



# UNIVERSITÀ DI PAVIA

Anno Accademico 2021/2022

## ELABORAZIONE DI DATI BIOMEDICI

<b>Anno immatricolazione</b>	2020/2021
<b>Anno offerta</b>	2021/2022
<b>Normativa</b>	DM270
<b>SSD</b>	ING-INF/06 (BIOINGEGNERIA ELETTRONICA E INFORMATICA)
<b>Dipartimento</b>	DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE
<b>Corso di studio</b>	BIOINGEGNERIA
<b>Curriculum</b>	PERCORSO COMUNE
<b>Anno di corso</b>	2°
<b>Periodo didattico</b>	Secondo Semestre (07/03/2022 - 17/06/2022)
<b>Crediti</b>	6
<b>Ore</b>	72 ore di attività frontale
<b>Lingua insegnamento</b>	Italiano
<b>Tipo esame</b>	SCRITTO E ORALE CONGIUNTI
<b>Docente</b>	MAGNI PAOLO (titolare) - 6 CFU
<b>Prerequisiti</b>	Il linguaggio usato in statistica è prevalentemente matematico. Occorrono alcune delle nozioni dei corsi di Analisi Matematica e Geometria e Algebra. In particolare saranno utili le nozioni di limite, di integrale e di derivata, di serie, di funzione di più variabili e di funzioni vettoriali, di massimizzazione/minimizzazione di funzione di una o più variabili oltre che la teoria degli insiemi ed elementi di logica.
<b>Obiettivi formativi</b>	Il corso si propone di fornire allo studente le competenze necessarie per padroneggiare i metodi di analisi statistica e probabilistica più usati nella letteratura medico scientifica. Tali metodi giocano un ruolo importante in settori di ricerca quali la bioinformatica, lo studio del genoma e delle reti metaboliche cellulari, la messa a punto di nuovi farmaci e la valutazione del loro effetto, l'individuazione di geni responsabili di malattie, lo studio della diffusione di epidemie, la medicina predittiva e così via. Verranno pertanto forniti gli strumenti per la corretta progettazione di uno studio e

per l'analisi efficace dei risultati. A tal fine, il corso fornisce dapprima gli strumenti di base di probabilità e statistica per poi addentrarsi in tecniche di analisi più complesse quali i molteplici test statistici proposti in letteratura per i diversi tipi di variabile e (da valutare anno per anno) regressione lineare semplice e multipla. Gli esempi che verranno illustrati durante il corso e gli esercizi proposti saranno prevalentemente di carattere biomedico. È opportuno sottolineare che la competenza nell'analisi statistica dei dati è un requisito sempre più importante in numerosi tipi di carriera, non solo in ambito biomedico, ma anche in altre aree del settore industriale e finanziario.

#### Programma e contenuti

- Introduzione alla biostatistica: cos'è?

- Statistica descrittiva

Vengono illustrate le principali tecniche con cui si possono estrarre informazioni di sintesi a partire da dati sperimentali

Tipi di dati: variabili qualitative/quantitative. Tipi di scale di misura: nominale/ordinale/ad intervalli/di rapporti. Matrice dei dati.

Strumenti di sintesi: distribuzione (tabelle) di frequenza per dati raggruppati e creazione delle classi.

Sintesi quantitativa: misure di tendenze centrale (media aritmetica/pesata/geometrica/armonica/quadratica, mediana, moda, intervallo medio, media interquartile), quantili (quartili/decili/percentili,frattile), misure di dispersione o variabilità (campo o intervallo di variazione/differenza interquartile/scarti della media/scarto medio assoluto/devianza o somma dei quadrati/varianza o quadrato medio/deviazione standard o scarto quadratico medio/coefficiente di variazione), Disuguaglianza di Markov, di Chebychev e di Cramer, momenti di ordine superiore, indici di forma (simmetria: skewness di Pearson, Gamma1 di Fisher, Beta1 di Pearson; curtosi: mesocurtica/leptocurtica/platicurtica, Gamma2 di Fisher, Beta2 di Pearson).

Sintesi qualitativa (grafici): istogrammi o poligoni/distribuzioni cumulate, diagrammi a rettangoli, ortogrammi, aerogrammi, pittogrammi, diagrammi polari, dotplot, boxplot, diagrammi di dispersione a due variabili, diagrammi cartesiani a due variabili).

- Gli studi statistici

Vengono illustrate le principali caratteristiche degli studi condotti in ambito biomedico.

Scopo di uno studio.

Progetto di uno studio. Campionamento: metodi probabilistici e non; campione di convenienza, a valanga, casuale semplice, pesato, sistematico, stratificato, a grappolo. Campioni a due o più stadi.

Epidemiologia: misure e indici specifici (prevalenza, incidenza, morbilità, morbosità, morbidità, letalità, mortalità, rischio relativo, riduzione del rischio assoluto, riduzione del rischio relativo), tassi grezzi, specifici e standardizzazione, rapporto tra proporzioni, rapporto tra odds. Tipi di studi: osservazionali (descrittivi/analitici - ecologici, trasversali, retrospettivi, prospettici longitudinali), sperimentali (trial clinici, sul campo, di popolazione). Studi clinici nelle diverse fasi di sviluppo di un farmaco.

Accuratezza, precisione e numero di cifre significative nella raccolta dati.

#### - Statistica matematica: elementi di probabilità

Vengono introdotti i concetti elementari della teoria della probabilità, il teorema di Bayes, e le più importanti classi di distribuzioni di probabilità. Eventi e spazio campionario, combinazione di eventi, calcolo combinatorio di raggruppamenti semplici (permutazioni, disposizioni, combinazioni).

Definizione di probabilità matematica o classica, frequentista e soggettiva, vari tipi di convergenza di successioni di variabili aleatorie, assiomi della probabilità, probabilità condizionate e indipendenza condizionale, teorema della probabilità totale e teorema di Bayes e sua applicazione ai test di screening (veri/falsi positivi, veri/falsi negativi, sensibilità, specificità, efficienza, valore predittivo positivo/negativo, curva ROC, calcolo prevalenza con test di screening).

Variabili casuali (discrete/continue), funzione di distribuzione cumulativa, funzione di densità, funzione di probabilità di massa, momenti di variabili casuali.

Variabili casuali congiunte, funzione di distribuzione cumulativa congiunta e di densità congiunta, distribuzione e densità marginale, probabilità di massa congiunte e marginali, distribuzioni condizionate, variabili casuali indipendenti, covarianza, correlazione, funzioni di variabili casuali (distribuzione, media, varianza e propagazione dell'incertezza).

Variabili casuali vettoriali.

Distribuzioni di probabilità di variabili discrete: uniforme, bernoulli, binomiale/multinomiale, Poisson, geometrica e Pascal, binomiale negativa, ipergeometrica.

Distribuzioni di probabilità di variabili continue: rettangolare, normale o gaussiana (approssimazione alla normale e teorema del limite centrale, lognormale, esponenziale (Erlang), gamma, gamma inversa, weibull, beta, dirichlet, chi<sup>2</sup>, t-student, F-fisher.

Quale distribuzione seguono i dati? I grafici di probabilità (qqplot).

Simulazione come strumento per l'investigazione dei dati.

#### - Statistica inferenziale: teoria della stima

Vengono introdotti i concetti basi della teoria della stima.

L'inferenza statistica e le distribuzioni campionarie.

Teoria della stima: stima puntuale e per intervallo, stima alla Fisher, stima bayesiana, stima parametrica e stima non parametrica (es. momenti campionari), stimatore e sue proprietà (polarizzazione, consistenza, efficienza), stimatori lineari, limite di Cramer-Rao e informazione di Fisher anche nel caso vettoriale (matrice di covarianza della stima), metodi per la costruzione di stimatori (metodo dei momenti, stima a massima verosimiglianza e sue proprietà, stima bayesiana, stimatori puntuali e distribuzioni coniugate), intervalli di confidenza.

Stima dei parametri di distribuzioni note: binomiale e proporzioni, Poisson e tassi, normale, esponenziale. Proprietà di questi stimatori.

Distribuzione campionarie e intervalli di confidenza dei conteggi di frequenza (proporzioni), della media, di differenza di medie, varianza e del rapporto di varianza.

Intervalli di confidenza, numerosità del campione e livello fiduciario.

Valutazione delle distribuzioni campionarie e degli intervalli di confidenza attraverso la simulazione.

- Statistica inferenziale: i test statistici

Vengono presentati i concetti alla base dei test statistici e presentati i principali test parametrici e non parametrici.

Definizione di un test (statistica del test e distribuzione della statistica del test) e relazione con gli intervalli di confidenza, ipotesi nulla (bilaterale/unilaterale) e ipotesi alternativa e regola di rifiuto (alfa), p-value, test parametrici e non parametrici, errore di tipo I e tipo II e protezione, potenza e significatività, fattori che incidono sulla potenza (alfa, delta,  $\sigma^2$ , n) e loro relazioni nella distribuzione z, potenza a priori (n) e a posteriori (beta).

Criteri che guidano nella scelta del test (tipo dati, scala di misura, simmetria/normalità della distribuzione, omoschedasticità dei diversi campioni. Confronto tra test: il rapporto potenza-efficienza.

[PROSEGUE IN "ALTRE INFORMAZIONI"]

#### Metodi didattici

Lezioni (ore/anno in aula): 22. Verranno presentati dal docente i principali concetti metodologici.

Esercitazioni (ore/anno in aula): 26. Verrà illustrata dal docente, anche con la collaborazione degli studenti, l'applicazione delle metodologie introdotte per la risoluzione di specifici problemi e casi di studio.

Attività pratiche (ore/anno in aula): 26. Gli studenti dovranno affrontare individualmente o in piccoli gruppi sotto la guida del docente e di tutor, ove disponibili, esercizi di simulazione ed analisi di dati reali, attraverso l'uso di pacchetti software e delle metodologie introdotte durante il corso. Attività di tutorato, quando possibile. Svolgimento guidato di ulteriori esercizi.

#### Testi di riferimento

Materiale distribuito dal docente agli iscritti alla mailing list del corso W. Navidi. Probabilità e statistica per l'ingegneria e le scienze.

McGraw-Hill. Libro di riferimento del corso.

Norman e Streiner. Biostatistica, Quello che avreste voluto sapere.

Casa Editrice Ambrosiana, Testo "divertente" di riepilogo.

W. W. Daniel. Biostatistica. EdiSES. Testo di approfondimento.

L. Soliani. Manuale di statistica per la ricerca e la professione.

<http://www.dsa.unipr.it/soliani>. I capitoli 1,2,3,4,5,6,7,8,9,10,11,12,15 sono alcuni degli argomenti del corso.

Laboratorio virtuale di probabilità e statistica.

[http://www.ds.unifi.it/VL/VL\\_IT/index.html](http://www.ds.unifi.it/VL/VL_IT/index.html). Sito con risorse interattive per studenti e docenti di probabilità.

#### Modalità verifica apprendimento

L'esame consiste in una prova scritta e in una prova orale in cui vengono valutate sia la conoscenza dei fondamenti teorici sia la capacità di risolvere esercizi.

E' inoltre possibile, durante il corso, svolgere a casa, in apposite finestre temporali comunicate di volta in volta, esercizi di verifica messi a disposizione sulla piattaforma kiro.

#### Altre informazioni

[PROSEGUE DA "PROGRAMMA E CONTENUTI"]

Variabile effetto misurata almeno su scala intervallare: 1 campione: ipotesi sulla media per popolazione normale o numerosa (test t e z) e calcolo della potenza a priori e a posteriori, ipotesi sulla varianza per

popolazione normale (test  $\chi^2$ ). 2 campioni indipendenti: ipotesi sulla differenza tra due medie per popolazioni normali o numerose (test t e z) e calcolo della potenza a priori e a posteriori, ipotesi sulla varianza di due popolazioni normali (test F). 2 campioni appaiati: ipotesi sulla differenza tra due medie per popolazioni normali o numerose (test t). Ipotesi sull'appartenenza di un'osservazione a un campione normale (test t). Più campioni indipendenti: ipotesi sulla varianza di più popolazioni normali (test Hartley, Cochran, Bartlett, Levene), ipotesi sulle medie di più popolazioni normali (test ANOVA una via), confronti multipli pianificati ortogonali e metodo dei polinomi ortogonali o post-hoc e correzione per confronti multipli (Bonferroni, Scheffé, LSD, HSD, Dunnett). Più campioni dipendenti: ipotesi sul confronto tra le medie (test ANOVA per misure ripetute). Più campioni indipendenti classificati secondo due fattori senza interazione (test ANOVA a due vie e quadrati latini), classificati secondo più fattori senza interazione (test ANOVA a più vie, quadrati greco-latini), classificati secondo più fattori con interazione (test ANOVA per esperimenti fattoriali). Quanti fattori considerare? L'efficienza relativa. Valutazione dell'effetto del trattamento tramite  $R^2$  e  $\eta^2$ .

Variabile effetto misurata su scala nominale: 1 campione: ipotesi su una proporzione (test z, binomiale), ipotesi sulla distribuzione e test di bontà di adattamento (test  $\chi^2$ , test G, test T2 di Freeman-Tukey). 2 campioni indipendenti: studio di fattori di rischio e tabelle di contingenza, test sulla differenza di due proporzioni (test z) e tabelle 2x2 (test  $\chi^2$ , test G), test esatto di Fisher, potenza a priori e posteriori, rischio relativo (test z e formula di Miettinen), odds ratio (test z e formula di Miettinen, test  $\chi^2$  di Mantel-Haenszel), rapporto di tassi (test z e formula di Miettinen). Test di indipendenza e di omogeneità e associazione tra variabili (coefficiente di contingenza di Pearson e  $\phi_{ch}$  di Cramer). 2 campioni dipendenti: test McNemar (variabili dicotomiche), estensione test McNemar o test di Bowker (variabili politomiche). Più campioni indipendenti: tabelle 2xN e MxN (test  $\chi^2$ , test G, metodo esatto). Più campioni dipendenti: test Q di Cochran.

Variabile effetto misurata su scala ordinale: 1 campione: ipotesi sulla casualità di un campione temporale o spaziale (test delle successioni), ipotesi sulla tendenza centrale (test del segno, test di Wilcoxon o dei ranghi con segno, test di casualizzazione), ipotesi sull'omogeneità di conteggi (test di Poisson e indice di dispersione), bontà di adattamento (test di Kolmogorov-Smirnov). 2 campioni dipendenti: ipotesi sulla tendenza centrale (test dei segni, test di Wilcoxon, test di casualizzazione). 2 campioni indipendenti: ipotesi sull'effetto ordine (test di Gart), ipotesi sulla tendenza centrale (test della mediana, test di Wilcoxon-Mann-Whitney, test U Mann-Whitney, test S di Kendall, test di casualizzazione), aderenza di due distribuzioni (test successioni o test di Wald-Wolfowitz, test di Kolmogorov-Smirnov), ipotesi sulle varianze (test di Siegel-Tukey). Più campioni: ipotesi sulla tendenza centrale (test della mediana, Kruskal-Wallis), ipotesi sulla varianza. Più campioni indipendenti classificati secondo due fattori (analisi della varianza per ranghi a due vie di Friedman), confronti multipli.

Correlazione e regressione lineare (valutato di anno in anno se svolgere questa parte)

Regressione semplice e multipla.

**Obiettivi Agenda 2030 per lo  
sviluppo sostenibile**

[\\$bl legenda sviluppo sostenibile](#)