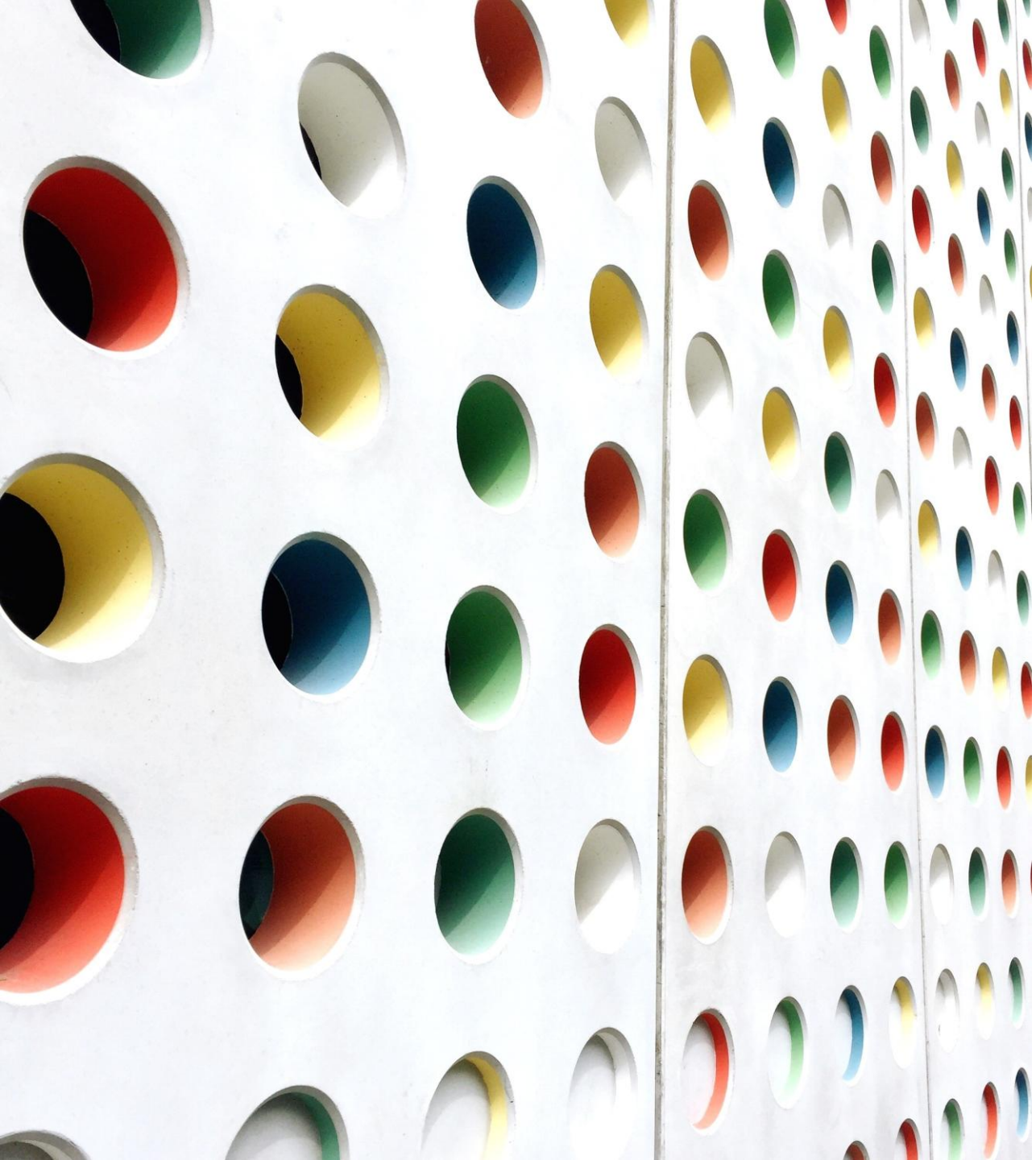


UniBa UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

DIEF DIPARTIMENTO DI ECONOMIA E FINANZA

ORIENTAMENTO CONSAPEVOLE

IMPARARE DAI DATI: LA STATISTICA COME STRUMENTO DELLA CONOSCENZA
A. A. 2024-2025

FACCIAMO STATISTICA! CONOSCERE LA REALTÀ CHE CI CIRCONDA ATTRAVERSO I NUMERI

Prof.ssa Crescenza Calculli crescenza.calculli@uniba.it
Prof.ssa Nunziata Ribecco nunziata.ribecco@uniba.it

Che cos'è la Statistica (1/2)

Alcune domande

- Qual è la **media dei voti in matematica** degli studenti della tua scuola?
- Qual è lo **sport preferito** dei giovani italiani?
- Qual è il **monumento più visitato** in Italia nell'ultimo anno?

La Statistica è una Scienza

Studia i fenomeni collettivi (e naturali, ambientali) allo scopo di:

- **descrivere** i fenomeni
- **individuare relazioni** tra le manifestazioni di uno o più fenomeni
- **estrarre informazioni** per supportare il processo decisionale

Questa foto di Autore sconosciuto è concesso in licenza da CC BY-NC

Che cos'è la Statistica (2/2)

Esempi di fenomeni collettivi

- nascite/morti, malattie
- consumo e/o utilizzo di determinati prodotti

Esempi di **fenomeni naturali e ambientali**

- terremoti, eruzioni
- inquinamento atmosferico, delle acque
- presenza di determinate specie

Nella nostra società dell'informazione e della comunicazione, la Statistica **e** il ragionamento statistico sono ovunque

- nei telegiornali compaiono ogni giorno **grafici, tabelle**, risultati di **sondaggi**
- le carte fedeltà registrano i **dati** relativi ai nostri **acquisti, abitudini** e **gusti**
- in farmacia acquistiamo **prodotti testati e sperimentati**

A cosa serve la Statistica

La Statistica serve quindi:

📊 a **comprendere i fenomeni**
Raccoglie, analizza e interpreta dati per studiare il mondo reale

📈 a **supportare le decisioni**
Valuta ciò che osserviamo e guida decisioni basate su dati (in diversi ambiti, come economia, medicina, scienze sociali e ingegneria)

🎲 a **gestire l'incertezza**
Affronta aspetti come *casualità*, *aleatorietà* e **incertezza**

🔬 a **sostenere il progresso scientifico**
Strumento chiave nelle scienze sperimentali, naturali ed empiriche

❓ a **modellare la probabilità (e non le certezze)**
Non stabilisce verità assolute, ma attribuisce **probabilità** alle affermazioni

Chi è lo Statistico

Lo Statistico è un esperto nell'analisi dei dati. Può essere anche definito come:

Data Analyst

- Si occupa di **esplorare** e **interpretare** dati per identificare tendenze
- Usa **strumenti** come Excel, SQL, Tableau, Power BI per analizzare dati strutturati
- Si concentra su **reportistica** e **visualizzazione dei dati**
- Ha **competenze statistiche di base**

Data Scientist

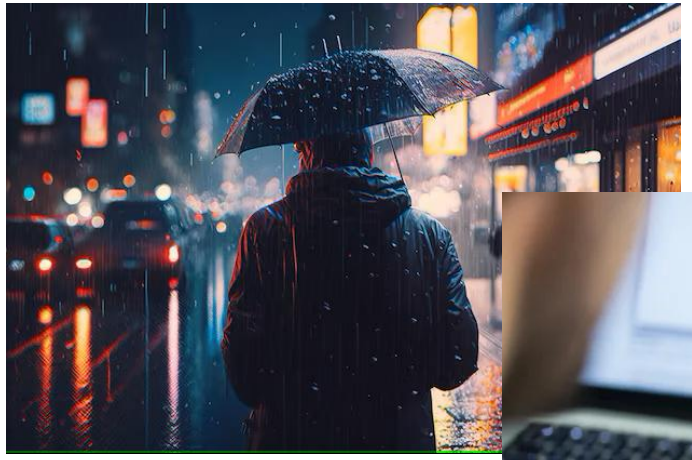
- Va oltre l'analisi descrittiva; **sviluppa modelli predittivi** e algoritmi di Machine Learning
- Usa **strumenti avanzati** come R, Python , Spark, Hadoop
- Lavora con **Big Data** e dati anche non strutturati (testi, immagini, sensori IoT)
- Ha un **forte background in statistica, matematica e programmazione**

Tuttavia…

Siamo tutti statistici!

Tutti noi, anche se **inconsapevolmente**, sfruttiamo la Statistica:

- consultiamo il meteo
- assumiamo un farmaco testato
- scommettiamo su un risultato sportivo



Ma perché tutto questo interesse verso questa figura professionale?

I **dati** sono sempre più **abbondanti** e **pervasivi** nella nostra società (si parla di esplosione di dati!) e, di conseguenza, la **capacità di trattarli e capirli**, di dare loro un significato diventa sempre più rilevante…

L’era dei Big Data

Dati in **crescita esponenziale**

- Già nel 2018, si stimava che ogni giorno vengono creati 2,5 quintilioni di byte di dati (fonte: Forbes, 2018) (1 quintilione=10^{30} ossia 1000000^5)

- Con AI, cloud computing e digitalizzazione di ogni settore, questa cifra è destinata ad aumentare

Cosa sono i **Big Data**?

Grandi volumi di dati, strutturati o non strutturati, che sono così vasti e complessi da non poter essere trattati con i tradizionali strumenti di gestione e analisi dei dati. Sono caratterizzati da (**3 V**)

- **Volume**: enorme quantità di dati da fonti come sensori, social media, transazioni online, dispositivi IoT (es. dispositivi medici di telemonitoraggio)
- **Velocità**: rapidità con cui i dati vengono generati ed elaborati, analizzati e trasformati in informazioni utili (es. dati in tempo reale come quelli delle transazioni finanziarie)
- **Varietà**: diversità dei dati, che può includere dati strutturati (come tabelle di database) e non strutturati (come testi, immagini, video)

Strumenti statistici per la trattazione dei Big Data

Le **tecniche statistiche** di elaborazione e trattazione dei Big Data sono fondamentali per **trasformare** enormi quantità di **dati grezzi** in informazioni utili e conoscenza

Poiché i big data sono complessi, eterogenei e in continua evoluzione, è necessario applicare metodi avanzati di analisi e manipolazione per ottenere risultati significativi

Alcune tecniche utilizzate:

- **Analisi descrittiva** (es. Statistiche di base, Distribuzioni, Rappresentazioni grafiche)
- **Preprocessing dei Dati** (pulizia, normalizzazione e standardizzazione, riduzione della dimensionalità, es. PCA)
- **Analisi Predittiva e Modelli Statistici** (modelli di dipendenza, es. regressione; classificazione)
- **Machine learning** (alberi decisionali, reti neurali, ecc.)
- **Text Mining** (es. Analisi del sentiment, utilizzate sui dati dei social media)
- **Data Mining** (es. individuazione di regole di classificazione che associano variabili indipendenti a variabili dipendenti, utile per predire comportamenti o categorie)

La classifica dei lavori più richiesti

Fonte: **World Economic Forum**, *The future of jobs report* (9 Gennaio 2024)

1. AI & **Machine Learning Specialists**: Creano algoritmi per automatizzare processi e fare previsioni
2. Sustainability Specialists: Sviluppano strategie ecologiche per le aziende
3. **Business Intelligence Analysts**: Analizzano i dati aziendali per supportare decisioni strategiche
4. Information Security Analysts: Proteggono i dati aziendali da minacce
5. Fintech Engineers: Sviluppano soluzioni tecnologiche per il settore finanziario
6. **Data Analysts** e **Data Scientists**: Analizzano i dati per supportare scelte aziendali e Creano modelli predittivi per supportare decisioni strategiche
7. Robotics Engineers: Progettano robot per vari settori
8. **Big Data Specialists**: Gestiscono grandi volumi di dati con modelli statistici avanzati
9. Agricultural Equipment Operators: Gestiscono attrezzature agricole avanzate
10. Digital Transformation Specialists: Guida la digitalizzazione aziendale

Percorso di studi per diventare uno Statistico

Corso di Laurea in Scienze Statistiche

(https://www.uniba.it/it/corsi/scienze-statistiche/l-41)

@ Dipartimento di Economia e Finanza
Università degli Studi di Bari

- Corso di Laurea della durata di **tre anni** (180 CFU = Crediti Formativi Universitari)
- Dopo la Laurea Triennale si può proseguire con la **Laurea Magistrale** (120 CFU)

Caratteristiche

- Laurea STEM (Science, Technology, Engineering e Mathematics)
- Accesso libero
- Test per gli OFA (Obblighi Formativi Aggiuntivi) dopo l'immatricolazione
- **Unico corso di Laurea** riguardante le discipline Statistiche in **Puglia e Basilicata**

Sedi di Lauree Triennali e Magistrali in Statistica

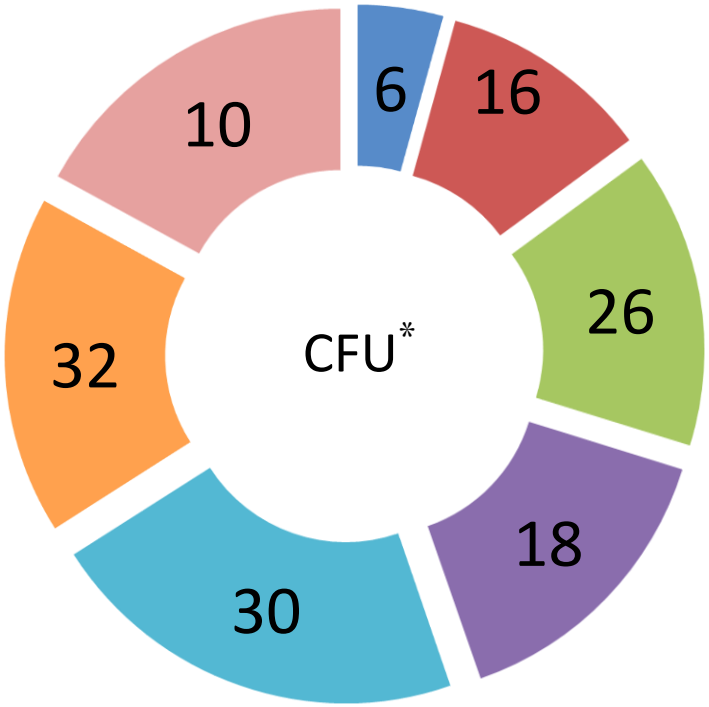
Sud Italia

Università degli Studi di Bari – LT in Scienze Statistiche (L-41)
Università degli Studi di Bari – LM Statistica e Metodi per l'Economia e la Finanza (LM- 83)

Altre sedi:

- Università degli Studi di Napoli Federico II
- Università della Calabria
- Università degli Studi di Salerno
- Università degli Studi del Sannio (Benevento)

Corso di Laurea in Scienze Statistiche: Cosa si studia



Ambiti disciplinari

- Inglese
- Economico
- Statistico Socio-Sanitario
- Statistico-Economico
- Statistico
- Matematico-Probabilistico
- Informatico

Piano di studi dettagliato @ https://www.uniba.it/it/corsi/scienze-statistiche/studiare/didattica/piano-di-studio-coorte-2024-2025/piano-di-studio-coorte-2024-2025

12 CFU a scelta dello studente (es. tirocinio opzionale)

24 CFU a scelta tra due liste di esami

3 CFU per abilità richieste dal Mondo del lavoro
3 CFU per la prova finale

* 1 CFU = 25 h di studio complessivo, di cui 7 di lezione frontale (es: corso da 8 CFU = 56 ore di lezioni frontali)

Le scelte di un laureato in Scienze Statistiche (1/2)

CONTINUA A STUDIARE

Corso di Laurea in Scienze Statistiche (3 anni)

+

Corso di Laurea Magistrale in Statistica e metodi per l'Economia e la Finanza (2 anni)

Scelto anche dai laureati in discipline economiche interessati ad approfondire la conoscenza dei metodi quantitativi

=

Dottore Magistrale in STATISTICA (5 anni: 3+2)

Le scelte di un laureato in Scienze Statistiche (2/2)

INIZIA A LAVORARE

Cosa "sa fare" un laureato in Scienze Statistiche

Valutazione (es. fattori di rischio per malattie), analisi di dati per il supporto alle decisioni (in finanza, politica, genetica, ecc.)

Saper fare un sondaggio d'opinione / indagine campionaria (es. preferenza sul partito politico, stima della presenza turistica)

Saper riassumere dati complessi / costruire indicatori (es. comunali, vendite, consumi, disoccupazione…)

Possibilità occupazionali

- **Compagnie assicurative**: Analisi del rischio, valutazione di trend per ottimizzare le polizze
- **Pubblica amministrazione** (uffici statistici): Raccolta e analisi di dati demografici, economici e sociali per supportare le decisioni politiche e amministrative
- **Istituti di ricerca privati** (medicina, marketing, sondaggi, ecc.): Analisi di dati sperimentali, studi di mercato, previsioni e modellizzazione statistica.
- **Enti pubblici** (turismo, formazione, servizi, ecc.): Elaborazione di dati per migliorare le strategie di sviluppo territoriale e valutare l'efficacia di interventi pubblici
- **Società di consulenza**: Analisi dei dati aziendali, supporto decisionale per strategie di mercato e ottimizzazione dei processi aziendali
- **Istituti di credito e banche**: Valutazione del rischio di credito, previsione delle tendenze finanziarie, analisi di mercato
- **Ospedali e ASL**: Analisi di dati sanitari per migliorare la gestione delle risorse ospedaliere e valutare l'efficacia dei trattamenti
- **Aziende farmaceutiche** (efficacia dei farmaci, ecc.): Sperimentazione clinica, analisi dell'efficacia dei farmaci, studio di dati epidemiologici
- **Aziende di produzione e di servizi** (controllo della qualità, customer satisfaction, ecc.): Analisi dei dati di produzione, controllo qualità, misurazione della soddisfazione dei clienti
- **Istituti di ricerca pubblici** (università, ISMETT, ecc.): Ricerca accademica e applicata, sviluppo di nuovi metodi statistici, analisi di grandi dataset

Possibilità occupazionali e iniziative

Deloitte.

Una delle quattro maggiori società di consulenza e revisione contabile al mondo, incontra gli studenti dei corsi di Statistica ed Economia presso il DiEF (11 marzo 2024)

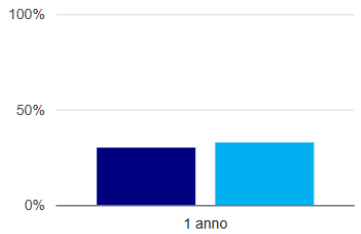
Profilo dei laureati e condizione occupazionale

Indagini del Consorzio

1) Profilo dei Laureati (Link all'indagine 2023)

2) Condizione Occupazionale dei laureati ad uno anno dalla laurea (Link all'indagine 2023)

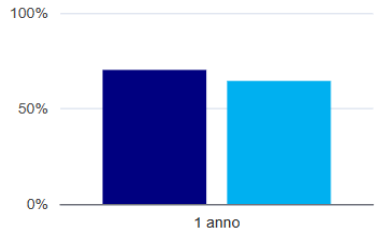
Tasso di occupazione



	1 anno
corso	30,0%
Ateneo	32,7%

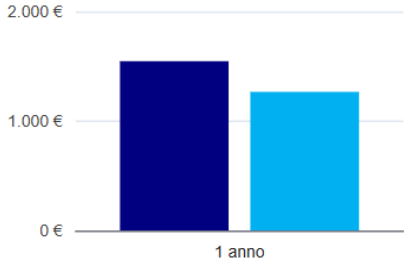
...ma

Laureati attualmente iscritti a un corso di laurea di secondo livello[1]



	1 anno
corso	70,0%
Ateneo	64,2%

Retribuzione mensile netta (medie, in euro)[2]



	1 anno
corso	1.542€
Ateneo	1.261€

Offerta formativa di secondo livello

Corso di Laurea Magistrale in **Statistica e Metodi per l'Economia e la Finanza (SMEF)**
(https://www.uniba.it/it/corsi/smef)

@ Dipartimento di Economia e Finanza
Università degli Studi di Bari

Il corso fornisce al laureato le conoscenze idonee per **analizzare** la realtà **economico-finanziaria** sviluppando una capacità critica e decisionale autonoma, tale da renderlo competitivo sul mercato del lavoro italiano ed estero

Obiettivo formativo: formare **specialisti** nei seguenti ambiti

- **metodologie quantitative** applicate al settore economico, assicurativo, previdenziale e finanziario
- **tecniche attuariali**, della finanza matematica dei mercati e dell'impresa

Profilo professionale

- ampio riscontro nel mercato del lavoro
- esperto, in posizioni di alta responsabilità
- può operare nei seguenti settori:
  - **Compagnie di assicurazione**, Enti del settore previdenziale e finanziario
  - **Istituti di vigilanza** come l'Ivass (per le assicurazioni) e la Covip (per i fondi pensione)
  - Azienda, nelle **Aree gestionali**, Informatica, Data science per l'analisi dei Big Data, Pianificazione e controllo, Gestione dei rischi
  - **Ricerca/Università/Scuola**
  - Libera professione di **Attuario**, previo esame di Stato e iscrizione all'Albo (Sez. A dell'Albo degli Attuari)

Informazioni utili

Coordinatore dei Corsi di Studio in
Scienze Statistiche
Statistica e metodi per l'Economia e la Finanza

Prof.ssa Nunziata Ribecco
nunziata.ribecco@uniba.it

Segreteria didattica DiEF (Dipartimento di Economia e Finanza)
didattica.dief@uniba.it

Sito web del Dipartimento
https://www.uniba.it/it/ricerca/dipartimenti/dse

https://www.instagram.com/dief.uniba/

https://www.facebook.com/p/Dipartimento-di-Economia-e-Finanza-100063473921544/?locale=it_IT

Introduzione ai concetti base della Statistica

La Statistica si occupa di quei fenomeni il cui studio richiede l'osservazione di una pluralità di manifestazioni individuali (fenomeni collettivi) e ne fa una SINTESI

Ma è proprio indispensabile fare una sintesi? Non è meglio guardare dettagliatamente tutte le informazioni a disposizione?

Esempio. In una scuola secondaria di primo grado, gli insegnanti hanno realizzato una indagine sui viaggi e le vacanze estive degli alunni. Hanno chiesto a ciascun ragazzo di esprimersi tra sei tipologie di possibili scelte: **nessun viaggio**, **mare**, **montagna**, **campagna**, **città d'arte**, e **altro**

Risposte

mare	città d'arte	mare	mare	mare
mare	altro	città d'arte	montagna	nessun viaggio
nessun viaggio	mare	altro	altro	nessun viaggio
montagna	mare	mare	montagna	mare
mare	nessun viaggio	nessun viaggio	mare	montagna
nessun viaggio	nessun viaggio	nessun viaggio	mare	campagna
campagna	montagna	nessun viaggio	altro	campagna
città d'arte	nessun viaggio	altro	città d'arte	montagna
nessun viaggio	mare	mare	montagna	altro
mare	mare	mare	nessun viaggio	città d'arte

Quanti sono?

Qual è la **meta preferita**?

Siamo in grado di rispondere in pochi secondi?

Introduzione ai concetti base della Statistica

Tipo di vacanza	Totale
mare	17
nessun viaggio	12
montagna	7
altro	6
città d'arte	5
campagna	3
Totale complessivo	**50**

I dati che prima erano in una successione ora sono ben riassunti da una **tabella**!

La meta preferita è il mare

Gli studenti intervistati in totale sono 50

Chiamiamo ora le cose con il loro nome (un po' di «**statistichese**»)

Lessico statistico

CARATTERE o VARIABILE: è una **caratteristica di interesse** che viene rilevata / misurata / osservata

MODALITÀ: sono **manifestazioni del carattere**, sono numeriche o non numeriche, e devono essere

- Esaustive: devono includere tutti i modi di essere del carattere considerato
- Non sovrapponibili: esattamente individuate o mutuamente escludenti

Nell'esempio precedente, il carattere/variabile rilevato/a dagli insegnanti è il TIPO DI VACANZA e le modalità sono: «nessun viaggio», «mare», «montagna», «campagna», «città d'arte», «altro»

UNITA' STATISTICA: il **singolo elemento** (anche detto unità elementare) su cui vengono osservati i caratteri/variabili
Può essere **naturale** (un essere umano, un'auto) o **convenzionale** (la famiglia, la città)
È sempre necessario definire in modo preciso l'unità statistica o in altre parole l'unità elementare di Indagine

Nell'esempio precedente, le unità statistiche sono i 50 studenti

Lessico statistico

Caratteri

Quantitativi

Esprimibili mediante numeri (valori) riferiti ad una unità di misura. Sono *misurabili*

Discreti

Possono assumere delle quantità numeriche distinte, come ad esempio dei numeri interi, preventivamente individuabili ed elencabili (es. numero di fratelli)

Continui

Possono assumere tutti i valori in un certo intervallo di numeri reali (es. peso, altezza)

Qualitativi

Esprimibili mediante attributi. *Non* sono *misurabili*

Ordinali (o carattere qualitativo rettilineo)

Ordine naturale di successione (es. livello di istruzione)

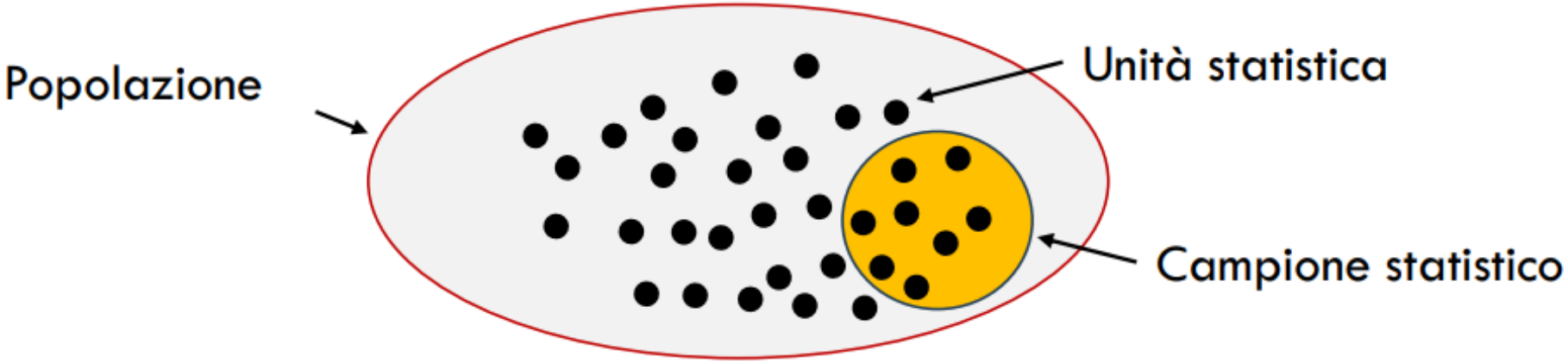
Sconnessi (o carattere qualitativo nominale)

Nessun ordine tra le modalità (es. genere, colore capelli)

Lessico statistico

POPOLAZIONE/COLLETTIVO: è l'**insieme delle unità statistiche** di interesse, omogenee rispetto a uno o più caratteri. Con *N* si indica il numero totale delle unità statistiche che compongono la popolazione

CAMPIONE: **sottoinsieme di unità estratto da una popolazione**, con lo scopo di fornire un'immagine **ridotta** ma **fedele** delle sue caratteristiche. Si definisce rappresentativo se riproduce in modo proporzionale la composizione della popolazione rispetto a variabili rilevanti, evitando distorsioni che possano compromettere l'analisi statistica. L'ampiezza del campione si indica generalmente con *n*



Lessico statistico

INDAGINE STATISTICA: processo di raccolta, elaborazione e analisi dei dati relativi ad un fenomeno per trarne informazioni utili

Fasi dell'indagine:

- **Progettazione dell'indagine**: si definiscono gli obiettivi, la popolazione di riferimento, il metodo di raccolta dati e il campione (se necessario)
- **Rilevazione dei dati**: si raccolgono le informazioni attraverso 1. metodi diretti o 2. indiretti (dove 1. implica la raccolta tramite questionario, intervista, sondaggio postale o telefonico; 2. implica la consultazione di fonti ufficiali come Istat, Eurostat, OCSE, Osservatori, Ministeri, e altre banche dati)
- **Elaborazione dei dati**: i dati vengono organizzati, puliti e analizzati mediante strumenti statistici
- **Presentazione dei dati**: le informazioni vengono rappresentate in tabelle, grafici o report per facilitarne la comprensione
- **Interpretazione dei dati**: si analizzano i risultati per trarre conclusioni e prendere decisioni informate

Si distinguono:

Indagine (o rilevazione) totale: vengono rilevate tutte le unità di un collettivo o popolazione (tutte le N unità). Ne è un esempio il **censimento**

Indagine (o rilevazione) parziale o campionaria: vengono rilevate solo alcune unità della popolazione (la cui numerosità si denota con n)

Lessico statistico

Esempio

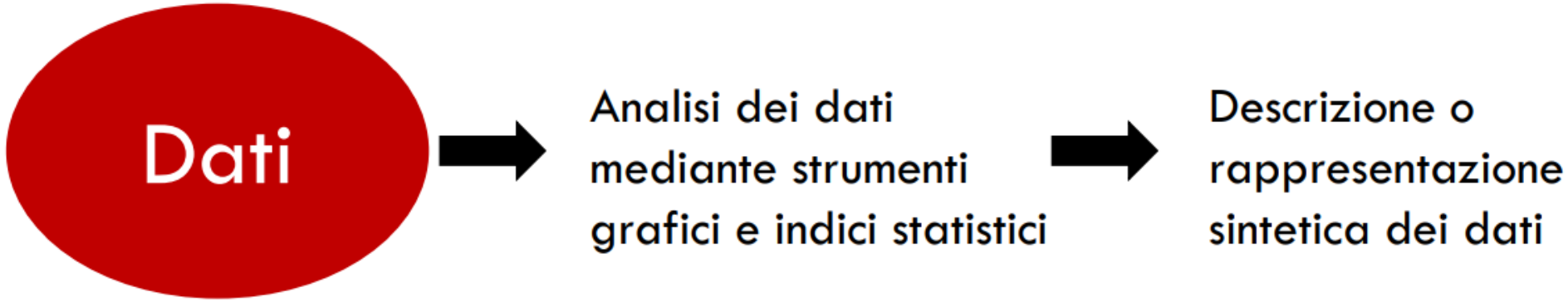
- **Popolazione**: individui residenti in Puglia affetti da diabete
- **Unità**: un individuo residente in Puglia affetto da diabete
- **Campione**: soggetti affetti da diabete in cura presso l'Azienda Ospedaliero Universitaria Consorziale Policlinico (AOUCP); pazienti diabetici di AOUCP registrati nel sistema dal 01 gennaio 2022
- **Variabili quantitative continue**: altezza, peso, pressione sanguigna, glicemia a digiuno, concentrazione di colesterolo nel sangue
- **Variabili quantitative discrete**: età in anni, numero di componenti del nucleo famigliare, anni dalla diagnosi di diabete
- **Variabili qualitative ordinali**: livello di istruzione (elementare, media, secondaria, ecc.), grado di ipertensione (bassa, normale, alta)
- **Variabili qualitative nominali**: colore degli occhi (marrone, azzurro, verde), stato coniugale (celibe/nubile, sposato/a, vedovo/a, divorziato/a)

Strumenti statistici per l'analisi dei dati

Statistica descrittiva

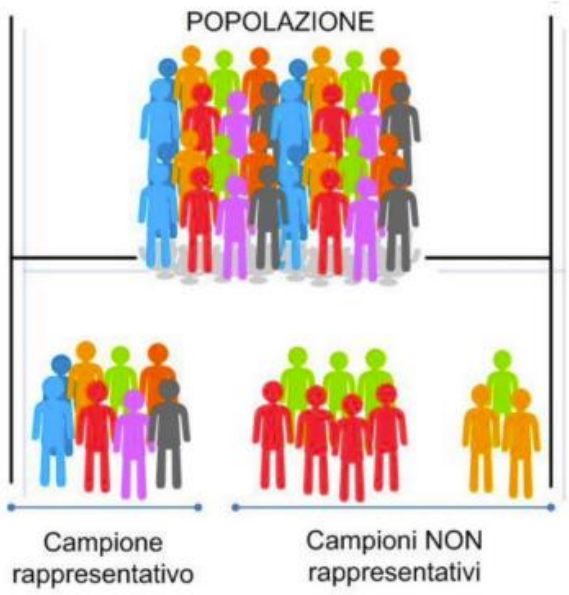
I dati vengono analizzati **senza fare assunzioni** esterne all'insieme di dati considerati. Lo scopo dell'analisi è quindi l'organizzazione dei **dati** in modo da **evidenziarne la struttura**, e di **rappresentarli in modo sintetico** ed efficace

- Indici statistici (indicatori di posizione come la media, di dispersione, come la varianza, di correlazione, di forma, come la curtosi e la skewness, ecc.)
- Strumenti grafici (diagrammi a barre, a torta, istogrammi, boxplot)



Statistica inferenziale o Inferenza statistica (1/2)

I **dati di un campione** statistico vengono analizzati al fine di **trarre**, o **inferire**, delle **caratteristiche della popolazione** di appartenenza. Importante che il campione sia rappresentativo della popolazione di interesse, altrimenti non è detto che le conclusioni tratte sulla base del campione siano applicabili alla popolazione



Statistica inferenziale o Inferenza statistica (2/2)

Per poter inferire delle caratteristiche della popolazione a partire dal campione osservato, si deve tener conto **dell'influenza del caso** sui dati osservati → alcune differenze osservate nel valore di una variabile tra due campioni potrebbero essere semplicemente casuali

Esempio: Immagina di voler sapere qual è l'altezza media degli studenti di una scuola, ma invece di misurarli tutti (cosa che richiederebbe molto tempo), misuri solo un piccolo gruppo scelto a caso (campione). Ora, il problema è che le misurazioni nel campione potrebbero variare per puro caso. Ad esempio, potresti scegliere per sbaglio un gruppo di studenti leggermente più alti o più bassi della media reale della scuola. Per evitare errori di interpretazione, entra in gioco la Statistica inferenziale, che ci aiuta a capire quanto possiamo fidarci delle nostre stime e se le differenze tra due gruppi (ad esempio maschi e femmine) sono dovute al caso o ad una reale differenza.

La Statistica inferenziale:

- definisce un **modello probabilistico del fenomeno studiato** → modello che descrive la probabilità che i dati osservati assumano certi valori
- utilizza quindi i dati per fare inferenze su queste probabilità (si basa sulla Teoria della probabilità)

Strumenti statistici per l'analisi dei dati

Quali sono i principali strumenti di Statistica descrittiva

- Tabelle di frequenza
- Rappresentazioni grafiche
- Indici di posizione
- Indici di dispersione
- Percentili
- Coefficiente di correlazione

Le Frequenze

Le **frequenze** rappresentano l'esito di un **conteggio** oppure l'**intensità** del fenomeno (dipende dal tipo di modalità)

La **tabella statistica** o **tabella di frequenza** (semplice) è uno schema o una rappresentazione concisa delle modalità (valori) di un carattere

- Con n_i si indicano le frequenze, e con x_i le modalità, con l'indice i=1,2,…,S dove S è il numero di modalità
- $\sum_{i=1}^{s} n_i = N$ equivale a $n_1+n_2+n_3+\ldots+n_i+\ldots+n_s=N$

Tabella di frequenza

Modalità del carattere (X_i)	Frequenze assolute (n_i)
x_1	n_1
x_2	n_2
x_3	n_3
.	.
x_i	n_i
.	.
x_s	n_s
Totale	N

Strumenti statistici per l'analisi dei dati

Si distinguono

Frequenze Assolute → $\sum_{i=1}^{s} n_i = N$

Frequenze Relative → $y_i = \frac{n_i}{N}$

Frequenze Percentuali → $P_i = \frac{n_i}{N} \times 100$

Richiamando l'esempio precedente

Tipo di vacanza	Frequenze assolute	Frequenze relative
mare	17	0,34
nessun viaggio	12	0,24
montagna	7	0,14
altro	6	0,12
città d'arte	5	0,1
campagna	3	0,06
Totale	**50**	**1**

Le **Frequenze Assolute**, di due serie di dati, anche dello stesso tipo, **non sono confrontabili** in quanto si riferiscono, in generale, ad un diverso numero di casi complessivi (N)

Strumenti statistici per l'analisi dei dati

Consideriamo due gruppi classificati per il sesso (M = Maschi e F = Femmine), attraverso il calcolo delle frequenze relative possiamo individuare il gruppo con la presenza maschile più elevata

Gruppo 1	Frequenze assolute	Frequenze relative
M	*2*	*0,333*
F	*4*	*0,667*
Tot.	*6*	*1*

Gruppo 2	Frequenze assolute	Frequenze relative
M	*12*	*0,207*
F	*46*	*0,793*
Tot.	*58*	*1*

Strumenti statistici per l'analisi dei dati

Come si procede se si vogliono analizzare le frequenze di una variabile quantitativa continua?

Dobbiamo definire delle
Classi di frequenza

- Il range di valori assunti dalla variabile viene diviso in intervalli, dette **classi**, e per ciascun intervallo si calcolano le frequenze
- il **numero di classi** deve essere equilibrato né troppe, né troppo poche
- le classi, in genere, hanno la **stessa ampiezza**

PESO (Kg)	N° STUDENTI (frequenze)
52	**1**
54	**1**
55	**2**
61	1
63	1
68	2
69	3
71	1
73	1
75	1
TOTALE	**14**

CLASSI DI PESO	N° STUDENTI (frequenze)
50 – 60 Kg	**4**
60 – 70 Kg	7
70 – 80 Kg	3
totale	14

Rappresentazioni grafiche

Obiettivo: illustrare mediante figure, linee o segmenti, superfici o aree, solidi, simboli convenzionali, una distribuzione di frequenze o di intensità in funzione delle modalità di uno o più caratteri

Perché si usano?

- una visualizzazione immediata dell'andamento del fenomeno e della distribuzione, che consente una efficace e globale descrizione dei dati;
- una sintesi con la possibilità, in poco spazio, di confrontare più distribuzioni (curve, spezzate ecc.);
- per finalità investigative:

  ✓mettere in risalto i casi "anomali" (particolari picchi grafici), imputabili a errori nei dati o a effettivi casi anomali, che vanno ulteriormente approfonditi
  ✓evidenziare le relazioni tra caratteri
  ✓permettere di individuare andamenti di fondo (trend) interpolabili con funzioni matematiche (ad es. curva normale, retta ecc.)

Rappresentazioni grafiche di caratteri qualitativi/mutabili

- Grafici a barre
- Diagrammi circolari
- Diagrammi in coordinate polari
- Cartogrammi, mappe tematiche

I Grafici a barre (o ortogrammi)

a) **a colonne**, se sono costituiti da una successione di colonne, segmenti verticali o rettangoli (a base uguale o arbitraria) equidistanti, in numero pari alle modalità del carattere, e hanno altezza uguale o proporzionale alla frequenza (assoluta o relativa) o all'intensità della modalità da rappresentare
b) **a nastri**, se sono costituiti da tanti nastri (segmenti orizzontali, rettangoli) sovrapposti ed equidistanti, in numero pari alle modalità del carattere, e hanno lunghezza uguale o proporzionale alla frequenza (assoluta o relativa) o all'intensità della modalità da rappresentare.

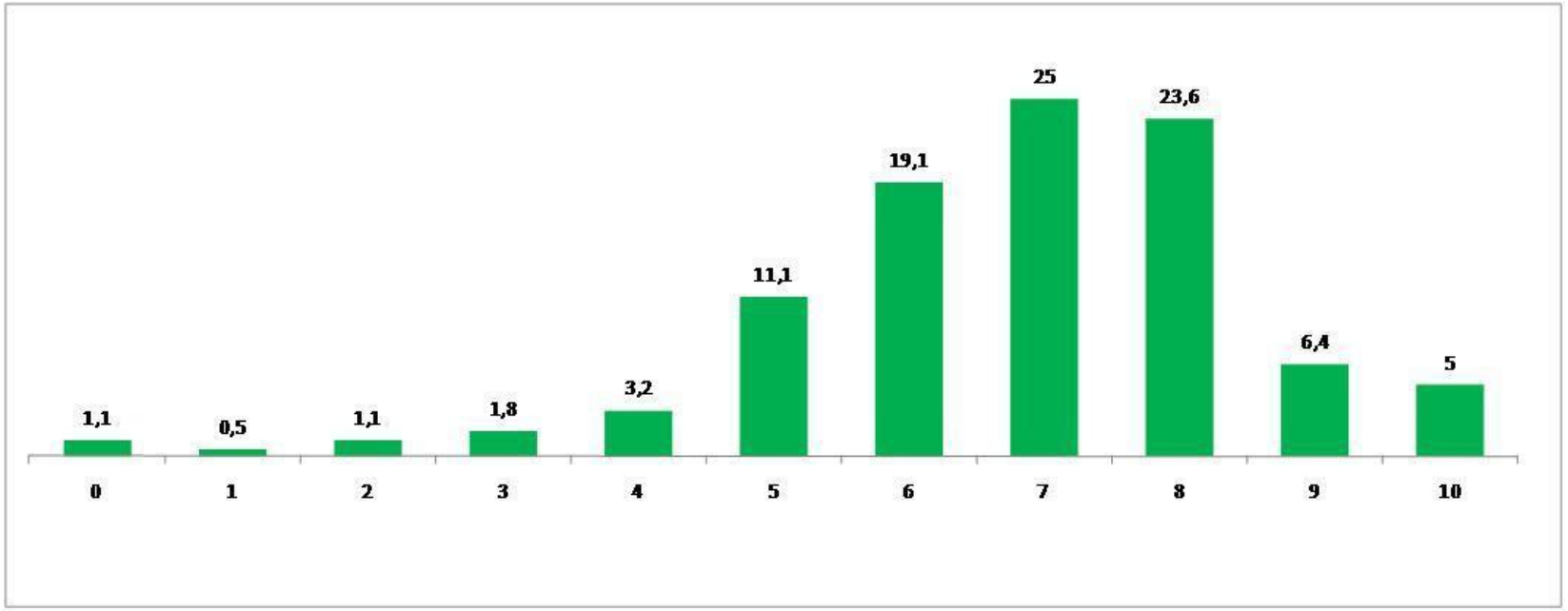
Caratteristiche

I grafici a barre sono pertanto caratterizzati dall'avere **un solo asse** (verticale nel caso a colonne, orizzontale nel caso a nastri) in scala graduata, secondo l'unità di misura che si è scelta per rappresentare le frequenze o intensità. Sull'altro asse, invece, figureranno le modalità (qualitative), per convenzione equidistanti

Strumenti statistici per l'analisi dei dati

a) Ortogramma a colonne

Persone di 14 anni e più per livello di soddisfazione per la vita nel complesso (a) in Italia
Anno 2013 (per 100 persone di 14 anni e più)

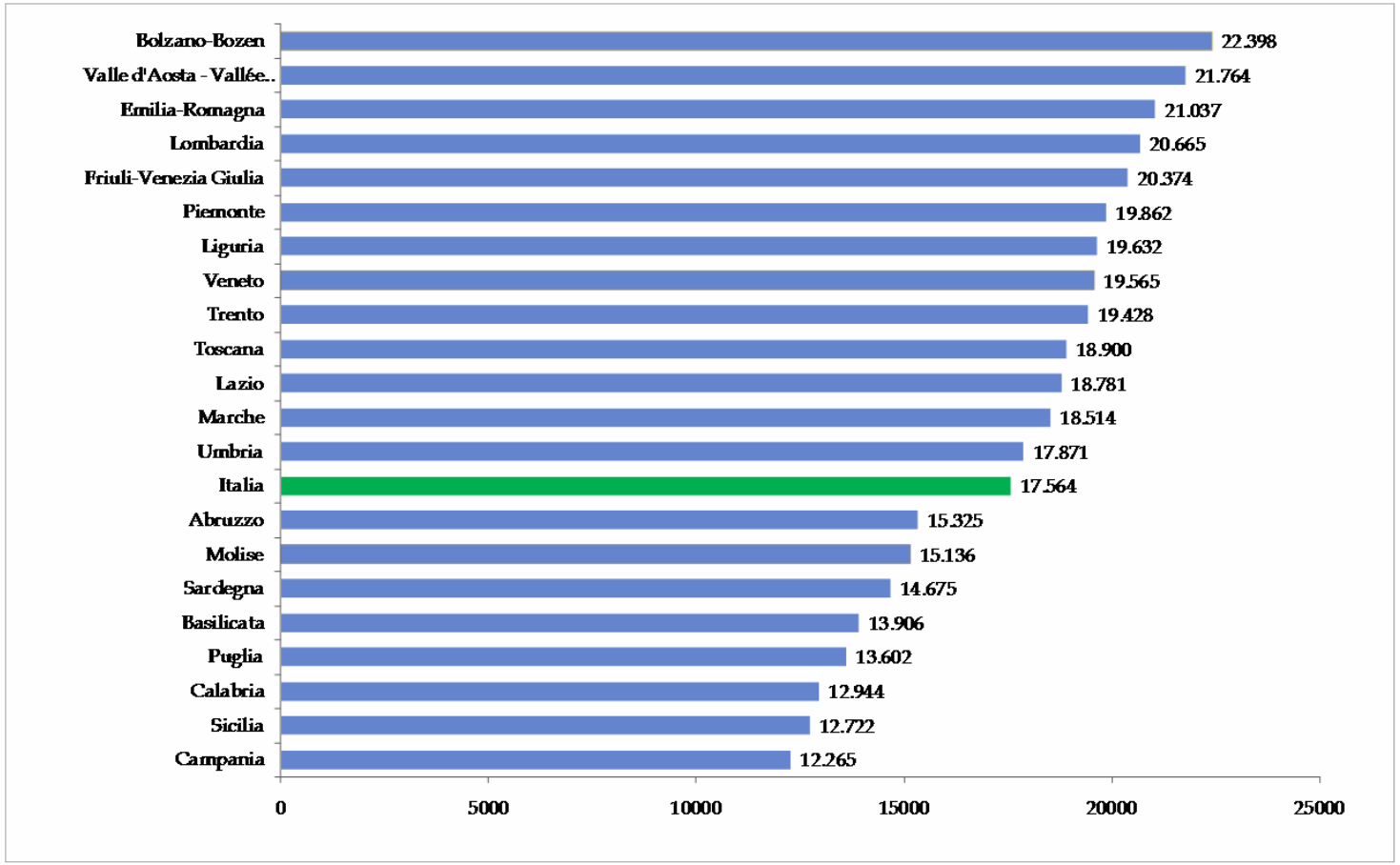


Se la rappresentazione grafica riguarda un carattere sconnesso, l'**ordine** in cui saranno poste le modalità è arbitrario; se si tratta invece di una serie rettilinea/ordinata (come in questo caso), le modalità saranno poste nell'ordine naturale che esse presentano nella serie.

Strumenti statistici per l'analisi dei dati

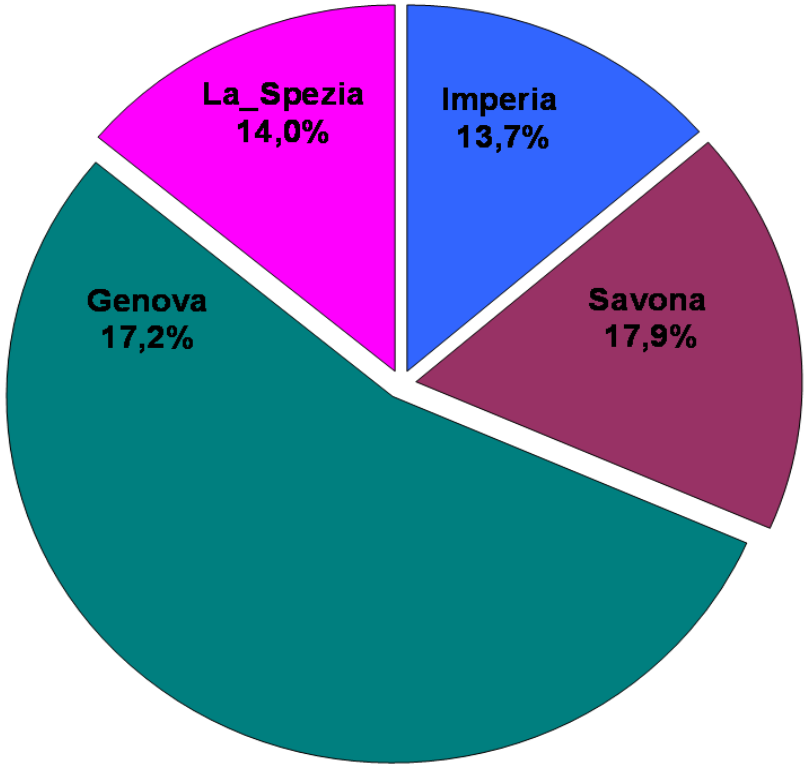
b) Ortogramma a nastri

Reddito disponibile delle famiglie per abitante per regione – Anno 2012 (*valori in euro*)



Diagrammi a settori circolari

I diagrammi a settori circolari (o **aereogrammi**) per la loro forma circolare, sono comunemente noti come "diagrammi a torta".
Sono particolarmente adatti alle serie sconnesse o rettilinee. **Sono efficaci per mettere in evidenza l'importanza relativa delle singole modalità rispetto al totale.**
Gli spicchi o settori, presentano angoli al centro corrispondenti alle frequenze assolute (o relative) delle singole modalità e raggio fisso



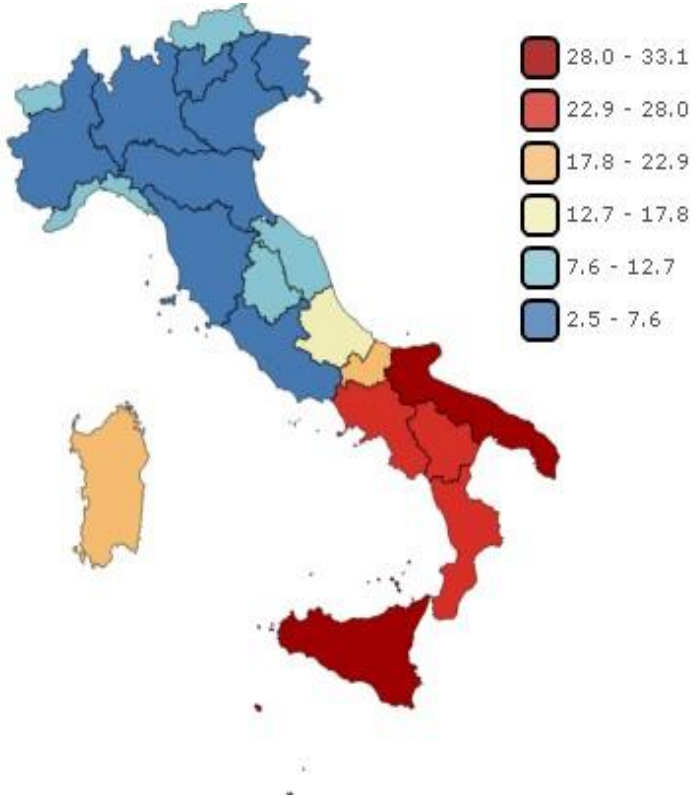
Strumenti statistici per l'analisi dei dati

Cartogrammi, mappe tematiche

I cartogrammi sono grafici utili per rappresentare **serie territoriali**.

Per costruire un cartogramma occorre disporre di una carta geografica o topografica in cui siano chiaramente delimitate le diverse zone, regioni, circoscrizioni (geografiche, politiche, amministrative) rispetto alle quali viene analizzata l'intensità o la frequenza di uno o più caratteri (es. nati, morti, reddito pro capite, secondo le Regioni, Province, Comuni).

Incidenza della povertà relativa per regione
Anno 2012 (valori percentuali)

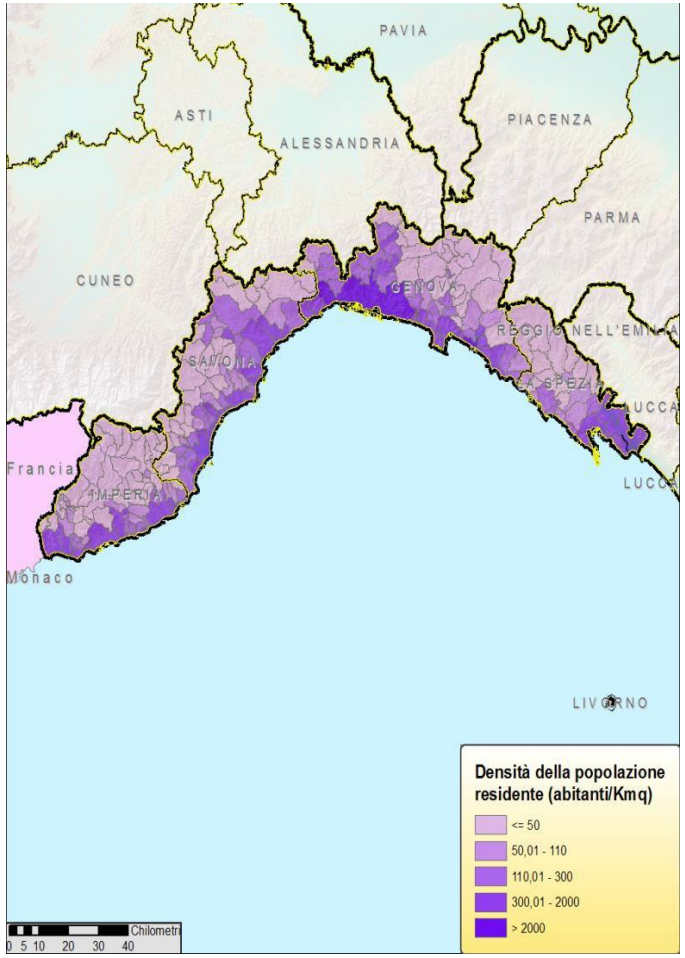


Strumenti statistici per l'analisi dei dati

Cartogrammi, mappe tematiche

Le mappe tematiche consentono, anche se molto empiricamente e in prima approssimazione, di tener conto dell'**autocorrelazione spaziale**, intesa come possibilità che un fenomeno che interessa un certo territorio sia influenzato nelle sue manifestazioni dalla contiguità spaziale esistente tra i luoghi in cui il fenomeno è osservato.

Per Esempio nella mappa della Liguria qui a destra, si nota come la densità della popolazione residente sia maggiore nei comuni costieri e in quelli circostanti i comuni capoluogo rispetto a quelli dell'entroterra.

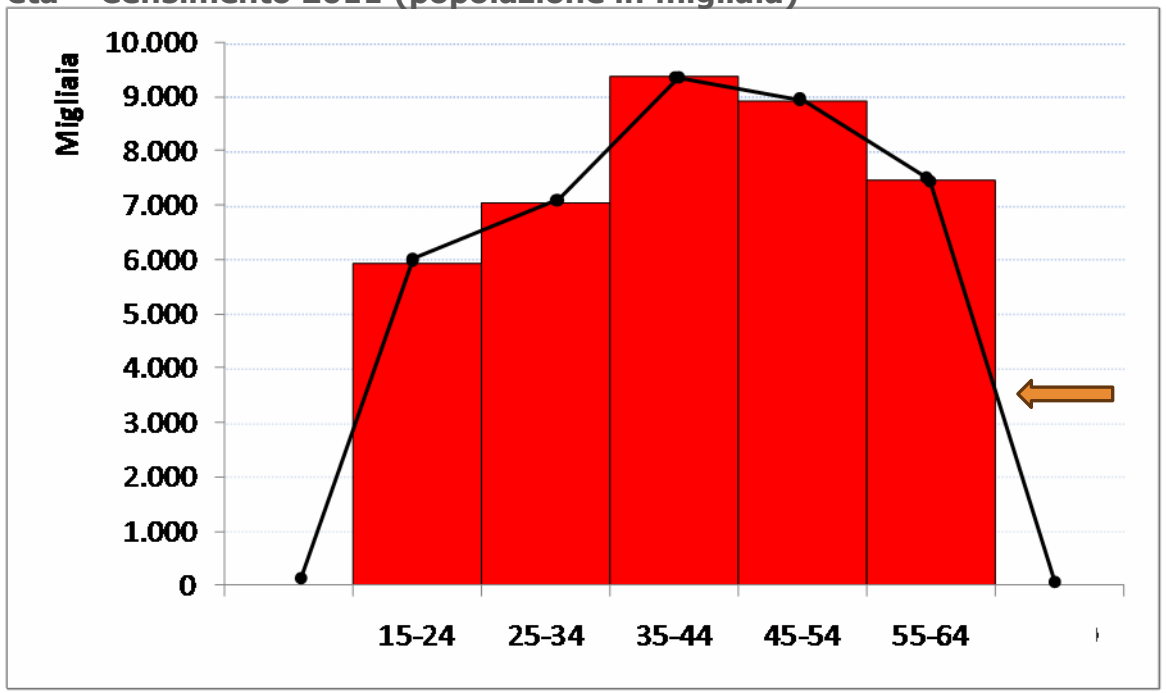


Rappresentazioni grafiche di caratteri quantitativi

ISTOGRAMMA Distribuzioni di frequenza di caratteri quantitativi per i quali le **modalità** sono costituite da **classi di valori**

Caso 1. Le classi di valori hanno uguale ampiezza. In questo caso avremo tanti rettangoli contigui, ciascuno avente base uguale all'ampiezza della classe e altezza uguale o proporzionale alla frequenza (assoluta o relativa) assunta nell'insieme delle unità della classe

Istogramma e poligono di frequenza della popolazione residente in Italia in età lavorativa per classi decennali di età – Censimento 2011 (popolazione in migliaia)



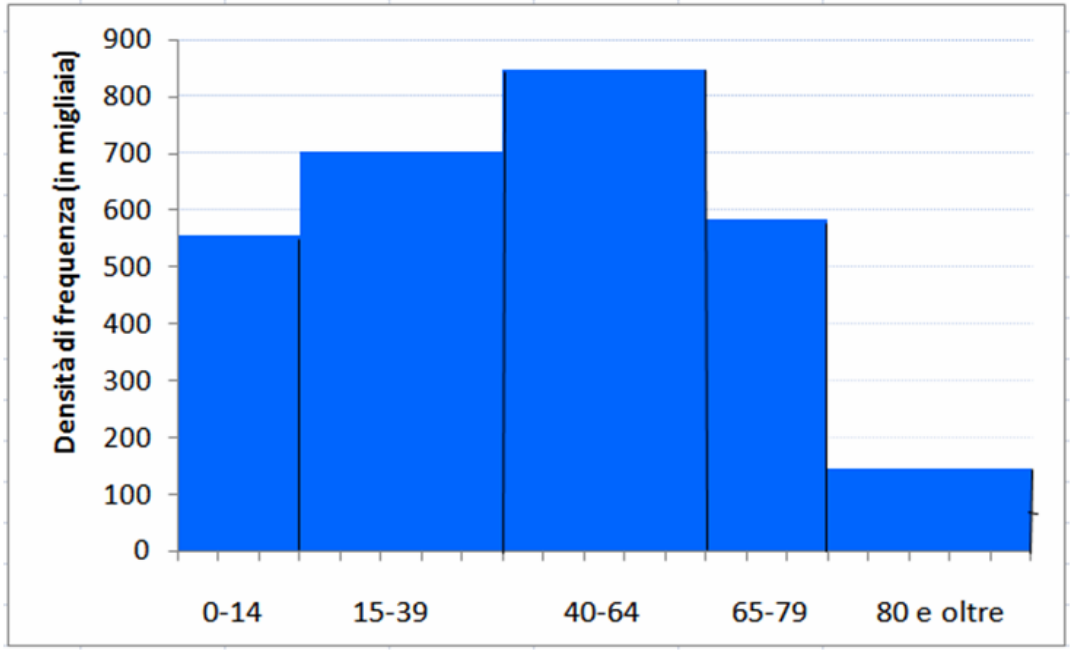
La linea che congiunge il centro delle classi è detta **Poligono di frequenza** (l'area all'interno del poligono equivale a quella dell'Istogramma)

Rappresentazioni grafiche di caratteri quantitativi

ISTOGRAMMA Distribuzioni di frequenza di caratteri quantitativi per i quali le **modalità** sono costituite da **classi di valori**

Caso 2. Le classi di valori hanno diversa ampiezza. In quest'altro caso avremo una serie di rettangoli aventi basi diverse uguali all'ampiezza delle classi e altezze da calcolarsi, in modo che le frequenze siano proporzionali alle aree dei rispettivi rettangoli.

Istogramma della popolazione residente in Italia per classi di età – Censimento 2011 (popolazione in migliaia)



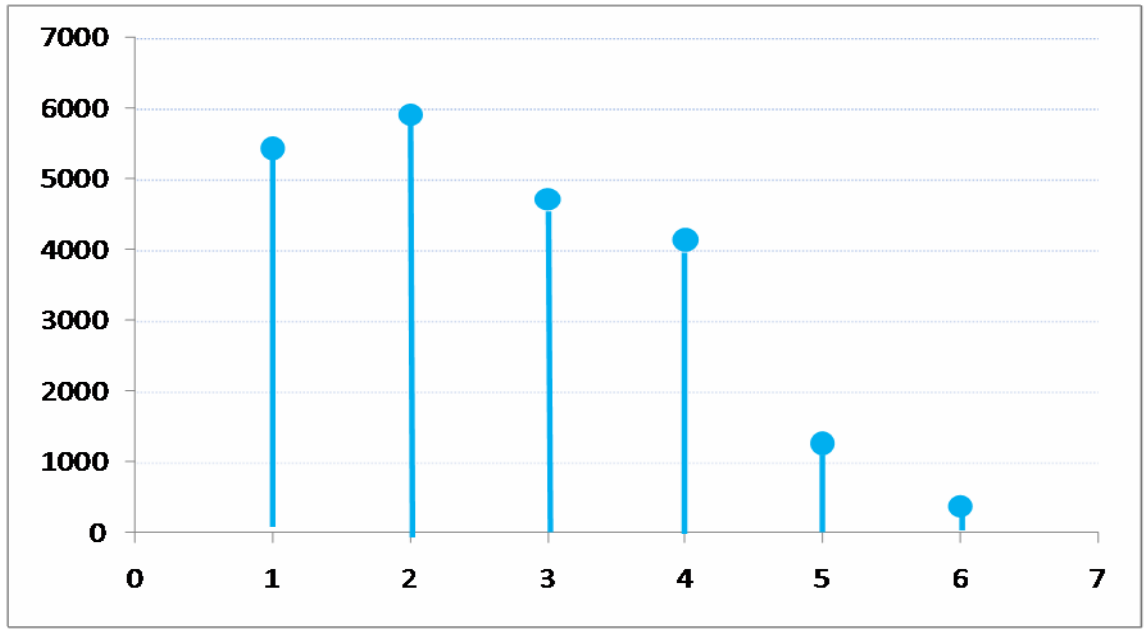
In ordinata si hanno le densità di frequenza, date dal rapporto tra le frequenze assolute o relative e l'ampiezza della relative classi

Rappresentazioni grafiche di caratteri quantitativi

Diagrammi cartesiani a segmenti

Distribuzioni di frequenza di caratteri quantitativi discreti, non divisi in classi, e possono configurarsi a segmenti verticali. Esempi: Numero dei componenti per famiglia, numero delle stanze delle abitazioni, numero di unità locali delle aziende

Distribuzione delle famiglie residenti in Italia per numero di componenti – Censimento 2001 (valori in migliaia)



Si costruiscono come gli usuali diagrammi cartesiani aventi due assi perpendicolari aventi origine comune in O. Ogni coppia ordinata di valori (xi, yi) determinerà un punto nel piano e l'insieme di tutte le coppie (xi = modalità quantitativa i-esima, yi = frequenza della modalità i-esima) determinerà l'insieme dei punti nel piano che costituiscono la rappresentazione grafica della distribuzione considerata. Per rendere maggiormente visibili tali punti, si tracciano dei segmenti verticali congiungenti l'ascissa (xi) con il punto del piano corrispondente all'ordinata (yi).

Rappresentazioni grafiche di caratteri quantitativi

Diagrammi cartesiani per serie storiche

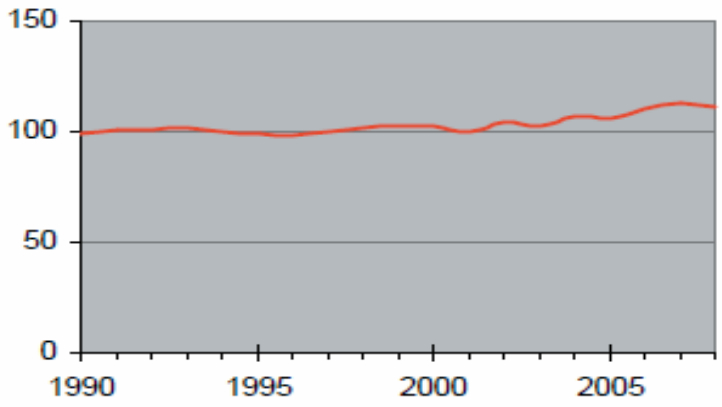
sono impiegati anche per rappresentare graficamente le serie storiche, per i caratteri che si suppone si modifichino con continuità nel tempo

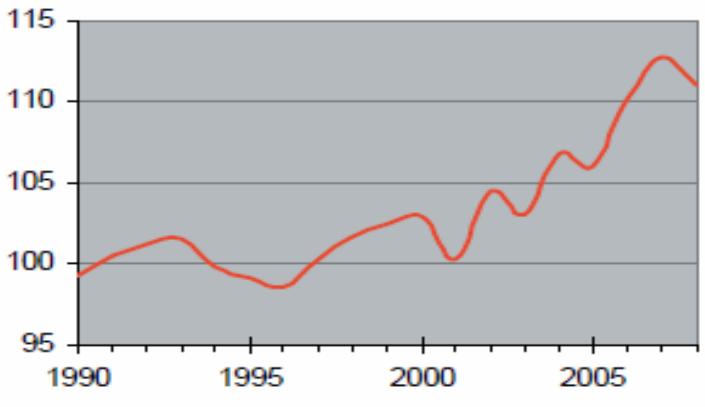
Si costruiscono in maniera analoga rispetto ai diagrammi ortogonali visti in precedenza. In questo caso tuttavia, una volta ottenute le coppie (xi, yi) nel piano cartesiano i punti vanno uniti mediante una spezzata.

Problema di scala

Il grafico a sx mostra un trend stabile stabile per circa 10 anni, seguito da una moderata crescita

Riducendo la scala (grafico a dx), gli stessi dati evidenziano variazioni più marcate intervenute nei primi dieci anni e sembra più marcata la crescita dell'ultimo periodo.





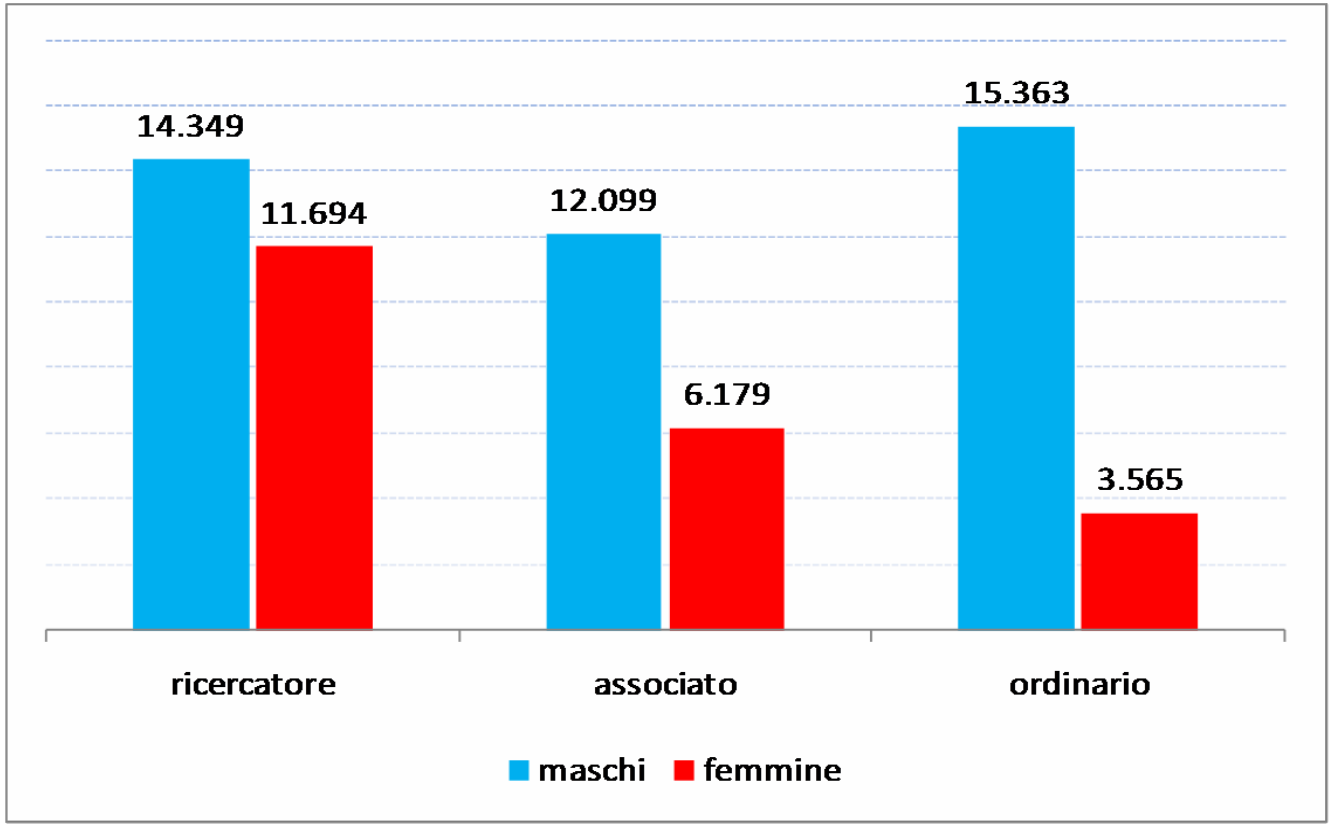
Rappresentazioni grafiche di due caratteri

Nel caso di distribuzioni statistiche doppie in cui vi sono due caratteri qualitativi o una variabile quantitativa e una qualitativa (**tabelle di contingenza**) esistono diverse possibili rappresentazioni grafiche, a seconda della natura dei due caratteri considerati e degli scopi della rappresentazione grafica

Tra le rappresentazioni più impiegate vi è quella già vista per le mutabili statistiche semplici, il diagramma a barre, anche se in questo caso sono rappresentati i due caratteri e le frequenze di associazione delle modalità dei medesimi: Diagramma a barre verticali e Diagramma a barre per impilamento

Rappresentazioni grafiche di due caratteri

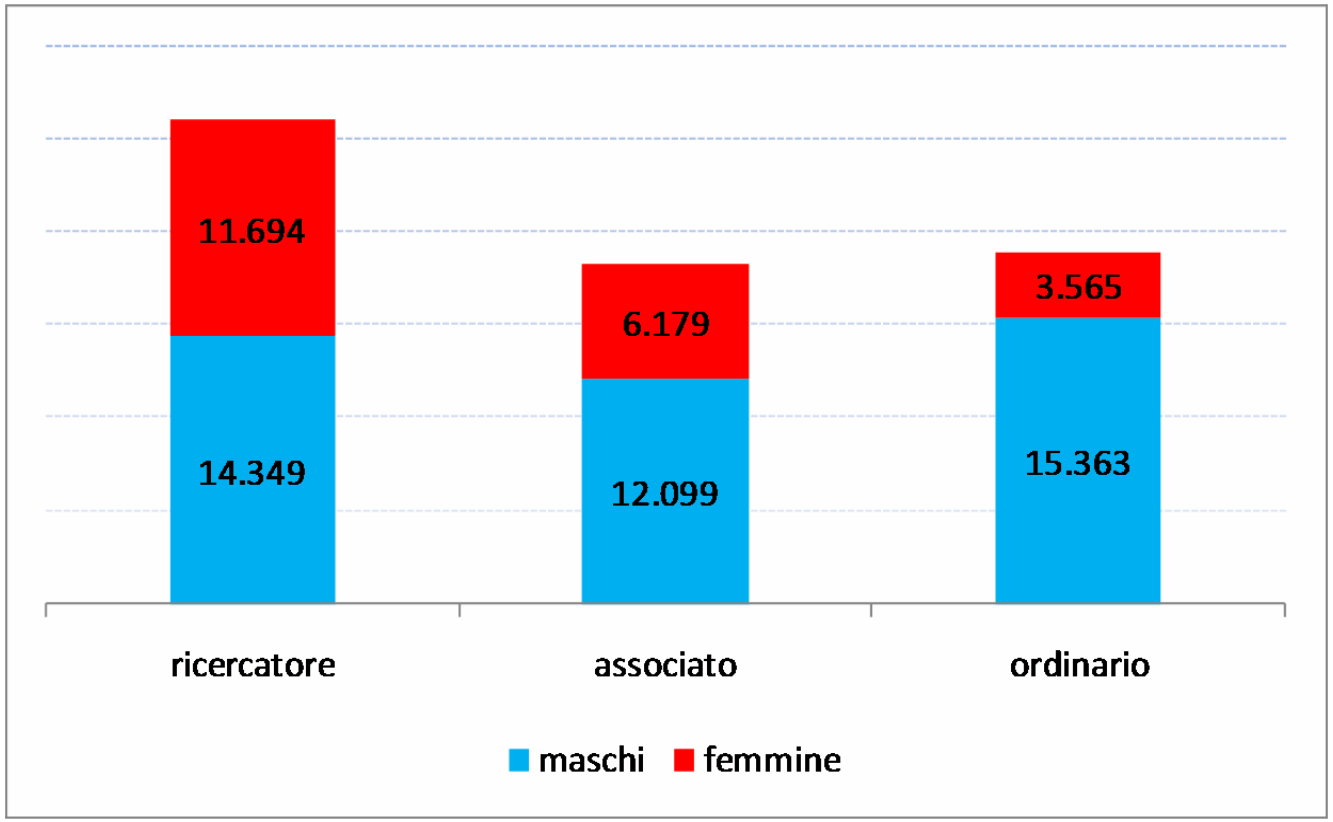
Docenti universitari per qualifica e genere – Italia – Anno 2008



Fonte: Istat

Rappresentazioni grafiche di due caratteri

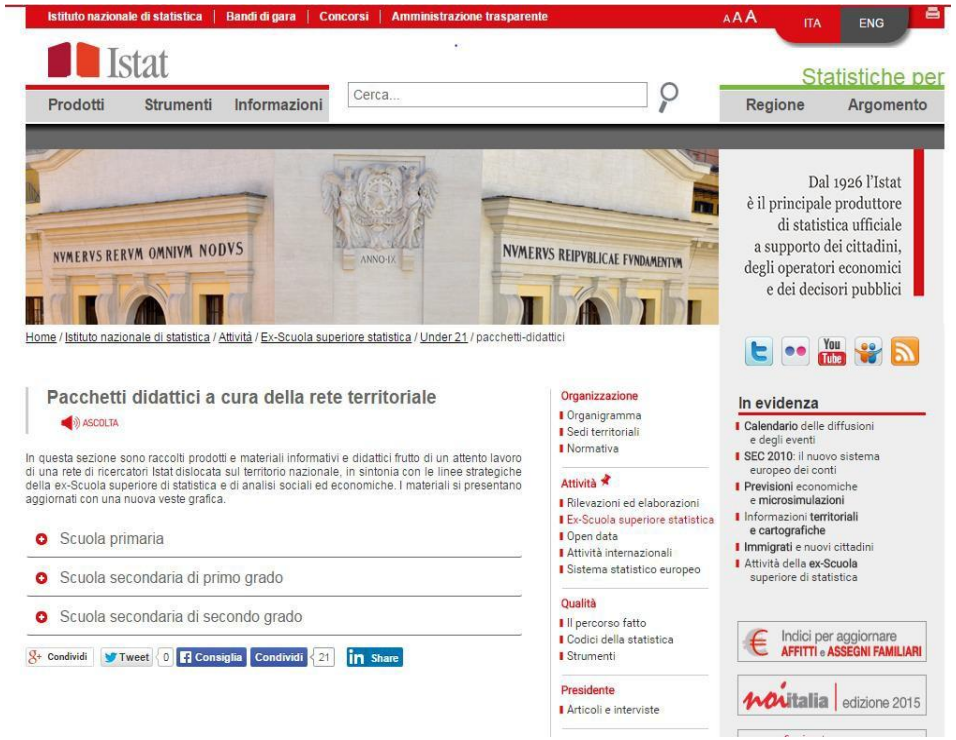
Docenti universitari per qualifica e genere – Italia – Anno 2008



Fonte: Istat

Strumenti statistici per l'analisi dei dati

Rappresentazioni grafiche di due caratteri - Approfondimenti



http://www.istat.it/it/istituto-nazionale-di-statistica/attivit%C3%A0/ex-scuola-superiore-statistica/under-21/pacchetti-didattici

Indici di posizione

A partire dai dati possiamo calcolare delle **statistiche sintetiche**, ovvero grandezze che rappresentato alcune caratteristiche dei dati

- Gli indici di posizione rappresentano il «**centro**» dell'insieme di dati, ovvero un valore attorno al quali i dati si distribuiscono
  - Media (semplice, ponderata)
  - Mediana
  - Moda

Effettuando il calcolo della Media si ottiene un valore di sintesi (cioè un unico valore). Ma questo valore è REALMENTE rappresentativo del fenomeno? Occorre valutare la capacità di sintesi della media aritmetica: più è alta la VARIABILITÀ, meno informativa risulta la Media aritmetica

Strumenti statistici per l'analisi dei dati

Supponiamo di avere un insieme di n dati $x_1, x_2, \ldots, x_n$

- Si dice **Media** semplice e si denota $\bar{x}$ la quantità

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

Esempio

Voti degli esami sostenuti da Mario Rossi: 24, 26, 30, 24, 28, 30, 26, 24, 27, 26

Media : $\bar{x} = \frac{1}{10} \times 265 = 26{,}5$

Strumenti statistici per l'analisi dei dati

Se nei dati compaiono k valori distinti $v_{1,}\, v_2, \dots, v_k$ con frequenze assolute $F_{1,}\, F_2, \dots, F_k$ e frequenze relative $f_{1,}\, f_2, \dots, f_k$, la Media può essere calcolata come **media pesata (ponderata)** dei valori assunti dai dati. Il peso di ciascun dato è la sua frequenza relativa

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} F_i \times v_i = \sum_{i=1}^{k} \frac{F_i}{n} \times v_i = \sum_{i=1}^{k} f_i \times v_i$$

Esempio:

Voti degli esami sostenuti da Mario Rossi: 24, 26, 30, 24, 28, 30, 26, 24, 27, 26

Media ponderata: $\bar{x} = \frac{1}{10} \times (24 \times 3 + 26 \times 3 + 27 \times 1 + 28 \times 1 + 30 \times 2) = 26{,}5$

Strumenti statistici per l'analisi dei dati

La **Mediana** rappresenta il **valore centrale** una volta che i dati sono stati ordinati dal più piccolo al più grande

Definizione formale: Assegnato un insieme di dati di ampiezza n, lo si ordini dal minore al maggiore

- se n è dispari, si dice Mediana il valore del dato in posizione $(n+1)/2$
- se n è pari, la Mediana è la media aritmetica tra i valori dei dati in posizione $n/2$ e $n/2+1$

Esempio 1.

Voti degli esami: 24, 26, 30, 24, 28, 30, 26, 24, 27, 27, 29

Voti ordinati: 24, 24, 24, 26, 26, **27**, 27, 28, 29, 30, 30

$n = 11 \rightarrow n$ è dispari $\rightarrow$ la Mediana è il valore in posizione 6 $\rightarrow$ mediana = **27**

Esempio 2.

Voti degli esami: 24, 26, 30, 24, 28, 30, 26, 24, 27, 27

Voti ordinati: 24, 24, 24, 26, **26, 27**, 27, 28, 30, 30

n = 10 $\rightarrow$ n è pari $\rightarrow$ la Mediana è la media aritmetica tra i voti in posizione 5 e 6 $\rightarrow$ mediana = (26+27)/2 = **26,5**

Media vs Mediana

- La Media fa uso di tutti i dati ed, in particolare, è influenzata in maniera sensibile da valori eccezionalmente alti o bassi, i cosiddetti **outlier**

- La Mediana, facendo leva sull'ordine delle osservazioni (interessano quelle che, in ordine, sono intorno alla metà) e non sul loro valore (il valore che entra il gioco è solo quello di uno o due campioni) non risente particolarmente dei valori estremi

Esempio

Valori di pressione sanguigna sistolica (in mmHg)

105, 108, 109, 110, 110, 111, 112, 113, 115, 150, 160

Media = 118.45 mmHg

Mediana = 111 mmHg

Moda

- La **Moda** di un insieme di dati, se esiste, è l'unico valore che assume **frequenza massima**
- Se i valori con frequenza massima sono più di uno, essi vengono detti valori modali
- La Moda, a differenza di Media e Mediana, ha senso solo per variabili discrete che assumono solo un numero limitato di valori

Esempio.
Voti degli esami: 24, 26, 30, 24, 28, 30, 26, 24, 27, 27, 29

Moda = 24

Voto	Frequenza assoluta
24	3
26	2
27	2
28	1
29	1
30	2

Interpretazione di Media, Mediana e Moda

Stipendio netto mensile dei dipendenti dell'azienda XYZ

Stipendio [euro]	Frequenza assoluta [n. dipendenti]
1300	2
1700	22
2200	19
2600	3
6500	2
9400	1
23000	1

- Media = 2700 euro
  - Se tutti i dipendenti avessero lo stesso stipendio netto, questo sarebbe di 2700 euro.
- Mediana = 2200 euro
  - Circa metà dei dipendenti percepisce meno di 2200 euro, e circa metà più di 2200 euro.
- Moda = 1700 euro
  - Lo stipendio netto più comune è 1700 euro.

La Variabilità (la vera ragione per cui esiste la Statistica!)

Il paradosso di Trilussa

Per ironizzare, la Statistica è spesso associata alla poesia
«..te tocca un pollo all'anno/ e, se nun entra nelle spese tue/
T'entra nel la statistica lo stesso/ perché c'è un antro che ne magna due…»

Costruiamo una Tabella di esempio

Polli (xi)	Consumatori (ni)
0	9
10	1
	N=10

La Media aritmetica (cfr. Trilussa) di 1 pollo a testa deriverebbe da (0*9+10*1)/10=1 ma non è rappresentativa perché il 90% dei soggetti consuma fino a 0 polli… questo è un caso particolare: la **massima variabilità**. Occorre approfondire….

Definizione di Variabilità

La Variabilità è l'attitudine del fenomeno quantitativo ad assumere **diverse modalità** o meglio la tendenza di ogni singola osservazione ad assumere **valori diversi rispetto al valore medio**

- La variabilità cresce all'aumentare della «diversità» tra le osservazioni
- Si definisce scarto la differenza tra la singola osservazione e la media

$$scarto = x_i - \bar{x}$$

Per misurare la variabilità presente in una serie di dati usiamo gli Indici di variabilità o dispersione

Indici di variabilità

Gli indici si distinguono in:

1. Indici di **Variabilità assoluta**, che sono espressi nella stessa unità di misura del fenomeno osservato e non sono né valutabili (è tanta o poca?) né confrontabili (è più o meno variabile?)
2. Indici di **Variabilità relativa**, che prescindono dall'unità di misura e consentono confronti

Indici di variabilità assoluta

Campo di Variazione, Differenza Interquartilica, Scarto Quadratico Medio, Varianza (è il quadrato dello s.q.m.), Devianza (è il numeratore della varianza)

Indici di variabilità relativa

Indice di variabilità relativo alla media (esempio Coefficiente di Variazione), Indice di variabilità relativo al massimo (variano tra 0 e 1), Concentrazione

Alcuni Indici di Variabilità assoluta

- **Campo di variazione o Range***: differenza tra il massimo ed il minimo valore nei dati

$$x_{min} = \min(x_1, x_2, \ldots, x_n)$$
$$x_{max} = \max(x_1, x_2, \ldots, x_n)$$

$$\Rightarrow \quad \text{range} = x_{max} - x_{min}$$

Mese	2011	2012	2013
gennaio	30	40	60
febbraio	30	20	10
marzo	30	30	20
totale	***90***	***90***	***90***
Campo variazione	0	20	50

2011→Campo variazione=30-30=0 quindi NON SI HA VARIABILITÀ

2013→Campo variazione=60-10=50

Le vendite hanno presentato più variabilità nel 2013!

*Il range è influenzato solo dai valori estremi dei dati, non tiene in considerazione come sono distribuiti gli altri dati

Alcuni Indici di Variabilità assoluta

Supponiamo di avere un insieme di n dati $x_1, x_2, \dots, x_n$. I seguenti indici di variabilità ci dicono quanto i dati sono dispersi attorno alla media $\bar{x}$

Devianza

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

Varianza

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

Scarto quadratico medio o Deviazione standard

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

Devianza, Varianza e Deviazione standard sono indici non negativi ed assumono valore nullo solo quando tutti i valori sono uguali

Alcuni Indici di Variabilità assoluta

Esempio: Consideriamo i seguenti due insiemi di dati aventi la stessa media

A: 3, 4, 6, 7, 10
B: -20, 5, 15, 24

I dati in A e in B presentano la stessa media pari a 6. Tuttavia, i dati in B risultano molto più dispersi attorno alla media, rispetto ai dati in A. Verifichiamo se questo si riflette sui valori degli indici di variabilità.

Insieme A

$$Dev = (3-6)^2 + (4-6)^2 + (6-6)^2 + (7-6)^2 + (10-6)^2 = 30$$

$$Var = \frac{Dev}{(5-1)} = \frac{30}{4} = 7.5$$

$$sqm = \sqrt{Var} = \sqrt{7,5} = 2,74$$

Insieme B

$$Dev = (-20-6)^2 + (5-6)^2 + (15-6)^2 + (24-6)^2 = 1082$$

$$Var = \frac{Dev}{(4-1)} = \frac{1082}{3} = 360,67$$

$$sqm = \sqrt{Var} = \sqrt{360,67} = 18,99$$

Un esempio di Indice di Variabilità relativo: Coefficiente di variazione

Dato un insieme di *n* dati $x_1, x_2, \dots, x_n$ con media $\bar{x} \neq 0$ e deviazione standard *s*, si dice Coefficiente di variazione, *CV*, la seguente quantità

$$CV = \frac{s}{|\bar{x}|}$$

- Il coefficiente di variazione è spesso espresso in termini percentuali

$$CV\% = \frac{s}{|\bar{x}|} \times 100$$

- Il coefficiente di variazione permette di quantificare quanto è grande la deviazione standard rispetto alla media

Un esempio di Indice di Variabilità relativo: Coefficiente di variazione

Esempio.
Consideriamo 3 neonati con peso 3, 4 e 5 Kg (media: 4 Kg, deviazione standard: 1 Kg) e 3 bambini con peso 10, 11 e 12 Kg (media: 11 Kg, deviazione standard: 1 Kg)

DOMANDA: E' maggiore la variabilità del peso dei neonati o quella dei bambini?

La deviazione standard è la stessa nei due casi, pari a 1 Kg.
Tuttavia, il gruppo dei neonati presenta **maggiore variabilità del peso** in rapporto alla sua media.

Questo risulta evidente calcolando il Coefficiente di variazione.
Neonati: CV = 1/4 = 0.25 → 25%
Bambini: CV = 1/11 = 0.09 → 9%

Percentili

Sia *k* un numero intero tra 0 e 100 inclusi. Il ***k*-esimo percentile** di un insieme di dati è quel **valore che risulta contemporaneamente maggiore di una percentuale k dei dati e minore di una percentuale 100-k dei dati**

- Più nello specifico, se esiste solo un dato che soddisfa la condizione precedente, il *k*-esimo percentile è il valore che assume quel dato; se invece ci sono 2 dati che soddisfano la condizione precedente, il *k*-esimo percentile è la media aritmetica di questi due valori
- Il 25-esimo percentile di un campione di dati viene detto primo quartile (**Q1**)
- Il 50-esimo percentile di un campione di dati è la sua mediana campionaria e viene anche detto secondo quartile (**Q2**)
- Il 75-esimo percentile di un campione di dati viene detto terzo quartile (**Q3**)

I quartili dividono i dati in quattro parti, ciascuna contenente circa il 25% dei dati

- La differenza tra il terzo e il primo quartile viene detta **range interquartile (IQR)**
- A volte si indica come range interquartile l'**intervallo [Q1, Q3]**

Esempio calcolo percentili

Consideriamo l'insieme di 10 dati

$$3, 5, \boxed{6}, 11, 15, 16, 18, 20, 21, 25$$

Troviamo il **25-esimo percentile**

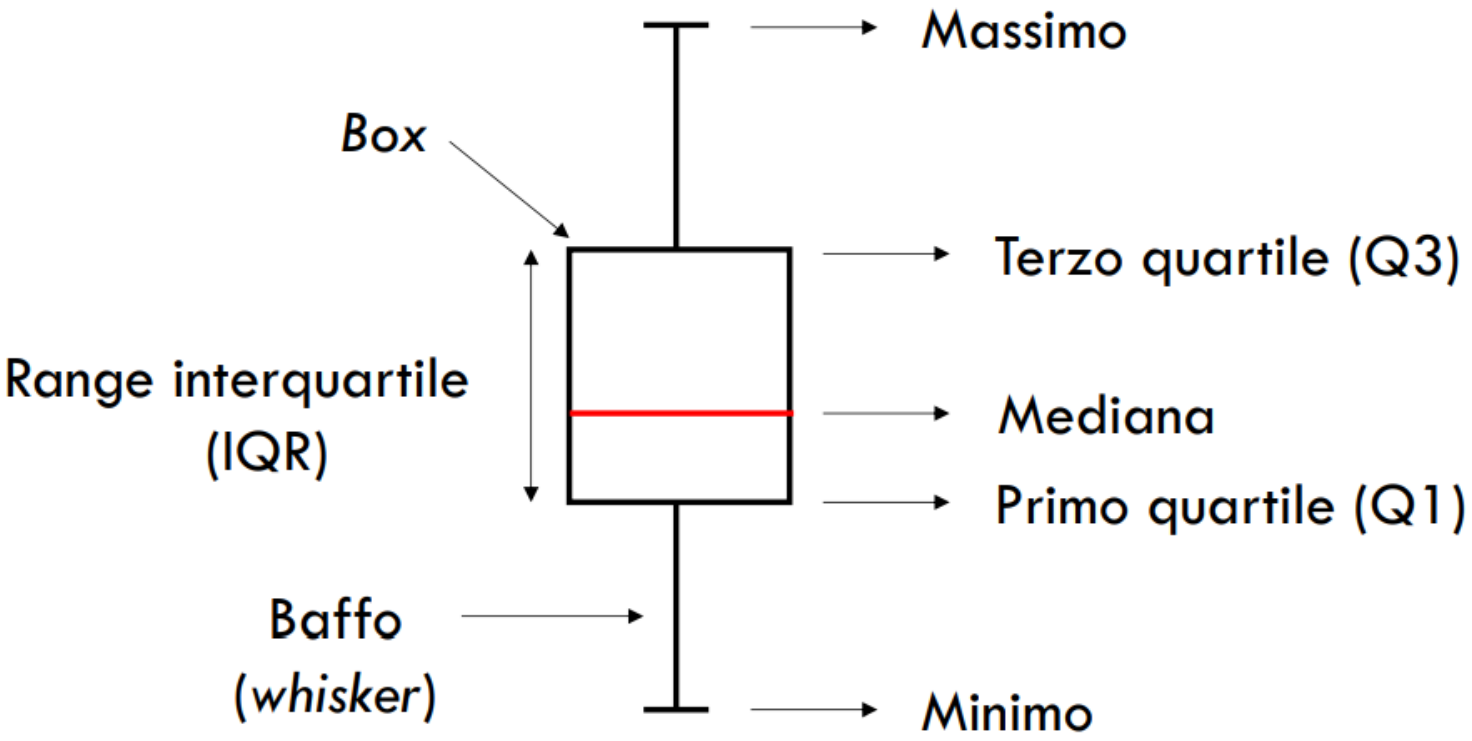
$p=0.25$, $n \cdot p = 2.5 \longrightarrow$ il 25-esimo percentile è il dato in posizione 3 $\longrightarrow$ 6

Troviamo il **40-esimo percentile**

$p=0.40$, $n \cdot p = 4 \longrightarrow$ il 40-esimo percentile è la media tra i dati in posizione 4 e 5 $\longrightarrow (11+15)/2=13$

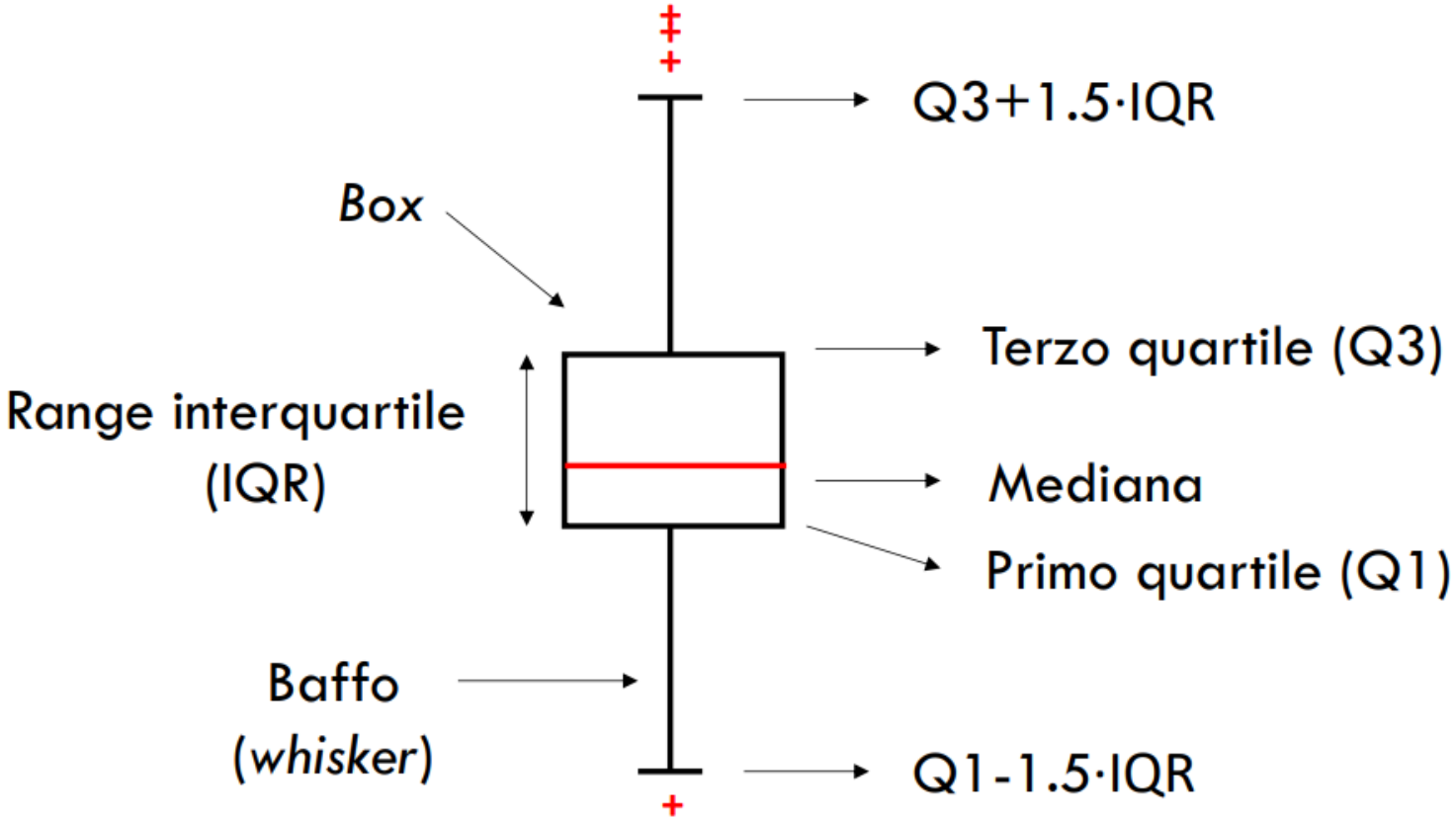
Boxplot (1/2)

Il boxplot è un tipo di grafico che consente di visualizzare alcune delle statistiche rappresentative della distribuzione di un insieme di dati



Boxplot (2/2)

Eventuali outlier, ovvero valori estremi nella distribuzione dei dati, vengono rappresentati oltre il limite dei baffi



Analisi bivariata

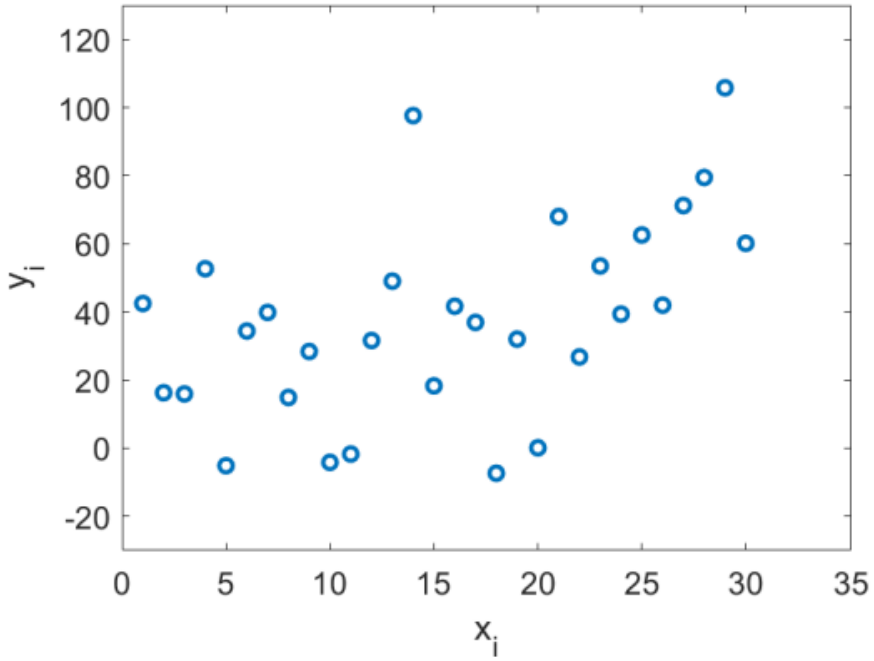
Spesso vengono analizzati coppie di dati tra i quali esiste una qualche relazione

- **Esempio**: vogliamo analizzare **congiuntamente** la temperatura ambientale e il numero di accessi al pronto soccorso

Esiste una relazione tra i due caratteri? Ovvero al crescere dei valori di un carattere aumentano anche i valori dell'altro o viceversa?

Analisi bivariata

- X= «temperatura ambientale», Y=«accessi al pronto soccorso»
- Si rappresentano le coppie $(x_i, y_i) \quad i = 1, \dots, n$
- Diagramma di dispersione (**Scatterplot**)
- La **Correlazione** studia l'associazione lineare esistente tra due variabili
- La misura della forza della associazione tra due variabili è il **Coefficiente di correlazione lineare**



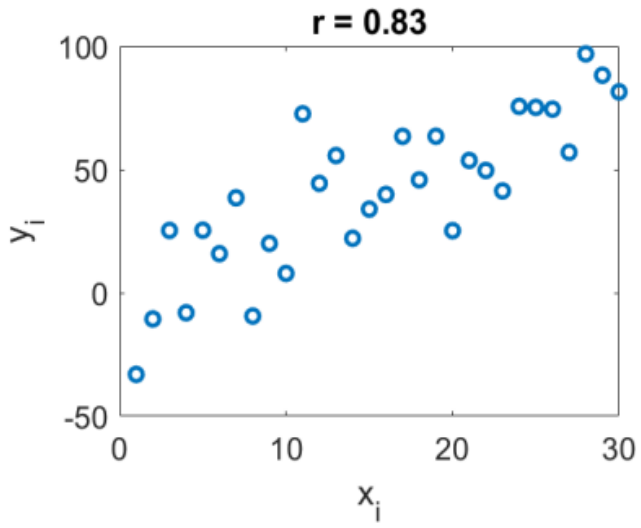
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

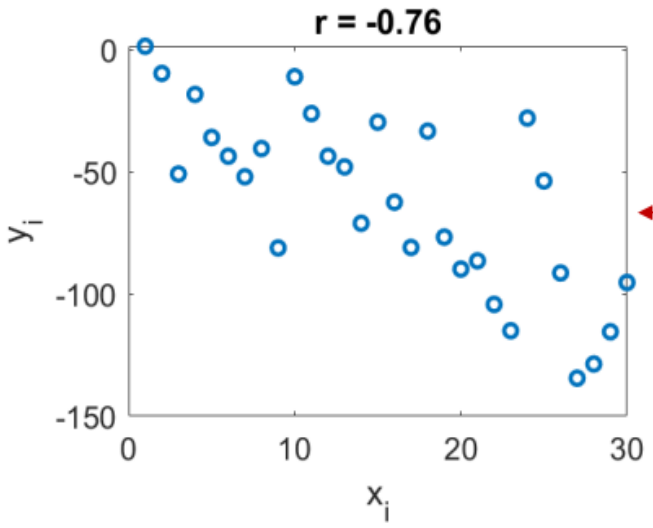
Analisi bivariata

- $-1 \leq r \leq 1$
- Se $r > 0 \rightarrow$ i dati x_i e y_i sono correlati positivamente, ovvero al crescere (decrescere) di x_i , cresce (decresce) y_i (e viceversa)
- Se $r < 0 \rightarrow$ i dati xi e y_i sono correlati negativamente, ovvero al crescere (decrescere) di x_i , decresce (cresce) y_i (e viceversa)
- Se $r = 0 \rightarrow$ i dati non sono correlati, ovvero al crescere (decrescere) di x_i , y_i varia in modo casuale (e viceversa)

Strumenti statistici per l'analisi dei dati

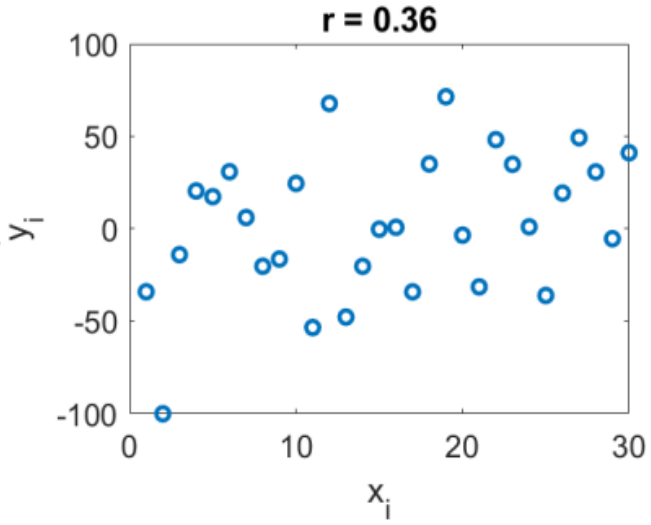
Forte correlazione positiva

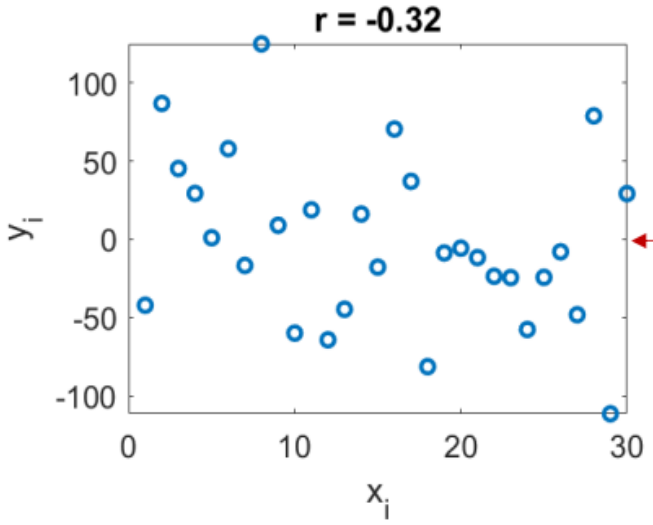




Forte correlazione negativa

Debole correlazione positiva





Debole correlazione negativa

…e adesso…
buon lavoro!