Capitolo 8

Probabilità: concetti di base

Concetti primitivi di probabilità

La prova → La prova è un esperimento che ha due o più possibili risultati (es. lancio del dado, estrazione e intervista dell'utente)

L'evento → Per evento si intende uno dei possibili risultati della prova (es. numero due del dado, grado di soddisfazione "buono")

La probabilità → La probabilità è un numero compreso tra 0 ed 1 che misura il grado di incertezza sul verificarsi di un evento

La probabilità è una estensione del concetto di frequenza relativa (semplice o percentuale) dei valori di una certa variabile di interesse in questo senso:

Esempio:

variabile: variazione dichiarata dai pazienti dopo il trattamento espressa in una scala che varia da peggioramento (0), lieve peggioramento (1), nessuna variazione (2), lieve miglioramento (3), miglioramento (4).

Se si osservano un certo numero di pazienti ad esempio 5 pazienti e si rilevano i valori di questa variabile è possibile calcolare la distribuzione di frequenza, ovvero contare il numero di volte che la variabile assume il valore 0 1 2 3 4. Supponiamo che i valori osservati siano 3 3 2 4 4 e costruiamo la distribuzione di frequenza:

Dolore	Frequenza assoluta	Frequenza relativa semplice	Frequenza relativa %
0	0	0	0
1	0	0	0
2	1	1 / 5 = 0,2	20 %
3	2	2 / 5 = 0,4	40 %
4	2	2 / 5 = 0,4	40 %
Totale	5	1	100 %

Il concetto di frequenza è riferito quindi all'osservazione e alla misurazione di una certa variabile in un insieme di unità statistiche.

Però può essere di interesse cercare di quantificare in generale la probabilità che il trattamento determini un netto miglioramento indicata con P(4) con riferimento a tutti i potenziali pazienti e non soltanto ad un campione osservato di pazienti.

Quindi parlo di frequenza relativa (semplice e quindi variabile tra 0 e 1 oppure %) dell'evento 'miglioramento' quando osservo un campione di pazienti, parlo invece di probabilità (variabile tra 0 e 1 oppure espressa in %) di 'miglioramento' quando considero la popolazione virtualmente infinita di pazienti trattabili.

Concezioni della probabilità

Frequentista

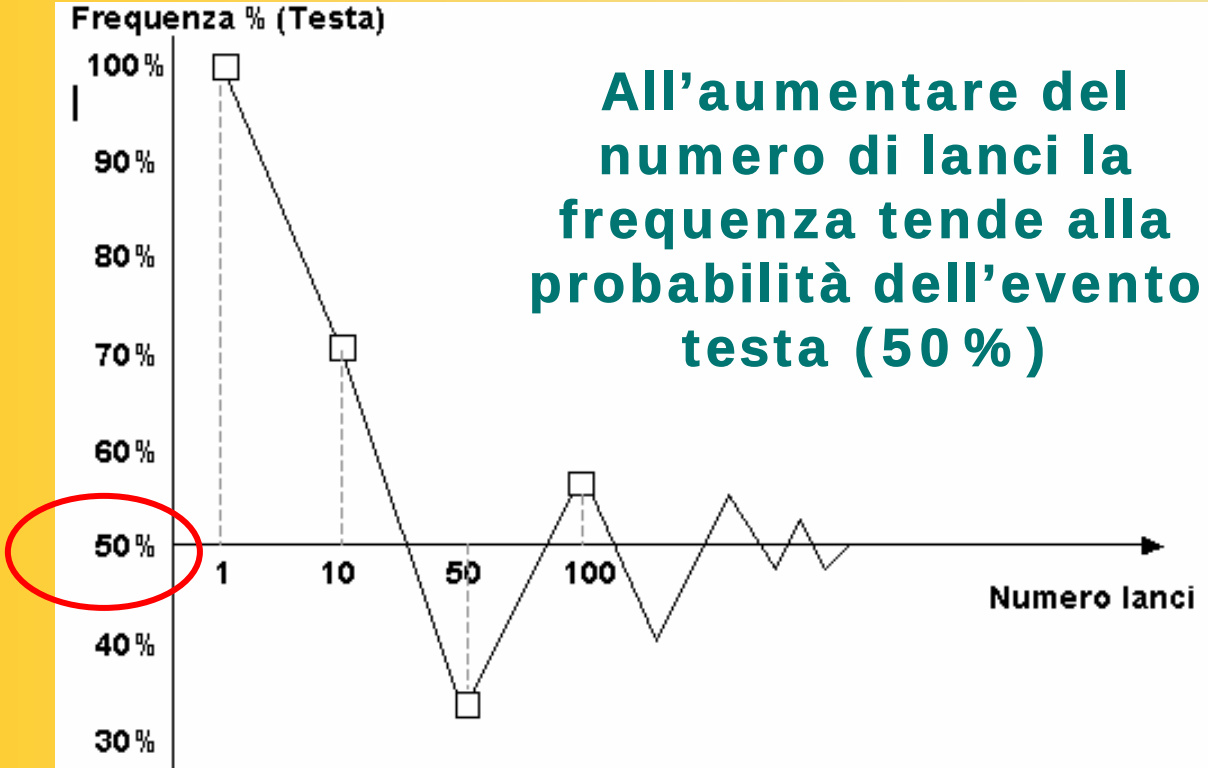
Basata sul Postulato empirico del caso: In un gruppo di prove, ripetute più volte nelle stesse condizioni, ciascuno degli eventi possibili compare con una frequenza quasi eguale alla sua probabilità; generalmente l'approssimazione migliora quando il numero delle prove cresce.

Faccia della moneta	Freq. assoluta	Freq. relativa %
Testa	1	100
Croce	0	0
Totale	1	100

Faccia della moneta	Freq. assoluta	Freq. relativa %
Testa	7	70
Croce	3	30
Totale	10	100

Faccia della moneta	Freq. assoluta	Freq. relativa %
Testa	17	34
Croce	33	66
Totale	50	100

Faccia della moneta	Freq. assoluta	Freq. relativa %
Testa	56	56
Croce	44	44
Totale	100	100



All'aumentare del numero di lanci la frequenza tende alla probabilità dell'evento testa (50 %)

Capitolo 9

Variabili casuali e Distribuzioni di Probabilità

Variabili casuali discrete

$P(X=x_i)$ → Probabilità che la v.c. X assuma il valore x_i

La funzione di probabilità di una variabile casuale discreta X associa ad ognuno dei valori x_i la corrispondente probabilità $P(X=x_i)$

Proprietà →

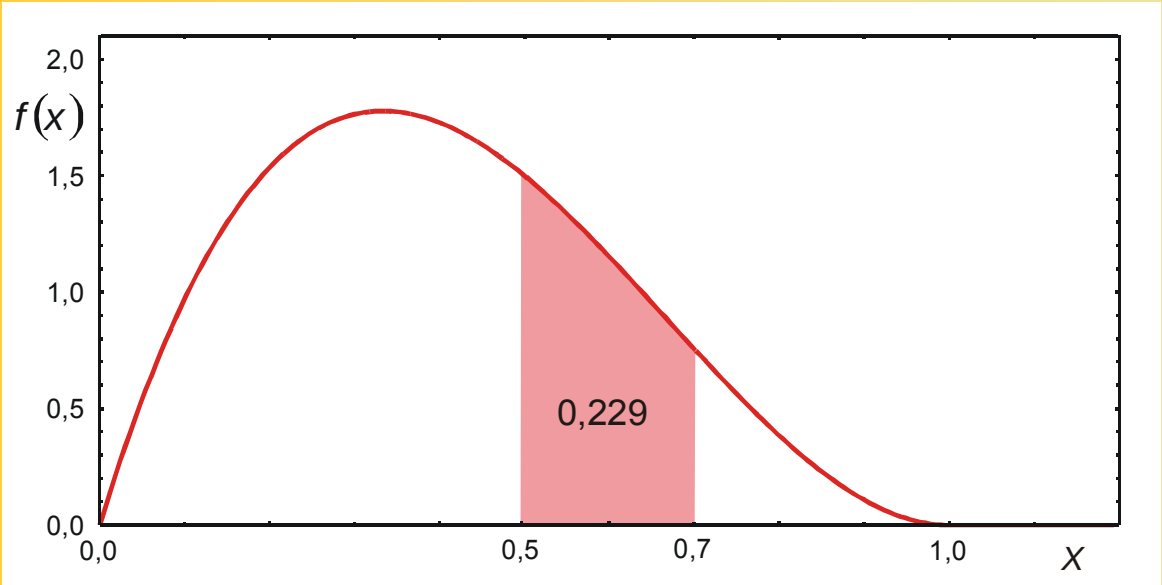
$$\sum_i P(x_i) = 1$$

→

$$P(x_i) \geq 0$$

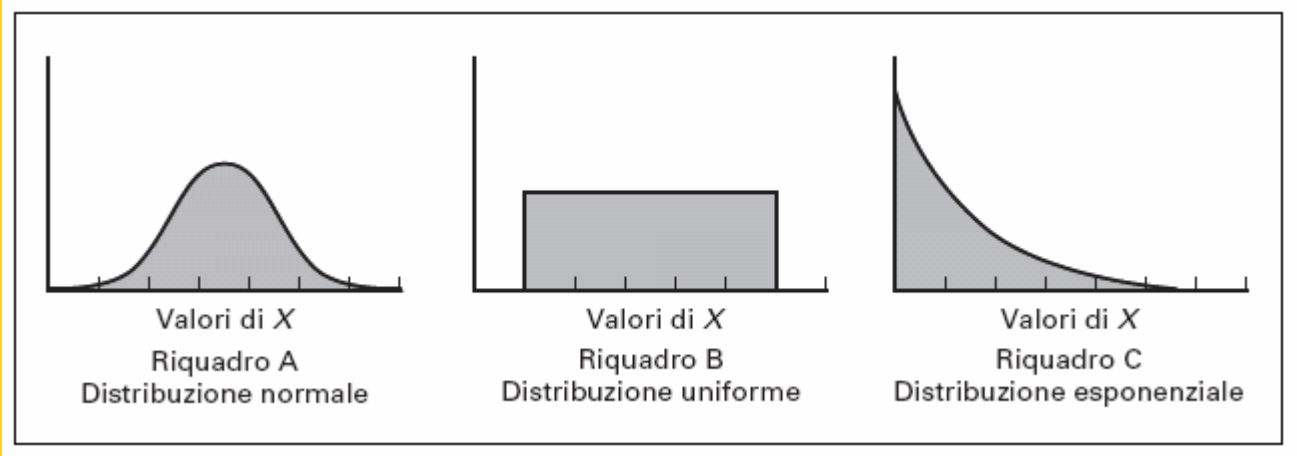
Variabili casuali continue

Chiameremo **Funzione di densità**, la funzione matematica f(x) per cui l'area sottesa alla funzione, corrispondente ad un certo intervallo, è uguale alla probabilità che X assuma un valore in quell'intervallo.



- Una **funzione di densità di probabilità continua** è un modello che definisce analiticamente come si distribuiscono i valori assunti da una variabile aleatoria continua
- Quando si dispone di un'espressione matematica adatta alla rappresentazione di un fenomeno continuo, siamo in grado di calcolare la probabilità che la variabile aleatoria assuma valori compresi in intervalli
- I modelli continui hanno importanti applicazioni in ingegneria, fisica, economia e nelle scienze sociali

- Alcuni tipici fenomeni continui sono l'altezza, il peso, le variazioni giornaliere nei prezzi di chiusura di un'azione, il tempo che intercorre fra gli arrivi di aerei presso un aeroporto, il tempo necessario per servire un cliente in un negozio
- La figura rappresenta graficamente tre funzioni di densità di probabilità: normale, uniforme ed esponenziale



Valore atteso e varianza di una v.c.

Il valore medio di una v.c. X, è indicato con:

$$E(X)$$

La varianza di una variabile casuale X è indicata con:

$$V(X) \text{ o } \sigma^2$$

La deviazione standard è indicata con:

$$SD(X) = \sigma = \sqrt{V(X)}$$

Distribuzione Normale

La distribuzione **normale** (o distribuzione *Gaussiana*) è la distribuzione continua più utilizzata in statistica.

La distribuzione normale è importante in statistica per tre motivi fondamentali:

1. Diversi fenomeni continui sembrano seguire, almeno approssimativamente, una distribuzione normale.
2. La distribuzione normale può essere utilizzata per approssimare numerose distribuzioni di probabilità discrete.
3. La distribuzione normale è alla base dell'*inferenza statistica classica* in virtù del *teorema del limite centrale.*

Distribuzione Normale

Funzione di densità di probabilità normale

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(1/2)[(X-\mu)/\sigma]^2}$$

dove e = costante matematica approssimata da 2.71828

π = costante matematica approssimata da 3.14159

μ = valore atteso della popolazione

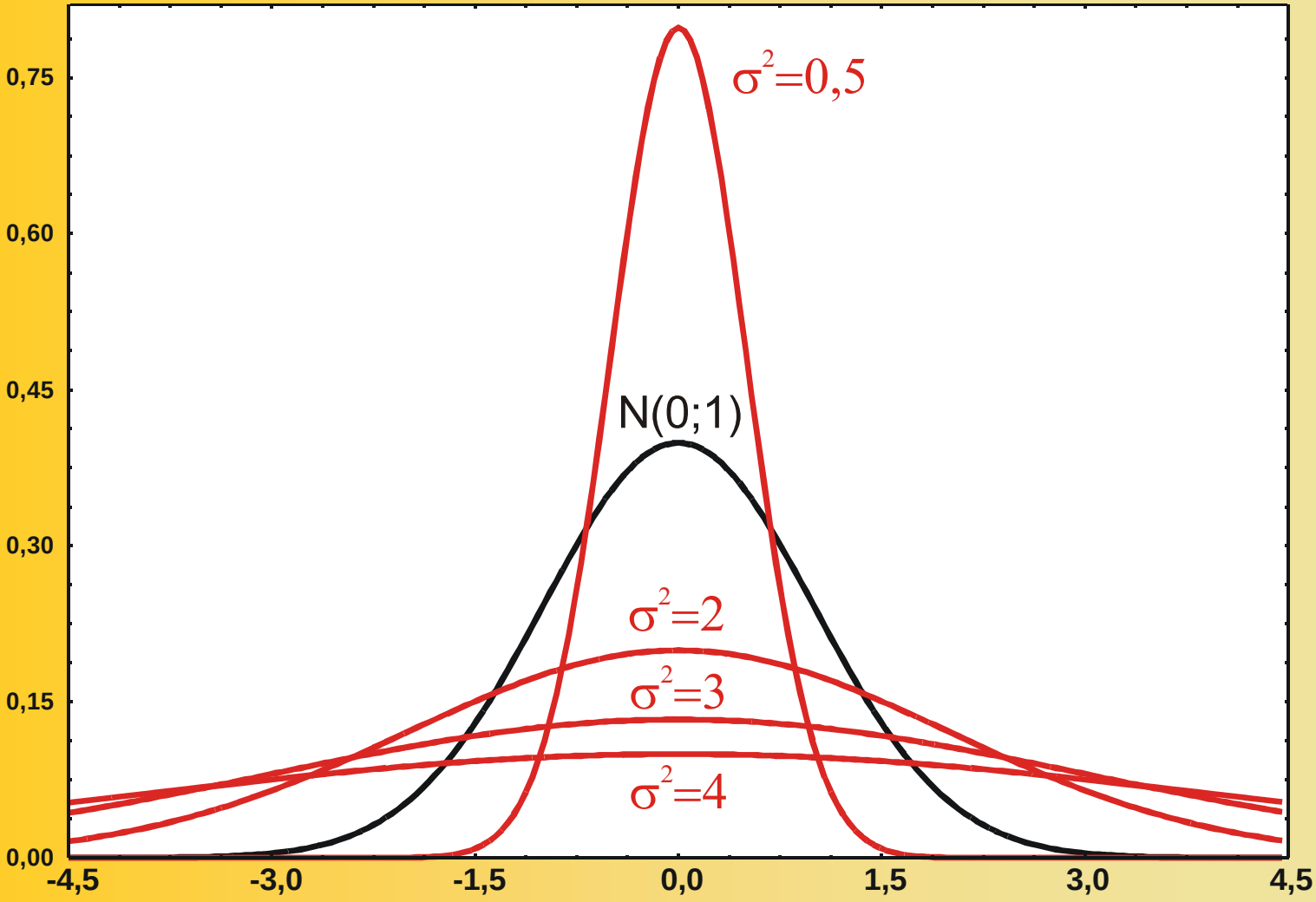
σ = scarto quadratico medio della popolazione

X = valori assunti dalla variabile aleatoria, $-\infty < X < +\infty$

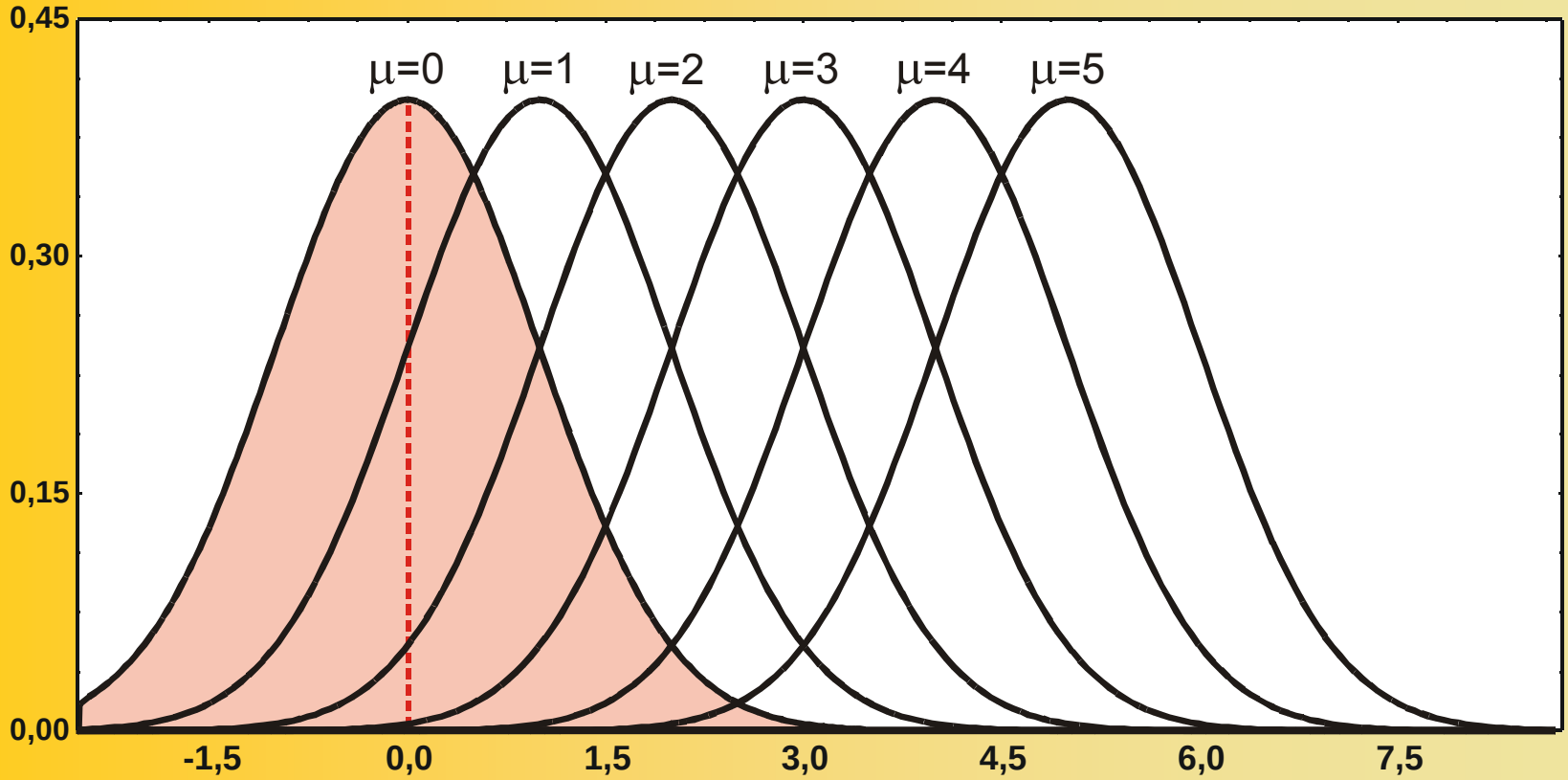
Notiamo che, essendo e e π delle costanti matematiche, le probabilità di una distribuzione normale dipendono soltanto dai valori assunti dai due parametri μ e σ.

Specificando particolari combinazioni di μ e σ, otteniamo differenti distribuzioni di probabilità normali.

Distribuzione Normale



Distribuzione Normale



Poiché esiste un numero infinito di combinazioni dei parametri μ e σ, introduciamo ora una formula di trasformazione delle osservazioni, chiamata standardizzazione, che consente di trasformare una generica variabile aleatoria normale in una variabile aleatoria normale standardizzata per la quale sono state derivate delle tavole che consentono di calcolare la probabilità associata a qualsiasi intervallo e viceversa.

La standardizzazione

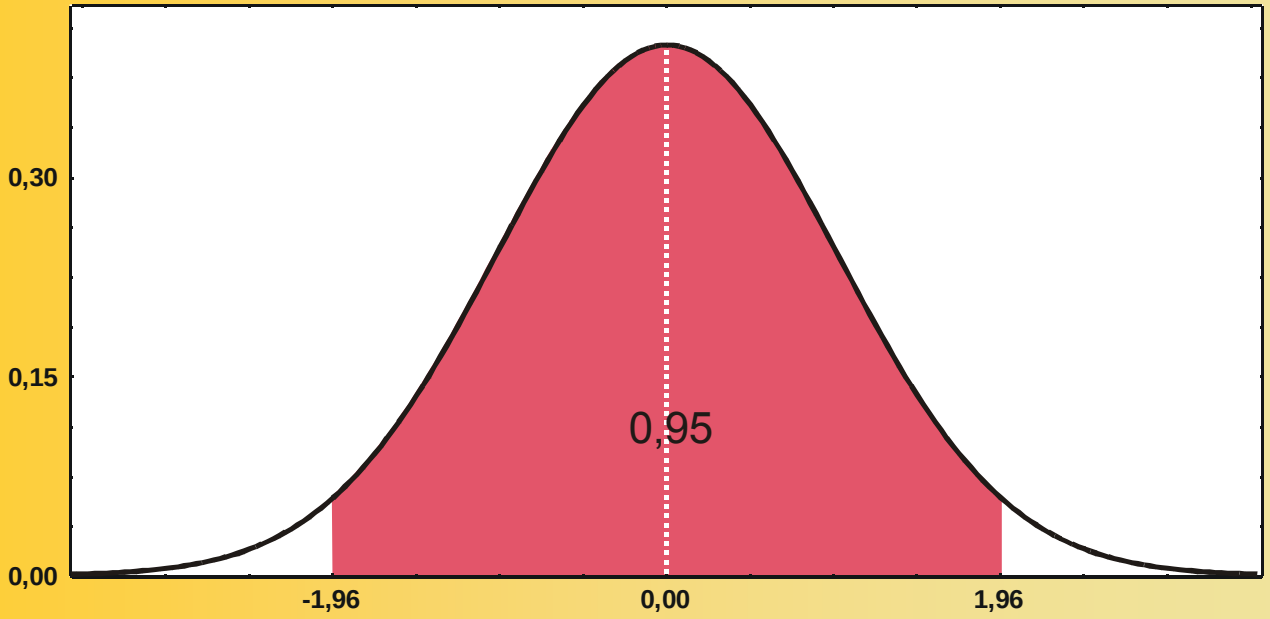
$$Z = \frac{X - \mu}{\sigma}$$

Z è la variabile ottenuta sottraendo ad *X* il suo valore atteso μ e rapportando il risultato allo scarto quadratico medio, σ.

La v.c. Normale Standardizzata Z

Se la variabile casuale X ha una distribuzione normale con parametri μ e σ^2, allora $Z= (X- \mu)/ \sigma$ è ancora una v.c. Normale con media nulla e varianza unitaria.

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$



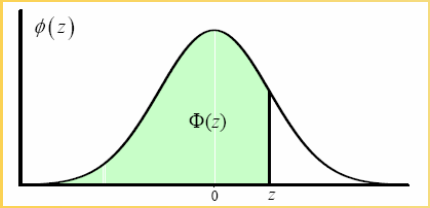
La distribuzione normale standardizzata

La variabile aleatoria standardizzata *Z* ha la caratteristica di avere valore atteso nullo ($\mu=0$) e scarto quadratico medio pari a uno ($\sigma=1$).

Le tavole della distribuzione normale standardizzata consentono di calcolare le probabilità associate ad intervalli e viceversa. In particolare:

$$P(-1{,}96 \leq Z \leq +1{,}96)=0{,}95 \ (95\%)$$

$$P(-2{,}58 \leq Z \leq +2{,}58)=0{,}99 \ (99\%)$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964

Campionamento e inferenza statistica

Nelle indagini campionarie si pone il problema di come generalizzare le informazioni tratte dal campione alla totalità della popolazione.

- **selezione di un campione rappresentativo della popolazione;**
- **valutare il grado di attendibilità delle generalizzazioni.**

Si parla di **inferenza statistica** con riferimento all'insieme delle procedure statistiche che consentono di:

- **estendere i risultati ottenuti da un campione alla popolazione;**
- **controllare e quantificare il grado di incertezza delle generalizzazioni in termini probabilistici.**

Nella fase inferenziale le misure introdotte finora a fini descrittivi (frequenze, medie, misure di variabilità, tassi, …) diventano oggetto di inferenza.

Uno degli scopi principali dell'analisi inferenziale consiste nell'uso di statistiche calcolate sui dati campionari, (come la media campionaria, la deviazione standard campionaria) per ottenere STIME o VERIFICARE IPOTESI sui corrispondenti parametri della popolazione da cui è stato tratto il campione.

Esempio di problema di STIMA: Un'azienda produttrice di cereali vuole stimare il contenuto medio in grammi delle scatole prodotte.

Esempio di problema di VERIFICA DI IPOTESI: l'azienda vuole verificare che il peso medio delle scatole contenenti i cereali sia pari a 368 grammi.

Inferenza statistica: la stima

La procedura di **stima** consiste nell'attribuzione di valori a quantità incognite (frequenze, medie, misure di variabilità, …) della popolazione, sulla base delle informazioni campionarie.

I criteri di stima sono di due tipi:

- **stima puntuale**: si attribuisce alla quantità incognita un unico valore;
- **stima per intervallo (intervalli di confidenza)**: si attribuisce alla quantità incognita un insieme di valori possibili costruendo un intervallo entro cui si ha 'fiducia' che sia compreso il valore vero della quantità incognita.

Stimatore puntuale: singola **statistica** che viene usata per stimare il vero valore di un **parametro** della popolazione. Ad esempio la media campionaria è uno stimatore puntuale della media della popolazione μ, la varianza campionaria è uno stimatore puntuale della varianza della popolazione σ^2, ecc.

Esempio: Si vuole stimare il peso medio in grammi delle scatole di cereali:

Informazione campionaria (media): 370 gr

Stima puntuale

370 gr

Stima intervallare (intervallo di confidenza)

(365 gr, 375 gr)

Intervalli di confidenza:

1. Intervallo di confidenza per la media (popolazioni normali e varianza nota)
2. Intervallo di confidenza per la media (popolazioni normali e varianza non nota)
3. Intervallo di confidenza per la media (popolazioni non normali)

1. Intervallo di confidenza per la media (varianza nota)

Dato un campione casuale estratto da una popolazione Normale con media ignota μ e varianza nota σ^2, l'intervallo di confidenza per la media della popolazione al livello di confidenza $1-\alpha$ è:

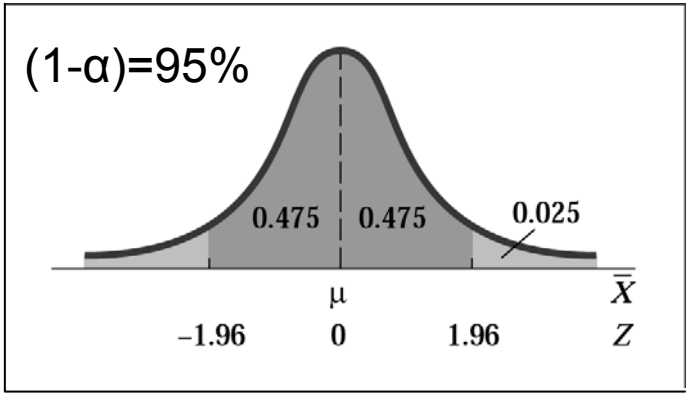
$$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

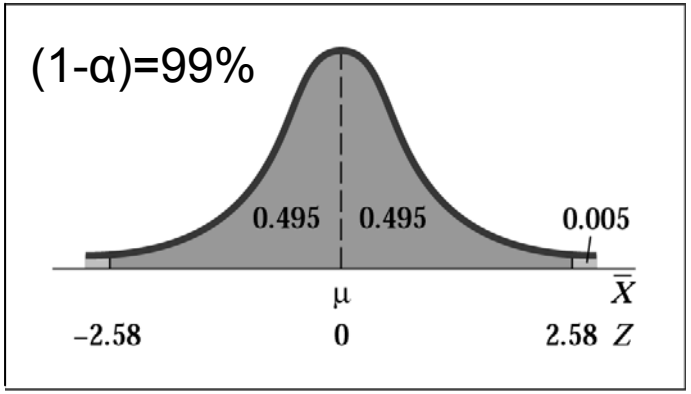
La probabilità che il parametro appartenga all'intervallo è detta **livello di confidenza**, generalmente indicato con (1−α)% dove α è la probabilità che il parametro si trovi al di fuori dell'intervallo di confidenza

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

Il livello di confidenza è fissato dal ricercatore e generalmente è pari al 90%, 95% e 99%. In alcuni casi risulta desiderabile un grado di certezza maggiore, ad es. del 99%, ed in altri casi possiamo accettare un grado minore di sicurezza, ad es. del 90%.

Il valore $Z_{\alpha/2}$ di Z che viene scelto per costruire un intervallo di confidenza è chiamato **valore critico**. A ciascun livello di confidenza (1−α) corrisponde un diverso valore critico.





Intervallo di confidenza stimato per la media (varianza nota)

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

Esempio

Siano $n = 10$ $\sigma^2 = 9$ $1 - \alpha = 0{,}99$ $\bar{x} = 4{,}924$

Dalle tavole della Normale standardizzata si ottiene:

$$z_{\alpha/2} = z_{0{,}005} = 2{,}576$$

e quindi l'intervallo di confidenza al 99%(IC 99%) sarà:

$$\left[4{,}924 \pm 2{,}576\frac{\sqrt{9}}{\sqrt{10}}\right] \Rightarrow [2{,}4802\,;\,7{,}3678]$$

Limite o estremo inferiore

Limite o estremo superiore

Se potessimo ripetere il campionamento infinite volte, nel 99% dei casi (per il 99% dei campioni) l'intervallo di confidenza includerebbe il valore vero della media incognita. Essendo elevata questa probabilità si ha fiducia che l'intervallo **stimato calcolato per il campione osservato**, contenga il valore incognito della media della popolazione.

Con riferimento al processo industriale di riempimento di scatole di cereali ci troviamo nel concreto a risolvere un esempio del tipo: si assuma che il peso X delle scatole sia $X \sim N(\mu; 15^2)$. Dato un campione casuale di $n=25$ scatole con peso medio 362.3 grammi si vuole costruire un intervallo di confidenza al 95% per μ. Applichiamo la formula per l'intervallo stimato:

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = \left[362.3 - 1.96\frac{15}{\sqrt{25}}, 362.3 + 1.96\frac{15}{\sqrt{25}}\right]$$

Nel caso specifico si ottiene (356.42; 368.18).

Se potessimo ripetere il campionamento infinite volte, nel 95% dei casi (per il 95% dei campioni) l'intervallo di confidenza includerebbe il valore vero della media incognita. Essendo elevata questa probabilità si ha fiducia che l'intervallo stimato calcolato per il campione osservato, contenga il valore incognito della media della popolazione.

La lunghezza dell'intervallo di confidenza si ricava dalla differenza tra estremo superiore e estremo inferiore:

$$\text{Lunghezza} = 2 z_{\alpha/2} \left(\sigma / \sqrt{n} \right)$$

Dipende da:

1. **la dimensione del campione**
2. **il livello di confidenza**
3. **la varianza della popolazione**

Intervenendo sulla dimensione del campione o sul livello di confidenza si può aumentare o diminuire la lunghezza dell'intervallo.

Minore è la lunghezza dell'intervallo, maggiore è la capacità informativa dell'intervallo.

Esempio:

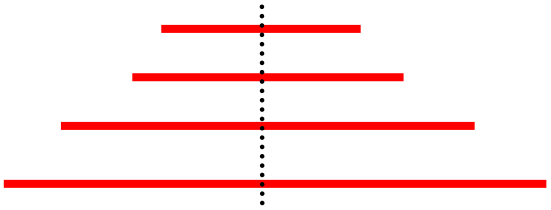
Fissato $1 - \alpha$

$n = 100$

$n = 70$

$n = 50$

$n = 10$



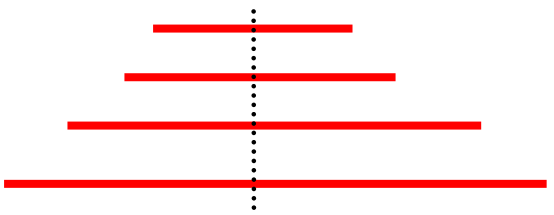
Fissato n

$1 - \alpha = 0,85$

$1 - \alpha = 0,90$

$1 - \alpha = 0,95$

$1 - \alpha = 0,99$



1. Intervallo di confidenza per la media (varianza nota)

$$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

2. Intervallo di confidenza per la media (varianza ignota)

$$\left[\overline{X} - t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}\right]$$

3. Intervallo di confidenza per la media (popolazioni non Normali, *n* sufficientemente grande, Teorema Limite Centrale)

$$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

$$\left[\overline{X} - t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}\right]$$

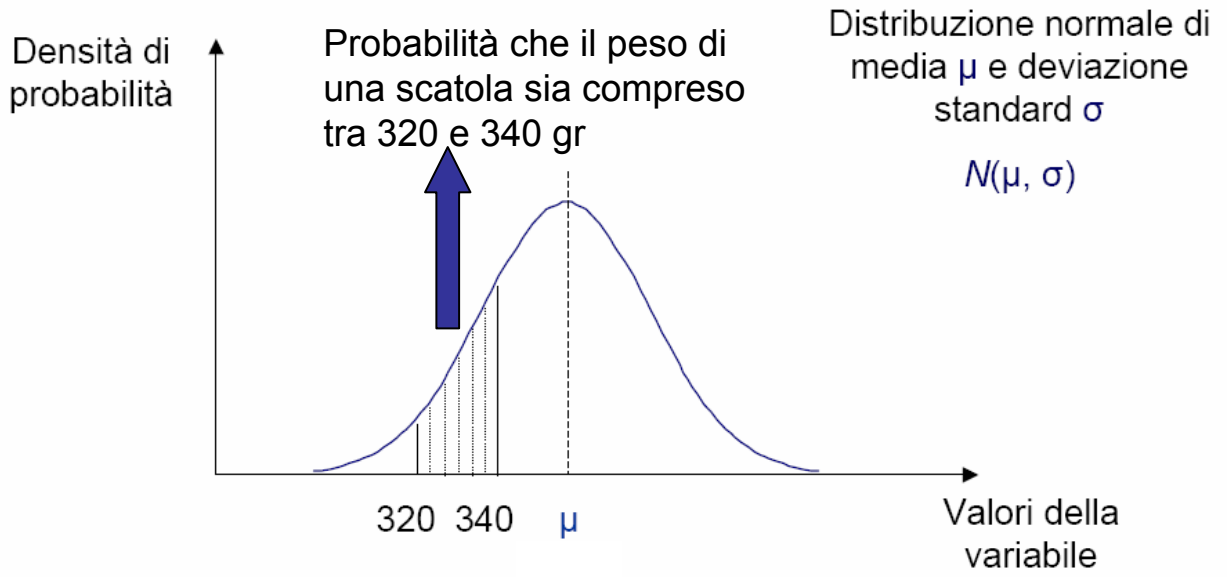
Perché queste formule?

La distribuzione campionaria della media campionaria

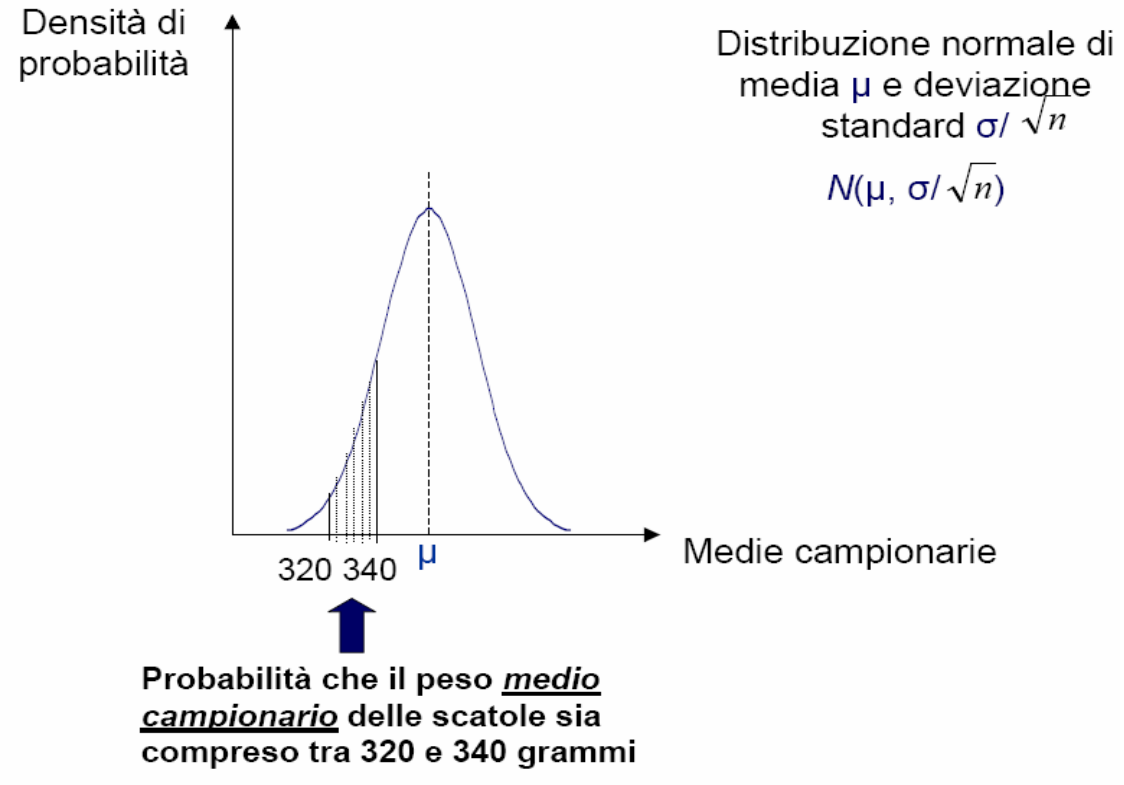
- La media campionaria – la media aritmetica degli elementi di un campione – viene utilizzata per stimare la media della popolazione
- La **distribuzione (campionaria) della media campionaria** è la distribuzione di tutte le possibili medie che osserveremmo se procedessimo all'estrazione di tutti i possibili campioni di una certa ampiezza fissata *n*.

Se un campione è estratto da una popolazione **normale** con media μ e scarto quadratico medio σ, la media campionaria ha distribuzione **normale** indipendentemente dall'ampiezza campionaria n, con media μ e scarto quadratico medio $\sigma / \sqrt{n}$.

Distribuzione della variabile casuale peso in grammi delle scatole di cereali nella popolazione



Distribuzione delle medie campionarie calcolate per tutti i campioni di ampiezza pari a *n*



In via ipotetica, per usare le statistiche campionarie con lo scopo di stimare i parametri della popolazione, dovremmo analizzare tutti i campioni di una certa ampiezza prestabilita che possono essere estratti da questa. Nella pratica, da una popolazione viene estratto a caso un solo campione, ma anche se non sappiamo quanto la media dell'unico campione osservato sia vicina alla media della popolazione, siamo sicuri che la media delle medie di tutti i campioni che potremmo selezionare coincide con la media della popolazione μ.

La media campionaria è caratterizzata da una minore variabilità rispetto ai dati originali. Le medie campionarie saranno quindi caratterizzate, in generale, da valori meno dispersi rispetto a quelli che si osservano nella popolazione. Lo scarto quadratico medio della media campionaria, detto **errore standard della media**, quantifica la variazione della media campionaria da campione a campione:

$$\sigma_{\bar{X}} = \sigma / \sqrt{n}$$

L'errore standard della media è uguale allo scarto quadratico medio della popolazione diviso $\sqrt{n}$.

La distribuzione campionaria della media campionaria

Sinora abbiamo analizzato la distribuzione della media campionaria nel caso di una popolazione con distribuzione normale. Tuttavia, si presenteranno spesso casi in cui la distribuzione della popolazione non è normale. In questi casi è utile riferirsi ad un importante teorema della statistica, il teorema del limite centrale, che consente di dire qualcosa sulla distribuzione della media campionaria anche nel caso in cui una popolazione non abbia distribuzione normale.

Il teorema del limite centrale

Quando l'ampiezza del campione casuale diventa sufficientemente grande, la distribuzione della media campionaria può essere approssimata dalla distribuzione normale. E questo indipendentemente dalla forma della distribuzione dei singoli valori della popolazione.

Si tratta, allora, di stabilire cosa si intende per “sufficientemente grande”, problema ampiamente affrontato dagli statistici. Come regola di carattere generale, molti sono concordi nell’affermare che quando il campione raggiunge un’ampiezza pari almeno a 30, la distribuzione della media campionaria è approssimativamente normale. Tuttavia, il teorema del limite centrale può essere applicato anche con campioni di ampiezza inferiore se si sa che la distribuzione della popolazione ha alcune caratteristiche che la avvicinano alla normale (ad esempio, quando è simmetrica).

Il teorema del limite centrale svolge un ruolo cruciale in ambito inferenziale, in quanto consente di fare inferenza sulla media della popolazione senza dover conoscere la forma specifica della distribuzione della popolazione.

Sulla base dei risultati ottenuti per le distribuzioni note (es. la normale, l'uniforme, l'esponenziale) possiamo trarre alcune conclusioni in merito al teorema del limite centrale:

- Per la maggior parte delle popolazioni, indipendentemente dalla forma della loro distribuzione, la distribuzione della media campionaria è approssimativamente normale, purché si considerino campioni di almeno 30 osservazioni.
- Se la distribuzione della popolazione è abbastanza simmetrica, la distribuzione della media campionaria è approssimativamente una normale, purché si considerino campioni di almeno 5-15 osservazioni.
- Se la popolazione ha una distribuzione normale, la media campionaria è distribuita secondo la legge normale, indipendentemente dall'ampiezza del campione.

Dalla distribuzione campionaria della media campionaria alla costruzione dell'intervallo di confidenza per la media

Standardizzazione della media campionaria per passare ad una variabile normale di media 0 e varianza 1:

$$\bar{X} \sim N\left(\mu, \sigma/\sqrt{n}\right) \quad \Rightarrow \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

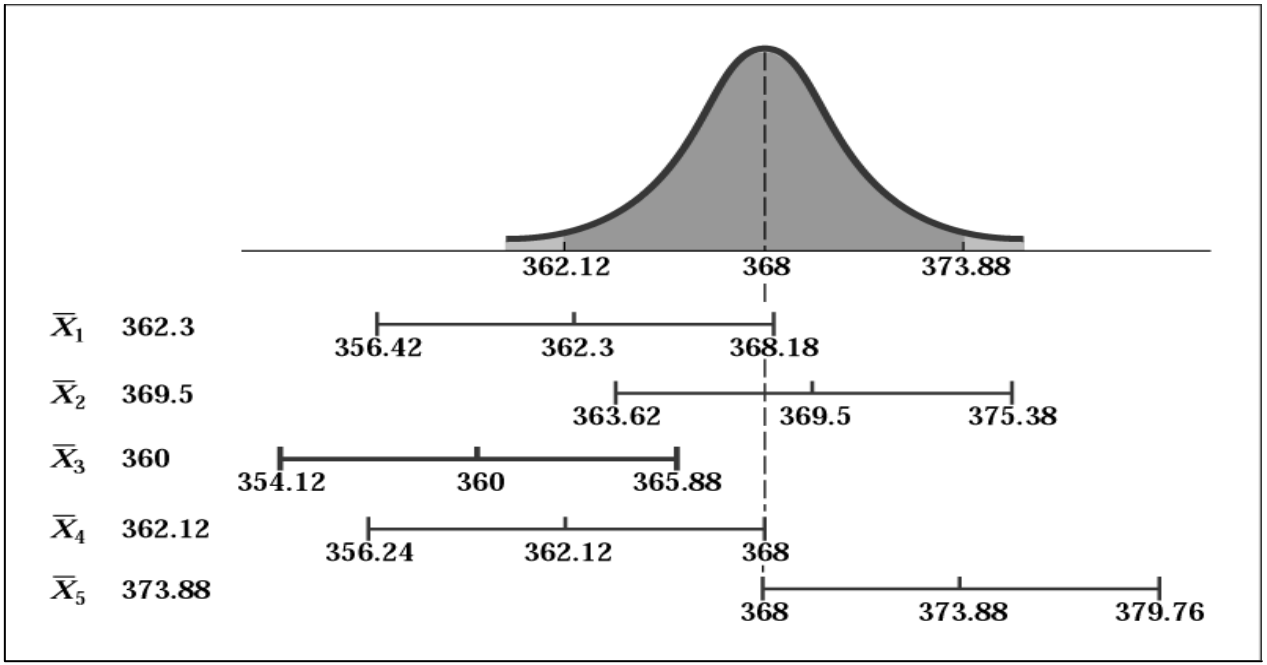
$$P\left(-z_{\alpha/2} \le Z \le +z_{\alpha/2}\right) = 1-\alpha$$

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

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

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

Nell'esempio del peso delle scatole di cereali, ipotizziamo che μ sia uguale a 368 gr. Per comprendere a fondo il significato della stima per intervallo e le sue proprietà è utile fare riferimento all'ipotetico insieme di tutti i possibili campioni di ampiezza n che è possibile ottenere e alla distribuzione campionaria della media.



Osserviamo che per alcuni campioni la stima per intervallo al 95% di μ è corretta, mentre per altri non lo è.

Se potessimo ripetere il campionamento infinite volte, nel 95% dei casi (per il 95% dei campioni) l'intervallo di confidenza includerebbe il valore vero della media incognita.

Capitolo 13

- La verifica di ipotesi e il test statistico
- Regione di accettazione e rifiuto
- Test per la media
- Il *p*-value
- Errori di I e II tipo
- Potenza di un test

La verifica di ipotesi è una procedura inferenziale che consiste nel fare un'ipotesi su una quantità incognita della popolazione (parametro) e nel decidere sulla base di campione casuale (probabilistico) (per mezzo di una statistica campionaria) se essa è accettabile o meno.

Es. Il processo produttivo di riempimento delle scatole di cereali può essere considerato appropriato (sotto controllo) se il **peso medio** μ (parametro di interesse: media) delle scatole è di 368 grammi. Possiamo ritenere vera questa ipotesi?

Es. Si ipotizza che una nuova terapia possa ridurre il tempo medio di risoluzione di una infezione (parametro di interesse: media). Possiamo ritenere vera questa ipotesi?

Es. Si ipotizza che la soddisfazione per il proprio lavoro (parametro di interesse: proporzione o frequenza) sia diversa in base alle classi di età. Possiamo ritenere vera questa ipotesi?

Obiettivo:

Attraverso un campione di osservazioni stabilire, con un certo grado di attendibilità, se rifiutare o meno l'ipotesi di interesse. Il problema quindi è di prendere una decisione sulla base dei dati campionari.

Es. Si effettuano 25 misurazioni del peso delle scatole prodotte. La media aritmetica campionaria (statistica campionaria) risulta pari a 372,5 gr.

Valore osservato della statistica campionaria

$\bar{x} = 372{,}5\ gr$

Ipotesi sul parametro

$\mu = 368\ gr$

Test di Ipotesi

Differenza statisticamente non significativa

La differenza riscontrata è solo casuale, legata al campionamento. La media del peso può ritenersi uguale a 368 gr. Il processo produttivo è sotto controllo.

Decisione

Differenza statisticamente significativa

La differenza riscontrata non è casuale. La media del peso non può ritenersi uguale a 368 gr. Il processo produttivo non è sotto controllo.

Il test statistico e il sistema di ipotesi

Si definisce **test di ipotesi** il procedimento che consente di rifiutare o non rifiutare l'ipotesi statistica.

L'impostazione data da J.Neyman e E.S.Pearson, nota come test d'ipotesi parametrico, prevede la formulazione di un'ipotesi nulla e un'ipotesi alternativa.

L'ipotesi statistica da verificare viene detta **ipotesi nulla** ed è indicata con H_0.
A fronte dell'ipotesi nulla risulta definita l'**ipotesi alternativa** indicata con H_1.

In generale l'ipotesi nulla ipotizza **l'assenza di differenze** del parametro rispetto ad un valore**, l'assenza di differenze** significative tra gruppi o **l'assenza di relazioni** tra variabili, a differenza dell'ipotesi alternativa che ipotizza l'esistenza di una differenza o di una relazione. L'ipotesi nulla così definita corrisponde ad un atteggiamento conservatore del ricercatore. L'ipotesi nulla è preesistente all'osservazione dei dati campionari, ritenuta vera fino a prova contraria. H_0 corrisponde all'ipotesi che si vorrebbe respingere attraverso l'indagine, ma che comunque si continua a ritenere vera a meno che non risulti una **forte evidenza contraria.**

1º passo: Formulazione delle ipotesi

Es. Il processo produttivo di riempimento delle scatole di cereali può essere considerato appropriato (sotto controllo) se il **peso medio μ** (parametro di interesse: media) delle scatole è di 368 grammi. Possiamo ritenere vera questa ipotesi?

Sistema di ipotesi

$$\begin{cases} H_0: \boldsymbol{\mu = 368\ gr} \\ H_1: \boldsymbol{\mu \neq 368\ gr} \end{cases} \rightarrow \text{Test di ipotesi bidirezionale o a due code}$$

Oppure

$$\begin{cases} H_0: \boldsymbol{\mu = 368\ gr} \\ H_1: \boldsymbol{\mu < 368\ gr} \end{cases}$$

Oppure

$$\begin{cases} H_0: \boldsymbol{\mu = 368\ gr} \\ H_1: \boldsymbol{\mu > 368\ gr} \end{cases}$$

Test di ipotesi unidirezionale o a una coda

2º passo: Scelta di una statistica test

La statistica test è una statistica campionaria (o una funzione di questa) la cui distribuzione campionaria (considerando l'insieme di tutti i possibili campioni di una data ampiezza *n* estraibili dalla popolazione) deve essere nota (es. distribuzione normale) e completamente specificata sotto l'ipotesi nulla (ad esempio per la distribuzione normale, si specificano i valori di μ e σ).

Es. Nell'esempio del processo produttivo di riempimento delle scatole di cereali l'ipotesi riguarda il **peso medio μ** delle scatole. Una statistica test appropriata per il test d'ipotesi è la media campionaria $\overline{X}$ o la media campionaria standardizzata Z di cui conosciamo la distribuzione campionaria. Se X è normale (con σ noto) o se siamo nelle condizioni di applicabilità del teorema limite centrale:

$$\overline{X} \sim N\left(\mu, \sigma/\sqrt{n}\right) \qquad Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Ponendo ad esempio $\sigma = 15$ gr e ricordando che $n = 25$

$$\overline{X} \sim N\left(\mu, 15/\sqrt{25}\right) \qquad Z = \frac{\overline{X}-\mu}{15/\sqrt{25}} \sim N(0,1)$$

3º passo: Regione di accettazione e regione di rifiuto

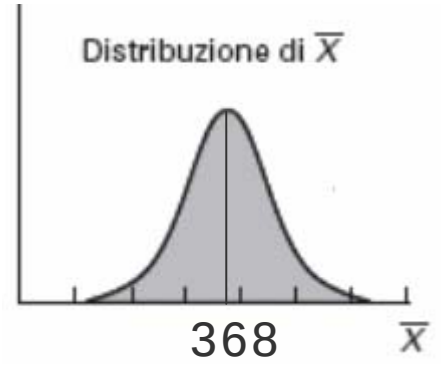
Il test statistico procede **ipotizzando vera** l'ipotesi nulla.

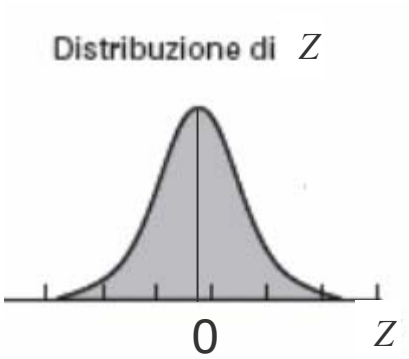
Es. Nell'esempio del processo produttivo di riempimento delle scatole di cereali, l'ipotesi nulla stabilisce che μ = 368 gr. Sotto questa ipotesi la distribuzione campionaria della statistica test sarà:

$$\bar{X} \sim N\left(368, 15/\sqrt{25}\right)$$

o in modo equivalente

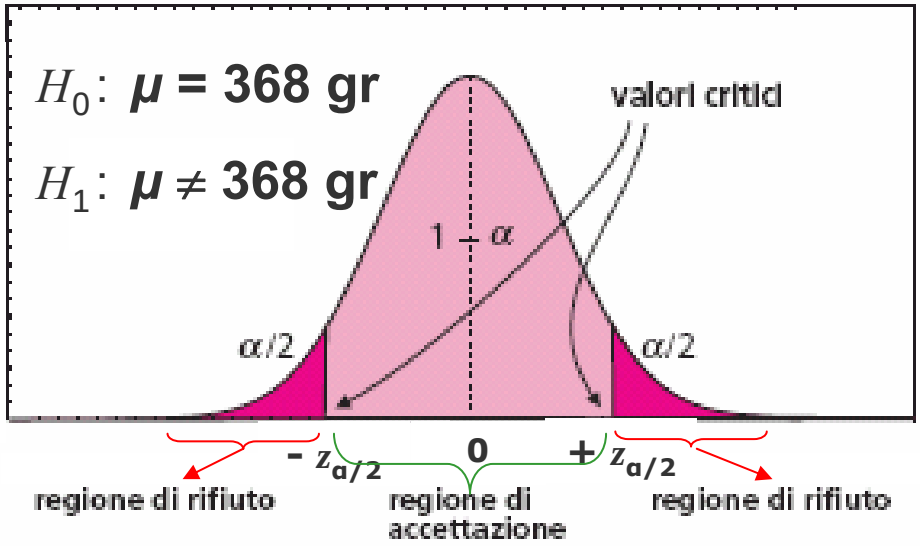
$$Z = \frac{\bar{X} - 368}{15/\sqrt{25}} \sim N(0,1)$$

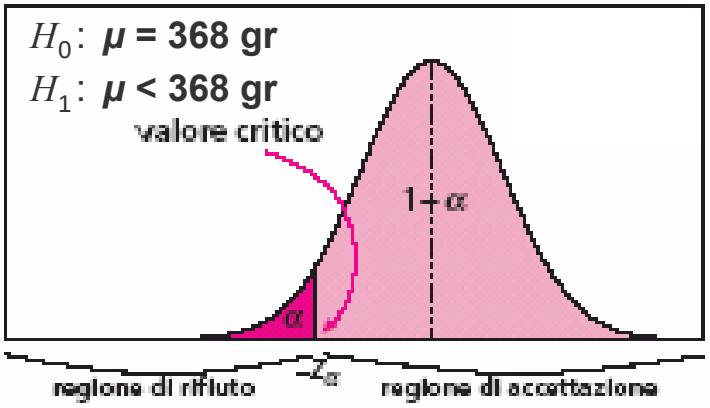


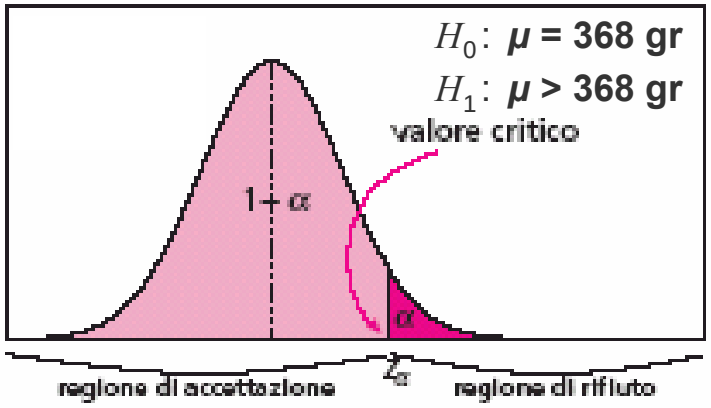


Su questa distribuzione vengono definite due regioni, la **regione di accettazione** e la **regione del rifiuto.**

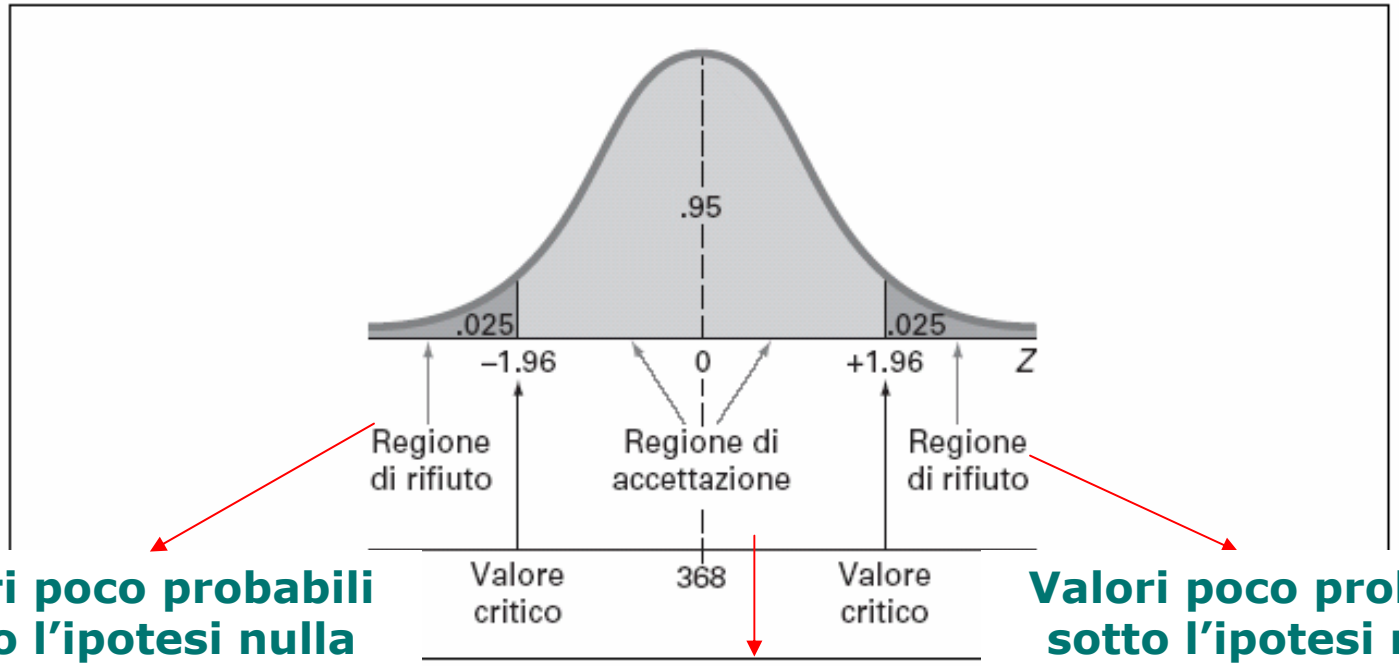
Le due regioni sono definite in corrispondenza dei cosiddetti valori critici. Tali valori dipendono dal livello di significatività α: maggiore è il suo valore, più ampia sarà la regione di rifiuto. La definizione delle due regioni dipende inoltre dal tipo di test (bidirezionale o unidirezionale). Considerando la distribuzione della statistica test Z:







Il livello di significatività α è generalmente posto pari al 5%, al 1% o al 0,1%. Ad esempio per α =5% si avrà:



Sotto questa ipotesi il test discrimina i **campioni** che portano all'accettazione dell'ipotesi nulla da quelli che portano al suo rifiuto secondo questa regola:

I valori della statistica test che cadono nella regione di accettazione portano **all'accettazione dell'ipotesi nulla** (perché molto probabili).

I valori della statistica test che cadono nella regione di rifiuto portano al **rifiuto dell'ipotesi nulla** (perché poco probabili).

4º passo: calcolo del valore della statistica test per il campione osservato

Se il campione osservato cade in una delle due zone di rifiuto potremmo prendere una delle due decisioni:

- **si è verificato un valore molto raro sotto l'ipotesi nulla, ma non rifiutiamo l'ipotesi nulla**

- **essendosi verificato un valore che è molto raro sotto l'ipotesi nulla, rifiutiamo l'ipotesi nulla**

Il test statistico si basa su questa regola

Pertanto la **regola decisionale del test di ipotesi** è la seguente:

Rifiutare H_0 se $z < -1.96$ o $z > +1.96$

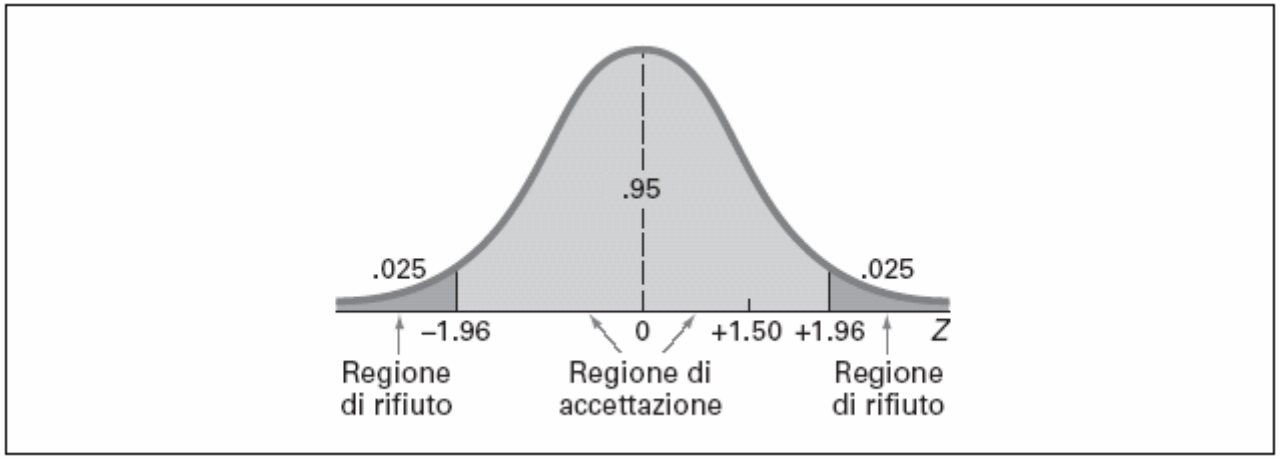
oppure in modo equivalente

se $\bar{x} < \mu - 1{,}96 \times \sigma/\sqrt{n}$ o $\bar{x} > \mu + 1{,}96 \times \sigma/\sqrt{n}$

Non rifiutare H_0 altrimenti

Es. Nell'esempio del processo produttivo di riempimento delle scatole di cerali, la media campionaria è risultata pari a 372,5:

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



Il valore osservato della statistica test cade nella regione di accettazione. Il test mi porta a non rifiutare l'ipotesi nulla. La differenza tra la media campionaria osservata (372,5 gr) e il valore ipotizzato per la media della popolazione (368 gr) non è statisticamente significativa (è dovuta al caso e al processo di campionamento). Il processo è sotto controllo.

Il *p*-value

Un altro modo per evidenziare il risultato del test è quello di calcolare il *p*-value (detto anche livello di significatività osservato).

p-value : probabilità di osservare un valore della statistica test uguale o più estremo del valore ottenuto dal campione, sotto l'ipotesi nulla.

E' una quantità che misura l'evidenza fornita dai dati contro l'ipotesi nulla: minore è il valore del p-value, più è forte l'evidenza contro l'ipotesi nulla.

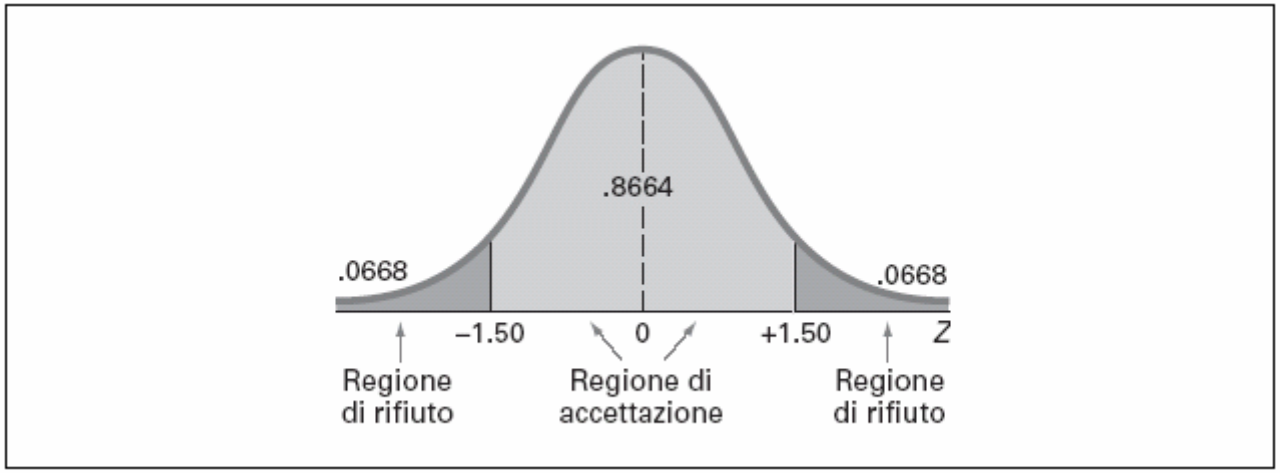
In base all'approccio del *p*-value, la regola decisionale per rifiutare H_0 è la seguente:

Se il *p*-value è $\geq \alpha$, l'ipotesi nulla non è rifiutata.

Se il *p*-value è $< \alpha$, l'ipotesi nulla è rifiutata.

Es. Torniamo ancora una volta all'esempio relativo alla produzione delle scatole di cereali. Nel verificare se il peso medio dei cereali contenuti nelle scatole è uguale a 368 grammi, abbiamo ottenuto un valore di Z uguale a 1.50 e non abbiamo rifiutato l'ipotesi, perché 1.50 è maggiore del valore critico più piccolo –1.96 e minore di quello più grande +1.96.

Risolviamo, ora, questo problema di verifica di ipotesi facendo ricorso all'approccio del *p*-value. Per questo test a due code, dobbiamo, in base alla definizione del *p*-value, calcolare la probabilità di osservare un valore della statistica test uguale o più estremo di 1.50.



Si tratta, più precisamente, di calcolare la probabilità che Z assuma un valore maggiore di 1.50 oppure minore di –1.50. In base alla Tavola della distribuzione normale standardizzata, la probabilità che Z assuma un valore minore di –1.50 è 0.0668, mentre la probabilità che Z assuma un valore minore di +1.50 è 0.9332, quindi la probabilità che Z assuma un valore maggiore di +1.50 è 1 – 0.9332 = 0.0668. Pertanto il *p*-value per questo test a due code è 0.0668 + 0.0668 = 0.1336. Essendo il *p*-value > α (0.05) non rifiutiamo H_0.

Errori di I e II tipo

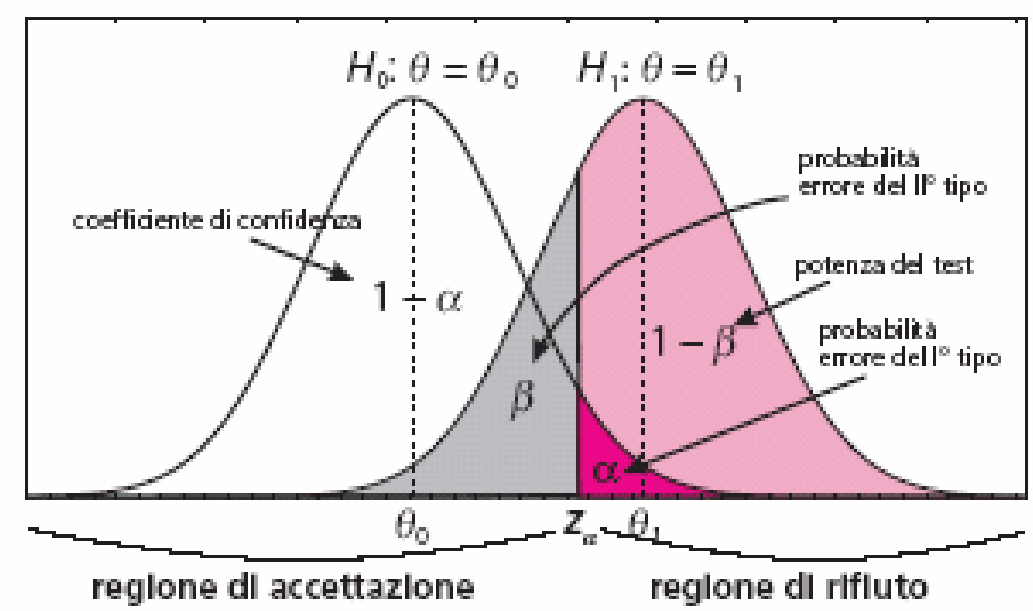
- **errore del I tipo: si rifiuta l'ipotesi nulla mentre questa è vera.**
- **errore del II tipo: si accetta l'ipotesi nulla mentre questa è falsa**

	Decisione	
	Accetto H_0	**Rifiuto** H_0
H_0 è vera	Corretta $1-\alpha$	Errore del I tipo α
H_0 è falsa	Errore del II tipo β	Corretta $1-\beta$

- α è la probabilità di commettere l'errore del I tipo, ovvero di rifiutare l'ipotesi nulla quando è vera (livello di significatività del test).
- $1-\alpha$ è detto coefficiente di confidenza del test.
- β è la probabilità di commettere l'errore del II tipo, ovvero di accettare l'ipotesi nulla quando è falsa.
- $1-\beta$ è la potenza del test e corrisponde alla probabilità di rifiutare l'ipotesi nulla quando questa è falsa.

Errori di I e II tipo

I diversi errori che si possono commettere:



Tra α e β sussiste una relazione inversa: minore è il valore di α, maggiore è il valore di β. Le probabilità di commettere gli errori corrispondono a delle aree.

In genere, si controlla l'errore di prima specie fissando il livello del rischio *α* che si è disposti a tollerare

Dal momento che il livello di significatività è specificato prima di condurre la verifica di ipotesi, il rischio di commettere un errore di prima specie *α* è sotto il controllo di chi compie l'analisi (in genere i valori assegnati ad *α* sono 0.05, 0.01 o 0.001)

La scelta di *α* dipende fondamentalmente dai costi che derivano dal commettere un errore di prima specie.

Un modo per controllare e ridurre l'errore di seconda specie consiste **nell'aumentare la dimensione del campione** perché un'elevata dimensione del campione consente di individuare anche piccole differenze tra la statistica campionaria e il parametro della popolazione.

Per un dato valore di α, l'aumento della dimensione campionaria determina una riduzione di β e quindi un aumento della potenza del test per verificare se l'ipotesi nulla H_0 è falsa.

Tuttavia per una data ampiezza campionaria dobbiamo tenere conto del trade-off tra i due possibili tipi di errori: possiamo fissare un valore piccolo per α, tuttavia al diminuire di α, β aumenta e pertanto una riduzione del rischio connesso all'errore di prima specie si accompagna a un aumento di quello connesso a un errore di seconda specie.

Passi da seguire nella verifica d'ipotesi

- Definizione del sistema d'ipotesi
- Scelta della statistica test
- Scelta del livello di significatività e della numerosità campionaria
- Definizione della regione di rifiuto
- Estrazione del campione
- Calcolo della statistica test
- Decisione

Test di ipotesi *Z* per la media (σ noto)

Statistica *Z* per la verifica d'ipotesi sulla media (σ noto)

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

Test di ipotesi *t* per la media (σ non noto)

Statistica *t* per la verifica d'ipotesi sulla media (σ non noto)

$$t = \frac{\overline{X} - \mu}{S / \sqrt{n}}$$

Il test di significatività non deve essere considerato come una regola automatica per prendere delle decisioni; il ricercatore deve aggiungere all'evidenza del test anche altri elementi di valutazione.

Per una corretta impostazione e scelta del test è bene considerare:

- **il tipo di variabile su cui si verifica una data ipotesi;**
- **ipotesi sulla distribuzione della variabile nella popolazione;**
- **caratteristiche delle unità statistiche (campioni dipendenti, non dipendenti);**
- **tipo di studio (osservazionale, sperimentale, …);**
- **dimensione del campione;**
- **significatività (clinico, …) dell'ipotesi vs significatività statistica.**

Vi sono vari tipi di test:

test normale sulla media
test t-student sulla media
test normale per una frequenza
test t-student per la differenza di due medie
test chi-quadro per la varianza
test chi-quadro per il confronto di frequenze
…

Legame tra intervalli di confidenza e verifica di ipotesi

Consideriamo i due elementi principali dell'inferenza statistica – gli intervalli di confidenza e la verifica di ipotesi. Sebbene abbiano una stessa base concettuale, essi sono utilizzati per scopi diversi: gli intervalli di confidenza sono usati per stimare i parametri della popolazione, mentre la verifica di ipotesi viene impiegata per poter prendere delle decisioni che dipendone dai valori dei parametri.

Tuttavia è importante sottolineare che anche gli intervalli di confidenza possono consentire di valutare se un parametro è minore, maggiore o diverso da un certo valore: ad esempio, anziché sottoporre a verifica l'ipotesi μ=368 gr, possiamo risolvere il problema costruendo un intervallo di confidenza per la media μ. In questo caso accettiamo l'ipotesi nulla se il valore ipotizzato è compreso nell'intervallo costruito, perché tale valore non può essere considerato insolito alla luce dei dati osservati. D'altronde, l'ipotesi nulla va rifiutata se il valore ipotizzato non cade nell'intervallo costruito, perché tale valore risulta insolito alla luce dei dati.

Legame tra intervalli di confidenza e verifica di ipotesi

Con riferimento al problema considerato, l'intervallo di confidenza è costruito ponendo: n = 25, $\overline{X}$ = 372.5 grammi, σ = 15 grammi.

Per un livello di significatività del 95% (corrispondente al livello di significatività del test α =0.05), avremo:

$$\overline{X} \pm Z_{\alpha/2} \cdot \sigma / \sqrt{n} \Rightarrow 372.5 \pm (1.96) \cdot 15 / \sqrt{25} \Rightarrow 366.6 \leq \mu \leq 378.4$$

Poiché l'intervallo comprende il valore ipotizzato di 368 grammi, non rifiutiamo l'ipotesi nulla e concludiamo che non c'è motivo per ritenere che il peso medio dei cereali contenuti nelle scatole sia diverso da 368 grammi.