Esercizi dal Foglio 9:

Esercizio D1

Un veicolo marcia per 50 km alla velocità v_1, e per altri 50 km alla velocità v_2.
La sua velocità sull'intero percorso di 100 km e' data da

1A La media aritmetica di v_1 e v_2
1B La media geometrica di v_1 e v_2
1C La differenza tra v_1 e v_2
1D La somma di v_1 e v_2
1E Nessuna delle precedenti **Risposta esatta**

SOLUZIONE

In questo caso la velocità media è la media armonica delle due velocità,

infatti posto d=50km v_1= e v_2= si ha che la velocità media è il percorso totale diviso il tempo totale, ossia $2d/(t_1+t_2)$ dove $t_1=d/v_1$ e $t_2=d/v_2$

e quindi la velocità media è

$2d/(t_1+t_2)=2d/[(d/v_1)+(d/v_2)]= 2/[(1/v_1)+(1/v_2)]$

che è proprio la media armonica tra v_1 e v_2.

RICORDIAMO CHE la MEDIA ARMONICA di

$\xi_1,\xi_2,\xi_3,\dots, \xi_{n-1}, \xi_n$, PER DATI STRETTAMENTE POSITIVI

è data da

h^{-1}(MEDIA ARITMETICA di $h(\xi_1), h(\xi_2), h(\xi_3),\dots,h(\xi_{n-1}), h(\xi_n)$)=

$= h^{-1}\big([h(\xi_1) + h(\xi_2) + h(\xi_3) +\dots+ h(\xi_{n-1}) + h(\xi_n)]/n \big)$

dove $h(x)=1/x$ e $h^{-1}(x)$ è la sua funzione inversa

e in questo caso, essendo $y=1/x$ SE E SOLO SE $x=1/y$, ossia $h^{-1}(y)=1/y$

$1/\big([(1/\xi_1) + (1/\xi_2) + (1/\xi_3) +\dots+ (1/\xi_{n-1}) + (1/\xi_n)]/n \big) =$

$= n/ [(1/\xi_1) + (1/\xi_2) + (1/\xi_3) +\dots+ (1/\xi_{n-1}) + (1/\xi_n)]$

Per n=2 viene appunto $2/ [(1/\xi_1) + (1/\xi_2)]$

Esercizio D2: Sono assegnati i seguenti dati numerici : 0, 3, 3, 3, 5, 5, 5, 8. Sia M la loro media aritmetica e s lo scarto quadratico medio. Si consideri l'intervallo (M - s, M + s).La differenza tra la lunghezza di tale intervallo e la distanza interquartile e'

2A 2,4 **Risposta esatta.**
2B 7,5
2C 1,5
2D 0,6
2E 2

SOLUZIONE: Vengono analizzati n=8 dati, che sono (in ordine crescente, o meglio NON DESCRESCENTE)

0,3,3,3,5,5,5,8 OSSIA

$\xi_{(1)}=0$, $\xi_{(2)}=3$, $\xi_{(3)}=3$, $\xi_{(4)}=3$, $\xi_{(5)}=5$, $\xi_{(6)}=5$, $\xi_{(7)}=5$, $\xi_{(8)}=8$,

I valori assunti sono 4 : $x_1=0$ $x_2=3$ $x_3=5$ $x_4=8$

le rispettive frequenze assolute sono : $f_1=1$ $f_2=3$ $f_3=3$ $f_4=1$

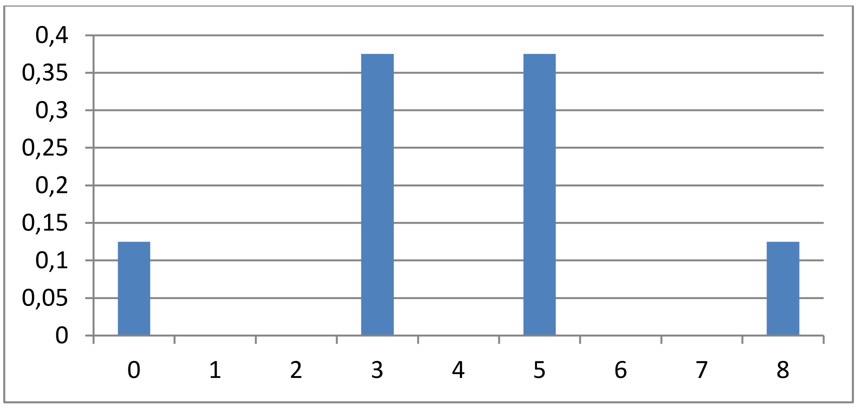
La media dei dati vale quindi

$M = x_{(1)} f_{(1)} + x_{(2)} f_{(2)} + x_{(3)} f_{(3)} + x_{(4)} f_{(4)} = 0\ (1/8) + 3\ (3/8) + 5\ (3/8) + 8\ (1/8) = 4$

OSSERVAZIONE La distribuzione (statistica) è data dalla tabella

x_i	0	3	5	8
f_i	1	3	3	1
f_i/n	1/8=0,125	3/8=0,375	3/8=0,375	1/8=0,125

L'ISTOGRAMMA DELLE FREQUENZE RELATIVE E'



L'istogramma è simmetrico rispetto al punto 4 e questo è sufficiente per garantire che la media aritmetica sia 4.

La varianza vale

$s^2 = (0-4)^2(1/8) + (3-4)^2 (3/8) + (5-4)^2(3/8)+(8-4)^2(1/8) = 2 (16/8) + 2 (3/8)= 19/4$

e quindi lo scarto quadratico medio (o deviazione standard vale

$s= \sqrt{(19)}/2$ = (circa) 4,35/2

e l'ampiezza dell'intervallo (M-s,M+s) vale 2s =(circa) 4,35.

Per calcolare la distanza interquartile, vanno calcolati il primo quartile q_1 e il terzo quartile q_3

La DISTANZA INTERQUARTILE vale q_3-q_1, (PER DEFINZIONE)

per ottenere q_1 calcoliamo n/4= 8/4=2 e quindi $q_1= [\xi_{(2)}+ \xi_{(3)}]/2= (3+3)/2=3$

per ottenere q_3 calcoliamo n(3/4)= 8(3/4)=6 e quindi $q_1= [\xi_{(6)}+ \xi_{(7)}]/2= (5+5)/2=5$

e quindi la distanza interquartile vale $q_3-q_1=5-3=2$

e la differenza tra l'ampiezza 2s e la distanza interquartile q_3-q_1 vale (circa) 4,35-2=2,35

RICORDIAMO CHE, se i dati sono $\xi_1,\xi_2,\xi_3,\dots, \xi_{n-1}, \xi_n$, e vengono riordinati in modo NON DECRESCENTE $\xi_{(1)} \leq \xi_{(3)} \leq \xi_{(3)} \leq \dots \leq \xi_{(n-1)} \leq \xi_{(n)}$,

IL PRIMO QUARTILE q_1 è definito come quel (se è unico) valore q_1 per il quale

la percentuale del 25% (=1/4) dei dati sono minori o uguali a q_1 , e invece il restante 75%=(3/4) dei dati è maggiore o uguale a q_1

ANALOGAMENTE

IL SECONDO QUARTILE q_2 è definito come quel (se è unico) valore q_2 per il quale

la percentuale del 50% (=2/4=1/2) dei dati sono minori o uguali a q_2 , e invece il restante 50%=(=2/4=1/2) dei dati è maggiore o uguale a q_2

IL SECONDO INTERQUARTILE COINCIDE CON LA MEDIANA

IL TERZO QUARTILE q_3 è definito come quel (se è unico) valore q_3 per il quale

la percentuale del 75% (=3/4) dei dati sono minori o uguali a q_3 , e invece il restante 25%=(1/4) dei dati è maggiore o uguale a q_3

NEL CASO IN CUI se n/4 è intero, tutti i punti tra $\xi_{(n/4)}$ e $\xi_{(n/4+1)}$ godono di questa proprietà e per rendere unico il valore di q_1 si prende il primo interquartile come la media aritmetica di $\xi_{(n/4)}$ e $\xi_{(n/4+1)}$

Ossia $q_1=[\xi_{(n/4)} + \xi_{(n/4+1)}]/2$ (ANALOGO DISCORSO PER il secondo e terzo interquartile)

Esercizio D5

Tre amici, Aldo, Bruno e Carlo scommettono sulle percentuali che otterranno 4 candidati alle elezioni comunali, X, Y, Z, e T. Il vincitore della scommessa sarà decretato in base al metodo dei minimi quadrati. Ad elezioni avvenute, risulta che, rispetto alle percentuali effettive: Aldo ha indovinato i voti di X, Y e Z e attribuito +2 a T. Bruno ha attribuito +0,5 sia a X che a Y, e -0,5 sia a Z che T. Carlo ha indovinato i voti di X e Z, ha dato +1 a Y e -1 a T.

Chi ha vinto la scommessa?

SOLUZIONE:

poste X_A,Y_A,Z_A,T_A, le previsioni di Aldo, X_B,Y_B,Z_B,T_B, le previsioni di Bruno, e X_C,Y_C,Z_C,T_C, le previsioni di Carlo e X,Y,Z,T le percentuali VERE ottenute dai candidati

dal problema sappiamo che

$X_A=X$, $Y_A=Y$, $Z_A=Z$ e $T_A=T+2\%=T+2/100$

$X_B=X+0,5\%$, $Y_B=Y+0,5\%$, $Z_B=Z-0,5\%$ e $T_B=T-0,5\%=T-0,5/100$

$X_C=X$, $Y_C=Y$, $Z_A=Z+1\%$ e $T_A=T-1\%=T-1/100$

quindi gli errori sono

$X_A-X=0$, $Y_A-Y=0$, $Z_A-Z=0$ e $T_A-T=2\%= 2/100$

$X_B-X=0,5\%$, $Y_B-Y=0,5\%$, $Z_B-Z=-0,5\%$ e $T_B-T=-0,5\%=-0,5/100$

$X_C-X=0$, $Y_C-Y=0$, $Z_A-Z=1\%$ e $T_A-T=-1\%=-1/100$

L'errore commesso da Aldo è quindi la somma dei quadrati degli errori ossia

$(X_A-X)^2 +(Y_A-Y)^2+ (Z_A-Z)^2+(T_A-T)^2=0+0+0+(2\%)^2= 4/100^2$

$(X_B-X)^2 +(Y_B-Y)^2+ (Z_B-Z)^2+(T_B-T)^2=(0,5\%)^2+(0,5\%)^2+(-0,5\%)^2+ (-0,5\%)^2= 4\ (1/2)^2/100^2 = 1/100^2$

$(X_C-X)^2 +(Y_C-Y)^2+ (Z_C-Z)^2+(T_C-T)^2=0+0+(1\%)^2+(-1\%)^2= 2/100^2$

e quindi, per il criterio usato nella scommessa il vincitore è Bruno

RICORDIAMO CHE IL METODO DEI MINIMI QUADRATI CONSISTE NEL MINIMIZZARE LA SOMMA DI QUADRATI. RIPRENDEREMO L’ARGOMENTO NEL CASO DELLA RETTA DI REGRESSIONE

ESERCIZIO D10 In una certa popolazione il rapporto tra il numero delle donne e quello degli uomini è di 6 a 5. Se l'età media delle donne è 40, e quella degli uomini è 45, qual è l'età media della popolazione?

10A 42
10B 42,80
10C 43
10D 43,05
10E 42,27 **Risposta esatta**.

OSSERVAZIONE: le risposte precedenti sono tutte plausibili, invece se ci fosse tra le risposte, ad esempio 50, allora sarebbe ovviamente una risposta da SCARTARE: L'ETA' MEDIA DEVE ESSERE UN NUMERO TRA 40 e 45!!

Se nella popolazione ci sono n_D donne di età $x^D{}_1, x^D{}_2, ..., x^D{}_{nD}$, ed n_U uomini di età $x^U{}_1, x^U{}_2, ..., x^U{}_{nU}$, allora posto $n = n_D + n_U$, e $x^D = 40$ e $x^U = 40$ le medie aritmetiche delle donne e degli uomini, rispettivamente, dal testo sappiamo che $n_D/n = 6/11$ e che $n_U/n = 5/11$

(dire che rapporto tra il numero delle donne e quello degli uomini è di 6 a 5 significa che esiste un numero m tale che $n_D = 6m$ e che $n_U = 5m$ e quindi $n = n_D + n_U = (5+6)m = 11m$ e quindi $n_D/n = 6m/(11m) = 6/11$

e quindi l'età media della popolazione vale $40\ (6/11) + 45\ (5/11) = 42{,}27$

Questo deriva dalla formula generale per cui $x = x^D\ (n_D/n) + x^U\ (n_U/n)$

INFATTI

la media aritmetica dell'età delle donne è $x^D = [x^D{}_1 + x^D{}_2 + ... + x^D{}_{nD}]/n_D$

la media aritmetica dell'età degli uomini è $x^U = [x^U{}_1 + x^U{}_2 + ... + x^U{}_{nU}]/n_U$

mentre l'età media della popolazione vale

$\underline{x} = ([x^D{}_1 + x^D{}_2 + ... + x^D{}_{nD}] + [x^U{}_1 + x^U{}_2 + ... + x^U{}_{nU}]) / n$ dove $n = n_D + n_U$.

si vede quindi facilmente che

$\underline{x} = \{[x^D{}_1 + x^D{}_2 + ... + x^D{}_{nD}]/n_D\}\ (n_D/n) + \{\ [x^U{}_1 + x^U{}_2 + ... + x^U{}_{nU}]/n_U\ \}\ (n_U/\ n) = x^D\ (n_D/n) + x^U\ (n_U/n)$.

OSSERVAZIONE: La MEDIA PESATA $\underline{x} = x^D\ (n_D/n) + x^U\ (n_U/n)$ è sicuramente un numero minore del massimo tra x^D e x^U, ed il minino tra x^D e x^U:

INFATTI $\min(x^D, x^U) \le x^D,\ x^U \le MAX(x^D, x^U)$ da cui, ricordando che $n = n_D + n_U$

$$\mathbf{x^D\ (n_D/n) + x^U\ (n_U/n) \le MAX\ (x^D, x^U)\ (n_D/n) + MAX\ (x^D, x^U)\ (n_U/n) \le MAX\ (x^D\ x^U)\ [(n_D/n) + (n_U/n)\ ] = MAX(x^D, x^U)}$$

ANALOGAMENTE

$$\mathbf{\min(x^D, x^U) = \min(x^D, x^U)\ (n_D/n) + \min(x^D, x^U)\ (n_U/n) \le x^D\ (n_D/n) + x^U\ (n_U/n)}$$

D. 57 Con riferimento ai dati dell'esercizio 10.4.3 del volume Matematica per Discipline Biomediche: in un gruppo di 5 adulti, la somministrazione di dosi diverse di un farmaco ha comportato le seguenti diminuzioni della pressione arteriosa

DOSE (in mg)	**DIMINUZIONE DELLA PRESSIONE (in mmHg)**
7	**10**
12	**18**
15	**20**
20	**25**
22	**25**

si calcoli l'equazione delle retta di regressione chiamando y la dose e x la diminuzione della pressione.
Con le consuete approssimazioni l'equazione è
57A y = x **-** 4 **Risposta esatta.**
57B y = x + 4
57C y = 4x + 1
57D y = 4x – 1
57E y = -x + 4

PRIMA di svolgere l'esercizio RICORDIAMO ALCUNI FATTI.

1) L'analisi dei dati relativi a SOLO 5 osservazioni SONO SOLO PER ESERCIZIO, MA NON AVREBBE SENSO in un esperimento reale
2) È più logico parlare della retta di regressione della variabile DIMINUZIONE DELLA PRESSIONE rispetto alla DOSE DEL FARMACO: **OSSIA chiameremo x la DOSE ed y la DIMINUZIONE e fare la regressione della diminuzione rispetto alla dose**.

QUINDI INIZIEREMO SVOLGENDO prima questo esercizio e considereremo invece la retta di regressione dell'esercizio in da ottenere la regressione di x rispetto ad y e quini la domanda e le risposte DEVONO ESSERE CAMBIATE NEL SEGUENTE MODO:

D. 57 (MODIFICATO) Con riferimento ai dati della tabella

x=DOSE (in mg)	**y=DIMINUZIONE DELLA PRESSIONE (in mmHg)**
7	**10**
12	**18**
15	**20**
20	**25**
22	**25**

(a) si calcoli la retta di regressione di y rispetto ad x
(b) si calcoli la retta di regressione di x rispetto ad y

57A x = y **-** 4 ovvero y=x+4 **Risposta esatta.**
57B x = y + 4 ovvero y=x-4
57C x = 4y + 1 ovvero y= (x-1)/4
57D x = 4y – 1 ovvero y= (x+1)/4
57E x = -y + 4 ovvero y=-x+4

CENNO AL PROBLEMA DELLA REGRESSIONE NEL CASO GENERALE:

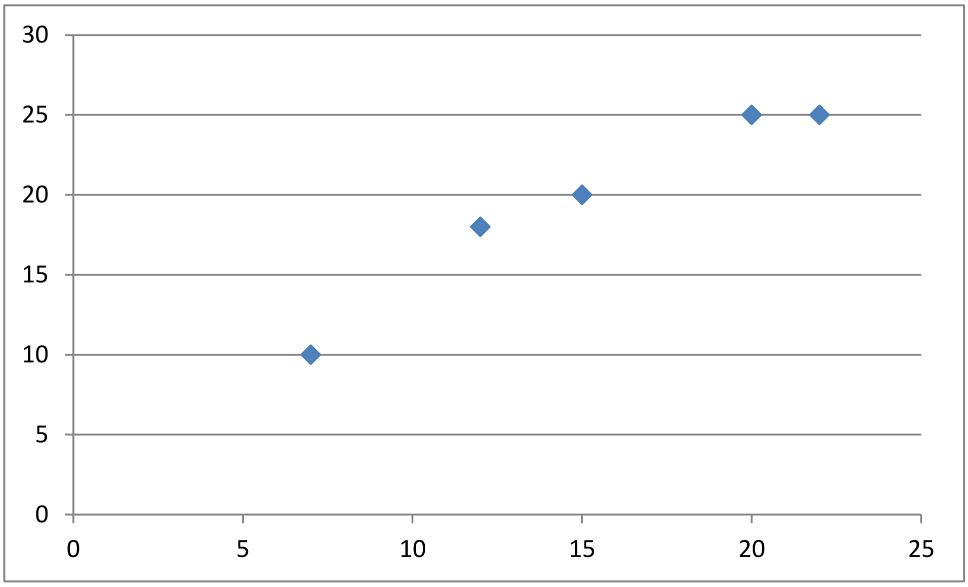
il problema della regressione si pone quando i DATI STATISTICI sono IN DUE DIMENSIONI:

a volte per ogni osservazione vengono forniti due numeri: ad esempio per n individui possiamo avere sia il dato del peso che la sua altezza.

I dati sono quindi del tipo n punti del piano (x_1, y_1), (x_2, y_2),...,(x_n, y_n).

A volte ci si può chiedere se esiste c'è una "specie di dipendenza lineare" tra i dati. In genere non c'è una precisa dipendenza lineare, cioè in genere non esiste un a e un b tali che y_i=a+bx_i per ogni i=1,2,...,n. TUTTAVIA ci si puo' chiedere se esiste una retta y*(x)=a*+b*x per la quale i dati y_i differiscano di poco dal valore y*(x_i)=a*+b*x_i.

PRIMA DI VEDERE LA TEORIA GENERALE SULLA RETTA DI REGRESSIONE, vediamo cosa significa nel caso dell'Esercizio:



Nel precedente grafico sono disegnati i punti (x_i,y_i) OSSERVATI, ossia i punti

(7,10)   (12,18)   (15,20)   (20,25)   (22,25)

Dai dati si "intuisce" che c'è una tendenza a crescere e ci si pone il problema se esiste una retta che "APPROSSIMA" bene i dati, ossia se esiste una relazione lineare tra i dati e se esiste una "LEGGE" lineare tra la DOSE e la DIMINUZIONE della pressione.

COSA VUOLE DIRE APPROSSIMA BENE?

Dobbiamo introdurre una MISURA QUANTITATIVA che ci dica quando i dati sono ben approssimati da una retta: IDEA MINIMI QUADRATI ATTENZIONE LA RETTA HA EQUAZIONE y=a+bx (E NON y=ax+b, come compare nella figura)

Data una retta $y=ax+b$ la misura della distanza dai dati è data dalla somma dei quadrati delle distanze dei punti (x_i,y_i) OSSERVATI dai punti $(x_i,a+bx_ia+bx)$ che appartengono alla retta e hanno le stesse ASCISSE.

È chiaro che, per ogni i il quadrato della distanza fra (x_i,y_i) e $(x_i,a+bx_i)$ vale

$$(x_i-x_i)^2+(y_i-a-bx_i)^2=(y_i-a-bx_i)^2.= (a+bx_i- y_i)^2$$

QUINDI la misura della distanza è una FUNZIONE di (a,b) e, quando si hanno n osservazioni (x_i,y_i), $i=1,\ldots,n$, tale misura diviene

$$H(a,b) = (a+bx_1- y_1)^2+(a+bx_2- y_2)^2+\ldots+(a+bx_{n-1}- y_{n-1})^2+(a+bx_n -y_n)^2.$$

Quindi, nel nostro caso, ad esempio, per la retta $y=2x-5$, ricordando che i dati osservati sono

(7,10) (12,18) (15,20) (20,25) (22,25)

mentre i punti sulla retta sono

(7,9) (12,19) (15,25) (20,35) (22,39)

la misura della distanza nel senso dei minimi quadrati è

$$H(2,-5) = (9-10)^2+(19-18)^2+ (25-20)^2+(35-25)^2+ (25-39)^2 = (-1)^2+ 1^2+5^2+10^2+14^2=$$

$$= 1 + 1 + 25 + 100 + 196 = 323$$

L'idea è trovare la retta per la quale H(a,b) è minima.

QUINDI PER RISPONDERE ALLA DOMANDA A RISPOSTA MULTIPLA POTREBBE BASTARE calcolare H(a,b) per i diversi valori di a e b proposti e scegliere quello che ha valore minore.

TUTTAVIA QUESTO METODO VALE SOLO SE LA DOMANDA E' A RISPOSTA MULTIPLA:

COME FARE PER IL CASO IN CUI SI DEBBA TROVARE LA RISPOSTA, senza nessun suggerimento?

SOLUZIONE GENERALE:

DATI OSSERVATI (x_i,y_i) **, per** $i=1,2,\ldots,n$

la retta di regressione è la retta che minimizza la somma dei quadrati delle distanze

$H(a,b) = \sum_{1\le i\le n} (a+bx_i - y_i)^2$

ossia la retta $y^*(x)=a^*+b^*x$ **per la quale**

$H(a^*,b^*) =\sum_{1\le i\le n} (a^*+b^*x_i - y_i)^2 \le \sum_{1\le i\le n} (a+bx_i - y_i)^2= H(a,b)$ **per ogni a,b**

da cui il nome del **METODO DEI MINIMI QUADRATI**

la retta di regressione ha la seguente proprietà IMPORTANTE:

passa per il punto $(\underline{x},\underline{y})$ **dove**

$\underline{x}$**=Media aritmetica dei dati** x_i $\underline{y}$**= Media aritmetica dei dati** y_i

e cioè è del tipo $y-\underline{y} = m\,(x-\underline{x})$

PER TROVARE il coefficiente angolare della retta di regressione (ossia a*) si procede così:

di definisce la COVARIANZA (simmetrica nei dati di tipo x e di tipo y) come

$Cov_{XY}= (1/n) \sum_{1\le i\le n} (x_i-\underline{x})(y_i-\underline{y}) = (1/n) \sum_{1\le i\le n} (x_i y_i) - (\underline{x})(\underline{y}) = \underline{xy} - (\underline{x})(\underline{y})$

ATTENZIONE nell'ultima riga abbiamo introdotto la notazione

$\underline{xy} = (1/n) \sum_{1\le i\le n} (x_i y_i)$

ossia la media aritmetica del prodotto dei valori osservati $x_i y_i$**, per** $i=1,2,\ldots,n$

IN CONLUSIONE LA RETTA DI REGRESSIONE E' LA RETTA

$$y-\underline{y} = [Cov_{XY} /(s^2_X)]\,(x-\underline{x})$$

dove s^2_X **è la VARIANZA dei dati OSSERVATI**

ossia $s^2_X =(1/n) \sum_{1\le i\le n} (x_i-\underline{x})^2$.

NELL'ESEMPIO DELL'ESERCIZIO, essendo i dati (x_i,y_i)

(7,10) (12,18) (15,20) (20,25) (22,25)

si ha

$\underline{x}$ = [7+ 12+15+20+22]/5= 76/5= 15,2

$\underline{y}$ = [10+18+20+25+25]/5 = 98/5=19,6

i valori

$x_i - \underline{x}$	7-15,2=-8,2	12-15,2=-3,2	15-15,2=-0,2	20-15,2=4,8	22-15,2=6,8
$y_i - \underline{y}$	10-19,6=-9,6	18-19,6=-1,6	20-19,6=0,4	25-19,6=5,4	25-19,6=5,4
$(x_i - \underline{x})(y_i - \underline{y})$	-8,2 (-9,6)= 78,72	-3,2 (-1,6)= 5,12	0,2 (0,4)= 0,08	4,8 (5,4)= 25,92	6,8 (5,4)= 36,72
$(x_i - \underline{x})^2$	$(-8,2)^2$= 67,24	$(-3,2)^2$= 10,24	$(0,2)^2$= 0,04	$(4,8)^2$= 23,04	$(6,8)^2$= 46,24

da cui

Cov_{XY}= (1/5) $\sum_{1\leq i\leq 5} (x_i-\underline{x})(y_i-\underline{y})$=[78,72 + 5,12 +0,08 +25,92+36,72]/5=146,56/5=29,312

s^2_X =(1/5) $\sum_{1\leq i\leq 5} (x_i-\underline{x})^2$ = [67,24 +10,24+ 0,04 + 23,04 +46,24]/5 = 146,8/5= 29,36

da cui la retta di regressione di y rispetto ad x è

$$y-\underline{y} = [Cov_{XY} /(s^2_X)] (x-\underline{x})$$

OSSIA, essendo $Cov_{XY} /(s^2_X)$= 29,312/29,36= 0,99836512261580381471389645776567

$$y-19,6 = 0,999 (x-15,2)$$

APPROSSIMANDO 29,312/29,36 con 1 viene la retta y-19,6 = x-15,2 ossia la retta

y = x-15,2+19,6 OSSIA circa la retta y = x +4,4

NOTA BENE: introducendo la notazione

$s^2_Y = (1/n) \sum_{1\leq i\leq n} (y_i-\underline{y})^2 = (1/n) \sum_{1\leq i\leq n} (y_i)^2 - (\underline{y})^2 = \underline{y\text{^}2} - (\underline{y})^2$

per la VARIANZA relativa ai dati y,

dove $\underline{y\text{^}2}$ = MEDIA ARITMETICA DEL QUADRATO DEI VALORI OSSERVATI PER LA y, ossia $\underline{y\text{^}2} = (1/n) \sum_{1\leq i\leq n} (y_i)^2$

s_Y è la radice quadrata di s^2_Y, cioè la deviazione standard per le ordinate, e analogamente per s_X

e introducendo **il coefficiente di correlazione**

$\boldsymbol{\rho_{XY}= Cov_{XY} /(s_X\, s_Y)}$ (ρ è la lettera greca "rho")

la retta di regressione $y-\underline{y} = [Cov_{XY} /(s^2_X)] (x-\underline{x})$

si può anche scrivere come (dividendo per s_Y)

$(y-\underline{y})/ s_Y = [Cov_{XY} /(s^2_X)] [(x-\underline{x})/ s_Y] = [Cov_{XY} /(s_X s_Y)] [(x-\underline{x}) / s_X]$

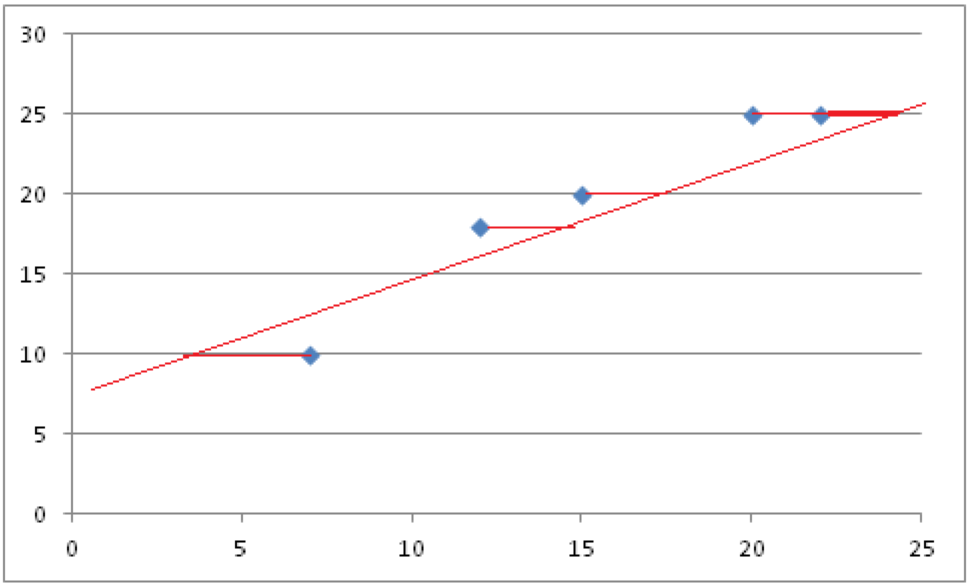
ovvero $\boldsymbol{(y-\underline{y})/ s_Y = \rho_{XY} [(x-\underline{x}) / s_X]}$

SI OSSERVI CHE il coefficiente di correlazione ρ_{XY} varia in [-1,1] .

INOLTRE si potrebbe dimostrare che $\rho_{XY} = \pm 1$, allora i dati sono perfettamente allineati, cioè per ogni i $y_i = y(x_i) = a^*+b^*x_i$

REGRESSIONE DI x RISPETTO AD y

Si tratta dello stesso tipo di problema in cui però si considera la x in funzione della y e quindi si cerca una retta del tipo $x=\alpha y+\beta$, e quindi si vuole invece minimizzare la distanza "orizzontale" ossia di minimizzare $H(\alpha,\beta)=\sum_{1\le i\le n} (\alpha+\beta y_i- x_i)^2$



A questo punto è chiaro che le due rette di regressione in generale sono diverse.

E la retta di regressione di x rispetto ad y è data da

$x-\underline{x} = [Cov_{XY} /(s^2_Y)] (y-\underline{y})$ ovvero , EQUIVALENTEMENTE $\boldsymbol{(x-\underline{y})/ s_X = \rho_{XY} [(y-\underline{y}) / s_Y]}$

CHE IN GENERE E' DIVERSA dalla retta di regressione di y rispetto ad x, TRANNE NEL CASO IN CUI IL COEFFICIENTE DI REGRESSIONE sia uguale a +1 o -1,

OSSIA TRANNE NEI DATI IN CUI I DATI SONO TUTTI SU UNA RETTA.

NELL'ESEMPIO DELL'ESERCIZIO, essendo i dati (x_i,y_i)

(7,10) (12,18) (15,20) (20,25) (22,25)

si ha di nuovo $\underline{x}$ **= [7+ 12+15+20+22]/5= 76/5= 15,2 e** $\underline{y}$ **= [10+18+20+25+25]/5 = 98/5=19,6**

i valori

$x_i - \underline{x}$	7-15,2=-8,2	12-15,2=-3,2	15-15,2=-0,2	20-15,2=4,8	22-15,2=6,8
$y_i - \underline{y}$	10-19,6=-9,6	18-19,6=-1,6	20-19,6=0,4	25-19,6=5,4	25-19,6=5,4
$(x_i - \underline{x})(y_i - \underline{y})$	-8,2 (-9,6)= 78,72	-3,2 (-1,6)= 5,12	0,2 (0,4)= 0,08	4,8 (5,4)= 25,92	6,8 (5,4)= 36,72
$(x_i - \underline{x})^2$	$(-8,2)^2$= 67,24	$(-3,2)^2$= 10,24	$(0,2)^2$= 0,04	$(4,8)^2$= 23,04	$(6,8)^2$= 46,24
$(y_i - \underline{y})^2$	$(-9,6)^2$= 92,16	$(-1,6)^2$= 2,56	$(0,4)^2$= 0,16	$(5,4)^2$= 29,16	$(5,4)^2$= 29,16

da cui

$Cov_{XY}= (1/5) \sum_{1\leq i\leq 5} (x_i-\underline{x})(y_i-\underline{y}) =[78,72 + 5,12 +0,08 +25,92+36,72]/5=146,56/5=29,312$

$s^2_y =(1/5) \sum_{1\leq i\leq 5} (y_i-\underline{y})^2 = [92,16 +2,56 + 0,16 + 29,16+29,16]/5 = 153,2/5= 30,64$

da cui la retta di regressione di x rispetto ad y è

$$x-\underline{x} = [Cov_{XY} /(s^2_Y)] (y-\underline{y})$$

OSSIA, essendo $Cov_{XY} /(s^2_y)= 29,312/30,64=0,956657963446475195822454308093 99$

$$x-15,2 = 0,957 (y-19,6)$$

OVVERO

$$y-19,6 = 1,045 (x-15,2)$$

E' CHIARO CHE QUESTA RETTA DIFFERISCE DALLA RETTA di regressione di rispetto ad x, calcolata precedentemente, ossia y-19,6 = 0,999 (x-15,2)

SE INVECE CONSIDERIAMO LE APPROSSIMAZIONI di queste due rette ottenute approssimando una volta 0,999 con 1 e una volta 1,045 con 1, la retta ossia y-19,6 = 0,999 (x-15,2) e la retta y-19,6 = 1,045 (x-15,2) vanno a coincidere con la retta y=x+4,4

ESEMPIO DEI TEST DIAGNOSTICI:

ESERCIZIO D38 del foglio RA2

si prende un campione " rappresentativo" di N persone appartenenti a una popolazione (ad esempio gli italiani tra 18 e 65 anni)

e li sottopone ad un test per diagnosticare una malattia.

Si indica

con M^+ l'insieme delle persone del campione che hanno la malattia

e con M^- l'insieme delle persone del campione che non hanno la malattia

con T^+ l'insieme delle persone del campione che sono risultate positive al test

e con T^- l'insieme delle persone del campione che sono risultate negative al test

(attenzione evidentemente c'è un modo per diagnosticare la malattia sicuro e forse "costoso" mentre il test non è sicuro ed "economico)

A questo punto la popolazione è divisa in 4 sottoinsiemi

$T^+ \cap M^+$ l'insieme dei veri positivi

$T^- \cap M^-$ l'insieme dei veri negativi

$T^+ \cap M^-$ l'insieme dei falsi positivi

$T^- \cap M^+$ l'insieme dei falsi negativi

Se prendiamo una persona a caso tra gli N sottoposti al test,

la probabilità di prendere una persona che ha malattia è

$P(M^+)=|M^+|/N$ OSSIA, espressa in percentuale, è LA **PREVALENZA** della malattia nel campione (la prevalenza di una malattia è la percentuale della malattia all'interno di una determinata popolazione)

la probabilità di prendere una persona che non ha la malattia è

$P(M^-)=|M^-|/N = (N-|M^+|)/N=1-|M^+|/N = 1-P(M^+)$

la probabilità di prendere (A CASO) una persona che è risultata positiva al test è

$P(T^+)=|T^+|/N$

la probabilità di prendere una persona che è risultata negativa al test è

$P(T^-)=|T^-|/N = (N-|T^+|)/N=1-|T^+|/N = 1-P(T^+)$

ed analogamente per

$P(T^+\cap M^+) = |T^+\cap M^+|/N$ è la probabilità di prendere un vero positivo

$P(T^-\cap M^-) = |T^-\cap M^-|/N$ è la probabilità di prendere un vero negativo

$P(T^+\cap M^-) = |T^+\cap M^-|/N$ è la probabilità di prendere un falso positivo

$P(T^-\cap M^+) = |T^-\cap M^+|/N$ è la probabilità di prendere un falso negativo

$P(T^+| M^+) = P(T^+\cap M^+)/P(M^+)= [|T^+\cap M^+|/N] / (|M^+|/N]=|T^+\cap M^+| / |M^+|$

è la probabilità che la persona sia risultata positiva al test sapendo che la persona ha la malattia

ed è detta LA **SENSIBILITA'** DEL TEST

$P(T^-| M^-) = P(T^-\cap M^-)/P(M^-)= [|T^-\cap M^-|/N] / (|M^-|/N]=|T^-\cap M^-| / |M^-|$

è la probabilità che la persona sia risultata negativa al test sapendo che la persona NON ha la malattia ed è detta LA **SPECIFICITA'** DEL TEST

Inoltre, di conseguenza,

$P(T^+| M^-) = 1- P(T^-| M^-)$

$P(T^-| M^+) = 1 - P(T^+| M^+)$

La probabilità che la persona sia risultata positiva al test (NON CONDIZIONATA) vale (per la probabilità totali)

$$\mathbf{P(T^+) = P(M^+)P(T^+|M^+) + P(M^-)P(T^+|M^-) = P(M^+)P(T^+|M^+) + [1- P(M^+)] [1-P(T^-|M^-)]}$$

quindi se sono noti la sensibilità $P(T^+|M^+)$ e la specificità $P(T^-|M^-)$ e $P(T^+)$ allora possiamo calcolare la prevalenza della malattia, che soddisfa una semplice equazione lineare

$$P(T^+) = P(M^+)P(T^+|M^+) + 1-P(T^-|M^-) - P(M^+) [1-P(T^-|M^-)]$$

$$= P(M^+) [P(T^+|M^+) -1 + P(T^-|M^-)] + 1-P(T^-|M^-)$$

da cui si può calcolare facilmente $P(M^+)$

INOLTRE (PER LA FORMULA DI BAYES)

possiamo calcolare la probabilità che una persona scelta a caso nel campione abbia la malattia **sapendo che la persona scelta è risultata positiva al test.**

INFATTI

$P(M^+|T^+) = P(M^+)P(T^+|M^+)/ [P(M^+)P(T^+|M^+) + P(M^-)P(T^+|M^-)]$

$= P(M^+)P(T^+|M^+) / [P(M^+)P(T^+|M^+) + [1-P(M^+)][1-P(T^-|M^-)]]$

IMPORTANTE se il campione è scelto in modo "rappresentativo" ed è abbastanza grande possiamo considerare che le probabilità precedenti (che sono in realtà pensate come frequenze relative) si possano prendere come le probabilità degli eventi relativi agli eventi del tipo

$P(M^+)=|M^+|/N$ la probabilità che una persona scelta a caso nella popolazione di cui il campione è stato scelto abbia la malattia

$P(T^+)=|T^+|/N$ la probabilità che una persona scelta a caso nella popolazione di cui il campione è stato scelto risulti positivo alla malattia

(NOTA BENE: questo approccio corrisponde ad usare l'impostazione frequentista delle probabilità)

e così via, e IMPORTANTE

$P(M^+|T^+) = P(M^+)P(T^+|M^+)/ [P(M^+)P(T^+|M^+) + P(M^-)P(T^+|M^-)]$

$= P(M^+)P(T^+|M^+) / [P(M^+)P(T^+|M^+) + [1-P(M^+)][1-P(T^-|M^-)]]$

si può considerare come la probabilità che una persona scelta a caso nella popolazione abbia effettivamente la malattia sapendo che la persona sia risultata positiva al test.

ESEMPIO

ESERCIZIO D38 del foglio RA2

D. 38 Un test diagnostico per la malattia M
ha specificità $P(T^-/M^-) = 80\%$, e sensibilità $P(T^+/M^+) = 90\%$.
Su 10000 soggetti, il test ha dato esito negativo in 7500 casi.

a) qual è, all'incirca, la prevalenza della malattia?
b) detti veri negativi i soggetti sani per i quali il test ha dato esito negativo, ovvero $T^-=T\cap M^-$, quanti veri negativi ci possiamo attendere?
c) Un individuo ha avuto test positivo, qual è la probabilità che abbia effettivamente la malattia?

DATI del PROBLEMA

$P(T^+|M^+)=90\% =90/100=9/10$, $P(T^-|M^-)=80\%=80/100=8/10$, $N=10000$,

$|T^-|=7500$

da cui $|T^+|=2500$

punto a) qual è, all’incirca, la prevalenza della malattia?

dai dati del problema possiamo affermare che

$P(T^+)=|T^+|/N= 2500/10000=1/4$

e d'altra parte

$P(T^+) = P(M^+)P(T^+|M^+) + P(M^-)P(T^+|M^-) = P(M^+)P(T^+|M^+) + [1- P(M^+)] [1-P(T^-|M^-)]$

$\quad = P(M^+) (9/10) + [1- P(M^+)][1-(8/10)] = P(M^+) (9/10) + [1- P(M^+)](2/10)$

e quindi

$1/4=P(M^+)[(9/10) - (2/10)] + (2/10)$

da cui la prevalenza della malattia vale

$P(M^+)= [(1/4)-(2/10)]/(7/10)= ([25-20]/100)*(10/7)= 5/70=_{(circa)} 0{,}71=7{,}1\%$

ATTENZIONE il libro di testo prevede anche un altro modo per risolvere questo tipo di esercizi, che è equivalente al precedente,(vi invito a pensare perché è equivalente) e che qui presento prendendo come incognita $|M^+|$ (invece il libro prende come incognita $x=|M^-|$):

Poiché la specificità $Sp=P(T^-/M^-)=|T^-\cap M^-|/|M^-|$ e ovviamente $|T^+\cap M^-|=|M^-| - |T^-\cap M^-|$, possiamo dire che

$|T^-\cap M^-| = Sp\ |M^-|$ e che $|T^+\cap M^-|=|M^-| -Sp\ |M^-|=(1-Sp)\ |M^-|$

Analogamente, poiché la sensibilità $Se=P(T^+/M^+)=|T^+\cap M^+|/|M^+|$

e ovviamente $|T^+\cap M^+|=|M^+| - |T^-\cap M^+|$, possiamo dire che

$Se\ |M^+|= |T^+\cap M^+|$ e che $|T^-\cap M^+|=|M^+| -Se\ |M^+|=(1-Se)\ |M^+|$

D'altra parte $|T^-|=|T^-\cap M^-| + |T^-\cap M^+|$ e $|M^-|=N-|M^+|$ quindi

$|T^-|=Sp\ |M^-| + (1-Se)\ |M^+|= Sp\ (N- |M^+|) + (1-Se)\ |M^+|$

da cui si può ricavare $|M^+|$ direttamente con una semplice equazione.

punto b)) detti veri negativi i soggetti sani per i quali il test ha dato esito negativo, ovvero $T^-=T\cap M^-$, quanti veri negativi ci possiamo attendere?

Essendo $P(T^-\cap M^-) = |T^-\cap M^-|/N$ e $P(T^-\cap M^-)=P(M^-)P(T^-| M^-)$ e $P(M^-)=1-P(M^+)$ ovviamente si ha che il numero dei veri negativi è

$|T^-\cap M^-|= P(T^-\cap M^-)\ N = P(M^-)P(T^-| M^-)\ N = (65/70)*(8/10)*10000=7428{,}57$ approssimato a 7429 (evidentemente la specificità e la sensibilà sono approssimate) analogamente si potrebbe ottenere

che il numero dei falsi positivi è

$|T^+ \cap M^-| = P(T^+ \cap M^-)\, N = P(M^-)P(T^+|\, M^-)\, N = (65/70)*(2/10)*10000 = 1857{,}14$ approssimato a 1857

che il numero dei veri positivi è
$|T^+ \cap M^+| = P(T^+ \cap M^+)\, N = P(M^+)P(T^+|\, M^+)\, N = (5/70)*(9/10)*10000 = 642{,}857$ approssimato a 643 (del resto 1857+643=2500, il numero delle persone risultate positive)

e infine che il numero dei falsi negativi è
$|T^- \cap M^+| = P(T^- \cap M^+)\, N = P(M^+)P(T^-|\, M^+)\, N = (5/70)*(1/10)*10000 = 71{,}428$ approssimato a 71 (del resto 7429 +71= 7500, il numero delle persone risultate negative)

punto c) Un individuo ha avuto test positivo, qual è la probabilità che abbia effettivamente la malattia?

scelta a caso una persona che è risultata positiva, la probabilità che abbia la malattia vale

$$P(M^+/T^+) = P(T^+ \cap M^+)/\, P(T^+) = P(M^+)P(T^+|M^+)/\, P(T^+)$$

$$= P(M^+)P(T^+|M^+)/\, [P(M^+)P(T^+|M^+) + P(M^-)P(T^+|M^-)\,]$$

$$= P(M^+)P(T^+|M^+) \,/\, [\, P(M^+)P(T^+|M^+) + [1-P(M^+)][1-P(T^-|M^-)]\ \,]$$

$$= (7{,}1/100)\,(9/10) \,/\, [(7{,}1/100)\,(9/10) + (91{,}9/100)(2/100)\,]$$

$$= 7{,}1*9/[7{,}1*9+92{,}9*2] = 0{,}2559 =_{(circa)} 26\%$$

Ovviamente, avremmo potuto utilizzare anche il fatto che

$P(T^+)=|T^+|/N= 2500/10000=1/4$ a denominatore invece della formula per cui

$$P(T^+)=P(M^+)P(T^+|M^+) + P(M^-)P(T^+|M^-)$$

(che tra l'altro, poiché abbiamo approssimato $P(M^+)= 5/70$ con il 71%, $P(T^+)$ ci è venuta uguale a $(7{,}1/100)\,(9/10) + (91{,}9/100)(2/100) = 249{,}7/1000 = 0{,}2497$, leggermente diversa da $0{,}25=1/4$, che è il valore preciso.

QUESTA FORMULA E' TUTTAVIA UTILE NEL CASO IN CUI LA PERSONA CHE HA EFFETTUATO IL TEST NON SIA SCELTA A CASO, MA SIA IN UN GRUPPO DI PERSONE A RISCHIO, come mostra il seguente ragionamento:

IMPORTANTE non siate meravigliati del fatto che $P(M^+/T^+)$, cioè la probabilità di avere effettivamente la malattia SAPENDO che il TEST è positivo, è venuta abbastanza piccola:

il punto è che abbiamo preso una persona a caso e NON ABBIAMO MOTIVI DI PENSARE CHE abbia la malattia. In genere chi fa il test di solito ha dei motivi che per cui NON E' GIUSTO USARE P(M+)=7,1% come prevalenza della malattia, in quanto appartiene a una sottopopolazione in cui la prevalenza della malattia è più alta, ad esempio se fosse che tale probabilità nella classe delle persone fosse del 50%=1/2 si otterrebbe invece

$P(M^+/T^+)= (1/2)\ (9/10)\ /\ [(1/2)\ (9/10) + (1/2)(2/100)\ ]= 9/(9+2)=9/11=_{(circa)}0{,}818= 81{,}8\%$

..

..

ωεχ<ξχ!∀≤