Sistemi di variabili casuali

0. Introduzione

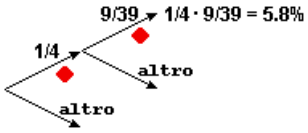
Abbiamo inizato a vedere, esaminando i concetti di dipendenza ed indipendenza stocastica, come si studino i rapporti tra variabili casuali riferite allo stesso fenomeno o a fenomeni in qualche modo collegati. In questa scheda approfondiremo questo studio, analizzando varie tecniche per visualizzare le relazioni tra due variabili casuali e per studiarle numericamente. Questi temi li riprenderai e approfondirai in un secondo tempo, quando disporrai di strumenti matematici adeguati.

1. Dipendenza e indipendenza stocastica

Richiamiamo come abbiamo introdotto concetti, prima di approfondirli. Partiamo da un esempio già considerato.
(**a**) *Qual* è la probabilità che *alzando* 2 volte un mazzo (nuovo) di carte da scopa ottenga sempre una carta di denari?
(**b**) *Qual* è la probabilità che *estraendo* 2 carte dal mazzo queste siano entrambe di denari?

- Nel caso della ***alzata***, avendo supposto il mazzo nuovo (e non truccato e mescolato bene) posso ritenere che, tagliandolo, le carte, e quindi (essendoci 10 carte per ogni seme) anche i semi, a valori in {♥, ♦, ♣, ♠}, escano con distribuzione uniforme: l'uscita di una carta di denari ha la stessa probabilità di quella di una di fiori o … Posso rappresentare queste due alzate col *grafo ad albero a fianco,* a due diramazioni. Ho 1/4 di probabilità di estrarre denari alla prima alzata ed 1/4 di estrarlo alla seconda. La probabilità cercata è dunque 1/4·1/4 = 1/16.

- Anche nel caso della ***estrazione*** posso ritenere equiprobabili le carte, e i semi, del mazzo. Ma mentre alla prima estrazione ho 1/4 di probabilità di estrarre una carta di denari, alla seconda estrazione la probabilità cambia. Le carte da cui effettuare l'estrazione sono, ora, una in meno, e, se ho estratto una carta di denari alla prima estrazione, le carte di denari rimaste sono 9. Il grafo a destra illustra la situazione. La probabilità cercata in questo caso è 1/4·9/39 = 3/4/13 = 0.0576923… = 5.8%.



Indichiamo con le variabili casuali S1 e S2 il seme della prima uscita e quello della seconda. Nel caso della ***alzata*** S1 e S2 *sono indipendenti:* qualunque seme abbia la 1ª carta, la probabilità che la 2ª abbia un certo seme è sempre la stessa. Ciò corrisponde al fatto che il grafo relativo all'alzata si riproduce allo stesso modo passando da una diramazione alla successiva. Per calcolare Pr(S1=♦ **and** S2=♦) posso fare direttamente Pr(S1=♦)·Pr(S2=♦) = 1/4·1/4 = 1/16.

Nel caso della ***estrazione*** S1 e S2 *non sono indipendenti:* ad es. Pr(S2=♦) (la probabilità che la 2ª carta sia di ♦) dipende dal valore assunto da S1 (cioè dal seme della 1ª carta). Ciò corrisponde al fatto che il grafo relativo alla estrazione non si riproduce allo stesso modo passando da una diramazione alla successiva: al primo arco "♦" è associata la probabilità 1/4, al secondo arco "♦" è associata la probabilità 9/39.

Due ***variabili casuali*** X e Y sono *probabilisticamente* ***indipendenti*** se sono indipendenti gli eventi **A** e **B** comunque prenda **A** evento relativo a X (condizione in cui compare solo la variabile X) e **B** evento relativo a Y (condizione in cui compare solo variabile Y): conoscere qualcosa su come si manifesta X non modifica le mie aspettattive sui modi in cui può manifestarsi Y, e viceversa. Altrimenti sono *probabilisticamente* ***dipendenti***. Esempio:

– sapere qualcosa a proposito del seme della 1ª carta estratta cambia le mie valutazioni sul seme che potrebbe avere la 2ª carta estratta: il seme della 1ª estrazione e quello della 2ª sono variabili casuali dipendenti.

Ricordiamo che il concetto di *dipendenza* ora introdotto è diverso da quello impiegato per esprimere il legame tra due grandezze quando una varia *in funzione* dell'altra. L'avverbio "probabilisticamente" (o l'equivalente avverbio "stocasticamente") evidenzia questa differenza. Se non ci sono ambiguità, questo avverbio viene omesso.

1 Vengono tirate tre palle contro un bersaglio. I tiri sono indipendenti. La probabilità di colpire il bersaglio è P. Qual è la probabilità che due dei tre tiri vadano a centro?

2. Le distribuzioni bivariate

Qui puoi rivedere i concetti di dipendenza ed indipendenza, su cui torneremo nel §3. Mettiamo, ora, a punto degli strumenti per studiare meglio il legame tra diverse variabili casuali.

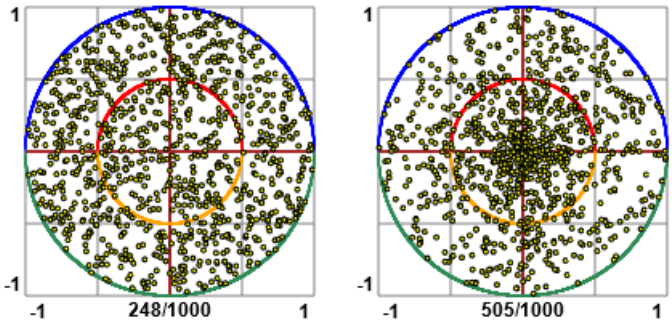
Esempio 1. Vogliamo stimare sperimentalmente la probabilità che, prendendo "a caso" un punto in un bersaglio composto da due cerchi concentrici con raggi uno doppio dell'altro, il punto cada nel cerchio centrale. Consideriamo due possibili procedimenti, mediante i quali viene generato un punto che cade nel cerchio di centro O=(0,0) e raggio 1 e verificato se la sua distanza da O è minore di 1/2.

Il primo consiste nel generare a caso, con distribuzione uniforme, ascissa e ordinata di un punto che sta nel quadrato a lati paralleli agli assi in cui è inscritto il cerchio, nel considerare solo i punti che stanno nel cerchio e nel contare quanti di questi stanno anche nel cerchietto di raggio 1/2. Gli esiti di un esperimento di tal genere, generando 1000 punti che cadono nel cerchio, sono illustrati sotto, a sinistra. È indicata anche la percentuale di quelli che cadono nel cerchietto.

Il secondo consiste nel generare a caso, con distribuzione uniforme, la distanza **r** (tra 0 ed 1) del punto da O e la direzione (tra 0 e 2π)

in cui esso sta. Sotto, a destra, è illustrato il procedimento.

Se vuoi, **qui** e **qui** trovi gli script con cui realizzare queste rappresentazioni.



Mentre col primo procedimento i proiettili si distribuiscono in modo uniforme nel cerchio, col secondo si distribuiscono concentrandosi maggiormente intorno a (0,0). Mentre col primo la frequenza tende a stabilizzarsi su 1/4, pari al rapporto tra area del centro e area del bersaglio, col secondo tende a stabilizzarsi su 1/2: il fatto che il punto generato disti meno di metà raggio dal centro dipende solo dal valore di **r**, che essendo con distribuzione uniforme in [0,1), ha il 50% di probabilità di essere minore di 1/2.

Si tratta di *due* cadute casuali con diverse *leggi di distribuzione.*

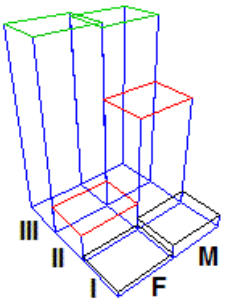
Esempio 2. La tabella seguente rappresenta la distribuzione delle variabili *Sesso* e *Settore di attività* (classificato in "agricoltura", "industria", "servizi") in cui una persona (in Italia nel 2012) era occupata: per ogni possibile coppia (Sesso, Settore) è indicata la corrispondente frequenza assoluta. Una tabella "a doppia entrata" come questa, in cui sono indicate le frequenze o le probabilità di due variabili casuali viene chiamata ***tabella di contingenza***.

(*1000)	Maschi	Femmine
Agricoltura	603	246
Industria	5051	1311
Servizi	7787	7901

%	Maschi	Femmine
Agricoltura	2.6	1.1
Industria	22.1	5.7
Servizi	34.0	34.5

% Maschi		
Agricol.	Industria	Servizi
4.5	37.6	57.9

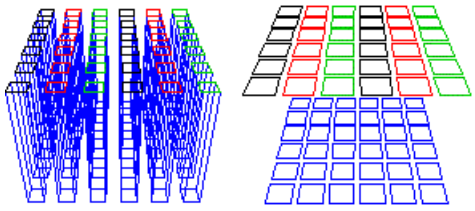
% Femmine		
Agricol.	Industria	Servizi
2.6	13.9	83.5



La tabella che sta sotto alla precedente contiene la ripartizione percentuale: 100 occupati, mediamente, come si distribuivano per sesso e per settore di occupazione. Le tabelle a destra contengono, invece, la distribuzione percentuale dei maschi e quella delle femmine nei tre settori (queste distribuzioni sono chiamate anche *profili colonna* della tabelle in quanto evidenziano come si ripartiscono i dati nelle colonne). L'***istogramma tridimensionale*** a destra rappresenta graficamente la tabella di contingenza: esso consente di interpretare visivamente le informazioni numeriche in essa contenute: maschi e femmine occupati nel settore "servizi" sono più o meno equinumerosi, nella "industria" e nella "agricoltura" sono invece molto più numerosi i maschi.

Sia nell'esempio 1 che nell'esempio 2 abbiamo esteso il concetto di legge di distribuzione dal caso di una variabile casuale U al caso di U = (X,Y) con X e Y variabili casuali. Questo viene chiamato anche caso ***bivariato***; si dice anche che U è un ***sistema*** di variabili casuali. Nel caso 1 si trattava di variabili X ed Y a valori reali, nel caso 2, invece, X ed Y avevano valori non mumerici. Nel caso 2 abbiamo visto come rappresentare la distribuzione con un istogramma tridimensionale. Vediamo, ora, come si rappresenta graficamente il caso in cui le due variabili siano numeriche.

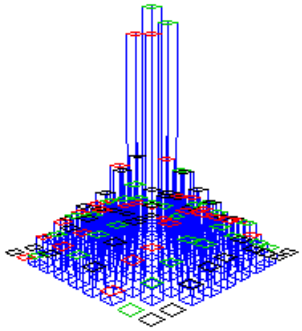
Nel caso di una singola variabile classificavo le uscite in intervallini; analogamente, nel nuovo caso, posso rappresentare le distribuzioni classificando le uscite in tanti rettangolini.

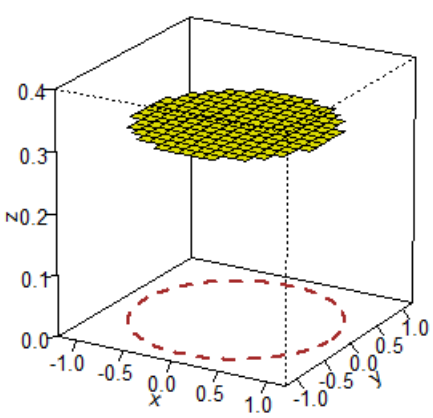


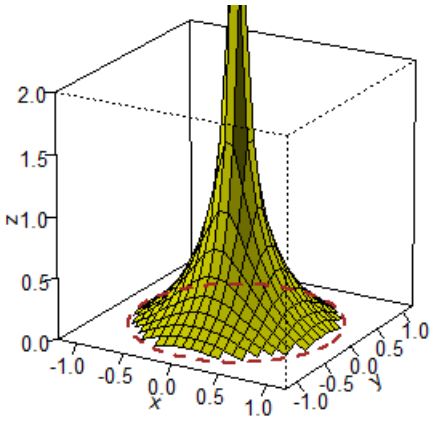
A destra è rappresentata la legge $U=(U_1,U_2)$, esito del lancio di due dadi equi; in questo caso le altezze degli istogrammi corrispondono alle probabilità $Pr(U=(i, j))$ con i e j in {1, 2, …, 6}; questa è una *distribuzione uniforme* finita: le probabilità $Pr(U = (i,j))$ sono tutte uguali, a 1/NumeroUscitePossibili = $1/(6\cdot6) = 1/36$. Il secondo istogramma rappresenta solo le parti superiori delle colonne, evidenziando meglio che si tratta di una distribuzione uniforme.

Nel caso dell'esempio iniziale, la caduta dei proiettili, siamo di fronte a un sistema (X,Y) di variabili casuali non discrete. Un'idea della distribuzione mi è fornita dal ***grafico di dispersione*** (scatter diagram), ossia dalla rappresentazione grafica delle coppie di uscite sperimentali. All'inizio del paragrafo abbiamo visto i grafici di dispersione dei due esempi di caduta.

Per una rappresentazione tridimensionale osserviamo che, come nel caso di una singola variabile a valori in un intervallo di numeri reali realizzavamo un istogramma classificando le uscite in intervallini, analogamente, ora, possiamo rappresentare le distribuzioni classificando le uscite in tanti rettangolini la cui unione copra il dominio delle uscite. Per il primo tipo di cadute otteniamo un istogramma che, all'aumentare del numero dei lanci e dei rettangolini, tende ad assumere una forma piatta, in quanto le uscite in rettangolini uguali tendono ad essere in numero eguale. Ecco, invece, sotto a sinistra, una possibile rappresentazione per il secondo tipo di cadute, in cui le colonnine sono state separate per facilitare la "lettura".



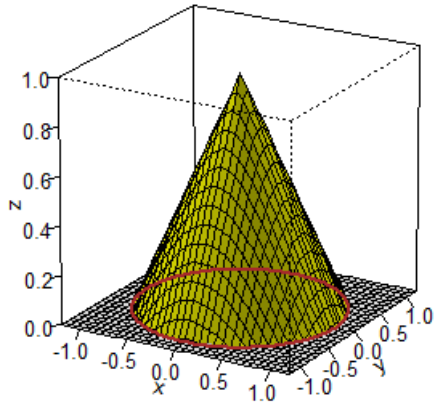




Nel caso di una variabile continua X all'aumentare delle prove e all'infittirsi della partizione il contorno superiore dell'istogramma sperimentale (normalizzato, in modo che sia di area 1) tende a stabilizzarsi su una curva. Analogamente nel caso di una variabile U = (X,Y), se all'aumentare delle prove e all'infittirsi dei rettangolini il contorno superiore dell'istogramma tridimensionale (in cui ogni colonnina abbia altezza pari alla frequenza relativa divisa per l'area del rettangolino di base, in modo che il volume complessivo sia 1) tende a stabilizzarsi su una superficie che sottende uno spazio di volume 1, il calcolo delle probabilità può essere ricondotto al calcolo di volumi. Sopra, a destra, sono raffigurate queste "superfici limite", che rappresentano le leggi di distribuzione dei due tipi di cadute. Analogamente al caso univariato, le funzioni di due variabili sul cui grafico tendono a stabilizzarsi gli istogrammi sperimentali si chiamano *funzioni di densità.*

Qual è la funzione densità della variabile U = (X,Y) corrispondente al primo tipo di caduta? È la funzione che a ogni (x,y) del cerchio associa il valore $1/\pi$ e agli altri punti associa 0. Il volume del cilindro verticale che essa delimita assieme al piano z = 0 ha volume 1.

2 A lato è raffigurato il grafico di una funzione densità **f**. Il cerchio di base ha raggio 1 [per (x,y) esterno al cerchio assumo **f**(x,y)=0].
Come sono fatte le curve di livello di **f**? ***Quanto*** vale **f**(0,0)?



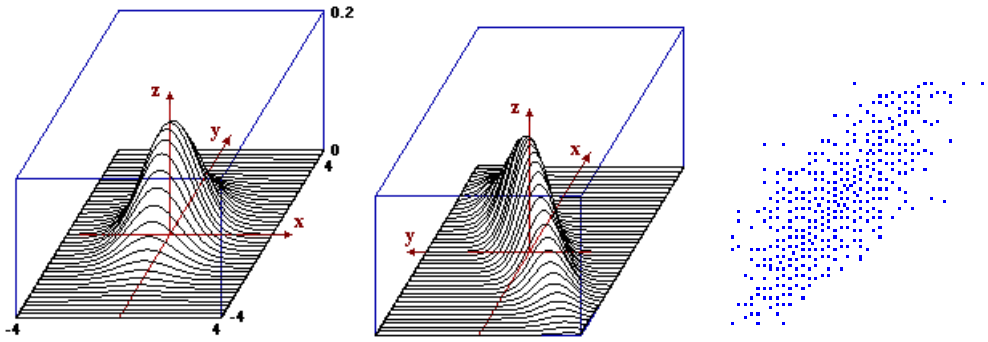
3. Dipendenza/indipendenza e forma del grafico della distribuzione

Come si fa a capire dal grafico della distribuzione di U=(X,Y) se X e Y sono variabili casuali indipendenti o no?
Nel caso → dell'istogramma di (Sesso,Settore) la riga di colonne che rappresenta la distribuzione dei maschi non ha andamento analogo a quella delle femmine, e questo ci fa capire che Sesso e Settore non sono indipendenti. Nel caso → del lancio di due dadi, invece, tutte le righe di colonne hanno andamento simile (anzi, uguale): le uscite di primo e secondo dado sono indipendenti. L'ipotesi di indipendenza corrisponde, nel caso finito (e nel caso sperimentale), alla *proporzionalità tra le righe o tra le colonne* della tabella a doppia entrata e, passando all'istogramma tridimensionale, corrisponde alla proporzionalità tra le altezze delle righe di colonnine o tra le altezze delle file di colonnine.

In generale, passando alle funzioni di densità al posto delle righe e delle file di colonnine si considerano le sezioni parallele al piano xz e le sezioni parellele al piano yz: due qualunque sezioni, ad esempio parallele al piano xz, devono essere ottenibili l'una dall'altra mediante una dilatazione/contrazione verticale. Nel caso → delle funzioni densità dei due esempi inziali, dei proiettili, ciò non accade mai: ad esempio la sezione determinata dal piano yz non ha lo stesso andamento (a meno di un fattore di scala) di nessuna delle altre sezioni ad essa parallela. Del resto è intuitivo che il valore di X e quello di Y sono tra loro condizionati: devono essere le coordinate di un punto che sta nel cerchio (X^2+Y^2 deve essere al più 1; se X è vicino ad 1 Y per forza deve essere vicino a 0).

3 ***Nel caso*** del sistema di variabili avente densità che ha per grafico un cono circolare retto, considerato sopra, X e Y sono indipendenti?

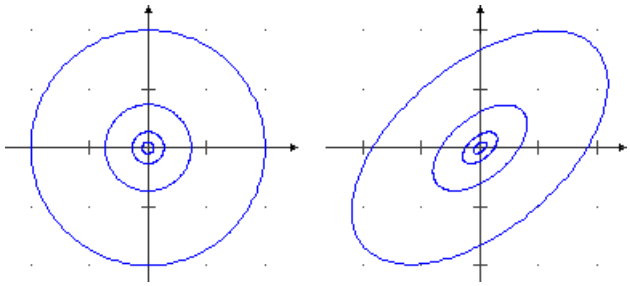
Altri due esempi. Sotto a sinistra è rappresentato un sistema (X,Y) con X e Y indipendenti: comunque sezioni la superficie con piani paralleli ai piani xz e yz ottengo grafici con andamenti simili: hanno massimo e punto di flesso collocati nella stessa posizione. Potrebbe avere una forma simile (anche se non centrata in (0,0) e con diverse unità su gli assi) la distribuzione di (X,Y) con X e Y altezze di un uomo e una donna sorteggiati a caso.
Invece, nel caso a destra (per cui abbiamo tracciato anche un possibile grafico di dispersione sperimentale) siamo di fronte a X e Y non indipendenti; ad es., è evidente che le sezioni parallele al piano yz sono curve con il punto di massimo che man mano si sposta verso destra (avanza lungo la direzione dell'asse y).



Potrebbe avere una forma simile la distribuzione di (X,Y) con X e Y altezze di marito e moglie di una coppia sorteggiata a caso: l'altezza di uomini spostati con donne di una certa altezza ha andamento più o meno gaussiano, ma la loro altezza media è maggiore di quella degli uomini sposati con donne più basse (uomini più alti tendenzialmente sposano donne più alte: non è affatto vero che l'amore è cieco!).

Ma la dipendenza tra X e Y in questo ultimo caso è in un qualche senso "più forte" di quella che c'era tra X e Y nel caso dei proiettili: là avevamo che i valori che poteva assumere una delle due variabili era condizionato da quello che assumeva l'altra, qui abbiamo qualcosa di più: al crescere di X anche Y tende a crescere. Su questo aspetto ci si sofferma nel prossimo paragrafo, sulla "correlazione".

Sotto sono tracciate alcune curve di livello delle due ultimi superfici considerate. Nel primo caso sono ellissi simmetriche rispetto agli assi x e y, nel secondo hanno assi di simmetria obliqui, a conferma del fatto che al crescere dell'uscita X l'uscita di Y tende a crescere anch'essa.



4. Covarianza e correlazione

Abbiamo visto esempi di variabili X e Y, come ascissa e ordinata dei punti di un bersaglio che vengono colpiti, che non sono stocasticamente indipendenti (il valore assunto da X condiziona i valori che può assumere Y), ma che, tuttavia, danno luogo a un diagramma di dispersione in cui non viene privilegiata alcuna direzione. Ne abbiamo visti altri, come quello considerato alla fine del paragrafo precedente, in cui i punti tendono a disporsi lungo una linea obliqua, ossia in cui, all'aumentare di X, Y tende ad aumentare più o meno proporzionalmente. Per distinguere queste situazioni, in entrambe delle quali X e Y sono dipendenti, si dice che X e Y nel secondo caso sono ***correlate***, nel primo no. Cerchiamo di *quantificare* quanto due variabili sono correlate.

Abbiamo visto che per avere un'idea numerica di come i dati sono dispersi attorno alla media si usa la *varianza,* ossia la *media dei "quadrati" degli scarti* dalla *media.* In altre parole, essa vale **M((X–M(X))²)**, ossia, per n dati X_1, …, X_n di *media* $\boldsymbol{m}$, $\Sigma_{i=1..n} (X_i - \boldsymbol{m})^2 / n$

Quanto più i dati sono vicini a **m** tanto più questo valore è vicino a 0. Il seguente valore, in cui invece di $(X–M(X))^2$ si considera $(X–M(X))\cdot(Y–M(Y))$, è invece un valore che è più vicino a 0 quanto più i punti sono disposti attorno al *baricentro,* cioè al punto le cui coordinate sono M(X) e M(Y):

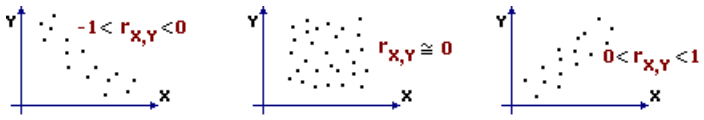
covarianza: $\boldsymbol{Cov(X,Y)} = M((X–M(X))\cdot(Y–M(Y)))$

Nel caso sperimentale, se $\boldsymbol{m_x}$ e $\boldsymbol{m_y}$ sono le medie di X_1, …, X_n e di Y_1, …, X_n, questo termine diventa: $\Sigma_{i=1..n} (X_i–\boldsymbol{m_x})(Y_i–\boldsymbol{m_y}) / n$

Questa formula, come puoi vedere nel §7, rappresenta un indicatore che assume un valore assoluto che scende quanto più i punti tendono a disporsi in modo da presentare una simmetria verticale o orizzontale e che cresce quanto più i punti tendono a disporsi lungo una retta obliqua. Per non tener conto delle unità di misura in cui sono espressi X e Y (e per passare da un'"area" a un numero puro) la covarianza viene normalizzata dividendo per gli s.q.m. di X e Y, introducendo il:

$$\textit{coefficiente di correlazione: } r_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\cdot\sigma(Y)}$$

Si può dimostrare che se X e Y sono dipendenti deterministicamente e legate da una relazione lineare $Y = aX + b$ il coefficiente di correlazione assuma il valore assoluto massimo. Vale 1 se l'andamento è crescente e −1 se è decrescente. Quindi, in generale, $-1 \le r_{X,Y} \le 1$.



Nota. Come abbiamo visto ➡ nella scheda sul "limite centrale", σ può essere stimato usando come denominatore "$n−1$" invece che "n". Una cosa analoga accade per la covarianza. Dato che il coefficiente di correlazione si ottiene come rapporto in cui compaiono a numeratore e a denominatore covarianza e "σ", per esso non ci sono ambiguità: sia in un caso che nell'altro si ottiene lo stesso valore.

Per rendere più semplice il calcolo del coefficiente di correlazione si può usare lo script **RegCorr** che, oltre a calcolare il coefficente di correlazione, individua anche la "retta di regressione", su cui ci soffermeremo nel prossimo paragrafo. Ecco, sotto, che cosa si deve fare per studiare la relazione tra X = (10, 20, 30, 40) e Y = (30, 40, 50, 60).

Linear regression and **correlation coefficient**. Enter x and y of the data and (possibly) the P through which the graph is constrained to pass (if you don't put anything, the centroid is taken)

10, 20, 30, 40

P: **x** separated by "," ↑ - **y** separated by "," ↓ regression C

30, 40, 50, 60

y = x +

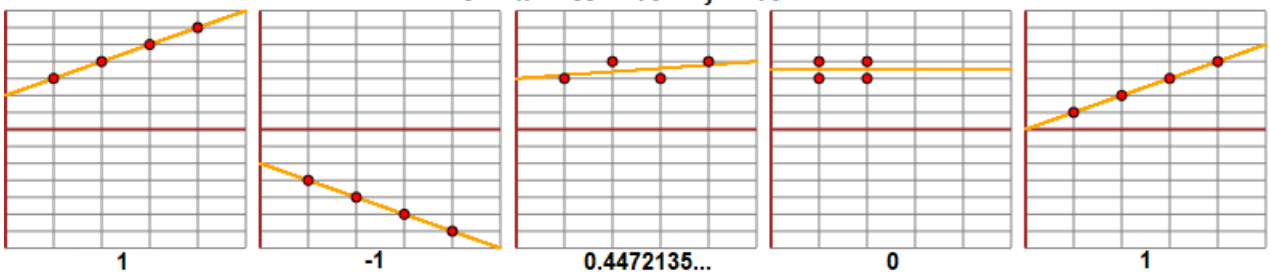
xM yM ← ↑ Round according to the context!

correlation coefficient

4 Calcola il coefficiente di correlazione usando lo script precedente tra X = (10, 20, 30, 40) e Y = (30, 40, 50, 60), tra X = (10, 20, 30, 40) e Y = (-30,-40,-50,-60), tra X = (10, 20, 30, 40) e Y = (30, 40, 30, 40), tra X = (10, 10, 20, 20) e Y = (30, 40, 30, 40), tra X = (10, 20, 30, 40) e Y = (10, 20, 30, 40).

Nelle figura seguente, relativa ai dati del quesito, le linee arancioni sono delle rette che approssimano i dati, ricavabili dallo stesso script, su cui (come già detto) ci soffermeremo successivamente.

0 <= x <= 50 -70 <= y <= 70



1 **-1** **0.4472135...** **0** **1**

L'impiego di un programma è praticamente indispensabile per analizzare tabelle di dati. Il software più usato (e gratuito) per le analisi statistiche (e per quasi ogni attività matemaatica) è **R**, che puoi vedere **qui** come scaricare e come impiegare. Per semplicità noi usermo degli script online.

Analizziamo i dati relativi a un'indagine ai 92 studenti di un corso universitario, tratta dal manuale del software *MiniTab*, raccolti nel file **battito**.

Se raccolti su una usuale tabella i dati assumerebbero l'aspetto qui a destra:

battito	64	58	62	...
bat.dopo corsa	88	70	76	...
fatta corsa	1	1	1	...
fumo	0	0	1	...
sesso	1	1	1	...
altezza	168	183	186	...
peso	64	66	73	...
attività fisica (0-3)	2	2	3	...

I dati sono stati rilevati durante una lezione di un corso universitario (almeno così viene detto in un manuale di *MiniTab* da cui essi sono stati tratti e parzialmente rielaborati – per presentarli nel sistema metrico decimale). La colonna "battiti dopo" si riferisce a un secondo rilevamento del battito cardiaco effettuato dopo che gli studenti a cui (lanciando una moneta) è uscito testa (1 nella colonna "corsa") hanno fatto una corsa di un minuto.

Uno strumento che ci serve, evidentemente, è quello che ci consenta estarre da una tabella i dati che soddisfino certe condizioni, ad esempio estrarre i dati realtivi al battito dopo la corsa solo in corrispondenza di chi fuma (il valore 1 della riga "fumo"). Lo script è **DaTabella**. Ecco come usarlo:

Two sequence are in **x** and **y**. I choose A and B and extract (and put in **z**) the values of **x** such that the corresponding values of **y** are included between A and B

```
88, 70, 76, 78, 80, 84, 84, 72, 75 118, 94, 96, 84, 76, 76, 58, 82, 72, 76, 80 106, 76 102,
94, 140, 100, 104, 100, 115, 112, 116, 118, 110, 98, 128, 62, 62, 74, 66, 76, 66, 56, 70,
74, 68, 74, 64, 84, 62, 58, 50, 62, 68, 54, 76, 84, 70, 88, 76, 66, 90, 94, 70, 70, 68, 84,
76, 66, 84, 70, 60, 92, 66, 70, 56, 74, 72, 80, 66, 76, 74, 78, 68, 68, 80, 76, 84, 92, 80,
68, 84, 76
```

1 <= y <= 1 **x** separated by "," ↑ - **y** separated by "," ↓ subtable C Cz

```
0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0,
0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1,
0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0,
0, 0
```

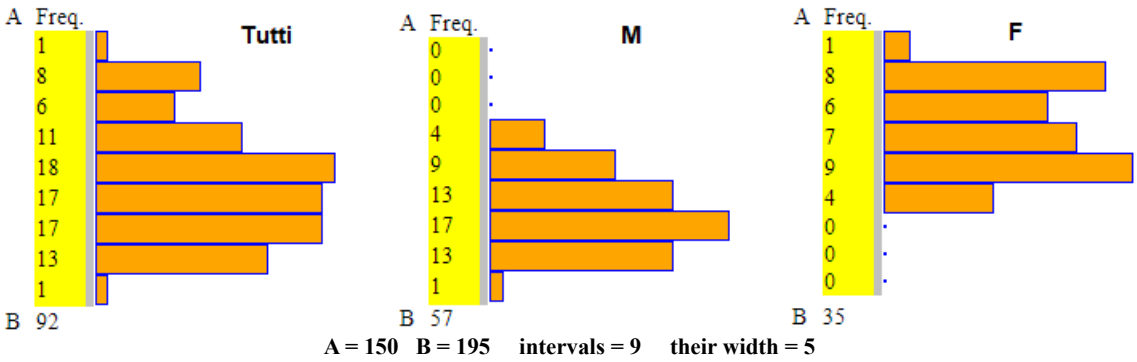
z: output ↓

```
76,  78,  96,  76,  72,  76,  80 106,  104,  112,  118,  62,  62,  66,  74,  68,  62,  50,
76,  84,  70,  90,  70,  68,  84,  56,  66,  80,  84
```

Con questo script posso anlizzare i dati relativi alle altezze scomponendoli in maschili e femminili. Poi posso anlizzarli con la **grande CT**. Ottengo:

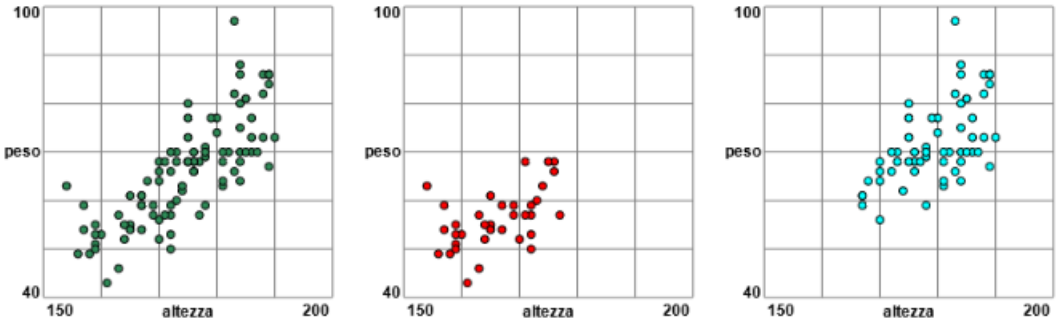
Tutti	**M**	**F**
n=92	n=57	n=35
min=154 max=190	min=167 max=190	min=154 max=177
median = 175	median = 181	median = 165
1^,3^ quartile: 167 183	1^,3^ quartile: 175 185	1^,3^ quartile: 159 172
mean = 174.43	mean = 179.60	mean = 166.03
experim. standard dev. = 9.34	experim. standard dev. = 6.61	experim. standard dev. = 6.64

Posso poi rappresentarli graficamente con lo script **Istogramma**:



Usando **RegCorr** posso analizzare la correlazione tra le diverse variabili. Ad esempio confrontando **Altezza** e **Sesso** (1: M, 2: F) ottengo **-0.709**, molto vicino a -1, a conferma che i maschi sono in genere più alti.

Analizzo analogamente la realzione tra **Altezza** e **Peso**. Ottengo **0.783**. Un valore molto alto. Se ci restringiamo a una sottopopolazione più omogenea (quella femminile o quella maschile, che hanno pesi e altezze con medie abbastanza diverse), mi potrei aspettare di ottenere un coefficiente maggiore. Ma se, dopo aver rappresentato graficamente la relazione tra altezza e peso, estraggo i maschi e estraggo le femmine, e rappresento la relazione anche in questi due casi ottengo:



Capisco che la forma allungata dell'insieme dei punti relativi all'intero campione è dovuta all'unione di due "nuvole" (quella dei maschi e quella delle femmine) centrate su baricentri disposti lungo una retta inclinatata.

Determinando i coefficienti di correlazione nei due casi troviamo effettivamente dei valori molto più bassi: per i maschi **0.590**, per le femmine **0.519**.

Questo esempio mette in luce come le ***statistiche*** che si ottengono sono spesso ***ingannevoli***. In casi come questo, abbastanza frequenti, il problema è dovuto alla presenza di due ***sottopopolazioni*** con caratteristiche differenti.

Poi, come già osservato ➡ introducendo i concetti probabilistici, occorre tener conto che quelle individuate sono solo relazioni statistiche, non di ***causa-effetto***. Ad esempio nel caso della correlazione tra le colonne "battito dopo" e "corsa" di "battito" c'è effettivamente una relazione causale (l'aver fatto la corsa influenza il battito cardiaco). Ma quando nel caso di uno studio statistico sulle condizioni delle famiglie è emersa una forte correlazione negativa fra il loro consumo di patate e la superficie dell'abitazione in cui vivono, essa non è da interpretare come conseguenza di una relazione di causa-effetto: è semplicemente dovuta al fatto che le famiglie benestanti abitano in genere in case di maggiori dimensioni e, nello stesso tempo, consumano meno patate delle altre famiglie privilegiando cibi più costosi, come la carne e il pesce. Purtroppo, specie nei campi medico e socio-psicologico, spesso si fanno collegamenti di questo genere.

Osserviamo, infine, che il coefficiente di correlazione è rilevante se i dati sono molti; basti pensare che avere tre punti più o meno allineati ha senz'altro un significato diverso dall'averne molti. Su questi aspetti ci soffermeremo in una prossima scheda.

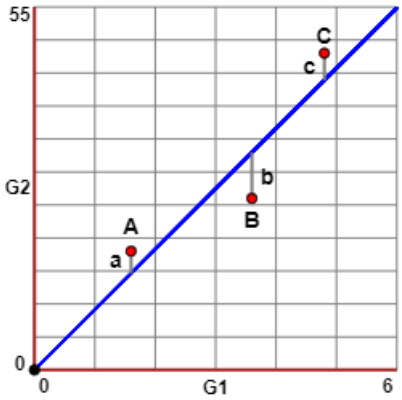
5 Studia, statisticamente, la relazione tra "battito prima" e "peso".

Nota. Le tabelle dotate di indici possono essere dotate di una particolare struttura algebrica; in questo ambito, vengono chiamate ***matrici***. Se interessato, vedi **qui**.

5. Rette di regressione

Di fronte a dati sperimentali relativi a un sistema (X,Y) per cui si ritiene che Y vari *in funzione* di X, si può cercare di trovare una funzione F tale che il suo grafico approssimi i punti sperimentali. Vediamo come procedere nel caso in cui X ed Y siano casuali. Si cerca di individuare il tipo di funzione (lineare, polinomiale, esponenziale, …) che si vuole utilizzare. Se si ipotizza che ci sia una relazione *lineare* che esprima Y in funzione di X, e non si hanno altre informazioni, la tecnica in genere usata è quella dei ***minimi quadrati***. Illustriamola su un semplice esempio.

Voglio approssimare la relazione tra due variabili casuali G1 e G2 con una relazione del tipo G2 = k G1. Devo determinare k. In tre esperimenti ottengo le coppie (G1,G2) rappresentate a lato con i punti A, B e C. Per determinare k decido di cercare, tra tutti i grafici del tipo G2 = k G1 (rette passanti per l'origine) come quello raffigurato a fianco, il grafico che rende *minima la somma dei quadrati degli scarti* tra i valori di G2 sperimentali e quelli che sarebbero stati associati ai valori G1 dalla relazione G2 = k G1. Ossia cerco la pendenza k che deve avere una retta passante per l'origine affinché la somma dei quadrati di **a**, **b** e **c** (vedi fig. a lato) sia minima.



$\mathbf{a} = |y_A - k\,x_A|$, $\mathbf{b} = |y_B - k\,x_B|$, $\mathbf{c} = |y_C - k\,x_C|$.

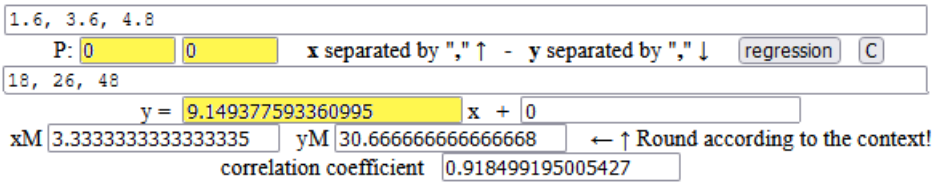
$\mathbf{a}^2 + \mathbf{b}^2 + \mathbf{c}^2 = (k\,x_A - y_A)^2 + (k\,x_B - y_B)^2 + (k\,x_C - y_C)^2$.

È un'espressione *polinomiale* in k di *secondo grado.* Assume valore minimo quando si annulla la sua derivata rispetto a k, ossia quando:

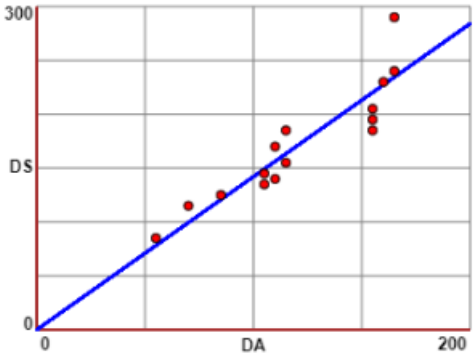
$2(k\,x_A - y_A)x_A + 2(k\,x_B - y_B)x_B + 2(k\,x_C - y_C)x_C = 0$

ossia: $k = (x_A y_A + x_B y_B + x_C y_C)/(x_A^2 + x_B^2 + x_C^2)$

I calcoli sono abbastanza facili. Comunque, per fare prima, e ridurre la possibilità di commettere errori, possiamo ricorrere a **RegCorr**. Se i punti sono (1.6,18), (3.6,26), (4.8,48), imponendo che la retta passi per (0,0), trovo y = 9.1494*x:



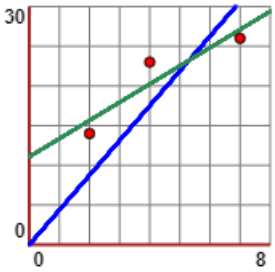
Un esempio concreto. Consideriamo il file terraria in cui sono contenute un po' di coppie (DA, DT) che rappresentano le *distanze chilometriche* da Genova *in linea d'aria* e *lungo la strada* di alcune città italiane (a lato è tracciato il grafico di dispersione). Prima di analizzare i dati cerchiamo di congetturare con un ragionamento teorico come potrebbe essere fatta una funzione che approssimi la relazione tra DT e DA: la distanza lungo la strada sicuramente è maggiore di quella in linea d'aria; supposto di essere in una regione dalle caratteristiche geografiche non molto diversificate e che non presenti territori invalicabili (a causa di catene montuose a picco, di insenature o laghi molto grandi) essa dovrebbe crescere più o meno proporzionalmente alla distanza in linea d'aria; per stimare il rapporto tra l'una e l'altra, tenendo conto delle curve supponiamo che esso sia circa pari al rapporto che c'è tra la strada per raggiungere due vertici opposti di un quadrato passando per il bordo o la strada diretta, cioè √2 circa.



6 Confronta con √2 il coefficiente della retta di regressione passante per (0.0) che trovi con l'opportuno script.

Negli esempi precedenti (vedi la risposta al precedente esercizio qui) abbiamo cercato una F: x → *a*x+*b* per cui sia minimo $(F(X_1)-Y_1)^2+(F(X_2)-Y_2)^2+\ldots+(F(X_n)-Y_n)^2$, cioè la somma dei quadrati degli scarti tra i valori Y_i sperimentali e quelli che si avrebbero applicando F ai corrispondenti X_i nel caso particolare in cui si imponeva che *b* fosse 0, ossia che la retta passasse per l'origine. In modo simile a quanto fatto sopra, ma in modo più complesso (che vedremo l'anno prossimo) si possono ricavare i valori di *a* e di *b* nel caso in cui non si pongano vincoli sulla retta cercata. Vediamo, per ora, solo come questi valori posso essere ottenuti con lo script precedente:

7 Dati i punti (2,14), (4,23), (7,26), stabilisci (usando lo script) qual è la corrispondente retta di regressione vincolata a passare per (0,0) e qual è quella senza vincoli (arrotondda i coefficienti a 4 cifre).



x → 4.3__ * x
x → 2.___ * x + 11.__

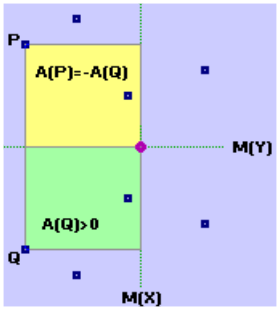
Abbiamo visto solo due metodi per approssimare punti sperimentali con funzioni lineari. La retta che approssima i punti viene chiamata ***retta di regressione*** e il coefficiente direttivo di essa è chiamato ***coefficiente di regressione***. Nel caso in cui non si vincoli la retta a passare per un punto, si ottiene che essa passa per il "baricentro" (M(X), M(Y)), come avrai già intuito osservando gli esiti dello script.

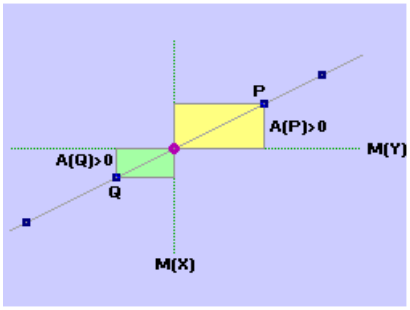
Vi sono altri modi per approssimare con funzioni dei dati. Ne vedremo alcuni il prossimo anno.

6. Approfondimenti

La formula ***Cov(X,Y)*** = M((X–M(X))·(Y–M(Y))) può essere motivata in vari modi. Ad es. si può interpretarla come un indicatore che assume un valore assoluto che scende quanto più i punti tendono a disporsi in modo da presentare una simmetria verticale o orizzontale e che cresce quanto più i punti tendono a disporsi lungo una retta obliqua. Infatti le componenti della sommatoria rappresentano aree "con segno" di rettangolini che hanno come dimensioni le distanze "con segno" delle coordinate dei punti dalle coordinate del baricentro. Nella figura sotto a sinistra (simmetria orizzontale) le componenti della sommatoria due a due si annullano, per cui la covarianza è nulla. Se schiaccio obliquamente la nuvola di punti la compensazione diventa solo parziale. Nella caso della figura a destra (X e Y in relazione lineare) non c'è alcuna compensazione (componenti tutte positive).

Il segno sarà uguale al segno della pendenza della retta lungo cui i punti tendono a disporsi.





Se X e Y sono indipendenti, ci aspettiamo che la covarianza sia nulla. E infatti:

$$\mathrm{Cov}(X,Y) = M((X-M(X))(Y-M(Y))) = M((X-M(X)))\cdot M((Y-M(Y))) = 0\cdot 0 = 0.$$

L'interpretazione geometrica ci facilita la comprensione che $M((X-M(X))(Y-M(Y)))$ è 0 anche nel caso in cui fissando diversi valori di X la Y continua a variare in modo "analogo" attorno sempre allo stesso valor medio, e, viceversa, fissando … . Cioè se all'aumentare di una delle due variabili l'altra non tende a modificarsi. Ovviamente, una persona "allenata" può capire ciò direttamente dalla formula.

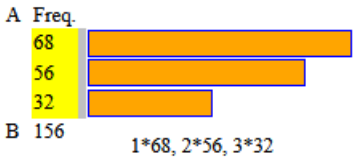
Un'altra possibile interpretazione è basata sull'osservazione che $\mathrm{Cov}(X,Y) = M(X\cdot Y)-M(X)\cdot M(Y)$: la covarianza è un indicatore dello scarto di $M(X\cdot Y)$ da $M(X)\cdot M(Y)$, cioè dal valore che $M(X\cdot Y)$ assumerebbe nel caso della indipendenza.

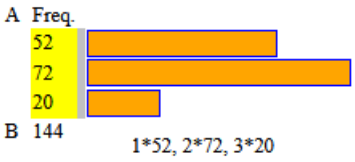
7. Esercizi

e1 Vengono scelti a caso 300 statunitensi, che vengono suddivisi per convinzioni politiche e per sesso. Si ottiene la seguente tabella di contingenza:

	democratici	repubblicani	indipendenti
femmine	68	56	32
maschi	52	72	20

Che cosa rappresentano i grafici a fianco? Come sono stati tracciati? Valuta l'indipendenza del sesso e dell'orientamento politico.

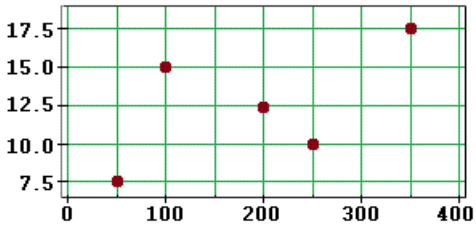




e2 In quali casi le seguenti variabili casuali X e Y sono praticamente indipendenti?

(1) X = altezza di un uomo adulto; Y = sua età.
(2) X = altezza di un bambino; Y = sua età.
(3) X = altezza di un uomo adulto; Y = suo peso.
(4) X = peso di una pietra; Y = sua massima lunghezza.

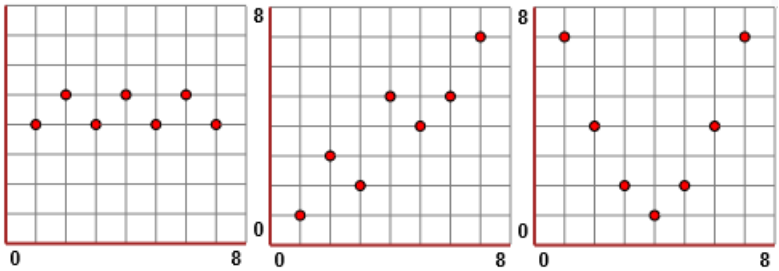
e3 Ecco rappresentati graficamente i dati relativi a due variabili casuali X e Y. Quanto vale il coefficiente di correlazione tra X e Y (svolgi il calcolo sia "a mano" che con del software, e confronta i risultati ottenuti).



e4 Studia, usando opportuni concetti statistici e i dati contenuta nella scheda, il legame tra battito cardiaco e sesso.

e5 Facendo riferimento ai dati considerati in **e3**, individua le retta di regressione di Y in funzione di X e quella di X in funzione di Y, e rappresentale graficamente sullo stesso sistema di riferimento.

e6 Sotto sono rappresentate graficamente tre sequenze di coppie di dati. Ad una corrisponde una correlazione pari a circa 0.9, ad un'altra una pari a 0, ad un'altra una pari ad una delle due precedenti. Associa ad ogni sequenza di dati la corrispondente correlazione. Controlla la tua risposta calcolando i tre coefficienti di correlazione.



e7 Per studiare il legame tra la temperatura ambientale e il numero di parti difettose che escono da una particolare linea di produzione un'azienda registra per circa una ventina di giorni le temperature massime e la quantità dei difetti riscontrati, ottenendo i dati allegati. Traccia il relativo diagramma di dispersione, calcola il coefficiente di correlazione lineare e valuta se puoi concludere qualcosa circa la relazione tra temperatura e parti difettose prodotte.

1) Segna con l'evidenziatore, nelle parti della scheda indicate, frasi e/o formule che descrivono il significato dei seguenti termini:

indipendenza probabilistica (§1), *tabella di contingenza* (§2), *funzioni di densità (bivariate)* (§2), *coefficiente di correlazione* (§3), *minimi quadrati* (§3).

2) Su un foglio da "quadernone", nella prima facciata, esemplifica l'uso di ciascuno dei concetti sopra elencati mediante una frase in cui esso venga impiegato.

3) Nella seconda facciata riassumi in modo discorsivo (senza formule, come in una descrizione "al telefono") il contenuto della scheda (non fare un elenco di argomenti, ma cerca di far capire il "filo del discorso").

script: piccola CT grande CT isto isto con % boxplot striscia 100 ordina Grafici GraficD divisori Indet distanza Triang eq.polinomiale eq.nonPolin sistemaLin moltPolin sempliciEq divisori fraz/mcd opFraz SumPro sin LenArc Poligono Circ3P Inscr3P IntegrPol Istogramma RandomNum IntGauss AB3dim TabFun Det3 DaTabella RegCorr ALTRO