

Metodi Statistici per l'Ingegneria anno 2017/2018

Docente: Giacomo Dimarco

**Dipartimento di Matematica e Informatica
Università di Ferrara**



VNIVERSITÀ
DEGLI·STVDI
DI·FERRARA

[https://sites.google.com/a/unife.it/giacomo-dimarco-home-page/
giacomo.dimarco@unife.it](https://sites.google.com/a/unife.it/giacomo-dimarco-home-page/giacomo.dimarco@unife.it)

Corso di Laurea in Ingegneria Elettronica ed Informatica

Programma del corso

Obiettivi Formativi

- Fornire i mezzi per comprendere le nozioni di probabilità e statistica e come queste si possano applicare per lo studio di vari fenomeni aleatori in diversi ambiti scientifici.
- La teoria della probabilità è propedeutica allo studio dei segnali e comunicazioni, della matematica discreta, dei linguaggi di descrizione dell'hardware dei sistemi di controllo digitale e dell'ingegneria dei sistemi web.
- Capacità di comprensione di argomenti di probabilità e statistica.
- Capacità di ragionamento induttivo e deduttivo nell'affrontare problemi coinvolgenti fenomeni casuali.
- Capacità di schematizzare un fenomeno aleatorio, di impostare un problema e di risolverlo.
- Autonomia di giudizio e di ragionamento critico.
- Capacità di esporre argomenti di natura probabilistico-statistica.

Programma del corso

- **Programma del corso:**

- Il corso ha una durata di 60 ore divise in lezioni frontali ed esercitazioni.
- **Statistica descrittiva.** 4 ore.
- **Elementi di Calcolo Combinatorio.** 6 ore.
- **Elementi di Probabilità.** 8 ore.
- **Variabili aleatorie discrete.** 8 ore.
- **Variabili aleatorie continue.** 8 ore.
- **Leggi congiunte di variabili aleatorie.** 8 ore.
- **Proprietà del valore atteso.** 6 ore.
- **Teoremi limite.** 6 ore.
- **Teoria della stima.** 4 ore.

- **Metodi didattici** Lezioni teoriche ed esercitazioni. Le esercitazioni sono svolte in concertazione con gli studenti. In particolare si richiede agli studenti frequentanti di proporre soluzioni agli esercizi proposti.
- **Orario di ricevimento:** Venerdì 9.30-11.30. Dipartimento di Matematica, Via Machiavelli 30.

Informazioni

- Modalità di verifica dell'apprendimento:

- Scritto

- Consiste nella risoluzione di alcuni esercizi del tipo di quelli svolti in aula o proposti come esempi e di alcune domande di teoria.
- È vietato l'uso di appunti e la consultazione del testo.
- L'esame si intende superato con voto uguale o superiore a 18. In questo caso l'esame orale è facoltativo. Se si ripete lo scritto e si consegna il compito, il voto precedente viene annullato.
- Si viene ammessi all'orale con un punteggio minimo di 15.

- Orale

- La prova orale è facoltativa per chi ha preso almeno 18 allo scritto.
- L'orale in genere modifica il voto dello scritto in un intervallo di $[-3, +3]$.
- In caso di non superamento della prova orale occorre ripetere lo scritto.
- Ci sono 6 appelli in un anno. Tre appelli nella sessione estiva, due appelli nella sessione invernale, un appello nel mese di Settembre.

- Testi di riferimento

- 1) S. M. Ross, Calcolo delle Probabilità, Apogeo 2014.
- 2) S. M. Ross, Probabilità e statistica per l'ingegneria e le scienze, Apogeo 2003.
- 3) R. Spiegel, Probabilità e statistica: 760 problemi risolti, collana Schaum teoria e problemi, ETAS libri.

Introduzione alla Statistica

- La **raccolta** dei dati e la loro **analisi** sono strumenti indispensabili per la comprensione della realtà.
- La **Statistica** si può definire come l'arte di apprendere dai dati.
- In particolare la statistica si occupa di
 - **Raccolta** dei dati.
 - **Descrizione** dei dati.
 - **Analisi** dei dati.
- L'**analisi** dei dati ci guida naturalmente sotto certe ipotesi verso delle **conclusioni**.

Raccolta e Analisi dei dati

- L'analisi statistica può partire da un **campione di dati** definito.
Esempi: quantità di precipitazioni, scosse di terremoto, tasso di disoccupazione, tasso di inflazione, etc.
- In altri casi il lavoro dello statistico parte **prima della raccolta**. Consiste quindi nell'**ideare un procedimento** per la raccolta.
- **Esempio**: Quale metodo di insegnamento è il più efficace ?
 - **Soluzione** : dividere la classe in due sottogruppi, usare un diverso metodo didattico e confrontare i punteggi ottenuti.
 - **Domanda**: Qualè il miglior modo di dividere il gruppo?
 - **Risposta**: la miglior soluzione è scegliere "**a caso**".
- Alla **fine dell'esperimento** i dati vengono raccolti ed analizzati.
- Quella parte della statistica che si occupa di illustrare e sintetizzare i dati è detta **Statistica Descrittiva**.

Inferenza Statistica, campioni e popolazione I

- Continuando con l'esempio precedente, dopo che i dati sono stati illustrati e sintetizzati, vorremmo poter trarre una **conclusione** su quale dei due metodi è più efficace.
- La parte della statistica che si occupa di questo aspetto è detta **Inferenza Statistica**.
- Per trarre conclusioni è
 - necessario **comprendere l'influenza del caso**.
 - necessario fare alcune **assunzioni** sulla probabilità che i dati assumano diversi valori.
 - L'insieme di queste ipotesi è detto **modello probabilistico**.
- **Esempio**: vogliamo scoprire la frazione di circuiti integrati difettosi dovuta a un nuovo metodo produttivo.
 - **Soluzione** : selezionare un certo numero di chip e testarli. Il numero di difettosi costituisce il dato sperimentale.
 - Se la scelta è fatta a caso è **ragionevole supporre** che ciascuno dei chip sia difettoso con probabilità $p = \text{frazione chip difettosi} / \text{numero di chip totali}$.

Inferenza Statistica, campioni e popolazione II

- **Problema:** Ci sono altre situazioni in cui il modello probabilistico appropriato per un certo campione di dati non è facile da scegliere.
- **Soluzione:** fare un ipotesi sul modello e fare altri esperimenti per validarlo.
- **Riassumendo:** L'inferenza statistica si basa sul presupposto che i fenomeni possano essere descritti in termini di probabilità ed utilizza i dati per fare inferenze su queste probabilità.
- La statistica è interessata ad ottenere informazioni su un **insieme completo detto popolazione**.
- Poichè spesso il numero di dati corrispondente a l'intera popolazione è troppo grande si cerca di **apprendere qualcosa analizzando dei sottogruppi detti campioni**. Il campione deve essere rappresentativo.
- **Esempio:** vogliamo conoscere la distribuzione di età degli abitanti di un comune.
- **Soluzione:** Chiedere alle prime **100** persone che entrano nella biblioteca è una buona strategia?
- I **campioni scelti a caso** sono normalmente più rappresentativi.

Breve storia della statistica I

- La **raccolta sistematica di dati** è cominciata a Venezia e Firenze durante il Rinascimento.
- Il **termine statistica** deriva dalla parola stato in quanto indicava una raccolta di dati fatta nell'interesse dello stato.
- L'idea si diffuse rapidamente in Europa e alla metà del XVI secolo era **consuetudine registrare nascite, decessi e matrimoni**.
- Nel **1562** a Londra si cominciò a **pubblicare i bollettini** di mortalità dovuti quasi interamente alla peste.
- Nel **1662 John Graunt (1620 – 1674)** pubblicò un libro dal titolo "Natural and Political Observation made upon the bills of mortality".
- Graunt pensò di **utilizzare i bollettini per stimare** la popolazione di Londra.

Breve storia della statistica II

- Questi dati furono utilizzati per **stabilire** quante persone potevano essere **coscritti** e il **numero di persone da tassare**.
- Graunt stabilì anche delle **stime di speranze di vita** utilizzate per le rendite vitalizie.
- **Halley (1656 – 1742)** usò le tabelle di Graunt per stabilire con che probabilità una persona sarebbe vissuta fino a una certa età.
- **Dopo** Graunt e Halley la **raccolta dati si diffuse** per tutto il XVII e XVIII.
- Il **termine statistica** che era usato come scienza descrittiva dello stato iniziò ad essere **associata ai numeri** nel **1800**.
- All'inizio del XIX secolo **l'uso della teoria della probabilità** (sviluppata da matematici come Bernoulli, Gauss, Laplace) per studiare fenomeni statistici era quasi **inesistente**. Non c'era interesse di fare inferenza sui singoli ma a studiare le intere popolazioni.

Breve storia della statistica III

- Negli ultimi anni dell' 800 la statistica iniziò ad occuparsi di inferire a partire da dati numerici. Fra i fautori ricordiamo Galton (1822 – 1891, cugino di Darwin) che si occupò di studi sull'ereditarietà dell'intelligenza e Pearson (1857 – 1936).
- Pearson sviluppò il test del chi-quadro per verificare la bontà di un fit e fu primo direttore dell'istituto Galton fondato nel 1904.
- Uno dei primi studenti dell'istituto Galton fu Gosset, celebre per la teoria del test t .
- I campi di maggiore importanza all'inizio del XX secolo erano la biologia e l'agricoltura per l'interesse di Pearson ed ai risultati di un altro pioniere Fisher.
- La teoria sviluppata era abbastanza generale da applicarsi a diversi problemi quantitativi e pratici. Per questo la statistica ha cominciato ad essere considerata come uno strumento capace di fornire risposte quantitative a problemi pratici.
- Attualmente gli ambiti della statistica sono ovunque : ricerca medica, ingegneria, marketing, qualità, economia, i sondaggi, le inchieste, le previsioni meteorologiche.

Organizzazione e descrizione dei dati

- I **risultati** numerici di una ricerca o di un esperimento devono essere presentati in maniera
 - **Chiara**
 - **Concisa**
 - Devono dare un **idea** più rapidamente possibile delle loro caratteristiche globali
- Esistono delle tecniche di rappresentazione tabellari e grafiche accettate universalmente
- Esse hanno il pregio di evidenziare aspetti come
 - **Supporto**
 - **Simmnetria**
 - **Grado di concentrazione dei dati**

Tabelle e grafici di frequenze I

Se i dati sono in **numero relativamente basso** possono essere rappresentati in una tabella tramite la loro frequenza

Stipendio Iniziale	Frequenza
27	4
28	1
29	3
30	5
31	8
32	10
34	5
36	2
37	3
40	1

Table: Redditi annuali iniziali Ingegneri Elettrici neolaureati. Dati in migliaia di dollari.

Tabelle e grafici di frequenze II

Per **rappresentare graficamente** la distribuzione di frequenze di un insieme di questo tipo si possono usare, fra le varie scelte, il cosiddetto **line graph**

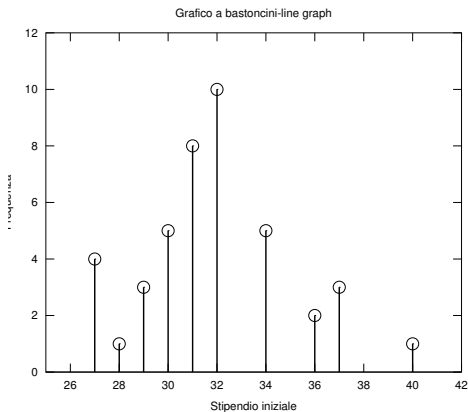


Tabelle e grafici di frequenze III

Sostituendo la linea con un rettangolo otteniamo il **bar graph**, mentre nel caso in cui i punti del grafico siano uniti da una spezzata abbiamo il **polygon graph**. In tutti i casi sull'asse delle **ascisse** sono indicati i dati, sulle **ordinate** la frequenza di ogni valore

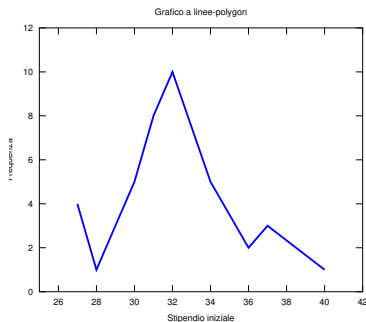
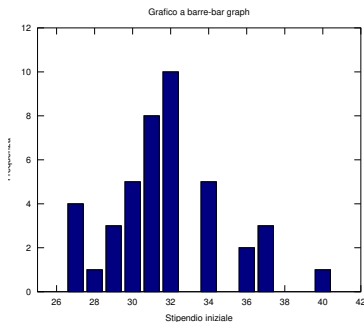


Tabelle e grafici di frequenze relative

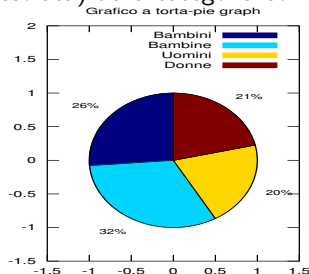
Considerando un insieme di n dati numerici, se f è la frequenza di uno dei valori allora il rapporto f/n si chiama **frequenza relativa**. Nella rappresentazione ci sarà un fattore di scala $1/n$.

Stipendio Iniziale	Frequenza relativa
27	$4/42 \simeq 0.0952 = 9.52\%$
28	$1/42 \simeq 0.0238 = 2.38\%$
29	$3/42 \simeq 0.0714 = 7.14\%$
30	$5/42 \simeq 0.1190 = 11.90\%$
31	$8/42 \simeq 0.1905 = 19.05\%$
32	$10/42 \simeq 0.2381 = 23.81\%$
34	$5/42 \simeq 0.1190 = 11.90\%$
36	$2/42 \simeq 0.0476 = 4.76\%$
37	$3/42 \simeq 0.0714 = 7.14\%$
40	$1/42 \simeq 0.0238 = 2.38\%$

Table: Redditi annuali iniziali. Dati in migliaia di dollari.

Tabelle e grafici di frequenze relative

Un altro tipo di rappresentazione grafica è il **grafico a torta** utile in particolare quando i dati sono categorici e non numerici. L'**angolo al centro** è proporzionale alla frequenza (relativa o assoluta) della categoria corrispondente.



Tipo	numero	Frequenza relativa
Bambini	123	26.82%
Bambine	150	33.20%
Uomini	95	21.37%
Donne	100	20.3%

Table: Clienti di una gelateria.

Raggruppamento dei dati

- Abbiamo visto come rappresentare dei dati che abbiano un **numero di valori distinti basso**.
- Quando il campione di dati non soddisfa questo requisito normalmente si **dividono i campioni in classi**.
- Si **rappresentano** poi con grafici e tabelle il numero di dati che **cadono in un certo intervallo**.
- La **scelta del numero di classi** è importante e dipende dal problema. Se si prendono **poche** classi si perdono informazioni mentre se se ne prendono **troppe** diventa difficile riconoscere la forma della distribuzione.
- I **bordi della classe** sono gli estremi dei suoi intervalli si adotta la convenzione di includere i bordi di sinistra e escludere quelli di destra.
- **Esempio**: con intervallo 20-30 si intendono tutti i valori x tali che $20 \leq x < 30$.

Raggruppamento dei dati: frequenze assolute per classi di valori

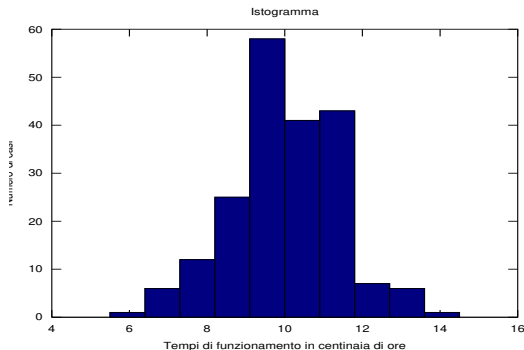
Riportiamo a [titolo di esempio](#) i tempi di funzionamento in ore di 200 lampadine ad incandescenza sintetizzati in 10 intervalli di classe di lunghezza 100 con il primo che comincia a 500.

Intervallo di classe	Frequenza
500-600	2
600-700	5
700-800	12
800-900	25
900-1000	58
1000-1100	41
1100-1200	43
1200-1300	7
1300-1400	6
1400-1500	1

Table: Tempo di vita in ore di 200 lampadine ad incandescenza

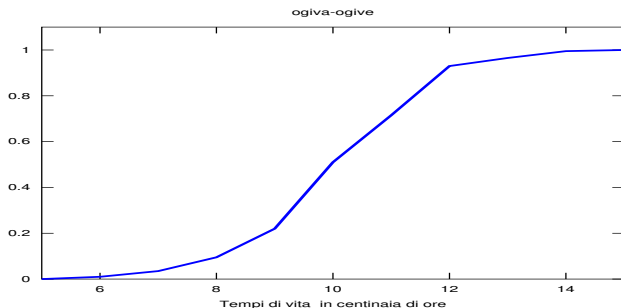
Raggruppamento dei dati: Istogrammi

Il **grafico a barre** delle frequenze, o delle frequenze relative, delle classi prende il nome di **istogramma**, o **istogramma delle frequenze relative**.



Raggruppamento dei dati: frequenze cumulative

Per **grafico delle frequenze cumulative** si intende una curva sul piano cartesiano per cui le ascisse rappresentano i possibili valori dei dati e le ordinate indicano il numero o la frazione di dati che sono minori o uguali ai valori in ascissa. Questo rappresentazione è **detta ogiva o ogive**.



Raggruppamento dei dati: diagramma stem and leaf

Una maniera efficiente di organizzare un **numero non troppo grande** di dati è il diagramma **stem and leaf** (**ramo-foglia**). Per costruirlo occorre dividere le cifre di ogni dato numerico in **due parti**, una **più significativa** (lo stem) e una **meno significativa** (la leaf). L'esempio riportato nella tabella fornisce le medie annuali delle temperature minime giornaliere in **35** città americane.

[illegible]

Table: temperature medie minime giornaliere in 35 città americane