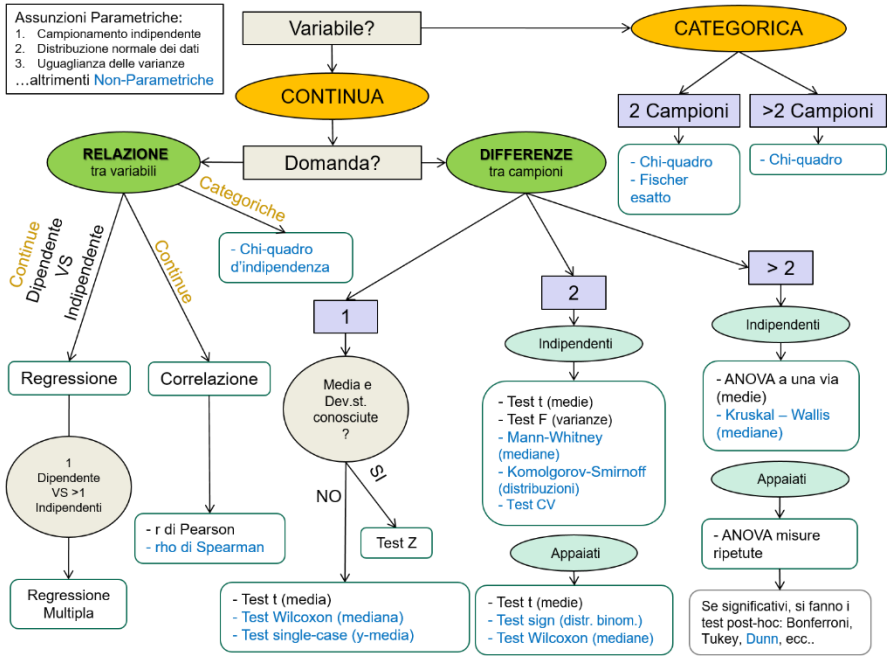
Corso di Laurea di Scienze Forestali ed Ambientali

Ecologia e Statistica per l'ambiente

a.a. 2020-2021 - MATTEO GARBARINO - matteo.garbarino@unito.it

STATISTICA INFERENZIALE



Rami della Statistica

Statistica Descrittiva

- Metodo deduttivo *(dal generale al particolare)*
- Raccolta dei dati
- Sintesi dei dati di popolazione o del campione
- Presentazione dei risultati *(Analisi esplorativa)*

Statistica Matematica

- Distribuzioni teoriche
- Calcolo delle probabilità

Statistica Inferenziale

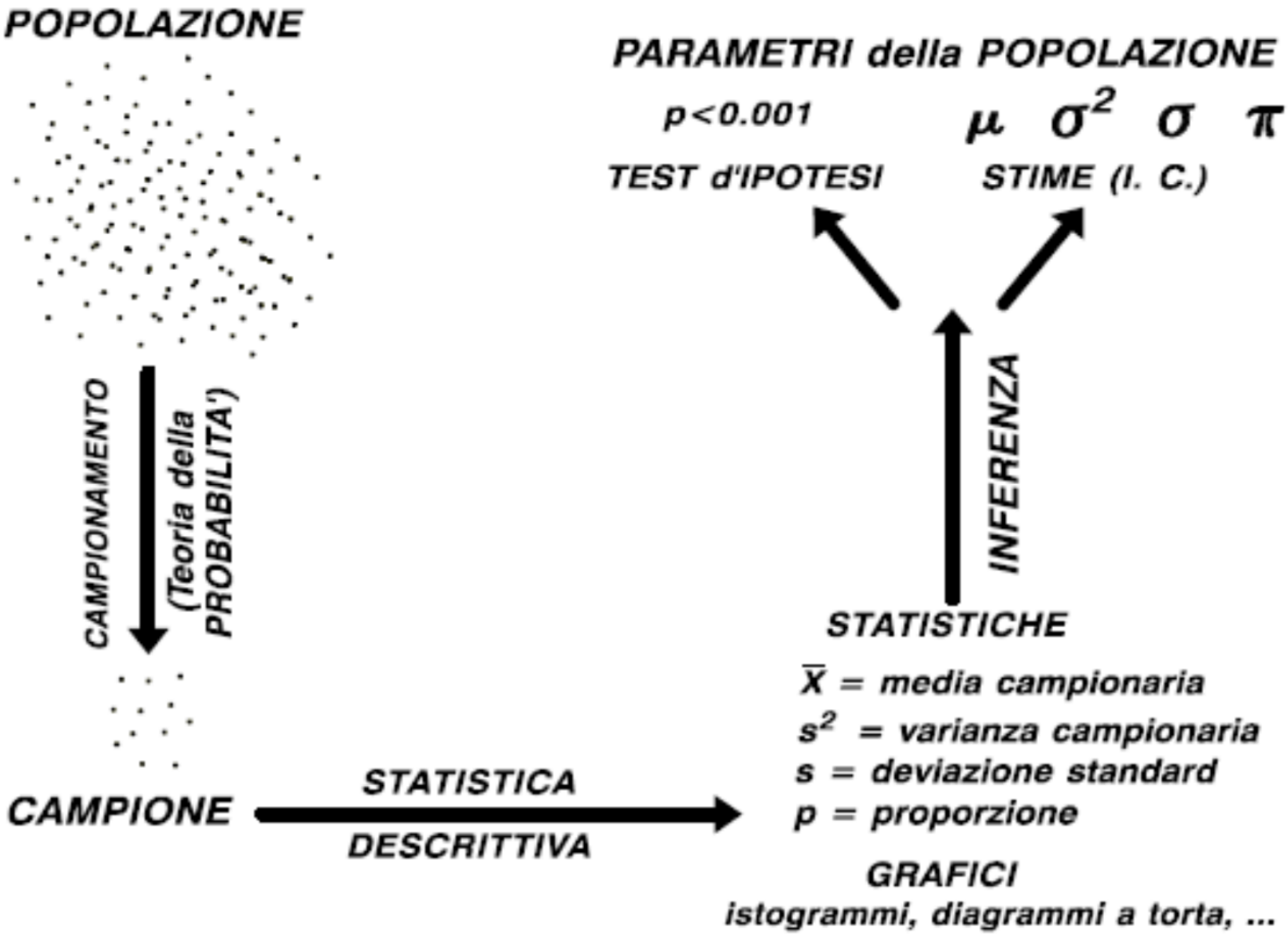
- Metodo induttivo *(dal particolare al generale)*
- Scelta del tipo di popolazione
- Rilevazioni parziali
- Stima dei parametri di popolazione
- Verifica delle ipotesi
- Previsione

Statistica Inferenziale

- La statistica descrittiva, pur aiutandoci a capire le proprietà dei dati in nostro possesso, non aggiunge nulla alle informazioni che già abbiamo. Le sue affermazioni, essendo relative a dati certi, sono certe.

- La statistica inferenziale, invece, si propone di fare nuove affermazioni a proposito di dati che non possediamo, per mezzo di una elaborazione matematica derivata dalla teoria delle probabilità. Le sue affermazioni, quindi, sono probabilistiche.

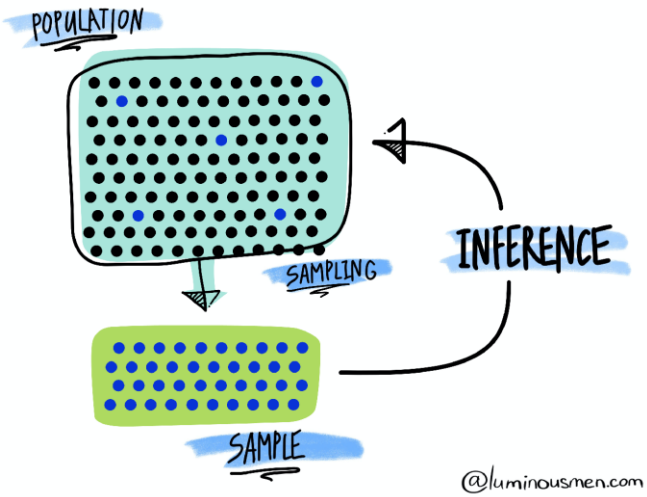
Schema logico nella statistica inferenziale

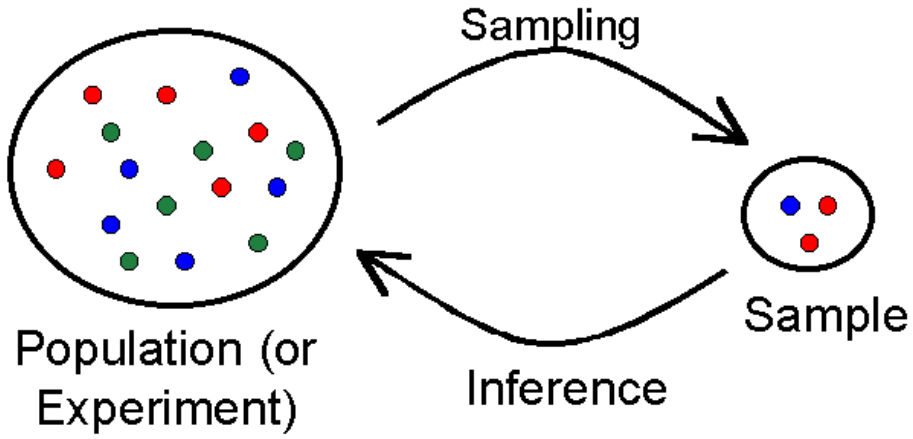


Inferenza: esempio

Data un'urna con composizione nota di 6 palline bianche e 4 palline rosse (**POPOLAZIONE NOTA**), utilizzando le regole del calcolo delle probabilità possiamo dedurre che se estraiamo una pallina a caso dall'urna, la probabilità che essa sia rossa è 0,4 (4 su 10).

Si ha invece un problema di inferenza statistica quando abbiamo un'urna di cui non conosciamo la composizione (**POPOLAZIONE SCONOSCIUTA**), *estraiamo n palline a caso, ne osserviamo il colore e, a partire da questo, cerchiamo di inferire la composizione dell'urna*.





Statistica Inferenziale

- Il concetto di verità delle affermazioni della statistica inferenziale deve essere ben compreso.
- Le affermazioni della statistica inferenziale sono matematicamente vere e rigorose (nell'ambito della validità del modello matematico che si adotta, e purchè, naturalmente, i calcoli vengano condotti correttamente), ma riguardano esclusivamente la *probabilità* della verità di altre affermazioni.
- In altre parole, la statistica inferenziale non ci fornisce certezze sull'argomento della nostra ricerca, ma solo certezze sulla probabilità che le nostre asserzioni su tale argomento siano vere.

Statistica Inferenziale

- I problemi che la statistica inferenziale cerca di risolvere sono essenzialmente di due tipi:

- 1) Problema della stima (per esempio stima di una media):
  - fornisce informazioni sulla media di una popolazione quando sono note media e deviazione standard di un campione della stessa.

- 2) Problema della verifica di ipotesi (per esempio confronto fra due o più campioni):
  - calcola la probabilità che due campioni, di cui siano note media e deviazione standard, siano campioni derivati da una stessa popolazione oppure da due popolazioni diverse.

1) Stima - Campionamento statistico

- Nell'ambito della statistica descrittiva abbiamo finora considerato strumenti per descrivere un'intera popolazione quando siano noti tutti i dati ad essa relativi. Ma nella ricerca, in genere, non si conoscono i dati dell'intera popolazione, ma solo quelli di un campione.

- Il campionamento si usa quando si vuole conoscere uno o più parametri di una popolazione, senza doverli misurare in ogni suo elemento. Il campionamento consiste nel selezionare un numero più piccolo di elementi fra tutti quelli che formano una popolazione. Può essere fatto in vari modi, ma deve sempre essere di tipo probabilistico (cioè garantire la casualità della selezione).

- Parleremo allora di numerosità, media e deviazione standard del campione, e dobbiamo porci il problema di che rapporto esista fra questi valori e la numerosità, la media e la deviazione standard dell'intera popolazione.

1) Stima – Campione VS Popolazione

- Immaginiamo di avere una popolazione rappresentata da mille persone (per esempio la popolazione degli abitanti maschi di un paese), e di volere conoscere la loro statura.
- Se conoscessimo la statura di ciascuno dei mille abitanti, potremmo descrivere la popolazione con assoluta precisione in termini di media e deviazione standard.

1) Stima – Campione VS Popolazione

- Se però non abbiamo le risorse per misurare la statura di mille abitanti, possiamo scegliere un campione casuale, per esempio di 30 abitanti. Avremo allora una media e una deviazione standard del campione, la cui numerosità è naturalmente 30.
- Che rapporto c'è fra questi valori e quelli dell'intera popolazione di mille abitanti?

1) Stima – Campione VS Popolazione

- Immaginiamo di ripetere l'operazione di campionamento 20 volte, ogni volta con un diverso campione casuale di 30 abitanti. Otterremo 20 medie diverse, e 20 DS diverse.
- Un concetto importante è che l'insieme di queste medie dei campioni tende ad assumere una distribuzione normale, anche se la popolazione di origine non è distribuita normalmente.
- In altre parole, il processo di campionamento casuale è di per sé un fenomeno che si distribuisce normalmente.

2) Test (verifica) delle ipotesi

- La verifica di ipotesi è il secondo tipo di problema affrontato dalla statistica inferenziale.
- L'ipotesi da verificare in questo caso è la cosiddetta "ipotesi nulla" (null hypothesis)

2) Test - Ipotesi Nulla

- L'ipotesi nulla (H_0) è un'ipotesi che il ricercatore fa riguardo a un parametro della popolazione oggetto della ricerca (in genere la media) e che viene confutata o non confutata dai dati sperimentali. Nel caso più comune, del confronto fra due campioni, la forma dell'ipotesi nulla è la seguente:

$$H_0: \mu_1 = \mu_2$$

Dove μ_1 e μ_2 sono le medie delle due popolazioni da cui sono stati tratti i due campioni.

Esempio: ho due popolazioni di alberi (lariceti dell'Alpe Veglia e lariceti dell'Alpe Devero). L'ipotesi nulla è che la media dei diametri delle due popolazioni (cavallettamento totale) sia identica.

2) Test - Ipotesi Nulla

- Molto spesso l'ipotesi nulla è l'opposto di ciò che si vorrebbe dimostrare.
- Come vedremo, l'ipotesi nulla viene rigettata oppure no a secondo del suo livello di "improbabilità".
- Se l'ipotesi nulla viene rigettata, questo è un dato a favore dell'ipotesi alternativa. In senso stretto, però, il test statistico non dice nulla sull'ipotesi alternativa H1, ma solo sulla probabilità dell'ipotesi nulla.

Riassumendo:

- Se H0 viene rigettata perché improbabile, questo è un dato a favore di H1
- Se H0 non viene rigettata, questo non vuol dire che H0 debba essere vera. Si può solo dire che, sulla base dei dati raccolti, non la si può considerare "abbastanza" improbabile.

Esempio: se l'ipotesi nulla viene rigettata è molto probabile che i lariceti dell'Alpe Devero siano diversi (in diametro) da quelli dell'Alpe Veglia.

2) Test - Il p-value (*p*)

- Ma che vuol dire "abbastanza" improbabile? Anche nel caso della verifica di ipotesi, è necessario decidere un "livello" di improbabilità che autorizzi a "rigettare" l'ipotesi nulla.
- Questo valore si chiama p-value, o soltanto p, e si può definire come la probabilità che il risultato ottenuto (per esempio la differenza fra le medie dei due campioni) sia dovuto al caso, se l'ipotesi nulla è vera, cioè se le medie delle popolazioni da cui i campioni sono tratti sono uguali.
- Il p si esprime come frazione dell'unità. Valori di p spesso usati come livello sono:

✓ $p < 0.05$	(cioè una probabilità < al 5%)	Abb.za Significativo *
✓ $p < 0.01$	(cioè una probabilità < al 1%)	Molto significativo **
✓ $p < 0.001$	(cioè una probabilità < allo 0,1%)	Estr.te significativo ***

Questi sono LIVELLI DI SIGNIFICATIVITÀ e vengono indicati spesso con dei simboli (*, **, ***).

2) Test - Errori di tipo I e II

SE il p è < a 0,01: l'ipotesi nulla viene rigettata, in favore di una possibile ipotesi alternativa.

(studio che ha successo)

Se però l'ipotesi nulla è vera, si commette un errore di tipo I.

La probabilità di commettere un errore di tipo I (detta α) è uguale al p-value. *Es.* $p = 0.01 \rightarrow 1\%$

SE il p è > a 0,01: l'ipotesi nulla non viene rigettata. Ciò non dimostra che essa sia vera.

(studio che non ha successo)

Se comunque l'ipotesi nulla è falsa, si commette un errore di tipo II.

La probabilità di commettere un errore di tipo II (detta β) spesso non è calcolabile.

La causa più frequente di errore di tipo II è la numerosità insufficiente dei campioni.

Campione non rappresentativo della popolazione

2) Test – Err. di tipo II e Potenza

- β è la probabilità di commettere un errore di tipo II, cioè di non riuscire a rigettare un'ipotesi nulla che è falsa (in altre parole, di non riuscire ad affermare la nostra ipotesi anche se è vera
- 1- β esprime la potenza di uno studio, cioè la probabilità di non commettere un errore di tipo II
- Se β è 0,20, la potenza dello studio sarà 0,80, in altre parole lo studio avrà l'80% di probabilità di riuscire a dimostrare la propria ipotesi, se questa è vera

2) Test - Da cosa dipende la potenza?

1. Dalla dimensione reale dell'effetto che si vuole dimostrare. In altre parole, quanto più il segnale da rivelare è grande, tanto più facile è, per uno studio, rivelarlo. *Es. Differenze tra 2 campioni*
2. Dal livello di significatività prefissato (soglia di p). In altre parole, quanto più bassa si pone la soglia di p, tanto più facile è che non si arrivi a quella soglia anche se l'ipotesi è vera. Uno studio che vuole essere più affidabile, sarà anche meno potente. *Se p è basso ho poca potenza, ma sono molto robusto!*
3. Dalla numerosità del campione. Più grande è N, più potente è lo studio. *Una delle poche cose su cui posso intervenire → aumento n*
4. Dalla varianza (o DS) della popolazione di origine. Più grande è la varianza, meno potente è lo studio
5. Da altri fattori: normalità della popolazione, tipo di test statistico adoperato *Normalità = test parametrici = più potenti*

Dimensionamento del campione

- Un campione troppo piccolo porta più facilmente ad errori di tipo II
- La numerosità del campione dipende però in modo critico dall'entità della differenza esistente fra le due popolazioni relativamente al parametro oggetto dello studio
- In uno studio RCT, quindi, è importante dimensionare in anticipo il campione, cioè decidere prima quanti soggetti dovranno essere arruolati per rispondere al quesito
- Il dimensionamento va fatto tenendo conto della differenza più piccola che si ha interesse a cogliere (grandezza del segnale minimo che si considera utile), e del livello di significatività statistica che si desidera raggiungere (cioè, della soglia fissata per il p)

RCT = Randomized Controlled Trial = confronto tra gruppo case e control

STIMA
Campioni, confronti, ipotesi

Da una lezione di Michele Scardi

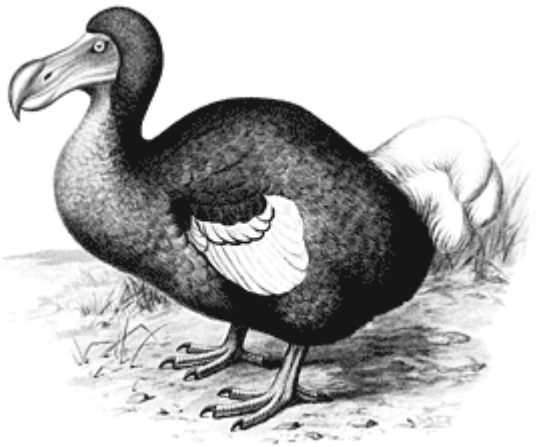
Due modi diversi di ragionare…

1. Inferenza deduttiva

2. Inferenza induttiva

Quanti campioni sono possibili?

Immaginiamo di essere tornati un po' indietro nel tempo e di aver potuto studiare la popolazione di Dodo prima della sua estinzione. Il nostro obiettivo è sapere quante uova deponeva in media ciascuna delle 6 femmine rimaste (non estinte) → **intera popolazione**.



Dodo	Uova
A	0
B	9
C	6
D	3
E	1
F	5

μ = 4 media

σ^2 = 9.33 varianza

σ = 3.06 dev.stan.

Quanti diversi campioni erano possibili per n = 3 ?

$$\frac{6!}{3!*3!} = 20$$

campionamento

Medie stimate dai 20 campioni possibili

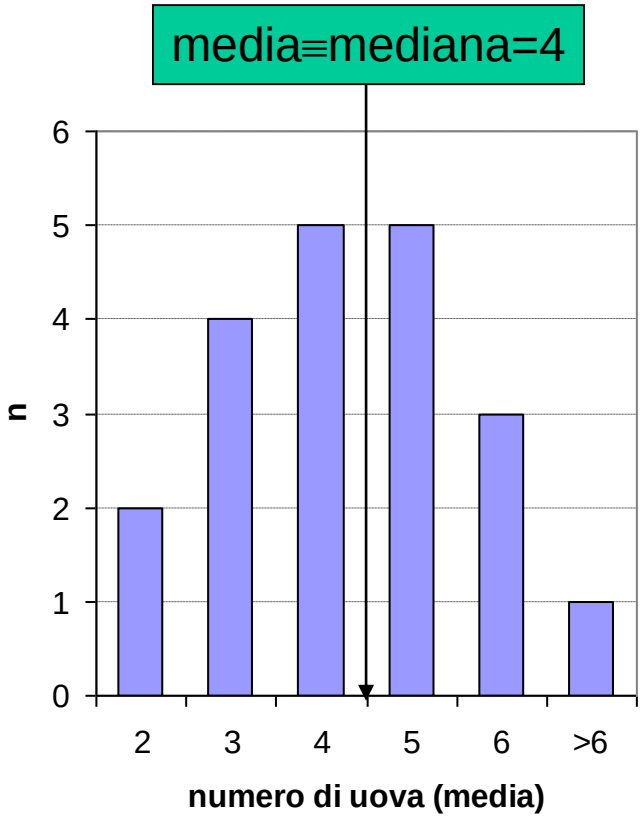
Risultato di tutte le combinazioni possibili tra i valori della tabella precendente

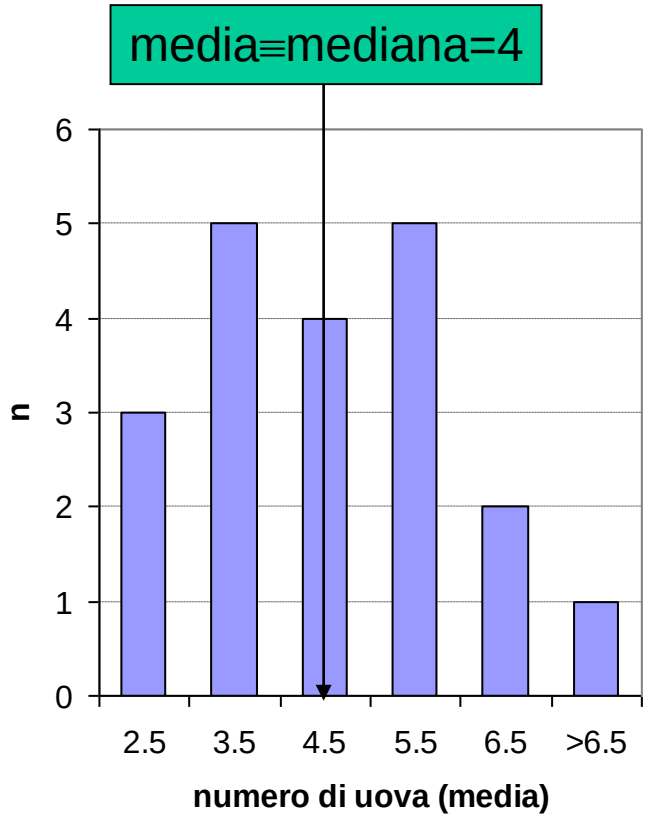
1

20

Dodo #1	Dodo #2	Dodo #3	Media del campione (m)
0	1	3	1.33
0	1	5	2.00
0	1	6	2.33
0	1	9	3.33
0	3	5	2.67
0	3	6	3.00
0	3	9	4.00
0	5	6	3.67
0	5	9	4.67
0	6	9	5.00
1	3	5	3.00
1	3	6	3.33
1	3	9	4.33
1	5	6	4.00
1	5	9	5.00
1	6	9	5.33
3	5	6	4.67
3	5	9	5.67
3	6	9	6.00
5	6	9	6.67

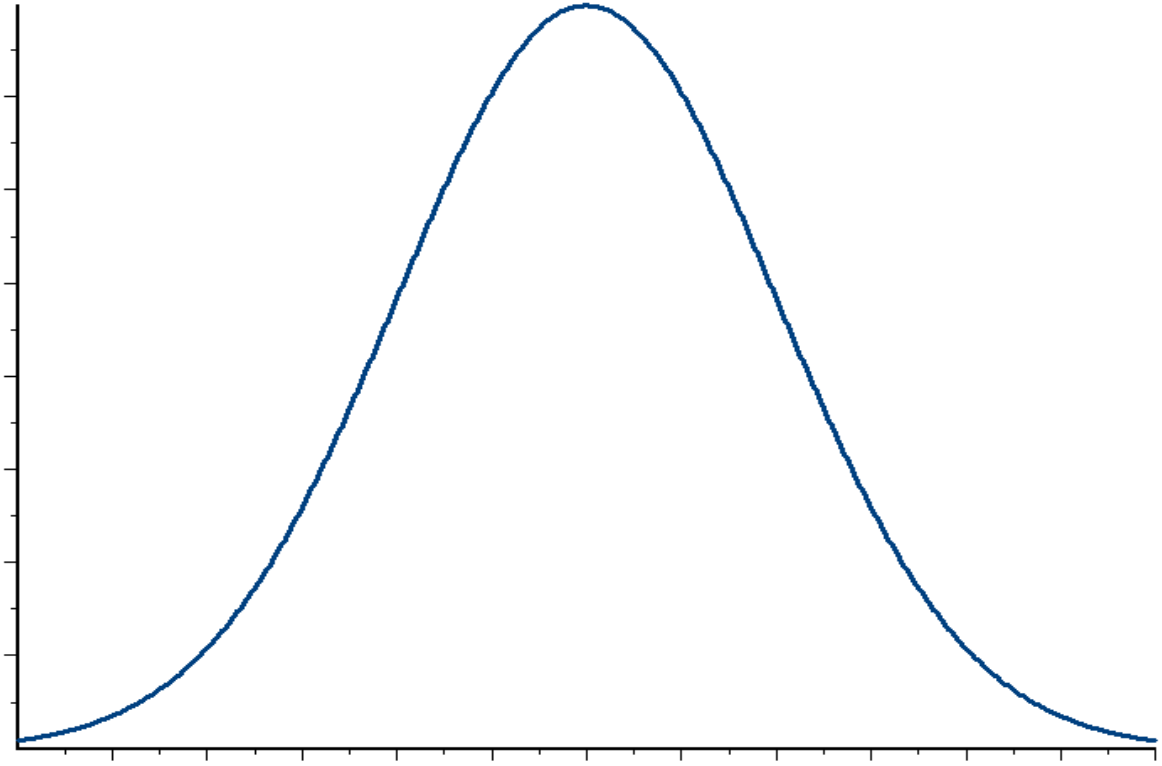
Distribuzione delle medie



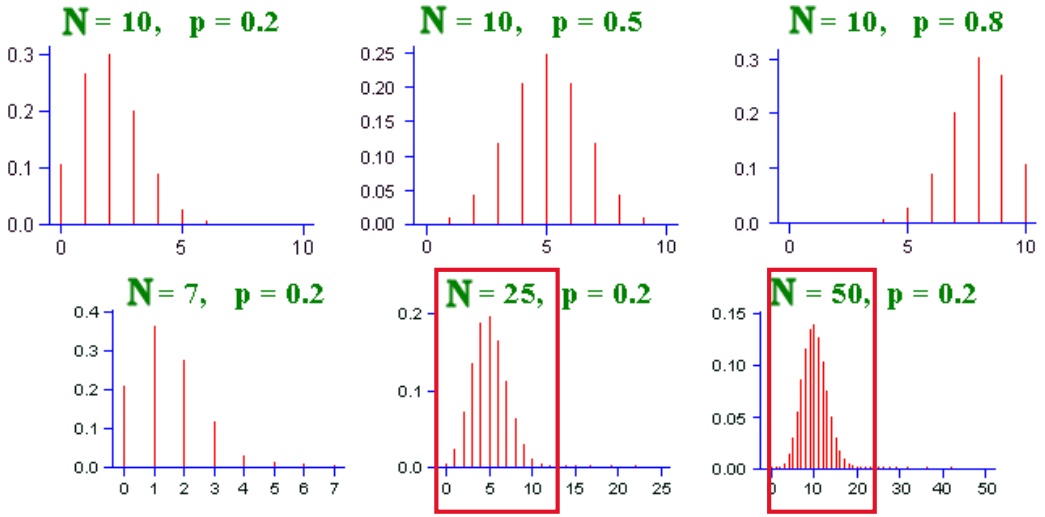


Differente solamente l’intervallo tra le classi degli istogrammi

La distribuzione normale



Per grandi numeri (num.tà campionaria), anche distribuzioni non normali tendono alla normale (teorema del limite centrale)



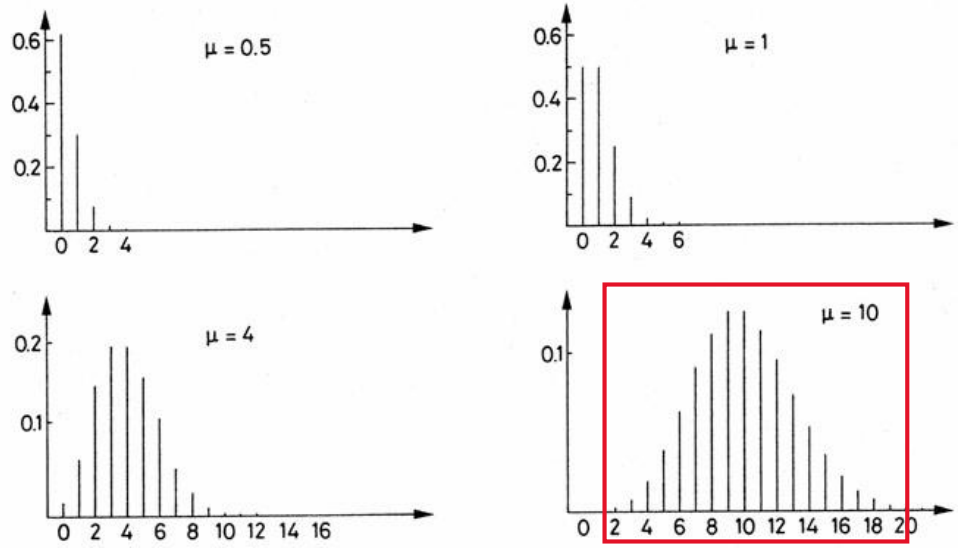
Distribuzione binomiale

$$P_p(n|N) = \binom{N}{n} p^n q^{N-n}$$

$$= \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

$$P_x = e^{-\mu} \left[\frac{\mu^x}{x!} \right]$$

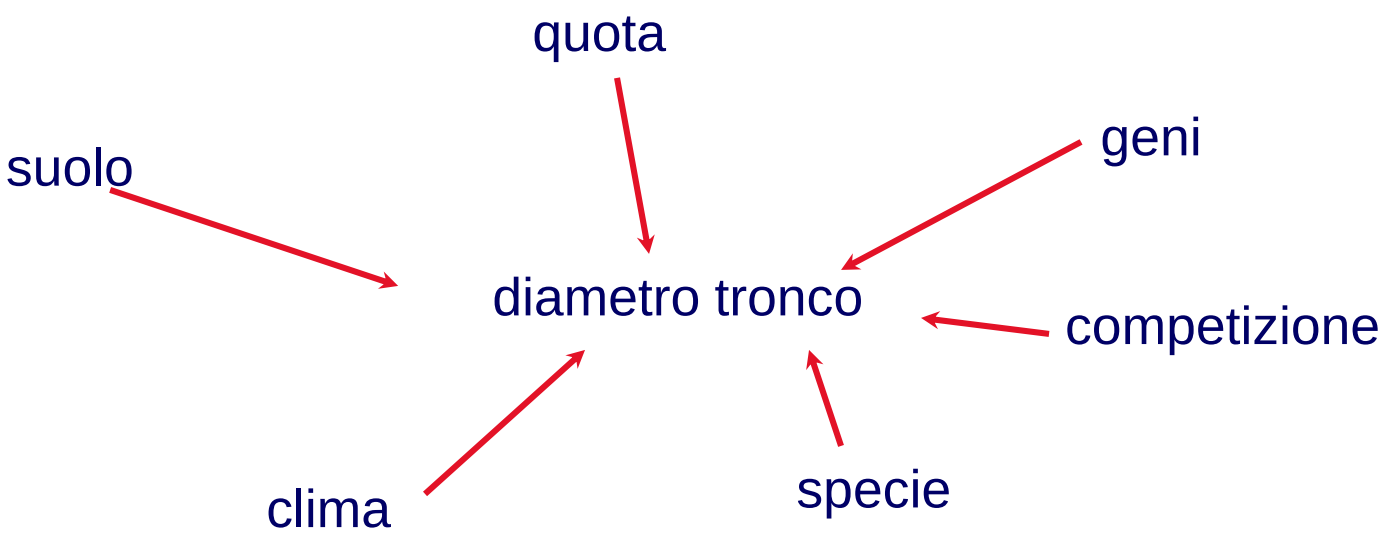
Distribuzione di Poisson



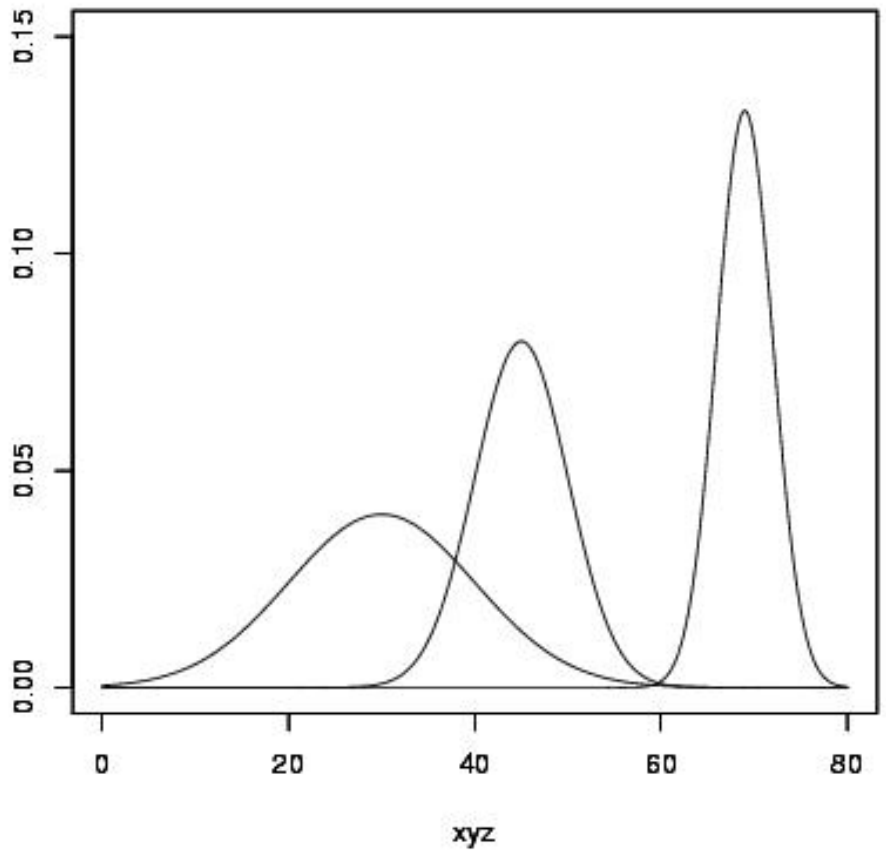
Una variabile casuale influenzata da numerosi fattori tende ad avere una distribuzione normale

Dati biometrici, tassi di vario tipo, misure fisiche in generale, etc.

Se i valori misurati sono influenzati da un numero elevato di eventi casuali, allora la distribuzione tenderà ad essere normale.



Le curve normali hanno forme variabili…



Quindi, **per comparare più distribuzioni normali tra loro**, dobbiamo **standardizzarle** in qualche modo…

Standardizzazione: la variabile Z

$$Z = \frac{\text{valore osservato var. casuale} - \text{media}}{\text{deviazione standard}}$$

ovvero

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

Questa operazione di trasformazione delle variabili è anche detta “ajust to standard deviate” ed è molto utilizzata in statistica multivariata per rendere confrontabili variabili che hanno unità di misura differenti (es. Quota e Diametri degli alberi).

Standardizzazione VS Normalizzazione

La **normalizzazione** riscala tutti i valori tra [0,1], utile nel caso si vogliano rendere positive tutte le variabili e si vogliano eliminare gli outlier.

$$X_{nuovo} = \frac{X - Xmin}{X_{max} - Xmin}$$

La **standardizzazione** invece riscala tutti i valori ad avere media $\mu = 0$ e deviazione standard $\sigma = 1$. per molte applicazioni è da preferirsi.

$$X_{nuovo} = \frac{X - \mu}{\sigma}$$

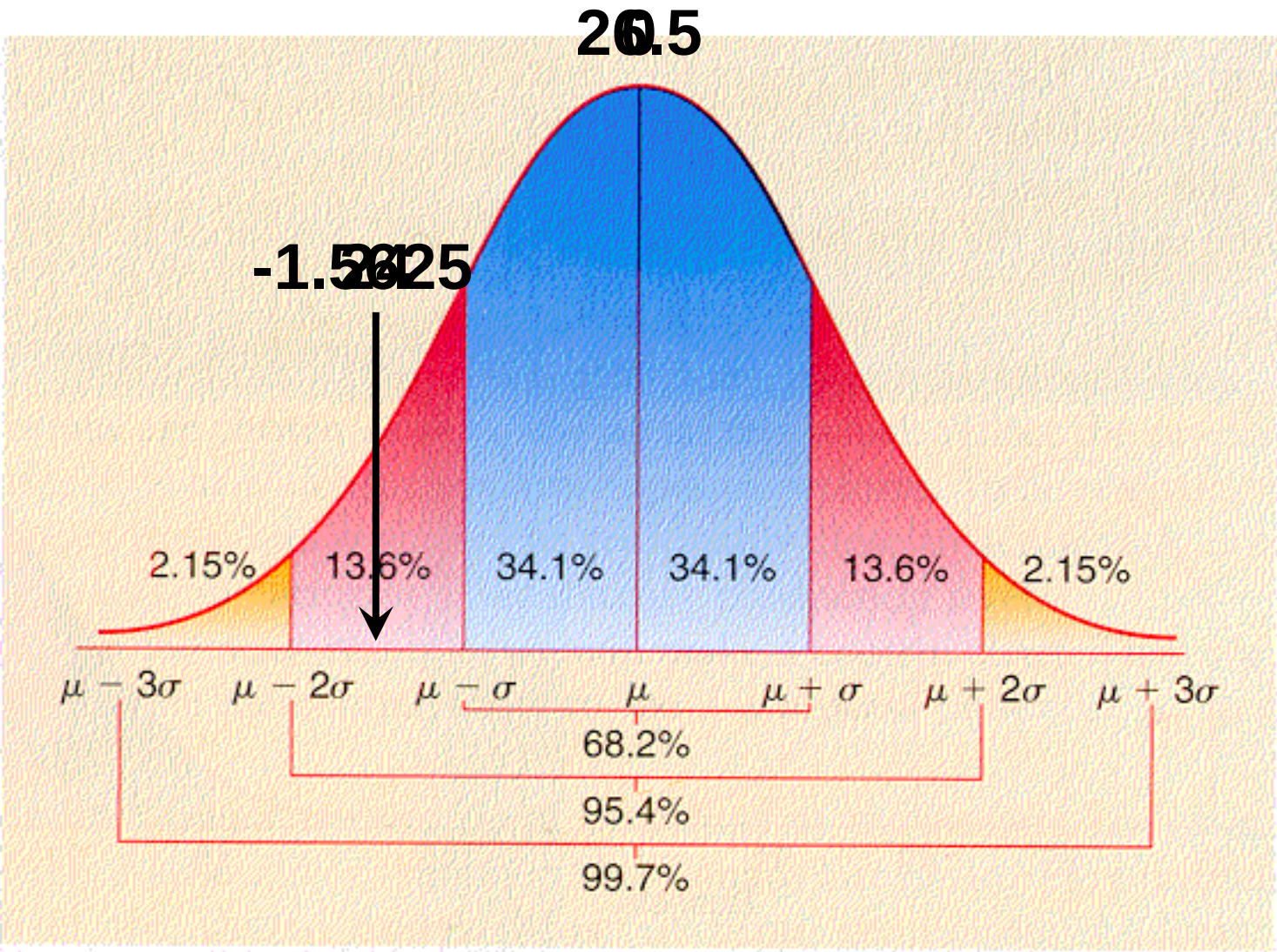
Esempio

Il voto medio del corso di Statistica è 26.5, mentre la deviazione standard è 1.6. Se hai avuto 24, qual'è stato il valore della variabile Z nel tuo caso?

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

$$Z = \frac{24 - 26.5}{1.6} = -1.5625$$

In pratica, Z ci dice di quante deviazioni standard un valore si scosta dalla media…



Ogni deviazione standard di scarto dalla media definisce un’area sotto la curva, che equivale a una certa percentuale di casi

Distribuzione delle medie

- Se una popolazione è molto più grande di quella del Dodo, non potrò calcolare tutte le medie possibili, né conoscere la media vera.
- Se raccolgo i dati relativi a un campione, posso stimare l'intervallo entro cui si trova la media vera con un certo livello di probabilità?
- Si, perché so che la distribuzione delle medie di tutti i campioni che posso estrarre è normale.
- Quello che mi serve è l'intervallo fiduciale della media.

Intervallo fiduciale o di confidenza della media

Per intervallo di confidenza di un parametro Θ della popolazione, intendiamo un intervallo delimitato da due limiti L_{inf} (limite inferiore) ed L_{sup} (limite superiore) che abbia una definita probabilità $(1-\alpha)$ di contenere il vero parametro della popolazione:

$$p(L_{inf} < \Theta < L_{sup}) = 1-\alpha$$

dove:

$1-\alpha$ = grado di confidenza

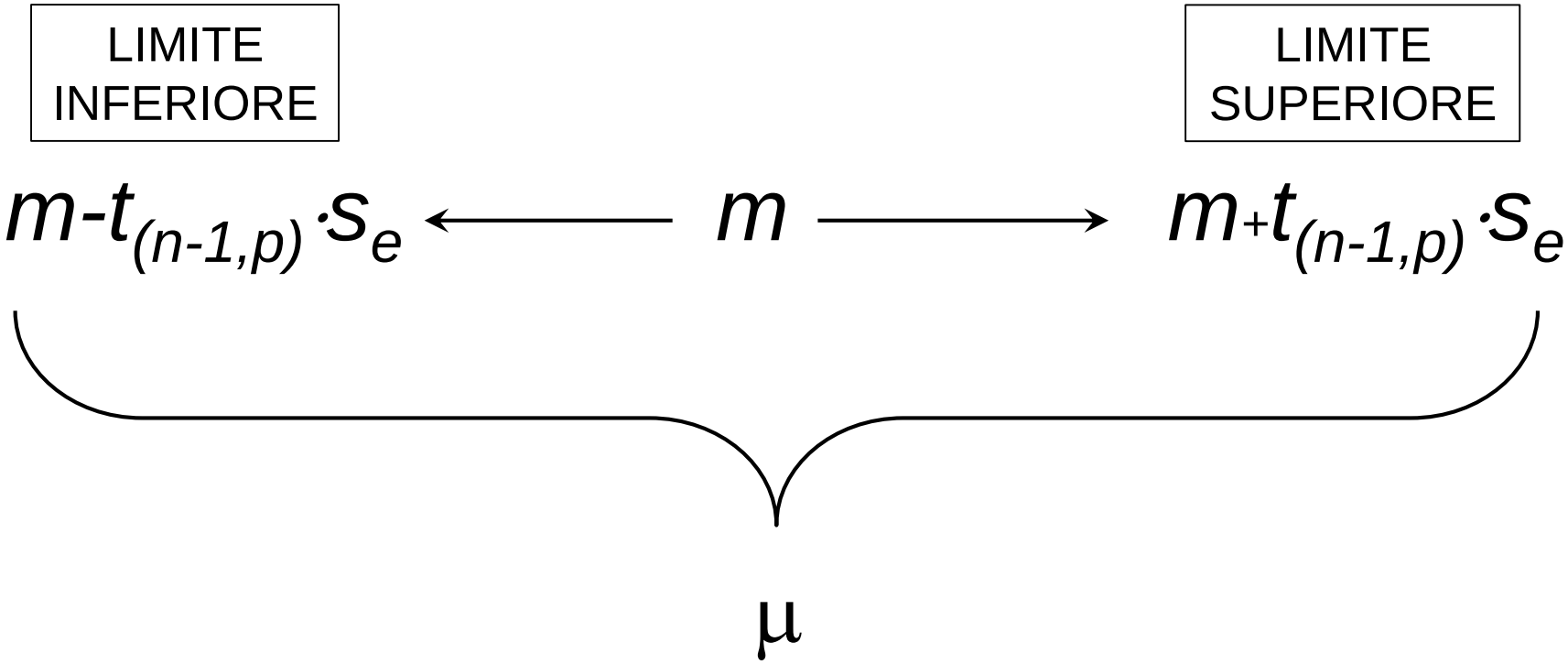
α = probabilità di errore

Intervallo fiduciale o di confidenza della media

- Calcolo la media
- Calcolo la deviazione standard
- Calcolo l'errore standard della media: $s_e = \frac{\sigma}{\sqrt{n}}$

- La **media della popolazione** μ sarà compresa nell'intervallo fra la **media campionaria** ***m*** meno $t_{(n-1,p)} \cdot s_e$ e la **media campionaria** ***m*** più $t_{(n-1,p)} \cdot s_e$ dove $t_{(n-1,p)}$ è il valore del ***t* di Student con *n-1*** gradi di libertà per il livello di **probabilità *p*** desiderato.

Intervallo fiduciale della media (in altre parole…)



(con una probabilità *p)*

Intervallo fiduciale della media (in altre parole…)

Dodo	Uova
A	0
B	9
C	6
D	3
E	1
F	5

$m=(9+3+1)/3=4.333$

x	x-m	$(x-m)^2$
9	4.667	21.778
3	-1.333	1.778
1	-3.333	11.111

$s^2=[\Sigma(x-m)^2]/(n-1) = 17.333$

$s=\sqrt{[\Sigma(x-m)^2]/(n-1)} = 4.163$

Intervallo fiduciale della media (in altre parole…)

Dodo	Uova
A	0
B	9
C	6
D	3
E	1
F	5

m=4.333 s=4.163

$s_e = s/\sqrt{n} = 4.163/\sqrt{3} = 2.404$

$t_{(n-1,p)} = t_{(3-1,0.975)} =$ 4.303

	Quantile (area to the left of t)									
df	0.600	0.700	0.800	0.900	0.950	0.975	0.980	0.990	0.995	0.9995
1	0.325	0.727	1.376	3.078	6.314	12.706	15.895	31.821	63.657	636.619
2	0.289	0.617	1.061	1.886	2.920	4.303	4.849	6.965	9.925	31.599
3	0.277	0.584	0.978	1.638	2.353	3.182	3.482	4.541	5.841	12.924
4	0.271	0.569	0.941	1.533	2.132	2.776	2.999	3.747	4.604	8.610
5	0.267	0.559	0.920	1.476	2.015	2.571	2.757	3.365	4.032	6.869
6	0.265	0.553	0.906	1.440	1.943	2.447	2.612	3.143	3.707	5.959
7	0.263	0.549	0.896	1.415	1.895	2.365	2.517	2.998	3.499	5.408
8	0.262	0.546	0.889	1.397	1.860	2.306	2.449	2.896	3.355	5.041
9	0.261	0.543	0.883	1.383	1.833	2.262	2.398	2.821	3.250	4.781
10	0.260	0.542	0.879	1.372	1.812	2.228	2.359	2.764	3.169	4.587

Intervallo fiduciale della media (in altre parole…)

Dodo	Uova
A	0
B	9
C	6
D	3
E	1
F	5

$m=4.333 \quad s=4.163$

$s_e=s/\sqrt{n}=4.163/\sqrt{3}=2.404$

$t_{(n-1,p)}=t_{(3-1,0.95)}= \quad 4.303$

$m-t_{(n-1,p)}\cdot s_e < \mu < m+t_{(n-1,p)}\cdot s_e$

$4.333-4.303\cdot 2.404 < \mu < 4.333+4.303\cdot 2.404$

$-6.011 < \mu < 14.677$ per p=0.975 (97.5%)