ANALISI STATISTICA

DATABASE

È essenziale creare una base di dati ben strutturata per ottenere risultati precisi e validi nelle analisi statistiche.
Il database è un foglio di lavoro elettronico che raccoglierà con precisione i dati dello studio e ci aiuterà ad esplorarli prima di affrontare un'adeguata analisi statistica. Esistono diversi programmi su cui creare un database, più comunemente viene utilizzato Excel.
Una volta creato il database consigliamo di farne una prima copia in cui si inizia a scremare i dati o ad ordinarli in maniera più idonea (la chiameremo database_statistica) e successivamente si andrà a fare una seconda copia che verrà modificata in modo da poter eseguire le diverse analisi statistiche e i diversi calcoli (la chiameremo database_calcoli).

Assicurati di includere variabili chiave e di utilizzare un formato coerente per evitare errori durante la fase di analisi.

SOFTWARE

E' importante la scelta del software che verrà utilizzato per eseguire le analisi statistiche. Esistono diversi tipi di software, i più diffusi sono STATA, SPSS e MedCalc (prova gratuita 14 giorni). SPSS si può scaricare da Unito.

VARIABILI

Esistono diversi tipi di variabili che per essere confrontate richiedono test statistici diversi.
Le variabili possono essere categoriali o continue.
Le variabili **categoriali** sono:

- dicotomiche: variabili che possono essere definite solo secondo due categorie (ad esempio: sì/no, presente/assente)
- nominali: in cui vi sono più di una categoria senza un ordine interno
- ordinali: in cui le categorie possono essere ordinate

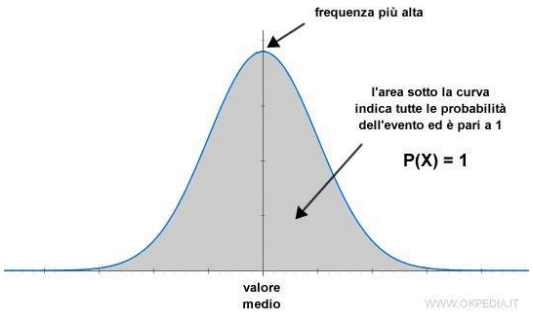
Le variabili categoriali sono definite solitamente mediante numeri e percentuale sul totale.
Le variabili **continue** sono ad esempio età, peso, altezza. Per queste esiste un'unità di misura che permette il confronto diretto tra due osservazioni. Vengono definite in base a ***Media±SD (deviazione standard)*** o ***Mediana e IQR (range interquartile).*** E' fondamentale conoscere la differenza tra media e mediana.

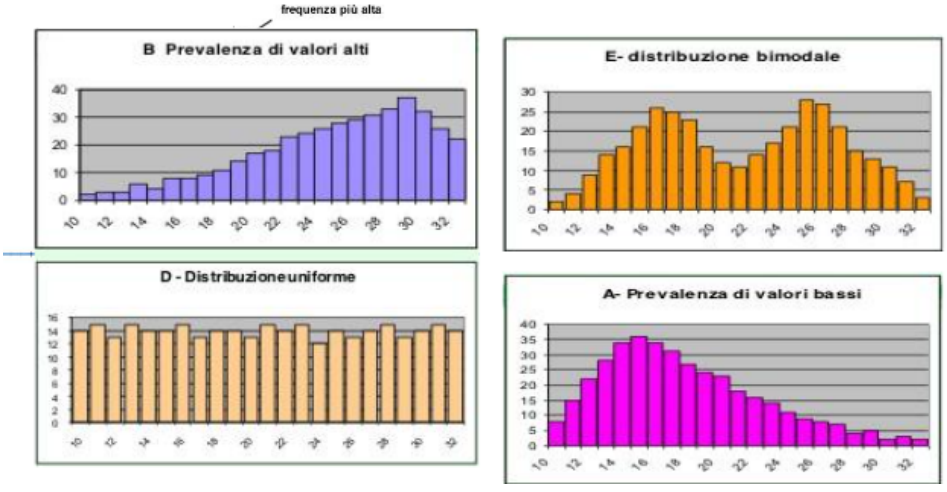
La media è la semplice media dei campioni presenti, in un foglio di calcolo Excel si esprime con la formula =Media(C2:C10). La deviazione standard è la misura della dispersione del campione (ovvero quanto i valori del campione si discostano dalla media). In un foglio di calcolo Excel si ottiene con la formula =dev.st(C2:C10).

La mediana è il valore assunto dalle unità statistiche che si trovano nel mezzo della distribuzione. In un foglio di calcolo Excel la mediana è ottenuta tramite la formula = Mediana(C2:C10).I quartili sono valori/modalità che ripartiscono la popolazione in quattro parti di uguale numerosità. La differenza tra il terzo ed il primo quartile è un indice di dispersione ed è definito come range interquartile.

DISTRIBUZIONI

Ogni variabile casuale ha una corrispondente distribuzione di probabilità.
Una variabile casuale può essere discreta (che assume solo un numero finito o numerabile di risultati) o continua (che può assumere qualsiasi valore all'interno di un intervallo).
Molte variabili continue seguono una distribuzione normale, nota anche come distribuzione a campana o gaussiana: sull'asse delle ascisse sono indicate le classi di valori, che può assumere la variabile in esame, mentre sulle ordinate sono rappresentate le frequenze della variabile. Questa curva rappresenta la funzione della densità di probabilità di trovare un determinato valore della variabile nella popolazione. molte misure mediche (peso, altezza, livelli ematici di acido urico…) seguono una distribuzione normale.
Tuttavia non sempre le variabili seguono una distribuzione normale, come i punteggi. Per quanto riguarda i punteggi, le misurazioni possono concentrarsi sul lato sinistro, sul lato destro, al centro, da nessuna parte o in due punti.





Si passa da un approccio parametrico ad uno non parametrico, ovviando così, senza perdita sostanziale di efficienza, alle limitazioni sopra accennate nei seguenti casi:

A. quando non è noto il modello distributivo
B. ci sono dati mancanti non a caso
C. la numerosità del campione è inferiore al numero delle variabili

Tra i dati che non si adattano alla distribuzione normale vi sono i punteggi (score) e le votazioni utilizzati da osservatori/osservatrici, come medici, psicologi/psicologhe, insegnanti, giudici di gara, ecc., per valutare fenomeni come l'intelligenza, la capacità di memoria, il rendimento a scuola, la produttività nel lavoro, la prestazione atletica, ecc.

In tutti questi casi la scala non è riferita a grandezze fisiche, bensì a diversi livelli qualitativi di espressione del fenomeno, trasformati numericamente solo in base a convenzione. Ad esempio, nei licei si attribuisce 6 per indicare la sufficienza, mentre all'università si attribuisce 18.

Questi concetti sono utili per comprendere qual è il corretto test da utilizzare ai fini dell'analisi statistica (vedi dopo) e quindi per fare **INFERENZA**, ovvero il processo in cui si inducono le caratteristiche della popolazione generale, mediante un campione selezionato in modo casuale.

TEST di IPOTESI

La stragrande maggioranza della statistica su cui si basano gli studi che noi leggiamo è di tipo frequentista e fondata sul test di ipotesi o *null hypothesis testing,* in cui si parte da un'assunzione iniziale (**ipotesi nulla**), che viene sottoposta al valio del test e, in base ai risultati ottenuti, si stabilisce se stiamo sperimentando un effetto casuale o si tratta di un risultato reale.

Ad esempio, vogliamo valutare se l'assunzione di aspirina riduca il rischio cardiovascolare rispetto al placebo nei soggetti con pregresso infarto.

Poiché abbiamo valutato l'effetto dell'aspirina su un campione di soggetti, dobbiamo chiederci **se l'effetto osservato è reale** (cioè è presente anche nella popolazione da cui proviene il campione).

Per fare ciò ci avvaliamo del processo di ipotesi nulla ovvero:

a) Assumiamo che l'ipotesi nulla $\boldsymbol{H_0}$ sia vera: non esiste nessun effetto nella popolazione
b) Valutiamo la probabilità di osservare la relazione che abbiamo visto nel nostro campione quando l' $\boldsymbol{H_0}$ è vera
c) Se la probabilità è molto bassa concludiamo che l'ipotesi nulla è errata e rigettiamo l'ipotesi nulla, accettando quella alternativa $\boldsymbol{H_1}$, ovvero che l'aspirina funzioni. In caso contrario manteniamo l'ipotesi nulla e non possiamo affermare che la relazione sia reale.

*Nel nostro caso l'ipotesi nulla $\boldsymbol{H_0}$ è che l'aspirina non sia efficace, l'ipotesi alternativa $\boldsymbol{H_1}$ è che l'aspirina sia più efficace del placebo.

Distinzione tra H0 e H1

L'ipotesi nulla (H0) è quella che si assume vera e che si vuole testare, e tipicamente afferma che "non c'è effetto" o "non c'è differenza." L'ipotesi alternativa (H1) è ciò che si accetta se l'ipotesi nulla viene rigettata.

Valore p: La probabilità (p-value) rappresenta la probabilità di osservare un effetto pari o più estremo di quello osservato nel campione, assumendo che l'ipotesi nulla sia vera. Se questa probabilità è molto bassa, suggerisce che l'effetto osservato è improbabile sotto l'ipotesi nulla.

Se p < 0,05 rifiutiamo l'ipotesi nulla H_0 a favore di quella alternativa H_1.
Se p > 0,05, manteniamo l'ipotesi nulla (test non significativo).

QUALE TEST UTILIZZARE?

OBIETTIVO	VARIABILI NORMALI	VARIABILI NON NORMALI	VARIABILI CATEGORICHE (DICOTOMICHE)	SOPRAVVIVENZA
Comparare due gruppi non appaiati	TEST T DI STUDENT	MANN WITHNEY TEST	CHI QUADRO (o Test esatto di Fisher)	LOG-RANK TEST (Kaplan-Meier)
Comparare tre o più gruppi non appaiati	ANOVA	KRUSKAL-W ALLIS TEST	CHI QUADRO	REGRESSIONE DI COX
Quantificare l'associazione tra due variabili	R DI PEARSON	Rho di SPEARMAN TAU-B di KENDALL		
Predire un valore in base a un'altra variabile misurata	REGRESSIONE LINEARE (UNIVARIATA)	REGRESSIONE NON PARAMETRICA	REGRESSIONE LOGISTICA SEMPLICE	REGRESSION DI COX
Predire un valore in base a più variabili misurate	REGRESSIONE LINEARE (MULTIVARIATA)		REGRESSIONE LOGISTICA MULTIPLA	REGRESSIONE DI COX

TEST CHI QUADRO

Il test chi-quadro è utilizzato per analizzare le associazioni tra variabili categoriali e si basa sulle tabelle di contingenza.
Il test valuta la differenza tra le frequenze osservate in ciascuna categoria della tabella e le frequenze attese, ovvero quelle che ci si aspetterebbe di trovare se l'ipotesi nulla (H0) fosse vera.

ESEMPIO**:**
Ad esempio, immaginiamo di voler valutare se l'utilizzo del casco protettivo in bicicletta riduca la frequenza di traumi cranici e per farlo confrontiamo il gruppo che usa il casco con quello che non lo utilizza.

Costruiamo, dunque, una tabella di contingenza.

Trauma cranico	utilizzo casco protettivo	non utilizzo casco protettivo	totale
Sì	17	218	235
No	130	428	558
Totale	147	646	793

Per esaminare l'efficacia del casco protettivo per bicicletta, vogliamo sapere se esiste un'associazione tra traumi cranici e uso del casco tra i soggetti coinvolti in un incidente. Pertanto, testiamo l'ipotesi nulla:

H_0: la proporzione di persone che hanno riportato traumi cranici tra coloro che indossavano il casco al momento dell'incidente è uguale alla proporzione di soggetti che hanno riportato traumi cranici tra coloro che non indossavano il casco.

contro l'ipotesi alternativa:

H_1: Le proporzioni di soggetti che hanno riportato traumi cranici tra coloro che indossavano il casco e coloro che non lo indossavano non sono uguali.

Eseguiamo il test ad un livello di significatività $\alpha = 0{,}05$.

I software di statistica attualmente disponibili permettono di calcolare agevolmente e automaticamente le frequenze. Ad esempio: i soggetti che hanno riportato traumi cranici nel gruppo che indossava il casco sono 17/147=11,56%. Invece, la proporzione di soggetti che ha riportato il trauma cranico nel gruppo che non indossava il casco è 218/646 = 33,74%. Per capire se le frequenze osservate sono **realmente diverse** e non per effetto del caso bisogna considerare la probabilità *p*, che viene fornita dai software e se è inferiore a 0,05 possiamo rifiutare l'ipotesi nulla.

Il test chi-quadro è appropriato quando il campione è sufficientemente grande e ogni cella ha una f**requenza attesa di almeno 5**.
Mentre quest'ultimo test è esatto solo asintoticamente per dimensioni molto grandi dei campioni, il **test esatto di Fisher** è, come dice il nome, sempre esatto. Quindi quando una cella ha valori inferiori a 5 si deve usare la *p* fornita dal test di Fisher.

TEST T DI STUDENT

Il test T di student è un test parametrico e per utilizzarlo assumiamo che la variabile **continua** in esame sia distribuita **normalmente** nella popolazione.
Ad esempio, supponiamo ora di avere le misurazioni del livello di ferro sierico per due campioni di bambini e bambine: un gruppo di sani/e e un gruppo affetto da fibrosi cistica. Le due popolazioni sono indipendenti e normalmente distribuite. Se la popolazione di bambini/e malati/e ha un livello medio di ferro sierico μ_1 e la popolazione di bambini/e sani/e ha una media μ_2 possiamo ancora una volta testare l'ipotesi nulla che le medie delle due popolazioni siano uguali.

H_0: $\mu_1=\mu_2$
Analogamente agli altri test statistici se la probabilità p di osservare quell'effetto per il caso è inferiore a 0,05 possiamo rifiutare l'ipotesi nulla.

TEST di MANN-WHITNEY

Il test di Mann-Whitney (anche noto come test U di Mann-Whitney o test di Wilcoxon) è utilizzato per selezionare due campioni da due popolazioni indipendenti ed è il corrispondente test non parametrico del test t di Student per due popolazioni. A differenza del test t, **non richiede che le due popolazioni in esame siano distribuite normalmente**, ma si basa sulla distribuzione per ranghi. Esso assume però che le distribuzioni abbiano la stessa forma generale.

Il test di Mann-Whitney verifica se c'è una differenza significativa tra le distribuzioni delle due popolazioni. Sebbene spesso si dica che il test confronta le mediane, questo è vero solo quando le distribuzioni delle due popolazioni hanno la stessa forma generale.

Ad esempio, consideriamo le distribuzioni dei punteggi delle età mentali normalizzate per due popolazioni di bambini/e con fenilchetonuria. I soggetti con questa malattia non sono in grado di metabolizzare la proteina fenilalanina. I bambini e le bambine del primo gruppo presentano livelli medi giornalieri di fenilalanina sierica inferiori a 10,0 mg/dl, quelli/e del secondo gruppo presentano livelli medi superiori o uguali a 10,0 mg/dl. Pertanto, vorremmo confrontare i punteggi delle età mentali normalizzate per le due popolazioni di bambini/e; tuttavia, non assumiamo che tali punteggi siano normalmente distribuiti nei bambini/e con questo disturbo.

Il procedimento è analogo ai test precedenti: si parte dall'ipotesi nulla che le mediane siano uguali nelle due popolazioni e si valuta se la p è inferiore a 0,05 per rifiutare l'ipotesi nulla.

Questi test sopra riportati sono utili per la parte iniziale (confronto tra i due gruppi) del nostro studio statistico. Nella maggior parte dei casi gli studi sono impostati nel seguente modo:

- la prima parte prevede il confronto tra determinate variabili tra i due campioni del vostro studio
- la seconda parte consiste nell'analisi univariata che analizza singolarmente l'effetto di una singola variabili sull'*outcome*
- la terza consiste nell'analisi multivariata che valuta contemporaneamente le variabili per valutare il loro reale impatto sull'outcome.

*Ovviamente questa non è una regola generale, tutto dipende da come è impostato il nostro studio e da cosa vogliamo dimostrare. Pertanto, consigliamo sempre di eseguire l'analisi statistica del nostro studio di tesi sotto la supervisione del/della proprio/a tutor o del/della proprio/a relatore/relatrice.

ANALISI DI REGRESSIONI

L'analisi della regressione è una tecnica che permette di analizzare se vi è una relazione tra la variabile dipendente (Y) e una o più variabili indipendenti (X). La decisione su quale sia la variabile dipendente viene posta a priori su base biologica. Va ricordato che identificare una variabile dipendente non significa comunque determinare una relazione causa-effetto tra variabile dipendente e variabile indipendente.

Le finalità dell'analisi di regressione sono:

- esplicative: descrivere se e come y dipende da x
- predittive: predire il valore di y in base a una data x o in base a una combinazione di variabili indipendenti.

Il modello di regressione dipende dalla variabile DIPENDENTE (Y).

Tipo di regressione	**Tipo di variabile**
Lineare	Quantitativo
Logistica	Dicotomica
Di Cox	Soppravvivenza

Indipendentemente dal tipo di modello, un'analisi di regressione può essere:

- Univariata/Semplice: UNA sola variabile indipendente X
- Multivariata: più variabili indipendenti X

REGRESSIONE LINEARE SEMPLICE

L'equazione y=a+bx è l'equazione della retta.
Occorre determinare il valore di a e b a partire dai dati campionari. È così possibile applicare in modo appropriato una linea retta ad un diagramma di dispersione e stimare il valore di una variabile sulla base dei valori di un'altra.

a: intercetta della retta sull'asse y (=valore medio di y quando x=0)
b: coefficiente angolare – pendenza della retta (variazione media di y al variare unitario di x).
Se b=0, allora non vi è associazione lineare fra y e x: l'equazione della linea si riduce a y=a, cioè y è costante, uguale ad a e non cambia al variare di x.

Se esiste una relazione lineare fra le due variabili in studio, significa che una aumenta o diminuisce al variare dell'altra e che quindi la retta ha una pendenza.
Per valutare la presenza di associazione fra le due variabili, cioè un valore di $b \neq 0$, si effettua un test statistico su b.
Si valuta cioè la probabilità di osservare il valore osservato sotto l'ipotesi nulla che la retta non abbia alcuna pendenza effettiva. Si esegue un test T di Student su b, si valuta intervallo di confidenza al 95% e il p-value che viene riportato dal software

statistico. Per scartare l'ipotesi nulla è importante che il p-value sia inferiore al valore impostato convenzionalmente a 0,05 e che IC 95% non comprenda lo 0.
L'IC 95% dà una stima di quanto varia il valore nella popolazione analizzata.

E' fondamentale poi valutare R^2; ci dice quanto è buono il nostro modello e quanto vadano bene le variabili indipendenti che abbiamo scelto. È un valore concettuale. Ovviamente, se abbiamo scelto bene le variabili indipendenti, man mano che le inseriamo il coefficiente R^2 dovrebbe aumentare. R^2 ci dice quanto siamo in grado di spiegare di quello che succede alla variabile dipendente con le variabili indipendenti che abbiamo messo nel modello.
La regressione multipla è simile alla regressione lineare solo che si analizzano più variabili indipendenti contemporaneamente.

REGRESSIONE LOGISTICA

La regressione logistica è un modello statistico utilizzato quando la variabile dipendente Y è dicotomica, cioè ha solo due possibili esiti (ad esempio, presenza/assenza dell'evento in studio). Le variabili indipendenti possono essere continue, categoriali o di qualsiasi altro tipo.

Questo modello permette di esplorare come ogni variabile esplicativa influenzi la probabilità che l'evento in studio si verifichi.
Il modello di regressione logistica calcola l'odds ratio (OR) per ogni variabile indipendente.
L'OR rappresenta il rapporto tra le probabilità che l'evento si verifichi (Y=1) e che non si verifichi (Y=0) per un'unità di variazione nella variabile indipendente.
Un OR maggiore di 1 indica che l'aumento della variabile indipendente è associato a una maggiore probabilità che l'evento si verifichi; un OR inferiore a 1 indica l'opposto.

Perché un OR sia considerato statisticamente significativo:

- L'OR deve essere diverso da 1.
- L'intervallo di confidenza al 95% (IC 95%) dell'OR non deve contenere il valore 1.
- Il p-value associato deve essere inferiore al livello di significatività ($\alpha = 0,05$).

ANALISI DI SOPRAVVIVENZA

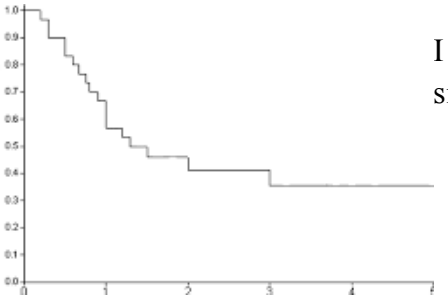
L'analisi di sopravvivenza è un insieme di tecniche statistiche utilizzate quando la variabile di interesse è il tempo che intercorre tra un'osservazione iniziale e l'insorgenza di un evento specifico (ad esempio, tempo tra la nascita e la morte, tra un trapianto e il rigetto dell'organo, o tra l'inizio di una terapia e la ricaduta della malattia). Questo periodo di tempo è chiamato "tempo di sopravvivenza

Sebbene le misurazioni dei tempi di sopravvivenza siano continue, raramente le relative distribuzioni sono normali, ma tendono ad essere asimmetriche a destra. L'analisi di questo tipo di dati si concentra sulla stima di probabilità che il/la paziente sopravviva per un determinato periodo di tempo.

Il metodo più utilizzato per calcolare la sopravvivenza è quello di **Kaplan-Meier**, anche detto metodo del prodotto limite. Si tratta di una tecnica **non parametrica** che utilizza gli esatti tempi

di sopravvivenza di ciascun soggetto di un campione. Infatti, quando analizziamo serie di dati di piccole dimensioni è possibile che alcuni intervalli non contengano alcun decesso per cui, in questi casi, è preferibile utilizzare il Metodo Kaplan-Meier.

Mediante i software statistici è possibile calcolare agevolmente la sopravvivenza. Il risultato è un grafico con la percentuale di soggetti vivi sulle ordinate e l'intervallo di tempo sulle ascisse.



I software statistici ci forniranno una tabella come quella a sinistra.

Tempo	q_t	$1-q_t$	$\hat{S}(t)$
0	0,0000	1,0000	1,0000
2	0,0833	0,9167	0,9167
3	0,0909	0,9091	0,8333
6	0,2000	0,8000	0,6667
7	0,1250	0,8750	0,5833
10	0,1429	0,8571	0,5000
15	0,3333	0,6667	0,3333
16	0,2500	0,7500	0,2500
27	0,3333	0,6667	0,1667
30	0,5000	0,5000	0,0833
32	1,0000	0,0000	0,0000

Nella prima colonna sono presenti gli intervalli di tempo, il momento esatto in cui si è verificato il decesso.
La seconda colonna contiene la percentuale di soggetti precedentemente vivi e morti in quel momento.
La terza colonna contiene la proporzione di pazienti che non sono morti/e al tempo t. Applicando il principio del prodotto della probabilità è possibile utilizzare la proporzione di soggetti che non sono ancora morti per stimare la funzione di sopravvivenza.
Ad esempio, supponiamo che la prima colonna contenga l'intervallo di sopravvivenza espresso in mesi e noi vogliamo sapere la percentuale di sopravvivenza a un anno.
In questo caso non è presente il dato a 12 mesi nella prima tabella ma è presente il dato a 10 mesi, per cui quando non è presente l'intervallo esatto si torna indietro a quello più vicino: in questo caso la sopravvivenza dei/delle pazienti a 1 anno nel nostro campione è del 50%.

LOG RANK-TEST

Il log-rank test è un test statistico utilizzato per confrontare le curve di sopravvivenza di due o più gruppi. Il suo obiettivo è determinare se esistono differenze significative nei tempi di sopravvivenza tra questi gruppi. A differenza dell'analisi dei tempi di sopravvivenza di un singolo gruppo, il log-rank test permette di confrontare la distribuzione dei tempi di sopravvivenza tra gruppi distinti, tenendo conto di eventuali dati censurati.

Se nessun gruppo include osservazioni troncate perse al follow-up, è possibile utilizzare il test di Mann-Whitney per confrontare la mediana dei tempi di sopravvivenza; se però sono presenti dati troncati è necessario utilizzare procedure diverse.

Uno dei metodi più utilizzati per testare l'ipotesi nulla che le distribuzioni dei due tempi di sopravvivenza siano uguali è una tecnica **non parametrica** nota come *log-rank test*.
È naturale aspettarsi una certa variabilità campionaria in queste stime, tuttavia la differenza nei tempi di sopravvivenza osservata nei due gruppi è maggiore rispetto a quanto si sarebbe potuto osservare per il caso?

Il log-rank test confronta il numero osservato di decessi con il numero atteso di decessi, quindi testa l'ipotesi nulla che la sopravvivenza nel gruppo A sia uguale a quella del gruppo B.

$\boldsymbol{H_0}$: $S_A(t) = S_B(t)$

I software statistici calcolano il p-value associato al log-rank test. Se questo p-value è inferiore al livello di significatività ($\alpha = 0{,}05$), si rifiuta l'ipotesi nulla e si conclude che i tempi di sopravvivenza sono significativamente diversi tra i gruppi. In altre parole, la differenza osservata tra le curve di sopravvivenza è troppo grande per essere spiegata dal caso.

Il log-rank test assume che il rapporto dei rischi tra i gruppi (hazard ratio) sia costante nel tempo. Se questa assunzione non è soddisfatta, potrebbe essere necessario utilizzare altri test come il test di Breslow o il test di Tarone-Ware.

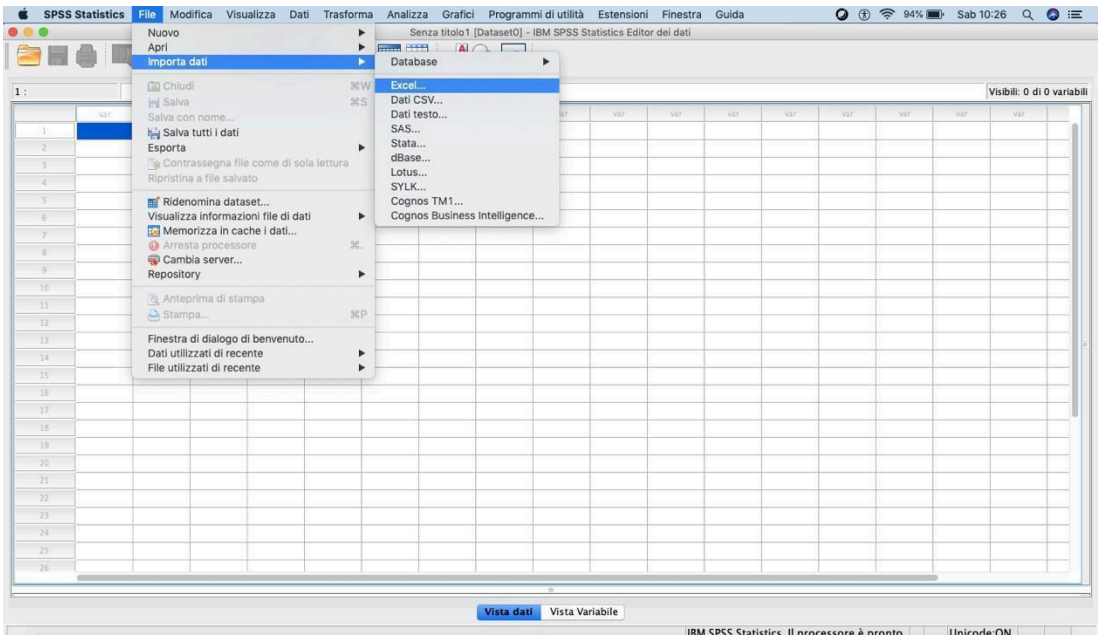
REGRESSIONE DI COX

Negli studi osservazionali, a differenza di quanto accade nei trial clinici, gli/le esposti/e ad un certo fattore di rischio possono differire dai non esposti/e per una serie di rilevanti caratteristiche cliniche (fattori confondenti), che possono alterare il rapporto tra l'esposizione oggetto dell'indagine da parte del/della ricercatore/ricercatrice e l'incidenza di una specifica malattia o esito clinico. Il modello di Cox è una particolare **tecnica di regressione multipla** che ci permette di analizzare il rapporto tra un fattore di rischio (per esempio il fumo) e l'incidenza di un determinato esito clinico (per esempio l'infarto del miocardio), correggendo per uno o più fattori di confondimento (quali l'obesità e l'ipertensione); infatti si presta per un'analisi multivariata.
La regressione di Cox si usa negli **studi di coorte**, sia prospettici che retrospettivi. Nella regressione di Cox, la variabile dipendente è il tasso di incidenza di un determinato evento, cioè il numero di eventi per persona - tempo.

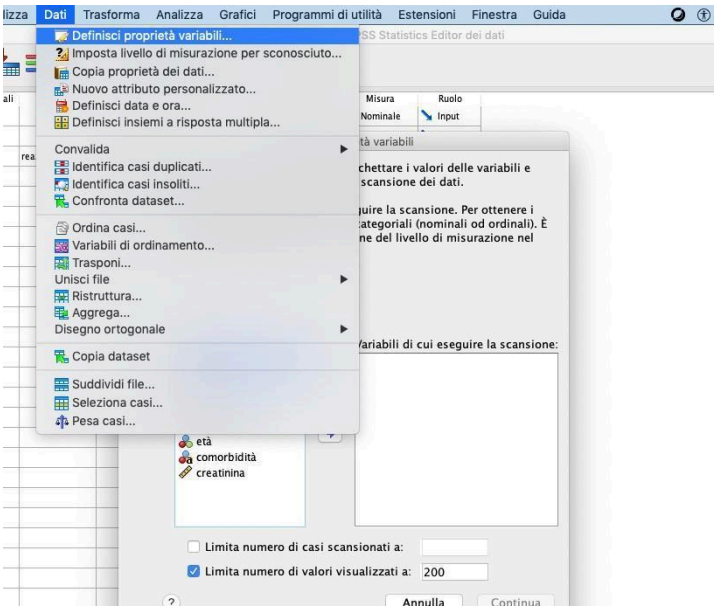
Ht=H0t*expb*Xi

Dove "**Ht**" è il tasso di incidenza dell'evento (stimato dal modello) al tempo **t**, "**H0t**" rappresenta il rischio di base (cioè il tasso di incidenza) dell'evento quando il fattore di rischio è assente), "**b**" è il coefficiente di regressione e "**Xi**" è il fattore di rischio.
Per convenienza di calcolo si utilizzano i logaritmi:
ln(Ht)=H0t+b*Xi
Il coefficiente di regressione (b) indica di quanto aumenta in media il logaritmo naturale del tasso di incidenza dell'evento negli esposti rispetto ai non esposti.

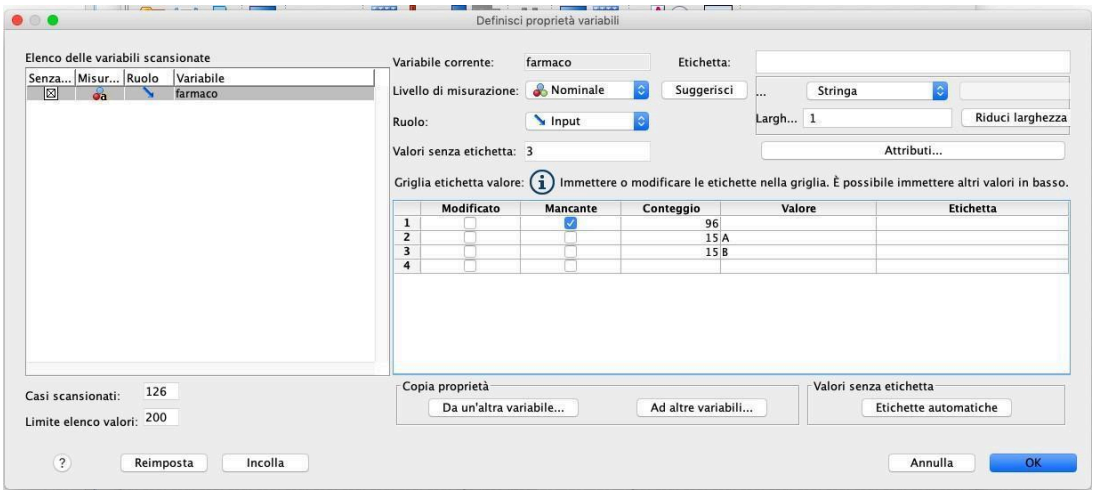
La regressione di Cox permette di ricavare l'***hazard ratio***, ovvero il rischio relativo tra gli/le esposti/e e i/le non esposti/e a un determinato fattore di rischio.
Attualmente, i moderni software statistici permettono di calcolare agevolmente, mediante la regressione di Cox, gli hazard ratio e le *p* che permettono di considerare la significatività statistica.



È molto importante, una volta creato il database, adattarlo ad SPSS: tutti i dati mancanti devono avere la casella vuota e non scritto mancante o n/a (non available) perché il tal caso SPSS non riuscirebbe a leggere l'informazione correttamente.
Una volta che avrete aperto il database con SPSS dovrete definire le proprietà delle vostre variabili. Per fare questo selezionare: dati definisci proprietà variabili e a questo punto selezionate la variabile di vostro interesse trascinandola nella casella.

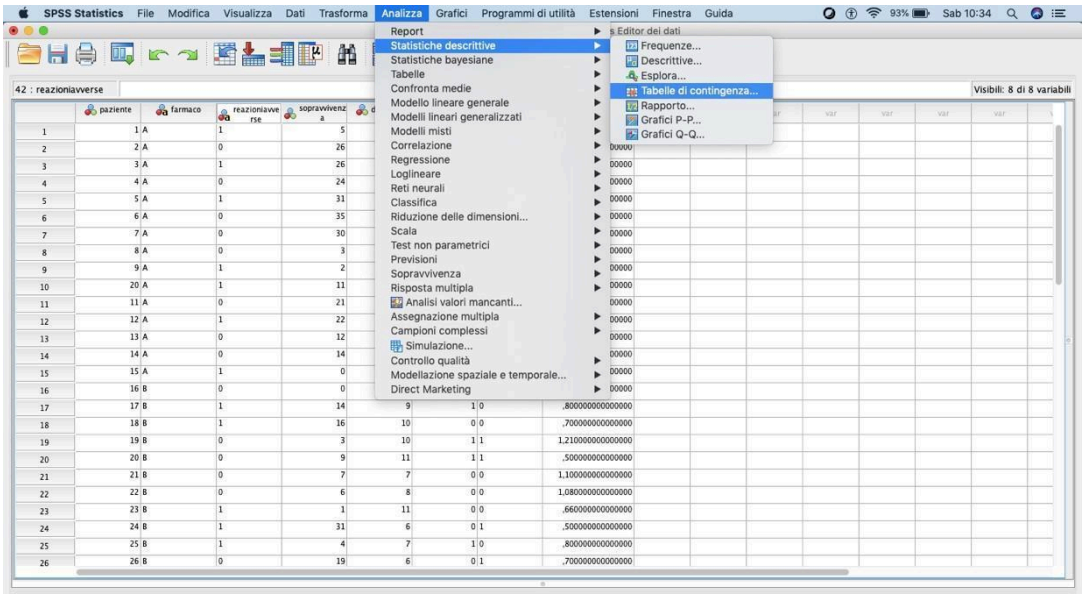


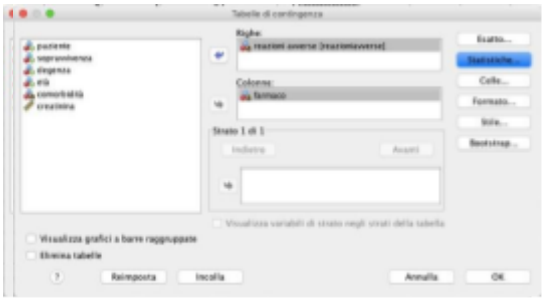
Una volta selezionata la variabile dovrete stabilire se è una variabile numerica o una variabile stringa, ovvero nominale. Inoltre, le caselle vuote dovranno essere indicate come "mancanti". Questa parte iniziale è fondamentale per eseguire i test statistici in modo corretto.

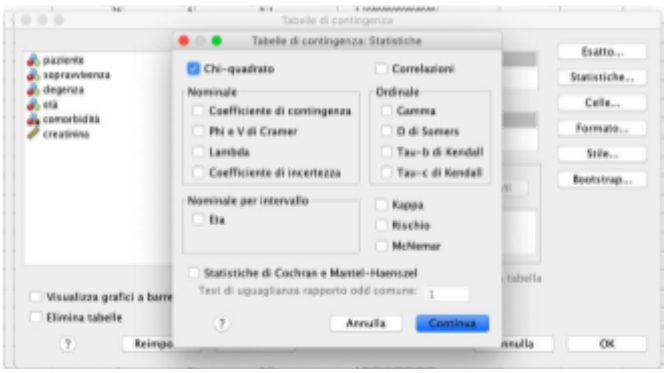


Dopodiché, cliccando su "analizza" è possibile eseguire numerosi test e calcolare automaticamente molti parametri statistici.
Ad esempio, quando si confrontano due gruppi per quanto riguarda le variabili dicotomiche, è possibile utilizzare il test del "Chi quadro" con le tabelle di contingenza.

Per farlo, cliccare su: analizza-> statistiche descrittive-> tabelle di contingenza.







Immaginiamo di confrontare il numero di reazioni avverse in due gruppi: nel primo gruppo è stato assegnato il farmaco A, nel secondo il farmaco B.

Pertanto, bisogna selezionare le variabili **“reazioni avverse”** e **“farmaco”.**

Successivamente è necessario cliccare su “statistiche” e selezionare il **test Chi quadro.**

Vi comparirà dunque una tabella di contingenza come quella a fianco e il livello di significatività statistica (freccia rossa). In questo caso p > 0,05% per cui non è possibile rifiutare l'ipotesi nulla.

Riepilogo elaborazione casi

	Casi					
	Valido		Mancante		Totale	
	N	Percentuale	N	Percentuale	N	Percentuale
reazioni avverse * farmaco	30	100,0%	0	0,0%	30	100,0%

Tavola di contingenza reazioni avverse * farmaco

Conteggio

		farmaco		
		A	B	Totale
reazioni avverse	0	8	10	18
	1	7	5	12
Totale		15	15	30

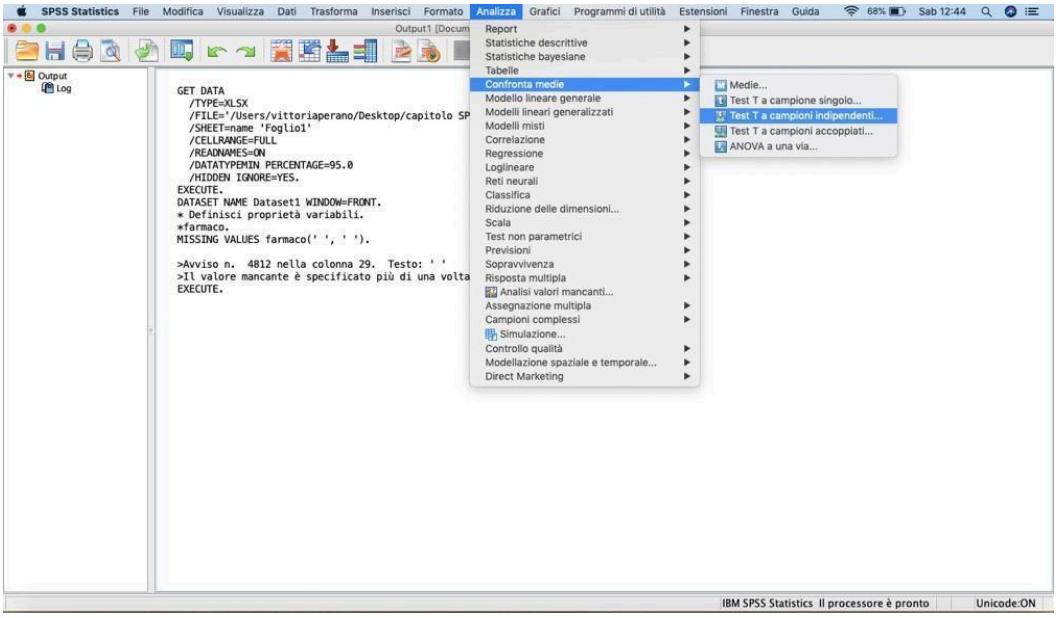
Test del chi-quadrato

	Valore	gl	Significatività asintotica (bilaterale)	Sign. esatta (bilaterale)	Sign. esatta (unilaterale)
Chi-quadrato di Pearson	,556[a]	1	,456		
Correzione di continuità[b]	,139	1	,709		
Rapporto di verosimiglianza	,558	1	,455		
Test esatto di Fisher				,710	,355
N di casi validi	30				

a. 0 celle (0,0%) hanno un conteggio previsto inferiore a 5. Il conteggio previsto minimo è 6,00.

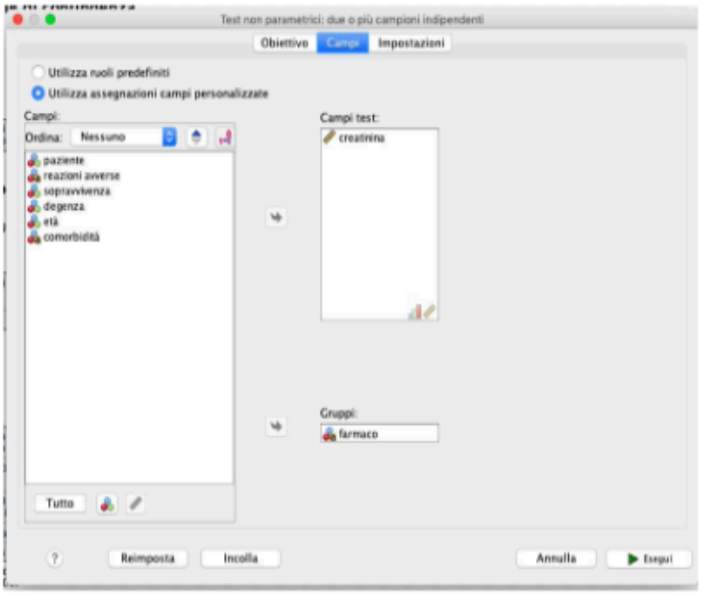
b. Calcolato solo per una tabella 2x2

Invece, per confrontare due gruppi sulla base di variabili continue si utilizzano altri test come il test T in caso di variabili distribuite normalmente. Per farlo selezionare: analizza confronta medie test T a campioni indipendenti.



Tuttavia, nella maggior parte dei casi, le variabili in studio non sono distribuite normalmente. Ad esempio, nel nostro studio che considera i/le pazienti ospedalizzati/e, vogliamo analizzare se il farmaco A faccia aumentare maggiormente i livelli di creatinina rispetto al farmaco B. Per farlo selezionare: analizza -> test non parametrici-> campioni indipendenti rispetto al farmaco B.

Successivamente, selezionare **la variabile per cui si vogliono confrontare i due gruppi.**

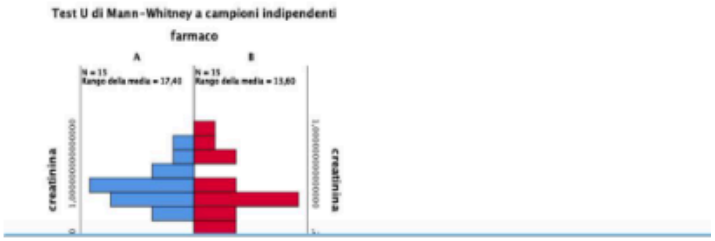


Test U di Mann-Whitney a campioni indipendenti

creatinina tra farmaco

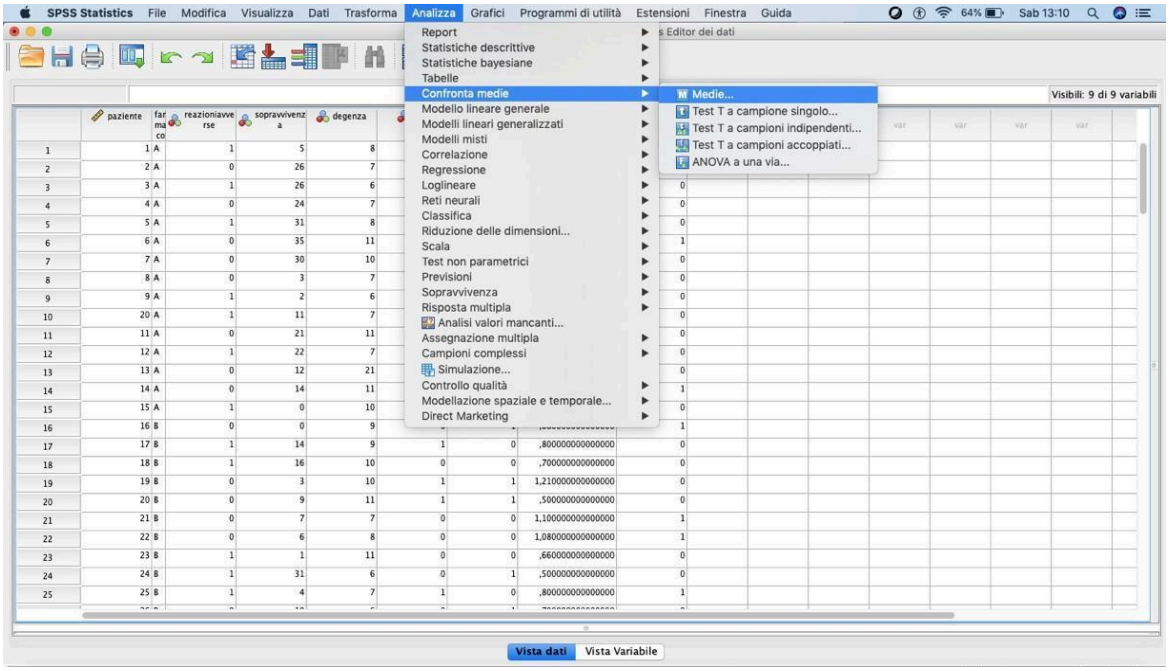
Riepilogo test U di Mann-Whitney a campioni indipendenti

Numero totale di casi	30
U di Mann-Whitney	84,000
W di Wilcoxon	204,000
Statistica del test	84,000
Errore standard	24,031
Statistica del test standardizzata	-1,186
Sign. asintotica (test a 2 vie)	,236
Sign. esatta (test a 2 vie)	,250

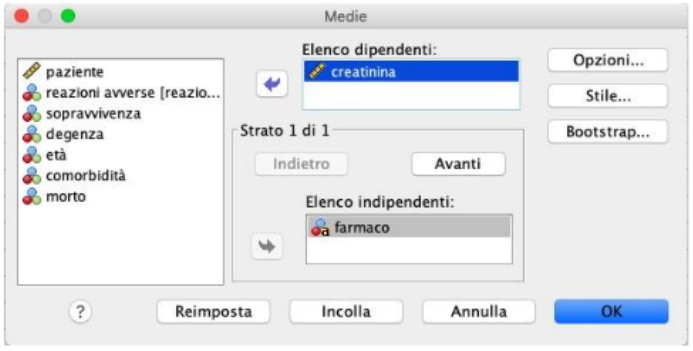


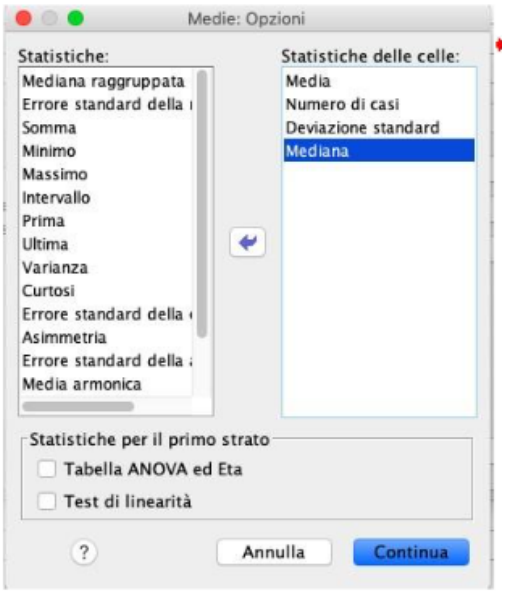
In questo caso “creatinina” e la inserisco nella casella “gruppi”, come indicato nella figura.
Vi comparirà una tabella come quella in esame, dove potrete valutare il livello di significatività statistica e decidere se rifiutare o accettare l’ipotesi nulla (freccia rossa).

Per ottenere la **mediana** della variabile nei gruppi da riportare nel vostro studio selezionare: analizza -> confronta medie -> medie.



Successivamente bisogna stabilire la **variabile dipendente**, ovvero la variabile per cui si vogliono confrontare i due gruppi (nel nostro caso la creatinina) e la variabile indipendente, ovvero il farmaco. selezionare dunque opzioni e stabilire cosa si vuole analizzare. Si ottiene così una tabella con tutti i parametri statistici di interesse.





Medie

[Dataset1]

Riepilogo elaborazione casi

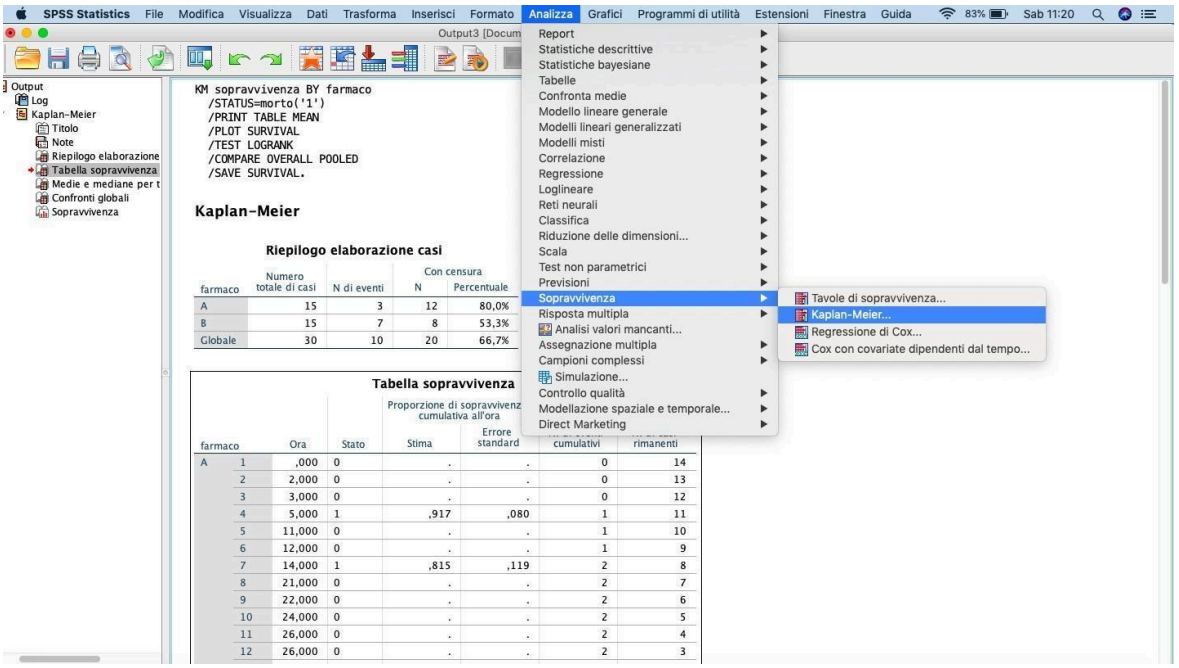
	Casi					
	Incluso		Escluso		Totale	
	N	Percentuale	N	Percentuale	N	Percentuale
creatinina * farmaco	30	23,8%	96	76,2%	126	100,0%

Report

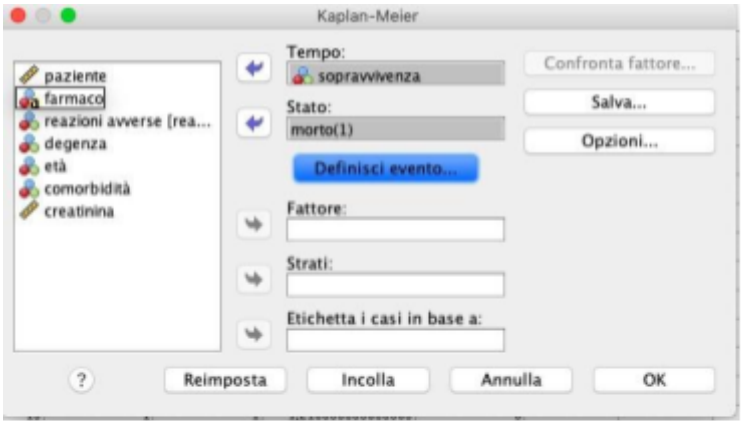
creatinina

farmaco	Media	N	Deviazione std.	Mediana	Minimo	Massimo
A	,814666667	15	,133034188	,820000000	,600000000	1,10000000
B	,788666667	15	,213570553	,700000000	,500000000	1,21000000
Totale	,801666667	30	,175324002	,785000000	,500000000	1,21000000

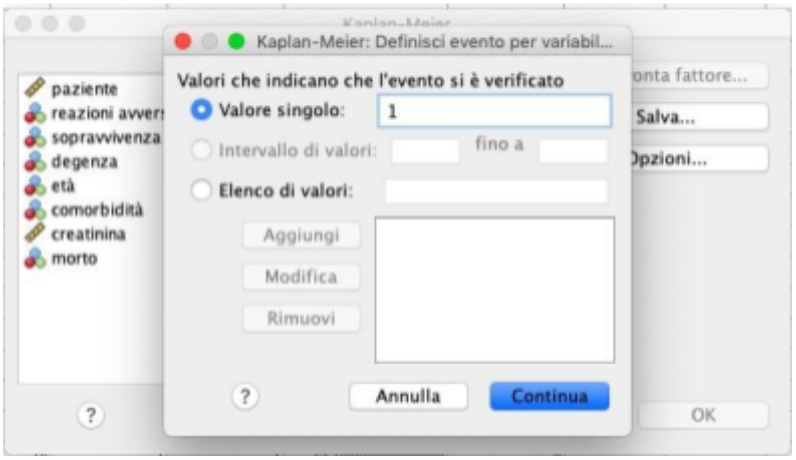
Per analizzare la sopravvivenza è possibile utilizzare il metodo Kaplan-Meier. Per farlo selezionare: analizza -> sopravvivenza -> Kaplan-Meier.



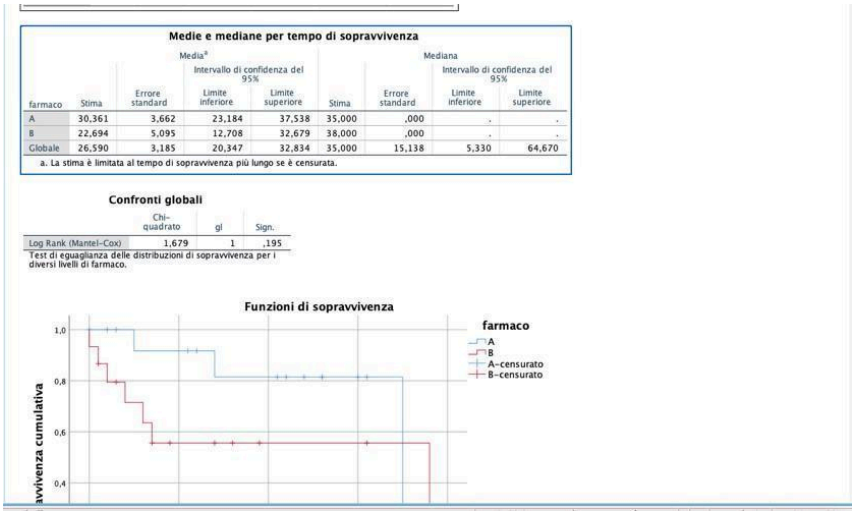
A questo punto è necessario definire la variabile **tempo** e lo **stato** (ovvero morto/a o vivo/a oppure libero/a da malattia o con recidiva tumorale ad es.). Selezionare dunque "definisci evento" e definire l'evento. Ad esempio, nel nostro database i soggetti deceduti erano stati indicati come 1, mentre i soggetti vivi come 0. Pertanto, in questo caso l'evento è definito come 1.



Nella casella "fattore" inserire la variabile di cui si vuole valutare l'impatto sulla sopravvivenza. Nel nostro caso è il farmaco utilizzato.



Si otterrà dunque un grafico e le relative mediane sopravvivenza. Inoltre, viene fornita la ***p*** (su SPSS, indicata come "Sign.") che permette di rifiutare o accettare l'ipotesi nulla.



Medie e mediane per tempo di sopravvivenza

farmaco	Media[a] Stima	Errore standard	Intervallo di confidenza del 95% Limite inferiore	Limite superiore	Mediana Stima	Errore standard	Intervallo di confidenza del 95% Limite inferiore	Limite superiore
A	30,361	3,662	23,184	37,538	35,000	,000	.	.
B	22,694	5,095	12,708	32,679	38,000	,000	.	.
Globale	26,590	3,185	20,347	32,834	35,000	15,138	5,330	64,670

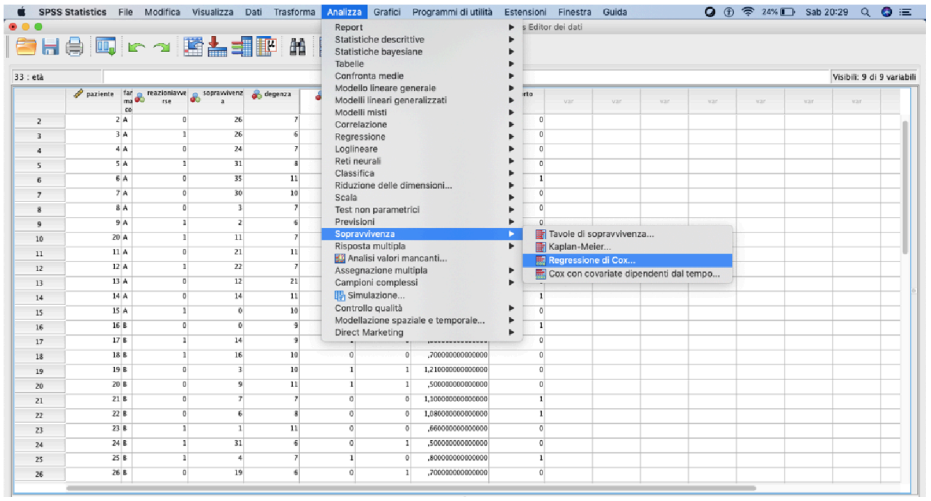
a. La stima è limitata al tempo di sopravvivenza più lungo se è censurata.

Confronti globali

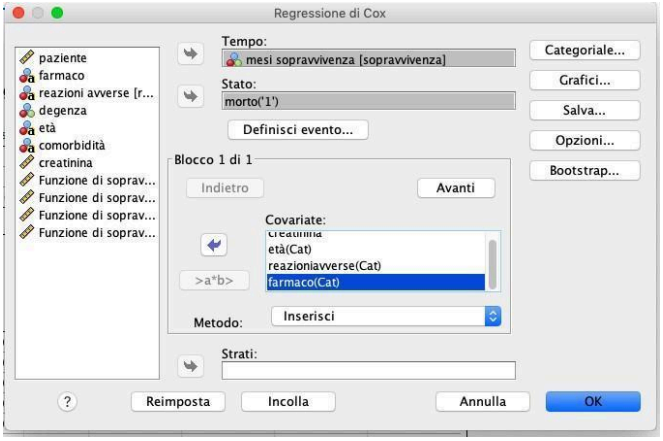
	Chi-quadrato	gl	Sign.
Log Rank (Mantel-Cox)	1,679	1	,195

Test di eguaglianza delle distribuzioni di sopravvivenza per i diversi livelli di farmaco.

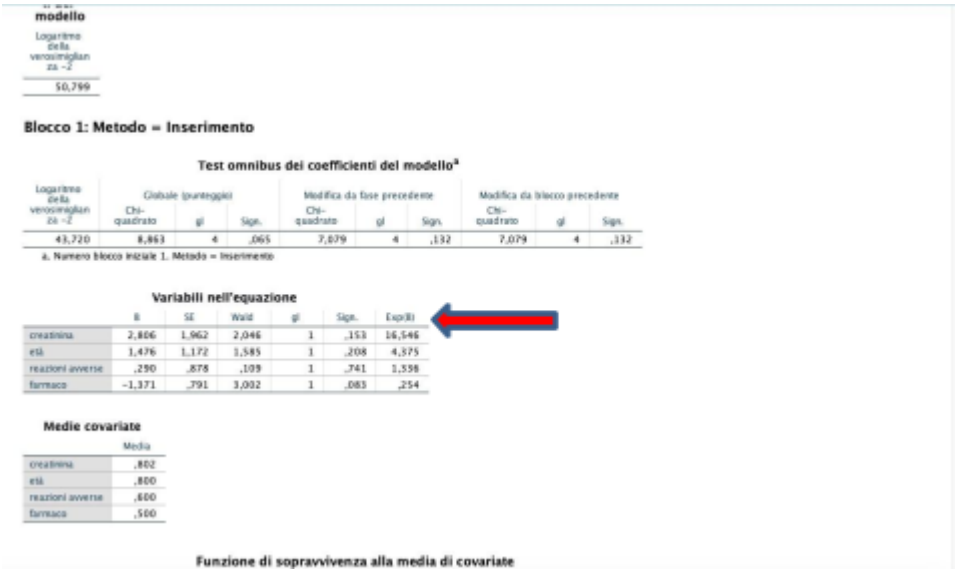
Per un'analisi di tipo multivariato è possibile eseguire una **Regressione di Cox selezionando**: analizza-> sopravvivenza -> regressione di Cox.



A questo punto è necessario selezionare la variabile che indica il **tempo** (nel nostro caso la sopravvivenza) e **l'evento** (vivo/a o morto/a) come viene fatto per l'analisi di Kaplan-Meier. Nella **casella covariate** bisogna inserire le variabili da analizzare.



Si ottiene dunque una tabella con l'odd ratio (freccia rossa, indicato come Exp(B)) e la p di significatività statistica.



modello

Logaritmo della verosimiglianza -2
50,799

Blocco 1: Metodo = Inserimento

Test omnibus dei coefficienti del modello[a]

	Globale (punteggio)			Modifica da fase precedente			Modifica da blocco precedente		
Logaritmo della verosimiglianza -2	Chi-quadrato	gl	Sign.	Chi-quadrato	gl	Sign.	Chi-quadrato	gl	Sign.
43,720	8,863	4	,065	7,079	4	,132	7,079	4	,132

a. Numero blocco iniziale 1. Metodo = Inserimento

Variabili nell'equazione

	B	SE	Wald	gl	Sign.	Exp(B)
creatinina	2,806	1,962	2,046	1	,153	16,546
età	1,476	1,172	1,585	1	,208	4,375
reazioni avverse	,290	,878	,109	1	,741	1,336
farmaco	-1,371	,791	3,002	1	,083	,254

Medie covariate

	Media
creatinina	,802
età	,800
reazioni avverse	,600
farmaco	,500

Funzione di sopravvivenza alla media di covariate