ESERCIZI DI STATISTICA RISOLTI
Federico Emanuele Pozzi

Risolverò solo un compito integralmente. Se avete domande sulla risoluzione di specifici esercizi postate nel forum, e le aggiungerò qui.
Qui presento solo i procedimenti, senza addentrarmi troppo nella teoria (che potete trovare nelle dispense).

26 Febbraio 2014

A) Data la distribuzione di frequenze della variabile "durata della gestazione in settimane" sotto riportata:

	X	f	fX	fX^2
1	37	11	407	15059
2	38	34	1292	49096
3	39	129	5031	196209
4	40	103	4120	164800
5	41	70	2870	117670
6	42	3	126	5292
TOTALI		350	13846	548126

Calcolare:

- **Media**
- **Ds**
- **Mediana**
- **10° centile**
- **range IQ**

Soluzione:

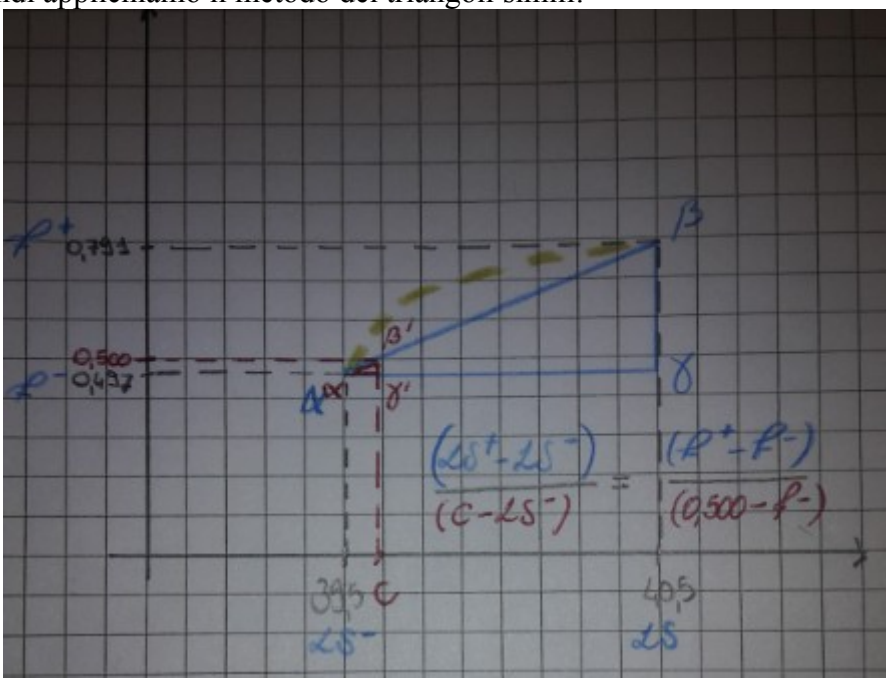
- Media: $\frac{\sum f_i \cdot X_i}{\sum f_i} = \frac{13846}{350} = 39.56$
- Deviazione standard: ci sono due modi per calcolarla (assolutamente equivalenti, dato che la formula del secondo modo è ricavabile dalla formula del primo)
  - $\sqrt{\frac{\sum f_i (X_i - \mu)^2}{\sum f_i}} = \sqrt{\frac{378{,}23}{350}} = 1{,}04$
  - $\sqrt{\frac{\sum f_i \cdot X_i^2 - \frac{(\sum f_i X_i)^2}{\sum f_i}}{\sum f_i}} = \sqrt{\frac{548126 - \frac{13846^2}{350}}{350}} = 1{,}04$
    - Perché è stato usato n e non n-1 nei denominatori? Perché in questo caso abbiamo una popolazione e non un campione; i dati ricavati sono oggetto di descrizione, non base per fare inferenza.
- Mediana: la mediana è quel valore che divide la popolazione in due metà – nel nostro caso ci saranno quindi 175 valori prima della mediana e 175 dopo.
  Osserviamo che per creare una tabella del genere, supponendo che la variabile in esame sia continua (è un tempo infatti, non è che tutte le madri partoriscono a 37, 38, 39 ecc settimane ESATTE) abbiamo diviso i valori in CLASSI, utilizzando come riferimento il valore centrale della classe.

Le nostre classi saranno ovviamente della forma [36,5:37,5] [37,5:38,5] e così via.
Il metodo da seguire è già stato spiegato nella parte teorica delle dispense (§4.1), qui lo riporterò brevemente:

Si crea la tabella delle frequenze cumulative di ciascuna classe:

Classe	Frequenza cumulativa
36,5 – 37,5	0,031
37,5 – 38,5	0,129
38,5 – 39,5	0,497
39,5 – 40,5	0,791
40,5 – 41,5	0,991
41,5 – 42,5	1,000

Ricerchiamo la classe che contiene il 50° centile (cioè la mediana).
È la classe 39,5-40,5.
Quindi applichiamo il metodo dei triangoli simili:



Come possiamo vedere dal disegno, i triangoli blu e rosso sono simili, pertanto vale la relazione scritta, che conduce alla formula

$$C = LS^{-} + (LS^{+} - LS^{-})\frac{(0{,}500 - F^{-})}{(F^{+} - F^{-})} = 39{,}5 + (40{,}5 - 39{,}5)\frac{(0{,}500 - 0{,}497)}{(0{,}791 - 0{,}497)} = 39{,}51 \quad .$$

Abbiamo usato l'approssimazione lineare per la distribuzione dei valori all'interno della classe (che tuttavia potevano anche seguire la curva verde tratteggiata tanto per fare un esempio).
Sostituendo 0,500 con qualsiasi valore di centile possiamo ricavare il valore corrispondente, e così faremo per determinare i centili quando abbiamo esercizi di STATISTICA DESCRITTIVA (quando cioè dobbiamo ricavare i centili a partire dai dati e non dai modelli).

- 10° centile: applichiamo il procedimento appena usato:
  La classe di interese è 37,5-38,5. La formula sarà
  $$C = 37{,}5 + (38{,}5 - 37{,}5)\frac{(0{,}100 - 0{,}031)}{(0{,}129 - 0{,}031)} = 38{,}2$$
- Range IQ: applichiamo il procedimento appena usato per ricavare 25° e 75° centile.

$$25\,°C = 38{,}5 + (39{,}5 - 38{,}5)\frac{(0{,}250 - 0{,}129)}{(0{,}497 - 0{,}129)} = 38{,}82$$

$$75\,°C = 39{,}5 + (40{,}5 - 39{,}5)\frac{(0{,}750 - 0{,}497)}{(0{,}791 - 0{,}497)} = 40{,}36$$

Per trovare il range basta ora fare la differenza: 40,36-38,82=1,54

B) Per una gaussiana in cui ¼ dei valori centrali sono compresi tra 170 e 180, calcolare:

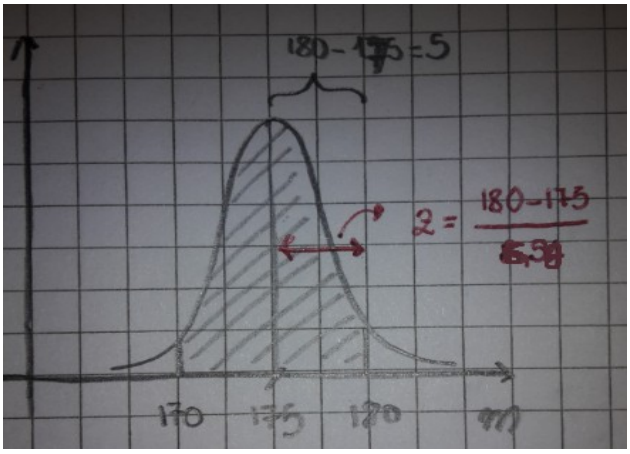
- **1) μ, σ, 10° centile, range IQ, 95° centile**
- **2) Probabilità che:**
  - **a) solo 1 di 8 valori estratti a caso sia compreso fra 170 e 180**
  - **b) più di 24 di 100 valori estratti a caso siano compresi tra 170 e 180**
  - **c) la media di 5 valori estratti a caso sia compresa tra 170 e 180**
  - **d) la differenza di 2 valori estratti a caso sia maggiore di 10**
  - **e) almeno 2 di 8 valori estratti a caso siano compresi tra 170 e 180**
  - **f) se il primo estratto non è compreso fra 170 e 180, lo sia il secondo**
- **3) Numero minimo di valori da estrarre per avere probabilità almeno del 95% di osservare:**
  - **g) almeno un valore compreso tra 170 e 180**
  - **h) che il minimo valore estratto sia inferiore e il massimo superiore a 180**

Soluzione:

- 1) La media è 175 (170 e 180 sono gli estremi di un intervallo di valori CENTRALI, quindi centrato sulla media, che deve essere a metà strada tra i due estremi).
  La deviazione standard si calcola a partire dalla guassiana standard.
  Tra 175 e 180 ho il 12,5% di valori (infatti è la metà dell'intervallo che contiene ¼, cioè il 25%, dei valori); dunque da 180 in poi ho il 37,5%. Ricavo lo z corrispondente dalla tabella – vale circa 0,32.
  A questo punto sostituendo nella relazione $z = \frac{(x_i - \mu)}{\sigma}$ ottengo $\sigma = (180 - 175)/0{,}32 = 15{,}62$.
  Per calcolare il 10° centile utilizziamo lo stesso procedimento. Ricaviamo lo z corrispondente a 10 (1,282) dalla tabella dell'integrale normale degli errori. Poiché il 10° centile è minore della media, la formula sarà $-1{,}282 = \frac{(x_{10} - 175)}{15{,}62} \rightarrow x_{10} = -1{,}282 \cdot 15{,}62 + 175 = 154{,}97$.
  Per calcolare il range interquartile dovrò calcolare il 25° e il 75° centile con il metodo appna visto (ricavando lo z corrispondente e osservando che una volta lo dovrò prendere positivo ed una volta negativo);
  $x_{25} = -0{,}6745 \cdot 15{,}62 + 175 = 164{,}46 \qquad x_{75} = 0{,}6745 \cdot 15{,}62 + 175 = 185{,}53$
  Quindi, faccio la differenza tra i due valori: range IQ=185,53-164,46=21,07
  Per calcolare il 95° centile farò la stessa cosa, trovando ovviamente lo z corrispondente a 5 (infatti la tabella dell'integrale normale degli errori vi dà la probabilità SOPRA un certo valore). Ottengo $x_{95} = 1{,}645 \cdot 15{,}62 + 175 = 200{,}69$.
- 2) Divido la soluzione per punti:
  a) In questo caso utilizzo la binomiale. La probabilità di successo è 0.25, i tentativi sono 8 e i successi 1; di conseguenza $P(1;8;0.25) = \frac{8!}{(8-1)!\,1!} \cdot 0{,}25^1 \cdot 0{,}75^7 = 0{,}27$
  b) In questo caso utilizzo l'approssimazione gaussiana della binomiale, con media nπ e deviazione standard $\sqrt{n\pi(1-\pi)}$. Osservo che la media è 0,25x100=25 e la deviazione standard 4,33.
  Devo ricordarmi ora di utilizzare la correzione per il continuo (siccome la variabile della

gaussiana è continua, ma i successi sono discreti, devo utilizzare le classi); pertanto, più di 24 successi significa da 24,5 in poi. Calcolo quindi lo z corrispondente a 24,5 come $(24{,}5-25)/4{,}33=-0{,}115$, cui corrisponde una probabilità dello 0,0438. Osserviamo che questa è la probabilità di osservare 24,5 o meno successi, pertanto la probabilità che cercavamo è 1-0,4522=0,5478.

c) Vedere il paragrafo sulla distribuzione di campionamento di m per la teoria. Poiché la distribuzione delle medie campionarie è gaussiana intorno alla media della popolazione (175), con deviazione standard pari all'errore standard della media (calcolato come $\frac{\sigma}{\sqrt{n}}$), calcolo il valore di z corrispondente al range indicato (170-180).



Come potete vedere facilmente dal disegno, il calcolo è $z=\frac{180-175}{\frac{15{,}62}{\sqrt{5}}}=0{,}71$. Dalla tabella dell'integrale normale degli errori ricavo la probabilità corrispondente: 0,238. Questa è la probabilità a DESTRA di z=0,71. Quindi quella compresa tra la media e z=0,71 sarà 0,5-0,238=0,262; siccome la gaussiana è simmetrica, l'area compresa tra due scarti di 0,71σ dalla media è 0,262x2=0,524.

d) Il procedimento è un po' laborioso. Tenete presente che negli esami della sessione 2014-2015 questo tipo di esercizi non ci sarà.

Possiamo considerare il tipo di problema presentato come l'estrazione di un campione di due elementi. Questo campione avrà una sua media m e una sua varianza s^2. Ragionandoci un po' su, vi renderete conto che la differenza tra i due valori può essere riscritta come (maggiore-media)+(media-minore), come (maggiore-media)=-(media-minore) per le proprietà della media.

Pertanto, dire che i valori estratti devono avere una differenza maggiore di 10 equivale a dire che (maggiore-media)+(media-minore)=10. Ciò comporta che maggiore-media=5, e media-minore=-5.

Se la differenza tra i due valori fosse esattamente 10, la devianza del campione sarebbe $5^2+5^2=50$; pertanto, è chiaro che a noi interessa trovare quanto è probabile estrarre un campione con una devianza MAGGIORE di 50 (o una varianza s^2 maggiore di 50, dato che in questo caso si divide per n-1=2-1=1).

Confrontate col paragrafo della teoria sulla distribuzione di campionamento della varianza; ciò equivale a trovare la probabilità di osservare un chi quadro pari o maggiore di $s^2\frac{(n-1)}{\sigma^2}$ con n-1 gradi di libertà.

Sostituendo i nostri valori il valore di chi quadro è $50\frac{1}{15{,}62^2}=0{,}204$. Se ora guardiamo la tabella del chi quadro osserviamo che a questo valore, per un grado di libertà, corrisponde una probabilità del 65%.

e) Per calcolare questa probabilità basta fare riferimento al punto a). Infatti possiamo

vederla come 1-(probabilità 1 successo)-(probabilità 0 successi). La probabilità di 1 successo già la conosciamo, e vale 0,27; quella di 0 successi la calcoliamo in modo analogo, e vale 0,10.

Quindi la probabilità di ottenere 2 o più successi è 1-0,37=0,63.

f) In questo caso si applica la probabilità condizionata P(A|B), con P(A)=0,25 e P(B)=0,75. Poiché i due eventi sono indipendenti, la loro intersezione è vuota e la probabilità condizionata si riduce a P(A). Quindi P(A|B)=0,25.

- 3) Divido per punti:

  g) Trovare il numero di tentativi per cui la probabilità di ottenere almeno un successo nel 95% dei casi equivale al trovare il numero di tentativi per cui la probabilità di non ottenere nemmeno un successo è del 5%.

  Poiché la probabilità di successo è dello 0.25, la probabilità di insuccesso è dello 0.75.

  Quindi vale la relazione $0{,}75^n < 0{,}05$, e per ricavare n passiamo ai logaritmi: $n \log 0{,}75 < \log 0{,}05$, che dà n>10,4. Pertanto per osservare almeno un successo con una probabilità del 95% dovremmo fare al minimo 11 estrazioni.

  h) 180 è il 62,5° centile. Quindi un valore estratto a caso ha il 37,5% di probabilità di essere superiore e il 62,5% di probabilità di essere minore di 180.

  Se devo avere almeno il 95% di probabilità di osservare almeno un valore inferiore e almeno un valore superiore a 180 significa che la probabilità di osservare SOLO valori superiori a 180 sommata alla probabilità di osservare SOLO valori inferiori a 180 deve essere minore o uguale al 5%.

  Quindi potrò scrivere $0{,}625^n + 0{,}375^n < 0{,}05$. In questo caso non posso applicare i logaritmi, perché compaiono somme tra esponenziali. Applicare il metodo di risoluzione per sostituzione non è efficace. Pertanto, possiamo procedere in due modi:

  - Per tentativi: provando con n=2, n=3, n=4 ecc finché la disequazione non è soddisfatta (metodo triste). Per n=7 otteniamo 0,038.
  - Vediamo che 0,375 è piccolo, e già elevato alla seconda dà 0,14. Quindi questo termine tenderà velocemente a 0 e possiamo trascurarlo. La disequazione APPROSSIMATA diventa $0{,}625^n < 0{,}05$, e possiamo usare i logaritmi; con $n \log 0{,}625 < \log 0{,}05$ otteniamo n>6,37, quindi n=7 come trovato col metodo precedente.

C) L'efficacia di due trattamenti A e B fu confrontata randomizzando 250 pazinti (A=127, B=123). Avendo osservato rispettivamente 77 e 99 guarigioni:

- **Stimare differenza di efficacia**
- **Calcolare NNT**

Soluzione:

- N.B.: in realtà l'esercizio C era più lungo, ma il programma svolto nell'anno 2014-2015 non permetteva di affrontare più argomenti.
- Per stimare la differenza di efficacia bisogna innanzitutto calcolare l'efficacia dei due trattamenti. È buona norma costruire la tabella di contingenza:

	A	B	Totali
Guariti	77	99	176
Non guariti	50	24	74
Totali	127	123	250

L'efficacia del farmaco A è data da $\frac{guariti}{totale\,curati\,con\,A} = \frac{77}{127} = 0{,}606$, mentre B analogamente è $\frac{99}{123} = 0{,}804$. La differenza di efficacia è pertanto 0,804-0,606=0,198.

- NNT è il number needed to treat, ovvero il minimo numero di pazienti da trattare col farmaco B per osservare un miglioramento significativo rispetto al farmaco A (è evidente che tanto più questo numero è basso, tanto più il farmaco B è efficace rispetto al farmaco A).
  Si calcola come il reciproco della differenza di efficacia: in questo caso $\frac{1}{0{,}198}=5{,}05$.
  Poiché non possiamo curare 5,05 pazienti, ma solo un numero naturale di pazienti, NNT=6 (si arrotonda sempre per eccesso).

D) In uno studio di accuratezza diagnostica, di 95 malati 88 sono risultati VP, mentre di 295 non malati 35 furono i FP. Stimare SE, SP (con i rispettivi intervalli di confidenza al 95%), LR+ e LR-.
Applicando questo test ad un paziente cui si assegni probabilità di malattia pre-test del 10%, qual è la stima di probabilità post test in caso di test positivo e in caso di test negativo?

Soluzione:

- Anche in questo caso ricaviamo la tabella di contingenza:

	Positivi	Negativi	Totali
Malati	88	7	95
Sani	35	260	295
Totali	123	267	390

La sensibilità è $\frac{VP}{VP+FN}=\frac{88}{95}=0{,}926$.
La specificità è $\frac{VN}{VN+FP}=\frac{260}{295}=0{,}881$.
Per calcolare gli intervalli di confidenza al 95% di proporzioni si utilizza come errore standard $\sqrt{\frac{\pi(1-\pi)}{n}}$, pertanto l'intervallo di confidenza sarà $[\pi-1{,}96\sqrt{\frac{\pi(1-\pi)}{n}}:\pi+1{,}96\sqrt{\frac{\pi(1-\pi)}{n}}]$.
Per la sensibilità $[0{,}926-1{,}96\sqrt{\frac{1-0{,}926}{95}}:0{,}926+1{,}96\sqrt{\frac{1-0{,}926}{95}}]=[87{,}3:98{,}0]$.
Per la specificità [0,844: 0,918].

- Il LR+ (likelihood ratio positivo) è uguale a $\frac{SE}{1-SP}=7{,}78$.
  Il LR- è $\frac{1-SE}{SP}=0{,}08$.
- Per calcolare la probabilità post test per il test positivo si calcola l'odds pre-test e lo si moltiplica per LR+, ricavando l'odds post test; quindi si ri-ricava la probabilità dall'odds. Analogamente per la probabilità post test per il test negativo usando LR-.
  $Odds\ pre-test=\frac{0{,}1}{0{,}9}=\frac{1}{9}$.
  $Odds\ post\ test+ =\frac{1}{9}7{,}78=0{,}86 \quad\rightarrow\quad Probabilità\ post\ test+ =\frac{Odds}{Odds+1}=\frac{0{,}86}{1{,}86}=0{,}46$.
  $Odds\ post\ test- =\frac{1}{9}0{,}08=0{,}008 \quad\rightarrow\quad Probabilità\ post\ test- =\frac{0{,}008}{1{,}008}=0{,}008$.

E) In un campione di 30 volontari è stata rilevata età (x) e pressione sistolica (y), ottenendo:
$\sum x=1354 \quad \sum x^2=67894 \quad \sum y=3935 \quad \sum y^2=518576 \quad \sum xy=179230$ **.**
Stimare:

- **Valore medio della pressione arteriosa nella popolazione campionata**
- **Coefficiente di correlazione lineare**
- **Coefficiente di regressione lineare**
- **Pressione sistolica attesa in un soggetto di 37 anni di età**

Soluzione:

- Il valor medio della pressione è dato da $\frac{\sum y}{30}=\frac{3935}{30}=131{,}16$.
- Il coefficiente di correlazione lineare r è dato da $\frac{D_{xy}}{\sqrt{D_{xx}D_{yy}}}=\frac{\sum xy-\frac{\sum x\sum y}{n}}{\sqrt{(\sum x^2-\frac{(\sum x)^2}{n})(\sum y^2-\frac{(\sum y)^2}{n})}}$, e sostituendo coi valori dati otteniamo $r=\frac{179230-\frac{1354\cdot 3935}{30}}{\sqrt{(67894-\frac{1354^2}{30})(518576-\frac{3935^2}{30})}}=0{,}40$.
- Il coefficiente di regressione lineare β è calcolato come $\frac{D_{xy}}{D_{xx}}=\frac{\sum xy-\frac{\sum x\sum y}{n}}{\sum x^2-\frac{(\sum x)^2}{n}}$, e nel nostro caso $\beta=\frac{178230-\frac{1354\cdot 3935}{30}}{67894-\frac{1354^2}{30}}=0{,}24$.
- Per calcolare la pressione sistolica attesa in un soggetto di 37 anni di età dobbiamo determinare la relazione tra età e pressione, per la quale ci manca ancora il coefficiente α, calcolabile come $\frac{\sum x^2\sum y-\sum x\sum xy}{n\sum x^2-(\sum x)^2}$, che nel nostro caso dà $\alpha=\frac{67894\cdot 3935-1354\cdot 179230}{30\cdot 67894-1354^2}=120{,}31$.
  La retta che descrive il variare delle pressioni in base all'età è quindi y=0,24x+120,31.
  Per calcolare la pressione attesa in un soggetto di 37 anni basta sostituire x=37, ottenendo quindi $y=0{,}24\cdot 37+120{,}31=129{,}19$.