# Statistica sociale - 10

*Prof. Antonio Mussino*

a. a. 2016-2017

SAPIENZA
UNIVERSITÀ DI ROMA

# L'elaborazione dei dati
(segue)

# 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

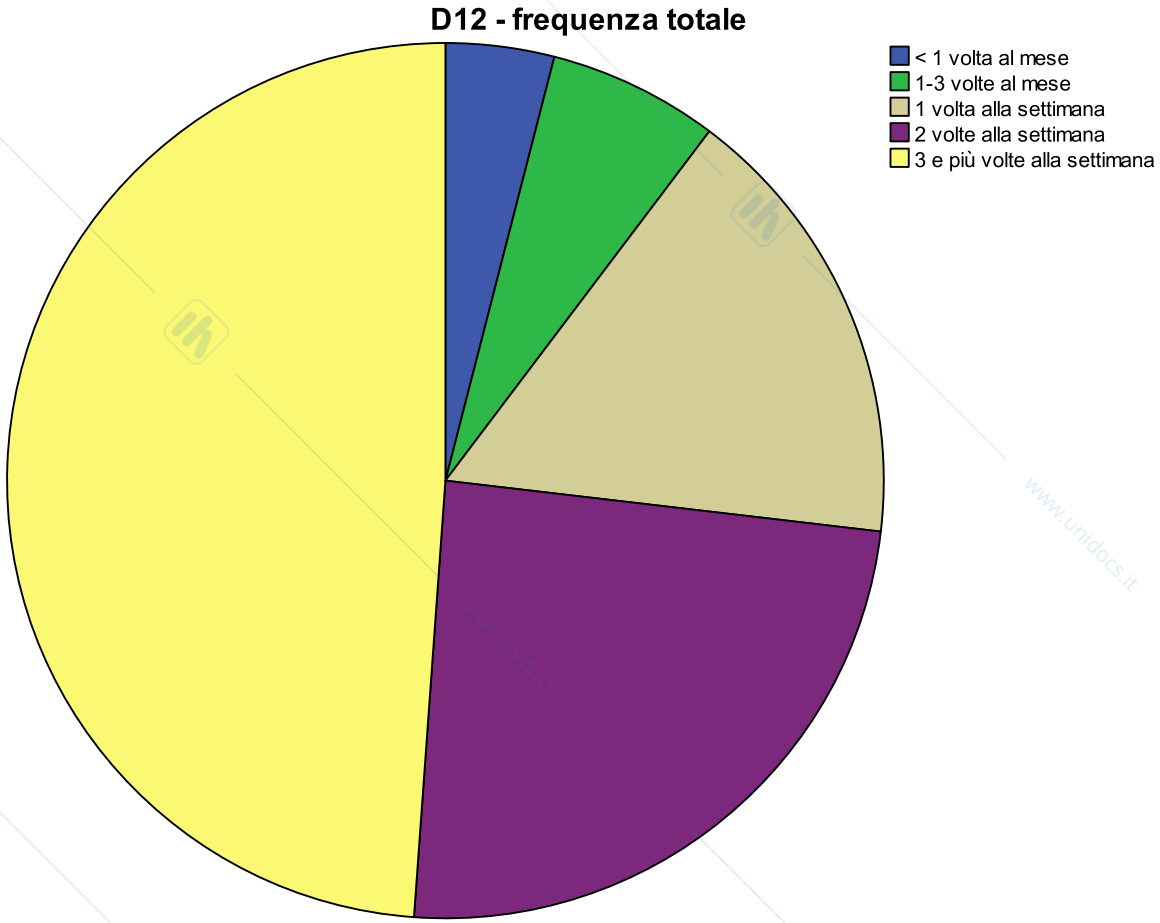
- La ***Percentuale cumulata*** è il risultato della somma progressiva delle Percentuali valide e ci segnala qual è l'ammontare del fenomeno fino al livello di pratica definito dalla categoria di riferimento:
  - ad esempio, il 26,9% pratica fino a "1 volta alla settimana", quindi anche meno frequentemente).
- 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

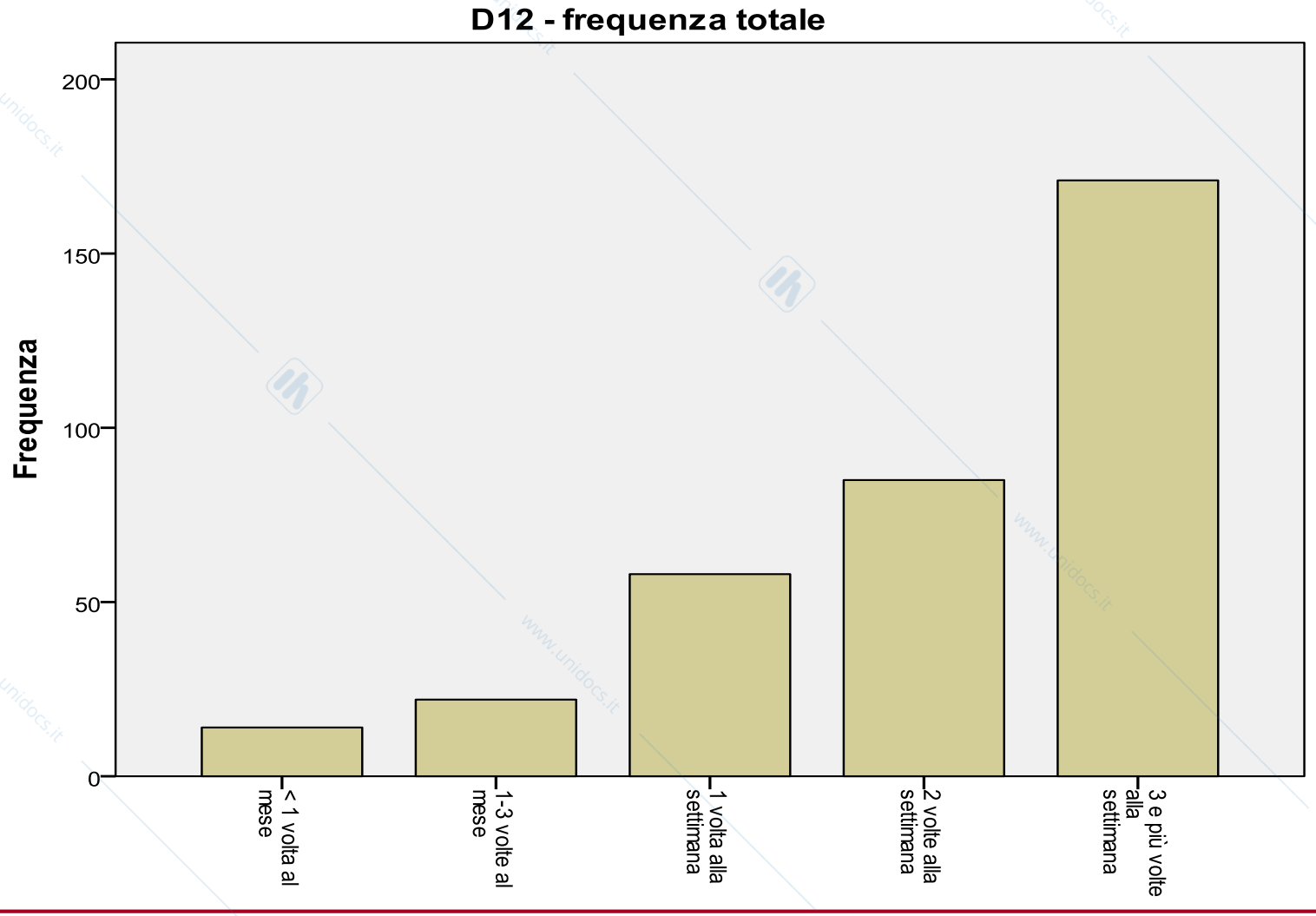
- Una presentazione più *accattivante*, ma sicuramente meno completa potrebbe essere quella grafica: di seguito una rappresentazione con ***grafico a torta*** e una con ***diagramma a barre***.

ALLEGATI 2 E 3

# Allegato 2



# Allegato 3



# 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

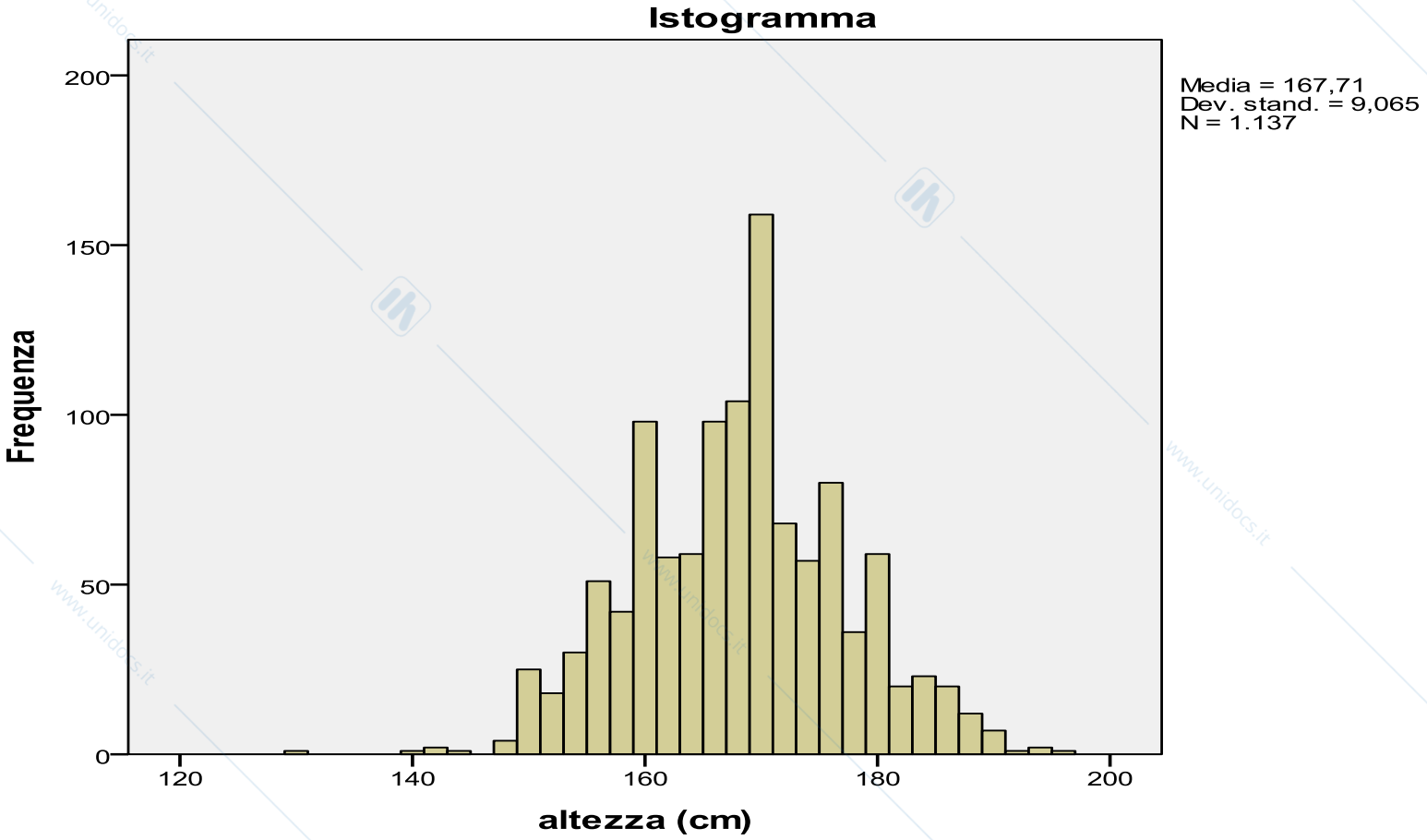
- Se, invece, volessimo descrivere e sintetizzare variabili **quantitative continue** (anche se da noi rese discrete nella codifica, con opportune aggregazioni), come ad esempio l'età o la statura, allora questa strategia non è percorribile:
  - le categorie sarebbero ben 50 per l'età e quasi altrettante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



# 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.

# Relazioni bivariate - 1

- Quando mettiamo in relazione due (o più) variabili siamo entrati nella fase ***esplicativa*** dell'analisi dei dati:
  - vogliamo vedere se esista o meno un legame fra di esse, ovvero se il variare di una comporti, e in che modo, quello dell'altra.

- La relazione che vogliamo studiare è prettamente ***statistica***, ovvero legata al concetto di "media" in senso lato.

# Relazioni bivariate - 2

- Se diciamo che c'è una relazione fra titolo di studio e frequenza della pratica sportiva, vuol dire che chi ha un titolo più alto in media pratica di più, ma ci possono essere laureati sedentari;
- se diciamo che c'è una relazione fra genere e disciplina praticata, vuol dire che tra gli uomini il calcio è lo sport più praticato, ma ci sono anche calciatrici, e tra le donne la ginnastica è la più praticata, ma ci sono molti uomini che vanno in palestra.
- In genere è più ***probabile*** che un laureato pratichi di più di un diplomato e tra gli uomini ci siano più calciatori.

# Relazioni bivariate - 3

- Questo per dire che la relazione non implica un nesso di ***causa-effetto***:
  - capire se e quale sia questo nesso esula dai compiti della Statistica e rientra in quelli del ricercatore che analizza i dati.

# Relazioni bivariate - 4

- Ad esempio, la relazione fra titolo di studio e pratica sportiva può essere spiegata considerando il fatto che:
  - chi ha un titolo più elevato ha un reddito più elevato e quindi più possibilità di spendere per praticare uno sport, quindi la relazione è indiretta e la causa della maggiore pratica è la maggiore capacità di spesa;
  - oppure si può considerare il fatto che l'attività sportiva rientra nella sfera culturale di un individuo e, in genere, più è alto il titolo di studio maggiore è il livello culturale.

# Relazioni bivariate - 5

- In questi semplici esempi abbiamo già visto come, anche se la relazione è solo statistica, il ricercatore tende ad assegnare alle due variabili un ruolo diverso:
  - una delle due è la possibile causa e l'altra l'effetto, ovvero la prima influenza il variare dell'altra.
- Allora la prima è definita ***indipendente*** e l'altra ***dipendente***.
- La scelta di quale ruolo giochino le variabili è fatta soggettivamente dal ricercatore e ci sono anche casi in cui questa scelta non è possibile perché le due variabili giocano un ruolo simmetrico nell'analisi.

# Relazioni bivariate - 6

- Semplificando al massimo la classificazione delle variabili, possiamo trovarci quindi di fronte a quattro situazioni:

  **a)** le due variabili sono entrambe (indipendente e dipendente) qualitative;

  **b)** le due variabili sono entrambe (indipendente e dipendente) quantitative:

  **c)** la variabile indipendente è qualitativa e quella dipendente quantitativa;

  **d)** la variabile indipendente è quantitativa e quella dipendente qualitativa.

# Relazioni bivariate - 7

- Entreremo ora nel dettaglio dell'analisi delle relazioni a seconda del ruolo e della natura delle variabili:
  - i più rilevanti sono i casi **a)** e **b)** e sono disponibili anche strategie importanti per trattare il caso **c)**;
  - non entreremo nel merito del caso **d)**, che in effetti si verifica molto raramente.
- In quest'ultimo caso, ma vedremo accade anche per il **c)**, si preferisce accorpare i valori della variabile quantitativa in classi trasformandola in qualitativa e tornando così al caso **a)**.

# Relazioni bivariate - 8

- La possibilità di ricondurre tutti gli altri al caso **a)**, ma soprattutto la netta prevalenza di variabili qualitative nell'area della Statistica sociale, ci spinge a iniziare e a trattare con maggiore accuratezza il caso della relazione fra variabili qualitative, che più precisamente definiremo ***associazione***.

# Il caso di variabili qualitative - 1

- Per affrontare questo argomento dobbiamo definire e costruire un nuovo modo di rappresentare i dati: la ***tabella*** (o *tavola*) ***di contingenza****.

- È, di fatto, una trasformazione della matrice originaria e anch'essa si può considerare una matrice nella quale le righe e le colonne sono le modalità delle variabili studiate (due per volta) e nelle celle c'è la frequenza delle volte in cui le modalità si presentano associate nel collettivo.

* Anche detta tabella a doppia entrata, incrocio, tabulazione incrociata.

# Il caso di variabili qualitative - 2

- Come esempio consideriamo
  - "età" (raggruppata in classi) e
  - "frequenza della pratica sportiva" (di fatto anch'essa raggruppata in classi),
- proprio per mostrare l'applicabilità di questa strategia anche con variabili quantitative.

ALLEGATO 8

# Allegato 8

**Tabella 3.8** - Tavola di contingenza fra “Frequenza totale nell'anno della pratica sportiva” e “Età”

| | | Età in classi | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 15-19 anni | 20-24 anni | 25-34 anni | 35-44 anni | 45-54 anni | 55-64 anni | Totale |
| Frequenza totale nell'anno della pratica sportiva | Mai | 65 | 104 | 205 | 167 | 139 | 107 | 787 |
| | Meno di 1 volta al mese | 2 | 1 | 7 | 2 | 1 | 1 | 14 |
| | 1-3 volte al mese | 3 | 5 | 10 | 4 | 0 | 0 | 22 |
| | 1 volta alla settimana | 3 | 9 | 19 | 15 | 7 | 5 | 58 |
| | 2 volte alla settimana | 23 | 15 | 24 | 10 | 10 | 3 | 85 |
| | 3 e più volte alla settimana | 39 | 37 | 39 | 31 | 18 | 7 | 171 |
| Totale | | 135 | 171 | 304 | 229 | 175 | 123 | 1137 |

# Il caso di variabili qualitative - 3

- Logicamente l' "età" è la variabile ***indipendente*** e la "frequenza della pratica" la ***dipendente***, per cui disponiamo convenzionalmente la prima sulle colonne e la seconda sulle righe.
- La frequenza assoluta 65 indica che nel collettivo ci sono 65 intervistati che hanno meno di 19 anni e che non hanno mai praticato nell'anno precedente e così via.
- Nell'ultima riga e nell'ultima colonna troviamo i "***totali marginali***" (che in realtà non fanno parte della matrice), che corrispondono alle distribuzioni di frequenza uni-variate rispettivamente delle variabili per colonna e per riga.

# Il caso di variabili qualitative - 4

- Il ricercatore utilizza questa presentazione perché vuole vedere se c'è ***dipendenza***, o meno, fra le due variabili e, in caso di dipendenza, che tipo e con quale intensità ci sia ***associazione*** fra di esse.

- Un altro caso, che presenta un numero minore di modalità e quindi è più immediato nel commento, è quello che mette in relazione la "frequenza della pratica" con il "genere" degli intervistati.

ALLEGATO 9

# Allegato 9

**Tabella 3.9** - Tavola di contingenza fra “Frequenza totale nell'anno della pratica” e “Genere”

| | | Genere | | Totale |
|---|---|---|---|---|
| | | Maschi | Femmine | |
| Frequenza totale nell'anno della pratica sportiva | Mai | 318 | 469 | 787 |
| | Meno di 1 volta al mese | 8 | 6 | 14 |
| | 1-3 volte al mese | 16 | 6 | 22 |
| | 1 volta alla settimana | 49 | 9 | 58 |
| | 2 volte alla settimana | 48 | 37 | 85 |
| | 3 e più volte alla settimana | 105 | 66 | 171 |
| Totale | | 544 | 593 | 1137 |

# Il caso di variabili qualitative - 5

- Ovviamente le frequenze proposte (quelle assolute) non sono utili per avere informazioni sulle eventuali associazioni, perché la numerosità dei gruppi di individui nelle diverse classi di età e nei diversi livelli di frequenza è differente:
- abbiamo bisogno di relativizzare l'informazione e, per far questo, calcoliamo le frequenze percentuali, che possono essere di tre tipi: ***percentuale*** *di* ***riga***, *di* ***colonna*** e *sul* ***totale***.

# Il caso di variabili qualitative - 6

- Nella prima tabella sono riportate le percentuali di riga, ovvero le percentuali in relazione alle diverse modalità della variabile “frequenza della pratica”:
  - è un’informazione poco utile, ci dice quale genere è prevalente all’interno dei diversi livelli di impegno sportivo e di sedentarietà.

ALLEGATO 10

# Allegato 10

**Tabella 3.10** - Tavola di contingenza fra "Frequenza totale nell'anno della pratica" e "Genere"

| *% di riga* | | Genere | | Totale |
|---|---|---|---|---|
| | | Maschi | Femmine | |
| Frequenza totale nell'anno della pratica sportiva | Mai | 40,4% | 59,6% | 100,0% |
| | Meno di 1 volta al mese | 57,1% | 42,9% | 100,0% |
| | 1-3 volte al mese | 72,7% | 27,3% | 100,0% |
| | 1 volta alla settimana | 84,5% | 15,5% | 100,0% |
| | 2 volte alla settimana | 56,5% | 43,5% | 100,0% |
| | 3 e più volte alla settimana | 61,4% | 38,6% | 100,0% |
| Totale | | 47,8% | 52,2% | 100,0% |

# Il caso di variabili qualitative - 7

- Per capire il motivo di tale affermazione consideriamo la "frequenza della pratica" solo per gli sportivi.
- In questo caso tutti gli intervistati hanno risposto alle due domande e non ci sono *mancate risposte*.
- Ma se riprendiamo in considerazione il questionario possiamo osservare come chi rispondeva di non praticare sport non doveva rispondere alle domande dalla numero 2 alla numero 11.

# Il caso di variabili qualitative - 8

- Si tratta del caso di ***risposte non dovute***, e abbiamo visto come si trattano nelle distribuzioni di frequenza univariate:
  - nel caso bivariato la gestione è più semplice, basta che manchi l'informazione per una delle due variabili che l'unità statistica non è contata nelle celle.
- Si può, infatti, notare come il totale generale sia pari a 350, ovvero gli intervistati che praticano sport.

ALLEGATO 11

# Allegato 11

**Tabella 3.11** - Tavola di contingenza fra "Frequenza totale nell'anno della pratica" e "Genere"

| | | Genere | | Totale |
|---|---|---|---|---|
| | | Maschi | Femmine | |
| Frequenza totale nell'anno della pratica sportiva | Meno di 1 volta al mese | 8 | 6 | 14 |
| | 1-3 volte al mese | 16 | 6 | 22 |
| | 1 volta alla settimana | 49 | 9 | 58 |
| | 2 volte alla settimana | 48 | 37 | 85 |
| | 3 e più volte alla settimana | 105 | 66 | 171 |
| Totale | | 226 | 124 | 350 |

# Il caso di variabili qualitative - 9

- Se calcoliamo le percentuali di riga rispetto a questa tabella, vediamo che le percentuali dei maschi sono ***sempre*** più alte:
  - pertanto, per capire se vi è una differente modalità di prevalenza è necessario comparare le percentuali di ciascuna riga (possiamo definirle i ***profili***) con quelle della riga del totale (64,6% tra i maschi e 35,4% tra le femmine).

ALLEGATO 12 e 13

# Allegati 12 e 13

**Tabella 3.12** - Tavola di contingenza fra "Frequenza totale nell'anno della pratica" e "Genere"

| % di riga | | Genere | | |
|---|---|---|---|---|
| | | Maschi | Femmine | Totale |
| Frequenza totale nell'anno della pratica sportiva | Meno di 1 volta al mese | 57,1% | 42,9% | 100,0% |
| | 1-3 volte al mese | 72,7% | 27,3% | 100,0% |
| | 1 volta alla settimana | 84,5% | 15,5% | 100,0% |
| | 2 volte alla settimana | 56,5% | 43,5% | 100,0% |
| | 3 e più volte alla settimana | 61,4% | 38,6% | 100,0% |
| Totale | | 64,6% | 35,4% | 100,0% |

**Tabella 3.13** - Tavola di contingenza fra "Frequenza totale nell'anno della pratica" e "Genere"

| Inclusi i non praticanti | % di colonna | Genere | | |
|---|---|---|---|---|
| | | Maschi | Femmine | Totale |
| Frequenza totale nell'anno della pratica sportiva | Mai | 58,5% | 79,1% | 69,2% |
| | Meno di 1 volta al mese | 1,5% | 1,0% | 1,2% |
| | 1-3 volte al mese | 2,9% | 1,0% | 1,9% |
| | 1 volta alla settimana | 9,0% | 1,5% | 5,1% |
| | 2 volte alla settimana | 8,8% | 6,2% | 7,5% |
| | 3 e più volte alla settimana | 19,3% | 11,1% | 15,0% |
| Totale | | 100,0% | 100,0% | 100,0% |

# Il caso di variabili qualitative - 10

- Più immediata è la lettura delle percentuali di colonna:
  - infatti i due sottoinsiemi che si mettono a confronto sono uniformati rispetto alla numerosità, in quanto si considera il risultato ogni 100 maschi e ogni 100 femmine.
- Vediamo così che le donne sono nettamente prevalenti se consideriamo la mancata pratica, ma quando si impegnano lo fanno con maggiore costanza e regolarità.
- Il confronto fra i ***profili colonna*** è estremamente efficace, rispetto al nostro obiettivo di scoprire le associazioni, in entrambe le tabelle.

# Il caso di variabili qualitative - 11

- La scelta di privilegiare i profili colonna sui profili riga è dovuta al fatto che la variabile indipendente è posizionata sulle colonne;
- se fosse posta sulle righe, ovviamente, bisognerebbe invertire la scelta.

(ALLEGATO 14)

# Allegato 14

**Tabella 3.14** - Tavola di contingenza fra "Frequenza totale nell'anno della pratica" e "Genere"

| Esclusi i non praticanti | % di colonna | Genere | | Totale |
|---|---|---|---|---|
| | | Maschi | Femmine | |
| Frequenza totale nell'anno della pratica sportiva | Meno di 1 volta al mese | 3,5% | 4,8% | 4,0% |
| | 1-3 volte al mese | 7,1% | 4,8% | 6,3% |
| | 1 volta alla settimana | 21,7% | 7,3% | 16,6% |
| | 2 volte alla settimana | 21,2% | 29,8% | 24,3% |
| | 3 e più volte alla settimana | 46,5% | 53,2% | 48,9% |
| Totale | | 100,0% | 100,0% | 100,0% |

# Il caso di variabili qualitative - 12

- L'ultima opportunità di calcolo di percentuali è relativa a quelle ***totali***;
  - come si può vedere dalla tabella seguente non vi è un'informazione aggiuntiva alla tabella originaria delle frequenze assolute:
  - queste percentuali (il totale 100% è relativo a tutta la tabella) non ci servono per studiare le eventuali associazioni.

# Il caso di variabili qualitative - 13

- Pertanto questa modalità non si utilizza mai, a meno che non si voglia confrontare la situazione in questo collettivo (con numerosità 1137), con quella ottenuta in un altro collettivo di numerosità diversa:
  - è come se calcolassimo la distribuzione di frequenze percentuali di una variabile ricostruita associando in tutti i modi possibili le modalità delle due variabili studiate.

ALLEGATO 15

# Allegato 15

**Tabella 3.15** - Tavola di contingenza fra "Frequenza totale nell'anno della pratica" e "Genere"

| % sul totale | | Genere | | Totale |
|---|---|---|---|---|
| | | Maschi | Femmine | |
| Frequenza totale nell'anno della pratica sportiva | Mai | 28,0% | 41,2% | 69,2% |
| | Meno di 1 volta al mese | ,7% | ,5% | 1,2% |
| | 1-3 volte al mese | 1,4% | ,5% | 1,9% |
| | 1 volta alla settimana | 4,3% | ,8% | 5,1% |
| | 2 volte alla settimana | 4,2% | 3,3% | 7,5% |
| | 3 e più volte alla settimana | 9,2% | 5,8% | 15,0% |
| Totale | | 47,8% | 52,2% | 100,0% |

# Il caso di variabili qualitative - 14

- Lo studio delle percentuali calcolate relativamente alle modalità della variabile indipendente è di per sé sufficiente per capire se e quali relazioni sono presenti nella tabella;
- si può, peraltro, cercare di sintetizzare con un ***indice*** il livello di associazione fra le due variabili.

# Il caso di variabili qualitative - 15

- Questo indice dovrebbe essere pari a ***0*** in caso di ***assenza di associazione*** fra le due variabili e avere un suo ***massimo*** nel caso di ***massima associazione*** possibile;
- questo massimo potrebbe essere ***normalizzato***, costruendo così un indice compreso fra ***0*** e ***1*** (o fra ***0*** e ***100*** e così via):
- è facile e immediatamente evidente quale è la situazione per il livello 0, in quanto corrisponde alla situazione di assoluta ***indipendenza*** fra le due variabili;
- più complessa è la rappresentazione della ***massima dipendenza*** possibile, perché è legata alla struttura della tabella.

# Il caso di variabili qualitative - 16

- Il caso di ***indipendenza*** si verifica quando tutti i profili per riga e per colonna sono uguali,
  - ad esempio se la composizione percentuale dei livelli di pratica è la stessa per maschi e femmine, indipendentemente dalla numerosità dei due gruppi (maschi e femmine)
  - e, simmetricamente, la percentuale di maschi e femmine è la stessa per ogni livello, uguale a quella presente nell'intero collettivo.

# Il caso di variabili qualitative - 17

- Questo risultato teorico si ottiene, tabella per tabella, moltiplicando i totali di riga per i totali di colonna di ogni coppia di modalità e dividendo per il totale generale delle unità.
- Così nell'ALLEGATO 16 troviamo le frequenza che ci aspetteremmo se non ci fosse alcuna associazione fra il "genere" e la "frequenza di pratica" degli sportivi (cfr. ALLEGATO 11).
- Si definiscono ***frequenze attese***: sono valori teorici, come si nota dal fatto che ci siano valori decimali, non ammissibili per le frequenze.
- Le frequenze marginali di riga e di colonna sono le stesse della tabella d'origine.

# Allegato 16

**Tabella 3.16** - Tavola di contingenza fra "Frequenza totale nell'anno della pratica" e "Genere"

| Frequenze attese | | Genere | | Totale |
|---|---|---|---|---|
| | | Maschi | Femmine | |
| Frequenza totale nell'anno della pratica sportiva | Meno di 1 volta al mese | 9,0 | 5,0 | 14,0 |
| | 1-3 volte al mese | 14,2 | 7,8 | 22,0 |
| | 1 volta alla settimana | 37,5 | 20,5 | 58,0 |
| | 2 volte alla settimana | 54,9 | 30,1 | 85,0 |
| | 3 e più volte alla settimana | 110,4 | 60,6 | 171,0 |
| Totale | | 226,0 | 124,0 | 350,0 |

# Il caso di variabili qualitative - 18

- Se le frequenze della tabella da noi ottenuta, che si definiscono ***frequenze osservate*** coincidessero con quelle *attese*, il valore dell'indice che stiamo cercando per sintetizzare la relazione fra le variabili sarebbe ovviamente uguale a 0!
- Più le due frequenze divergono più forte si può considerare l'associazione fra le variabili!

# Il Chi-quadrato - 1

- Un indice che risponde a queste caratteristiche è il ***Chi-quadrato***, che si ottiene sommando gli scarti fra frequenze osservate e attese al quadrato, rapportati per normalizzarli alle frequenze attese.
- In realtà il Chi-quadrato ha un minimo teorico pari a 0, ma il suo massimo dipende dalle caratteristiche della tabella:
  - esso sarà infatti uguale al più piccolo dei seguenti valori:
    - dimensione del collettivo per numero delle righe meno 1 e
    - dimensione del collettivo per numero delle colonne meno 1.

# Il Chi-quadrato - 2

- Pertanto per normalizzarlo ed avere un massimo pari a 1 dobbiamo dividerlo per questo massimo.
- Otteniamo così un nuovo indice che è la ***V di Cramer***.
- IL Chi-quadrato  gioca un ruolo prioritario nel Test di ipotesi.

# Il Chi-quadrato - 3

- La formula esatta della V di Cramer è la seguente:

$$V = Sqrt(\chi^2/N*(min\ (r,c) - 1))$$

- Nella nostra tabella il Chi-quadrato è pari a **14,217**, N è **350**, il minimo fra r e c è **2** e quindi

$$V = .202.$$

# Il Chi-quadrato - 4

- La formula esatta della V di Cramer è la seguente:

$$V = \mathrm{Sqrt}(\chi^2/N*(\min(r,c) - 1))$$

- Nella nostra tabella il Chi-quadrato è pari a **14,217**, N è **350**, il minimo fra r e c è **2** e quindi

$$V = .202.$$

# Il Chi-quadrato - 5

- Molti altri sono i coefficienti che possono essere calcolati per misurare l'associazione fra le variabili, a seconda se la relazione possa essere trattata
  - come **simmetrica** (***Phi quadro***, ***Phi***, ***Coefficiente di contingenza***) o
  - come **asimmetrica**, ovvero con lo studio solo dell'effetto di una variabile sull'altra;
- a seconda se le variabili abbiamo modalità
  - **ordinabili** o
  - **non ordinabili**.

# Il Chi-quadrato - 6

- Qui si è voluto proporre solo i due coefficienti, il Chi quadrato e la V di Cramer, emblematici dell'approccio inferenziale e di quello descrittivo sintetico: lasciamo ai corsi di Statistica metodologica la descrizione delle caratteristiche e dell'utilizzabilità degli altri indici.

# Il Chi-quadro nel test di ipotesi -1

- Come è possibile utilizzare il Chi quadrato in un'ottica inferenziale?
- Negli *output* dei principali *software*, accanto al valore vengono indicati i gradi di libertà* e il *p-value*, ovvero la probabilità di ottenere i valori osservati presenti nella tabella di contingenza, se fosse vera l'ipotesi di indipendenza.
- Se il *p-value* è pari a 0,007, vuol dire che, se fosse vera l'ipotesi di indipendenza, il risultato rappresentato dai valori osservati si verificherebbe in soli 7 campioni su 1000!

# Il Chi-quadro nel test di ipotesi -2

- Si dice allora che il *p-value* è significativo: questo porta a rifiutare l'ipotesi nulla e ad affermare che fra le due variabili c'è una qualche associazione, con un rischio di sbagliare in tale affermazione pari proprio a 7 per 1000.
- Ovviamente, se volessimo dire qual è la misura di tale associazione, potremmo usare la già nota V di Cramer.
- In genere si accetta un errore al massimo dello 0,05. Se si vuole essere più sicuri dello 0,01.

# I gradi di libertà

- I gradi di libertà («*df*» *degrees of freedom*) sono un parametro della distribuzione del Chi-quadrato, che si calcola sottraendo al numero delle osservazioni quello dei vincoli della tavola: le osservazioni sono le celle (righe per colonne), i vincoli sono le celle marginali (righe più colonne meno uno, il totale generale).
- In altre distribuzioni il calcolo è diverso, ma il significato è analogo.

# La distribuzione campionaria del Chi-quadrato

- Tale distribuzione è nota ed è nelle appendici di ogni testo di Statistica, con df= (r-1)*(c-1) (con r = righe e c = colonne).
- Se non avessimo il *p-value*, potremmo consultare le tavole e confrontare il valore ottenuto con quello di riferimento per il livello di fiducia prescelto.
- Se quello ottenuto sui dati osservati è maggiore di quello tabulato si rifiuta l'ipotesi nulla, altrimenti non è possibile farlo.

# Caratteristiche della tabella

- Un'importante cautela da considerare nell'uso del Chi-quadrato con fini inferenziali è quella di osservare la percentuale di celle che hanno una frequenza attesa inferiore a 5.
- Se questa percentuale supera il 20, il risultato del Chi-quadrato non è accettabile, perché il coefficiente è troppo condizionato dalle basse frequenze.
- In questi casi è necessario accorpare le modalità delle variabili finché non si raggiunga una numerosità adeguata.