

Organizzazione e utilizzo di dataset di dati
anamnestici ed esiti diagnostici:
raccolta ed archiviazione dei dati

ROMA

7/8 giugno 2016

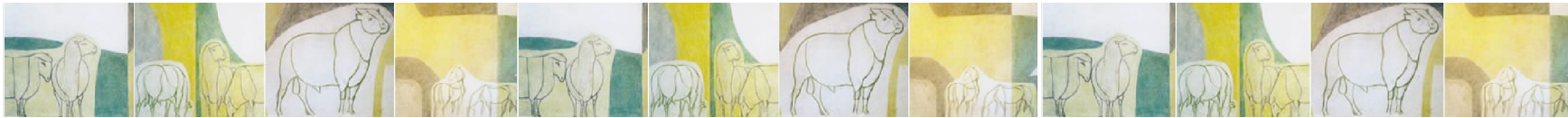


Tipologia di dati e di variabili

Francesca Iaconi

Osservatorio Epidemiologico

Istituto Zooprofilattico Sperimentale Lazio e Toscana



Unità statistica

è il soggetto elementare dell'indagine statistica per la sua appartenenza ad una popolazione di interesse

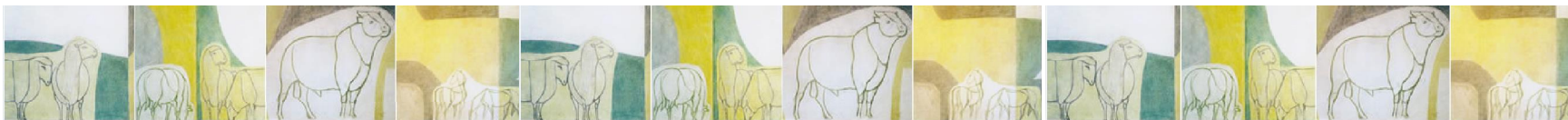
É *Individuo*

É *Gruppo*

É *Allevamento*

É *Azienda*

al quale si riferisce la raccolta dati



I dati

Per rispondere a domande specifiche:

- quanto latte produce in media una vacca alla prima lattazione nelle aziende del Lazio?
- quale è il gruppo di vacche che produce più latte nella mia azienda?
- quale è il gruppo di ovini del mio gregge con maggiori problemi di brucellosi?



RACCOLTA DATI





I dati

Il materiale di base della statistica è costituito dai DATI :

- *Numeri*
- *Conteggi*
- *Misure*
- *Espressioni di caratteristiche*



La variabile

Quando raccogliamo un dato, stiamo determinando una caratteristica di una unità statistica. Raccogliendo il dato su più unità statistiche...

Variabile: qualsiasi caratteristica (attributo) si presenti con modalità diverse, da unità a unità o nella stessa unità da un momento all'altro

Tutti i possibili valori di una variabile vengono chiamati **modalità**



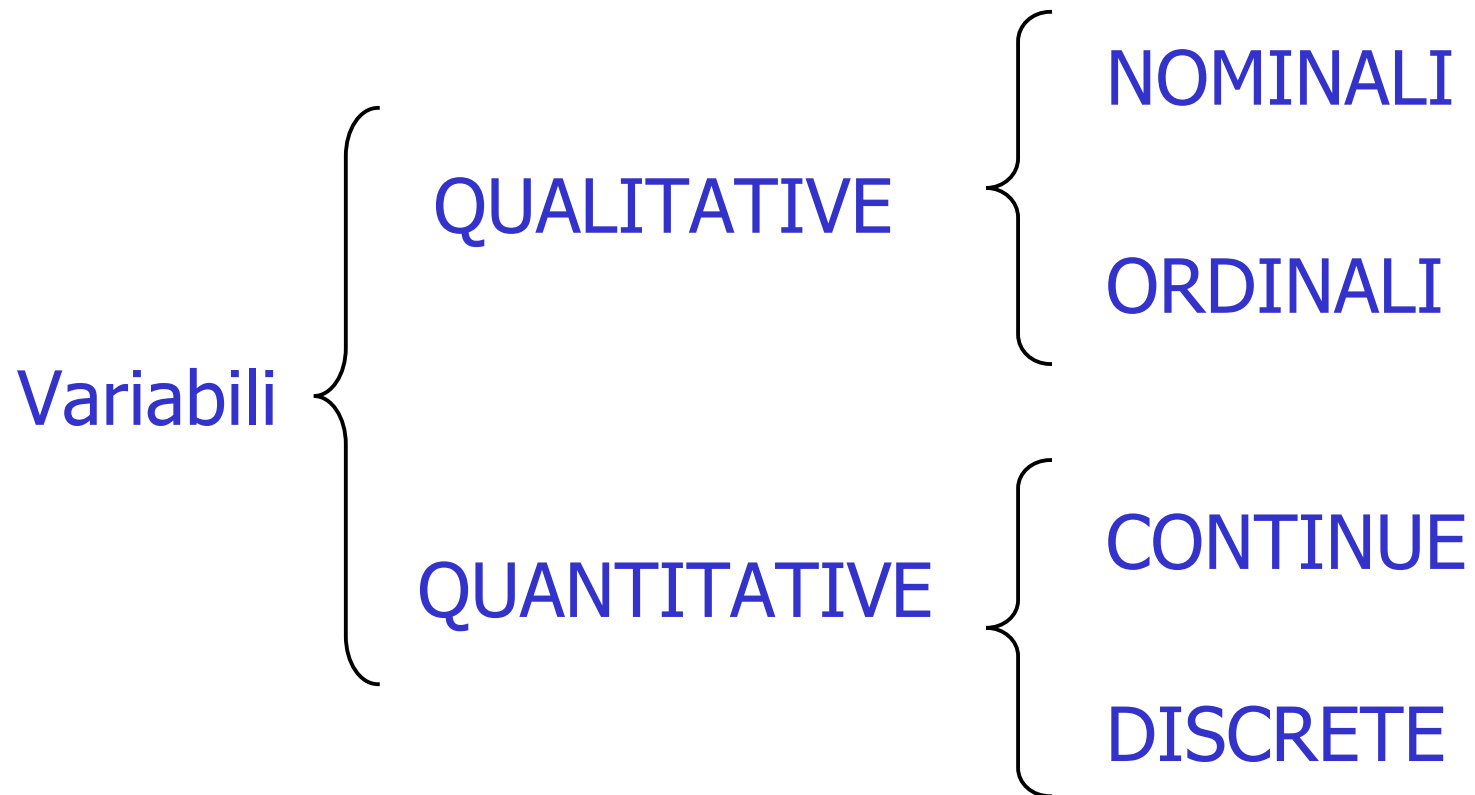
La variabile

Tale caratteristica può assumere valori diversi nei diversi individui della popolazione o negli stessi individui da un momento all'altro.

- *età*
- *razza*
- *numero di cellule somatiche nel latte*
- *durata dell'interparto*
- *causa della morte/riforma precoce*

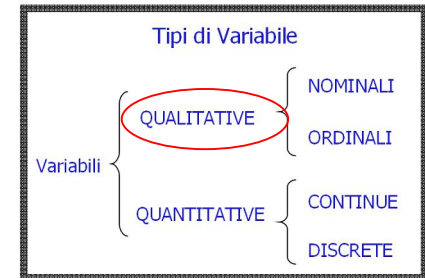


Tipi di Variabile





Tipi di Variabile



Variabili **QUALITATIVE**

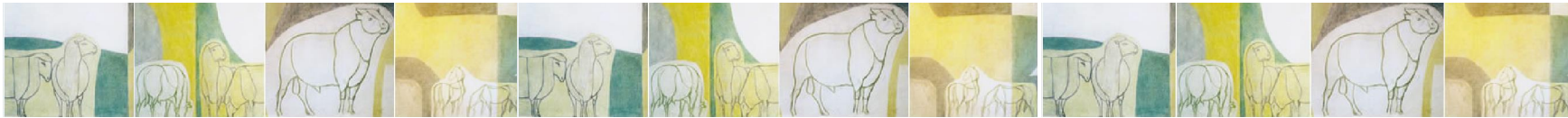
Assumono forma verbale e non un valore numerico



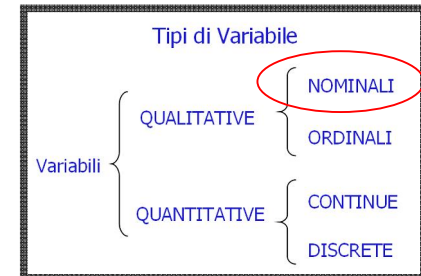
NOMINALI

ORDINALI

(dette anche sconnesse)



Tipi di variabile



QUALITATIVE – NOMINALI

Non possono essere poste in un sistema di ordinamento

Sesso: maschio, femmina, castrone

Razza: Frisona, Jersey – Sarda, Comisana...

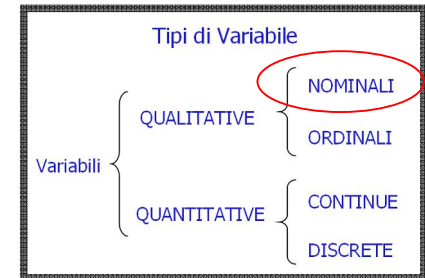
Gruppo: asciutta, fresca, gravida

Provenienza: Italia, Estero – Toscana, Lazio...

Codice azienda: 01VT12, 15RM56...



Tipi di variabile



QUALITATIVE – NOMINALI (caso particolare)

DICOTOMICHE

Positivo/Negativo

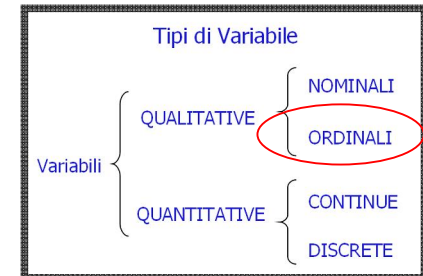
Vivo/Morto

Maschio/Femmina

Successo/Fallimento



Tipi di variabile



QUALITATIVE – ORDINALI

Possiedono un ordine "naturale"

E' possibile ordinare secondo un ordine crescente o decrescente

Parità (Nullipara, Primipara, Pluripara)

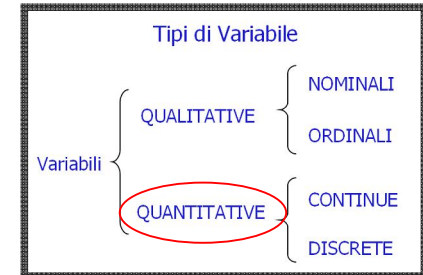
Indici e punteggi clinici

Caratteristiche di razza

Classe 1	Classe 2	Classe 3	Classe 4	Classe 5
Molto Magro	Magro	Normale	Grasso	Molto Grasso



Tipi di variabile



QUANTITATIVE

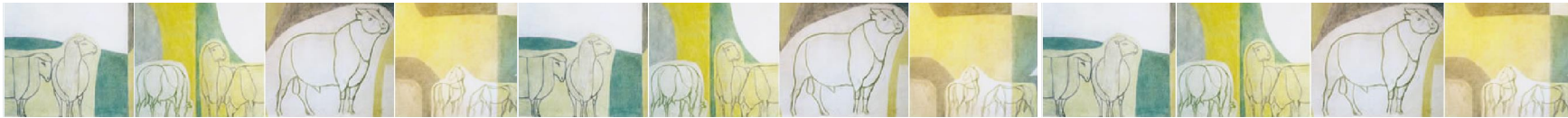
assumono valore numerico



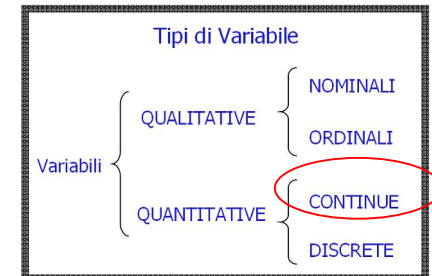
CONTINUE



DISCRETE



Tipi di variabile



QUANTITATIVE – CONTINUE

Possono assumere un qualsiasi valore all'interno di un ragionevole range – numeri reali.

Altezza al garrese: 117,5 cm

Peso: 198,3 Kg

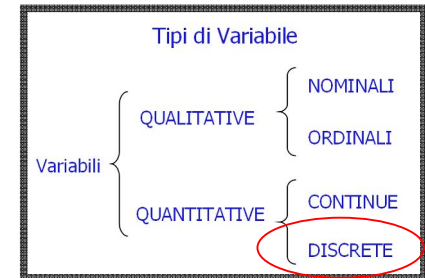
Esito: 0,005 mg/kg

In genere sono quantità che si misurano





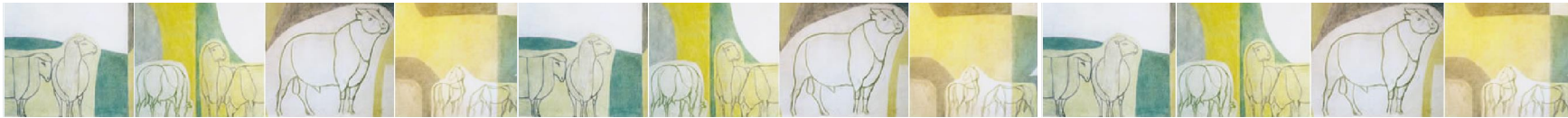
Le Variabili



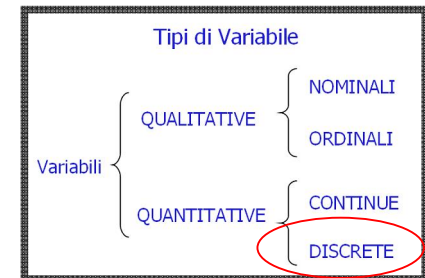
Variabili **QUANTITATIVE – DISCRETE**

Assumono valori fissi all'interno dei numeri naturali (non sono frazionabili)

Es. i conteggi sono variabili quantitative discrete

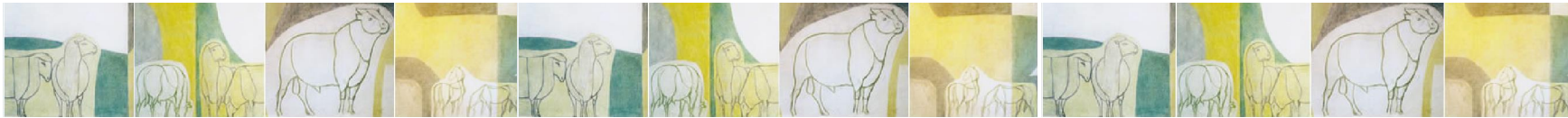


Le Variabili

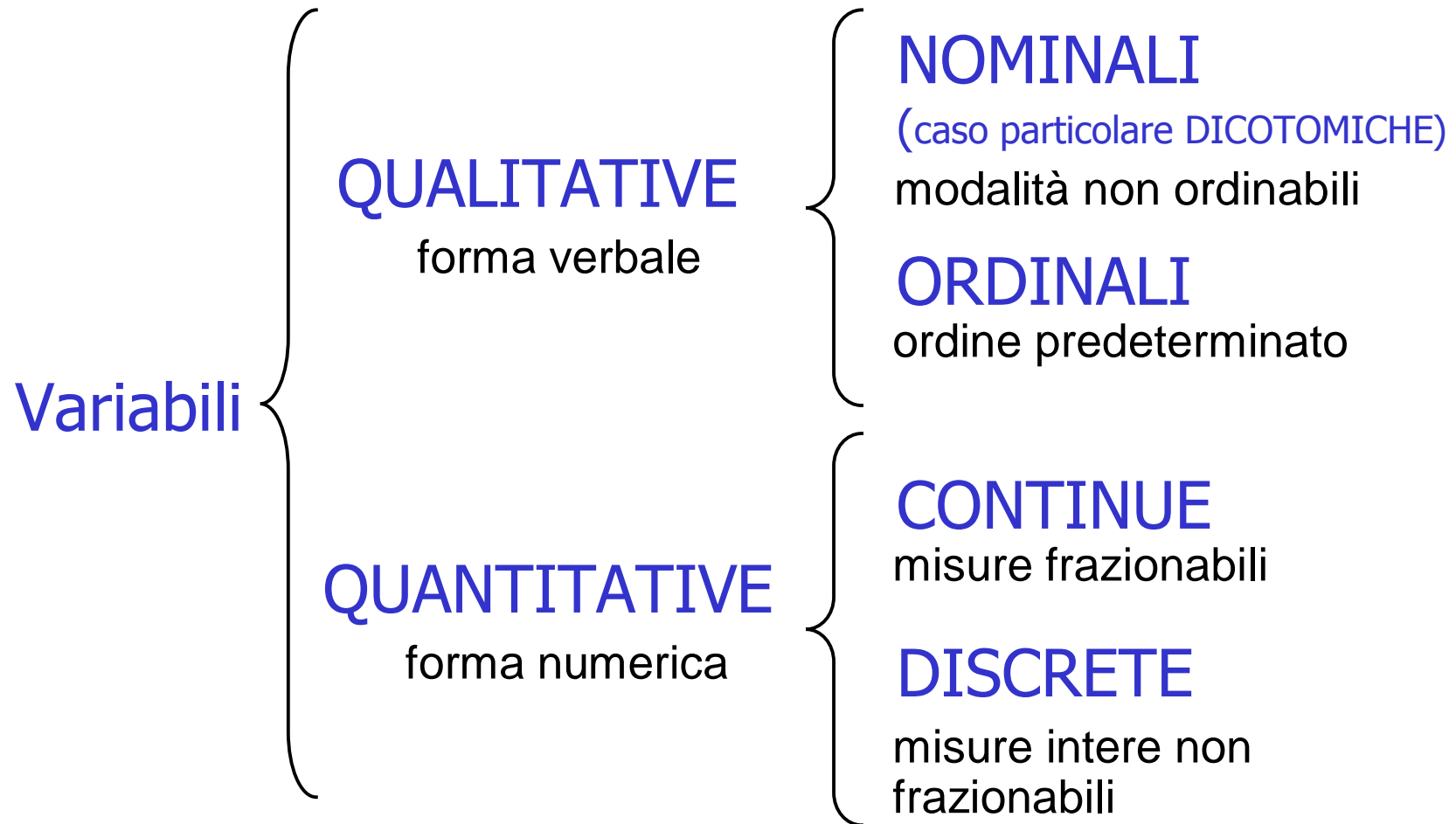


Es. variabili quantitative discrete

- *Consistenza dell'allevamento (10, 50, 100 ...)*
- *Numero di aborti/capo*
- *Numero di morti in allevamento*
- *Numero di casi di malattia*
- *Numero di parti (0, 1, 2 ...)*
- *Numero di lattazioni (0, 1, 2 ...)*



Tipi di Variabile





Variabili di esposizione

La definizione dell'esposizione varia a seconda dell'ipotesi di ricerca.

La variabile identificata come **variabile di esposizione** esprime i valori del fattore di cui si intende misurare l'associazione con l'esito (outcome).

Può quindi ipoteticamente causare un certo effetto e di conseguenza condizionare l'esito

		polmonite		Esito (Outcome)
		+	-	
Variabile di esposizione	chiuso	240	230	470
	aperto	160	1070	1230
		400	1300	1700



Variabili di raggruppamento

Nel dataset, può essere utile creare una variabile che discrimini due o più gruppi (per sesso, trattamento...) che devono essere confrontati in funzione di un outcome (positivi/negativi).

Tali gruppi rappresentano i diversi livelli (CLASSI) di esposizione

...ne parleremo più avanti



Variabili di raggruppamento- Esempio



GRUPPO	Incremento medio giornaliero (gr)
Trattato	639
Trattato	646
Trattato	650
Controllo	631
Controllo	650
Controllo	633

→ Variabile di raggruppamento

Si misura l'incremento medio giornaliero di peso in due gruppi (trattato/controllo)

Possiamo confrontare l'incremento medio giornaliero del gruppo trattato vs il gruppo di controllo.

La variabile di raggruppamento può essere quantitativa (discreta o continua suddivisa in classi) oppure qualitativa.



E' l'obiettivo che fa la variabile!

Polmonite	SESSO	Peso (Kg)
si	M	1095
no	F	700
no	M	990
no	M	800
si	F	850
si	F	730
no	F	780

Obiettivo 1: Vogliamo vedere la differenza di peso tra i sessi



Variabile di esposizione= SESSO
Variabile di outcome = PESO

Obiettivo 2: Vogliamo vedere le differenze di stato sanitario in base al peso



Variabile di esposizione= PESO
Variabile di outcome = polmonite



Statistica descrittiva

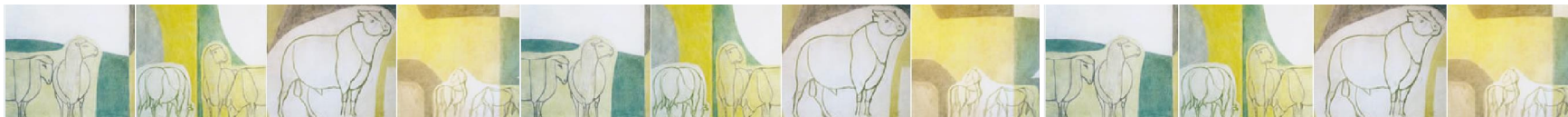
L'obiettivo principale della statistica descrittiva è quello di fornire chiavi di lettura dei fenomeni osservati di rapida ed immediata interpretazione.

Gli **indici di posizione** rappresentano uno degli strumenti più utilizzati per questo scopo e sono in grado di riassumere in un unico valore l'andamento generale dell'intera distribuzione



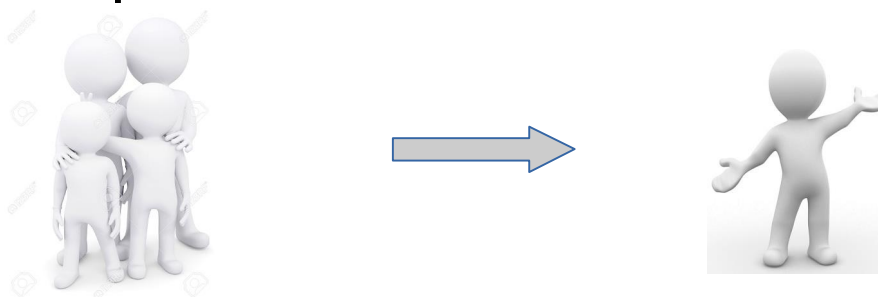
Statistica descrittiva

La sintesi dei dati comporta una **perdita di informazioni**, deve quindi essere privilegiato l'indice di posizione che minimizza la perdita e rappresenta nel modo più corretto l'insieme dei dati osservati

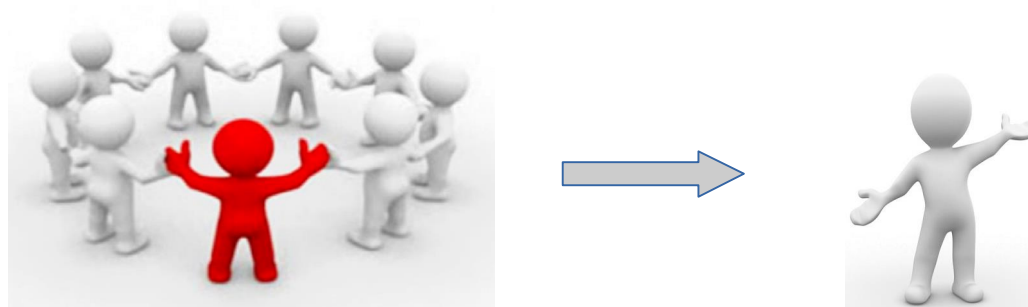


Indici di sintesi

deve essere compreso tra il dato più piccolo ed il dato più elevato della distribuzione



deve identificarsi con i valori più frequenti



Misure di “tendenza centrale” o di posizione



La media aritmetica

La media aritmetica è quel valore che rileva la tendenza centrale della distribuzione.

Rappresenta la parte del totale del fenomeno in esame che spetterebbe a ciascuna unità statistica.

$$Media = \frac{\sum_{i=1}^N x_i}{N}$$



La media aritmetica- Esempio

Unità statistiche /progressivo capo	peso (Kg)
1	1095,2
2	990,3
3	734,2
4	998,5
5	550
6	592,4
7	1111,5
8	754,6
9	882,1
10	950,3
11	925,4
12	1250,6
13	754,8
14	950,8
15	835,4
16	555,7
17	890,8
18	900
19	591,8
20	899,3
<u>Somma</u>	17213,7

Totale peso (Kg)	17213,7
Totale osservazioni	20

$$Media = \frac{17213,7}{20} = 860,68$$



La media aritmetica- Proprietà

- **Unicità:** posto un insieme di dati vi è una sola media aritmetica
- **Semplicità:** la media aritmetica è facile da capire e calcolare
- **La media è influenzata dai valori estremi**



La media aritmetica- Esempio

Parcella di un veterinario per visita in allevamento.

5 veterinari (valori in EURO)

75, 75, 80, 80, 280

Media = $(75+75+80+80+280)/5 = 118$ Euro

Sovrastima dovuta al valore estremo 75, 75, 80, 80, 280

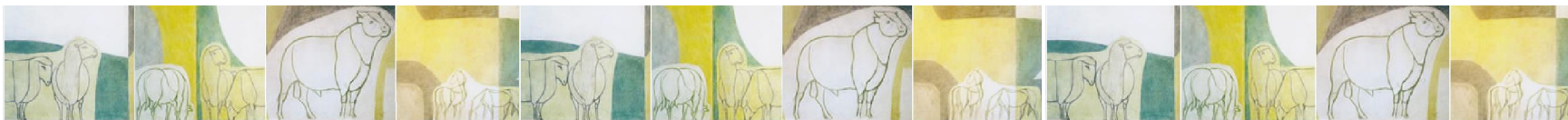


La mediana

In una serie finita e ordinata di osservazioni è il valore che divide la serie in 2 parti uguali

il numero di osservazioni uguale o minore alla mediana è uguale al numero di osservazioni uguale o maggiore alla mediana

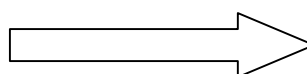
il 50% delle osservazioni sta sotto la mediana ed il 50% sta sopra la mediana



La mediana- Esempio

Unità statistiche /progressivo capo	peso (Kg)
1	1095,2
2	990,3
3	734,2
4	998,5
5	550
6	592,4
7	1111,5
8	754,6
9	882,1
10	950,3
11	9845
12	925,4
13	1250,6
14	754,8
15	950,8
16	835,4
17	555,7
18	890,8
19	900
20	591,8
21	899,3

Ordino in senso crescente



Unità statistiche /progressivo capo	peso (Kg)
5	550
17	555,7
20	591,8
6	592,4
3	734,2
8	754,6
14	754,8
16	835,4
9	882,1
18	890,8
21	899,3
19	900
12	925,4
10	950,3
15	950,8
2	990,3
4	998,5
1	1095,2
7	1111,5
13	1250,6
11	9845

50%

mediana

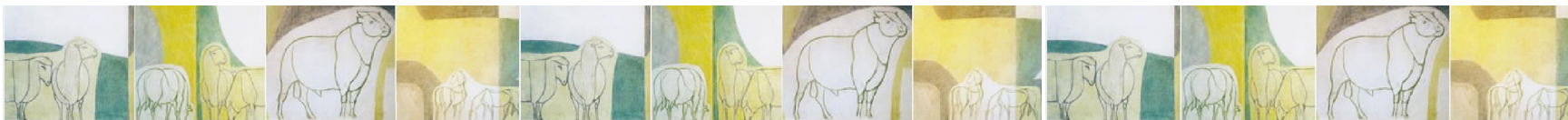
50%

NB. Posizione della mediana = $(21+1)/2 = 11$ esima osservazione



La mediana- Calcolo

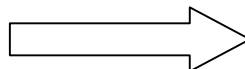
- se il numero di osservazioni è pari non vi è un solo valore mediano
- ci sono 2 osservazioni mediane
- la mediana è la media di queste 2 osservazioni mediane



La mediana-Esempio 2

Unità statistiche /progressivo capo	peso (Kg)
1	1095,2
2	990,3
3	734,2
4	998,5
5	550
6	592,4
7	1111,5
8	754,6
9	882,1
10	950,3
11	9845
12	925,4
13	1250,6
14	754,8
15	950,8
16	835,4
17	555,7
18	890,8
19	900
20	591,8

**Ordino in
senso
crescente**

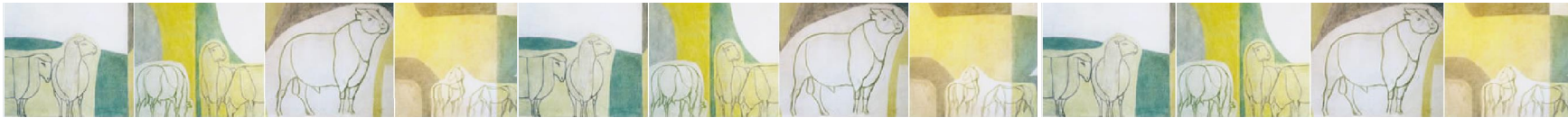


Unità statistiche /progressivo capo	peso (Kg)
5	550
17	555,7
20	591,8
6	592,4
3	734,2
8	754,6
14	754,8
16	835,4
9	882,1
18	890,8
19	900
12	925,4
10	950,3
15	950,8
2	990,3
4	998,5
1	1095,2
7	1111,5
13	1250,6
11	9845

$(890,8 + 900) / 2 = 908,1$

mediana

NB. Posizione della mediana = tra la $(20/2)$ e $(20/2)+1$ osservazione



La mediana- Calcolo

- É Ordino le N osservazioni in modo crescente
- É Attribuisco un progressivo crescente ad ogni osservazione (1,2,...,N)
- É Applico le seguenti formule generali:
 - É A: se N dispari
 $[(n+1)/2]$ esima osservazione
 - É B: se N pari, calcolo la media tra:
 $(n/2)$ e $(n+1)/2$ esime osservazioni



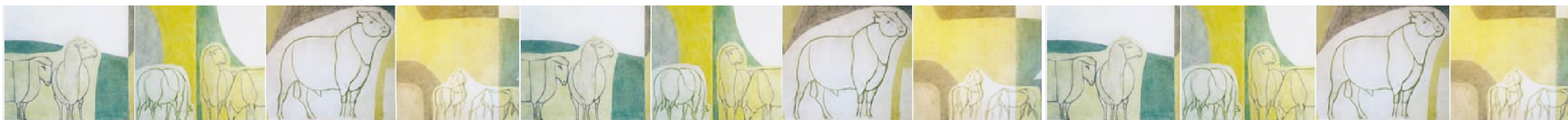
La mediana- Proprietà

- **Unicità:** posto un insieme di dati vi è una sola mediana
- **Semplicità:** è facile da capire e calcolare
- Non è influenzata da valori estremi



La moda

- in una serie finita di osservazioni è il valore che si verifica più frequentemente
- se tutti i valori sono diversi tra loro, non esiste una Moda
- un insieme di valori può avere più di una Moda (distribuzione bi-, tri- modale)



La Moda- Esempio

A

Incremento ponderale	Frequenza
Basso	8
Medio	10
Alto	2

B

Incremento ponderale	Frequenza
Basso	8
Medio	8
Alto	4



La moda- Proprietà

- **NON Unicità.** posto un insieme di dati vi possono essere più mode
- **Semplicità.** è facile da capire e calcolare
- non è sempre in grado di discriminare sufficientemente la distribuzione della variabile
- Utilizzata raramente come misura descrittiva
- Non viene MAI usata nella statistica analitica



Misure di posizione

	DEFINIZIONE	VANTAGGI	SVANTAGGI
MEDIA	$\frac{\text{somma dei dati}}{\text{numero dei dati}}$	adatta a manipolazioni matematiche	molto influenzata dai valori estremi
MEDIANA	livello di misura al di sotto del quale cade la metà dei dati	non influenzata dai valori estremi	non adatta a manipolazioni matematiche
MODA	valore che ricorre con maggiore frequenza	di significato facilmente intuibile	possibili distribuzioni bi-, tri-modali ecc.

www.quadernodiepidemiologia.it



Misure di posizione (di tendenza centrale)

il tipo di variabile statistica con cui si sta lavorando
pregiudica la scelta degli indici di posizione.

Indice di posizione	Variabile qualitativa nominale	Variabile qualitativa ordinale	Variabile quantitativa discreta	Variabile quantitativa continua
Media			✓	✓
Mediana		✓	✓	✓
Moda	✓	✓	✓	✓



Misure di posizione (di tendenza centrale)

Per tutte le variabili è possibile calcolare le statistiche descrittive di tendenza centrale

..che sono alla base de TUTTE le analisi descrittive di base effettuabili sul dataset

..il loro controllo, verifica e rettifica è quindi fondamentale.



Distribuzioni di frequenza

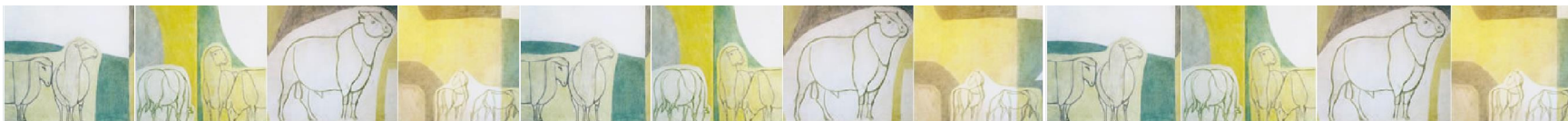
- Data una lista di tutti i valori di una variabile osservata sugli n individui della popolazione indagata, è possibile sintetizzare i dati in una distribuzione di frequenza.
- Quest'operazione sposta l'attenzione dalle singole unità statistiche (capi) ai valori rilevati e al numero di soggetti che li hanno manifestati



Distribuzioni di frequenza

una distribuzione di frequenza è semplicemente la successione delle modalità di una variabile in una tabella e il numero di volte che ogni modalità è stata osservata

Il numero di volte che ogni valore è stato osservato viene chiamata
Frequenza assoluta



Distribuzioni di Frequenza

Unità statistiche	X
	Sesso
1	Maschio
2	Maschio
3	Femmina
4	Maschio
5	Femmina
6	Femmina
7	Maschio
8	Femmina
9	Femmina
10	Maschio
11	Maschio
12	Maschio
13	Femmina
14	Maschio
15	Femmina
16	Femmina
17	Femmina
18	Maschio
19	Femmina
20	Femmina

LISTA DI DATI

Variabile SESSO:
Qualitativa nominale-dicotomica

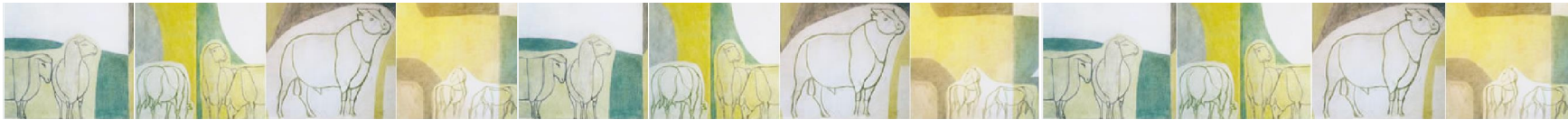
Sesso	Frequenze Assolute
Femmina	11
Maschio	9
Somma Σ	20



Distribuzioni di frequenza

- 1) Dividendo la frequenza assoluta di ogni valore di variabile per il numero totale di osservazioni si ottiene la **frequenza relativa**
- 2) Moltiplicando per 100 la frequenza relativa si ottiene la **frequenza relativa percentuale**

Sesso	Frequenze assolute	Frequenze relative	Frequenze relative percentuali
Femmina	11	0,55	55
Maschio	9	0,45	45
Somma Σ	20	1	100



Statistica descrittiva 1

Distribuzione di frequenza es.

Le altre variabili

Incremento ponderale	Frequenze assolute	Frequenze relative	Frequenze relative percentuali
Basso	8	0,4	40
Medio	5	0,25	25
Alto	7	0,35	35
Somma Σ	20	1	100

Qualitativa Ordinale

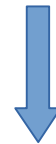
età / mesi	Frequenze assolute	Frequenze relative	Frequenze relative percentuali
24	1	0,05	5
25	1	0,05	5
28	1	0,05	5
29	3	0,15	15
30	3	0,15	15
32	4	0,2	20
33	1	0,05	5
34	3	0,15	15
38	2	0,1	10
41	1	0,05	5
Somma Σ	20	1	100

Quantitativa discreta

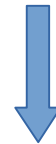


Distribuzioni di frequenza

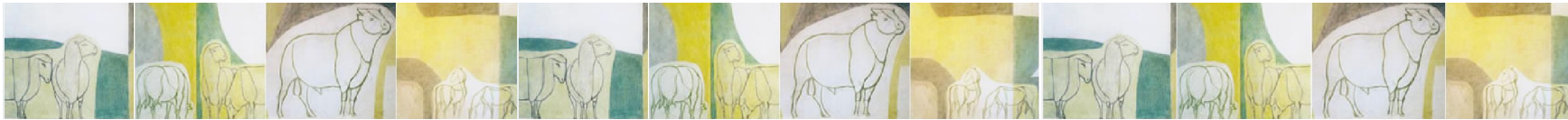
Per le variabili quantitative continue le frequenze assolute assumono valore unitario per tutte le modalità: nessuna misurazione si manifesta per più di un'unità statistica



Divisione in classi



distribuzione di frequenza su **N classi** anziché su n valori



Distribuzioni di frequenza

Frequenza cumulativa

Es. distribuzione aziende ovine positive per paratubercolosi Lazio e Toscana

classe prev %	Frequenze assolute	Frequenze relative	Frequenze relative percentuali	Frequenza cumulativa
a: 0	254	0,46	45,8	45,8
b: 0,1-5	85	0,15	15,3	61,2
c: 5,1-10	103	0,19	18,6	79,8
d: 10,1-20	76	0,14	13,7	93,5
e: 20,1-30	20	0,04	3,6	97,1
f: >30	16	0,03	2,9	100,0
Totale complessivo	554	1,00	100,0	

$$=45,8+15,3$$
$$=61,2+18,6$$

È utile per avere una idea immediata della soglia di prevalenza che comprende la maggior parte delle aziende

In questo caso si può dire che circa 80% delle aziende ha una prevalenza di paratubercolosi $\leq 10\%$

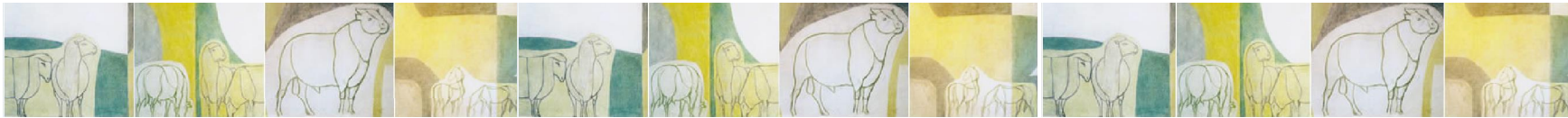


Distribuzioni di frequenza

La frequenza cumulativa

Osservando la tabella della distribuzione di frequenza è possibile avere un colpo d'occhio immediato sulla situazione generale della popolazione (azienda) rispetto ad una determinata variabile e...

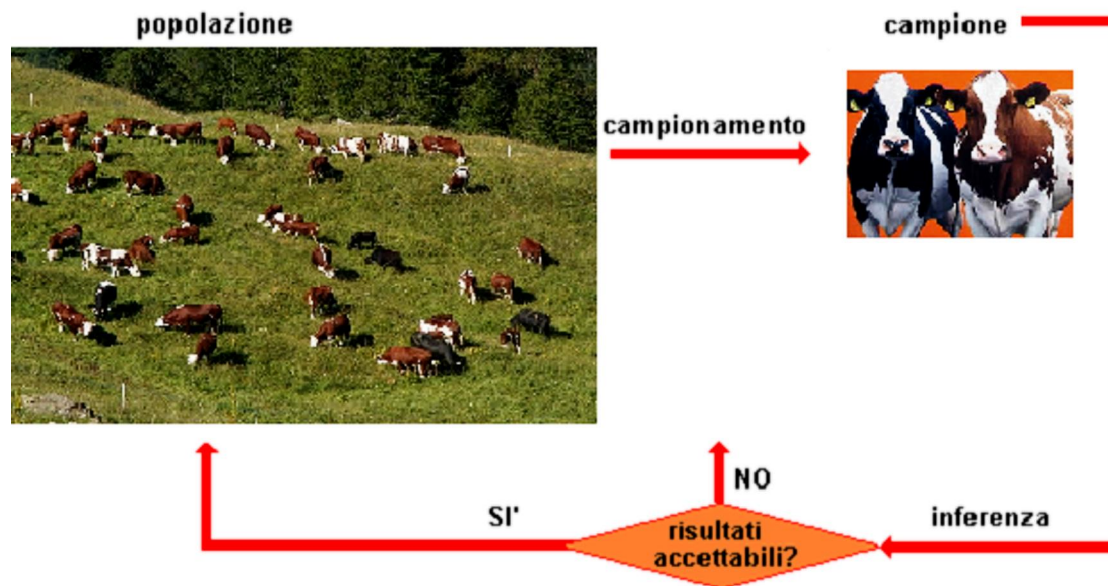
Valutare macroscopicamente aspetti positivi e negativi legati alla variabile

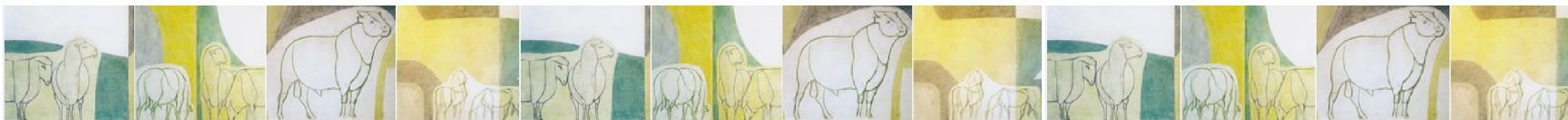


A cosa serve...

Descrittivo: descrive i dati osservati

Inferenziale: la generalizzazione dei risultati alla popolazione





Grazie per l'attenzione