

Esercizi di statistica descrittiva - soluzioni

1. Calcola la deviazione standard di 0, 1, 1.5, 2, 3

2. " 3, 3, 6, 6, 6, 12

Le formule che si possono usare per rispondere a questa domanda sono due, entrambe conducono alla risposta corretta

$$\sigma = \sqrt{\text{media dei quadrati degli scarti}}$$

$$\sigma = \sqrt{\text{media dei quadrati} - \text{media}^2}$$

Se vogliamo applicare la prima formula all'esercizio 1, calcoliamo la media

$$\mu = \frac{0 + 1 + 1.5 + 2 + 3}{5} = 1.5$$

a questo punto gli scarti sono

$$0 - \mu = -1.5 \quad 1 - \mu = -0.5 \quad 1.5 - \mu = 0 \quad 2 - \mu = 0.5 \quad 3 - \mu = 1.5$$

quindi

$$\sigma = \sqrt{\frac{(-1.5)^2 + (-0.5)^2 + 0^2 + 0.5^2 + 1.5^2}{5}} = \sqrt{1} = \underline{1}$$

Usando, invece, la seconda formula, otteniamo

$$\sigma = \sqrt{\frac{0^2 + 1^2 + 1.5^2 + 2^2 + 3^2}{5} - 1.5^2} = \sqrt{3.25 - 2.25} = \underline{1}$$

L'esercizio 2 si può affrontare alla stessa maniera

$$\mu = \frac{3+3+6+6+6+12}{6} = 6 \quad \text{ripetizioni}$$

$$\sigma = \sqrt{\frac{(3-6)^2 \cdot 2 + (6-6)^2 \cdot 3 + (12-6)^2}{6}} = \sqrt{\frac{18+0+36}{6}} = \underline{\underline{3}}$$

A volte, calcoli come questi possono essere abbreviati dalle seguenti osservazioni:

- Sommando una costante A a tutti i numeri di un certo insieme, si somma A alla loro media, mentre la deviazione standard non cambia.
- Moltiplicando tutti i numeri di un certo insieme per la medesima costante B , sia la loro media sia la loro deviazione standard si moltiplicano per B .

Sfruttando queste osservazioni, possiamo, per esempio, risolvere l'esercizio 2 così. Sommiamo -6 ai numeri dati ottenendo

$-3 \quad -3 \quad 0 \quad 0 \quad 0 \quad 6$

Questo non altera la deviazione standard, ora moltiplichiamo per $\frac{1}{3}$ ottenendo

$-1 \quad -1 \quad 0 \quad 0 \quad 0 \quad 2$

Questo moltiplica la deviazione standard per $\frac{1}{3}$.

Abbiamo quindi ottenuto che

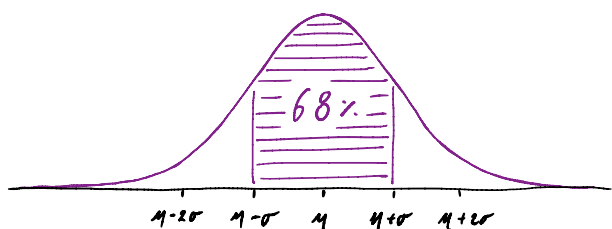
$$\begin{aligned}\frac{1}{3} \sigma &= \text{deviazione standard di } -1, -1, 0, 0, 0, 2 \\ &= \sqrt{\frac{(-1)^2 \cdot 2 + 0^2 \cdot 3 + 2^2}{6} - \left(\frac{(-1) \cdot 2 + 0 \cdot 3 + 2}{6} \right)^2} = \\ &= \sqrt{1 - 0} = 1\end{aligned}$$

Da cui ricaviamo $\sigma = 3$

3. In una dist. gaussiana, che % della popolazione ricade entro una deviazione standard dalla media?
4. IDOM entro due d.st. dalla media?
5. IDOM fra la media e la media più una d. st.?
6. IDOM fra media meno due d.st. e media?

A queste domande possiamo rispondere ricordando due fatti mnemonici:

- che il 68% della popolazione ricade entro una d.st.
- che il 95% della popolazione ricade entro due d.st.



Con entrambe le percentuali arrotondate all'intero più vicino.

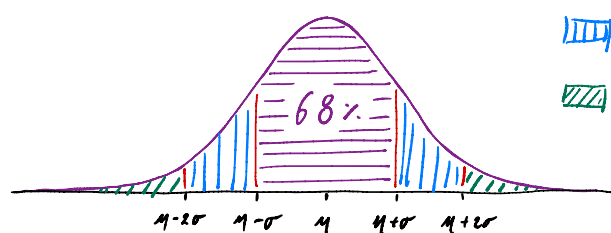
Questi due numeri ci forniscono, in maniera diretta, la

risposta alle prime due domande

3. In una dist. gaussiana, che % della popolazione ricade entro una deviazione standard dalla media? 68%

4. 100% entro due d.st. dalla media? 95%

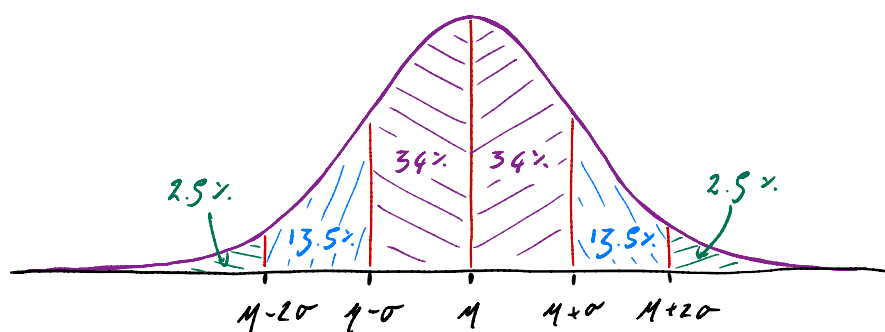
Adesso, sottraendo le due aree fra di loro, e dal totale di 100% possiamo ottenere



$$95\% - 68\% = 27\%$$

$$100\% - 95\% = 5\%$$

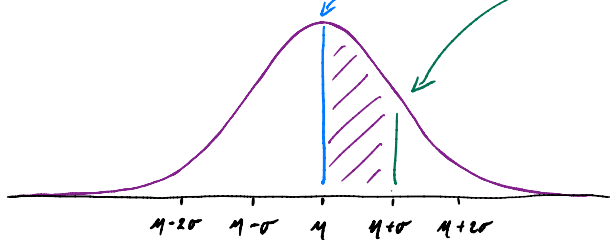
Da cui, sfruttando la simmetria della gaussiana



Questa figura permette di rispondere immediatamente alle rimanenti domande — non conviene tuttavia memorizzare le cifre in questa figura, perché si possono ricavare in maniera elementare, come abbiamo visto, dal 68% e dal 95%, che invece conviene sapere.

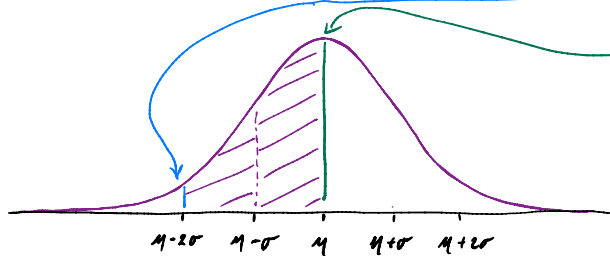
Tornando alle domande

5. IDOM fra la media e la media più una d. st. ?



Risposta: 34%

6. IDOM fra media meno due d. st. e media ?



Risposta: $13.5\% + 34\% = \underline{47.5\%}$

7. La mediana dei numeri $7, -7, 1, 2, -8, 15, 0$ vale?

Basta mettere i numeri in ordine e prendere quello centrale

$-8 \quad -7 \quad 0 \quad \underline{1} \quad 2 \quad 7 \quad 15$

8. A che percentile equivale il primo quartile?

Per definizione, il primo quartile di una variabile statistica è il valore di quella variabile al di sotto del quale cade un quarto della popolazione. Questo è il 25-esimo percentile, ossia il valore sotto cui cade il 25% della popolazione.

9. A che percentile equivale la mediana?

Ragionando come prima, metà della popolazione è sotto

la mediana, per definizione. Quindi, la mediana è il 50-esimo percentile.

Nelle pagine seguenti, si trovano le soluzioni dagli esercizi A, B, C, D. Per ciascuno di questi troverai, prima, una pagina con i risultati, e, a seguire, una illustrazione del procedimento.

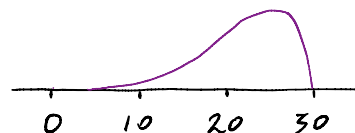
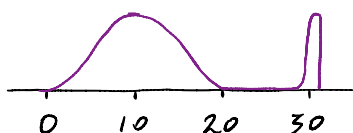
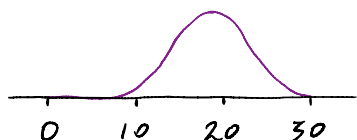
Es A.

Gli schizzi seguenti rappresentano gli istogrammi dei voti di latino ottenuti da tre classi di studenti:

A (104 studenti)

B (97 studenti)

C (100 studenti)



Completa le frasi seguenti:

La classe con la media più alta è C.

" mediana " è C.

" % di 30 " è B.

" d. standard " è B.

La media della classe B è maggiore della mediana.

" C è minore "

" A è circa uguale alla mediana

Nella classe B il 20% degli studenti ha preso ≥ 18

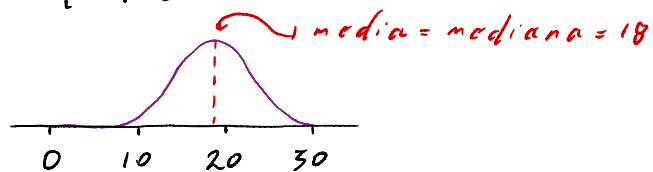
" A il 50% "

" C il 70% "

Gli studenti nel decile di voto più alto provengono principalmente dalla classe B. Quelli nel decile più basso da B. Quelli al di sopra della mediana da C. La mediana di tutti i voti vale circa 18.

Per giustificare queste risposte, iniziamo col tracciare una approssimazione delle medie e mediane delle tre classi.

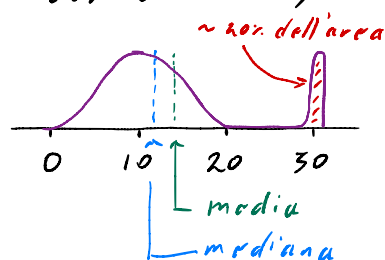
A (109 studenti)



Questa è la più facile: la distribuzione è simmetrica, per cui media e mediana coincidono

con l'ascissa dell'asse di simmetria della distribuzione.

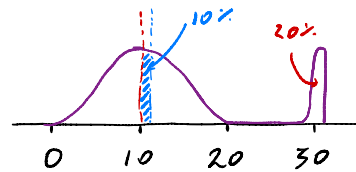
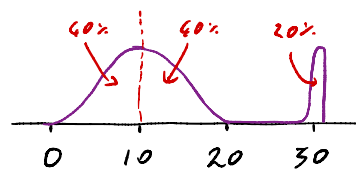
B (97 studenti)



In questa classe c'è un 20% di studenti (più o meno) che ha preso 30, il restante 80% ha preso, in media, circa 10. È importante, infatti ricordare che il numero di

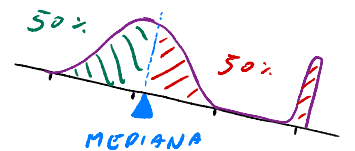
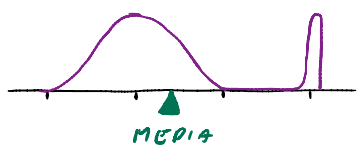
individui è l'area sotto l'istogramma. Possiamo stimare la media, quindi, calcolando $media \approx \frac{20 \cdot 30 + 80 \cdot 10}{100} = 14$

Per avere una stima della mediana, possiamo ragionare così: vediamo come si divide la popolazione rispetto alla soglia dei 10 punti. Questa soglia spacca la campana dell'80% in due metà simmetriche. Essa, quindi, divide la popolazione in un 40% sotto la soglia ed un 60% sopra. Di quanto dobbiamo spostare la soglia perché questa divisione diventi 50% e 50%? Dobbiamo lasciarci a sinistra un 10% in aggiunta al 40% che già c'è, e questo è uno spostamento piccolo: corrispon-

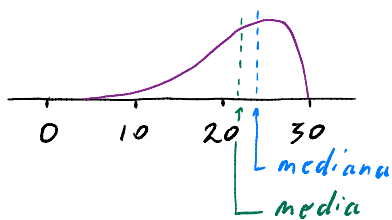


a metà dell'area della colonna sul 30. La mediana potrebbe essere 11, ed è certamente minore della media 14.

Possiamo, in altre parole, spiegare questo fenomeno pensando che la media è il punto dal quale si potrebbe sospendere l'istogramma mantenendolo bilanciato. Cosa avverrebbe se sospendessimo, invece, il medesimo istogramma sulla mediana? La mediana divide la massa dell'istogramma in metà, però il lato destro è distribuito a maggior distanza dal fulcro, quindi quest'altalena penderà verso destra. Di conseguenza, la media è a destra della mediana.



C (100 studenti)

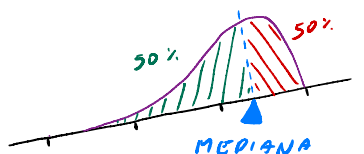


Nel caso di questo istogramma, è chiaro che media e mediana sono entrambe fra 20 e 25.

Quale di queste quantità è più grande?

Se ragioniamo come nel caso

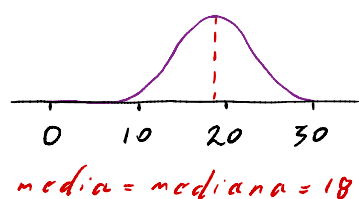
precedente, la coda della distribuzione a sinistra è più lunga di quella a destra, quindi la leva a sinistra ha un braccio maggiore, per cui la media è minore della mediana. Questo è un fatto generale:



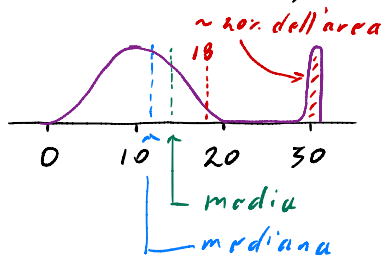
- $media < mediana$ indica una coda lunga a sinistra
- $mediana < media$ indica una coda lunga a destra
- $media \approx mediana$ suggerisce una distribuzione simmetrica.

Mettendo insieme le informazioni accumulate finora

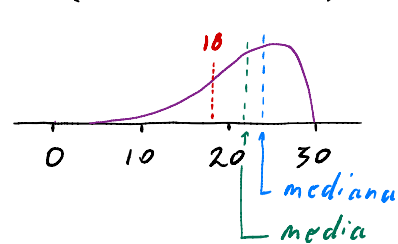
A (109 studenti)



B (97 studenti)



C (100 studenti)



possiamo facilmente rispondere a queste domande

La classe con la media più alta è C.

" mediana " è C.

" % di 30 " è B.

La media della classe B è maggiore della mediana.

" C è minore "

" A è circa uguale alla mediana

Nella classe B il 20% degli studenti ha preso ≥ 18

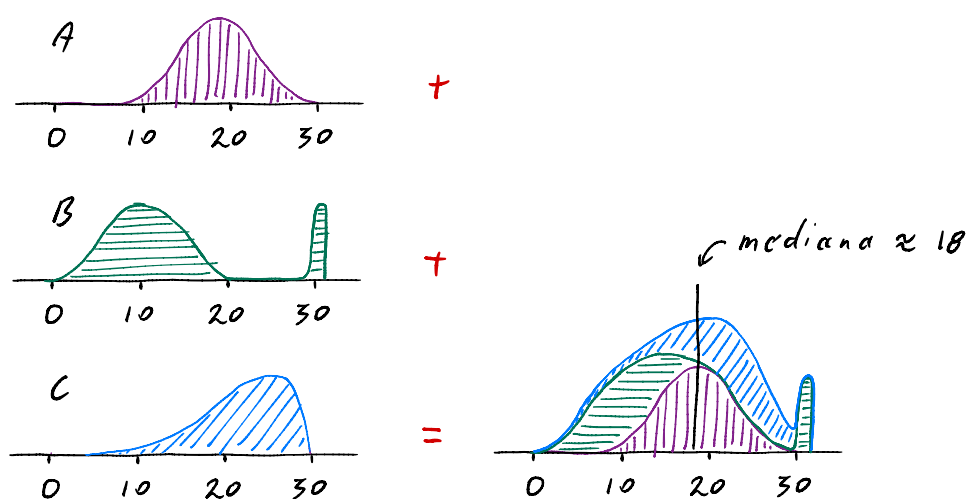
" A il 50% "

" C il 70% "

Per quanto riguarda la deviazione standard, dobbiamo ricordare che, grossolanamente, questa quantità misura la semi-ampiezza dell'intervallo attorno alla media in cui ricade la maggioranza degli individui. Per via del 20% di studenti sul 30, l'istogramma B è più disperso degli altri, quindi:
La classe con d. standard più alta è B.

Le ultime quattro domande "Gli studenti nel decile..." si

riferiscono all'istogramma del totale degli studenti. Qual è, ragionevolmente, la forma di questo istogramma? Per ogni voto dobbiamo sommare il numero di studenti provenienti da ciascuna delle tre classi. Siccome le tre classi hanno, grosso modo, il medesimo numero di studenti, i tre istogrammi contribuiscono circa in ugual misura.



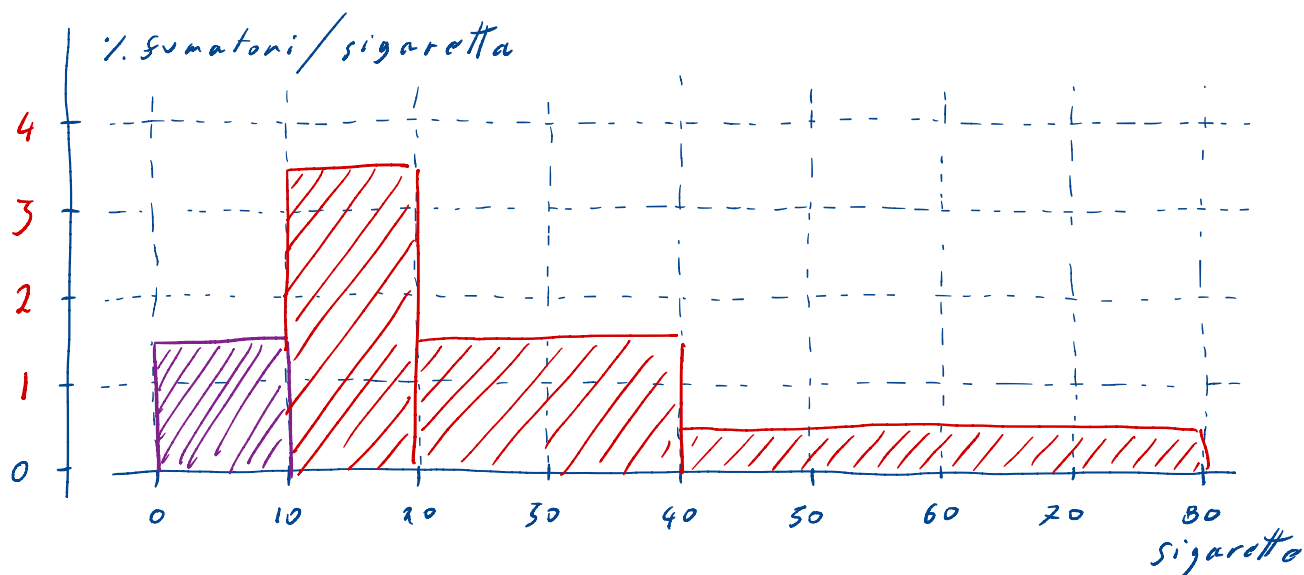
Gli studenti nel decile di voto più alto sono il 10% di studenti più a destra nell'istogramma, e questi provengono principalmente dalla classe **B**. Viceversa, il decile più basso è costituito dal 10% di studenti più a sinistra, e proviene principalmente dalla classe **B**. La mediana potrebbe valere circa **18** — i grafici qui non sono così precisi, e, per esempio, 20 sarebbe una risposta ugualmente accettabile. Al di sopra di questa mediana c'è circa metà della classe **B**, ma ben più di metà di **C**.

E5 B

In un campione di fumatori, il numero di sigarette consumate al giorno si distribuisce come segue.

# sigarette	# fumatori
$0 < - \leq 10$	225
$10 < - \leq 20$	525
$20 < - \leq 40$	450
$40 < - \leq 80$	300
$80 < -$	0

Completa il seguente istogramma



Qual è il numero mediano di sigarette fumate al giorno? ²⁰
Approssimativamente, che percentuale del totale dei fumatori consuma più di 30 sigarette al giorno? ^{35%}
~~Qual è il numero medio di sigarette al giorno? Dai una stima del numero medio di sigarette al giorno spiegando il tuo ragionamento. ~27~~

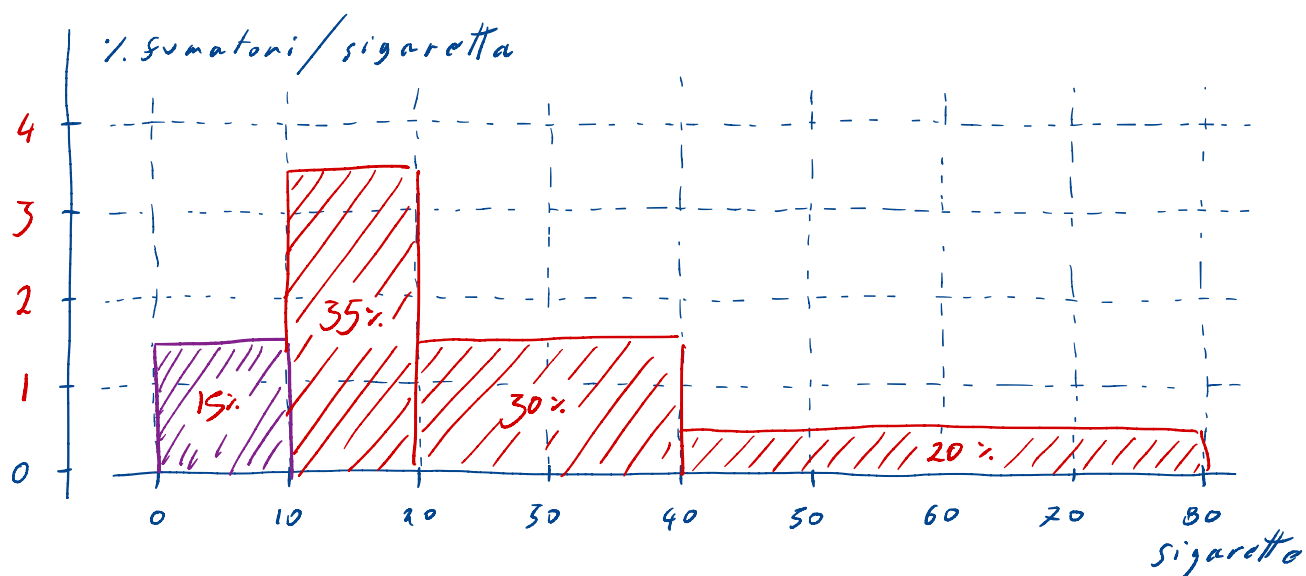
La prima richiesta è di completare l'istogramma. Vediamo che l'asse delle ascisse è graduato in sigarette e l'asse delle ordinate è in percentuale di fumatori a sigaretta, quindi le aree delle colonne dell'istogramma rappresentano percentuali di fumatori. Calcoliamo le percentuali corrispondenti ai dati forniti nella tabella.

# sigarette	# fumatori	% fumatori	
$0 < - \leq 10$	225	15 %	$= \frac{225 \cdot 100}{225 + 525 + 450 + 300}$
$10 < - \leq 20$	525	35 %	
$20 < - \leq 40$	450	30 %	
$40 < - \leq 80$	300	20 %	

L'area della colonna già tracciata è quindi il 15%. Il lato orizzontale di questa colonna è lungo 10 sigarette, di conseguenza la colonna è alta $\frac{15\%}{10 \text{ sigarette}} = 1.5 \text{ \% / sigaretta}$. Con questa informazione possiamo facilmente graduare l'asse delle ordinate. Allo stesso modo calcoliamo:

# sigarette	# fumatori	% fumatori	altezza colonna [%/sig.]
$0 < - \leq 10$	225	15 %	$\frac{15}{10-0} = 1.5$
$10 < - \leq 20$	525	35 %	$\frac{35}{20-10} = 3.5$
$20 < - \leq 40$	450	30 %	$\frac{30}{40-20} = 1.5$
$40 < - \leq 80$	300	20 %	$\frac{20}{80-40} = 0.5$

L'istogramma ha quindi l'aspetto seguente

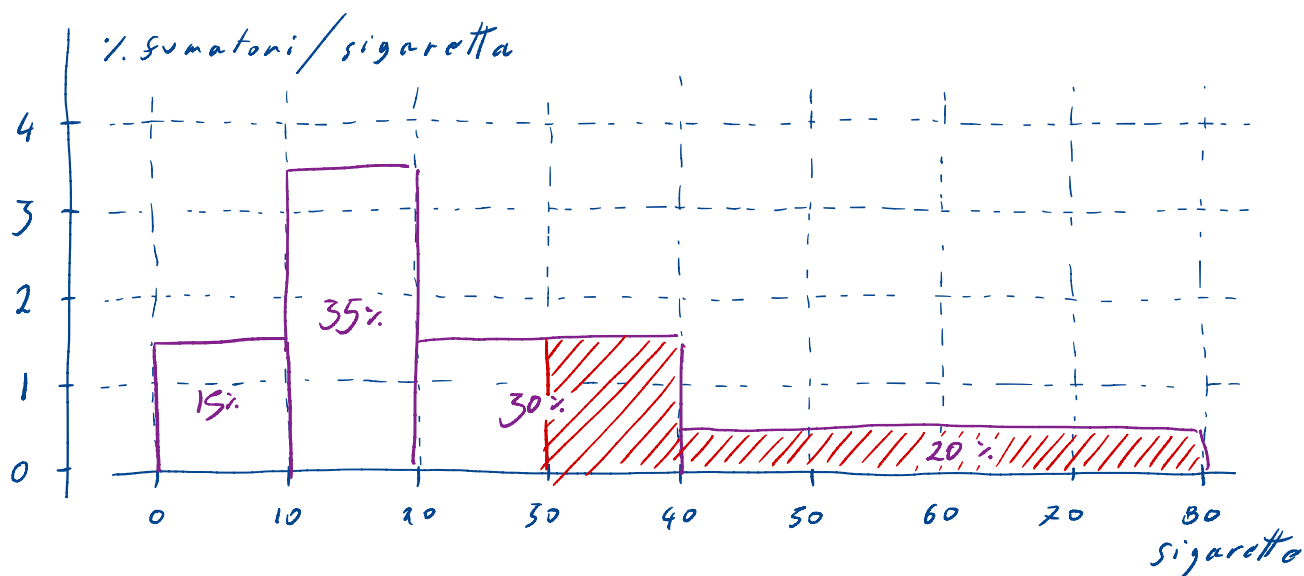


La mediana del numero di sigarette è la soglia che divide a metà l'area dell'istogramma. Si vede immediatamente che il 50% dei fumatori si trova sotto le **20** sigarette al giorno e il 50% sopra, questa è quindi la mediana.

Nota: il ragionamento precedente è leggermente impreciso, perché non considera che c'è un certo numero di fumatori che consumano esattamente venti sigarette al giorno. Questi, nella nostra tabella, sono contati nell'intervallo 10-20, e non nell'intervallo 20-40: abbiamo dunque una leggera asimmetria. Per gli scopi del nostro corso, possiamo ignorare questo dettaglio.

Passando alla seconda domanda, per determinare la percentuale dei fumatori che consuma più di 30 sigarette al giorno,

il meglio che possiamo fare è stimare l'area dell'istogramma che si trova a sinistra dell'ascissa *30 sigarette*



Questa area comprende metà della colonna del 30% e tutta quella del 20%, si tratta quindi del $\frac{1}{2} \cdot 30\% + 20\% = 35\%$ dei fumatori.

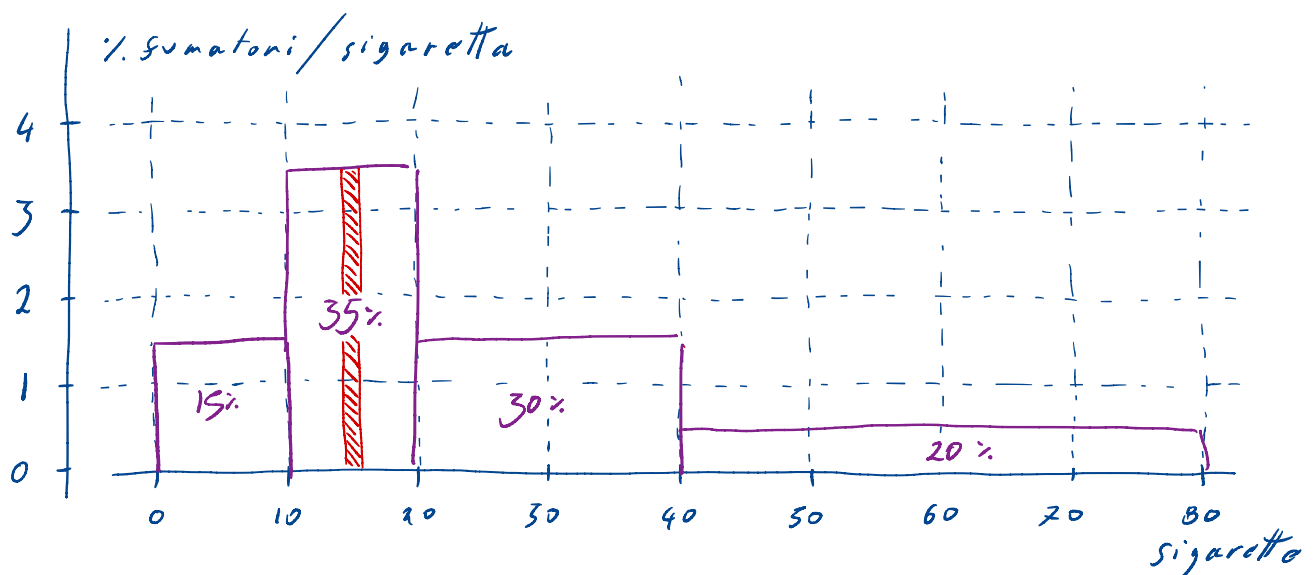
La domanda di quale percentuale dei fumatori consumi esattamente 15 sigarette al giorno è mal posta — colpa mia che me ne sono accorto troppo tardi... Tuttavia ci consente una discussione interessante. Il problema è che la variabile in esame — il numero di sigarette consumate al giorno — potrebbe essere *discreta* o *continua*, e il testo dell'esercizio non dice in che caso siamo. Una variabile si dice *discreta* se può assumere solo alcuni valori ben spaziali fra di loro, per esempio solo valori interi.

Se, viceversa, una variabile può assumere valori, anche frazionari, arbitrariamente vicini fra loro, allora si dice *continua*. Quando abbiamo una variabile discreta, immaginiamo, per esempio, il numero di Grattelli; ci sarà una certa percentuale della popolazione che ne ha 0, una altra percentuale che ne ha 1, e così via. Però, se prendiamo una variabile continua, come per esempio il peso, cosa significa chiedersi quante persone pesino 50 kg? Significa chiedersi quante siano fra 49.5 e 50.5? O fra 49.95 e 50.05? O esattamente a 50.000000 pesate al milligrammo? Si capisce che non ha senso chiedere quante persone pesino esattamente 50.0000... kg con precisione assoluta, e comunque la risposta sarebbe "nessuna". Nel caso di una variabile continua, si possono solo contare gli individui che ricadono in un certo intervallo, per esempio di peso compreso fra 45 kg e 55 kg.

La variabile del nostro esercizio, il numero di sigarette consumate al giorno, è discreta o continua? Se pensiamo che si tratti del numero di sigarette consumate *in un certo giorno* allora è discreta. Però il testo non fa riferimento ad un giorno specifico, per cui è probabile che il numero sia stato calcolato dividendo il numero di siga-

rette totali consumate durante un certo periodo di osservazione per la lunghezza del periodo, dando così luogo ad una variabile continua.

Come semplice esempio, facciamo, tuttavia, finta che la variabile sia discreta e rispondiamo alla domanda sotto questa ipotesi, anche se è poco realistica. La domanda è quale percentuale dei fumatori consumi esattamente 15 sigarette al giorno. In termini di istogramma, ci stiamo chiedendo quale sia l'area di una colonnina larga una sigaretta posta in corrispondenza dell'ascissa 15.



L'altezza di questa colonnina è $3.5\%/\text{sig.}$. l'area è pertanto $\text{area} = 3.5\%/\text{sig.} \cdot 1 \text{ sig.} = 3.5\%$.

Veniamo infine alla stima del numero medio di sigarette consumate al giorno. Il meglio che possiamo fare con i dati

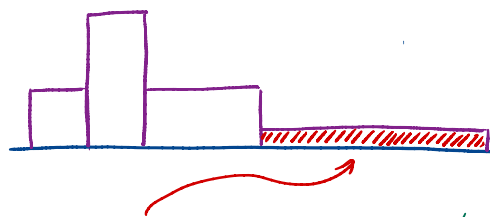
a nostra disposizione è supporre che i fumatori in ciascuna delle colonne dell'istogramma siano equidistribuiti. Per cui i fumatori nella colonna più a destra fumeranno in media 5 sigarette al giorno (la media fra 0 e 10). I fumatori nella colonna successiva fumeranno in media 15 sigarette al giorno (la media fra 10 e 20), etc. Qual è la media globale? Ragionando su un totale di *conto* di 100 fumatori, il numero di sigarette consumate al giorno da questi individui è

$$15 \cdot \frac{0+10}{2} + 35 \cdot \frac{20+10}{2} + 30 \cdot \frac{40+20}{2} + 20 \cdot \frac{80+40}{2} = 2700$$

il 15% fra 0 e 10 il 35% fra 10 e 20 il 30% fra 20 e 40 il 20% fra 40 e 80

Dal momento che queste sono le sigarette consumate al giorno da 100 fumatori, in media ciascun fumatore consuma $2700/100 = 27$ sigarette al giorno.

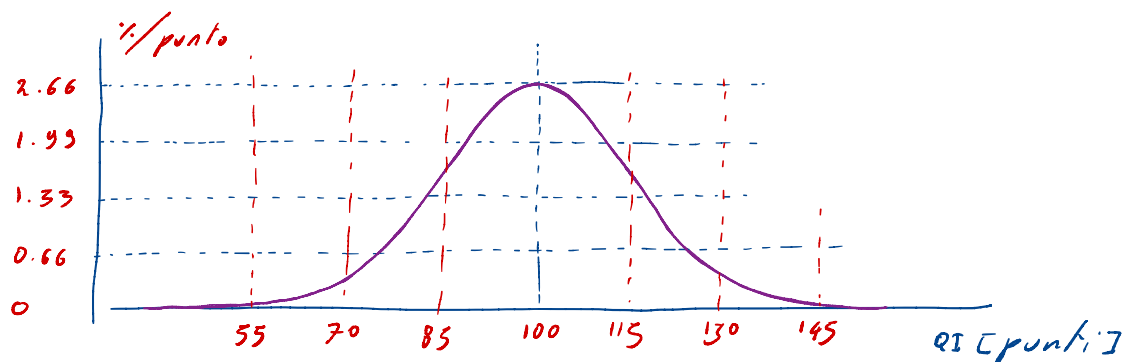
Nota: la media, 27, è venuta maggiore della mediana, che avremmo stabilito essere 20. Questo era atteso perché l'istogramma ha una *coda* verso destra.



E5 C

Il quoziente di intelligenza (QI) è costruito in modo tale che risulta distribuito in maniera gaussiana con media 100 e deviazione standard 15.

Riporta sul diagramma seguente la graduazione corretta, inclusa l'unità di misura sull'asse delle ordinate.



Completa le frasi seguenti:

Il 50 % delle persone ha un QI maggiore di 100.

Il 25 % " " 110.

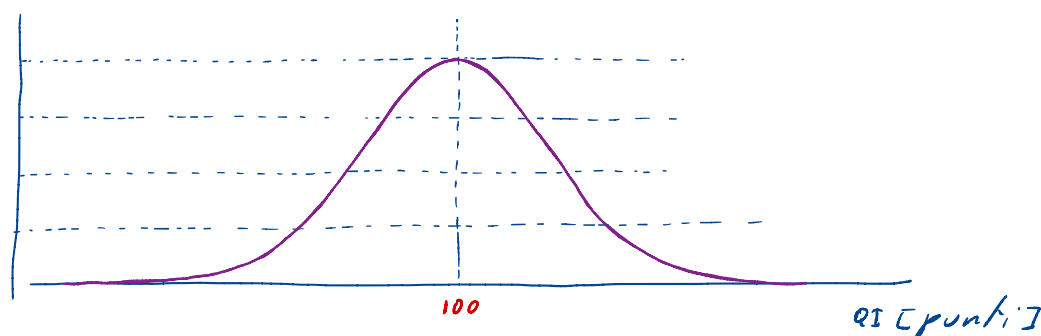
Il 9 % " " 120

Le persone con un QI minore di 95 sono tante quante quella con un QI maggiore di 105.

Solo una persona su 100 ha un QI maggiore di 135.

La differenza fra il 75-esimo ed il 25-esimo percentile (detta differenza interquartile) è di 20 punti, ed in questo intervallo di QI cade il 50 % delle persone.

La prima richiesta è di graduare l'istogramma. Per cominciare, possiamo segnare il valor medio, 100, sullo asse delle ascisse, in corrispondenza del picco della curva.



L'area sotto l'istogramma rappresenta una percentuale della popolazione. Siccome l'asse delle ascisse è misurato in punti, e il prodotto deve essere una percentuale, l'asse delle ordinate è misurato in **%/punto**. Che percentuale per punto indica l'altezza della curva al centro? Una tabella

z	Height	Area
0.00	39.89	0
0.05	39.84	3.99
0.10	39.69	7.97
0.15	39.45	11.92

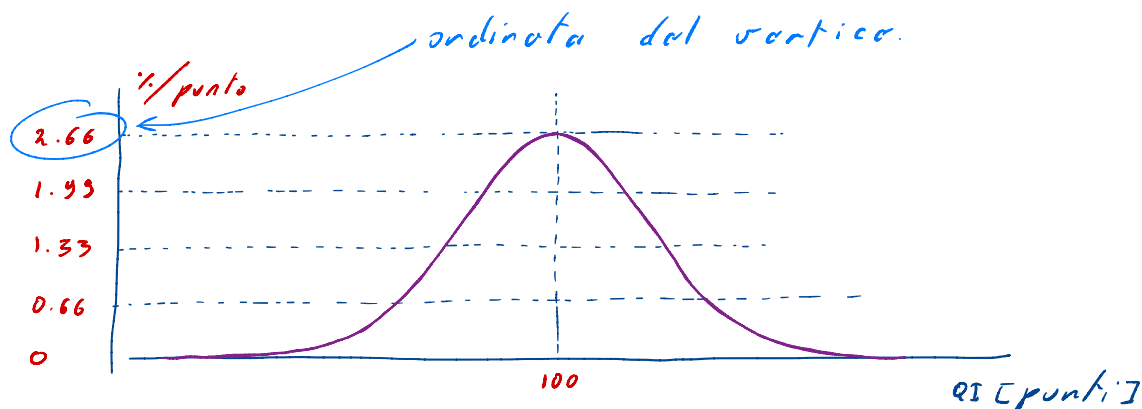
ci dica che, in unità standard, l'altezza della curva, nel picco, è **39.89 %/deviazione st.**

La deviazione standard della nostra curva è di 15 punti, otteniamo quindi

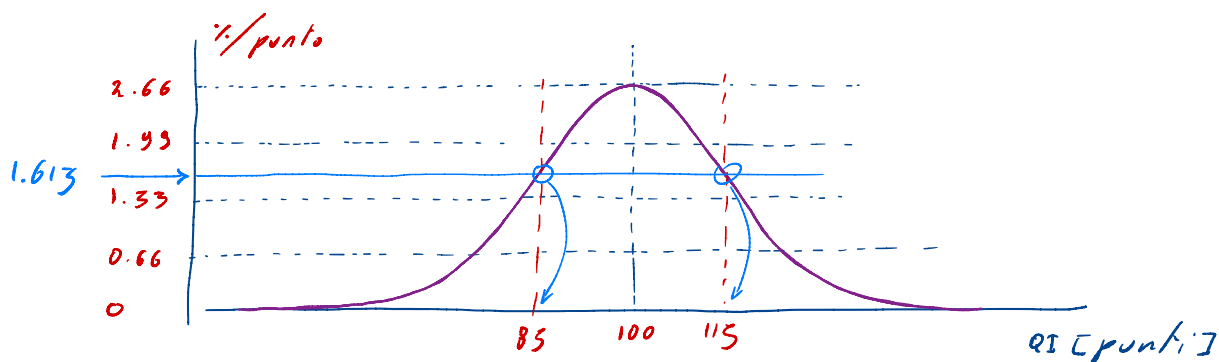
$$\text{altezza } (z=0) = 39.89 \frac{\%}{15 \cdot \text{punto}} = \frac{39.89}{15} \frac{\%}{\text{punto}} = 2.659 \frac{\%}{\text{punto}}$$

Il numero di cifre significative con cui indicare questo dato nel grafico è, in parte, questione di gusti. Almeno due servono, per avere un'indicazione utile, e più di

quattro non avrebbe senso perché il valore è calcolato usando una tabella che offre quattro sole cifre di precisione. Nel mio disegno scelgo di usarne tre. Stabilita l'ordinata del vertice della curva, è facile graduare l'asse delle ordinate proporzionalmente.



Infine, possiamo aggiungere al nostro istogramma alcune suddivisioni dell'asse delle ascisse. Il 115 è una deviazione standard a sinistra del 100, ma quanto è lunga una deviazione standard sul disegno? Facendo, nuovamente, riferimento alla tabella, vediamo che l'altezza della curva ad una deviazione standard dalla media ($z=1$) è $24.20 \text{ \%/d.st.}$ ossia $\frac{24.20}{15} \text{ \%/punto} = 1.613 \text{ \%/punto}$. Possiamo quindi determinare l'ampiezza di una deviazione standard tracciando la retta orizzontale di ordinata $1.613 \text{ \%/punto}$: i punti in cui questa retta interseca la curva si trovano una deviazione standard a destra e a sinistra della media.



Ora è immediato espandere la graduazione dell'asse della ascisse in modo proporzionale, e si ottiene la figura alla prima pagina della soluzione.

Possiamo alle domande successive.

Il — % delle persone ha un QI maggiore di 100.

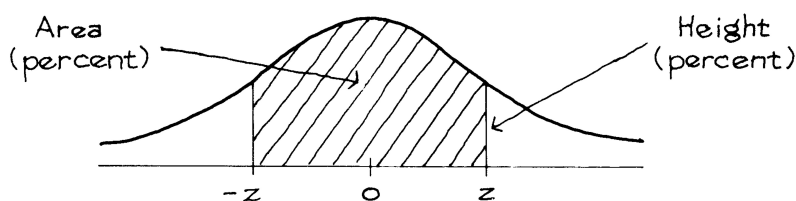
La risposta è 50, perché 100 è la mediana della distribuzione (che, per una distribuzione gaussiana, coincide con la media).

Il — % delle persone ha un QI maggiore di 110.

Per rispondere a questa domanda è necessario fare riferimento ai valori tabulati dell'area sotto la distribuzione normale. In unità standard, 110 punti corrispondono a

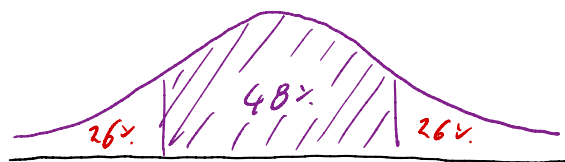
$$z = \frac{110 \text{ punti} - \text{media}}{\text{dev. standard}} = \frac{110 \text{ punti} - 100 \text{ punti}}{15 \text{ punti}} = 0.6667.$$

In prima approssimazione, possiamo arrotondare questo numero



z	Height	Area
0.50	35.21	38.29
0.55	34.29	41.77
0.60	33.32	45.15
<u>0.65</u>	32.30	<u>48.43</u>
0.70	31.23	51.61

al valore tabulato più vicino: 0.65. Dalla tabella, leggiamo che il 48.43% degli individui ha QI compreso fra 90 e 110 punti. Quelli con QI maggiore di 110 saranno quindi la metà della rimanenza:



$$\frac{100\% - 48.43\%}{2} \approx 26\%$$

Il risultato così ottenuto, sebbene essenzialmente corretto, si può migliorare ricavando un valore più preciso dell'area mediante la tecnica dell'interpolazione lineare.

z	Height	Area
0.50	35.21	38.29
0.55	34.29	41.77
0.60	33.32	45.15
$x_1 = 0.65$	32.30	48.43 = y_1
$x_2 = 0.70$	31.23	51.61 = y_2

Procedendo così otteniamo che l'area fra 90 e 110 punti è

$$\text{Area} = (0.6667 - x_1) \cdot \frac{y_2 - y_1}{x_2 - x_1} + y_1$$

$$= 49.49\%$$

Da cui si ricava la percentuale $\frac{100\% - 49.49\%}{2} = 25.26\%$.

Possiamo indicare questo risultato con la precisione del centesimo di 1%, che corrisponde alla precisione della tabella, ma è anche ragionevole arrotondare semplicemente all'intero più vicino: **25%**.

Il — % delle persone ha un QI maggiore di 120.

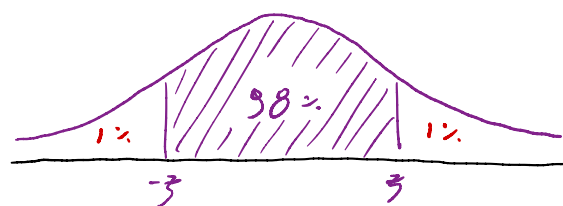
A questa domanda, si risponde applicando precisamente il medesimo procedimento della precedente. La risposta è **9.13%** o semplicemente **9%**.

Le persone con un QI minore di 95 sono tante quante quelle con un QI maggiore di —.

Grazie alla simmetria della gaussiana rispetto alla media, in questo caso di 100 punti, vediamo che il valore cercato è il simmetrico di 95 rispetto a 100, ossia **105**.

Solo una persona su 100 ha un QI maggiore di —.

Applicando a ritroso il procedimento delle domande precedenti, dobbiamo trovare z in modo tale che l'area a destra di questa soglia contenga l'1% della popolazione. Di conseguenza, l'area fra $-z$ e z conterrà il **98%**.



z	Height	Area
2.25	3.17	97.56
2.30	2.83	97.86
<u>2.35</u>	2.52	<u>98.12</u>
2.40	2.24	98.36
2.45	1.98	98.57

della popolazione. Grazie alla solita tabella, troviamo $z \approx 2.35$. Il valore cercato è pertanto

$$Q1 \approx \text{media} + z \cdot \text{dev. standard} = 100 + 2.35 \cdot 15 \approx 135$$

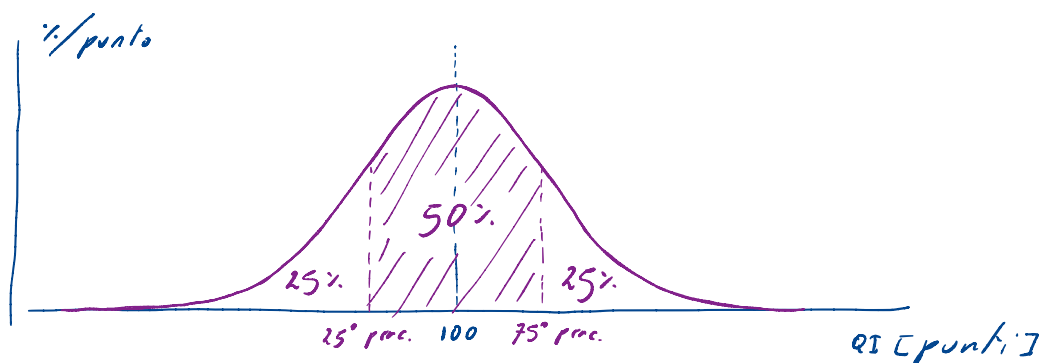
Nuovamente, possiamo migliorare il procedimento usando, nel calcolo di z , l'interpolazione lineare.

$$z = (98 - 97.86) \frac{2.35 - 2.30}{98.12 - 97.86} + 2.30 = 2.327$$

Da cui $Q1 = 134.9 \approx 135$.

La differenza fra il 75-esimo ed il 25-esimo percentile (detta differenza interquartile) è di — punti, ed in questo intervallo di $Q1$ cade il —% della persone.

Per definizione l' x -esimo percentile è la soglia sotto la quale si colloca lo $x\%$ della popolazione. La situazione è quindi come nella figura seguente



È quindi immediata la risposta alla seconda domanda: il 50% delle persone si trova fra il 25-esimo ed il 75-esimo percentile.

Per rispondere alla prima domanda, determiniamo, con la solita tabella, che il 50% dell'area sotto la gaussiana è compreso fra $-z$ e z con $z = 0.6747$. Ora

$$25^{\circ} \text{ percentile} = \text{media} - z \cdot \text{dev. standard} = 100 - z \cdot 15$$

$$75^{\circ} \text{ percentile} = \text{media} + z \cdot \text{dev. standard} = 100 + z \cdot 15$$

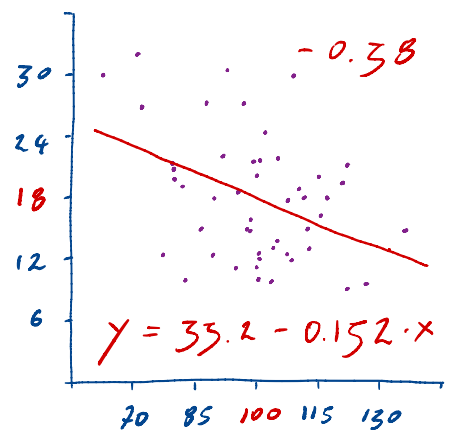
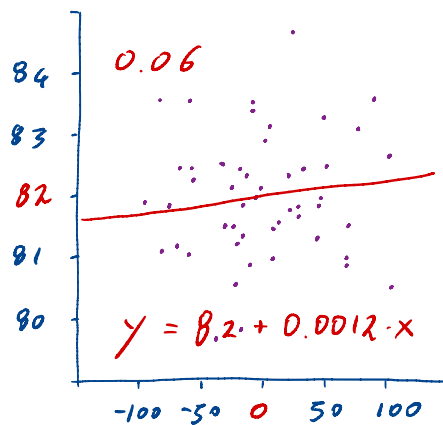
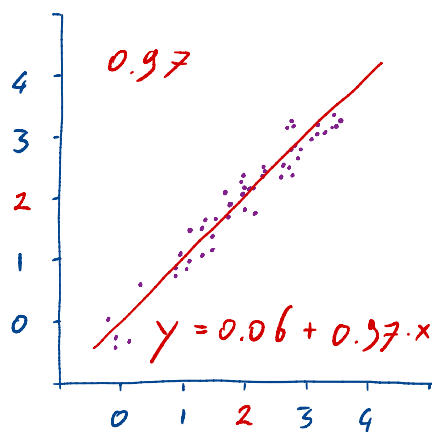
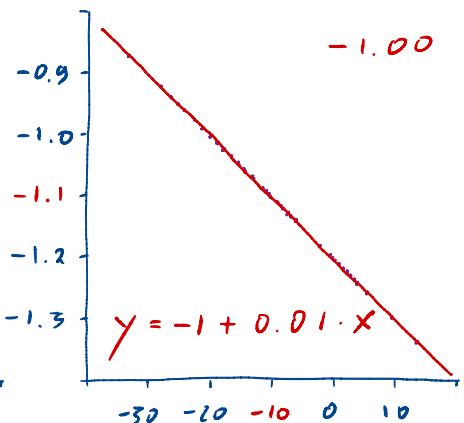
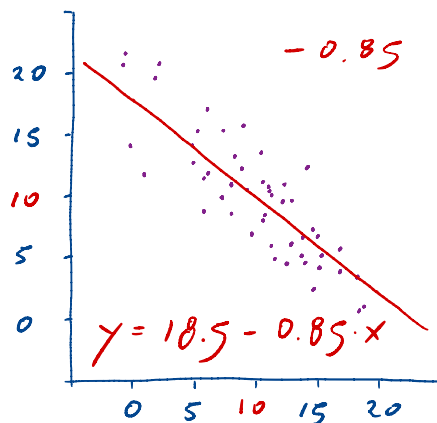
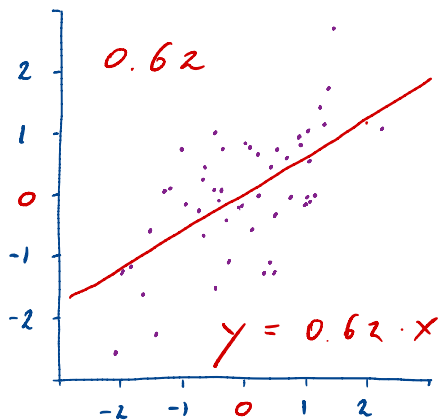
$$\begin{aligned} \text{diff. interquartile} &= 75^{\circ} \text{ percentile} - 25^{\circ} \text{ percentile} \\ &= z \cdot 30 = 20.24 \approx 20. \end{aligned}$$

E_s D

I coefficienti di correlazione corrispondenti ai seguenti diagrammi di dispersione sono

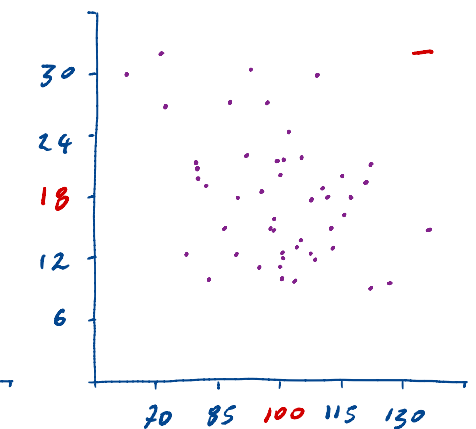
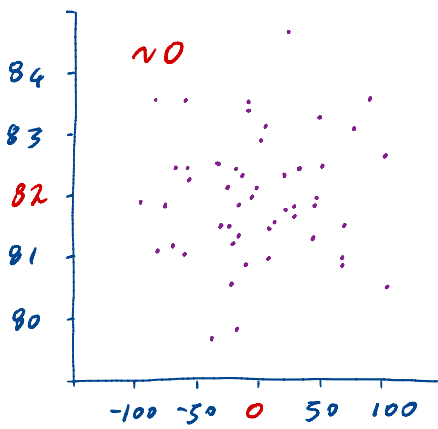
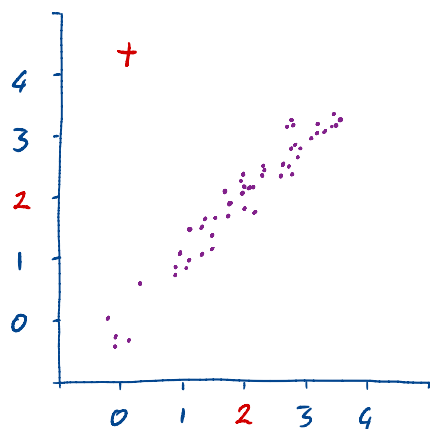
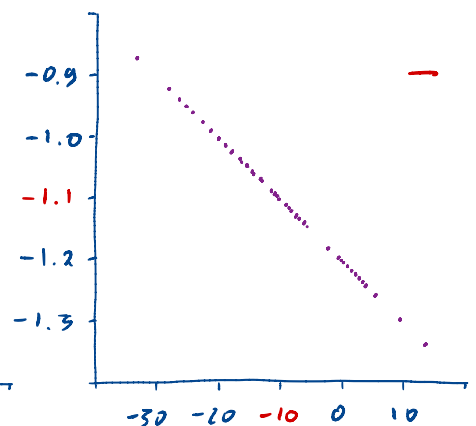
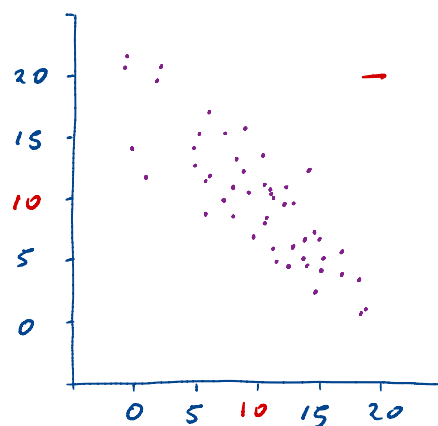
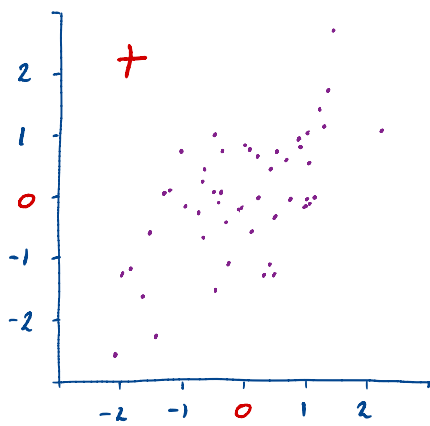
-0.85 -0.38 -1.00 0.06 0.97 0.62

Associa ad ogni diagramma il suo coefficiente

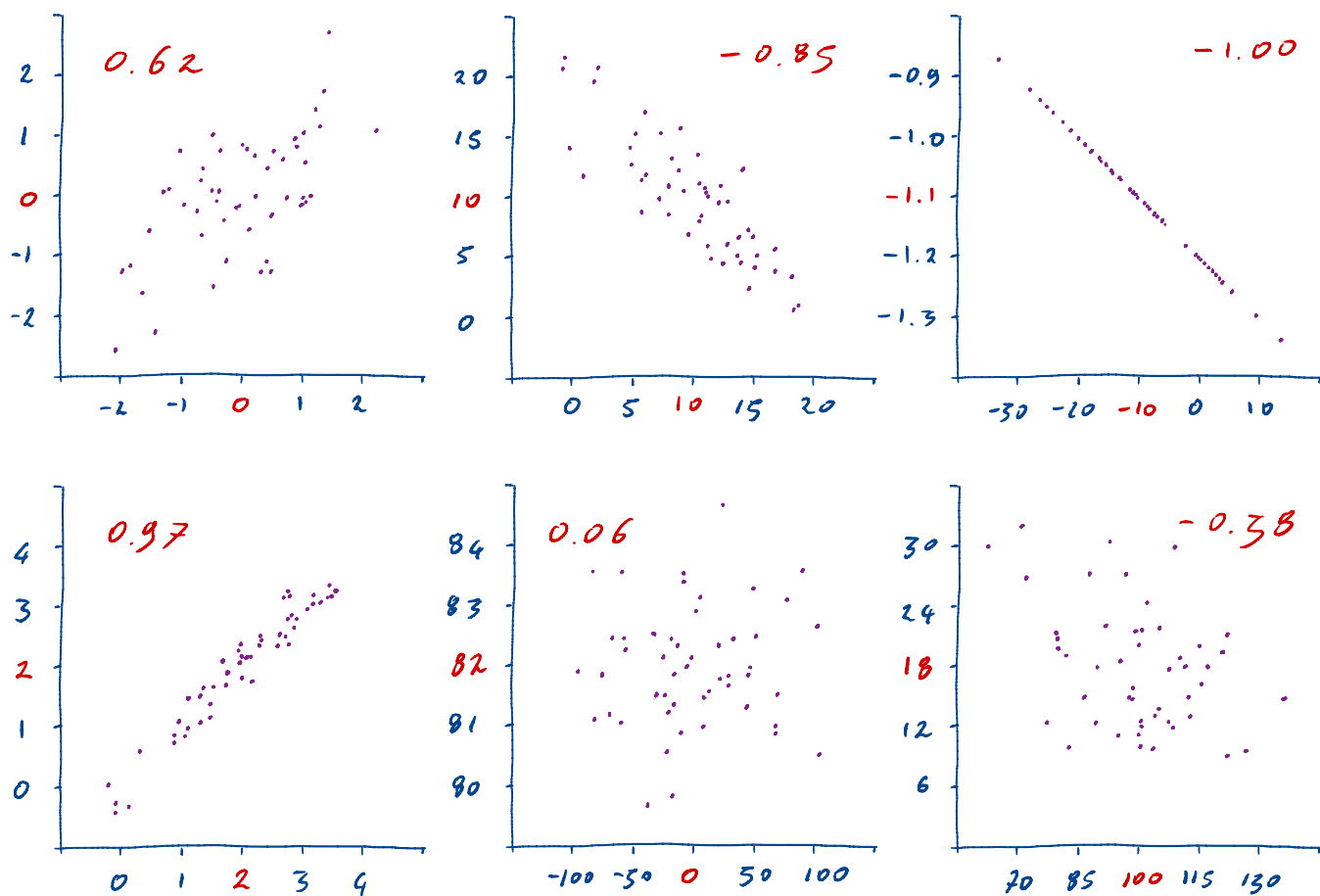


I valori indicati sugli assi sono spazati di una deviazione standard delle rispettive variabili, con la media posta al centro. Per ciascuno dei diagrammi, calcola l'equazione della retta di correlazione e traccia questa retta sul diagramma stesso.

Vediamo, inizialmente, che tre dei valori di correlazione dati sono negativi, e tre positivi, di cui uno praticamente zero. Una correlazione positiva indica che i punti esigono un andamento crescente, una correlazione negativa indica un andamento decrescente, la correlazione zero è associata all'assenza di un chiaro andamento. Abbiamo quindi questi segni:



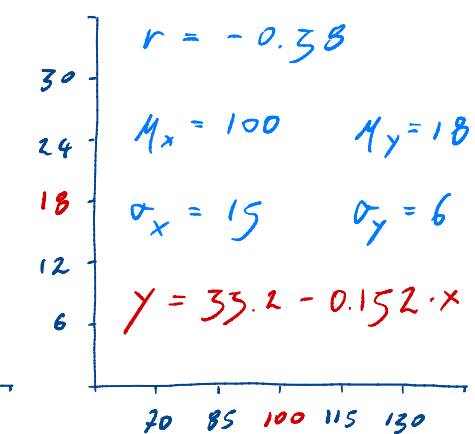
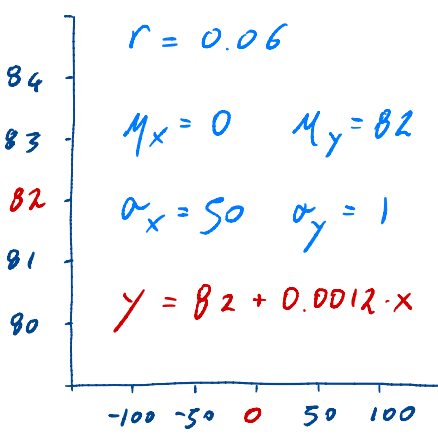
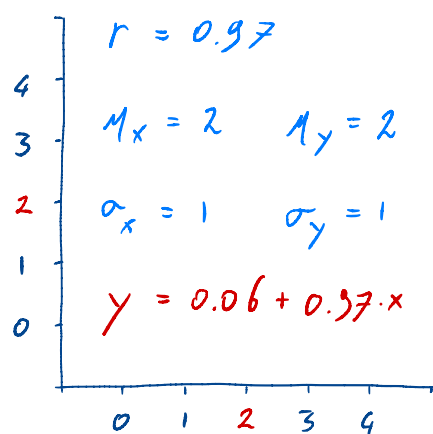
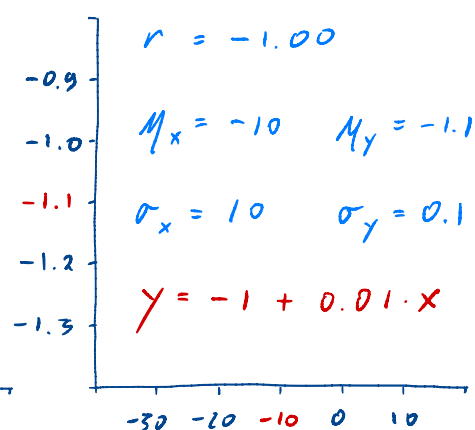
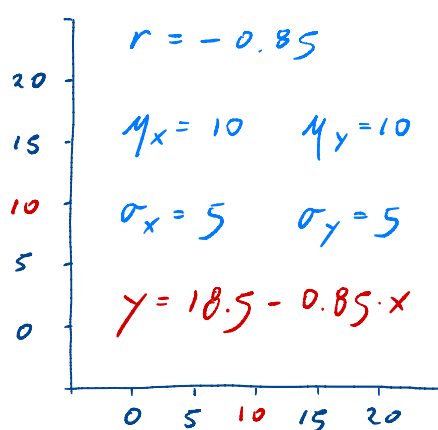
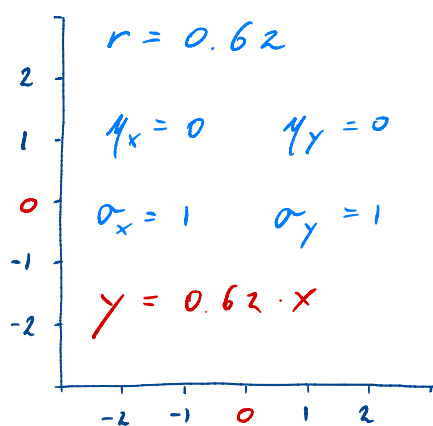
Il valore assoluto del coefficiente di correlazione indica è tanto più vicino a 1 quanto più l'alligse dei punti è schiacciato sul suo asse maggiore. Usando questa informazione arriviamo alla seguente assegnazione dei coefficienti:



Passiamo adesso alla tracciatura delle rette di regressione. L'equazione della retta di regressione è

$$y = \mu_y + (x - \mu_x) \cdot r \cdot \frac{\sigma_y}{\sigma_x}$$

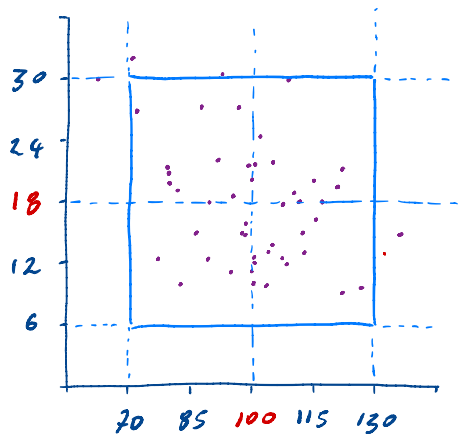
dove x e y sono le variabili in ascissa e in ordinata rispettivamente, μ_x e μ_y le loro medie, σ_x e σ_y le loro deviazioni standard, e r è il coefficiente di correlazione. Medie e deviazioni standard si ricavano dai diagrammi dati come indicato nel testo. Otteniamo così le seguenti equazioni.



Adesso è possibile tracciare la retta a partire dalle equazioni, ma è un po' scomodo. C'è un trucco che permette di farlo più agevolmente basandosi sui fatti seguenti:

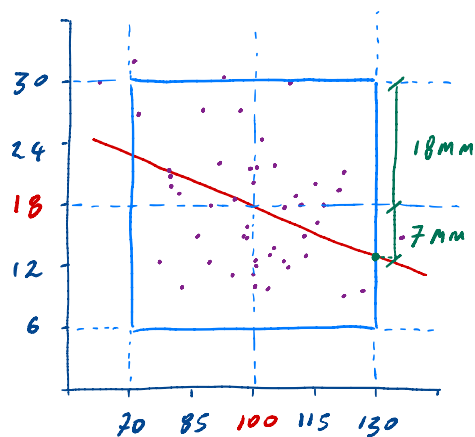
- la retta di regressione passa per il punto della media (μ_x, μ_y)
- spostandosi verso destra da questo punto di un certo numero a di deviazioni standard (σ_x) la retta sale di $a \cdot r$ deviazioni standard (σ_y).

Nel caso del sesto diagramma, per fare un esempio, possiamo facilmente immaginare un rettangolo centrato



nel punto delle medie che si estende a destra e a sinistra di questo punto per due deviazioni standard della x , e sopra e sotto questo punto per due deviazioni standard della y .

Ora, misurando queste due deviazioni standard della y , sulla mia figura, ottengo che corrispondono a circa 18 mm. Di conseguenza, la retta di regressione intercetta il lato destro del rettangolo 18 mm $\cdot r = 18 \text{ mm} \cdot (-0.38)$



$\approx -7 \text{ mm}$ al di sopra della quota della media — cioè 7 mm sotto la media. Posso così tracciare la retta.

Le rette tracciate alla prima pagina sono state ottenute ripetendo l'esercizio su ciascuno dei diagrammi.

Suggerimento Studia i diagrammi di questo esercizio, cercando di formarti un'idea dell'aspetto che ha la distribuzione dei punti in funzione del coefficiente di correlazione.