ipotesi

Nell'approccio classico alla verifica di ipotesi, si individuano i seguenti elementi chiave:

1. **L'ipotesi nulla H_0 si riferisce sempre a un valore specifico del *parametro della popolazione* (ad esempio μ), e non a una *statistica campionaria* (ad esempio *la media campionaria).***
2. **L'ipotesi nulla contiene *sempre* un segno di eguale relativo al valore specificato del parametro della popolazione (ad esempio H_0: μ=368 mm).**
3. **L'ipotesi alternativa non contiene *mai* un segno di eguale relativo al valore specificato del parametro della popolazione (ad esempio H_1: μ?368 mm).**

La verifica di ipotesi

- **Se la statistica campionaria prescelta si avvicina al valore ipotizzato nell'ipotesi nulla accettiamo H_0, altrimenti rifiutiamo H_0 a favore dell'ipotesi alternativa H_1.**
- **La teoria della verifica di ipotesi fornisce una regola su cui basare il processo decisionale.**

- **Questo risultato viene ricavato determinando prima la distribuzione campionaria della statistica di interesse (statistica test) e quindi calcolando il valore assunto per il particolare campione considerato.**

- **La distribuzione campionaria della statistica test spesso è una distribuzione statistica nota, come la Normale o la t di Student, e quindi possiamo ricorrere a queste distribuzioni per sottoporre a verifica un'ipotesi nulla.**

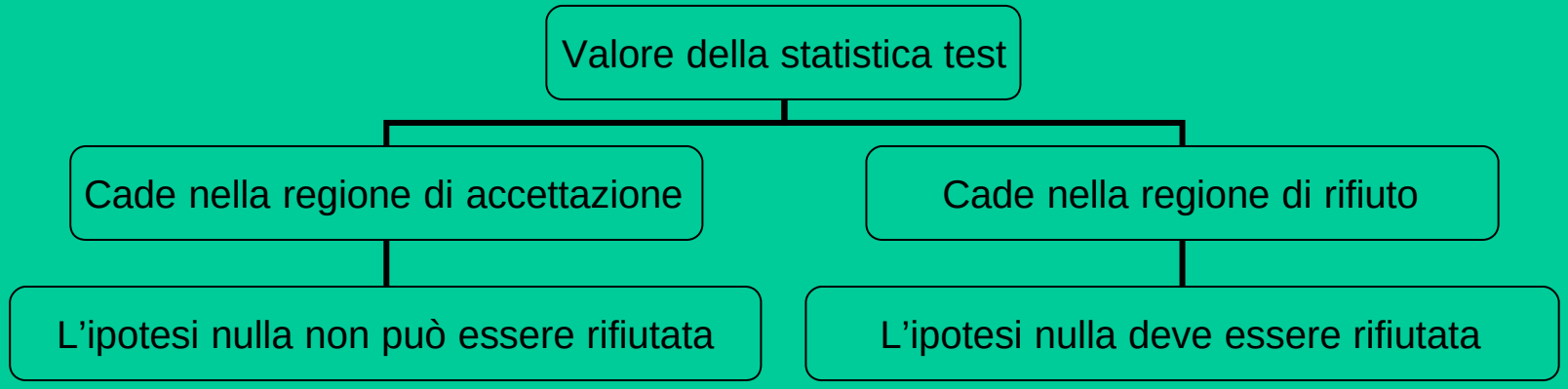
La verifica di ipotesi

La distribuzione campionaria della statistica test è divisa in due regioni:

- una regione di rifiuto (o regione critica)
- una regione di accettazione

<u>Regione di rifiuto</u>: insieme dei valori della statistica test che non è probabile si verifichino quando è vera H_0 ed è probabile si verifichino quando H_0 è falsa.

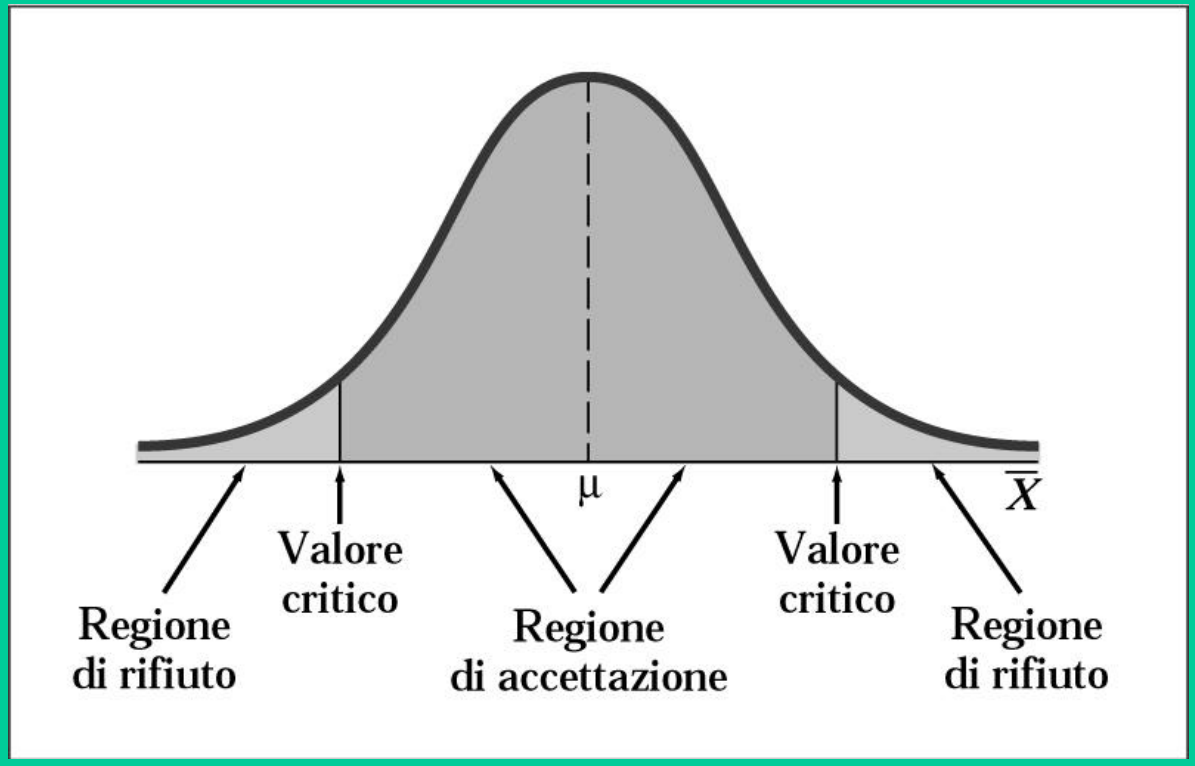
La regola decisionale è:



La verifica di ipotesi

Per prendere una decisione sull'ipotesi nulla, dobbiamo determinare il valore critico della statistica test.

Tale valore separa la regione di accettazione dalla regione di rifiuto.



Test Z per la media (σ noto)

Per verificare l'ipotesi che la media della popolazione sia uguale ad un certo valore μ, contro l'ipotesi alternativa che la media differisca da tale valore, conoscendo σ, si ricorre alla statistica Z:

La statistica test Z per la media (σ noto)

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad (6.1)$$

X è distribuita come una normale => sotto H_0 Z è distribuita come una normale standardizzata

Se Z assume valori vicini allo zero siamo portati ad accettare H_0, altrimenti si propende per rifiutare H_0 (test a due code).

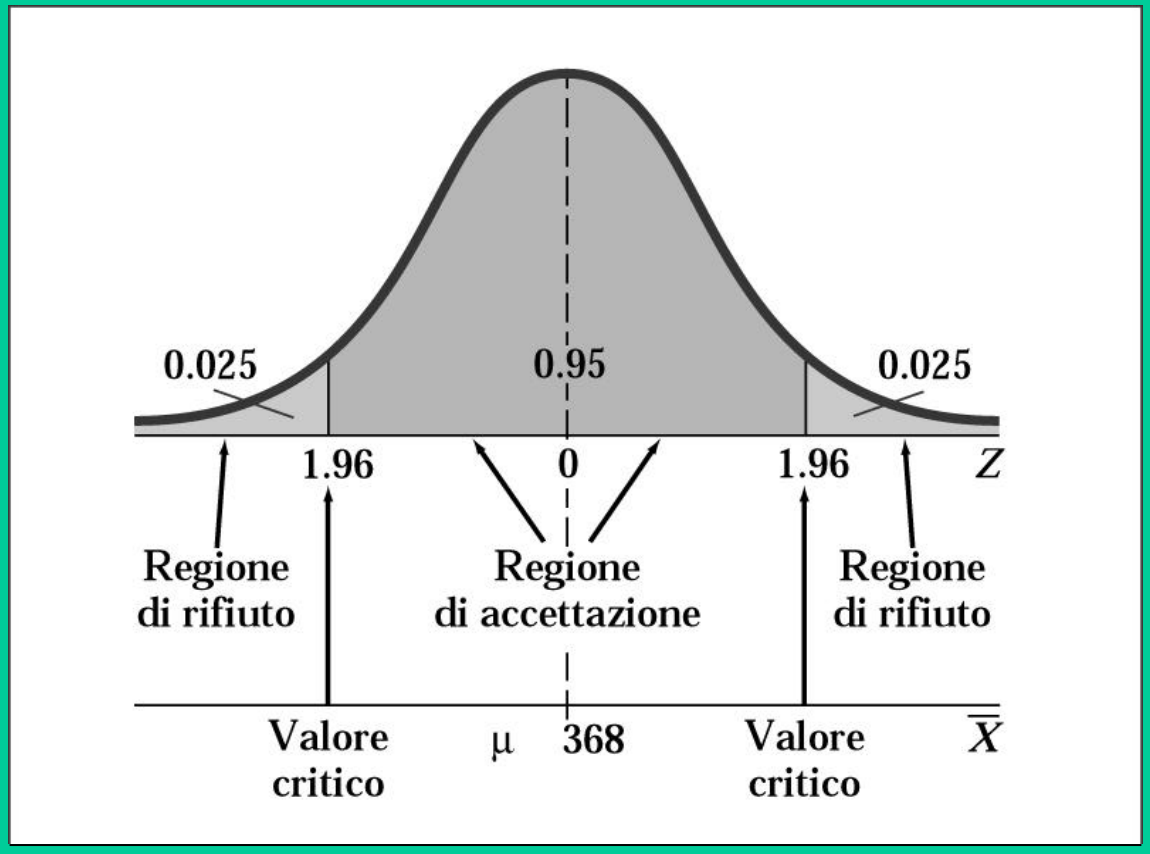
Test Z per la media (s noto)

Approccio del valore critico (livello di significatività di 0,05)

Regola decisionale:

Rifiuto H_0 se Z>+1,96 o se Z<-1,96

Accetto H_0 altrimenti



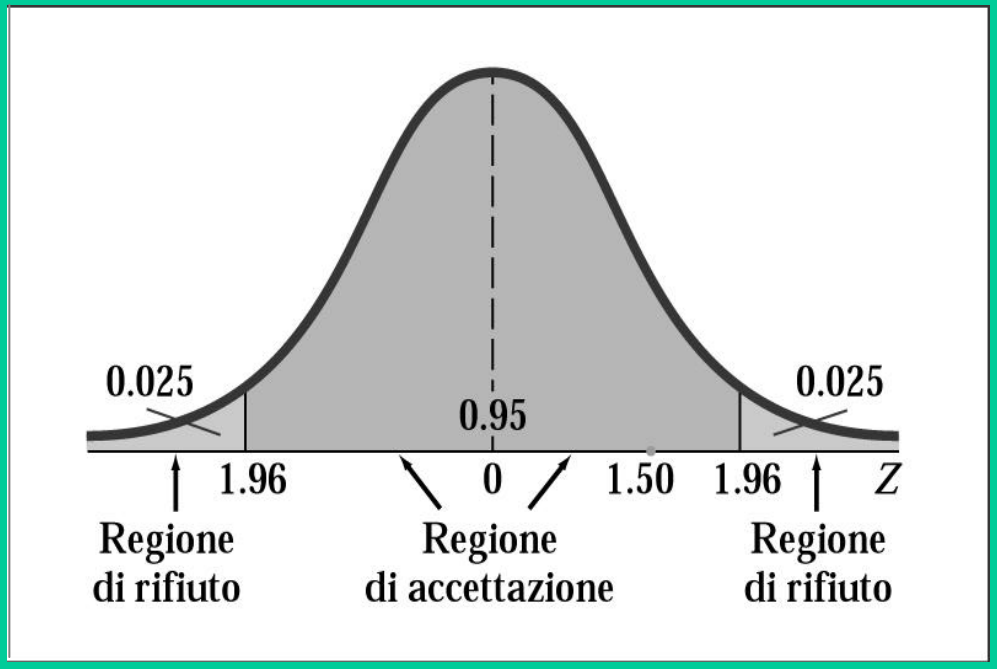
Test Z per la media (s noto)

Esempio: l'azienda che produce scatole metalliche intende valutare se il processo produttivo opera in modo tale da garantire che la lunghezza del lato maggiore sia pari a 368 mm. Viene estratto un campione di 25 scatole. Lo scarto quadratico medio della popolazione è pari a 15 mm e la media campionaria assume il valore 372,5 mm.

H_0: m = 368

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{372.5 - 368}{\frac{15}{\sqrt{25}}} = +1.50$$

Il valore della statistica test mi porta ad accettare H_0.



L'approccio del p-value

Negli ultimi anni, anche grazie all'ampia diffusione di pacchetti statistici e fogli elettronici, si è affermato un altro approccio alla verifica di ipotesi: l'approccio del p-value.

Il ***p*-value** rappresenta la probabilità di osservare un valore della statistica test uguale o più estremo del valore che si calcola a partire dal campione, quando l'ipotesi H_0 è vera.

Il p-value è anche chiamato livello di significatività osservato.

Regola decisionale:

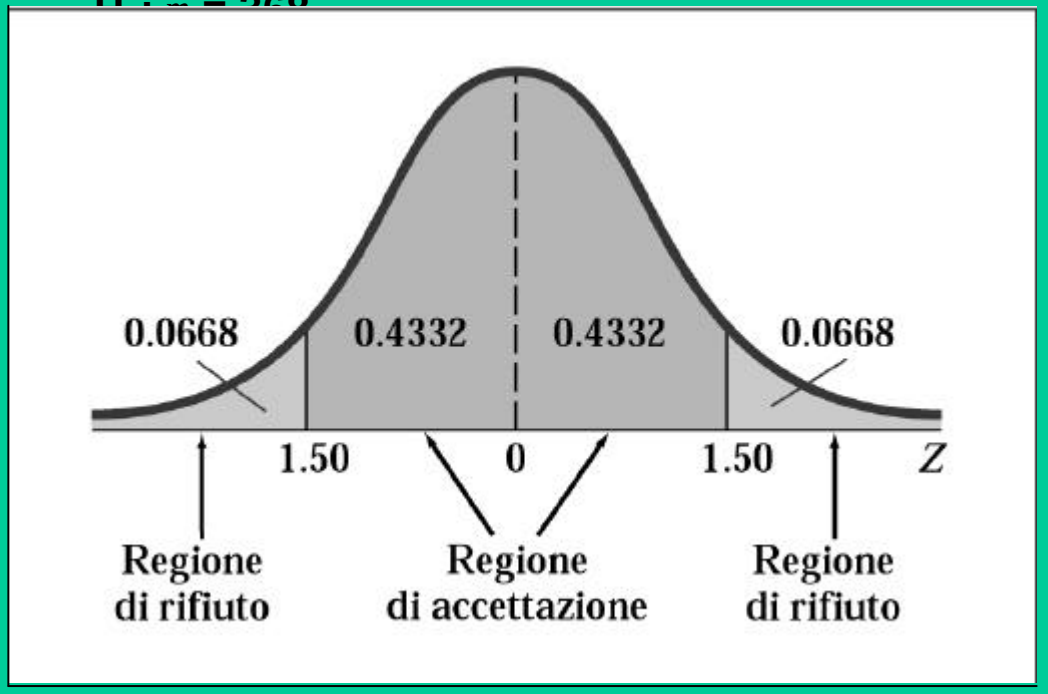
- **se il p-value è maggiore o uguale ad** a**, l'ipotesi nulla viene accettata**
- **se il p-value è minore di** a**, l'ipotesi nulla è rifiutata**

L’approccio del p-value

Esempio: l’azienda che produce scatole metalliche intende valutare se il processo produttivo opera in modo tale da garantire che la lunghezza del lato maggiore sia pari a 368 mm. Viene estratto un campione di 25 scatole. Lo scarto quadratico medio della popolazione è pari a 15 mm e la media campionaria assume il valore 372,5 mm.

Il valore della statistica test è pari a 1,5. La probabilità che Z assuma valori uguali o più estremi di 1,5 coincide con la probabilità che assuma valori maggiori di 1,5 o minori di -1,5 (test a due code). Le due probabilità sono pari a 0,0668 e la loro somma è perciò 0,1336.

Tale valore è maggiore di α **=0,05 perciò accetto** H_0**.**



I test a una coda

In alcune situazioni l'ipotesi alternativa suppone che il parametro sia maggiore o minore di un valore specificato (ci si focalizza in una direzione particolare).

In tal caso H_0 viene rifiutata per valori grandi della statistica Z se l'ipotesi alternativa suppone che il parametro sia maggiore di un certo valore, oppure per valori piccoli della stessa se l'ipotesi alternativa suppone che il parametro sia minore di un certo valore.

Nell'esempio dell'azienda che costruisce scatole di metallo, si potrebbe ritenere che solo se la lunghezza eccede i 368 mm si è in presenza di un problema, altrimenti il processo funziona correttamente. In tal caso

$$H_0: \mu = 368$$

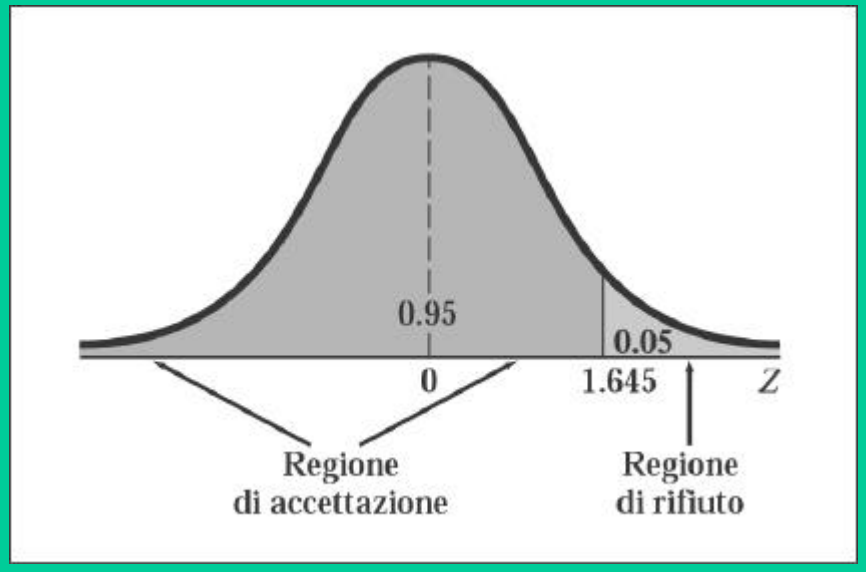
$$H_1: \mu > 368$$

La regione critica è contenuta nella coda destra della distribuzione e corrisponde ad un'area di 0,05. Il valore critico lascia perciò a sinistra un'area pari a 0,95 corrispondente al valore +1,645. Il valore della statistica Z è +1,5 < +1,645. Perciò accetto H_0.

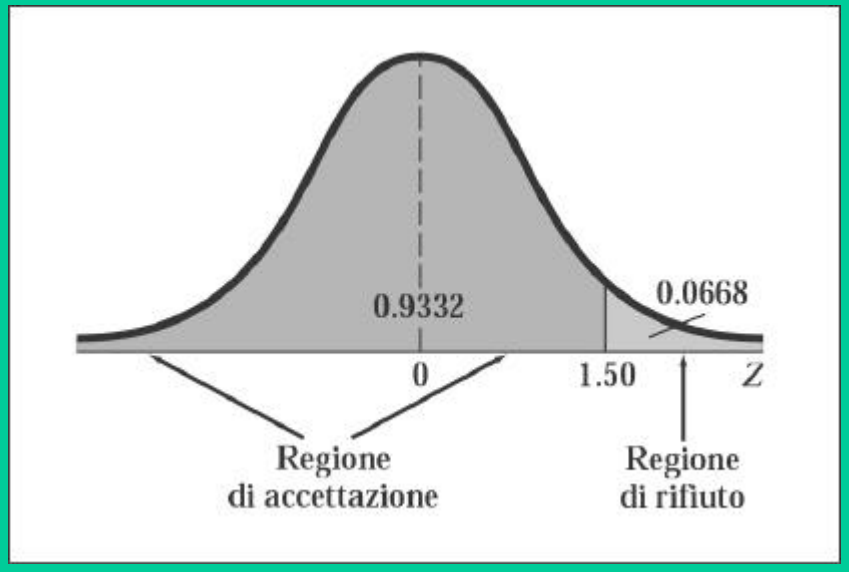
I test a una coda

Adottando l'approccio del p-value si ha che la probabilità che Z assuma valori maggiori di +1,5 è 1-0,9332 = 0,0668 (p-value). Poiché 0,0668 > 0,05 accetto H_0.

Approccio del valore critico



Approccio del p-value



I test a una coda

La specificazione dell'ipotesi nulla e dell'ipotesi alternativa nei test a una coda deve seguire le seguenti regole:

1. **L'ipotesi nulla H_0 è l'ipotesi sottoposta a verifica.**
2. **L'ipotesi alternativa H_1 è specificata come ipotesi opposta a quella nulla e rappresenta la conclusione sostenuta se l'ipotesi nulla è rifiutata.**
3. **L'ipotesi nulla H_0 si riferisce sempre a un *parametro della popolazione* (come μ) non a una *statistica campionaria* (come la media campionaria).**
4. **L'ipotesi nulla contiene *sempre* un segno di uguale riferito a un valore specificato del parametro della popolazione (H_0: μ = 368 mm).**
5. **L'ipotesi alternativa non contiene *mai* un segno di eguale riferito a un valore specificato del parametro della popolazione (H_1: μ > 368 mm).**

Il test t per la media (σ non noto)

Spesso lo scarto quadratico medio della popolazione σ non è noto ma è stimato mediante S.

In tal caso si può ricorrere al test t:

La statistica test *t* per la media (σ non noto)

$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \qquad (6.2)$$

dove la statistica t ha una distribuzione t con $n - 1$ gradi di libertà.

Anche in questo caso si può procedere secondo l'approccio del valore critico o quello del p-value. L'unica differenza consiste nel ricorrere alle tavole della distribuzione t di Student anziché a quelle della normale.

Il test t per la media (s non noto)

Esempio di verifica di ipotesi sulla media con s ignoto, con livello di significatività 0,05 e 11 gradi di libertà.

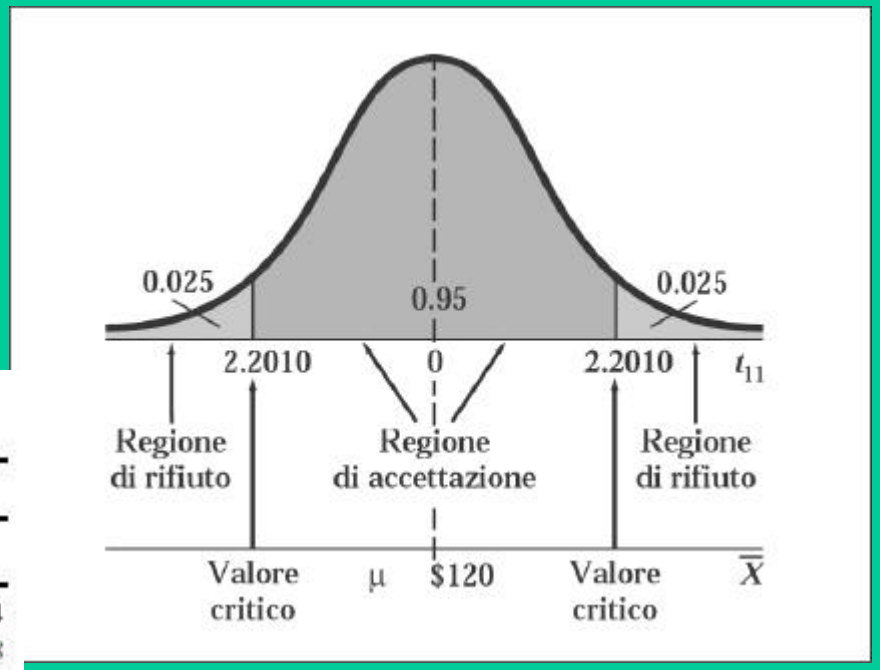
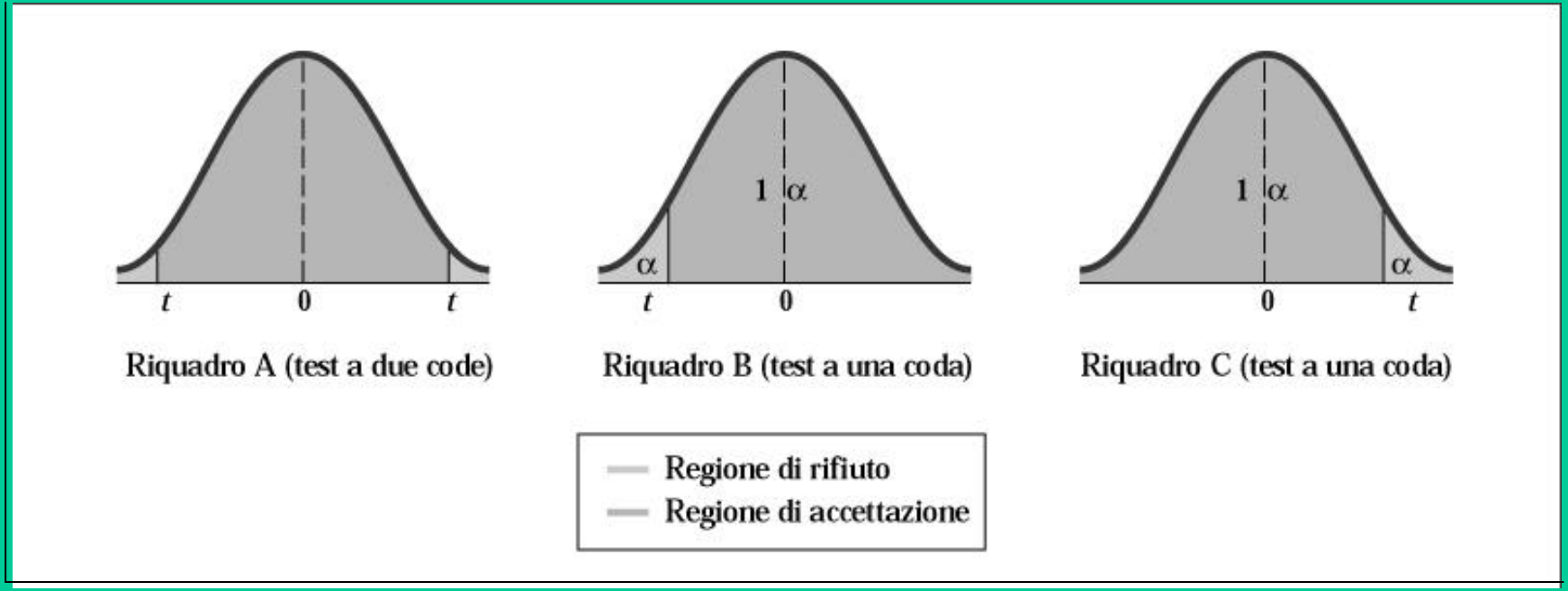


Tabella 6.3 *Ricerca del valore critico della statistica test* t *con* $\alpha = 0{,}05$ *e 11 gradi di libertà*

Gradi di libertà	Area nella coda destra					
	0.25	0.10	0.05	0.025	0.01	0.005
1	1.0000	3.0777	6.3138	12.7062	31.8207	63.6574
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0322
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9980	3.4995
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058

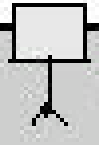
Il test t per la media (s non noto)

Regioni di accettazione e di rifiuto per il test t per la media



Verifica dell'ipotesi di normalità

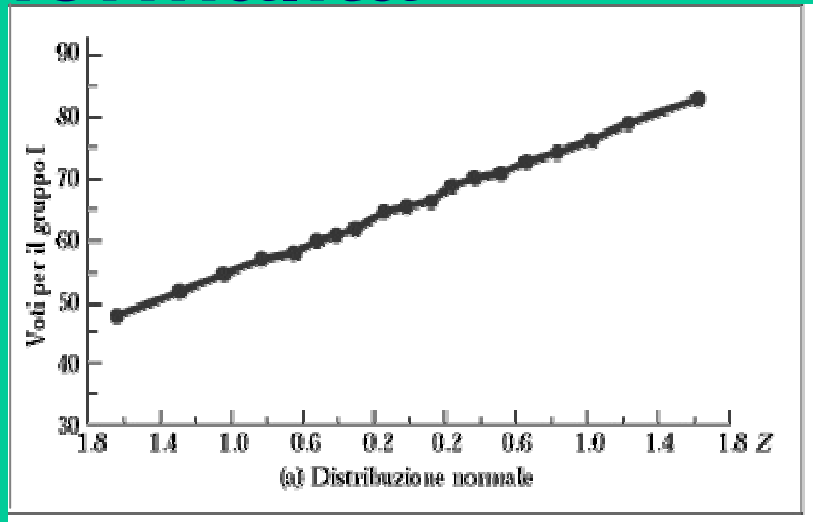
Un **normal probability plot** è un grafico a due dimensioni in cui le osservazioni sono riportate sull'asse verticale e a ciascuna di esse viene fatto corrispondere sull'asse orizzontale il relativo quantile di una distribuzione normale standardizzata.



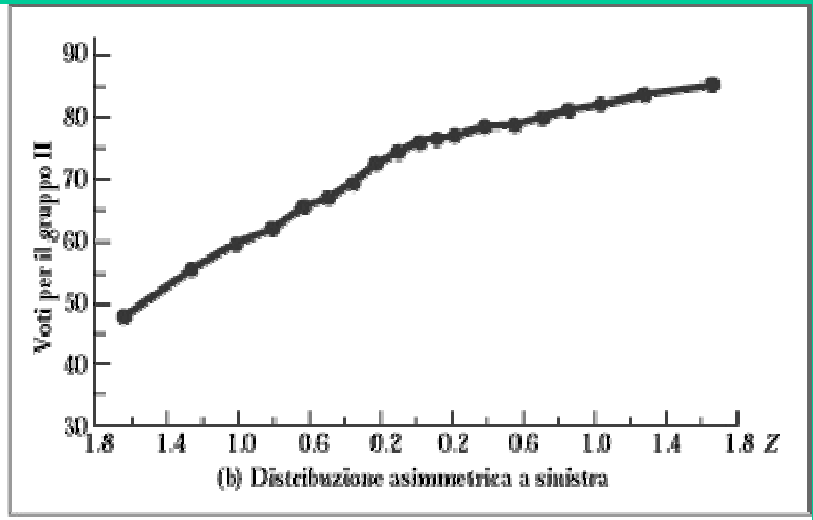
Riquadro 4.6 Come si costruisce un normal probability plot

- ✓ **1.** Ordinare le osservazioni
- ✓ **2.** Determinare i corrispondenti quantili della distribuzione normale standardizzata, i cosiddetti quantili normali standardizzati
- ✓ **3.** Costruire un diagramma di dispersione avente le osservazioni sull'asse verticale e i corrispondenti quantili normali standardizzati sull'asse orizzontale.
- ✓ **4.** Verificare la normalità delle osservazioni controllando se i punti del diagramma si trovano approssimativamente su una linea retta.

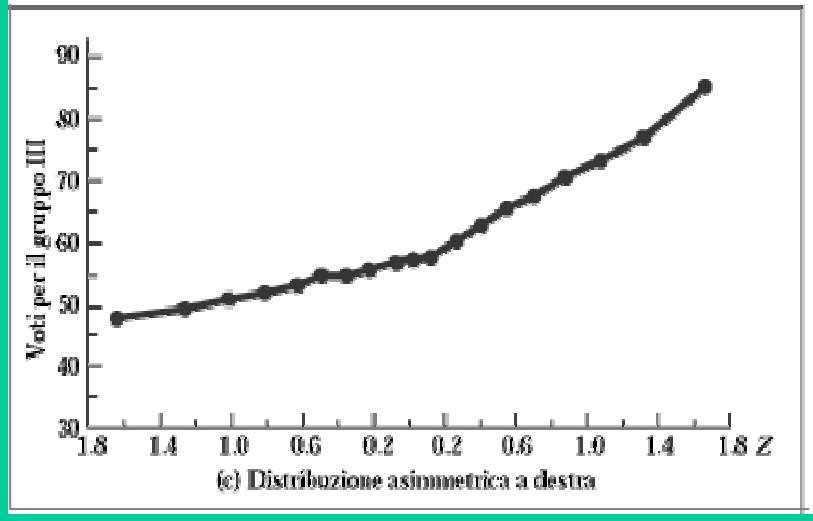
Verifica dell'ipotesi di normalità



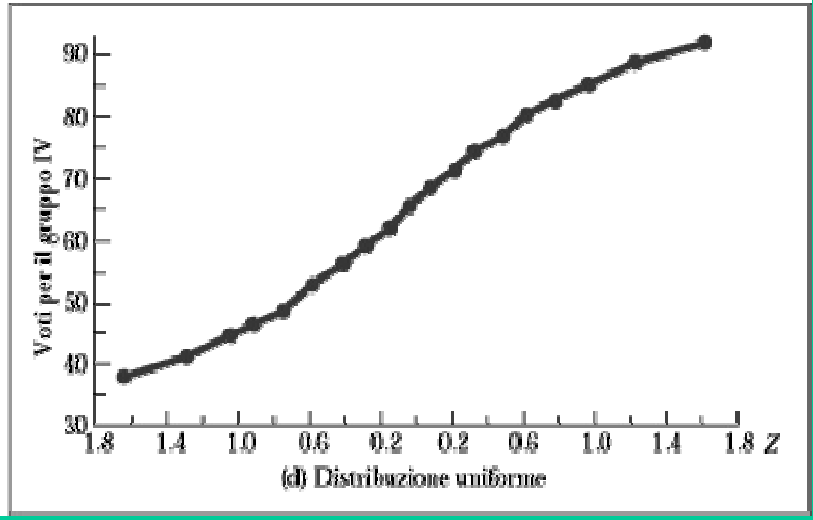
(a) Distribuzione normale



(b) Distribuzione asimmetrica a sinistra



(c) Distribuzione asimmetrica a destra



(d) Distribuzione uniforme

Il test Z per la proporzione

Se il numero di successi X e il numero di insuccessi (n-X) sono entrambi uguali almeno a 5, la distribuzione della proporzione di successi $p_s = X / n$ può essere approssimata dalla distribuzione normale.

La statistica di *Z* per la proporzione

$$Z \simeq \frac{p_s - p}{\sqrt{\frac{p(1-p)}{n}}} \qquad (6.3)$$

dove:

$$p_s = \frac{\text{numero di successi nel campione}}{\text{ampiezza del campione}}$$

p = proporzione di successi ipotizzata

La distribuzione della statistica test Z è approssimativamente normale

La statistica di *Z* per la proporzione

$$Z \simeq \frac{X - np}{\sqrt{np(1-p)}} \qquad (6.4)$$

Il test Z per la proporzione

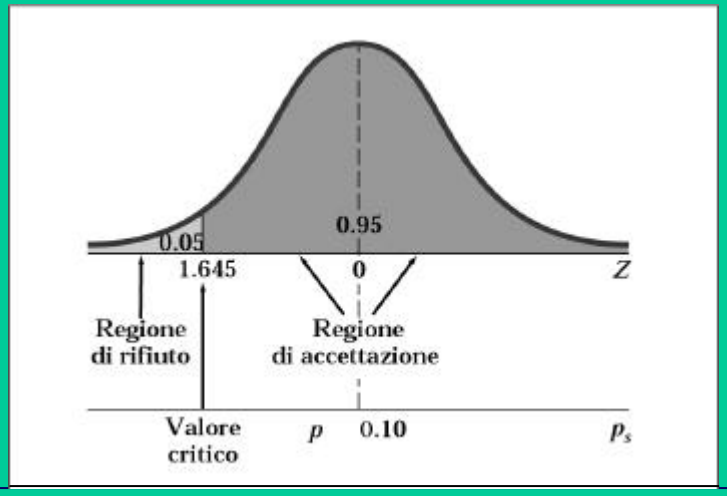
Esempio: supponiamo che il manager operativo dell'azienda che produce scatole metalliche sia interessato a valutare la percentuale di scatole non conformi. Nel passato il 10% delle scatole non è risultata conforme. Si sperimenta un nuovo sistema di produzione e, dopo un giorno di prova, si estrae un campione di 200 scatole, di cui 11 non risultano sigillate in maniera adeguata. Si vuole stabilire se la proporzione di scatole difettose è diminuita:

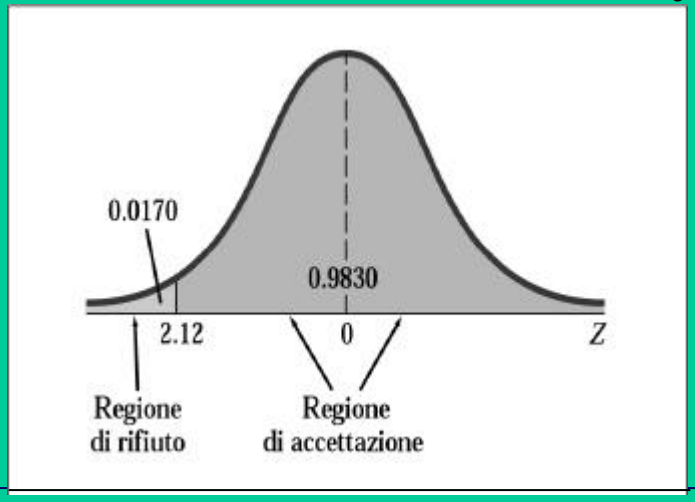
$H_0: p = 0,10$

$H_1: p < 0,10$

Rigetto l'ipotesi nulla se $Z < -1,645$. Se $p_s = 11/200 = 0,055$, $n = 200$ e $p = 0,10$ si ricava $Z = -2,12$. Quindi rifiuto H_0.

Alternativamente il p-value associato a $Z = -2,12$ è $0,0170 < 0,05$ => rifiuto H_0.





VERIFICHE DI IPOTESI SULLA BASE DI DUE CAMPIONI

Test t per campioni indipendenti

Esistono procedure di verifica per confrontare valori di sintesi calcolati su due o più campioni di dati numerici.

Obiettivo: fare inferenza sulle differenze tra i parametri caratteristici di due o più popolazioni.

Esempio: Il responsabile del processo di imbottigliamento di una bevanda analcolica presso uno stabilimento è interessato a ottenere la massima uniformità possibile nei pesi delle bottiglie prodotte. A tal fine vengono sperimentati due diversi strumenti di taratura delle bottiglie, basati su differenti carichi di pressione. Lo scopo è quello di verificare se vi è una differenza significativa tra gli scostamenti medi dei pesi delle bottiglie dal target stabilito (1

litro), quando queste vengono prodotte con lo strumento a maggiore piuttosto che a minore

pressione.

Test t per campioni indipendenti

Siano date due popolazioni indipendenti, con medie e varianze rappresentate nella seguente tabella:

Popolazione 1	Popolazione 2
μ_1, σ_1	μ_2, σ_2

Interessa verificare l'ipotesi nulla

$$H_0: m_1 = m_2 \ (m_1 - m_2 = 0)$$

contro l'ipotesi alternativa

$$H_1: m_1 \ ? \ m_2 \ (m_1 - m_2 \ ? \ 0)$$

La statistica test per verificare queste ipotesi si baserà sulla differenza tra le medie campionarie. Per valori piccoli di tale differenza in valore assoluto accetterò H_0, altrimenti rifiuterò H_0.

Test t per campioni indipendenti

In virtù del teorema del limite centrale la seguente statistica si distribuisce asintoticamente secondo una normale standard

Test *Z* per la differenza fra due medie

La statistica test in questo caso è:

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (7.1)$$

dove:

$\overline{X}_1$ = media degli elementi del campione estratto dalla popolazione 1
μ_1 = media della popolazione 1
σ_1^2 = varianza della popolazione 1
n_1 = ampiezza del campione estratto dalla popolazione 1
$\overline{X}_2$ = media degli elementi del campione estratto dalla popolazione 2
μ_2 = media della popolazione 2
σ_2^2 = varianza della popolazione 2
n_2 = ampiezza del campione estratto dalla popolazione 2

Test t per campioni indipendenti

Se, come spesso capita, non sono noti i valori s_1 e s_2, si può ricorrere al test t seguente:

Test *t* per la differenza fra due medie basato sulle varianze campionarie ponderate

La statistica test in questo caso è:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad (7.2)$$

dove

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

e

S_p^2 = varianza ponderata
$\bar{X}_1$ = media degli elementi del campione estratto dalla popolazione 1
S_1^2 = varianza degli elementi del campione estratto dalla popolazione 1
n_1 = ampiezza del campione estratto dalla popolazione 1
$\bar{X}_2$ = media degli elementi del campione estratto dalla popolazione 2
S_2^2 = varianza degli elementi del campione estratto dalla popolazione 2
n_2 = ampiezza del campione estratto dalla popolazione 2

Si dimostra che la statistica test t sotto l'ipotesi nulla si distribuisce secondo una t di Student con $n_1 + n_2 - 2$ gradi di libertà.

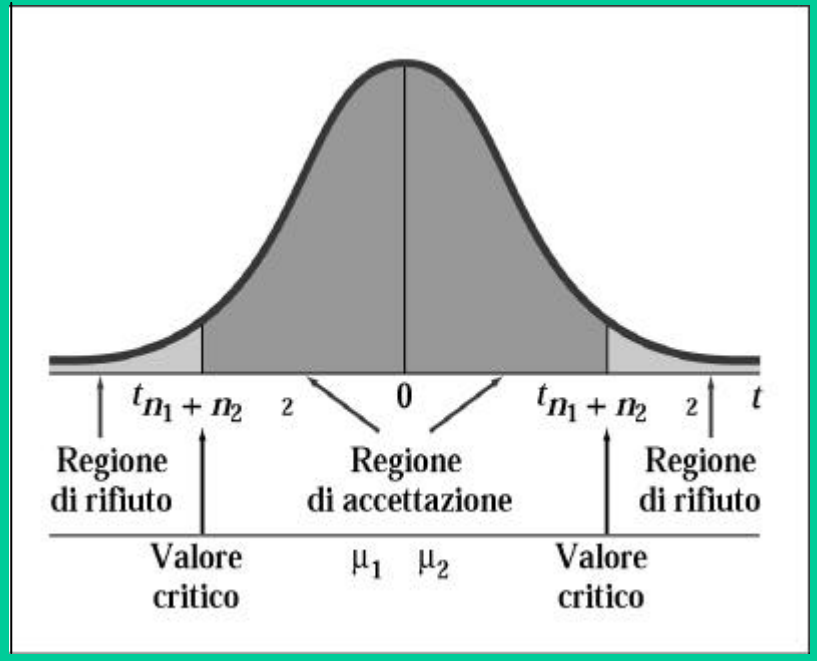
Test t per campioni indipendenti

Regola decisionale:

rifiuto H_0 se $t > t_{n1+n2-2}$ oppure $t < -t_{n1+n2-2}$

accetto H_0 altrimenti

dove $t_{n1+n2-2}$ indica il valore della distribuzione t di Student con n_1+n_2-2 gradi di libertà che lascia a destra una probabilità pari ad $\alpha/2$.



Test t per campioni indipendenti

Tabella 7.1 *Confronto fra gli scostamenti dal target dei pesi delle bottiglie prodotte da uno strumento a bassa pressione (25 psi) e uno strumento a pressione più elevata (30 psi)*

25 PSI				30 PSI			
−2.8	1.2	1.4	0.9	0.2	0.4	1.6	3.4
−1.6	−2.0	3.4		2.1	1.7	4.0	
0.2	−1.0	0.6		2.6	3.3	2.7	

Nota: uno scostamento negativo dal target denota una bottiglia "troppo vuota", mentre uno scostamento positivo dal target una bottiglia riempita eccessivamente (gli scostamenti sono misurati in millilitri).

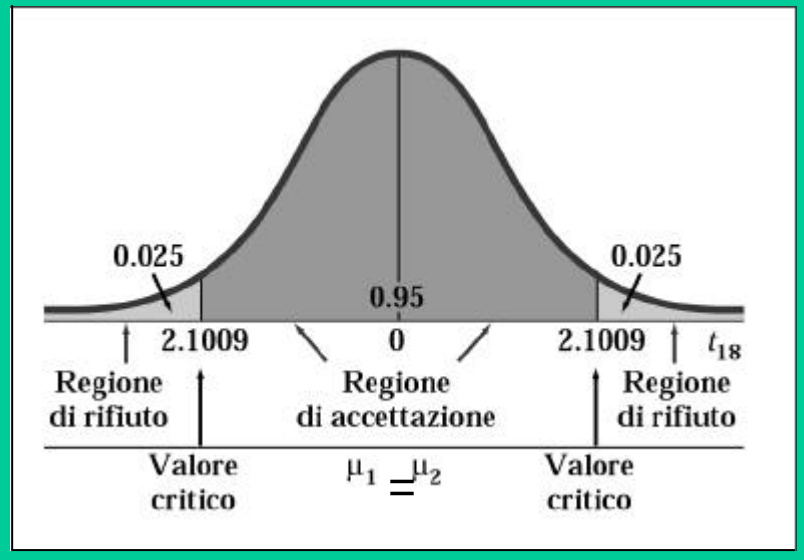


Tabella 7.2 *Alcune statistiche di riepilogo sui due campioni di bottiglie*

25 PSI		30 PSI	
$n_1 = 10$	$S_1^2 = 3.5068$	$n_2 = 10$	$S_2^2 = 1.5733$
$\bar{X}_1 = 0.03$	$S_1 = 1.873$	$\bar{X}_2 = 2.200$	$S_2 = 1.254$

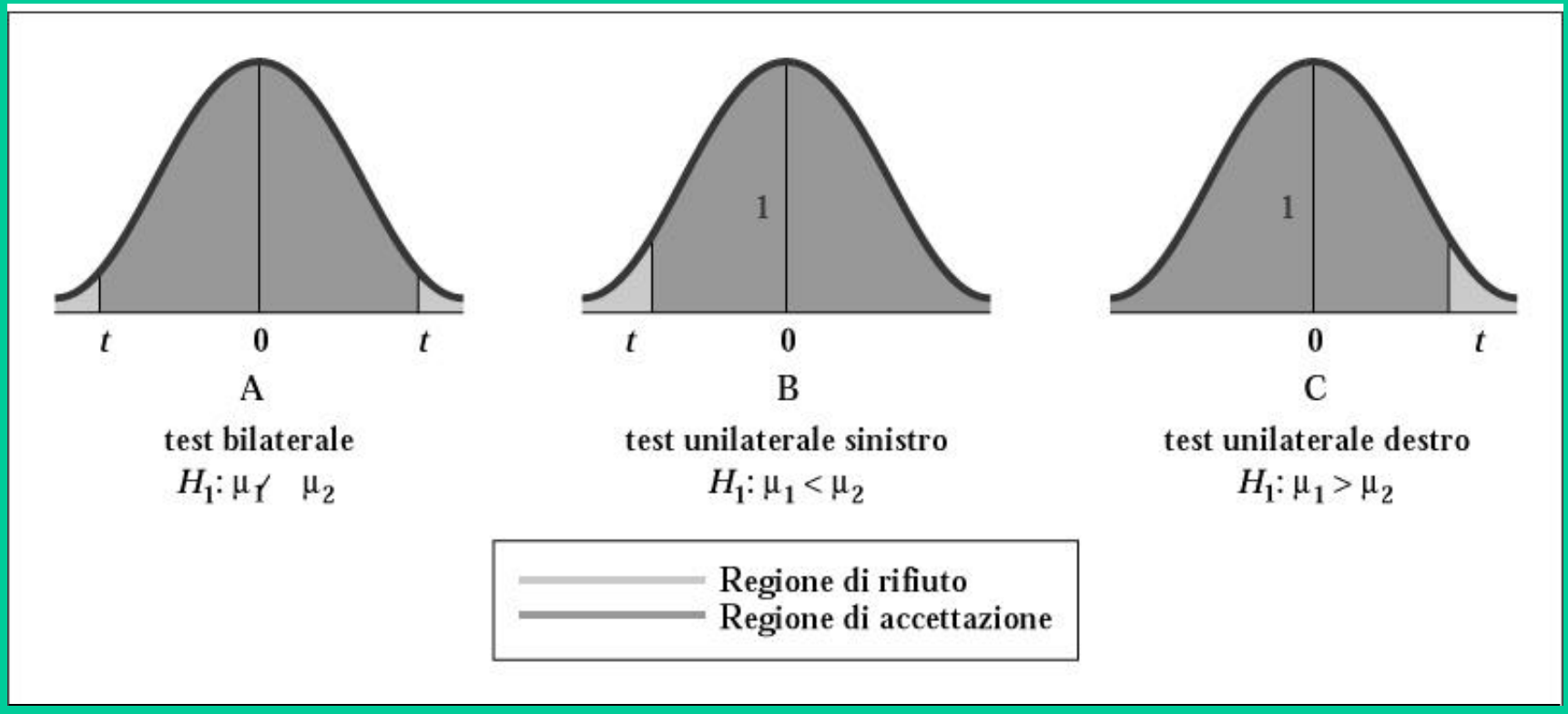
o, poiché $n_1 = n_2$,

$$S_p^2 = \frac{3.5068 + 1.5733}{2} = \frac{5.0801}{2} = 2.5401$$

di modo che:

$$t = \frac{0.03 - 2.20}{\sqrt{2.5401\left(\frac{1}{10} + \frac{1}{10}\right)}} = \frac{-2.17}{\sqrt{0.5080}} = -3.04$$

Test t per campioni indipendenti



<u>Condizioni per applicare il test</u>: per applicare il test t sulla differenza tra le medie di due popolazioni indipendenti occorre assumere che le popolazioni sono distribuite normalmente e hanno stessa varianza.

Il test è comunque robusto rispetto all'assunzione di normalità per campioni abbastanza grandi.

LA REGRESSIONE LINEARE SEMPLICE E LA CORRELAZIONE

Regressione e correlazione

Esistono molti metodi di inferenza statistica che si riferiscono ad una sola variabile statistica.

Obiettivo della lezione: studio della relazione tra due variabili.

Tecniche oggetto di studio:

regressione → **Costruire un modello attraverso cui** prevedere **i valori di una** variabile dipendente **o** risposta **(quantitativa) a partire dai valori di una o più** variabili indipendenti **o** esplicative

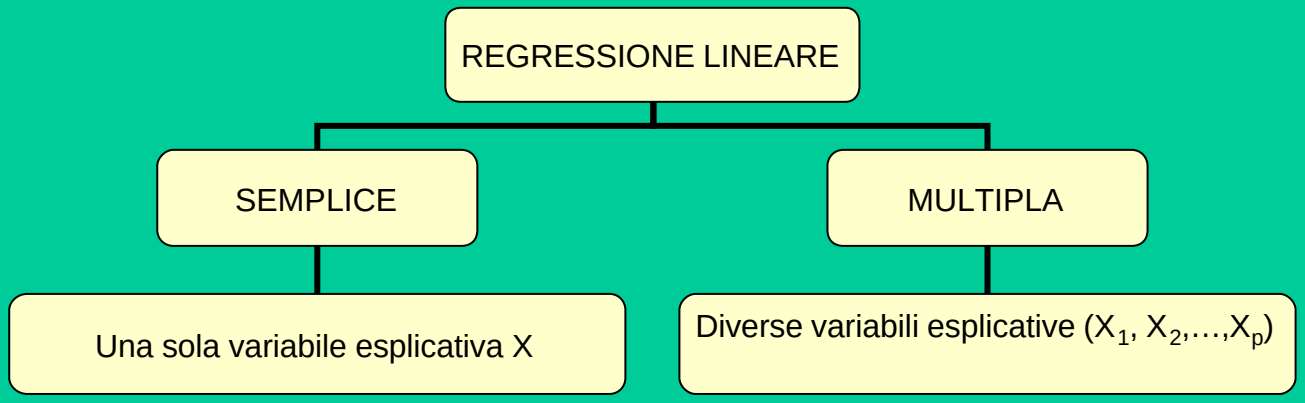
correlazione → **Studio della associazione tra variabili quantitative**

Regressione lineare

Solitamente nel modello di regressione si indica con

Y la variabile dipendente

X la variabile esplicativa

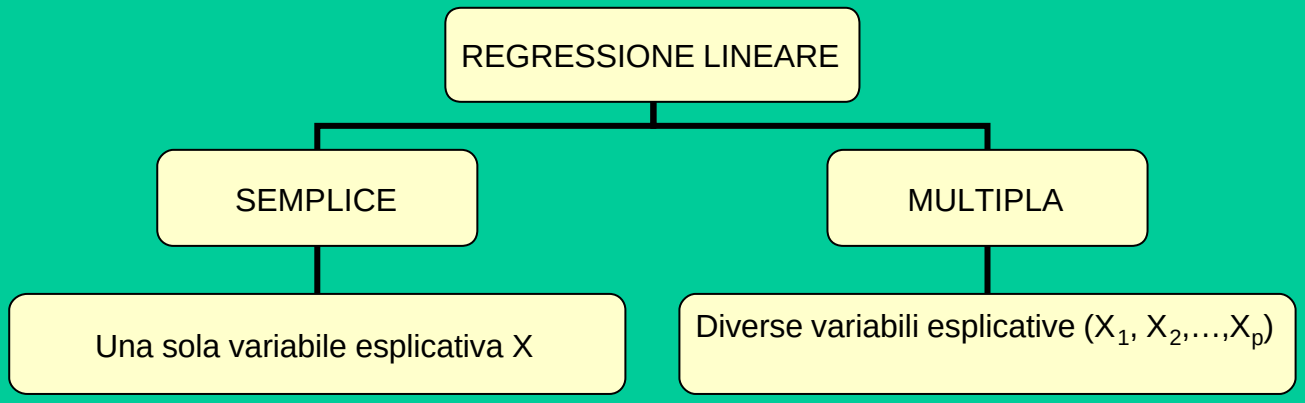


Regressione lineare

Solitamente nel modello di regressione si indica con

Y la variabile dipendente

X la variabile esplicativa



Il modello di regressione

Per studiare la relazione tra due variabili è utile il diagramma di dispersione in cui si riportano i valori della variabile esplicativa X sull'asse delle ascisse e i valori della variabile dipendente Y sull'asse delle ordinate.

La relazione tra due variabili può essere espressa mediante funzioni matematiche più o meno complesse tramite un modello di regressione.

Il modello di regressione lineare semplice è adatto quando i valori delle variabili X e Y si distribuiscono lungo una retta nel diagramma di dispersione.

Il modello di regressione lineare semplice

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad (9.1)$$

dove

β_0 = l'intercetta per la popolazione

β_1 = l'inclinazione per la popolazione

ϵ_i = l'errore casuale in Y corrispondente all'i–esima osservazione

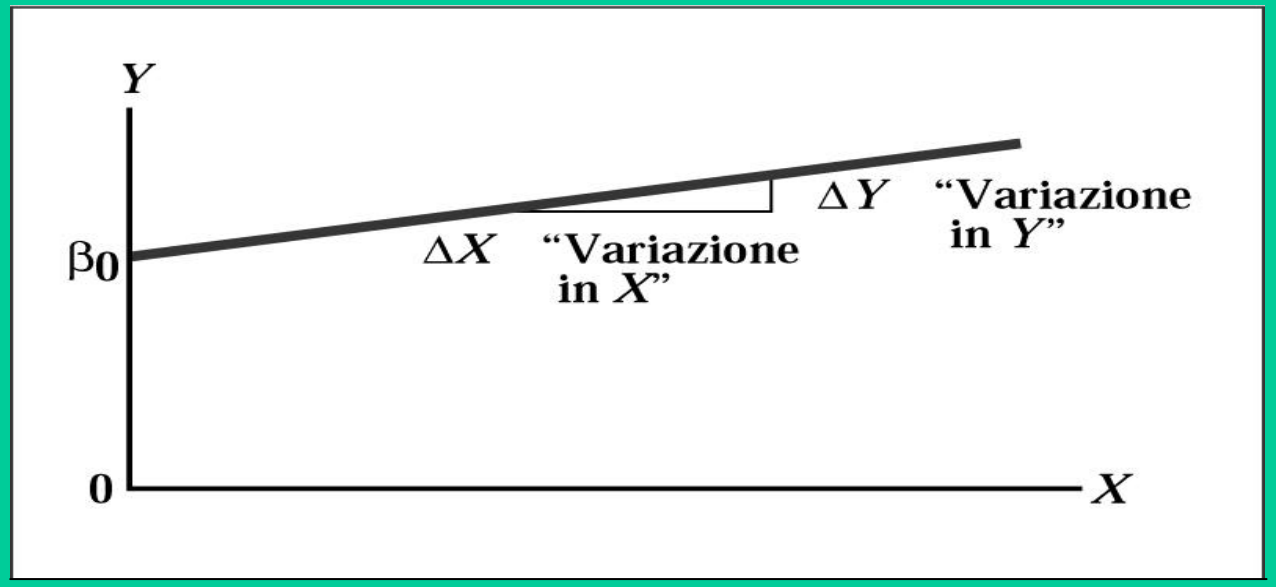
Il modello di regressione

L'inclinazione b_1 indica come varia Y in corrispondenza di una variazione unitaria di X.

L'intercetta b_0 corrisponde al valore medio di Y quando X è uguale a 0.

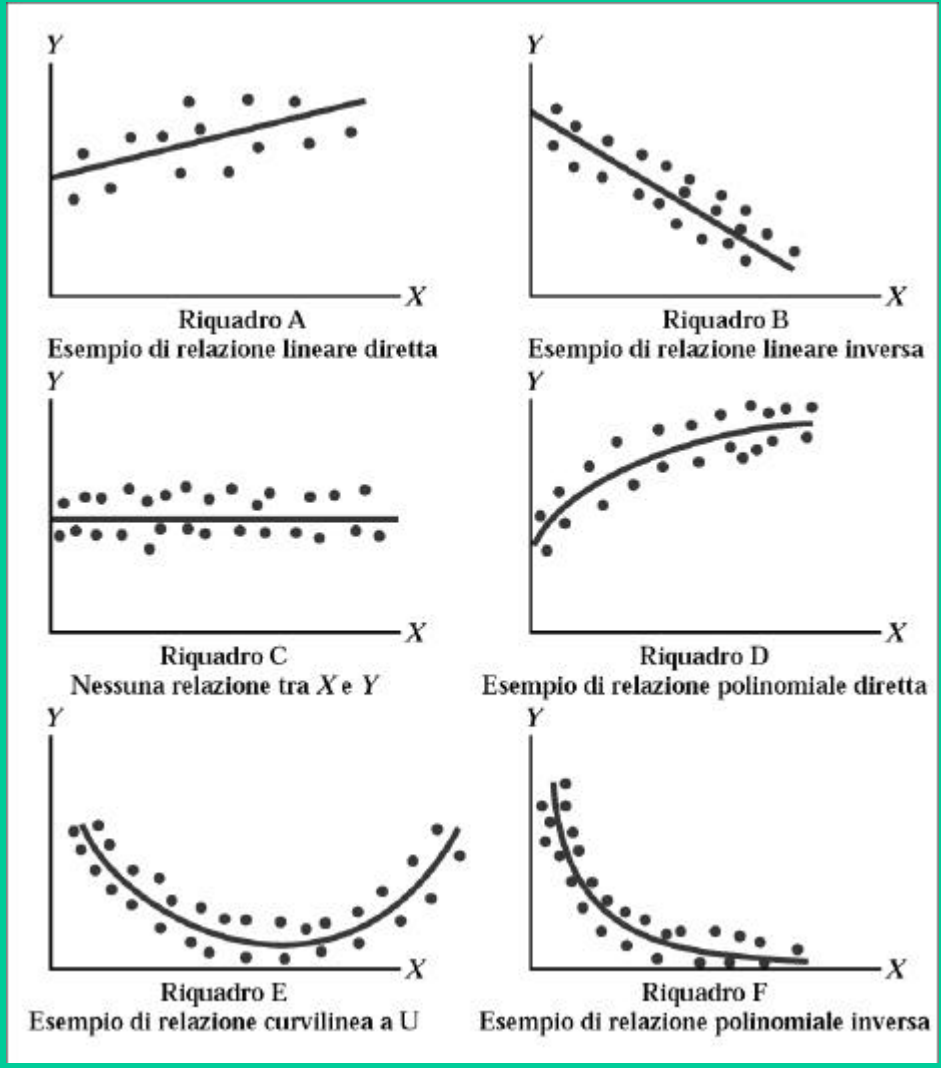
Il segno di b_1 indica se la relazione lineare è positiva o negativa.

Esempio di relazione lineare positiva



Il modello di regressione

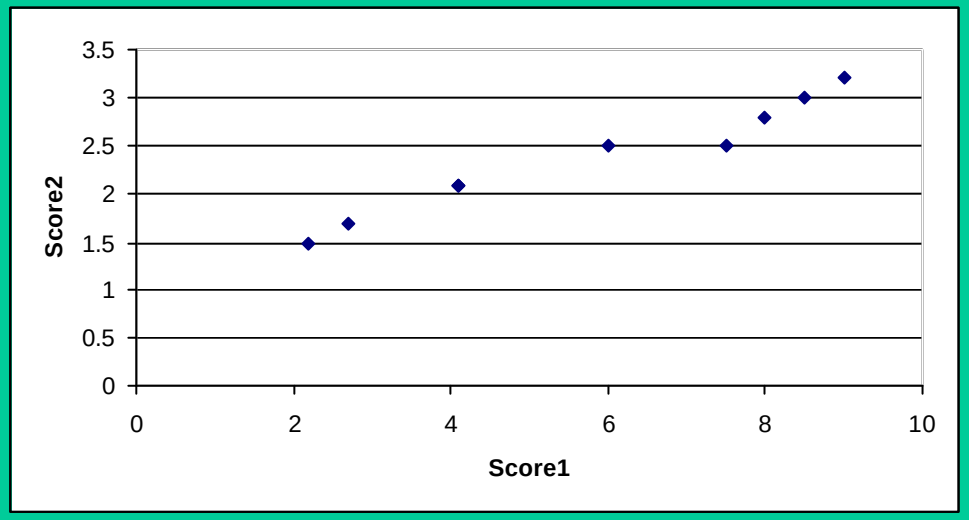
La scelta del modello matematico appropriato è suggerita dal modo in cui si distribuiscono i valori delle due variabili nel diagramma di dispersione



Il modello di regressione

Esempio: un produttore desidera ottenere una misura della qualità di un prodotto ma la procedura è troppo costosa. Decide allora di stimare questa misura (score 2) a partire dall'osservazione di un'altra misura (score 1) più semplice meno costosa da ottenere.

Unità di prodotto	Score1	Score2
1	4.1	2.1
2	2.2	1.5
3	2.7	1.7
4	6	2.5
5	8.5	3
6	4.1	2.1
7	9	3.2
8	8	2.8
9	7.5	2.5



Equazione della retta di regressione

Si dimostra che sotto certe ipotesi i parametri del modello b_0 e b_1 possono essere stimati ricorrendo ai dati del campione. Indichiamo con b_0 e b_1 le stime ottenute.

L'equazione campionaria del modello di regressione lineare

La previsione di Y in base al modello di regressione lineare è data dalla somma tra l'intercetta campionaria e il prodotto tra il valore di X e l'inclinazione campionaria

$$\hat{Y}_i = b_0 + b_1 X_i \qquad (9.2)$$

dove

$\hat{Y}_i$ = previsione di Y per l'osservazione i

X_i = valore di X per l'osservazione i

La regressione ha come obiettivo quello di individuare la retta che meglio si adatta ai dati.

Esistono vari modi per valutare la capacità di adattamento

Il criterio più semplice è quello di valutare le differenze tra i valori osservati (Y_i) e i valori previsti (Y_i)

Equazione della retta di regressione

Il metodo dei minimi quadrati consiste nel determinare b_0 e b_1 rendendo minima la somma dei quadrati delle differenze tra i valori osservati Y_i e i valori stimati Y_i.

si tratterà di *minimizzare* la somma dei loro quadrati:

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

dove

Y_i = il vero valore di Y per l'osservazione di i

$\hat{Y}_i$ = il valore previsto di Y per l'osservazione di i

Dal momento che in base al modello proposto $\hat{Y}_i = b_0 + b_1 X_i$, si tratta di minimizzare la seguente espressione:

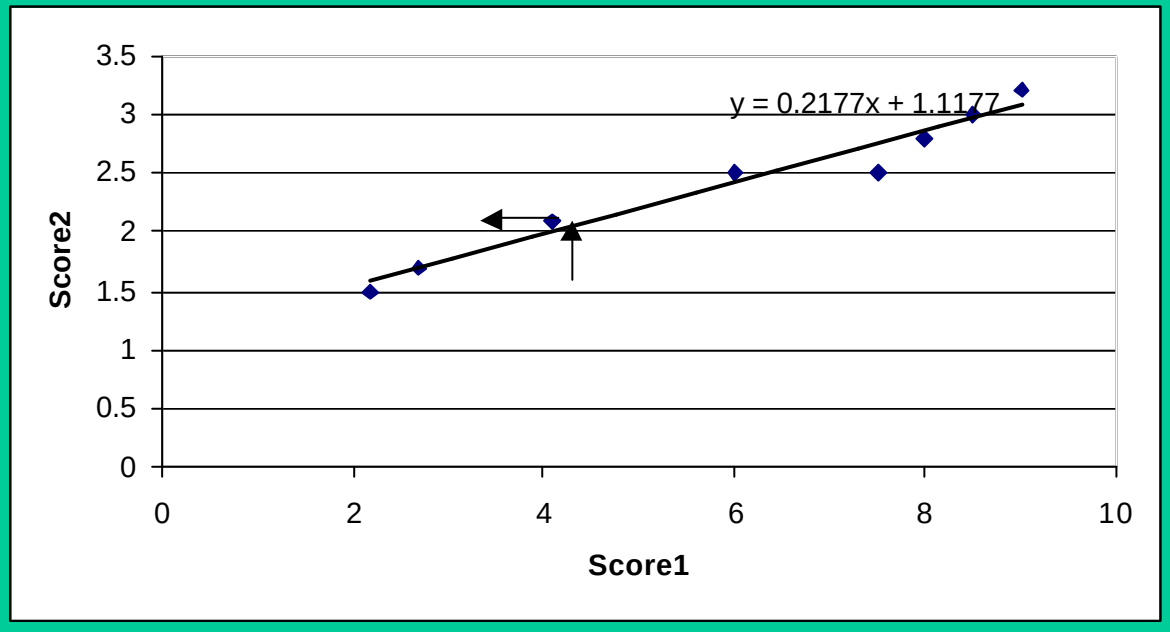
$$\sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$$

rispetto alle due incognite b_0 e b_1.

I valori b_0 e b_1 sono chiamati coefficienti di regressione.

Equazione della retta di regressione

Nell'esempio precedente in cui si intendeva prevedere il valore di una misura di qualità score2 in funzione di un'altra misura score1, applicando il metodo dei minimi quadrati si ottiene la seguente retta di regressione:



Risulta:

b_1 = 0,2177

b_0 = 1,1177

Perciò se aumenta di un'unità il valore di score1, il valore previsto di score2 subisce un incremento di 0,2177.

Se score1 assume valore 0, il valore previsto per score2 è pari a 1,1177.

Tramite l'equazione score2 = 1,1177 + 0,2177 score1 è possibile prevedere i valori di score2 in funzione di quelli osservati di score1. Se ad esempio osservassimo un valore di score1 pari a 4,5 il valore stimato di score2 sarebbe 2,1.

Misure di variabilità

Il coefficiente di determinazione è una misura utile per valutare il modello di regressione

Esso misura la parte di variabilità di Y spiegata dalla variabile X nel modello di regressione.

Il coefficiente di determinazione

Il coefficiente di determinazione è dato dal rapporto tra la somma dei quadrati della regressione e la somma totale dei quadrati.

$$r^2 = \frac{SQR}{SQT} \qquad (9.7)$$

Nell'esempio precedente risulta $r^2 = 0,96$

Analisi dei residui

Il residuo e_i è una stima dell'errore che commetto nel prevedere Y_i tramite Y_i.

Il residuo

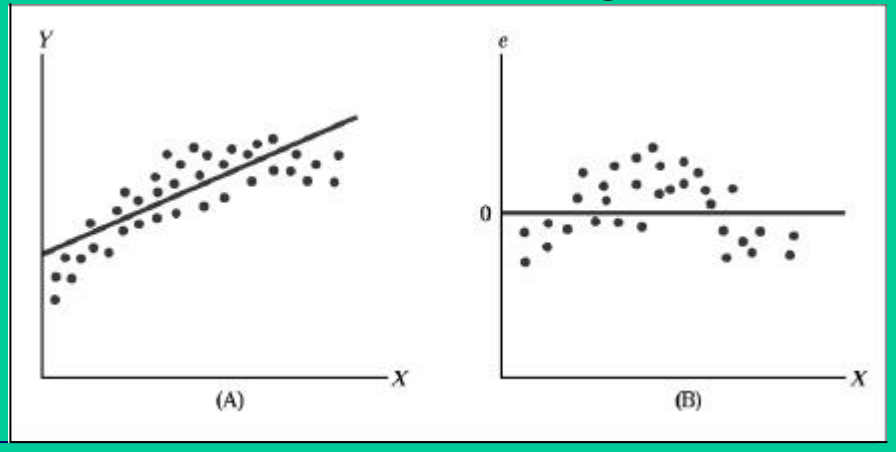
Il residuo è uguale alla differenza tra valore osservato e il valore previsto di Y:

$$e_i = Y_i - \hat{Y}_i \qquad (9.9)$$

Per stimare la capacità di adattamento ai dati della retta di regressione è opportuna una analisi grafica → grafico di dispersione dei residui (ordinate) e dei valori di X (ascisse).

Se si evidenzia una relazione particolare il modello non è adeguato.

Nell'esempio a lato il modello di regressione lineare non sembra appropriato. Il grafico a destra evidenzia lo scarso adattamento ai dati del modello (lack of fit). Quindi il modello polinomiale è più appropriato.

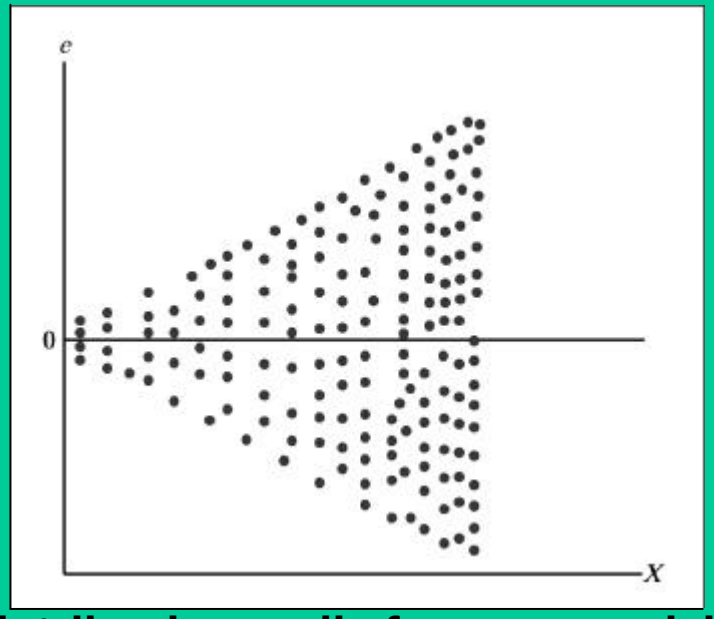


Analisi dei residui

Valutazione delle ipotesi:

- **Omoschedasticità: il grafico dei residui rispetto a X consente di stabilire anche se la variabilità degli errori varia a seconda dei valori di X**

Il grafico a lato evidenzia ad esempio che la variabilità dei residui aumenta all'aumentare dei valori di X.

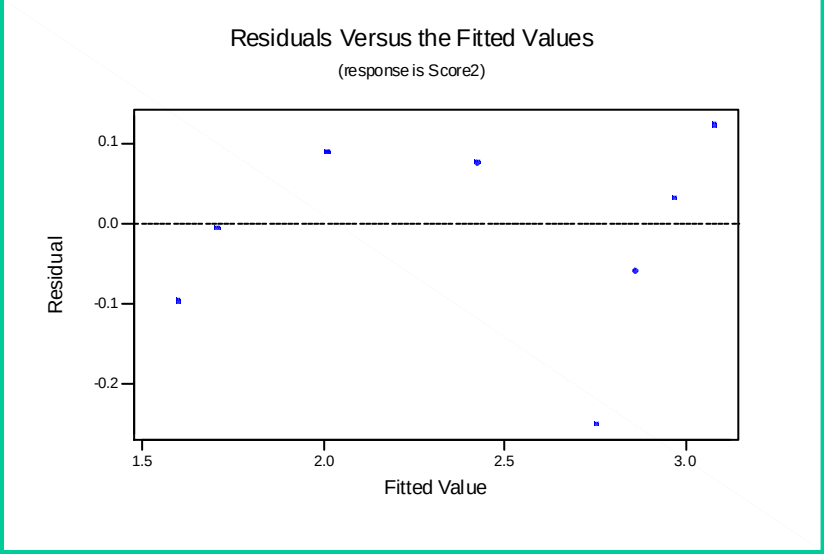


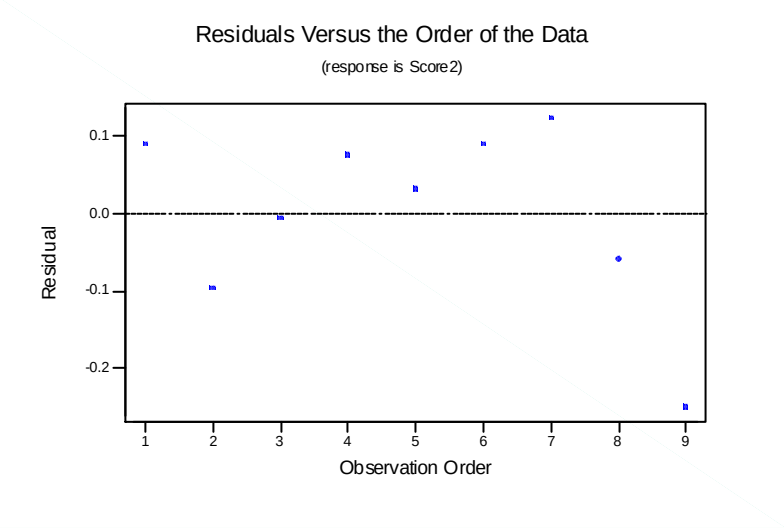
- **Normalità: rappresentazione della distribuzione di frequenze dei residui (es. istogramma)**
- **Indipendenza: rappresentando i residui nell'ordine con cui sono stati raccolti i dati emerge un'eventuale autocorrelazione tra osservazioni successive → statistica di Durbin-Watson.**

Analisi dei residui

Dall'esempio precedente risulta che i residui non si distribuiscono in modo regolare al variare delle stime della variabile dipendente (e quindi anche al variare della X). Il modello quindi non è ben specificato.

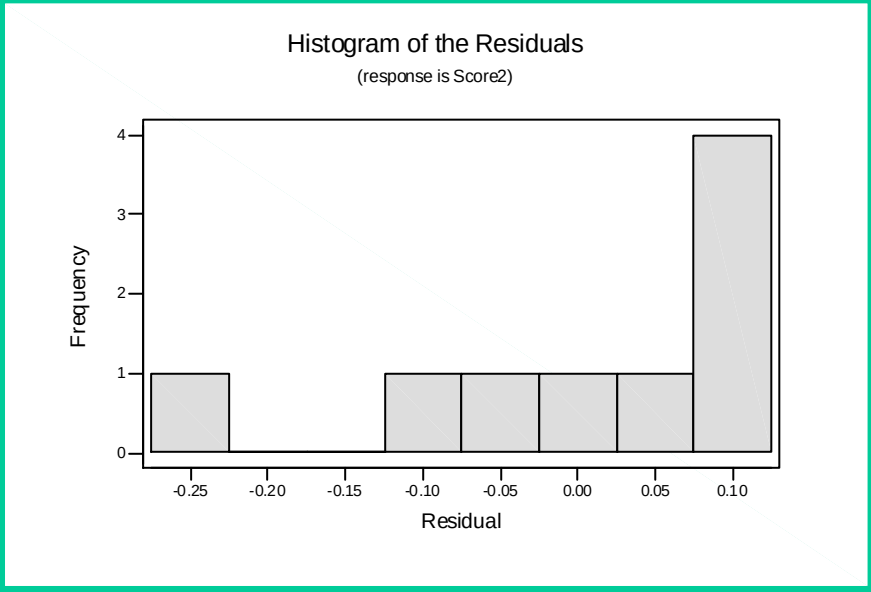
Il grafico dei residui rispetto al tempo non sembra evidenziare l'esistenza di autocorrelazione dei primi.

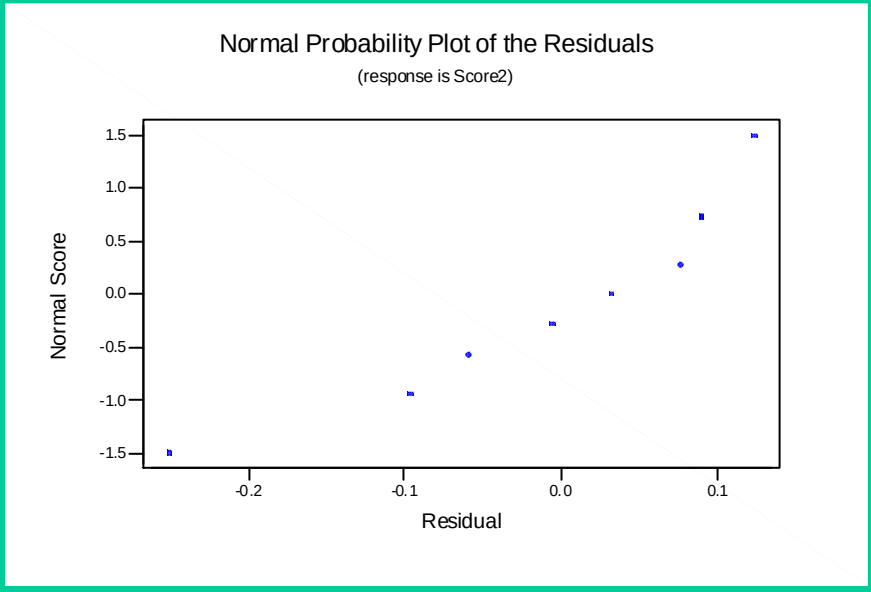




Analisi dei residui

Per quanto riguarda la normalità dei residui, l'istogramma delle frequenze e il normal probability plot ci portano ad escludere che la condizione sia verificata.





Inferenza sull'inclinazione della retta di regressione

Possiamo stabilire se tra le variabili X e Y sussiste una relazione lineare significativa sottoponendo a verifica l'ipotesi che b_1 (inclinazione della popolazione) sia uguale a zero.

H_0: $\beta_1 = 0$ (non vi è una relazione lineare)

H_1: $\beta_1 \neq 0$ (vi è una relazione lineare)

Il test *t* per la verifica di ipotesi sull'inclinazione β_1

La statistica t è data dalla differenza tra l'inclinazione campionaria e l'inclinazione ipotizzata della popolazione, il tutto diviso per l'errore standard dell'inclinazione.

$$t = \frac{b_1 - \beta_1}{S_{b_1}}$$

dove

(9.11)

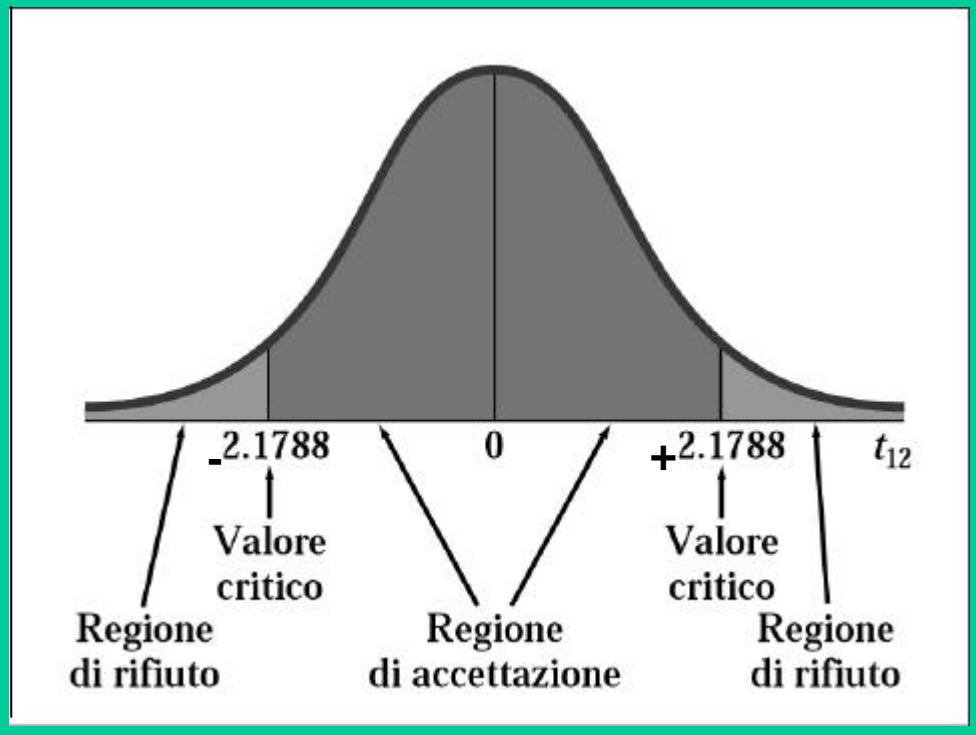
$$S_{b_1} = \frac{S_{YX}}{\sqrt{SQX}}$$

$$SQX = \sum_{i=1}^{n} (X_i - \overline{X})^2$$

La statistica t ha una distribuzione t di Student con $n - 2$ gradi di libertà.

Inferenza sull'inclinazione della retta di regressione

Se ad esempio a=0,05 e n=14, allora le regioni di accettazione e di rifiuto sono definite come segue:



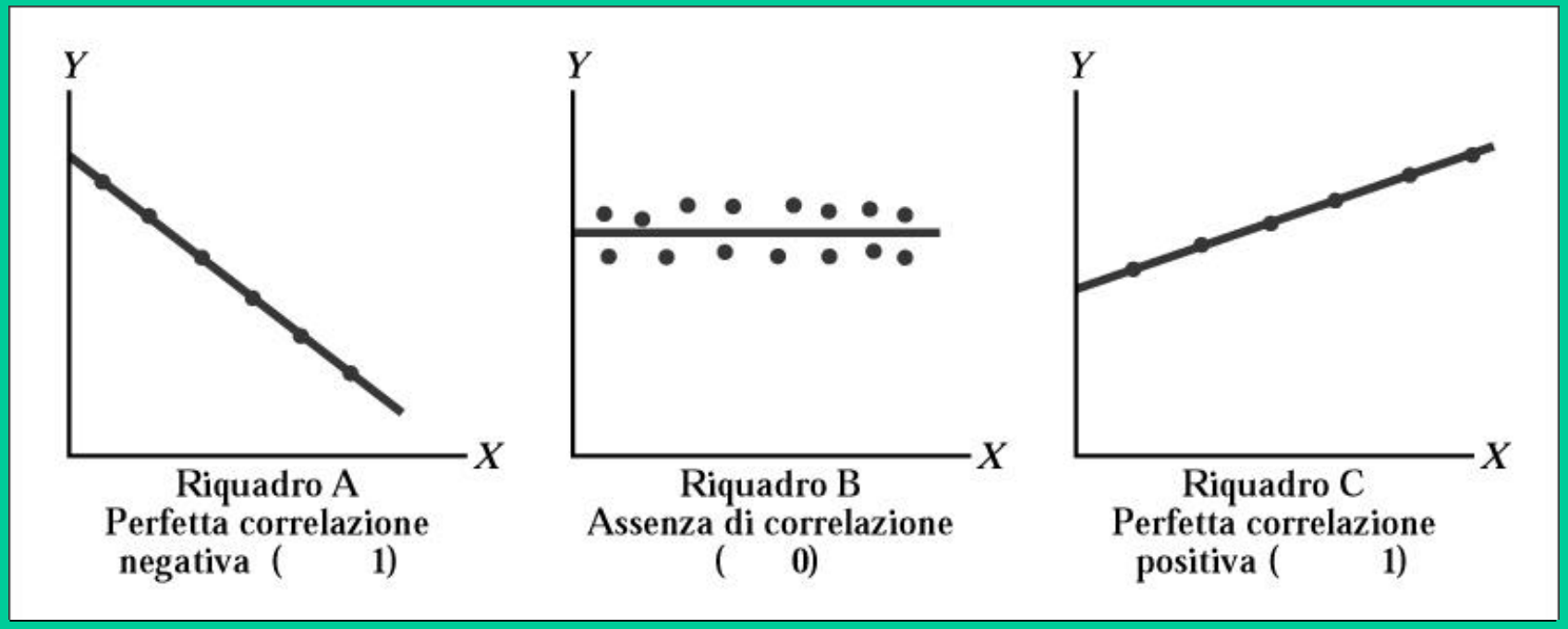
Nell'esempio del modello di regressione in cui score1 è variabile esplicativa e score2 variabile dipendente abbiamo che b_1=0,2177 n=8

$t=b_1/S_{b1}=12,51>t_6 = 2,45$

perciò rigetto l'ipotesi che l'inclinazione sia nulla a favore dell'ipotesi che esista inclinazione significativa.

La correlazione

L'intensità della relazione (associazione) tra due variabili di una popolazione viene misurata in genere mediante il coefficiente di correlazione r**.**



La correlazione

Il coefficiente di correlazione r può essere stimato mediante il coefficiente di correlazione campionaria:

Il coefficiente di correlazione

$$r = \frac{SQXY}{\sqrt{SQX}\sqrt{SQY}} \qquad (9.23)$$

dove

$$SQXY = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

$$SQX = \sum_{i=1}^{n} (X_i - \overline{X})^2$$

$$SQY = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

La correlazione

Il coefficiente di correlazione campionaria r può essere impiegato per verificare se tra due variabili sussista una associazione significativa nella popolazione.

H_0: $\rho = 0$ (non vi è correlazione tra le variabili)

H_1: $\rho \neq 0$ (vi è correlazione tra le variabili)

Statistica test per la verifica dell'esistenza della correlazione

$$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} \qquad (9.24)$$

Ha una distribuzione t di Student con $n - 2$ gradi di libertà.

Nel nostro esempio risulta che tra score1 e score2 esiste una correlazione positiva e significativa. Infatti r=0,978 e il p-value del test vale 0,000.

I MODELLI DI REGRESSIONE MULTIPLA

Regressione multipla

Nell'equazione del modello di regressione multipla sono presenti due o più variabili esplicative.

Ipotizzando una relazione lineare tra la variabile dipendente e ciascuna delle variabili esplicative:

Il modello di regressione multipla con *p* variabili indipendenti

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_p X_{pi} + \epsilon_i \qquad (10.1)$$

dove

β_0 = intercetta

β_1 = inclinazione di Y rispetto alla variabile X_1 tenendo costanti le variabili $X_2, X_3, \ldots, X_p$.

β_2 = inclinazione di Y rispetto alla variabile X_2 tenendo costanti le variabili $X_1, X_3, \ldots, X_p$

β_3 = inclinazione di Y rispetto alla variabile X_3 tenendo costanti le variabili $X_1, X_2, X_4, \ldots, X_p$.

β_p = inclinazione di Y rispetto alla variabile X_p tenendo costanti le variabili $X_1, X_2, X_3, \ldots, X_{p-1}$.

ϵ_i = errore in corrispondenza dell'osservazione i.

Regressione multipla

Consideriamo il caso a due variabili esplicative:

Il modello di regressione multipla con due variabili indipendenti

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \qquad (10.2)$$

dove

β_0 = intercetta

β_1 = inclinazione di Y rispetto alla variabile X_1 tenendo costante la variabile X_2

β_2 = inclinazione di Y rispetto alla variabile X_2 tenendo costante la variabile X_1

ϵ_i = errore in corrispondenza dell'osservazione i

Nel modello di regressione semplice con variabile esplicativa X_1, il coefficiente b_1 rappresenta la variazione di Y in corrispondenza di una variazione unitaria di X_1.

Nel modello di regressione multipla con due variabili esplicative b_1 tiene conto anche degli effetti della variabile X_2. Esso indica di quanto varia Y in seguito ad una variazione unitaria di X_1 rimanendo costante il valore di X_2.

Regressione multipla

Il coefficiente di determinazione $r^2_{Y.12}$ rappresenta la proporzione di variabilità di Y spiegata dalle variabili esplicative.

Il coefficiente di determinazione

Il coefficiente di determinazione è dato dal rapporto tra la somma dei quadrati della regressione e la somma totale dei quadrati.

$$r^2_{Y.12} = \frac{SQR}{SQT} \qquad (10.4)$$

L' r^2 corretto tiene conto dei gradi di libertà e quindi del numero di variabili esplicative del modello. E' utile per confrontare modelli che spiegano la stessa variabile dipendente con un numero diverso di variabili esplicative.

L'r^2 corretto

$$r^2_{adj} = 1 - \left[(1 - r^2_{Y.12\ldots p}) \frac{n-1}{n-p-1} \right] \qquad (10.5)$$

Dove p = numero delle variabili esplicative incluse nel modello.

Test di significatività del modello

La verifica sulla "bontà" del modello di regressione multipla si può tradurre nel seguente sistema di ipotesi:

H_0: $\beta_1 = \beta_2 = 0$ (Non vi è una relazione lineare tra la variabile dipendente e le variabili esplicative.)

H_1: Almeno un $\beta_j \neq 0$ (Vi è una relazione lineare tra la variabile dipendente e almeno una delle variabili esplicative.)

Anche in questo caso la verifica d'ipotesi si basa sul concetto di variabilità ed è realizzata attraverso un test F:

Il test *F* sull'intero modello nel modello di regressione multipla

La statistica F è data dal rapporto tra la media dei quadrati della regressione (MQR) e la media dei quadrati dell'errore (MQE)

$$F = \frac{MQR}{MQE} \qquad (10.6)$$

dove

p = numero delle variabili esplicative nel modello di regressione

F = la statistica test F avente una distribuzione F con p e $n - p - 1$ gradi di libertà

Inferenza sui coefficienti di regressione

Per verificare l'ipotesi che un certo coefficiente di regressione b_k sia uguale a 0 contro l'alternativa che il parametro sia diverso da 0 si ricorre al test t:

Il test *t* per la verifica di ipotesi sull'inclinazione nel modello di regressione multipla

$$t = \frac{b_k}{S_{b_k}} \qquad (10.7)$$

dove

p = numero di variabili esplicative

b_k = inclinazione di Y rispetto alla variabile k tenendo costanti le altre variabili

S_{b_k} = errore standard del coefficiente di regressione b_k

t = statistica test con distribuzione t con $n - p - 1$ gradi di libertà.

La stima intervallare relativa al coefficiente di regressione bk è la seguente:

Stima per intervallo di confidenza per l'inclinazione

$$b_k \pm t_{n-p-1} S_{b_k} \qquad (10.8)$$

Il modello di regressione quadratica

Il modello di regressione quadratica è un caso particolare di modello di regressione multipla, in cui le variabili esplicative sono X e X^2.

In base a questo modello infatti tra Y e X non esiste una relazione lineare bensì una relazione quadratica.

Il modello di regressione quadratica

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \epsilon_i \qquad (10.15)$$

dove

β_0 = Y intercetta

β_1 = coefficiente relativo all'effetto lineare su Y

β_2 = coefficiente relativo all'effetto quadratico su Y

ϵ_i = l'errore casuale in Y corrispondente all'i-esima osservazione

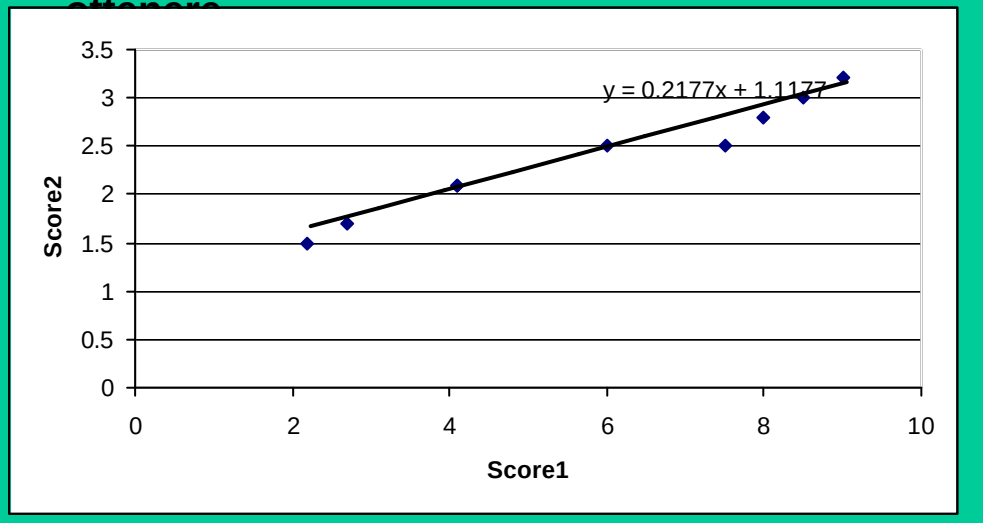
Dopo aver stimato i coefficienti di regressione si perviene alla seguente:

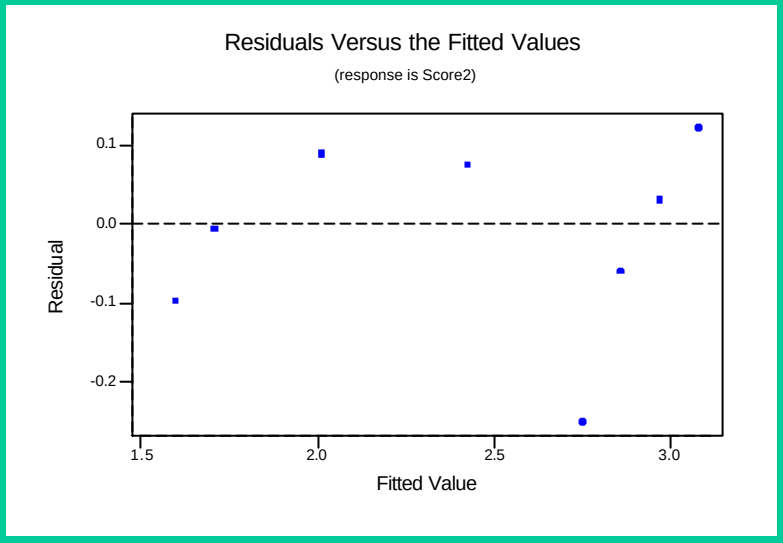
Espressione campionaria del modello di regressione quadratica

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{1i}^2 \qquad (10.16)$$

Il modello di regressione polinomiale

Esempio: un produttore desidera ottenere una misura della qualità di un prodotto ma la procedura è troppo costosa. Decide allora di stimare questa misura (score 2) a partire dall'osservazione di un'altra misura (score 1) più semplice meno costosa da ottenere





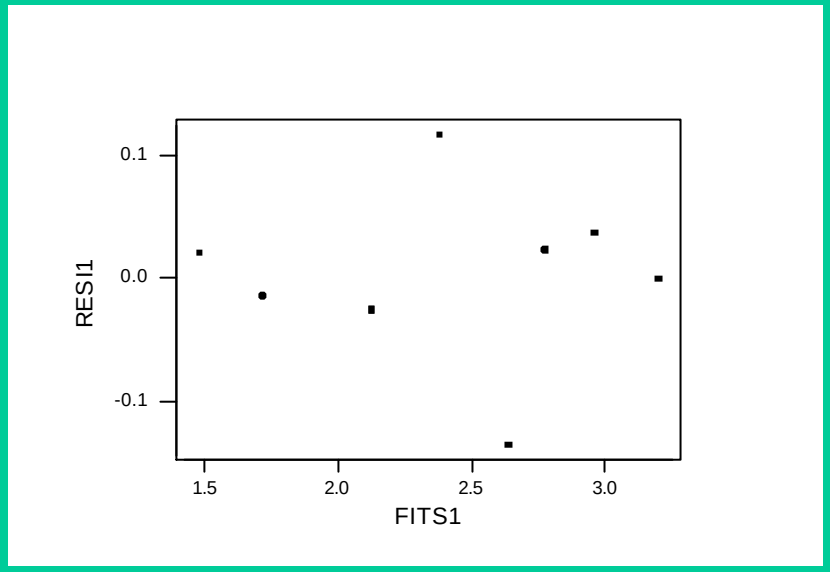
La regressione lineare ci ha restituito una retta di regressione che ben si adatta ai dati:

R^2=0,96 test F e test t riportano un p-value pari a 0 portandoci a rigettare l'ipotesi che il coefficiente di regressione di X non sia significativo.

L'analisi dei residui e della retta di regressione induce il sospetto che il modello sia ulteriormente migliorabile attraverso una specificazione di tipo polinomiale.

Il modello di regressione polinomiale

Specificando un modello polinomiale di terzo grado e procedendo alle stime dei minimi quadrati si ottiene:



Il modello ottenuto è sicuramente migliore del primo perché R^2=0,99. La curva di regressione sembra adattarsi bene ai dati e anche i residui appaiono distribuiti più casualmente intorno allo zero.

La conferma del miglioramento si ha osservando il test t relativo al coefficiente di X^3 (p-value = 0,022). Se tale coefficiente risulta significativamente non nullo il coefficiente di X^2 non è invece significativo (p-value=0,863).

Modelli con variabili dummy

Il modello di regressione stimato risulta $\hat{Y}_i = 50.09 + 16.186X_{1i} + 3.853X_{2i}$

Ovvero

se non si utilizza la materia prima fornita da Tizio vale $\hat{Y}_i = 50.09 + 16.186X_{1i}$

in quanto X_2=0;

Per i prodotti in cui si utilizza la materia prima fornita da Tizio si ha $\hat{Y}_i = 53.943 + 16.186X_{1i}$

in quanto X_2=1.

In pratica il valore della qualità misurato da Y è maggiore di 3,853 se si utilizza materia prima del fornitore Tizio.

In base ai p-value dei test t entrambi i coefficienti risultano significativi.

Per poter usare il modello considerato dobbiamo accertarci che il coefficiente di regressione per X_1 non dipenda dall'uso della materia prima di Tizio. Bisogna studiare allora l'interazione tra le due variabili definita dal prodotto: $X_3 = X_1 \times X_2$

Il test t applicato per verificare la significatività di b_3 riporta un p-value di 0,166 > 0,05 => accetto l'ipotesi H_0: $b_3 = 0$ rispetto all'alternativa H_1: b_3 ? 0. Il termine di interazione non dà un contributo significativo al modello.

La multicollinearità

Uno dei problemi che si può presentare nello studio di un modello di regressione multipla è la multicollinearità → presenza di un elevata correlazione tra le variabili esplicative.

Conseguenze della multicollinearità:

- **le variabili non forniscono informazioni aggiuntive ed è difficile valutare l'effetto di ciascuna di esse**
- **le stime dei coefficienti di regressione presentano elevata variabilità**