Statistica e analisi dei dati (UNISA 2022/2025)

**INTRODUZIONE**

Nella ricerca la statistica si divide in:

* Raccolta dei dati
* Elaborazione dei dati
* Descrizione

La statistica nella ricerca si divide in due tipi di analisi: Descrittiva e inferenziale.

L’analisi descrittiva si occupa della presentazione e della sintesi dei dati.

Inferenziale permette di trasferire informazioni ottenute su un campione all’intera popolazione.

La variabile è ciò che viene osservato e può assumere diversi valori.

Cos’è la statistica descrittiva?

E’una branca della statistica che si concentra sulla raccolta, sull'organizzazione, sull'analisi e sulla presentazione dei dati in modo da riassumere e descrivere le principali caratteristiche di un insieme di dati.

(L'obiettivo principale della statistica descrittiva è quello di fornire una comprensione chiara e sintetica dei dati)

Quindi la statistica descrittiva dato un fenomeno osservabile ci permette di analizzare il suo comportamento e le sue caratteristiche.

Possiamo misurare,contare/osservare i valori assunti da queste variabili e stabilire una loro caratteristica.

In statistica nell’analisi descrittiva le variabili possono essere di tre tipi: Qualitative,Quantitative e Ordinabili.

**L’AMBIENTE GRAFICO ‘R’**

R è dotato di un sofisticato ambiente grafico che permette di creare grafici per illustrare i risultati di elaborazioni statistiche e ne permette la totale modifica con anche annessa esportazione.

Comandi principali:

plot(): funzione che permette di illustrare l’andamento dei valori assunti dal vettore rispetto ai relativi indici.

line(): permette con l’opzione ‘*type= l’* di creare linee interconnesse e creare una **serie storica dei valori**

**inoltre** il comando line può permetterci di aggiungere linee ad un grafico già esistente (usata insieme alla funzione plot).

Rappresentazione grafici per vettori numerici…

**SERIE TEMPORALI:**

Una serie temporale è una sequenza di osservazioni di una variabile raccolte o registrate grazie ad intervalli di tempo regolari, l'obiettivo è analizzare i dati per identificare pattern o tendenze che possano aiutare a fare previsioni sul futuro (forecasting).

Una serie temporale **univariata** analizza una sola variabile misurata nel tempo. (Dinamica temporale di una singola quantità).

Una serie temporale **multivariata** considera più variabili misurata nel tempo (attenzione sulla dinamica temporale di più quantità).

Una serie temporale (serie storica) può essere memorizzata in un vettore se i dati sono univariati oppure in una matrice o in un data frame nel caso di dati multivariati.

R dispone di una funzione specifica, denominata **ts**() (time series) per rappresentare i valori di una serie temporale insieme ad alcune altre caratteristiche (inizio, fine, eventuale periodicità stagionale, . . .).

La *funzione window()* è utile per lavorare con serie temporali quando si vuole isolare o analizzare un sottoinsieme di dati temporali, viene particolarmente usata da oggetti di tipo **ts** e si applica come: window(x,start,end) dove x è l’oggetto di tipo time series.

TSstudio: è una libreria avanzata per R specializzata dell’analisi e visualizzazione di serie temporali. Crea grafici ts\_plot, ts\_heatmap.

ALTRI TIPI DI GRAFICI

BarPlot:

Rappresenta graficamente uan barra di lunghezza della variabile per mostrare le frequenze o altre misure quantitative di categorie specifiche di dati.

Il BarPlot viene utilizzato per:

* Comparazione: confronta i valori su diverse tipologie.
* Visualizzazione di dati aggregati: medie,conteggi o somme.
* Rappresentazione di distribuzioni: distribuzioni categorizzabili.

ScatterPlot: sono usati per coppie di variabili es di campione: {(x1,y1),(x2,y2)…} osservazioni di n coppie di tipo quantitativo.

Queste coppie formano punti in un piano euclideo e possiamo distinguere la variabile dipendente (ordinate->Y) da quella dipendente(ascisse->X).

Si ottiene un grafico che può portare a capire se ci sono irregolarità delle relazioni tra le variabili.

La funzione pairs() è in grado di visualizzare in un’unica finestra grafica una pluralità di grafici per punti ottenuti mettendo in relazione tutte le coppie di variabili quantitative definite all’interno di un data frame (o di una matrice).

**DISTRIBUZIONI DI FREQUENZA**

La distribuzione di frequenza, consideriamo una variabile X e indichiamo con z1 , z2 , . . . , zk le modalità distinte da essa assunte

*Se la variabile X:*

- è qualitativa le modalità indicano delle qualità distinte degli individui.

- è quantitativa le modalità sono dei numeri reali distinti.

Se indichiamo con n il numero di volte in cui ciascuna osservazione zi è presente nel campionela frequenza assoluta con cui essa appare nel campione viene chiamata **distribuzione di frequenza.**

Se non esistono dati mancanti, la somma delle frequenze assolute è sempre uguale alla numerosità del campione, ossia n = n1 + n2 + . . . + nk.

FREQUENZA RELATIVA

La frequenza relativa di un evento è il rapporto tra la frequenza assoluta di quell’evento e il numero totale di osservazioni o prove:

Viene espressa come frazione o percentuale e indica la proporzione con cui un determinato evento si verifica rispetto al totale:

Frequenza relativa: ----> (i=1….k)

La funzione table() in R crea una tabella di contingenza che mostra la frequenza (assoluta) degli elementi unici di un vettore o la combinazione di livelli di più variabili

Una tabella di contingenza è una tabella che riassume la distribuzione congiunta di due o più variabili categoriali, mostrando le frequenze (assolute o relative) con cui si verificano le combinazioni dei livelli delle variabili.

Il comando rep() è usato per ripetere più volte le varie modalità nel vettore

La funzione table() riporterà le modalità della variabile qualitativa che presentano un valore di frequenza assoluta diverso da zero.

***Ricorda: se vuoi ordinare devi farlo prima di calcolare la frequenza***

La frequenza assoluta cumulata rappresenta il numero totale di osservazioni che hanno un valore inferiore o uguale a un determinato valore in un insieme di dati (E’ LA SOMMA DELLE FREQUENZE RELATIVE FINO AD ORA OTTENUTE)

Gestione dei dati mancanti

Ribilanciamento dei dataset

La funzione cut() raggruppa i dati relativi ad un vettore in intervalli elencando nel parametro breaks gli estremi degli intervalli (aperti a sinistra e chiusi a destra).

Encoding di variabili categoriali: Le frequenze relative delle categorie possono essere utilizzate per trasformare variabili categoriali in valori numerici, un processo chiamato target encoding.

E’ possibile rappresentare anche un grafico a bastoncini utilizzando il comando plot(table(x)):

Un altro tipo di rappresentazione si ottiene mediante i diagrammi a torta che permettono di attribuire ciascuna modalità della variabile qualitativa in esame ad un settore circolare di un cerchio, la cui ampiezza è proporzionale alle frequenze. I diagrammi a torta sono quindi utili quando i dati non sono numerici ma categorici

**TABELLE DI CONTINGENZA**

Per vedere se esistono legami tra due variabili X e Y rilevati congiuntamente sugli stessi individui occorre costruire una tabella a doppia entrata detta tabella di contingenza.

Consideriamo un campione (x1 , y1 ), (x2 , y2 ), . . . , (xn , yn ) costituito da n osservazioni di (X, Y)

L’elemento nij in posizione (i,j) della tabella di contingenza, detto frequenza congiunta, rappresenta il numero di volte in cui una particolare coppia di modalità (zi ,wj ) si presenta nel campione.



Dalla tabella è possibile ottenere la distribuzione di frequenza marginale di X:



Mentre la distribuzione di frequenza marginale di Y è:

In R per determinare le distribuzioni di frequenza **bivariate** il comando da utilizzare è sempre table()

- Per estrarre la riga i-esima basta utilizzare il comando nomeTable[i, ] (distribuzione frequenza marginale X)

- Per estrarre la colonna j-esima basta utilizzare il comando nomeTable[,j]. (frequenza marginale Y)

È anche possibile considerare le frequenze relative associate ad una tabella di contingenza

Immagine che contiene testo, linea, Carattere, diagramma

Descrizione generata automaticamenteLe frequenze relative congiunte misurano la proporzione di osservazioni che appartengono simultaneamente a specifiche categorie di due variabili (i=1,2,…,n; j=1,2,…,n)

Le frequenze relative marginali rappresentano la distribuzione di una singola variabile senza considerare l'altra variabile - Vengono calcolate come la somma delle frequenze relative congiunte per ciascuna categoria di una variabile, indipendentemente dalle categorie dell'altra variabile

In R, per calcolare la distribuzione delle frequenze relative congiunte si può utilizzare il comando prop.table(nomeTabella)

Da questa si possono ricavare le distribuzioni delle frequenze relative marginali utilizzando margin.table()

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamente

Una tabella di contingenza può essere rappresentata mediante un grafico con rettangoli costituito da h × k rettangoli. L’area totale dei rettangoli è proporzionale al numero di casi

La **tabella di contingenza** può essere rappresentata mediante un **grafico a barre sovrapposte (Stacked).**

Il numero di barre è pari al numero delle modalità delle colonne della tabella di contingenza - Inoltre, all’interno di ciascuna barra sono rappresentate una sopra l’altra in altezza le frequenze di ciascuna modalità delle righe della tabella di contingenza.

Per grafici a *frequenza* *marginale* si possono usare barplot o grafici a stecca.

Il CAP3 si sofferma su comandi grafici di legenda/abbellimento o funzioni.

DA ISTOGRAMMI A BARPLOT AD INTAGLIO

ISTOGRAMMA…

hist genera istogrammi utilizzando un vettore numerico

hist(x, probability=TRUE) le colonne rappresentano frequenze relative invece che assolute

Per realizzare un istogramma in base alle frequenze relative, occorre impostare nella funzione hist() il parametro freq = FALSE

KERNEL DENSITY PLOT

I kernel density plot possono essere utilizzati in alternativa al tradizionale istogramma

Sono grafici basati sulla stima kernel di densità o

La stima kernel di densità è il metodo non parametrico impiegato per realizzare i kernel density plot o Si traccia una curva continua determinata da un fattore K (kernel), e da un parametro h (ampiezza della banda o bandwidth)

La densità descrive la distribuzione di una variabile continua (è rappresentata dall’area sottesa al grafico).

I kernel density plot forniscono una stima continua della densità dei dati

*A differenza degli istogrammi*, che segmentano i dati in intervalli discreti, i kernel density plot utilizzano una funzione kernel per creare una rappresentazione più fluida della distribuzione dei dati.

Immagine che contiene testo, schermata, Carattere, algebra

Descrizione generata automaticamente

Il parametro ‘ h ‘ è importante in quanto:

Un valore troppo vicino allo 0 rende la stima irregolare e la funzione di kernel applicata a ciascun punto dati avrà un picco molto stretto.

Una stima irregolare si presenta quando: ogni punto dati influisce sulla densità stimata in modo molto localizzato (anche piccole variazioni nei dati possono provocare grandi cambiamenti nella stima della densità.).

Una varianza Elevata si verifica quando la stima della densità è estremamente sensibile ai dati individuati, in questo caso la varianza della stima aumenta. -> comporta l’inaffidabilità nella densità stimata.

Un valore troppo grande fa si che la funzione kernel applicata a ciascun punto dati è "spalmata" su un'area molto più ampia.

Distorsione della stima: una larghezza di banda eccessiva provoca una smussatura della stima della densità con la rispettiva perdita di dati importanti.

Perdita di informazioni: aree con molte concentrazioni di dati possono apparire come una densità uniforme portando a conclusioni errate o fuorvianti.

Immagine che contiene testo, schermata, Carattere

Descrizione generata automaticamente

La regola di silverman è implementata di default in R con la funzione density()

Immagine che contiene testo, schermata, Carattere

Descrizione generata automaticamenteCHIEDI SE DEVI IMPARARE I TIPI DI KERNEL

QUARTILI

I quartili sono strumenti fondamentali in data science e analisi dei dati perché aiutano a comprendere la distribuzione e la dispersione dei dati, specialmente quando si lavora con dataset grandi o complessi.

**Aiutano a identificare la distribuzione.**

**Rilevano gli outlier** (valori anomali) ( l'intervallo interquartile (IQR), è la differenza tra il terzo quartile (Q3) e il primo quartile (Q1), I valori che si trovano al di fuori dell'IQR possono essere considerati outlier, poiché si distanziano dalla distribuzione centrale dei dati)

Considerando un campione xi per i che va da 1,…,n si procede a rappresentare il campione in ordine crescente:

* Primo quartile: : il valore per il quale il 25% dei dati sono alla sua sinistra 𝑄1
* Secondo quartile: il valore per il quale il 50% dei dati sono alla sua sinistra 𝑄2 (detto anche mediana)
* Terzo quartile: il valore per il quale il 75% dei dati sono alla sua sinistra 𝑄3
* 𝑄0 e 𝑄4 forniscono in minimo ed il massimo dei valori del campione

La centralità viene espressa con la media, la forma si deduce da Q1 e Q3 e anche la dispersione.

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

I boxplot ad intaglio (notched boxplots) sono una alternativa grafica dei boxplot

Sono utilizzati per visualizzare la distribuzione dei dati, ma forniscono un'informazione aggiuntiva sulla significatività statistica delle differenze tra mediane di gruppi, I boxplot ad intaglio sono spesso utilizzati per confrontare gruppi • Se si effettua un test statistico sulla differenza tra le mediane dei due gruppi con un livello di significatività del 5%, per grandi campioni le mediane dei due gruppi differiscono statisticamente se gli intervalli di confidenza dei due boxplot non si sovrappongono.

Da qui si ricavano gli intervalli di confidenza, il confronto tra questi (dato dalle mediane) può farci dedurre:

Sovrapposizione delle tacche : Se le tacche si sovrappongono tra due gruppi, significa che non ci sono prove sufficienti per affermare che le mediane dei gruppi sono significativamente diverse o Questo suggerisce che non c'è evidenza di una differenza significativa tra le mediane.

Non sovrapposizione delle tacche :Se le tacche non si sovrappongono, c'è una forte indicazione che le mediane dei due gruppi sono significativamente diverse o Questo suggerisce che è molto probabile che la differenza osservata tra le mediane non sia dovuta al caso, ma rappresenti una differenza reale tra i gruppi.

**DIAGRAMMA DI PARETO**

Permette di analizzare fenomeni qualitativi, grazie a il diagramma di Pareto infatti è possibile analizzare un insieme di dati in modo da determinare le poche variabili tra le tante prese in esame che influenzano in modo significativo i risultati finali.

Il diagramma è rappresentato tramite barre verticali ordinate in modo decrescente sulla frequenza relativa.

Le frequenza relative sono anche viste nella loro forma cumulata mediante dei segmenti crescenti.

{

Tracciare la cosiddetta linea dei valori cumulati o linea cumulativa: Questa linea risulta utile quando si vogliono individuare le percentuali cumulate di più colonne

Segmento 1: congiunge il punto che indica la percentuale dei difetti del primo tipo con quello ad altezza pari alla somma delle percentuali dei difetti del primo e secondo tipo

Altri segmenti: si ripete il procedimento per i segmenti successivi; o Ultimo segmento: terminerà nel punto più alto della scala percentuale, corrispondente al 100% dei difetti.

}

**GRAFICI DI FUNZIONE**

La funzione: curve(expr, from, to)

disegna una curva sulla base dell’espressione indicata in expr, nell’intervallo [from, to]

GRAFICI MULTIPLI…

***CORRELAZIONI BARPLOT***

La correlazione è una misura statistica che descrive la relazione tra due variabili.

Essa indica quanto e in che modo due variabili tendono a variare insieme o La correlazione può essere utilizzata per identificare se esiste una relazione tra le variabili e la direzione di tale relazione

Correlazione Positiva o Si verifica quando un aumento in una variabile corrisponde a un aumento nell'altra variabile

Correlazione Negativa o Si verifica quando un aumento in una variabile corrisponde a una diminuzione nell'altra variabile

Assenza di Correlazione o Non c'è una relazione chiara tra le due variabili; i cambiamenti in una variabile non influenzano l'altra

Il pacchetto corrplot è utilizzato per visualizzare matrici di correlazione, fornisce una rappresnetazione grafica intuitiva delle varie correlazioni degli elementi.

Il metodo corrplot() del pacchetto corrplot è utilizzato per visualizzare le matrici di correlazione

La firma del metodo e i principali parametri che puoi utilizzare:

corrplot(m, method = "circle", type = "full", addCoef.col = NULL, tl.col = "black", tl.srt = 45, diag = TRUE, main = NULL)

MAPPE INTERATTIVE:

leaflet è un pacchetto R che consente di creare mappe interattive, facilmente personalizzabili.

SI POSSONO AGGIUNGERE: poligoni per sezioni da evidenziare, marker, colori per attributi, linee evidenziabili, cerchi per indicare grandezza di valori.