



UMG
Dubium sapientae initium

appunti di biostatistica

per la Scuola di Specialità in Chirurgia Orale

massimo.borelli@unicz.it

Copyright © 2024 Massimo Borelli

UNIVERSITÀ *Magna Græcia* DI CATANZARO

[GITHUB.COM/UMGSTAT/MAXILLO](https://github.com/UMGSTAT/MAXILLO)

Licensed under the Creative Commons Attribution-NonCommercial 4.0 License (the “License”). You may not use this file except in compliance with the License. You may obtain a copy of the License at <https://creativecommons.org/licenses/by-nc-sa/4.0>. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

Prima edizione, Maggio 2024

Contents

1	Per iniziare	5
1.1	Di cosa parleremo in questo corso	5
1.2	Libri di testo e collane di articoli	5
1.2.1	Libri di testo	5
1.2.2	Collane di articoli	5
1.3	Quali software useremo	8
2	Descrivere i dati	9
2.1	La basi della statistica descrittiva	9
2.2	Riconoscere i fenomeni aleatori	11
2.2.1	La distribuzione normale, o gaussiana	11
2.2.2	Altre importanti distribuzioni statistiche	11
2.3	Frequenze, rischi e predittività	11
2.4	Descrivere l'affidabilità: le basi dell'inferenza statistica	12
3	Inferenza statistica univariata	13
3.1	L'esempio storico del T-test	13
3.2	Testare differenze tra due gruppi	13
3.3	L'approccio bayesiano al T-test	14
3.4	Testare differenze tra più gruppi: l'Anova	14
3.5	Altri possibili approcci: permutazione e ricampionamento	14
3.6	La retta di regressione	15
4	Inferenza statistica univariata	17
4.1	L'Ancova ed i modelli lineari	17
4.2	I modelli lineari generalizzati	17
	Bibliography	19
	Articles	19

	Books	19
	Index	21
	Appendices	23
A	Appendix Chapter Title	23
A.1	Appendix Section Title	23
B	Appendix Chapter Title	25
B.1	Appendix Section Title	25

1. Per iniziare

1.1 Di cosa parleremo in questo corso

Presentazione 1.1 <https://bit.ly/43XbcmD>

Per iniziare il corso, passiamo in rassegna alcune immagini tratte dalla rete, e riflettiamo se quando usiamo il simbolo \pm di 'tolleranza' di una posizione o di un angolo, nel senso della gnatologia, questo abbia un preciso senso statistico, legato per esempio al concetto di 'deviazione standard', SD, e se si possa sovrapporre correttamente il significato di tolleranza con quello di incertezza o di variabilità. Osserviamo anche come spesso in letteratura le informazioni statistiche vengano accompagnate con dei giudizi probabilistici ('p-Value'), molto spesso evidenziati con un asterisco in vari tipi di grafici, apparentemente molto eterogenei tra loro.

1.2 Libri di testo e collane di articoli

1.2.1 Libri di testo

Presentazione 1.2 <https://github.com/umgstat/maxillo>

Abbiamo messo a disposizione del nostro spazio Google Drive alcuni libri da consultare:

- <https://drive.google.com/drive/folders/1daqfCDXMuqXmyIeyJP5gbCpANgJ4IrIC>
J. S. Kim, R. J. Dailey. *Biostatistics for oral healthcare*. Si tratta di un libro di impostazione classica ed abbastanza approfondito, non mancano gli esercizi e le 'formule matematiche'. Lo teniamo come eventuale testo di consultazione e di riferimento.
- <https://drive.google.com/file/d/1d6QQS6ogIKliHX843SpM0fxEptpgPaYS/view>

1.2.2 Collane di articoli

Presentazione 1.3 <https://go.nature.com/3xvejpu>

Alcune riviste scientifiche, come ad esempio *Nature*, dedicano molta attenzione alla metodologia statistica: attraverso la rubrica denominata *Point of significance*, in una quarantina di articoli di poche pagine si discutono dei singoli temi da un punto di vista 'pratico', evidenziando vantaggi e criticità. Ve ne sono una decina che offrono una conoscenza di base e che vale la pena di sfogliare:

1. **Importance of being uncertain.** L'articolo, di tipo molto introduttivo, si concentra sul ruolo che la statistica ha nel condurre il processo decisionale e nell'interpretare l'incertezza,

sottolineando l'importanza di cosa significhi fare inferenza a livello di popolazioni ed a livello di un campione. Si discute inoltre sul concetto di distribuzione statistica, della media e della deviazione standard come misure che caratterizzano la posizione e la variabilità di una popolazione. L'articolo inoltre enfatizza l'uso del campionamento per stimare parametri della popolazione come la media e la deviazione standard, accennando ai concetti di distribuzione campionaria e del teorema del limite centrale, il quale afferma che la distribuzione delle medie campionarie diventa sempre più vicina a una distribuzione normale all'aumentare della dimensione del campione. Il documento infine evidenzia un'applicazione pratica delle statistiche nel delineare l'incertezza tramite gli intervalli di fiducia ('confidenza') e nella visualizzazione delle barre di errore.

2. **Error Bars.** Qui si affronta la possibile errata interpretazione delle barre di errore e il significato della loro sovrapposizione nell'analisi dei dati, evidenziando i misconcetti ed i fraintendimenti che compaiono nelle pubblicazioni scientifiche: nonostante l'uso comune infatti, in letteratura sono frequenti le interpretazioni errate delle barre di errore e della loro relazione con la significatività statistica. L'articolo sottolinea la necessità di chiarezza nella rappresentazione dell'incertezza, distinguendo tra diversi tipi di barre di errore: la deviazione standard (sd), l'errore standard della media (sem) e l'intervallo di confidenza (CI), sottolineando l'importanza della distinzione tra questi tipi di barre e fornendo esempi di come la loro interpretazione possa variare. Il documento fornisce inoltre criteri pratici per interpretare diversi tipi di barre di errore, in particolare nel contesto della dimensione del campione e del livello di significatività.
3. **Significance, P values and t-tests.** L'articolo spiega il significato pratico del 'valore di probabilità P' ed il modo in cui esso viene interpretato in un contesto biomedico. Dopo aver introdotto il concetto di significatività statistica e di test statistico, viene discusso il cosiddetto 't-test per un campione'. Gli autori descrivono il processo di costruzione della 'distribuzione nulla' per calcolare la probabilità (valore P) che un'osservazione si possa considerare accidentale. L'articolo mette in guardia dall'interpretare erroneamente il valore P come la probabilità che l'ipotesi nulla sia vera. L'articolo discute poi il t-test per un campione, spiegando come esso venga utilizzato per determinare se i campioni provengano da una distribuzione con una media specificata, e per costruire intervalli di fiducia. Inoltre, confronta le distribuzioni t e normale, sottolineando infine che la significatività statistica non implica la significatività biomedica.
4. **Power and sample size.** L'articolo traccia l'importanza del concetto di 'potenza' nei test di verifica delle ipotesi e sottolinea le conseguenze di una potenza 'bassa' negli esperimenti con confronti multipli (come accade in particolare negli studi -omici), relativamente al tasso di falsi positivi, falsi negativi, e delle relative implicazioni etiche ed economiche. Inoltre, il testo esplora la relazione tra specificità e potenza, errori di inferenza, dimensione del campione, dimensione dell'effetto e potenza e offre indicazioni su come ottenere una potenza sufficiente negli esperimenti.
5. **Visualizing samples with box plots.** L'articolo discute l'uso dei box plot come metodo di visualizzazione, confrontandolo con gli istogrammi. È comunemente accettato il fatto che i box plot siano maggiormente informativi nei casi di dimensioni del campione più piccole, e che essi forniscano informazioni più dettagliate nelle code della distribuzione rispetto agli istogrammi. Caratterizzano i campioni utilizzando quartili, essi sono meno sensibili ai valori outlier, rendendoli preferiti rispetto alla media e alla deviazione standard per distribuzioni asimmetriche o di forma irregolare. Viene spiegata in dettaglio la costruzione di un box plot, inclusa la lunghezza e la larghezza della scatola, l'estensione dei baffi e le tacche per l'intervallo di confidenza. Inoltre, il documento scoraggia l'uso di grafici a barre con barre di errore ('dynamite plot') e suggerisce di integrare i tradizionali grafici a dispersione di media ed errore in grafici a scatola per una visualizzazione più informativa dei dati campione.

L'articolo si conclude fornendo un collegamento a uno strumento online per la creazione di box plot ed elenca i riferimenti per ulteriori letture.

6. **Comparing samples - part I.** L'articolo discute il confronto di coppie di campioni indipendenti o correlati utilizzando diversi approcci al test t. Sottolinea la necessità di metodi per confrontare quantitativamente i campioni per giudicare se le differenze nei dati supportano l'esistenza di un effetto nelle popolazioni che rappresentano, evidenziando le differenze nella gestione di campioni indipendenti e correlati. L'articolo spiega anche le incertezze e le misure statistiche come la varianza e la deviazione standard coinvolte nel confronto dei campioni. Si discute inoltre del test t accoppiato per esperimenti con campioni abbinati e della robustezza del test t a due campioni per confrontare le medie, affrontando le ipotesi e le considerazioni per le prestazioni ottimali.
7. **Comparing samples - part II.** L'articolo ha discusso l'interpretazione dei valori P quando vengono condotti un gran numero di test. Ha sottolineato che i valori P devono essere interpretati in modo diverso quando viene monitorato un gran numero di risultati sperimentali, poiché ci si aspetta che risultati rari si verifichino per caso. L'uso dei valori P è fuorviante in tali scenari e per mitigare questo problema sono necessari metodi di correzione con test multipli. L'articolo illustra diversi approcci e sottolinea l'influenza di questi metodi sulla proporzione delle inferenze. Ha spiegato l'impatto dei metodi di correzione a test multipli sulla percentuale di falsi positivi e false scoperte, mirando al contempo a mantenere il potenza statistico. L'articolo ha confrontato e valutato metodi di correzione popolari come quelli di Bonferroni, Benjamini-Hochberg (BH) attraverso una simulazione di esperimenti di espressione omica. Ha sottolineato l'importanza di riportare il valore q come equivalente FDR del valore P in scenari con test multipli e ha fornito linee guida per la scelta del metodo di correzione appropriato in base alla tolleranza per i falsi positivi e al numero di confronti.
8. **Nonparametric tests.** L'articolo discute i test non parametrici, come i test del segno e della somma dei ranghi di Wilcoxon, che sono più adatti per dati distorti o classificati rispetto ai test parametrici come il test t. Questi test non parametrici allentano le ipotesi di distribuzione e sono più facili da giustificare, ma hanno una sensibilità inferiore a causa della minore informazione inerente alle loro ipotesi. Gli autori confrontano il t-test per un campione con il test dei segni non parametrico, dimostrando che il test dei segni è meno sensibile e necessita di campioni di dimensioni maggiori o di più prove a supporto rispetto al t-test. Viene anche discusso il test della somma dei ranghi di Wilcoxon per confrontare due campioni e verificare se provengono da distribuzioni con la stessa mediana. Si noti che i metodi non parametrici rappresentano un approccio più cauto e generalmente hanno una potenza inferiore, ma possono raggiungere una potenza maggiore e sono una scelta migliore rispetto al test t per dati molto distorti. Il documento evidenzia inoltre le implicazioni dell'utilizzo di test non parametrici in scenari con test multipli e presenta un confronto tra le prestazioni dei test t e Wilcoxon in diverse distribuzioni di campionamento e dimensioni del campione.
9. **Analysis of variance and blocking.** L'articolo discute l'importanza della misurazione riproducibile degli effetti del trattamento e introduce l'uso dell'analisi della varianza (ANOVA) come mezzo per distinguere in modo affidabile tra gli effetti sistematici del trattamento e il rumore derivante dalla variazione biologica e dall'errore di misurazione. Gli autori forniscono una spiegazione dettagliata dell'ANOVA, della sua applicazione nei disegni sperimentali e del calcolo della statistica F come test per differenze sistematiche tra le medie di trattamento. Viene introdotto il concetto di partizionamento della varianza, con particolare attenzione alla varianza all'interno del gruppo come errore sperimentale e alla varianza tra gruppi come variazione biologica. Gli autori evidenziano l'integrazione del blocco per isolare la variazione biologica e migliorare la sensibilità nel rilevare gli effetti del trattamento. Vengono discussi i vantaggi e le considerazioni del blocco, insieme a una dimostrazione del suo impatto sull'efficienza sperimentale. Nel complesso, il documento sottolinea l'importanza di un buon

disegno sperimentale nel mitigare gli errori sperimentali e i fattori non oggetto di studio.

10. **Bayesian statistics.** L'articolo introduce la statistica bayesiana e il suo contrasto con la statistica classica (frequentista), sottolineando il ruolo della conoscenza pregressa, delle osservazioni e delle distribuzioni di probabilità. Discute il teorema di Bayes e la sua applicazione alla stima iterativa delle probabilità, utilizzando esempi di lanci di monete distorti e previsioni di malattie basate sulla presenza di marcatori. L'articolo illustra l'uso del teorema di Bayes nell'aggiornamento delle previsioni man mano che le prove si accumulano, fornendo una spiegazione dettagliata dei calcoli coinvolti nella derivazione delle probabilità a posteriori. Sottolinea inoltre l'importanza di tenere adeguatamente conto delle conoscenze pregresse e dei tassi di prevalenza nel fare previsioni informate sul fenomeno in studio. Gli autori evidenziano il ragionamento intuitivo e il calcolo conveniente forniti dal teorema di Bayes, ponendo le basi per il confronto con l'inferenza frequentista nelle discussioni future.

1.3 Quali software useremo

Presentazione 1.4 <https://bit.ly/3QtauYN>

Anche Chat GPT conferma che vi sono molti validi software per eseguire le analisi statistiche, e tra essi il software R primeggia, tallonato dal linguaggio Python che è particolarmente versatile nella gestione dei dati web. Tuttavia il tempo di apprendimento richiesto per diventare esperti in questi settori può essere demotivante. Per fortuna, possiamo sfruttare l'affidabilità di R senza conoscerne la sintassi, avvalendoci di un'interfaccia grafica molto intuitiva: il software JASP, <https://jasp-stats.org/>.

Attività 1.1 — Prendiamo confidenza con JASP. Proviamo a vedere assieme come funziona JASP.

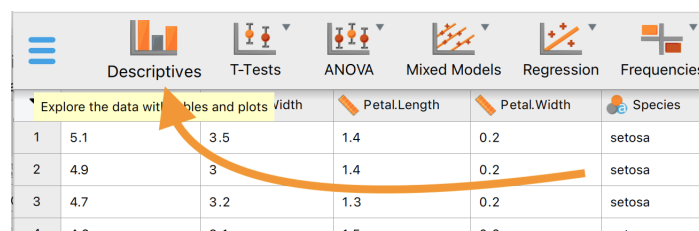
1. colleghiamoci a <https://github.com/umgstat/maxillo/tree/main/DATASET>
2. scarichiamo nel nostro computer il dataset `partecipanti.ods`
3. importiamolo in JASP
4. Rispondendo al questionario <https://forms.gle/8yQhKHFRjQqujJyz8> effettuiamo alcune semplici analisi statistiche.

2. Descrivere i dati

2.1 La basi della statistica descrittiva

Presentazione 2.1 <https://bit.ly/4aUFCZB>

Per iniziare dalle basi della statistica descrittiva ci esercitiamo con il dataset *iris*; vogliamo infatti imparare a descrivere le caratteristiche quantitative o qualitative di un dataset indicando la **posizione** dei dati e la loro **variabilità**. Vediamo ad esempio che nel dataset sono raccolte informazioni 'biometriche' riguardanti 3 diverse specie di fiori iris (*setosa*, *versicolor* e *virginica*): petal length, petal width, sepal length e sepal width. Impariamo a conoscere i contenuti del menu Descriptives:



The screenshot shows the JASP software interface. At the top, there is a menu bar with icons for Descriptives, T-Tests, ANOVA, Mixed Models, Regression, and Frequencies. Below the menu bar, there is a table with 6 columns: 'Explore the data with variables and plots', 'width', 'Petal.Length', 'Petal.Width', and 'Species'. The table contains 4 rows of data, all of which are 'setosa' species. An orange arrow points from the 'Descriptives' menu icon to the 'Explore the data with variables and plots' column header.

	Explore the data with variables and plots	width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa

ed imparare a distinguere:

- le **misures di tendenza centrale**, o di **posizione**
- le **misure di forma** o di **dispersione**
- il concetto di **quantile**
- quando un dataset si dice **bilanciato**
- quando un dataset si dice **completo**

E, come si suol dire, 'A picture is worth a thousand words': impariamo a rappresentare i seguenti grafici: i *dot plots*, i *distribution plots* ed i *boxplots*, precisando i concetti di quantili e di outliers. Impariamo inoltre a rappresentare gli *scatter plots*.

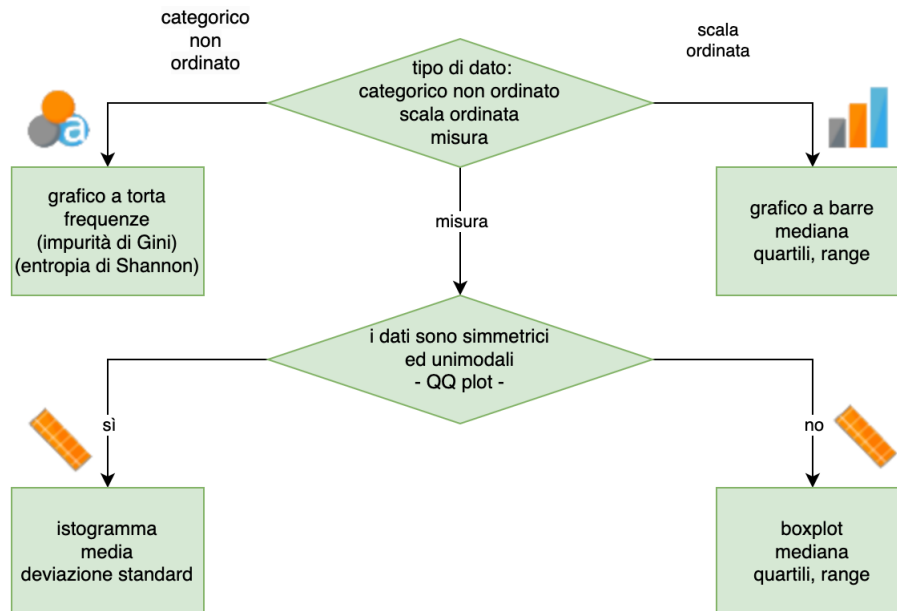
Attività 2.1 — Descrivere un dataset con JASP. Proviamo a vedere assieme come funziona JASP.

1. colleghiamoci a <https://github.com/umgstat/maxillo/tree/main/DATASET>
2. scarichiamo nel nostro computer il dataset *cefalometria.ods*

3. importiamolo in JASP

4. Rispondiamo al questionario <https://forms.gle/ShjTVxQqL1JgpkvX7>.

Per illustrare un dataset in modo appropriato, suggerisco di tenere a mente questo algoritmo:



Conviene poi tenere a mente che ci serviamo della statistica non per ragioni 'cosmetiche', per farci belli; ma per ragioni comunicative, 'descrittive' appunto: attenzione all'uso del simbolo \pm , ed attenzione a non essere inutilmente ridondanti. Infine, un cenno alla necessità di dover recuperare le informazioni descrittive mancanti come ad esempio, stimare la media μ o la deviazione standard σ conoscendo la mediana m , i quartili Q_1 , Q_3 oppure il range $[a, b]$ dei dati effettuati su un certo campione di dimensione n : sono questioni che spesso si incontrano nella pianificazione di uno studio o di una ricerca, quando per esempio ci serve individuare un opportuno *sample size*. Ecco tre utili relazioni per stimare la media:

$$\mu \approx \frac{a + 2m + b}{4} + \frac{a - 2m + b}{4n} \approx \frac{a + 2m + b}{4}$$

$$\mu \approx \frac{a + 2Q_1 + 2m + 2Q_3 + b}{8}$$

$$\mu \approx \frac{Q_1 + m + Q_3}{3}$$

Invece per la deviazione standard, le formule dipendono anche da due 'coefficienti' ξ e η , che sono

$$\sigma \approx \frac{b - a}{\xi(n)}$$

$$\sigma \approx \frac{Q_3 - Q_1}{\eta(n)}$$

$$\sigma \approx \frac{1}{2} \left(\frac{b - a}{\xi(n)} + \frac{Q_3 - Q_1}{\eta(n)} \right)$$

Per maggiori dettagli, e per calcolare automaticamente i coefficienti della tabella, vi rimandiamo a <https://arxiv.org/abs/2310.12550>.

2.4 Descrivere l'affidabilità: le basi dell'inferenza statistica

[illegible]

3. Inferenza statistica univariata

3.1

[illegible]

3.2

[illegible]

Attività 3.1 — Esercitarsi a testare differenze tra due gruppi. In preparazione, in preparazione.

3.6 La retta di regressione

[illegible]

4. Inferenza statistica univariata

4.1 L'Ancova ed i modelli lineari

[illegible]

4.2 I modelli lineari generalizzati

[illegible]

Bibliography

Articles

Books

Index

- Descrivere i dati, 9
 - Descrivere l'affidabilità: le basi dell'inferenza statistica, 12
 - Frequenze, rischi e predittività, 11
 - Le basi della statistica descrittiva, 9
 - Riconoscere i fenomeni aleatori, 11
- Inferenza statistica univariata, 13, 17
 - I modelli lineari generalizzati, 17
 - L'Ancova ed i modelli lineari, 17
 - L'approccio bayesiano al T-test, 14
 - L'esempio storico del T-test, 13
 - La retta di regressione, 15
 - Testare differenze tra due gruppi, 13
 - Testare differenze tra più gruppi: l'Anova, 14
 - Un altro approccio: i metodi di permutazione e di ricampionamento, 14
- Per iniziare, 5
 - Di cosa parleremo in questo corso, 5
 - Libri di testo e collane di articoli, 5
 - Collane di articoli, 5
 - Libri di testo, 5
 - Quali software useremo, 8

A. Appendix Chapter Title

A.1 Appendix Section Title

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam auctor mi risus, quis tempor libero hendrerit at. Duis hendrerit placerat quam et semper. Nam ultricies metus vehicula arcu viverra, vel ullamcorper justo elementum. Pellentesque vel mi ac lectus cursus posuere et nec ex. Fusce quis mauris egestas lacus commodo venenatis. Ut at arcu lectus. Donec et urna nunc. Morbi eu nisl cursus sapien eleifend tincidunt quis quis est. Donec ut orci ex. Praesent ligula enim, ullamcorper non lorem a, ultrices volutpat dolor. Nullam at imperdiet urna. Pellentesque nec velit eget est euismod pretium.

B. Appendix Chapter Title

B.1 Appendix Section Title

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam auctor mi risus, quis tempor libero hendrerit at. Duis hendrerit placerat quam et semper. Nam ultricies metus vehicula arcu viverra, vel ullamcorper justo elementum. Pellentesque vel mi ac lectus cursus posuere et nec ex. Fusce quis mauris egestas lacus commodo venenatis. Ut at arcu lectus. Donec et urna nunc. Morbi eu nisl cursus sapien eleifend tincidunt quis quis est. Donec ut orci ex. Praesent ligula enim, ullamcorper non lorem a, ultrices volutpat dolor. Nullam at imperdiet urna. Pellentesque nec velit eget est euismod pretium.