Statistica sociale - 10

Prof. Antonio Mussino

a. a. 2022-2023

SAPIENZA
UNIVERSITÀ DI ROMA

L’elaborazione dei dati

La codifica e l'input - 1

- Riprendiamo in considerazione la ***matrice dei dati*** introdotta precedentemente.
- Come caso di studio che ci accompagnerà in questa parte del corso utilizzeremo la matrice ottenuta dalla registrazione delle risposte di un campione di cittadini brasiliani ad un questionario sulla partecipazione sportiva*.

* Indagine pilota per il progetto Diagnostico Nacional do Esporto del Ministero dello Sport brasiliano, svolta nella città di Aracaju, Stato Federale di Sergipe nel 2012.

La codifica e l'input - 2

- Dal questionario, come caso di studio, estrarremo solo alcune domande, che evidenziano diversi approcci per la codifica e per l'input.
- Le 11 domande enucleate dal questionario originario, che ne propone 27 sulla partecipazione sportiva e 9 sulle caratteristiche strutturali dell'intervistato, rappresentano esempi di:
  - variabili quantitative e qualitative,
  - risposte precodificate e da codificare,
  - indicatori semplici da combinare in un indice di sintesi (cfr. COMPASS),
  - e così via.

La codifica e l'input - 3

- Nell'operazione di memorizzazione dei dati (input) dai questionari al foglio elettronico (Excel, SPSS, Dbase, Access e così via), ci possiamo trovare di fronte a domande le cui risposte sono:
  - codificate come variabili qualitative, di cui
    - alcune già completamente codificate (tipo a: 1,6,8,9, A1, A6),
    - altre parzialmente codificate (tipo b: 4) e
    - altre non codificate (tipo c: 3);
  - codificate come variabili quantitative (tipo d: 2, A2, A7, A8).

La codifica e l'input - 4

- **1**. Nel 2011, nel suo tempo libero (fuori dall'orario di lavoro e da quello scolastico), Lei ha praticato qualche sport?

  1. Si 0. No - ***Andare alla domanda 11***

- **2**. Quanti sport Lei ha praticato nel 2011? _____
- **3**. Indichi quali sono gli sport che Lei ha praticato nel 2011, in ordine di importanza in relazione al tempo e allo sforzo a loro dedicati? Un massimo di tre.
  - 1° sport_____
  - 2° sport_____
  - 3° sport_____

La codifica e l'input - 5

- **4.** Ci dica se qualcuno di questi sport Lei lo ha praticato come membro (tesserato, affiliato) a uno degli enti/associazioni qui citate (per la scuola non deve segnalare l'attività curricolare):
  - 0- No
  - 1- Si, a un club/società
  - 2- Si, a una federazione/ente
  - 3- Si, a una associazione scolastica/ universitaria
  - 4- Si, a un altro ente; indichi quale:_____

La codifica e l'input - 6

- **6**. Qual è il livello di competizione più alto al quale Lei ha partecipato, nel 2011?
  - 1. Nazionale/internazionale
  - 2. Statale
  - 3. Municipale
  - 4. Locale non ufficiale (torneo tra amici, nel quartiere, a scuola, nel club, etc.)
  - 5. Non ha partecipato ad alcuna competizione

La codifica e l'input - 7

- **8.** Qual è il motivo principale per il quale Lei pratica lo sport? Indicare solo un motivo.
    1. Per migliorare il fisico
    2. Per migliorare l'armonia corpo/mente
    3. Per rilassarmi nel tempo libero
    4. Per competere con gli altri e/o con me stesso
    5. Per stare insieme ai miei amici e/o farmene di nuovi
- **9.** Considerando tutti gli sport praticati nel 2011, con quale frequenza Lei li ha praticati?
    1. meno di una volta al mese (1-11 volte all'anno)
    2. 1-3 volte al mese
    3. 1 volta alla settimana
    4. 2 volte alla settimana
    5. 3 volte o più alla settimana

La codifica e l'input - 8

- **A1.** Sesso
  - 1. maschio 2. femmina
- **A2.** Età in anni compiuti: _____
- **A6.** Colore della pelle :
  - 1. Bianca 2. Gialla 3. Marrone 4. Nera
- **A7.** Peso in kg _____
- **A8.** Altezza in cm _____

La codifica e l'input – 3 (replica)

- Nell'operazione di memorizzazione dei dati (input) dai questionari al foglio elettronico (Excel, SPSS, Dbase, Access e così via), ci possiamo trovare di fronte a domande le cui risposte sono:
  - codificate come variabili qualitative, di cui
    - alcune già completamente codificate (tipo a: 1,6,8,9, A1, A6),
    - altre parzialmente codificate (tipo b: 4) e
    - altre non codificate (tipo c: 3);
  - codificate come variabili quantitative (tipo d: 2, A2, A7, A8).

La codifica e l'input - 9

- La dimensione di riga "n" è pari a 1137 cittadini fra i 15 e i 65 anni;

- si tratta di un campione individuato attraverso una procedura a due stadi, areale nel primo e ***random walk sample***;

- le quote sono state individuate per età e sesso, basandosi sui risultati del Censimento della popolazione del 2011.

La matrice dei dati

		Variabili					
		X_1	X_2	X_3	....	....	X_P
Casi	1	x_{11}	x_{12}	x_{13}	....	....	x_{1p}
	2	x_{21}	x_{22}	x_{23}	....	....	x_{2p}
	3	x_{31}	x_{32}	x_{33}	....	....	x_{3p}
	.	....	....	....	....	....	....
	.	....	....	....	....	....	....
	.	....	....	....	....	....	....
	n	x_{n1}	x_{n2}	x_{n3}	....	....	x_{np}

La codifica e l'input - 10

- Quando si incontrano le domande di tipo ***a*** è facile riportare sulla matrice il codice numerico che corrisponde alla modalità scelta dall'intervistato;
- per le domande di tipo ***b*** e ***c*** è necessaria una operazione di codifica a posteriori, ossia vengono letti i questionari (o un campione di essi, se sono molti) e si propone una codifica per le voci rilevate, cercando di accorpare tali voci.
- Il caso della domanda di tipo ***b*** è molto semplice e la risposta codificata con 4 potrebbe rimanere tale, con l'individuazione delle altre tipologie di organizzazione eventualmente dimenticate nella precodifica, comunque la frequenza di questa modalità si prevede residuale.

La codifica e l'input - 11

- Ben diverso è il caso della domanda di tipo ***c***:
- in questo caso potrebbe essere di aiuto una lista di sport, proposta da esperti, o basata sulle forme di organizzazione delle varie discipline (ad esempio: sport di squadra e sport individuali; sport acquatici; sport con la palla; attività svolte in palestra o all'aria aperta e così via).
- Si codificheranno le risposte in base a questa prima lista, salvo poi accorpare quelle modalità che presenteranno frequenze ridotte.

La codifica e l'input - 12

- Questa fase non è normalizzabile;
- si deve, infatti, tener conto delle varie specificità territoriali e temporali della ricerca, essendo diverse le tipologie di attività sportiva che vengono praticate nei differenti paesi, e nei territori all'interno di questi paesi.
- In caso si avesse, invece, la necessità di operare confronti internazionali, si dovrebbero utilizzare le classificazioni previste dalle fonti internazionali.
- In questo caso il CIO (Comitato Internazionale Olimpico) che ha sue liste di discipline codificate secondo la giurisdizione delle verie Federazioni Sportive internazionali.

La codifica e l'input - 13

- Nel caso di studio l'obiettivo era quello di avere un quadro di riferimento su quali fossero le discipline, e le tipologie di discipline, più praticate nel territorio di Aracaju, per cui si è definita la seguente postcodifica:

Codice	Tipologia di attività	Esempi di discipline comprese
1	Calcio	Calcio, Calcio a otto, Calcetto, Beach soccer
2	Ginnastica	Ginnastica, Posturale, Yoga, Pesistica
3	Nuoto	Nuoto, Immersione, Nuoto pinnato
4	Sport di combattimento	Arti marziali, Pugilato, Lotta
5	Pallavolo	Pallavolo, Beach volley
6	Corsa	Corsa in strada, jogging
7	Danza	Danza, balli vari
8	Walking	Camminare, Trekking
9	Altri sport di squadra	Basket, Rugby, Palla a mano
10	Altri sport individuali	Tennis, Equitazione, Pattinaggio, Vela, Scacchi

La codifica e l'input - 14

- Per questa domanda vi è un'ulteriore complessità da superare: le discipline indicate potevano essere più di una, fino a un massimo di tre.
- Le possibili strategie per risolvere questo problema sono due:
  - considerare una variabile* per la prima risposta, ovvero per il primo sport, una per il secondo e una per il terzo; ovviamente chi pratica un solo sport risulterà non praticante nella seconda e terza colonna, chi ne pratica due non riempirà la terza;

*ricordiamo sempre che ad ogni variabile corrisponde una colonna della matrice dei dati!

La codifica e l'input - 15

- scomporre la risposta in dieci variabili ***binarie*** (*dummy*), corrispondenti a ciascuna delle dieci modalità di risposta previste, che possono assumere valore "1" se quella tipologia di attività è praticata e "0" se non lo è; complessivamente gli "1" nella tabella saranno tanti quanti sono gli sport praticati dall'intervistato; se non pratica ci saranno tutti "0".

La codifica e l’input - 16

- Queste ultime considerazioni ci aiutano a introdurre un’ulteriore importante elemento della codifica: come dobbiamo trattare il caso di un intervistato che non vuole rispondere a una domanda?

La codifica e l'input - 17

- Si deve prevedere un codice specifico per questa situazione (***missing value***):
  - in genere si usa il codice "0", ma bisogna fare attenzione al caso in cui l'intervistato ***non debba*** rispondere, come per coloro che dichiarano di "non praticare sport" e quindi non devono rispondere alle domande sulle modalità della pratica.
  - In questo caso si suggerisce di usare un altro codice (ad esempio il numero corrispondente alla modalità più alta più uno, oppure "9", "99" e così via), per poter distinguere le due situazioni.

La codifica e l'input - 18

- Nel caso precedente della risposta multipla potremmo avere svariate situazioni:
  - se l'intervistato non pratica sport, egli non deve indicare quali sport e la stringa sulla matrice sarà costituita da dieci "9";
  - se l'intervistato dichiara di praticare sport, ma non dice quanti e/o quali, la stringa sulla matrice sarà costituita da dieci "0";
  - se l'intervistato dichiara di praticare uno sport e lo indica, la stringa sulla matrice sarà costituita da un "1" e nove "0";

La codifica e l'input - 19

- se l'intervistato dichiara di praticare due sport e li indica, la stringa sulla matrice sarà costituita da due "1" e otto "0";
- se l'intervistato dichiara di praticare tre sport e li indica, la stringa sulla matrice sarà costituita da tre "1" e sette "0".

La codifica e l'input - 20

- Le operazioni di codifica (e postcodifica) ci consentono di effettuare l'input delle modalità di risposta, tramite i codici, nel caso di variabili **qualitative**.

- Le risposte alle domande di tipo d, invece, possono essere registrate direttamente essendo **quantitative**: è necessario definire l'unità di misura per stabilire se c'è la necessità di utilizzare cifre decimali o meno per la registrazione.

La codifica e l'input - 21

- Nel caso di studio le variabili considerate sono tutte registrabili con ***numeri interi***, in quanto per l'età si è chiesto di esprimerla in anni compiuti, per il peso in chilogrammi e l'altezza in centimetri.
- Se l'altezza fosse stata registrata in metri, avremmo ovviamente, avuto bisogno di due cifre ***decimali***.
- Nel caso in cui un intervistato avesse espresso le variabili peso e altezza con una o più cifre decimali, si registrerebbe il valore arrotondato (con cifre dopo la virgola da 50 a 99 arrotondamento all'unità superiore, altrimenti taglio delle cifre decimali).

Distribuzioni di frequenza

Distribuzioni di frequenza - 1

- Il primo passo per un'analisi statistica dei dati è quello della **descrizione** e della **sintesi** delle informazioni contenute nelle colonne della matrice dei dati, ovvero delle risposte alle domande del questionario.
- Questo passo è rappresentabile dall'operazione di ***conteggio*** di quante unità statistiche hanno scelto una delle ***modalità*** di risposta (altrimenti definite ***categorie***).

Distribuzioni di frequenza - 2

- La sintesi è effettiva, in quanto le 1137 risposte sono rappresentabili su un numero ridotto, variabile da 2 a 10 categorie, associando ad ogni categoria la sua ***frequenza assoluta***.
- Ad esempio, proponiamo una tavola con la distribuzione di frequenza della variable corrispondente alla "frequenza totale della pratica sportiva".
- Oltre a quella assoluta sono proposte altre frequenze utili per il nostro obiettivo.

ALLEGATO 1

Allegato 1

		Frequenza	Percentuale	Percentuale valida	Percentuale cumulata
Validi	< 1 volta al mese	14	1,2	4,0	4,0
	1-3 volte al mese	22	1,9	6,3	10,3
	1 volta alla settimana	58	5,1	16,6	26,9
	2 volte alla settimana	85	7,5	24,3	51,1
	3 e più volte alla settimana	171	15,0	48,9	100,0
	Totale	350	30,8	100,0	
Mancanti	non praticante	787	69,2		
Totale		1137	100,0		

Distribuzioni di frequenza - 3

- La (frequenza) ***Percentuale***, ovvero la ***frequenza relativa*** (pari alla frequenza assoluta divisa per il totale delle unità) moltiplicata per 100, è utile per normalizzare il risultato in caso di confronto fra collettivi di numerosità diversa.
- La ***Percentuale valida*** è una percentuale calcolata solo su chi ha risposto alla domanda: in questo caso il 69,2% delle unità non dovevano rispondere perché si erano dichiarati "non praticanti", mentre non ci sono state "risposte mancanti" tra i praticanti;

Distribuzioni di frequenza - 4

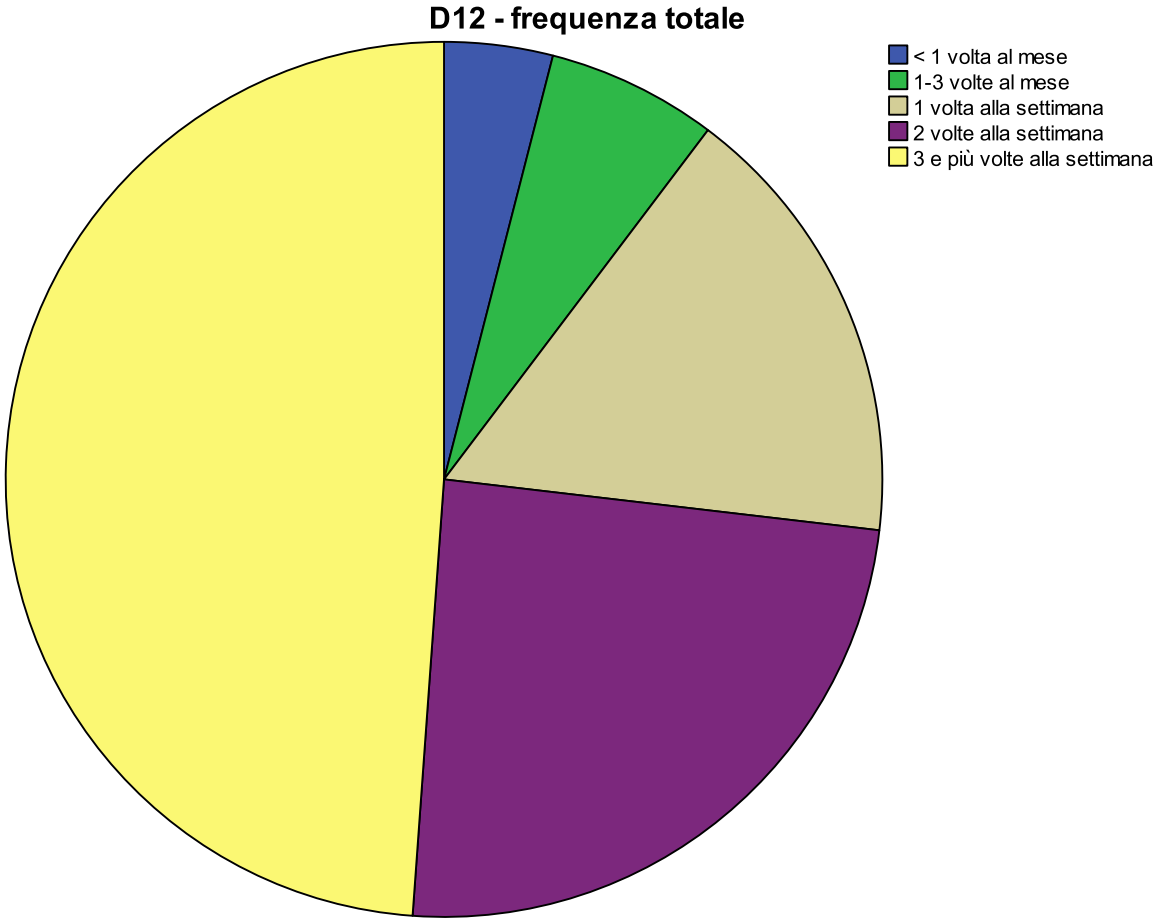
- La ***Percentuale cumulata*** è il risultato della somma progressiva delle Percentuali valide e ci segnala qual è l'ammontare del fenomeno fino al livello di pratica definito dalla categoria di riferimento:
  - ad esempio, il 26,9% pratica fino a "1 volta alla settimana", quindi anche meno frequentemente).
- È ovvio che, perché l'informazione abbia senso, è necessario che le categorie siano gerarchicamente ordinabili dal livello più basso al più alto:
  - non ha senso, ad esempio, calcolarla per le variabili "colore della pelle", "sesso", "tesseramento", "motivo della pratica" e così via.

Distribuzioni di frequenza - 5

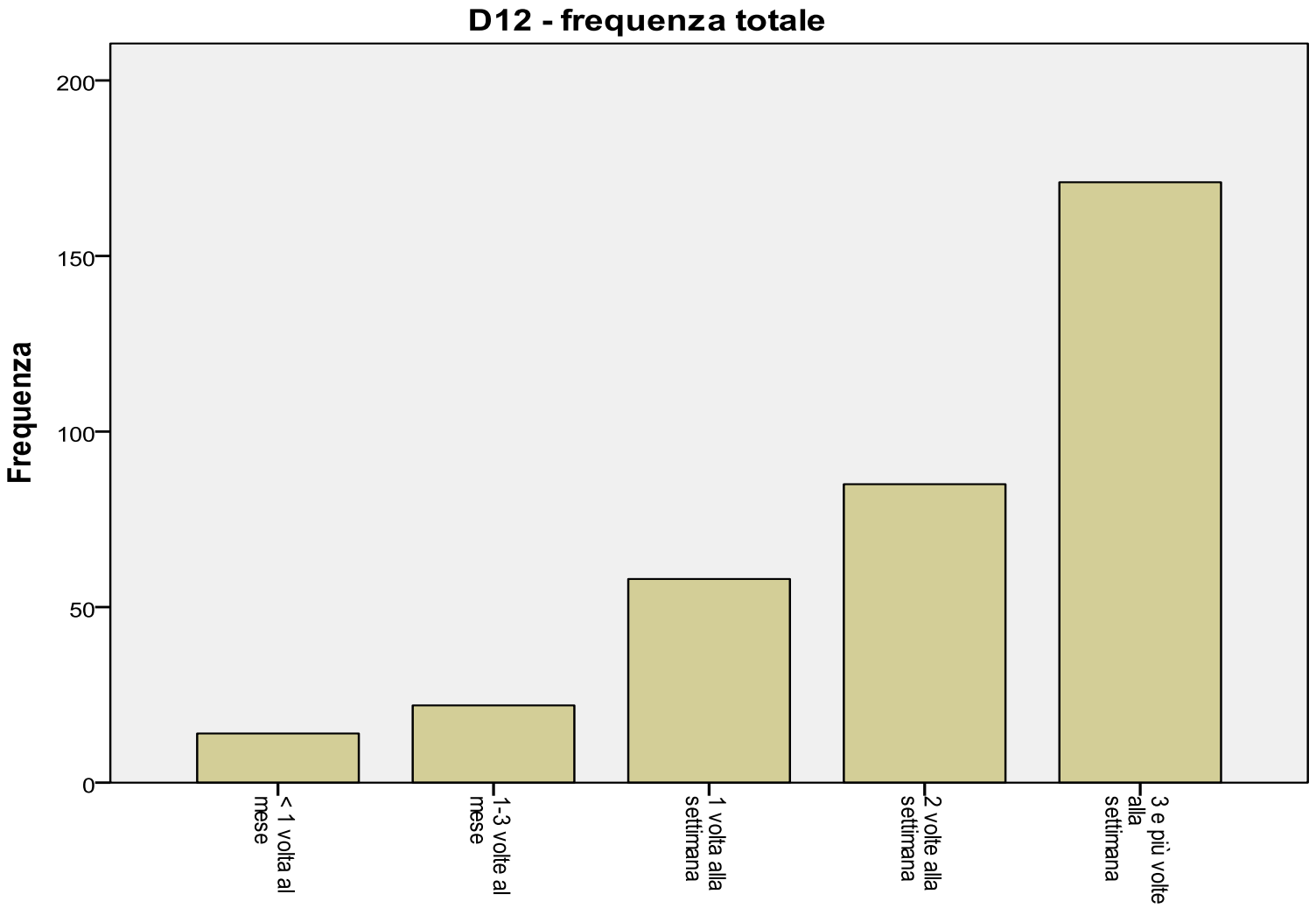
- Una presentazione più *accattivante*, ma sicuramente meno completa potrebbe essere quella grafica: di seguito una rappresentazione con ***grafico a torta*** e una con ***diagramma a barre***.

ALLEGATI 2 E 3

Allegato 2


D12 - frequenza totale
< 1 volta al mese
1-3 volte al mese
1 volta alla settimana
2 volte alla settimana
3 e più volte alla settimana

Allegato 3


D12 - frequenza totale
Frequenza
200
150
100
50
0
< 1 volta al mese
1-3 volte al mese
1 volta alla settimana
2 volte alla settimana
3 e più volte alla settimana

Distribuzioni di frequenza - 6

- Se le variabili sono
  - **qualitative**, **ordinabili o meno** (escludendo in questo secondo caso le Percentuali cumulate), o
  - sono **quantitative** ma **discrete** e con un numero di modalità ridotte (ad esempio il numero di sport praticati)

  le rappresentazioni proposte sono efficienti ed efficaci.

Distribuzioni di frequenza - 7

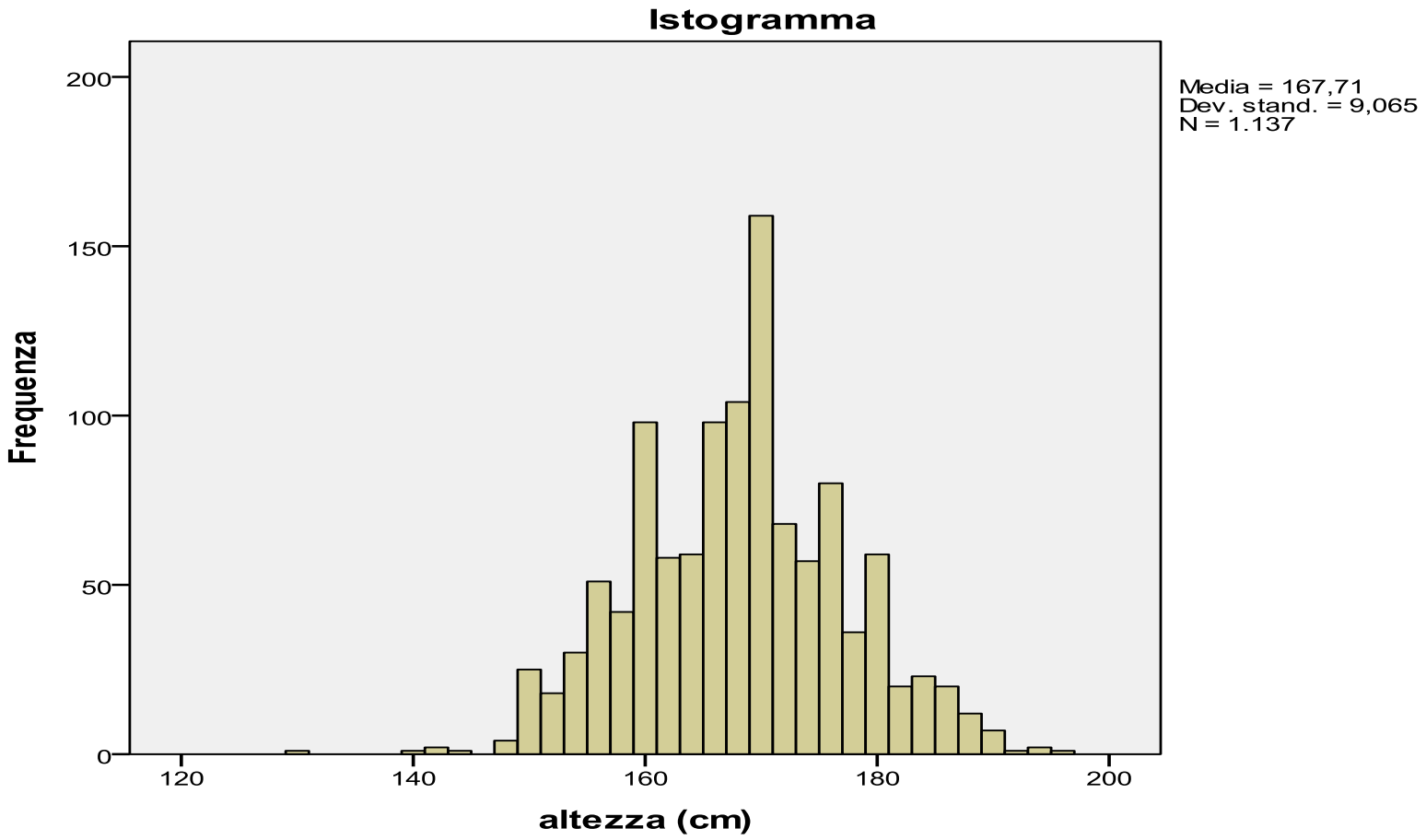
- Se, invece, volessimo descrivere e sintetizzare variabili **quantitative continue** (anche se da noi rese discrete nella codifica, con opportune aggregazioni), come ad esempio l'età o la statura, allora questa strategia non è percorribile:
  - le categorie sarebbero ben 50 per l'età e quasi altrettante per la statura.

Distribuzioni di frequenza - 8

- La rappresentazione di una variabile continua dovrebbe essere nel ***continuum***, come nel grafico che segue (***istogramma***), e potrebbe avere un significato accorpando i valori in un numero di classi ridotte;
- nel grafico questo accorpamento è fatto automaticamente e ogni classe ha la stessa ampiezza.

ALLEGATO 4

Allegato 4


Istogramma
Media = 167,71
Dev. stand. = 9,065
N = 1.137
Frequenza
200
150
100
50
0
120
140
160
180
200
altezza (cm)

Distribuzioni di frequenza - 9

- In realtà è più efficace accorpare logicamente le determinazioni della variabile;
  - nel caso dell'età, potremmo, infatti, considerare: gli "adolescenti"(da 15 a 19 anni); i "giovanissimi" (da 20 a 24 anni); i "giovani" (da 25 a 34 anni); gli "adulti" (da 35 a 54 anni); i "maturi" (da 55 a 65 anni).
- L'ampiezza delle classi non sarebbe la stessa, e quindi non si potrebbe utilizzare una rappresentazione grafica automatica.
- Un'efficiente ed efficace rappresentazione tabellare della distribuzione di frequenza delle età potrebbe essere la seguente. ALLEGATO 5

Allegato 5

Tabella 3.5 - Distribuzione di frequenza della variabile "Età", raggruppata in classi.

		Frequenza	Percentuale	Percentuale valida	Percentuale cumulata
Validi	15-19 anni	135	11,9	11,9	11,9
	20-24 anni	171	15,0	15,0	26,9
	25-34 anni	304	26,7	26,7	53,6
	35-54 anni	404	35,5	35,5	89,2
	55-64 anni	123	10,8	10,8	100,0
	Totale	1137	100,0	100,0	

Distribuzioni di frequenza - 10

- Si può notare come la colonna delle Percentuali valide sia uguale a quella delle Percentuali, in quanto tutti dovevano rispondere, e tutti hanno risposto, a questa domanda.
- Graficamente, si dovrebbe costruire un istogramma *ad hoc*:
  - la logica dell'istogramma (e di tutti i diagrammi) è che l'area deve essere proporzionale alla frequenza (assoluta o percentuale);
  - quindi se le basi sono uguali possiamo attribuire loro una dimensione ***unitaria*** e quindi l'area è uguale all'altezza, mentre se le basi sono diverse le altezze si devono calcolare caso per caso.

Distribuzioni di frequenza - 11

- Una volta elaborati i dati e calcolate le distribuzioni di frequenza, è possibile ottenere una sintesi ancora più efficace delle variabili studiate, calcolandone le ***misure di tendenza centrale*** e di ***variabilità***.
- Le prime permettono di sintetizzare con un unico valore la distribuzione, le seconde tengono conto della ***dispersione*** intorno a questo valore, che infatti potrebbe essere diversa di caso in caso.

Distribuzioni di frequenza - 12

- Nella tabella che segue è riportato il caso di una variabile quantitativa, quale è la statura.
- Poiché non ci sono valori mancanti la ***media*** *(**aritmetica**)* è calcolata su tutti gli intervistati;
- le altre due misure di tendenza centrale sono
  - la ***mediana***, media di posizione che corrisponde al 50 Percentile, e
  - la ***moda***, che corrisponde al valore con la frequenza più alta e che ha scarsa validità in distribuzioni continue come questa.

Distribuzioni di frequenza - 13

- La dispersione può essere misurata in modo analitico con
  - la ***deviazione standard*** (lo ***scarto quadratico medio***) rispetto alla media aritmetica, oppure
  - con la ***differenza interquartilica***, ovvero la differenza fra il 3^ e il 1^ quartile (ovvero il 75^ e 25^ percentile) rispetto alla mediana.
  - Infine, può essere interessante valutare il ***range*** della distribuzione, ovvero la differenza fra il valore più alto e quello più basso.

ALLEGATO 6

Allegato 6

Tabella 3.6 - Statistiche di sintesi per la variabile "Altezza".

N	Validi	1137
	Mancanti	0
Media		167,71
Moda		170
Deviazione standard		9,065
Minimo		130
Massimo		195
Percentili	25	161,00
	50	168,00
	75	173,00

Distribuzioni di frequenza - 14

- Queste misure sono differenti a seconda della natura delle variabili studiate.

- Nel caso di una variabile **qualitativa ordinabile** si possono utilizzare la mediana e valutare la differenza interquartilica,

- mentre poche opportunità ci sono per le variabili **qualitative non ordinabili**.

Distribuzioni di frequenza - 15

- In realtà spesso si trovano sintetizzate con le misure analitiche anche variabili qualitative ordinabili, come le scale di ***Likert*** e di ***Cantril***.

- Tecnicamente è una soluzione non corretta, ma può essere utilizzata, anche se con cautela, per la sua efficacia informativa e comparativa.

Distribuzioni di frequenza - 16

- Se riprendiamo in considerazione, ad esempio, le scale proposte per valutare la piscina del CUS Roma nella sede di Tor di Quinto, si vede come l'utilizzo della media aritmetica sia piuttosto efficace per evidenziare gli item per i quali c'è soddisfazione e quelli più criticati dagli utenti:
  - la percezione di insoddisfazione per le docce fornita dal punteggio medio 4,24, come pure quella di soddisfazione per gli istruttori (6,97) è molto efficace e fa superare le critiche metodologiche.

ALLEGATO 7

Allegato 7

Tabella 3.7 - Statistiche descrittive di sintesi per gli item di valutazione degli impianti del CUS Roma (sede di Tor di Quinto).

item	N	Media	Deviazione standard	Mediana	Moda
Pulizia spogliatoi	421	5,05	2,155	5	6
Comfort spogliatoi	422	4,68	1,988	5	6
Armadietti	339	4,69	2,310	5	6
Docce	409	4,24	2,027	5	5
Attrezzi	352	5,70	2,173	6	6
Istruttori	247	6,97	2,727	8	10
Pulizia piscina	400	6,64	1,831	7	7
Spazio acqua	403	6,59	1,849	7	7
Corsi	267	6,60	2,226	7	7
Temperatura acqua	413	6,90	1,839	7	8
Casi validi (*listwise*)	178				

Distribuzioni di frequenza - 17

- A tale proposito è interessante osservare come la gerarchia proposta dalla media aritmetica sia più discriminante delle, pur concordanti, graduatorie proposte dalla mediana, e anche dalla troppo rozza moda.