



UNIVERSITÀ DEGLI STUDI DI MILANO

DIPARTIMENTO DI SCIENZE AGRARIE E AMBIENTALI
PRODUZIONE, TERRITORIO, AGROENERGIA

Metodologia Sperimentale Agronomica / Metodi Statistici per la Ricerca Ambientale

Marco Acutis

marco.acutis@unimi.it

www.acutis.it

a.a. 2018 - 2019

CdS Scienze della Produzione e Protezione delle Piante (g59)

CdS Biotecnologie Vegetali, Alimentari e Agro-Ambientali (g61)

CdS Scienze Agro-Ambientali (g57)

Lezione 01 - Sommario

- ❑ Introduzione
- ❑ Pre-requisiti:
 - Scale di misura
 - Statistica Descrittiva
- ❑ Statistica Inferenziale
 - Introduzione
 - Obiettivi
 - Distribuzioni
 - Stima puntuale dei parametri
 - Stima per intervalli dei parametri



Introduzione

Struttura del Corso

Lezioni frontali: 32 ore - 4 CFU

Esercitazioni in aula informatizzata: 32 ore - 2 CFU
(utilizzo del software di analisi statistica SPSS)

Modalità d'Esame

- prova scritta con domande a risposta chiusa
- presentazione di una relazione scritta
- prova orale

Testi consigliati

- Soliani L., Statistica applicata alla ricerca biologica e ambientale, UNI. NOVA Parma, 2003
- Quinn G.P. and Keough M.J., Experimental Design and Data Analysis for Biologists, Cambridge University Press, New York, 2002



Scale di misura delle variabili

Qualitative: nominali o ordinali

- l'unico parametro valutabile è la proporzione

Quantitative: intervalli o rapporti

- possono essere eseguiti dei calcoli, i parametri valutabili sono molti (statistiche descrittive numeriche: misure di posizione e di dispersione)
- possono essere discrete o continue

Statistica Descrittiva

Obiettivo: descrivere e sintetizzare i dati osservati attraverso grafici (es. distribuzioni di frequenza), indici di posizione e dispersione



Indici di posizione (1/3)

Indicano la tendenza centrale di un insieme di dati

Media aritmetica

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

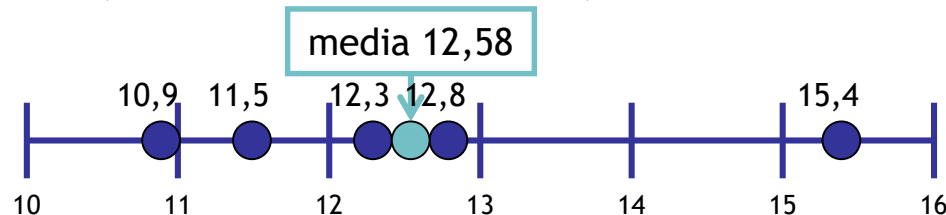
Proprietà della media aritmetica:

1) la sommatoria degli scarti di ogni dato dalla media (momento di I ordine) è nulla;

$$\sum_{i=1}^n (x_i - \bar{x}) = 0 \quad (x_1 - \bar{x}) + \dots + (x_n - \bar{x}) = \sum_{i=1}^n x_i - n\bar{x} = 0$$

2) la sommatoria del quadrato degli scarti (momento di II ordine) è minima (ovvero non esiste alcun altro punto che, sostituito alla media, dia un valore inferiore).

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \min$$



Indici di posizione (2/3)

se i dati sono espressi come frequenze:

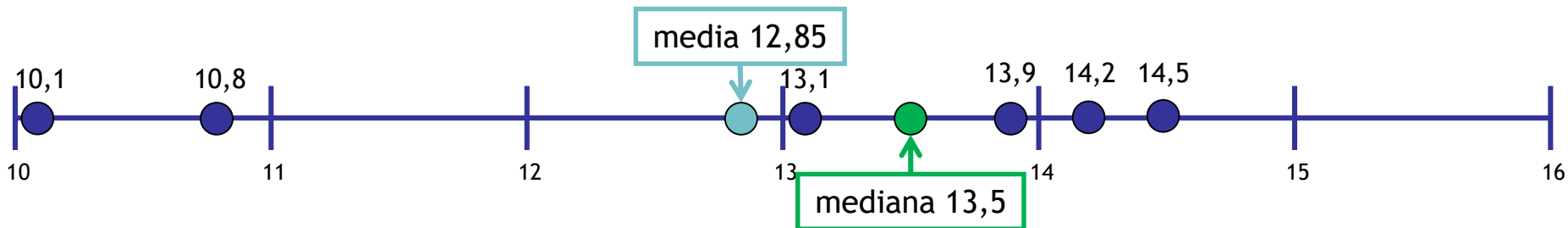
$$\bar{x} = \frac{\sum_{i=1}^n f_i x_i}{\sum_{i=1}^n f_i} \quad \text{media aritmetica ponderata}$$

se i dati sono espressi come proporzioni:

$$\bar{x} = \sum_{i=1}^n p_i x_i$$

Indici di posizione (3/3)

Mediana: divide la serie ordinata in due parti di uguale numerosità



Moda: è il valore della classe a cui corrisponde la maggiore frequenza

Media armonica: è il reciproco della media dei reciproci, idonea a mediare rapporti tra 2 variabili

$$m_h = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

Media geometrica: è la radice ennesima del prodotto di n dati, idonea per mediare tassi

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^n x_i}$$

Indici di tendenza centrale resistenti

Trimmed mean: media aritmetica nella quale non vengono considerate le code della distribuzione (es. il 5% dei dati)

M-estimators (Maximum likelihood estimators): media aritmetica pesata, dove il peso è funzione della distanza dal valore centrale; si differenziano per la funzione di assegnazione dei pesi

Indici di dispersione (1/4)

Quantili: misure di posizione non centrale, sono valori che dividono la serie ordinata in un certo numero di parti di uguale numerosità

➔ **Percentili:** dividono la serie ordinata in 100 parti uguali; il p-esimo percentile di una distribuzione è quel valore con p% dei valori inferiori ad esso (in statistica inferenziale sono interessanti 1, 2,5, 5, 95, 97,5 e 99 esimo percentile)

➔ **Quartili:** dividono la serie ordinata in 4 parti uguali: 25 esimo, 50 esimo (la mediana) e 75 esimo percentile

N.B. L'intervallo tra il 25 esimo e il 75 esimo percentile si chiama **distanza interquartile**

➔ **Decili:** dividono la serie ordinata in 10 parti uguali: 10, ..., 90 esimo percentile

Indici di dispersione (2/4)

Campo di variazione o Range: $x_{max} - x_{min}$

Scarti dalla media

1. Devianza o Sum of Squares: $SS = \sum_{i=1}^n (x_i - \bar{x})^2$

2. Varianza o Mean Square o Quadrato Medio:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

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

nel caso in cui i dati siano frequenze:

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

nel caso in cui i dati siano proporzioni:

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

Indici di dispersione (3/4)

3. Deviazione Standard o Standard Deviation:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

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

4. Coefficiente di Variazione o CV: $CV = \frac{s}{\bar{x}} 100$

Indici di dispersione (4/4)

Teorema di Tchebysheff: indipendentemente dalla distribuzione, fissata una costante K , l'intervallo $\bar{x} \pm Ks$ (dove s è la deviazione standard) contiene almeno $\left(1 - \frac{1}{K^2}\right)$ dati.



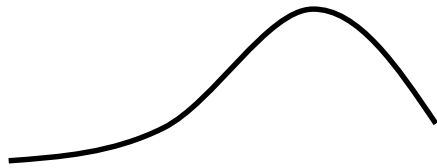
se $K=2$, allora l'intervallo contiene almeno il 75% dei dati
se $K=3$, allora l'intervallo contiene almeno l'89% dei dati

Approssimativamente, in una distribuzione simmetrica e «a campana»:
l'intervallo $\bar{x} \pm s$ contiene il 68% dei dati;
l'intervallo $\bar{x} \pm 2s$ contiene il 95% dei dati;
l'intervallo $\bar{x} \pm 3s$ contiene quasi il 100% dei dati.

Pre-requisiti - Statistica Descrittiva

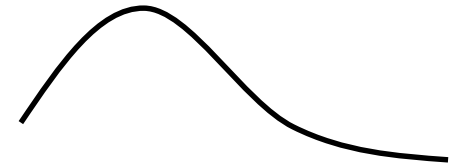
Indici di forma

Asimmetria o Skewness:
$$\frac{\sum_{i=1}^n (x_i - \bar{x})^3}{n \sigma^3}$$



negativa

positiva

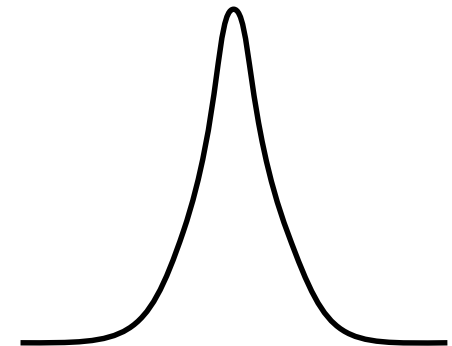


Curtosi o Kurtosis:
$$\frac{\sum_{i=1}^n (x_i - \bar{x})^4}{n \sigma^4}$$



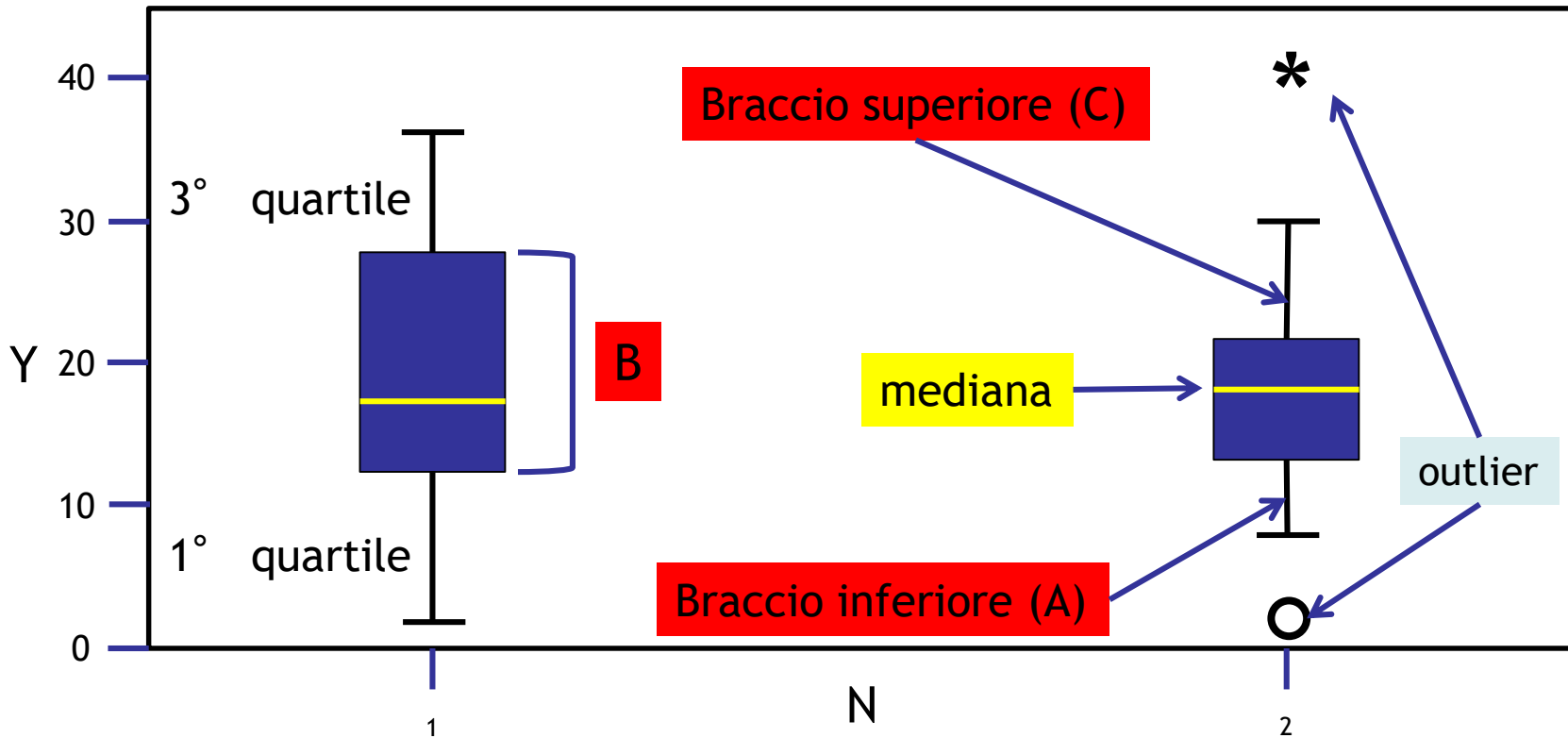
platicurtica

leptocurtica



Analisi Esplorativa dei Dati: Box Plot

Tra i più comuni strumenti grafici della Exploratory Data Analysis (oltre ai grafici a barre e agli istogrammi) ricordiamo il Box Plot.



Convenzioni Box Plot (1/2)

- ✓ rappresentazione grafica della distribuzione dei dati (per data set sufficientemente numerosi)
- ✓ sull'asse delle ordinate (Y) sono riportati in scala i valori assunti dalla variabile in esame
- ✓ i dati compresi tra il 1° e il 3° quartile sono rappresentati su piano cartesiano da un rettangolo
- ✓ il rettangolo è tagliato da una linea che rappresenta la mediana (o 2° quartile) a cui corrisponde una frequenza cumulata pari al 50%
- ✓ il braccio (o baffo) inferiore (A) rappresenta la distanza tra il valore minimo della serie di dati e il 1° quartile
- ✓ B è la distanza interquartile tra il 1° e il 3° quartile
- ✓ il braccio superiore (C) rappresenta la distanza tra il valore massimo della serie di dati e il 3° quartile

Pre-requisiti - Statistica Descrittiva

Convenzioni Box Plot (2/2)

- ✓ un braccio (A o C) può avere una lunghezza massima pari a $1.5 \times B$
- ✓ se A è maggiore di $1.5 \times B$, allora il valore minimo viene posto nel grafico fuori dal braccio e rappresentato come un dato «outlier»
- ✓ se il valore del 1° quartile è anche il valore minimo dei dati, allora il braccio non è rappresentato
- ✓ se C è maggiore di $1.5 \times B$, allora il valore massimo viene posto nel grafico fuori dal braccio e rappresentato come un dato «outlier»
- ✓ se il valore del 3° quartile è anche il valore massimo dei dati, allora il braccio non è rappresentato

○ dato inferiore (superiore) rispetto al valore del 1° (3°) quartile diminuito (sommato) di un valore compreso tra 1.5 e 3 moltiplicato per B

* dato inferiore (superiore) rispetto al valore del 1° (3°) quartile diminuito (sommato) di un valore maggiore di 3 moltiplicato per B



Popolazione e Campione

POPOLAZIONE

- ✓ insieme di tutte le manifestazioni relative a un certo fenomeno
- ✓ può essere finito o infinito
- ✓ in genere ci si occupa di popolazioni molto grandi

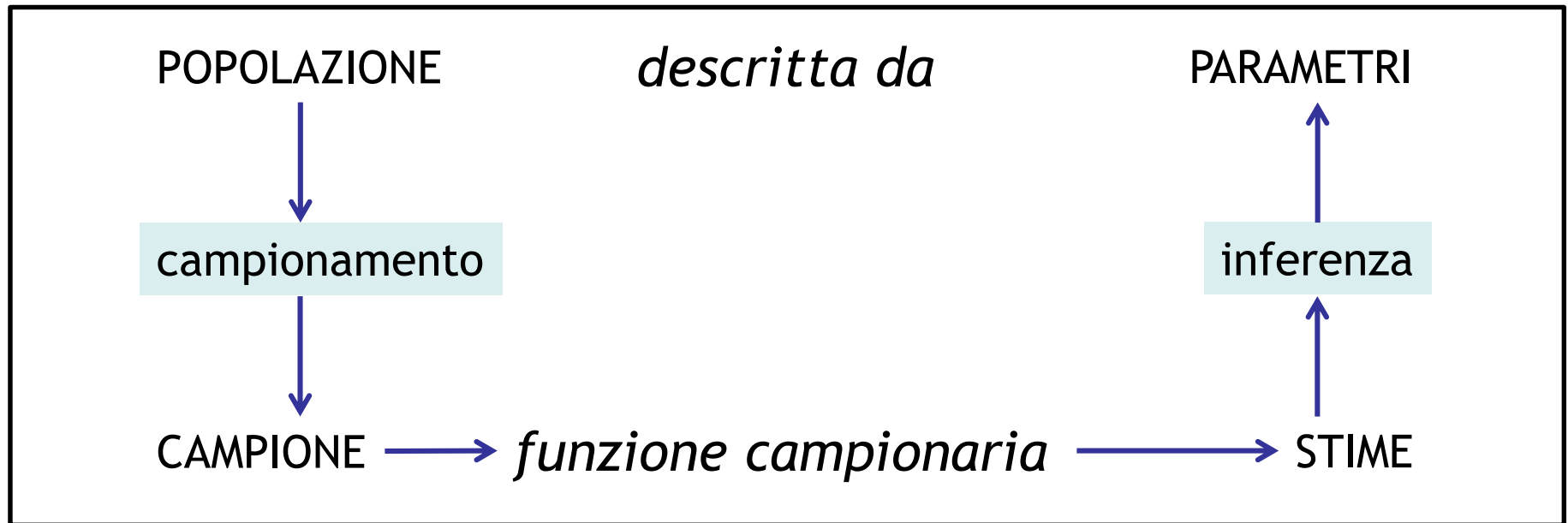
CAMPIONE

- ✓ sottoinsieme della popolazione
- ✓ se estratto casualmente, rappresenta la popolazione in esame



Obiettivi dell'Inferenza Statistica

1. Stima dei parametri della popolazione
2. Test delle ipotesi



Distribuzioni di probabilità

Variabile Casuale: variabile che assegna un valore a ciascuna realizzazione di un esperimento; può essere discreta o continua

Distribuzione di probabilità: funzione che rappresenta la probabilità associata a ciascun valore della variabile casuale

→ a VC discrete/continue si associano distribuzioni di probabilità discrete/continue

→ la distribuzione di probabilità è la distribuzione teorica della popolazione, **i cui parametri si intendono indagare**

→ la media di una distribuzione di probabilità è detta valore atteso della variabile casuale

Distribuzioni di probabilità d'interesse

- Distribuzione Binomiale
- Distribuzione Normale
- Distribuzione del t di Student
- Distribuzione F di Fisher
- Distribuzione del X^2 (chi quadro)

- Distribuzione di Poisson
- Distribuzione del Q
- Distribuzione Binomiale Negativa
- Distribuzione Gamma
- Distribuzione Beta
- Distribuzione di Cauchy
- Distribuzione di Gumbel
- Distribuzione di Weibull
- Distribuzione Log-Normale
- ...



Distribuzione Normale o Gaussiana (1/2)

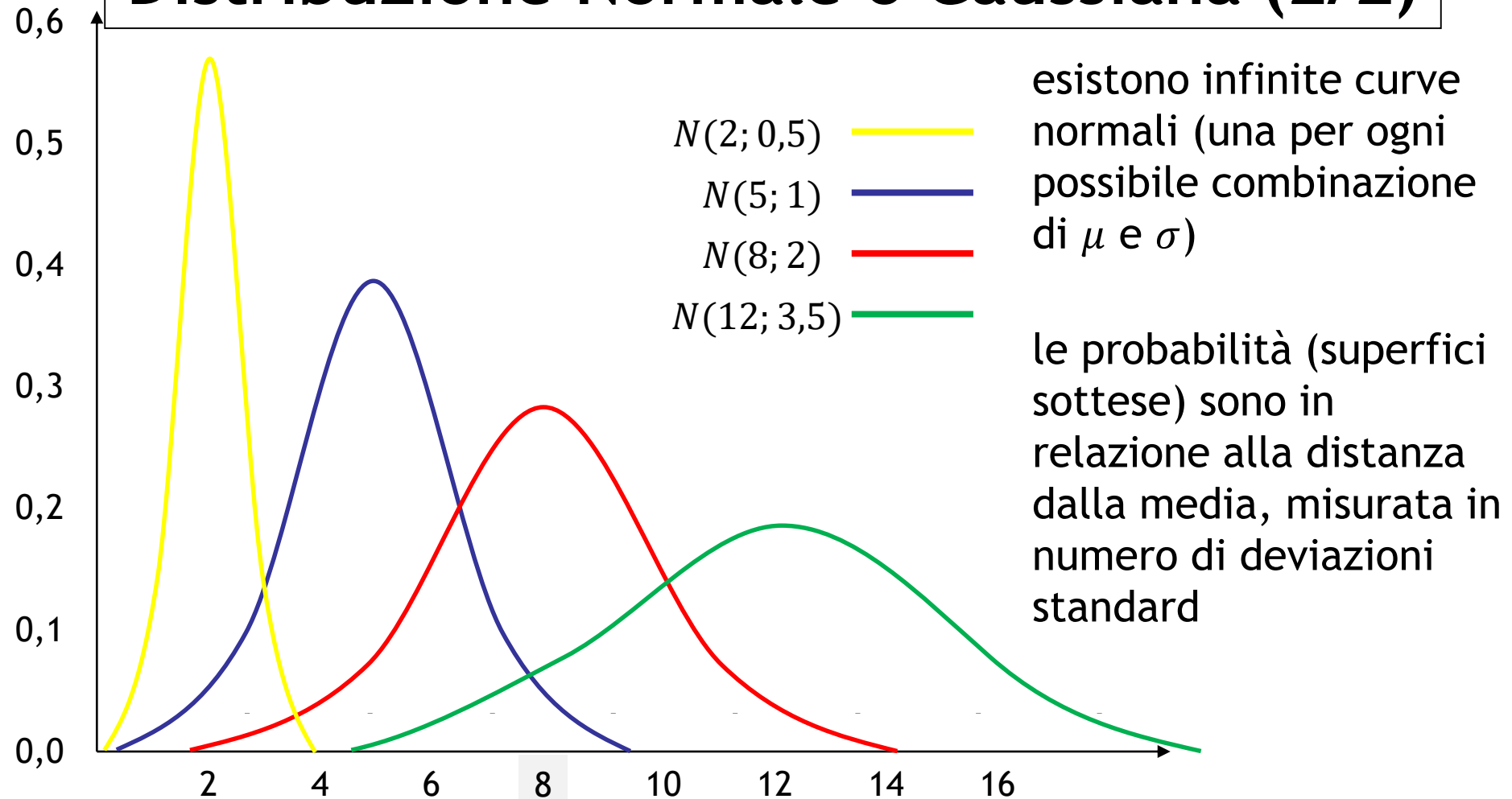
$$y = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Principali proprietà

- la variabile casuale x è compresa tra $-\infty$ e $+\infty$
- è completamente definita da 2 parametri (**media** e **varianza**) e viene sinteticamente indicata con $N(\mu; \sigma^2)$
- è simmetrica intorno alla media ed è a forma di campana
- ha il massimo in $x = \mu$ e 2 flessi in $x = \mu \pm \sigma$

N.B. L'integrale di $N(\mu; \sigma^2)$ tra x e $+\infty$ fornisce la probabilità che un'unità sperimentale abbia un valore superiore a x

Distribuzione Normale o Gaussiana (2/2)



Distribuzione Normale Standardizzata

- ✓ Tra le curve normali, si fa spesso riferimento alla cosiddetta “Normale Standardizzata”, che si indica con $N(0; 1)$ e quindi ha media = 0 e varianza = deviazione standard = 1.
- ✓ Tutte le normali possono essere ricondotte alla normale standardizzata, sottraendo a ogni dato la media e dividendo per la deviazione standard $\left(z = \frac{x - \mu}{\sigma} \right)$.
- ✓ La distribuzione normale standardizzata è detta anche Distribuzione Z .
- ✓ L'integrale della normale $N(\mu; \sigma^2)$ tra x e $+\infty$ è calcolabile, ma con notevole difficoltà, mentre l'integrale di Z è tabulato.

Statistica Inferenziale - Stima puntuale dei parametri

Stimatori (1/2)

Uno **stimatore** è una statistica ottenuta da un campione che stima un parametro della popolazione. Si indica con lettera latina, mentre i parametri della popolazione si indicano con lettera greca.

La media campionaria \bar{x} è uno stimatore della media della popolazione μ .

La varianza campionaria s^2 è uno stimatore della varianza della popolazione σ^2 .

La deviazione standard campionaria s è uno stimatore della deviazione standard della popolazione σ .



Statistica Inferenziale - Stima puntuale dei parametri

Stimatori (2/2)

Proprietà

Non distorsione (accuratezza): la media di tutti i possibili valori dello stimatore è uguale al valore del parametro della popolazione.

Consistenza: all'aumentare della dimensione del campione lo stimatore tende al valore del parametro.

Efficienza (precisione): è più efficiente, tra tutti gli stimatori non distorti, quello che ha minore varianza campionaria.



Stimatori di media e varianza

Il miglior stimatore della media di una popolazione è la media del campione:

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

Il miglior stimatore della varianza di una popolazione è: $s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$

N.B. Se si divide per n anziché per $n-1$, lo stimatore risulta distorto!

Non vi sono stimatori non distorti della deviazione standard: è per questo che si usa molto la varianza.

Statistica Inferenziale - Stima puntuale dei parametri

dimostrazione...

Campioni		Statistiche				
x_1	x_2	media	var(/n)	var(/n-1)	dev.st(/n)	dev.st(/n-1)
2	2	2,0	0,00	0,00	0,00	0,00
2	3	2,5	0,25	0,50	0,50	0,71
2	5	3,5	2,25	4,50	1,50	2,12
2	6	4,0	4,00	8,00	2,00	2,83
3	2	2,5	0,25	0,50	0,50	0,71
3	3	3,0	0,00	0,00	0,00	0,00
3	5	4	1,00	2,00	1,00	1,41
3	6	4,5	2,25	4,50	1,50	2,12
5	2	3,5	2,25	4,50	1,50	2,12
5	3	4,0	1,00	2,00	1,00	1,41
5	5	5,0	0,00	0,00	0,00	0,00
5	6	5,5	0,25	0,50	0,50	0,71
6	2	4,0	4,00	8,00	2,00	2,83
6	3	4,5	2,25	4,50	1,50	2,12
6	5	5,5	0,25	0,50	0,50	0,71
6	6	6,0	0,00	0,00	0,00	0,00
Media stimatori		4	1,25	2,50	0.88	1,24
Varianza stimatori		1,25	1,844	7,375	0,484	0,969
Deviazione Standard stimatori		1,118	1,358	2,716	0,696	0,984

Universo	{2, 3, 5, 6}
Media	4
Deviazione Standard	1,6
Varianza	2,5

Possibili campioni di numerosità 2 ottenibili per estrazione casuale con re-immisione



Teorema del Limite Centrale

Una variabile che derivi dalla somma di altre tende a essere distribuita normalmente. Tante più variabili concorrono alla somma, tanto più l'approssimazione è buona.



Le medie campionarie, anche se i campioni sono tratti da popolazioni con distribuzioni diverse dalla normale, tendono ad essere distribuite normalmente.
L'approssimazione è tanto maggiore quanto maggiore è la numerosità campionaria.

Distribuzione campionaria delle medie

la distribuzione campionaria della media di un campione di numerosità n , estratto casualmente da una popolazione con media μ e varianza σ^2 , ha:

media = μ (stimatore non distorto)

$$\text{varianza} = \frac{\sigma^2}{n}$$

$$\text{deviazione standard} = \frac{\sigma}{\sqrt{n}}$$

Inoltre, per il teorema del limite centrale, se n è sufficientemente grande, la distribuzione delle medie campionarie è normale.

Errore Standard della media

La deviazione standard della distribuzione delle medie campionarie, più piccola di σ di un fattore $\frac{1}{\sqrt{n}}$, si chiama

Errore Standard o deviazione standard della media o errore di campionamento della media.

$$es = \frac{s}{\sqrt{n}}$$

N.B. Se siamo interessati alla variabilità delle misurazioni, usiamo la deviazione standard. Se invece vogliamo mettere in evidenza l'errore che si commette stimando la media della popolazione a partire dalla media campionaria, calcoliamo l'errore standard della media. Va da sé che ogni volta che estraiamo un campione da una popolazione, la sua media varia.

Stime per intervalli (1/2)

Oltre al valore puntuale di una stima, è interessante conoscere qual è il **margin di errore connesso alla stima stessa**.

Si possono stabilire dei limiti entro i quali si ha una certa «confidenza» $(1 - \alpha)$ che sia compreso il vero valore del parametro nella popolazione.

Questi limiti si chiamano **LIMITI FIDUCIALI** e l'intervallo che definiscono si definisce **INTERVALLO FIDUCIALE** o **INTERVALLO DI CONFIDENZA**.

Stime per intervalli (2/2)

La stima di un parametro fatta a partire da un campione, corredata dai suoi limiti fiduciali, è detta **STIMA PER INTERVALLI**.

I valori usuali di α sono 0,01, 0,05 e 0,1, che danno luogo, rispettivamente, agli intervalli fiduciali o intervalli di confidenza del 99%, 95% e 90%.

Per definire un intervallo di confidenza si utilizzano le distribuzioni campionarie.



Statistica Inferenziale - Stima per intervalli dei parametri

Intervallo di confidenza di una media (σ noto) (1/4)

Data una popolazione di cui si conosce la deviazione standard σ e di cui si vuole stimare la media, si estrae da essa un campione di numerosità n .

Facendo riferimento alla distribuzione delle medie campionarie, sappiamo che la media del campione appartiene alla popolazione di medie campionarie, la quale ha:

- distribuzione normale;
- stessa media della popolazione di partenza;
- deviazione standard = $\frac{\sigma}{\sqrt{n}}$.

Si tratta, in questa distribuzione normale, di individuare l'intervallo che esclude $\alpha/2$ per lato. In tal modo questo intervallo avrà probabilità $(1 - \alpha)$ di includere la vera media della popolazione.



Statistica Inferenziale - Stima per intervalli dei parametri

Intervallo di confidenza di una media (σ noto) (2/4)

Se σ è noto, si fa riferimento alla distribuzione $Z = N(0; 1)$

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \quad \text{con} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Definito un grado di confidenza α , si ha:

$$P\left[\mu - z_{\alpha/2}(\sigma/\sqrt{n}) \leq \bar{x} \leq \mu + z_{\alpha/2}(\sigma/\sqrt{n})\right] = (1 - \alpha)$$

e di conseguenza

$$P\left[\bar{x} - z_{\alpha/2}(\sigma/\sqrt{n}) \leq \mu \leq \bar{x} + z_{\alpha/2}(\sigma/\sqrt{n})\right] = (1 - \alpha)$$



Statistica Inferenziale - Stima per intervalli dei parametri

Intervallo di confidenza di una media (σ noto) (3/4)

Fissando, ad esempio, il grado di confidenza $(1 - \alpha) = 0,95$ e conoscendo il valore tabulato $z_{\alpha/2} = 1,96$, ne consegue che l'intervallo di confidenza della media sarà:

$$P[\bar{x} - 1,96(\sigma/\sqrt{n}) \leq \mu \leq \bar{x} + 1,96(\sigma/\sqrt{n})] = 0,95$$

Valori critici usuali di $z_{\alpha/2}$ sono:

$$z_{0,05} = 1,64 \quad (\text{per confidenza del } 90\%);$$

$$z_{0,025} = 1,96 \quad (\text{per confidenza del } 95\%);$$

$$z_{0,005} = 2,57 \quad (\text{per confidenza del } 99\%).$$



Statistica Inferenziale - Stima per intervalli dei parametri

Intervallo di confidenza di una media (σ noto) (4/4)

$z_{\alpha/2}(\sigma/\sqrt{n})$ è la quantità che viene aggiunta e sottratta alla media campionaria per avere l'intervallo. Si chiama **massimo errore di stima**, ed è un indicatore della precisione della stima.

A parità di σ , i limiti fiduciali si restringono all'aumentare di:

1) α (e quindi al diminuire del grado di confidenza)

→ si esclude un'area di curva maggiore, ma aumenta la possibilità che i limiti non contengano il vero valore di μ

2) n

→ non vi sono controindicazioni, se non il costo o l'onere di un campione più grande



Statistica Inferenziale - Stima per intervalli dei parametri

Intervallo di confidenza di una media (σ ignoto) (1/2)

Ipotizziamo che dal campione si debbano stimare sia la media che la deviazione standard. In questo caso, non si può usare la distribuzione di Z , poiché per usare Z occorre conoscere σ . Si deve pertanto ricorrere alla distribuzione **t di Student**.

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} \quad \text{con} \quad s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Analogamente a quanto visto in precedenza, i limiti fiduciali per una confidenza $(1 - \alpha)$ saranno dati da:

$$\bar{x} \pm t_{\alpha/2}(s/\sqrt{n}) \quad \text{dove si considera una distribuzione di t con } n - 1 \text{ gradi di libertà}$$

Gli intervalli fiduciali saranno più “larghi” di quelli con σ nota, poiché vi sono due stime (\bar{x} e s) soggette a fluttuazioni campionarie.



Statistica Inferenziale - Stima per intervalli dei parametri

Intervallo di confidenza di una media (σ ignoto) (2/2)

ESEMPIO

Avendo rilevato produzioni di un pascolo, si sono ottenuti i seguenti valori ($t \text{ ha}^{-1}$ di sostanza secca): 3,6; 4,3; 4,8; 3,3; 3,2; 2,8; 4,1; 4,8; 3,3. Calcolare la produzione media ed i suoi limiti fiduciali al 90%, al 95% e al 99%.

SOLUZIONE

Limiti fiduciali della media		
α	Limite inferiore	Limite superiore
0,1	3,35	4,25
0,05	3,24	4,36
0,01	2,99	4,61

Media	3,800	Valori t tabulati	
Devianza	4,240	α	t_{9-1}
Stima varianza	0,53	0,1	1,860
Stima deviazione standard	0,728	0,05	2,306
Stima errore standard	0,243	0,01	3,355

