MasterGi www.hackmed.org

# Statistica Medica

Premessa: il seguente testo cerca di riassumere e rendere in forma comprensibile ai non esperti in matematica e statistica le nozioni e le procedure necessarie a svolgere gli esercizi dell'esame di Statistica Medica.

Verranno presi in esame:

- Il calcolo di Media aritmetica, geometrica, armonica; moda e mediana di un campione, coefficiente di variazione e limiti fiduciali della media aritmetica campionaria (intervallo di confidenza).

- L'applicazione dei test statistici: test di Student (con approssimazione di Cochran in caso di varianze non omogenee); test del t. paired (ad una o a due code); test del Chi quadrato di Paerson (con la correzione di Yates quando vi è un solo grado di libertà); test di Analisi della varianza (ad un criterio); test sulla regressione lineare semplice.

All'interno del programma di Statistica Medica sono anche presenti altri due test statistici, ovvero il test di analisi della varianza a 2 criteri ed il test sulla correlazione lineare semplice, tuttavia non ho inserito la sintesi delle procedure di questi test perché a quanto ne so è molto improbabile che possano capitare all'esame di Statistica Medica degli esercizi in cui sia necessario applicare questi ultimi 2 test statistici; in ogni caso se dovesse essere necessario si potrebbero sempre aggiungere in seguito le parti riguardanti questi test statistici.

Il consiglio che posso dare a tutti quelli che devono affrontare questo esame è quello di cercare di imparare a risolvere gli esercizi facendo molti esercizi di prova, ovvero di prendere gli esercizi dal libro di Statistica Medica oppure da Internet e cercare di risolverli applicando le procedure che ho riassunto in questo file, noterete che non è difficile come sembra all'inizio…

Non essendo io un matematico né uno statistico, e non essendo stati questi appunti revisionati da un matematico né da uno statistico, ritengo opportuno ricordare che possono essere presenti degli errori in qualsiasi parte del testo; tuttavia questo testo è basato sugli appunti che ho preso frequentando le lezioni di Statistica Medica, studiando sul libro di Statistica Medica ed andando a cercare spesso delucidazioni in Internet per le parti che mi risultavano poco chiare… applicando queste procedure sono riuscito a superare l'esame!!

# Media, moda e mediana

- Media aritmetica: Somma di tutti i valori di un gruppo diviso la numerosità del gruppo stesso.

Ad esempio abbiamo questi numeri: 1 ; 4 ; 6 ; 5

La media aritmetica sarà: [(1 + 4 + 6 + 5) : 4] = (16 : 4) = 4

- Media Geometrica: La radice ennesima (elevata alla numerosità) delle singole modalità elevate alle rispettive frequenze.

Ad esempio:

| Modalità (Evento) | Frequenza |
|---|---|
| 3 | 2 |
| 4 | 3 |
| 1 | 2 |

$$\sqrt[7]{3^2 \times 4^3 \times 1^2} = \sqrt[7]{576}$$

=

Per calcolare la radice elevata alla ennesima (numerosità totale) potenza devo elevare quel numero stesso al reciproco della numerosità totale:

$$\sqrt[7]{576} = 576^{1/7}$$

Per fare ciò devo scrivere sulla calcolatrice:

$$576 \wedge (1:7) = 2{,}4$$

- Media Armonica (Mh): Il numero totale delle osservazioni (numerosità totale) diviso la sommatoria del reciproco delle singole modalità per le rispettive frequenze:

$$Mh = \frac{n}{\left(\frac{1}{x_1}\right) \times n_1 + \left(\frac{1}{x_2}\right) \times n_2 \ldots}$$

- Moda: La modalità (il valore) che presenta la massima frequenza; possono essere presenti più di un valore di moda (distribuzioni bimodali, trimodali… plurimodali).

Ad esempio abbiamo questi valori: 4; 3; 5; 2; 6; 8; 1; 4

Per calcolare la moda mi sarà utile riscrivere tutti i numeri in ordine crescente:

1; 2; 3; 4; 4; 5; 6; 8

La moda sarà il numero 4 perché è il valore che si presenta il maggior numero di volte.

Facciamo ora un esempio di moda plurimodale:

Abbiamo questi valori: 2; 4; 5; 9; 7; 6; 7; 5; 7; 2; 1; 5; 4; 2

Ora riscriverò tutti i valori in ordine crescente:

1; 2; 2; 2; 4; 4; 5; 5; 5; 6; 7; 7; 7; 9

In questo esempio avremo tre mode: 2, 5, 7, perché questi valori hanno la frequenza massima riscontrata, ovvero ognuno di questi valori si presenta tre volte.

- Mediana: Di una distribuzione di modalità (valori) ordinati in modo crescente o decrescente, è data dalla modalità centrale nel caso in cui la numerosità totale sia un numero dispari o dalla semisomma delle modalità centrali nel caso la numerosità totale sia un numero pari.

[ Semisomma di A e B = (A + B) / 2 ]

Facciamo un esempio in cui la numerosità totale è un numero dispari;

Abbiamo questi valori: 5; 3; 7; 7; 8; 4; 9

Per calcolare la mediana mi sarà utile riscrivere tutti i valori in ordine crescente:

3; 4; 5; 7; 7; 8 ; 9

In questo esempio la mediana sarà 7

Facciamo un esempio in cui la numerosità totale è un numero pari;

Abbiamo questi valori: 5; 3; 7; 2; 7; 8; 4; 5;

Per calcolare la mediana mi sarà utile riscrivere tutti i valori in ordine crescente:

2; 3; 4; 5; 5; 7; 7; 8

In questo esempio, essendo la numerosità totale un numero pari, la mediana sarà data dalla semisomma delle modalità centrali: [(5 + 5) : 2] = 5

- Approssimazioni per arrotondamento:

(esempi)

2,342 → 2,34

2,347 → 2,35

2,3453 → 2,35

2,3450 → 2,34 o 2,35

- Potenze negative

$$10^{-n} = 1 : 10^{n}$$

---

# Devianza, varianza, deviazione standard, codevianza e coefficiente di variazione

Devianza, varianza, deviazione standard e codevianza sono degli "indici di dispersione" che servono per calcolare: i limiti fiduciali della media aritmetica campionaria (intervallo di confidenza) e per lo svolgimento dei test statistici: test di Student, test del t. paired, test del Chi quadrato di Paerson, test di Analisi della varianza e test sulla regressione lineare semplice.

- Devianza: In un gruppo di valori è la somma dei quadrati degli scarti dalla media aritmetica di ogni valore per le rispettive frequenze (ogni valore va sottratto alla media aritmetica e poi moltiplicato per sé stesso; poi tutti i valori vanno sommati).

(Cosa significa lo scarto di un valore dalla media aritmetica? Significa semplicemente sottrarre quel valore alla media aritmetica.)

Ad esempio abbiamo un gruppo di valori: 2,8 / 0 / 0,5 / 6,2 / 3,6 / 3,8 / -2,2

Innanzi tutto devo calcolare la media aritmetica di questo gruppo di valori:

[2,8 + 0 + 0,5 + 6,2 + 3,6 + 3,8 + (-2,2)] : 7 = 2,1

Ora faccio la somma dei quadrati degli scarti della media aritmetica:

$(2,8 – 2,1)^2 + (0 – 2,1)^2 + (0,5 – 2,1)^2 + (6,2 – 2,1)^2 + (3,6 – 2,1)^2 + (3,8 – 2,1)^2 + (-2,2 – 2,1)^2 =$
0,49 + 4,41 + 2,56 + 16,81 + 2,25 + 2,89 + 18,49 = 47,9

- Varianza: La devianza fratto i gradi di libertà (numerosità meno 1 se è un unico gruppo; numerosità meno 2 se sono 2 gruppi; ovvero numerosità meno numero di gruppi).

Riprendendo l'esempio precedente abbiamo questo gruppo di valori:

2,8 / 0 / 0,5 / 6,2 / 3,6 / 3,8 / -2,2

In questo caso la devianza (come abbiamo calcolato sopra) è 47,9 quindi la varianza sarà data dalla devianza diviso i gradi di libertà del gruppo, in questo caso i gradi di libertà saranno 6 (numerosità meno 1), quindi la varianza sarà 47,9 : 6 = 7,98

- Deviazione standard: La radice quadrata della varianza.

Riprendendo ancora l'esempio precedente, abbiamo questo gruppo di valori:

2,8 / 0 / 0,5 / 6,2 / 3,6 / 3,8 / -2,2

In questo caso la varianza è 7,98 quindi la deviazione standard si calcolerà semplicemente eseguendo √7,98 = 2,82

- Codevianza: Per calcolare la codevianza devo avere una tabella con 2 variabili (x ed y) in cui le 2 variabili siano appaiate appartenendo allo stesso soggetto (oggetto). Ad esempio in un campione di 4 bambini sono stati misurati il peso e la statura

| | VARIABILE X<br>PESO (Kg) | VARIABILE Y<br>Statura (Cm) |
|---|---|---|
| Bambino 1 | 20,5 | 100 |
| Bambino 2 | 21,8 | 103 |
| Bambino 3 | 23,6 | 105 |
| Bambino 4 | 22,3 | 106 |

Ora devo trovare la media aritmetica della variabile x e quella della variabile y

Media aritmetica di x = 22,5

Media aritmetica di y = 103,5

Ora devo ricompilare la tabella di prima ma inserendo per ogni valore di x ed y il suo valore di scarto dalla corrispettiva media aritmetica, quindi:

x → 20,5 – 22,5 = -2 // 21,8 – 22,5 = -0,7 // 23,6 – 22,5 = 1,1 // 22,3 – 22,5 = -0,2

y → 100 – 103,5 = -3,5 // 103 – 103,5 = -0,5 // 105 – 103,5 = 1,5 // 106 – 103,5 = 2,5

Ora potrò ricompilare la tabella:

| | VARIABILE X PESO (Kg) | VARIABILE Y Statura (Cm) |
|---|---|---|
| Bambino 1 | -2 | -3,5 |
| Bambino 2 | -0,7 | -0,5 |
| Bambino 3 | 1,1 | 1,5 |
| Bambino 4 | -0,2 | 2,5 |

Ora devo fare il prodotto tra lo scarto dalla media di x ed il corrispettivo scarto dalla media di y:

→ (-2) × (-3,5) = 7

→ (-0,7) × (-0,5) = 0,35

→ 1,1 × 1,5 = 1,65

→ (-0,2) × 2,5 = -0,5

Ora devo sommare quei valori che ricavo dal prodotto tra lo scarto dalla media di x ed il corrispettivo scarto dalla media di y; il risultato della sommatoria di questi valori sarà la CODEVIANZA

In questo caso: 7 + 3,5 + 1,65 + (-0,5) = 8,5

- Coefficiente di variazione (C.V.):

$$\frac{\text{Deviazione standard}}{\text{Media aritmetica}} \times 100$$

# Limiti fiduciali della media aritmetica campionaria

## (Intervallo di confidenza)

I limiti fiduciali della media aritmetica campionaria ci forniscono un intervallo di valori all'interno del quale con una probabilità data (del 90% 95% o del 99%) si troverà il valore della media aritmetica dell'intero universo statistico; ovvero estrapolato un campione da una popolazione intera (in assenza di errori sistematici), noi calcoleremo la media aritmetica di quel campione per un dato parametro; non è però detto che la media aritmetica di quel campione per quel dato parametro (ad esempio temperatura corporea) sia uguale alla media aritmetica dell'intera popolazione da cui il campione è stato preso, per quello stesso parametro (temperatura corporea).

Tramite il calcolo dei limiti fiduciali, noi possiamo trovare un intervallo di valori situati "ai lati" della media aritmetica campionaria all'interno del quale noi avremo, con una probabilità data (del 90% 95% oppure del 99% solitamente) il valore della media aritmetica dell'intera popolazione da cui il campione è stato preso.

La formula per il calcolo dei limiti fiduciali della media aritmetica campionaria è:

$$\mu \pm \left[ t \times (\text{dev. stand.}) : \sqrt{\text{numerosità}} \right]$$

$\mu$ = Media aritmetica campionaria

t = Valore ricavato dalla tabella del t di Student per un dato valore di $\alpha$ con i gradi di libertà corrispondenti al valore della numerosità del gruppo (del campione) meno 1

dev. stand. = è la deviazione standard; si trova calcolando la devianza, poi la varianza ed infine facendo la radice quadrata della varianza (per una spiegazione più esaustiva vedi sopra)

$$(\text{dev. stand.}) : \sqrt{\text{numerosità}}$$

→ è chiamato l'ERRORE STANDARD

Valori usuali di α:

α = 0,01 => 99%

α = 0,05 => 95%

α = 0,1 => 90%

---

# Test t di Student

## (Il confronto tra due medie)

Serve quando bisogna sapere se la differenza tra due medie sia significativa ossia se si possa affermare che la differenza non sia dovuta al caso ma esista invece una reale diversità tra le medie delle due popolazioni da cui i campioni stessi derivano.


Se il valore che risulta dal t di Student è inferiore al valore critico => DIFFERENZA NON SIGNIFICATIVA => CIOE' CASUALE (Il valore critico è quel valore che trovo nella tabella dei valori del t di Student con $\alpha \leq 0,05$ o $\alpha \leq 0,01$ ed i gradi di libertà corrispondenti ai gradi di libertà totali => numerosità – 2).

Se il valore che risulta dal t di Student è superiore al valore critico => DIFFERENZA SIGNIFICATIVA => CIOE' NON CASUALE.

Svolgimento dell'esercizio

Ho due gruppi di numeri (i valori non sono appaiati, ovvero non appartengono allo stesso soggetto, oggetto…) devo calcolare per ognuno: la media aritmetica, la devianza, la varianza e la deviazione standard (Il calcolo di media aritmetica, devianza, varianza, deviazione standard è stato spiegato precedentemente);

poi applico la formula del t di Student

$$t = \frac{X_A - X_B}{\sqrt{\frac{dev_A + dev_B}{n_A + n_B - 2}}} \times \sqrt{\frac{n_A \times n_B}{n_A + n_B}}$$

$X_A$ = Media aritmetica di A

$X_B$ = Media aritmetica di B

$dev_A$ = devianza di A

$dev_B$ = devianza di B

$n_A$ = numerosità di A

$n_B$ = numerosità di B

$n_A + n_B - 2$ = gradi di libertà totali

---

Studio dell'omogeneità delle varianze in un t di Student

Ogni volta che faccio un t di Student devo verificare l'omogeneità delle varianze:

Devo dividere la varianza maggiore per quella minore; così ottengo un valore, ora devo confrontare questo valore ad un altro valore che trovo sulla tabella ANOVA. Il valore ANOVA va cercato sulla tabella $\alpha \leq 0,05$ quando il t di Student va eseguito per significatività $\alpha \leq 0,05$ e viceversa sulla tabella ANOVA $\alpha \leq 0,01$ quando ho eseguito il t di Student con $\alpha \leq 0,01$. [ La tabella ANOVA è la tabella dei punti critici della distribuzione F (rapporto tra varianze)]. Il valore tabulare lo trovo intersecando il valore dei gradi di libertà del numeratore che è dato dal valore dei gradi di libertà del gruppo con la numerosità maggiore con il valore dei gradi di libertà del denominatore che è dato dal valore dei gradi di libertà del gruppo con la numerosità minore. Il valore che si interseca in questo modo va confrontato con quello precedente (varianza maggiore diviso varianza minore) e se il valore tabulare (quello trovato sulla tabella ANOVA) è superiore => VARIANZE OMOGENEE!! Se invece il valore tabulare è inferiore => LE VARIANZE NON SONO OMOGENEE, così dovrò applicare l'approssimazione di Cochran!

## Approssimazione di Cochran

Quando le varianze NON SONO OMOGENEE dovrò confrontare il valore del t (ricavato con il test di Student) non più con i valori tabulari semplici => valore critico dato da a $\alpha \leq 0,05$ oppure $\alpha \leq 0,01$ e gradi di libertà totali (numerosità totale – 2); ma il valore a cui dovrò confrontare il t (del t di Student) è il t' che ricaverò dalla formula dell'approssimazione di Cochran

$$t' = \frac{\left(t_A \times \frac{VAR_A}{n_A}\right) + \left(t_B \times \frac{VAR_B}{n_B}\right)}{\frac{VAR_A}{n_A} + \frac{VAR_B}{n_B}}$$

$t_A$ e $t_B$ = Sono i valori tabulari ricavati dalla tabella dei punti critici del t di Student e non sono altro che il valore che deriva dall'intersecazione della colonna $\alpha \leq 0,05$ oppure $\alpha \leq 0,01$(a seconda del mio $\alpha$...) e dei gradi di libertà del gruppo stesso (ad esempio $t_A$ per $\alpha \leq 0,05$ con il gruppo A avente 6 gradi di libertà è: 2,45)

$VAR_A$ = Varianza di A

$VAR_B$ = Varianza di B

$n_A$ = numerosità di A

$n_B$ = numerosità di B

Ora se t sarà maggiore di t' => DIFFERENZA SIGNIFICATIVA

Se invece t è minore di t' => DIFFERENZA NON SIGNIFICATIVA (CASUALE)

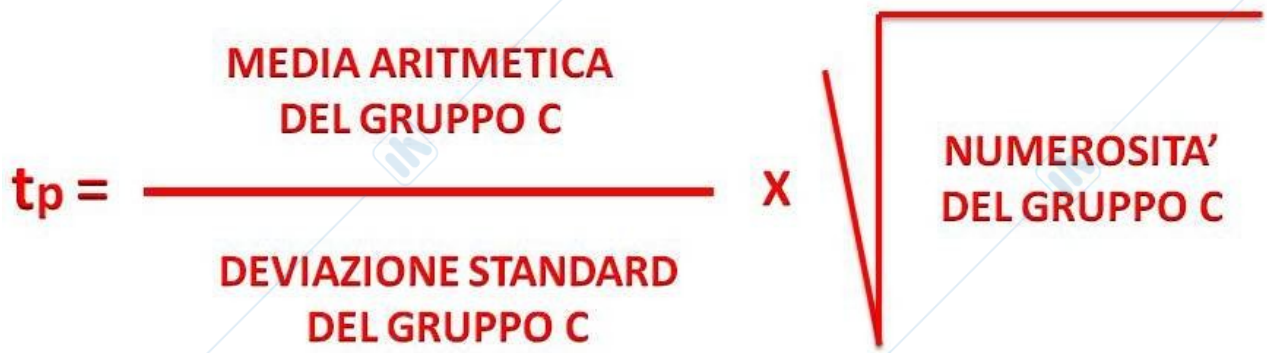
# T. Paired

## (Significatività della differenza tra 2 medie con dati non indipententi)

Test statistico utile a confrontare 2 medie quando i dati non sono indipendenti, ovvero abbiamo per esempio 10 soggetti a cui viene misurata la pressione arteriosa prima e dopo la somministrazione di un farmaco e noi vogliamo sapere se le medie sono significativamente diverse, quindi abbiamo almeno 2 valori derivanti dallo stesso soggetto (o oggetto…) per ogni soggetto (o oggetto…)

Per fare l'esercizio sul t paired devo calcolare per ogni soggetto (oggetto, elemento…) la differenza tra il 1° valore ed il 2° valore (chiamerò il gruppo dei valori derivanti dalla sottrazione tra il 1° ed il 2° valore di ogni elemento => GRUPPO C). Dovrò calcolare la MEDIA ARITMETICA, LA DEVIANZA, LA VARIANZA E LA DEVIAZIONE STANDARD DEL GRUPPO C.

Ora potrò applicare la formula del t paired




Troverò così un certo valore di $t_p$, ora dovrò confrontare questo valore con 2 punti critici diversi a seconda che il test sia ad una coda oppure a 2 code, i gradi di libertà saranno dati dalla numerosità del gruppo C - 1.

Il TEST E' AD UNA CODA quando la media di uno dei due gruppi può essere ESCLUSIVAMENTE MAGGIORE oppure ESCLUSIVAMENTE MINORE della media dell'altro gruppo ovvero la media di un gruppo non può essere sia maggiore che minore rispetto alla media dell'altro gruppo!! In questo caso dovrò cercare il punto critico per esempio di un test con $\alpha \leq 0,05$, nella tabella del t di Student con $\alpha \leq 0,10$ ovvero $\alpha$ moltiplicato per 2. In modo analogo al test a 2 code, i gradi di libertà saranno dati dalla numerosità del gruppo C meno 1.

Il TEST E' A DUE CODE quando la media di uno dei due gruppi può essere SIA MAGGIORE CHE MINORE RISPETTO ALLA MEDIA DELL'ALTRO GRUPPO. Troverò il punto critico nella tabella dei valori del t di Student con il corrispettivo valore di $\alpha$ ($\alpha \leq 0,05$ oppure $\alpha \leq 0,10$) ed i gradi di libertà dati dalla numerosità del gruppo C meno 1.

# Il Chi-Quadrato di Paerson

Serve a verificare la significatività delle differenze tra le distribuzioni di frequenze ovvero permette di confrontare due o più campioni per verificare, stabilita una certa probabilità di prima specie, se le differenze tra le distribuzioni di frequenza assolute, riferite alle diverse modalità del carattere studiato, siano significative.

Il metodo del chi-quadrato non può essere applicato a frequenze relative (percentuali), inoltre è applicabile quando il valore contenuto in ogni cella sia > 5 ed il numero totale delle osservazioni sia > 30, in caso contrario bisogna usare altri test (ad esempio test esatto di Fisher)

Inoltre quando vi è un solo grado di libertà è opportuno applicare la correzione di Yates.

Solitamente abbiamo 2 caratteri in un campione di un tot di unità statistiche; ogni carattere può avere più classi. Abbiamo cioè un tot di soggetti che vengono raggruppati in base a 2 criteri;

all'interno di questi criteri ci possono essere 2 o più classi; un criterio è posto nelle righe della tabella e l'altro criterio è posto nelle colonne della tabella.

Il 1° criterio, quello delle righe è il criterio INDIPENDENTE, mentre il 2° criterio, quello delle colonne, è il criterio DIPENDENTE. Con il Chi-quadrato di Paerson noi vogliamo sapere se il criterio INDIPENDENTE influenza in maniera significativa (con un certo valore di $\alpha$) i valori del criterio DIPENDENTE, ovvero se c'è una correlazione significativa tra il variare delle frequenze del criterio INDIPENDENTE ed il variare delle frequenze del criterio DIPENDENTE, ovvero se il criterio INDIPENDENTE influenza significativamente il criterio DIPENDENTE.

Esempio: Abbiamo 2 criteri: REDDITO e RENDIMENTO SCOLASTICO. Prendiamo un tot di studenti e chiediamo ad ogni studente il REDDITO della sua famiglia ed il suo RENDIMENTO SCOLASTICO; così possiamo costruire una tabella mettendo ogni studente in una data cella della tabella in base al suo REDDITO ed al suo RENDIMENTO SCOLASTICO. L'indagine viene fatta per appurare se il livello del reddito della famiglia influisce significativamente sul livello del rendimento scolastico; per questo il criterio REDDITO è il criterio INDIPENDENTE, mentre il criterio RENDIMENTO SCOLASTICO è il criterio DIPENDENTE (dal REDDITO secondo l'indagine).

(Criterio dipendente)
RENDIMENTO SCOLASTICO

(Criterio indipendente)
REDDITO

| | BUONO | DISCRETO | SCARSO | TOTALE |
|---|---|---|---|---|
| BASSO | 7 | 5 | 7 | 19 |
| MEDIO | 12 | 7 | 6 | 25 |
| ALTO | 15 | 8 | 3 | 26 |
| TOTALE PARZIALE | 34 | 20 | 16 | 70 |

Ora per verificare se il RENDIMENTO SCOLASTICO (criterio dipendente) dipende in maniera significativa dal REDDITO (criterio indipendente) poniamo come ipotesi (Ho) che questo non sia vero e che la differenza tra i vari rendimenti scolastici sia solo CASUALE; così calcolo la percentuale sul totale di studenti con rendimento scolastico BUONO, DISCRETO e SCARSO; troverò che 34 su 70 hanno rendimento BUONO, 20 su 70 hanno rendimento DISCRETO e 16 su 70 hanno rendimento SCARSO, quindi:

34 : 70 = 0,5 => (50%) rendimento BUONO

20 : 70 = 0,3 => (30%) rendimento DISCRETO

16 : 70 = 0,2 => (20%) rendimento SCARSO

Questi sono i valori percentuali di rendimento che sarebbe lecito attendersi se il rendimento non fosse correlato al livello di reddito familiare (IPOTESI Ho). Per verificare questa ipotesi costruisco una nuova tabella con i valori attesi casuali:

Devo moltiplicare la numerosità di ogni criterio indipendente (nell'esempio 19 // 25 // 26) per quei valori che poco fa ho trovato (nell'esempio 0,5 // 0,3 // 0,2).

Quindi dovrò fare:

RENDIMENTO BUONO => 19 x 0,5 = 9,5 => arrotondo = 9 //

25 x 0,5 = 12,5 => arrotondo = 12 //

26 x 0,5 = 13

(Se ho valori con le virgole devo approssimare).

Devo ripetere questo procedimento per ogni classe di variabile dipendente, così potrò costruire la tabella dei valori attesi casuali:

(Criterio dipendente)
RENDIMENTO SCOLASTICO

(Criterio indipendente)
REDDITO

| | BUONO | DISCRETO | SCARSO | TOTALE |
|---|---|---|---|---|
| BASSO | 9 | 6 | 4 | 19 |
| MEDIO | 12 | 8 | 5 | 25 |
| ALTO | 13 | 8 | 5 | 26 |
| TOTALE PARZIALE | 34 | 22 | 14 | 70 |

Questi sono i valori attesi che avremmo se il numero di studenti con rendimento scolastico buono, discreto e scarso non fosse correlato con il reddito familiare.

Ora possiamo applicare la formula del chi – quadrato di Paerson ($\chi 2$)

$\chi^2 = (7 - 9)^2 : 9 + (12 - 12)^2 : 12 + (15 - 13)^2 : 13 + (5 - 6)^2 : 6 + (7 - 8)^2 : 8 + (8 - 8)^2 : 8 + (7 - 4)^2 : 4 + (6 - 5)^2 : 5 + (3 - 5)^2 : 5 = 4,29$

Ovvero devo sottrarre al valore della 1° tabella, il corrispettivo valore della 2° tabella, poi devo elevare il valore ottenuto al quadrato ed infine devo dividere il valore ottenuto al valore della 2° tabella (tabella dei valori attesi) che ho pocanzi usato; il valore che ho ottenuto con questa procedura lo devo sommare ad ogni altro valore che deriva dalla stessa procedura per ogni casella della tabella (è più difficile a dirsi che a farsi…)

$$\chi^2 = \sum \frac{(\text{numero osservato} - \text{numero atteso})^2}{\text{numero atteso}}$$

Ora potrò concludere l'esercizio confrontando questo valore di $\chi^2$ con il punto critico che trovo nella tabella dei punti critici della distribuzione del chi – quadrato di Paerson per un dato valore di $\alpha$ con i gradi di libertà che calcolo facendo il prodotto tra il numero di criteri della variabile indipendente meno 1 per il numero di criteri della variabile dipendente meno 1:

[(numero di colonne – 1) × (numero di righe – 1)].

Nell'esempio di prima abbiamo 3 criteri per la variabile dipendente (RENDIMENTO SCOLASTICO => BUONO, DISCRETO, SCARSO) e 3 criteri per la variabile indipendente (REDDITO => ALTO, MEDIO, BASSO): Quindi i gradi di libertà saranno (3-1) × (3-1) = 4

Nella tabella dei punti critici della distribuzione del chi – quadrato di Paerson, per $\alpha = 0,1$ e per 4 gradi di libertà il valore sarà 7,78

Confrontando il valore di $\chi^2$ di 4,29 con il valore del punto critico di 7,78 avrò che:

4,29 < 7,78 = Ho

Ovvero non c'è una correlazione significativa (al 90% delle probabilità) tra il reddito familiare ed il rendimento scolastico.

---

Correzione di Yates (Quando vi è un solo grado di libertà)

All'interno della formula del chi – quadrato di Paerson il valore osservato più alto deve essere ridotto di 0,5 (ad esempio 40 – 0,5 = 39,5) ed il valore più basso deve essere aumentato di 0,5 (ad esempio 10 + 0,5 = 10,5) per il resto il procedimento da applicare per eseguire questo test statistico è esattamente lo stesso.

# Analisi della varianza ad un criterio

Questa procedura permette di stabilire, sempre ad una certa probabilità di α, se due o più campioni derivano dallo stesso universo statistico (Ho), ovvero se i criteri con cui si è stabilito che un elemento appartiene ad un campione piuttosto che ad un altro, costituiscono una fonte di variabilità sistematica, tale da poter affermare che gli stessi campioni provengano da differenti popolazioni statistiche ($H_1$) si tratterà allora di verificare quanto della variabilità complessiva sia dovuta all'errore e quanto al trattamento, se quest'ultimo abbia prodotto, sui campioni esaminati, una variabilità maggiore di quella casuale cioè se il trattamento ha avuto o meno un effetto significativamente diverso dal caso.

Come riconoscere uno di questi esercizi?

Solitamente il quesito dell'esercizio da svolgere con l'analisi della varianza ad 1 criterio recita: VALUTARE LA SIGNIFICATIVITA' DELLE DIFFERENZE TRA L'ATTIVITA' DELLE DUE VARIABILI (es. effetto dei 2 diversi geni, oppure effetto dei 2 diversi antibiotici…)

Svolgimento dell'esercizio:

Ho vari gruppi: (valori non appaiati) es. gruppo A, B, C; devo calcolare la media aritmetica di ogni gruppo e la devianza di ogni gruppo (mi servirà per calcolare la devianza entro gruppi).

Ora però devo calcolare la devianza totale: Per ogni gruppo devo fare la sommatoria del quadrato di ogni valore, otterrò così un valore per ogni gruppo (simbolo => $\Sigma\chi_2 i$)

Ora devo sommare i valori ottenuti da tutti i gruppi tra di loro (simbolo => $\Sigma\Sigma\chi_2 i$)

Ora dovrò trovare un altro valore, ovvero la sommatoria dei valori di tutti i gruppi (somma di tutti i valori del gruppo A, sommati alla somma di tutti i valori del gruppo B, sommati alla somma di tutti i valori del gruppo C) (simbolo => $\Sigma\chi i$)

Ho bisogno di solo un altro valore, ovvero la numerosità totale (di tutti i gruppi insieme) (Simbolo => $n_{tot}$)

Per calcolare la devianza totale dovrò applicare una semplice formula:

$$\Sigma\Sigma\chi_2 i - \frac{(\Sigma\Sigma\chi i)^2}{n_{tot}}$$

=> Così troverò la DEVIANZA TOTALE

Ora dovrò trovare la DEVIANZA ENTRO GRUPPI e la DEVIANZA TRA GRUPPI:

- DEVIANZA ENTRO GRUPPI: Si calcola semplicemente sommando la devianza di A + devianza di B + devianza di C ( $DEV_A + DEV_B + DEV_C$ )

- DEVIANZA TRA GRUPPI : Si calcola semplicemente sottraendo alla devianza totale la devianza entro gruppi ( $DEV_{tot} - DEV_{entro}$)

Ora dovrò trovare la VARIANZA TOTALE, la VARIANZA ENTRO GRUPPI e la VARIANZA TRA GRUPPI:

- VARIANZA TOTALE: Devianza totale diviso il numero di gradi di libertà dati dalla numerosità totale meno 1.

- VARIANZA ENTRO GRUPPI: Devianza entro diviso i gradi di libertà dati dalla somma dei gradi di libertà di ogni gruppo (ovvero numerosità del gruppo A – 1 sommato a numerosità del gruppo B – 1 sommato a numerosità del gruppo C – 1) => questi sono i gradi di libertà entro gruppi.

- VARIANZA TRA GRUPPI: Devianza tra diviso i gradi di libertà totali meno i gradi di libertà entro gruppi (vedi sopra)

Ora dovrò dividere il valore della VARIANZA TRA GRUPPI con quello della VARIANZA ENTRO GRUPPI e ricaverò un valore: F

$$F = \frac{VAR_{tra}}{VAR_{entro}}$$

Il valore di F sarà il valore da confrontare al valore TABULARE

Per ricavare il valore tabulare dovrò andare a trovare sulla tabella F (con α uguale al mio α → α=0,05 oppure α=0,01); quel valore che deriva dall'intersecazione dei GRADI DI LIBERTA' DEL NUMERATORE → GRADI DI LIBERTA' TRA GRUPPI ed i GRADI DI LIBERTA' DEL DENOMINATORE → GRADI DI LIBERTA' ENTRO GRUPPI.

Se il valore di F sarà maggiore a quello del punto critico → DIFFERENZA SIGNIFICATIVA

Se il valore di F sarà minore di quello del punto critico → DIFFERENZA NON SIGNIFICATIVA

# La regressione lineare semplice

Con la regressione lineare semplice analizziamo la dipendenza di una variabile (variabile dipendente: y) da un'altra variabile (variabile indipendente: x) ovvero analizziamo il tipo e l'intensità della relazione che sussiste tra le due variabili.

Si ricorre all'analisi della regressione quando dai dati campionari si vuole ricavare un modello statistico che predica i valori della variabile dipendente (y) a partire dalla variabile indipendente (x) ovvero dalla variabile causa e quindi determinare se esista una relazione lineare tra le variabili per un dato valore di α.

Solitamente in un esercizio sulla regressione lineare semplice ho un campione (ad esempio 10 persone, oppure 10 prelievi di sangue, oppure 10 diverse categorie lavorative…) da ogni singolo esponente del campione sono state misurate (ricavate) 2 variabili; si cerca di sapere se una variabile (la variabile dipendente) sia influenzata in maniera significativa (per un dato valore di α) dall'altra variabile (la variabile indipendente).

Il quesito dell'esercizio sarà quindi:

CALCOLA LA SIGNIFICATIVITA' DELLA RELAZIONE ESISTENTE TRA LA VARIABILE INDIPENDENTE E LA VARIABILE DIPENDENTE (OPPURE VERIFICA SE I VALORI DELLA VARIABILE Y SIANO DIPENDENTI DAI VALORI DELLA VARIABILE X).

Svolgimento dell'esercizio:

Devo calcolare la DEVIANZA DELLA VARIABILE X e la DEVIANZA DELLA VARIABILE Y

Ora devo calcolare la CODEVIANZA (XY): Per fare questo compilo una tabella con i valori degli scarti dalla media dei valori di x ed i valori degli scarti dalla media dei valori di y appaiati e moltiplico gli uni per gli altri; ricavo così dei valori; ora devo sommare questi valori ed il risultato che otterrò sarà la CODEVIANZA (XY).

Ora posso calcolare la DEVIANZA DELLA REGRESSIONE facendo:

$$\frac{[\text{codevianza (xy)}]^2}{\text{devianza (x)}} = \text{devianza della regressione}$$

La devianza totale non sarà altro che la DEVIANZA DI Y.

Ora posso calcolare la DEVIANZA DALLA REGRESSIONE =

$$\text{devianza dalla regressione} = \text{devianza totale (devianza di y)} - \text{devianza della regressione}$$

Ora dovrò calcolare le VARIANZE DELLA REGRESSIONE e VARIANZA DALLA REGRESSIONE, per fare questo però dovrò conoscere i gradi di libertà totali, i gradi di libertà della regressione ed i gradi di libertà dalla regressione:

- GRADI DI LIBERTA' TOTALI = Numerosità di y – 1
- GRADI DI LIBERTA' DALLA REGRESSIONE = Numerosità di y – 2
- GRADI DI LIBERTA' DELLA REGRESSIONE = Gradi di libertà totali – gradi di libertà dalla regressione = 1

Così potrò calcolare le VARIANZE DELLA e DALLA REGRESSIONE:

- VARIANZA DELLA REGRESSIONE = Devianza della regressione : Gradi di libertà della regressione → Varianza della regressione = Devianza della regressione

- VARIANZA DALLA REGRESSIONE = Devianza dalla regressione : Gradi di libertà dalla regressione → Varianza dalla regressione = [Devianza dalla regressione : (numerosità di y – 2)]

Finalmente potrò calcolare il valore di F:

$$F = \frac{\text{Varianza della regressione}}{\text{Varianza dalla regressione}}$$

Dovrò confrontare F con il punto critico della tabella dei punti critici della distribuzione F (rapporto tra varianze → ANOVA) con un dato valore di α ed il valore che si interseca dai GRADI DI LIBERTA' AL NUMERATORE = Gradi di libertà DELLA regressione che prima ho trovato ed i GRADI DI LIBERTA' AL DENOMINATORE = Gradi di libertà DALLA regressione che prima ho trovato.

Se il valore di F sarà maggiore del punto critico dovrò rifiutare l'ipotesi Ho, viceversa la dovrò accettare se il valore di F sarà minore di quello del punto critico.

MasterGi