COME ESEGUIRE UNO STUDIO DI ASSOCIAZIONE GENETICA IN R

https://link.springer.com/book/10.1007/978-3-319-14475-7

In seguito ad un precedente esperimento di mappatura genetica, si ipotizza che il peso corporeo in bovini da carne sia controllato geneticamente da un determinato tratto di un cromosoma. In pratica, il peso di un bovino, a parità di tutti gli altri fattori "ambientali" (età, alimentazione, stato sanitario ecc.) dipenderebbe dalla particolare sequenza nucleotidica a 5 *loci* che si trovano in quella stessa regione del genoma.
Un *locus* che controlla il valore fenotipico di un carattere quantitativo (come il peso corporeo, la quantità di latte, il diametro della fibra di lana ecc.) prende il nome di *Quantitative Trait Locus* (QTL).
Sebbene la localizzazione precisa di tali *loci* non sia nota, sono stati identificati 5 marcatori microsatelliti ognuno dei quali è associato ad uno dei *loci* che influenzano il peso corporeo.
Ciò significa che ogni locus microsatellite è adiacente ad un determinato QTL; per es.:

_____QTL1_m1_____QTL2_m2_____QTL3_m3___…

Siccome ***loci* associati vengono ereditati insieme come singolo aplotipo** dal momento che la **ricombinazione** (*crossing over*[1]) non li separa, ogni locus microsatellite si comporta come una sorta di “marcatore genetico” (*genetic marker*) del QTL a cui è associato.
Allora, se, per es., fosse possibile stabilire una correlazione (legame statistico) fra un certo allele[2] al marcatore m1 e il peso corporeo, si potrebbero selezionare come riproduttori gli individui portatori di tale allele per aumentare il peso corporeo dei vitelloni nelle generazioni successive.

[1]Vedi <**RICOMBINAZIONE_03.10.2021.pdf**>.
[2]Individuato attraverso il sequenziamento.

MICROSATELLITI

Se si sottopongono a migrazione elettroforetica[3] i prodotti della digestione con endonucleasi di restrizione di un campione di DNA nucleare di dimensioni cospicue, il profilo che si ottiene è rappresentato da una strisciata: le diverse taglie dei frammenti non possono essere separate perché costituiscono un continuum, mentre le uniche bande visibili corrispondono alle cosiddette "**sequenze satellite**". Queste ultime furono così definite poiché, in esperimenti di ultracentrifugazione attraverso gradienti continui di CsCl (cloruro di cesio), si riscontrava la presenza di un picco "satellite" del picco principale, che fu ad esse attribuito. Le sequenze satellite, altamente ripetitive, sono organizzate in raggruppamenti (*clusters*) a livello centromerico e telomerico, ossia nelle regioni eterocromatiche dei cromosomi. Esistono però anche sequenze satellite disperse, in maniera apparentemente casuale, lungo tutto il genoma: secondo la lunghezza, si distinguono sequenze LINE (Long INterspersed Elements) e sequenze SINE (Short INterspersed Elements). In analogia con le sequenze satellite vere e proprie, vengono identificate altre due categorie eterogenee di sequenze ripetute, i minisatelliti ed i microsatelliti. I minisatelliti si trovano, negli organismi eucarioti, nelle regioni eucromatiche dei cromosomi e sono costituiti dalla ripetizione testa-coda (a coprire un'estensione di 0,5-3 Kb) di un modulo formato da un numero di paia di basi compreso fra 9 e 100 e in media 15 bp. I microsatelliti, invece, sono costituiti da brevi moduli[4] (2-6 bp) ripetuti testa-coda un numero moderato di volte. Il genoma umano contiene almeno 30.000 loci di microsatelliti, localizzati nelle regioni eucromatiche. Minisatelliti e microsatelliti costituiscono un'importante classe di markers denominati, nel loro insieme, "*Variable Number of Tandem Repeats*" (numero variabile di gruppi di sequenze ripetute). I VNTR sono ipervariabili: in ogni specie animale, per ciascun locus VNTR, infatti, esistono numerosissimi alleli, la frequenza dei quali varia nelle diverse popolazioni. Ognuno di essi è identificato dal numero delle sequenze ripetute (moduli). Ad ogni locus VNTR gli alleli vengono ereditati in modo mendeliano, cioè ogni individuo possiede soltanto due alleli ad ogni locus VNTR (uno di origine materna e l'altro di origine paterna), ma il numero di alleli presenti nella popolazione è così elevato che la maggior parte degli individui è eterozigote. Dal

[3] Vedi <**1_GEN_LAB_20.01.2023.pdf**>

[4] Per questo motivo sono talvolta chiamati *Short Tandem Repeats* (STRs) o *Simple Sequence Repeats* (SSRs).

punto di vista medico-legale (ma anche negli studi sui rapporti di parentela fra individui di popolazioni animali naturali o allevate), ciò consente sia di escludere che uno specifico soggetto sia figlio di due determinati individui (presunti genitori) sia, con un'alta probabilità, di assegnare la paternità di quel soggetto ad un'altra coppia di genitori. Inoltre, se i due alleli per locus vengono moltiplicati per le decine di migliaia di loci VNTR presenti nel genoma di un mammifero, si ottiene un numero pressoché infinito di combinazioni possibili[5]. Ogni singolo individuo di una determinata specie animale potrà quindi essere inequivocabilmente identificato sulla base dell'assetto di alleli, ad un certo numero di loci VNTR, di cui è portatore: tale metodo di tipizzazione genetica prende il nome di DNA-*fingerprintig* ("impronta digitale" genetica). L'elevato numero di loci microsatellitari esistenti e di alleli per locus possono essere sfruttati, oltre che per applicazioni diagnostiche, medico-legali e forensiche, anche per lo studio della variabilità genetica delle popolazioni animali o vegetali naturali. Dal punto di vista operativo, la tecnica della PCR può essere vantaggiosamente applicata anche per l'amplificazione dei loci VNTR, poiché questi ultimi sono in genere preceduti e seguiti da sequenze invarianti, sulle quali dovranno essere costruiti i primers. Una certa conoscenza delle sequenze che fiancheggiano i loci VNTR sarà dunque necessaria, e ciò rappresenta un'importante differenza rispetto alla tecnica RAPD, per l'applicazione della quale non è richiesta alcuna informazione preliminare sulle sequenze da amplificare o su quelle che le fiancheggiano. Ciononostante, la disponibilità di un grande numero di primers quasi universali riduce notevolmente il lavoro preliminare e l'impegno che deve essere profuso quando si inizia a studiare una specie nuova. Riassumendo, l'amplificazione mediante PCR dei loci VNTR offre a considerare i seguenti aspetti:

1. L'amplificazione di loci VNTR produce profili elettroforetici individuali;
2. I diversi alleli per i singoli loci sono direttamente osservabili su di un gel di agaroso o di poliacrilammide;
3. Se le condizioni di reazione sono compatibili, numerosi loci possono essere co-amplificati nella stessa miscela di reazione (multiplex PCR).

[5] Per n differenti alleli ci sono n(n+1)/2 possibili genotipi.

In **R**:

```
> getwd()
[1] "/home/piero"
> setwd('/home/piero/GONDRO/capitolo2')
> getwd()
[1] "/home/piero/GONDRO/capitolo2"
> figli = read.table('dati_figli.txt', header=TRUE, sep='\
t', skip=7)
> str(figli)
'data.frame':    400 obs. of  14 variables:
 $ id  : chr  "id1" "id2" "id3" "id4" ...
 $ toro: chr  "toro1" "toro1" "toro1" "toro1" ...
 $ sex : chr  "F" "M" "M" "M" ...
 $ peso: chr  "520.28" "594.38" "602.64" "607.94" ...
 $ m11 : chr  "M5" "M1" "M2" "M2" ...
 $ m12 : chr  "M6" "M4" "M3" "M3" ...
 $ m21 : chr  "M2" "M3" "M2" "M2" ...
 $ m22 : chr  "M5" "M1" "M6" "M1" ...
 $ m31 : chr  "M3" "M4" "M3" "M4" ...
 $ m32 : chr  "M6" "M5" "M1" "M6" ...
 $ m41 : chr  "M2" "M4" "M4" "M2" ...
 $ m42 : chr  "M5" "M3" "M3" "M3" ...
 $ m51 : chr  "M4" "M2" "M4" "M4" ...
 $ m52 : chr  "M4" "M2" "M3" "M5" ...

> figli$peso = as.numeric(as.character(figli$peso))
# convertiamo il tipo di dato della colonna "peso" da carattere a
numerico

> str(figli)
'data.frame':    400 obs. of  14 variables:
 $ id  : chr  "id1" "id2" "id3" "id4" ...
 $ toro: chr  "toro1" "toro1" "toro1" "toro1" ...
 $ sex : chr  "F" "M" "M" "M" ...
 $ peso: num  520 594 603 608 509 ...
 $ m11 : chr  "M5" "M1" "M2" "M2" ...
 $ m12 : chr  "M6" "M4" "M3" "M3" ...
 $ m21 : chr  "M2" "M3" "M2" "M2" ...
 $ m22 : chr  "M5" "M1" "M6" "M1" ...
 $ m31 : chr  "M3" "M4" "M3" "M4" ...
 $ m32 : chr  "M6" "M5" "M1" "M6" ...
 $ m41 : chr  "M2" "M4" "M4" "M2" ...
 $ m42 : chr  "M5" "M3" "M3" "M3" ...
 $ m51 : chr  "M4" "M2" "M4" "M4" ...
 $ m52 : chr  "M4" "M2" "M3" "M5" ...
```

```
> summary(figli$peso)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  501.2   544.1   586.4   584.6   623.4   851.5       1

> tori = read.table('dati_tori.txt', header=TRUE, sep='\t',
skip=7)
> str(tori)
'data.frame':    10 obs. of  12 variables:
 $ id  : chr  "toro1" "toro2" "toro3" "toro4" ...
 $ peso: num  592 646 680 620 634 ...
 $ m11 : chr  "M2" "M3" "M2" "M2" ...
 $ m12 : chr  "M1" "M2" "M4" "M1" ...
 $ m21 : chr  "M3" "M3" "M2" "M1" ...
 $ m22 : chr  "M2" "M2" "M4" "M2" ...
 $ m31 : chr  "M3" "M2" "M3" "M4" ...
 $ m32 : chr  "M4" "M3" "M2" "M3" ...
 $ m41 : chr  "M4" "M2" "M3" "M2" ...
 $ m42 : chr  "M2" "M4" "M4" "M1" ...
 $ m51 : chr  "M4" "M3" "M1" "M4" ...
 $ m52 : chr  "M2" "M1" "M4" "M3" ...
> summary(tori$peso)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  590.9   623.4   636.9   632.4   644.9   680.4
> plot(figli$peso,main='grafico XY del peso nei figli',
xlab='figlio', ylab='peso', col='blue')
>
```

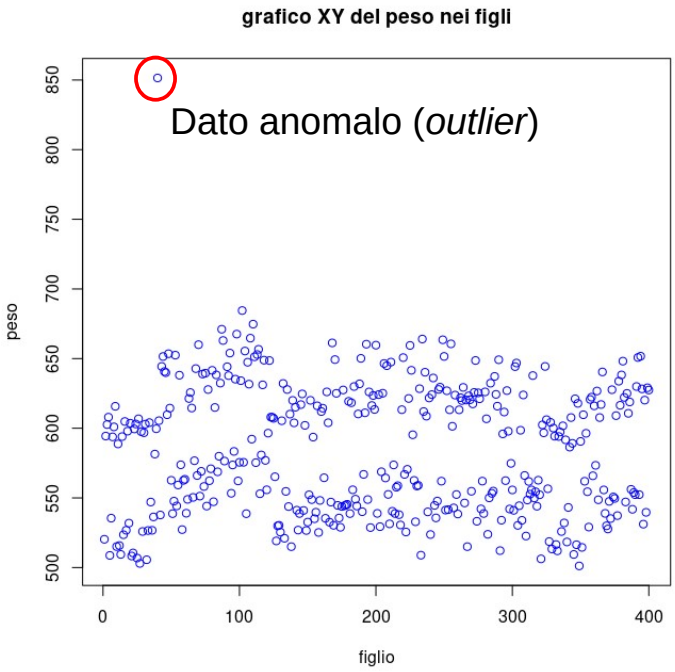


FIGURA 1. Grafico a dispersione del peso dei figli: **perchè i valori della variabile dipendente si concentrano in due gruppi, scarsamente sovrapposti**?

il dato anomalo è, verosimilmente, un errore, quindi lo eliminiamo:

```
> figli=figli[-which(figli$peso>800),]
> summary(figli$peso)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  501.2   544.0   584.9   583.9   623.0   684.5       1
> index=grep('m', names(figli))
```

la funzione **grep** cerca la lettera *m* nei nomi che formano l'intestazione (*header*) dell'oggetto figli e salva nell'oggetto **index** le posizioni in cui la trova

```
> figli
```

					locus m1		locus m2		locus m3		locus m4		locus m5	
	1	2	3	4	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
INTEST.	id	toro	sex	peso	**m11**	**m12**	**m21**	**m22**	**m31**	**m32**	**m41**	**m42**	**m51**	**m52**
1	**id1**	toro1	F	**520.28**	**M5**	**M6**	M2	M5	M3	M6	M2	M5	M4	M4
2	id2	toro1	M	594.38	M1	M4	M3	M1	M4	M5	M4	M3	M2	M2
3	id3	toro1	M	602.64	M2	M3	M2	M6	M3	M1	M4	M3	M4	M3
4	id4	toro1	M	607.94	M2	M3	M2	M1	M4	M6	M2	M3	M4	M5
5	id5	toro1	F	508.71	M1	M3	M2	M4	M4	M6	M4	M5	M2	M3

...

M5 M6 = genotipo della femmina id1 (figlia del toro1)
M5 = allele 1 al *locus* m1
M6 = allele 2 al *locus* m1
assumendo che il motivo ripetuto sia **aattag**
il **genotipo** di id1 sarà
aattagaattag**aattag**aattag**aattag** **aattag**aattag**aattag**aattag**aattag**aattag

ALLELE 1 ALLELE 2

DOMANDA: qual è il fenotipo di id1?
RISPOSTA: _____

```
> index
 [1]  5  6  7  8  9 10 11 12 13 14
> missing=numeric()
```

creiamo una nuova variabile chiamata **missing** di tipo numerico (conterrà dei numeri e non dei caratteri)

```
> length(index)
[1] 10
```

questo è il numero delle colonne di **figli** che contengono gli alleli

```
> for (i in 1:length(index))
+ missing=c(missing,which(figli[,index[i]]=='-'))
```

per *i* compreso fra 1 e 10, trova in quale posizione index contiene un trattino e memorizzala nell'oggetto **missing**
i è un **contatore**, il cui valore viene incrementato di 1 ad ogni reiterazione del ciclo for

```
> print(missing)
 [1] 69 69 69 69 69 69 69 69 69 69
```

c'è un trattino in tutte le colonne degli alleli nella riga N° 69

```
> missingU=unique(missing)
```

la funzione unique restituisce l'unico valore presente in **missing**

```
> print(missingU)
[1] 69
```

andiamo ad ispezionare la 69-esima riga, ossia la riga N° 70
la riga 1 è la 0-esima
la riga 2 è la prima
la riga 3 è la seconda
la riga **70** è la **69-esima**

```
> print(figli[missingU,])
     id  toro sex   peso m11 m12 m21 m22 m31 m32 m41 m42 m51 m52
70 id70 toro2   M 659.98   -   -   -   -   -   -   -   -   -   -
> boxplot(figli$peso~figli$toro,
+ col=1:length(levels(figli$toro)),
+ main='Boxplot del peso dei figli di ogni toro')
```

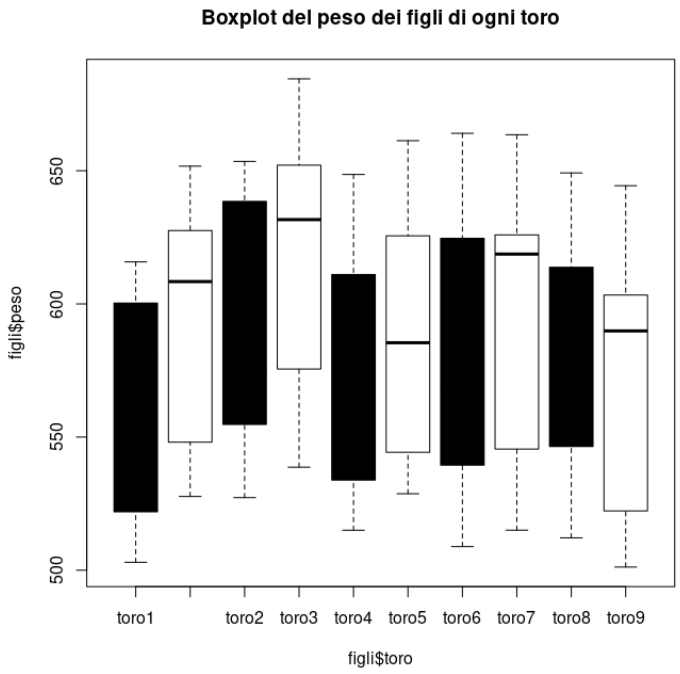


FIGURA 2. Boxplot del peso dei figli di ogni toro.

In statistica descrittiva, un **boxplot** (diagramma a scatola e baffi) è un metodo per rappresentare graficamente la variabilità di dati numerici attraverso i loro **quartili**.

☑ APPROFONDIMENTO

STATISTICA

Branca della matematica che si occupa della raccolta, analisi, interpretazione e presentazione di dati numerici, per es. di **dati sperimentali** ossia dei risultati di un esperimento scientifico.

Probabilità. Significato corrente: facilità con cui un evento futuro pensiamo possa verificarsi. Il concetto di probabilità è necessario se si lavora con sistemi biologici che generano osservazioni (dati) le quali non possono essere predette con certezza. Per es., la ***capacità di prevedere le caratteristiche della prole*** ottenuta da una determinata coppia di riproduttori assume una notevole rilevanza dal punto di vista del miglioramento dei caratteri delle specie animali allevate. Per i caratteri quantitativi non è possibile fare una previsione esatta ma soltanto una **previsione statistica**: se, per es., una bovina lattifera ad alta produzione (BLAP) ha prodotto mediamente 10 ton di latte nella prime 3 lattazioni, non è dato di conoscere quale sarà la produzione esatta di una sua figlia. Per citare un altro esempio, non è possibile sapere se, dopo aver inseminato una scrofa, un determinato ovocita verrà fecondato oppure no, se un embrione si impianterà o meno e se nasceranno 10 o 11 suinetti. Tutto ciò vale anche per molti fenomeni fisici: per es., non sapremo mai dopo esattamente quanto tempo un ponte inizierà ad incrinarsi.

FIGURA 3. Il ponte sul Po di Cardè prima che venisse ristrutturato.

Cos'hanno in comune tutti questi fenomeni apparentemente così ≠? La loro **natura stocastica** o **casuale** non significa che essi si verifichino con totale irregolarità; infatti, se si esegue una lunga serie di prove (confronti fra la quantità di latte prodotta da bovine e dalle loro madri, diagnosi di gravidanza, registrazione del n° di suinetti nati vivi e svezzati ecc.) si scopre che il n° di volte (**frequenza relativa**) con cui un evento si verifica sul totale delle prove effettuate è

relativamente stabile. Tale frequenza relativa calcolata ripetendo un esperimento molte volte (per es., n° di volte in cui esce testa lanciando una moneta 1000 volte) ci dà un'idea intuitivamente attendibile della facilità con cui riteniamo possa verificarsi un evento casuale se dobbiamo fare una previsione del risultato di un'ulteriore prova. Può essere utile descrivere un esperimento tenendo conto della probabilità con cui si ritiene che ciascuno dei suoi possibili risultati possa verificarsi, ossia costruire un **modello probabilistico** dell'esperimento stesso.
Esempio di fenomeno casuale: prendere 6 uova dal frigorifero ed immergerle delicatamente in acqua in pre-ebollizione; il guscio di alcune di esse si romperà lasciando fuoriuscire un po' di albume. La P di rottura del guscio può essere stimata calcolando il rapporto fra il n° di uova che si rompono ed il totale: se, per es., si rompono 4 uova su 6, P = 2/3 $\cong$ 0.66. Possiamo, così, concludere che P = 66%? No. Perché? In un'altra replica dello stesso esperimento, si potrebbero rompere 5 uova su 6, 3 su 6 oppure anche tutte o, viceversa, nessuna. Quali conclusioni possiamo trarre dopo aver ripetuto l'esperimento più volte?

1. La P di rottura non ha un unico valore, ma può assumere tutti i valori compresi nell'intervallo [0, 1/6, 1/3, 1/2, 2/3, 5/6, 1], ossia [0, 0.17, 0.33, 0.5, 0.66, 0.83, 1]. Possiamo, allora, introdurre la **VARIABILE CASUALE *X = n° di uova che si rompe***, che può assumere uno $\forall$ (qualsiasi) dei predetti valori.
2. Ripetendo l'esperimento molte volte (per es. 1000), possiamo contare quante volte, sul n° totale di prove fatte, si rompe un uovo su 6, quante volte se ne rompono 2, 3 e così via e costruire una **distribuzione di frequenza relativa**; quest'ultima può essere rappresentata graficamente da un istogramma, disegnato riportando sull'asse orizzontale i possibili valori della variabile casuale X e costruendo, sopra ogni valore, un rettangolo di h $\propto$ al n° di volte F(X) in cui si è osservato un determinato valore di X (FIGURA 4).

Il risultato delle 1000 prove effettuate è riportato nella TABELLA 1.

RISULTATO (**X**)	**NUMERO DI OSSERVAZIONI = F**(**X**)
0	3
1/6	45
1/3	85
1/2	246
2/3	394
5/6	173
1	54

TABELLA 1. Numero di volte in cui una determinata variabile casuale assume uno specifico valore.

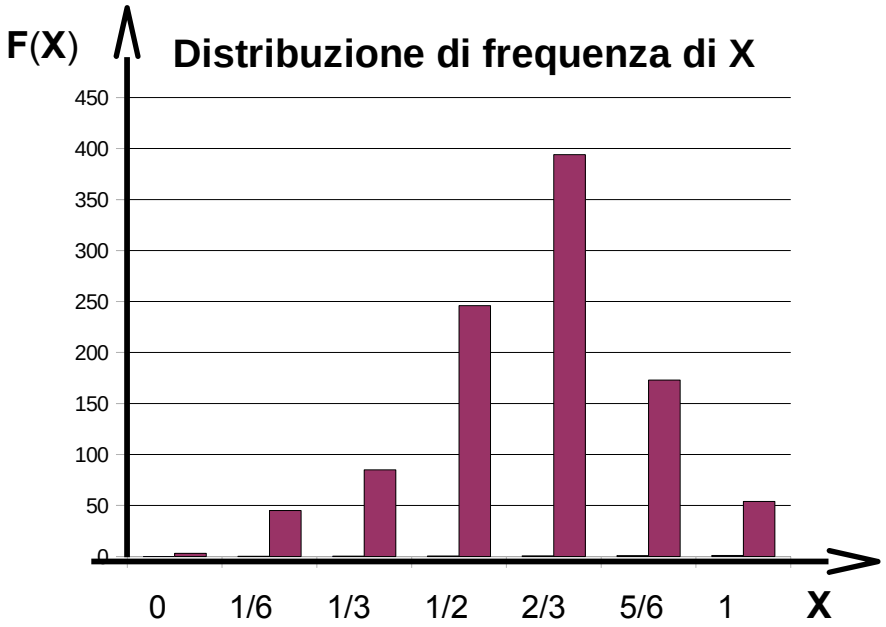


FIGURA 4. Esempio di distribuzione di probabilità di una variabile casuale.

Facciamo un altro esempio. Un giocatore d'azzardo che voglia sapere se un dado è truccato oppure no dovrebbe lanciarlo un n° ∞ di volte. Se il dado fosse perfettamente bilanciato, 1/6 delle «osservazioni» in questa popolazione sarebbe rappresentata da facce n° 1, 1/6 da facce n° 2, 1/6 da facce n° 3 e così via (FIGURA 5).

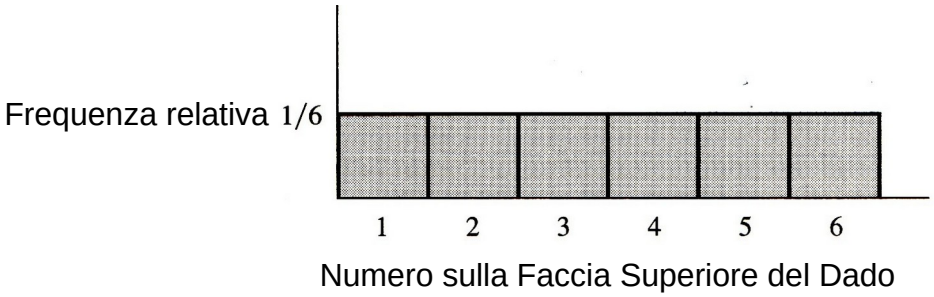


FIGURA 5. Distribuzione di frequenza dei possibili risultati del lancio di un dado, ripetuto un numero illimitato di volte.

Applicando il metodo scientifico, lo scommettitore propone l'ipotesi che il dado sia bilanciato e per falsificare tale ipotesi raccoglie osservazioni che la contraddicano. Un campione di 10 lanci viene estratto dalla popolazione, lanciando il dado 10 volte. Da tutti i 10 lanci si ottiene il n° 1. Lo scommettitore, deluso, conclude che la sua ipotesi non è in accordo con l'osservazione e che, quindi, il dado non è bilanciato. L'ipotesi viene confutata non perché sia impossibile ottenere 10 volte 1

in 10 lanci di un dado bilanciato ma perché ciò è estremamente improbabile. Sebbene lo scommettitore potrebbe non sapere come si calcola la P di 10 uno in 10 lanci, egli intuitivamente ritiene che tale risultato sarebbe molto improbabile se il dado fosse bilanciato.

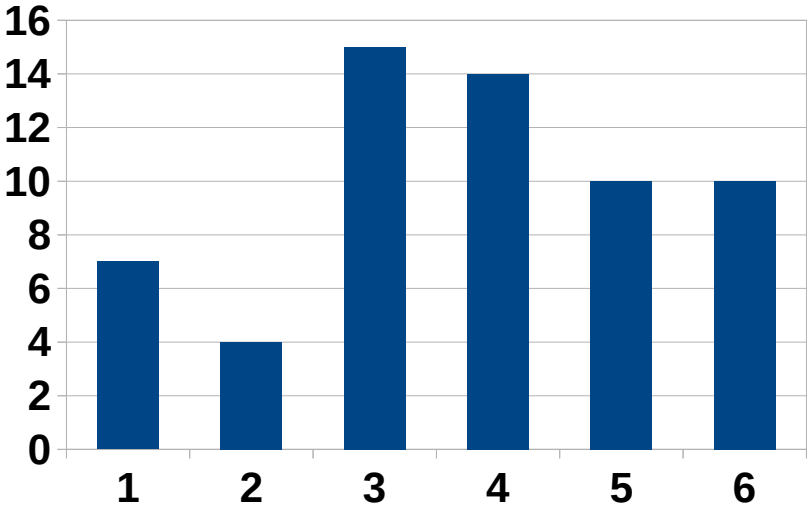


FIGURA 6. Se lanciassi un dato 60 volte potrei ottenere un risultato di questo tipo.

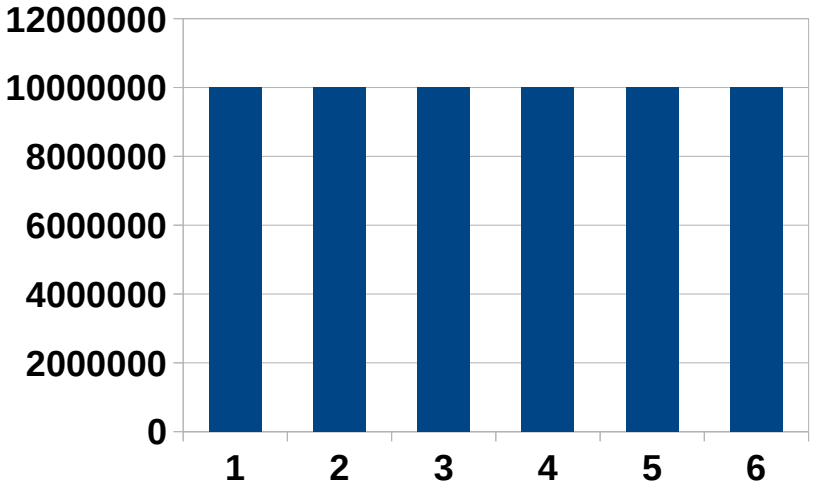


FIGURA 7. Ma se lo lanciassi 60 milioni di volte il risultato sarebbe ben diverso!

Esperimento. Processo che genera un'osservazione (per es., replicando in lab un determinato fenomeno naturale come il lancio di un dado o di una moneta).
Evento. Ognuno dei possibili risultati di un esperimento (contrassegnati da lettere maiuscole).
Evento elementare. Evento che non può essere decomposto e che corrisponde ad un «punto campione» indivisibile.
Spazio campione associato ad un esperimento. Insieme di tutti i possibili risultati di un esperimento (punti campione). È contrassegnato dal simbolo Ω (FIGURA 8).

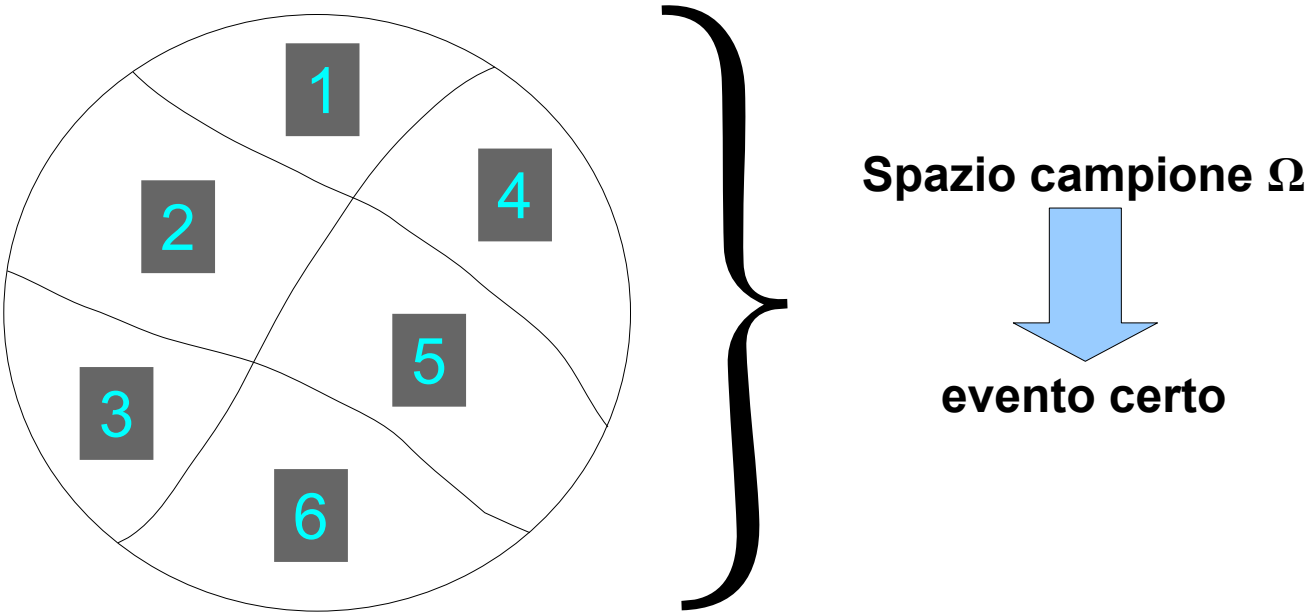


FIGURA 8. Spazio campione associato al lancio di un dado.

Si supponga che **S** sia uno spazio campione associato ad un esperimento. Ad ogni evento A in S (A è un sottoinsieme di S) è possibile assegnare un numero, il cui valore è ≥ 0 e ≤ 1, che esprime la probabilità che A si verifichi, P(A).
Se A_1, A_2, A_3,... formano una sequenza di eventi mutuamente esclusivi (incompatibili fra loro), allora

$$\mathbf{A_i \cap A_j = 0} \text{ se } \boldsymbol{i \neq j}$$

e

$$\mathbf{P(A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_n)} = \sum_{i=1}^{n} P(\mathbf{A_i})$$

Per es., nel caso in cui lanciando il dado si vince se esce 2 o se esce 3, la P di vincere sarà >, essendo data, poiché si tratta di eventi incompatibili, dalla somma della P(2) e della P(3):

$$\mathbf{P(A_2 \cup A_3)} = \sum_{i=2}^{3} P(\mathbf{A_i}) = \mathbf{1/6 + 1/6 = 1/3}$$

Infatti

$$\text{P di vincere} = \frac{\text{n° di casi in cui si vince}}{} = \frac{2}{} = 1/3$$

n° di casi possibili 6

TUTTI GLI EVENTI NON ELEMENTARI SI POSSONO SEMPRE CONSIDERARE COME UNIONE (∪) O INTERSEZIONE (∩) DI EVENTI ELEMENTARI.

Nel caso del lancio del dado, se $1 \leq j \leq 6$, allora

$$\mathbf{P(A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_6) = \sum_{i=1}^{6} P(A_i) = 1}$$

Ricapitolando:

- Spazio campione (Ω) = ∑ di 6 eventi elementari equiprobabili ed incompatibili fra loro.
- La P associata allo spazio campione è uguale all'unità.

$$\mathbf{P_\Omega = 1}$$

Eventi incompatibili. Si escludono a vicenda, poiché il verificarsi di uno dei 2 impedisce il contemporaneo verificarsi dell'altro (per es., se ad un lancio del dado esce la faccia n° 1 non può uscire contemporaneamente la faccia n° 2).

1. **Eventi equiprobabili**.

Se l'evento A e l'evento B sono incompatibili, l'evento A ∩ B (A intersezione B) è impossibile, poiché il verificarsi di A impedisce il contemporaneo verificarsi di B, e viceversa (A ∩ B = 0).
L'evento C = A ∪ B (A unione B) si verifica se si realizza l'evento A oppure l'evento B. La P_C (P dell'evento C) è la somma delle probabilità degli eventi elementari A e B, ma SOLO SE QUESTI SONO INCOMPATIBILI.

$$P_C = P_A + P_B = \frac{n_A + n_B}{n_\Omega}$$

n_A = numero di casi in cui si verifica l'evento A;
n_B = numero di casi in cui si verifica l'evento B;
n_Ω = numero di casi possibili.

Se il n° di casi favorevoli ad A (in cui si verifica l'evento elementare A) viene rapportato al n° delle prove effettuate, si ottiene la frequenza F_A, ossia la probabilità del verificarsi di A nell'esperimento effettuato

$$n_A$$

$$P_A = \frac{}{n} = F_A$$

n = numero di prove effettuate.

Se non si conoscono le regolarità del fenomeno studiato, l'unica cosa che si può fare è contare quante volte si verifica un determinato evento e calcolare il rapporto fra il n° di successi ottenuti ed il n° di prove effettuate.

2. **Eventi non equiprobabili**.

$P_A = 3/4 \Rightarrow$ nasce un capretto nero;
$P_B = 1/4 \Rightarrow$ nasce un capretto pezzato.
Esercizio: calcolare la P di nascita di 2 capretti neri e un capretto pezzato (nella stessa figliata).

- Sono eventi incompatibili (ogni capretto è nero oppure pezzato) ed indipendenti (il colore e la pezzatura del mantello di un capretto non influenzano le caratteristiche del mantello di un altro capretto).
- Ogni evento elementare ha P = 1/2 di verificarsi.
- I possibili esiti di un parto trigemellare sono 8 (n° delle **disposizioni con ripetizione** di n = 2 oggetti di classe k = 3, ossia n° delle combinazioni possibili di 2 oggetti (nero e pezzato) prendendone 3 alla volta:

$$D_{n,k} = n^k \Rightarrow D_{2,3} = 2^3 = 8$$

L'insieme di queste 8 combinazioni costituisce lo spazio degli eventi Ω, avente probabilità 1 ⇒ la probabilità di ogni tipo di figliata è 1/8.

- Ciascun tipo di figliata è dato dal contemporaneo verificarsi di 3 eventi elementari, indipendenti fra loro. Ogni evento elementare è rappresentato dalla nascita di un capretto nero o pezzato; la P che ciascun capretto sia nero o pezzato è del 50% (1/2). La P del contemporaneo verificarsi di 2 o più eventi elementari è data dall'intersezione degli eventi stessi:

P(1 capretto nero e 2 pezzati) = P(1 capretto nero ∩ 1 capretto pezzato ∩ 1 capretto pezzato)

NEL CASO DI EVENTI INDIPENDENTI, LA P DELLA LORO INTERSEZIONE É DATA DAL PRODOTTO DELLE PROBABILITÀ DEI SINGOLI EVENTI:

$$P(\text{1 capretto nero} \cap \text{1 capretto pezzato} \cap \text{1 capretto pezzato}) =$$
$$P(\text{1 capretto nero}) \cap P(\text{1 capretto pezzato}) \cap P(\text{1 capretto pezzato}) = (1/2)^3 = 1/8.$$

Eventi non incompatibili.

La P che si verifichi uno tra i 2 eventi:

1. almeno 2 capretti pezzati su 3;
2. tutti i capretti sono =

è data dalla somma delle P dei 2 eventi.

$$P_{(\text{ALMENO 2 PEZZATI})} = P_{mnm} + P_{mmn} + P_{nmm} + P_{mmm} = 1/8 + 1/8 + 1/8 + 1/8 = 4/8 = 1/2$$

$$P_{(\text{TUTTI UGUALI})} = P_{nnn} + P_{mmm} = 1/8 + 1/8 = 2/8 = 1/4$$

$$P_{(\text{ALMENO 2 PEZZATI O TUTTI UGUALI})} = 4/8 + 2/8 - 1/8 = 5/8$$

dove m stà per “macchiato” ed n per “nero”.

Infatti, P_{mmm} è stato contato 2 volte. Gli eventi “tutti uguali” e “almeno 2 macchiati” non sono incompatibili (nel caso in cui i capretti macchiati siano 3), perciò possono verificarsi entrambi contemporaneamente. Hanno un'intersezione, data dall'evento mmm, che li realizza entrambi; quest'intersezione dev'essere sottratta (FIGURA 65).

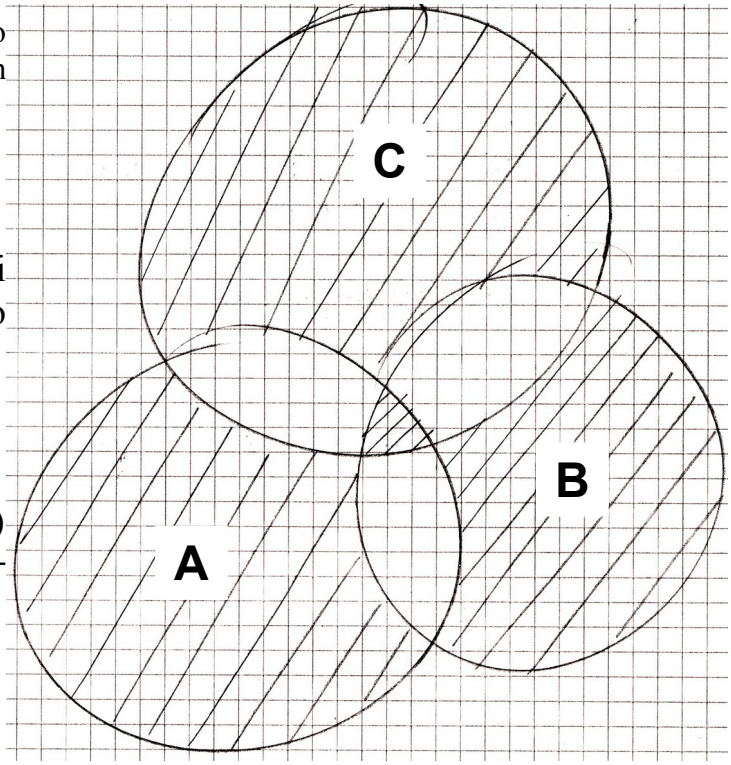
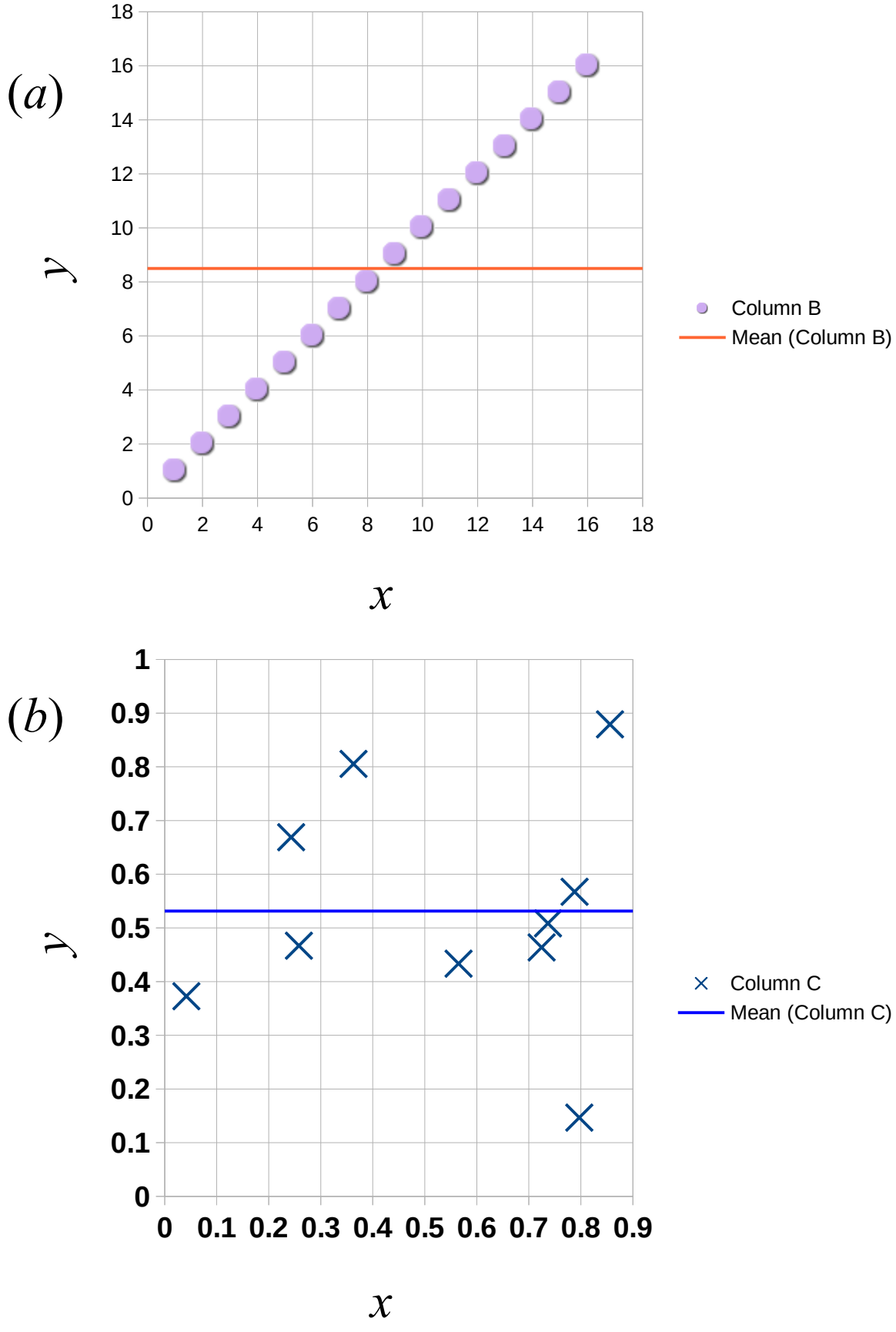


FIGURA 65. Spazio campione di tre eventi non incompatibili.

Probabilità che si verifichi l'evento A, o l'evento B o l'evento C:

$$\mathbf{P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + 2P(A \cap B \cap C)}$$



Le variabili X = *larghezza della groppa* di una vacca e Y = *facilità di parto* sono legate da una relazione di dipendenza: all'aumentare di X aumenta anche Y (FIGURA 67)(*a*): tutti i

FIGURA 66. Osservazioni dipendenti (*a*) e indipendenti (*b*) per la coppia di variabili *x*, *y*.

punti cadono lungo una linea retta. Al contrario, in (*b*) non si riscontra alcuna dipendenza fra le due variabili.
Si supponga di conoscere il valore medio di X (μ_1) e di Y (μ_2) e di indicarne le coordinate sul grafico; ora si localizzi un punto in (*a*), di coordinate (x, y), e si immagini di misurare la differenza fra ciascuna di esse e la sua media, ossia $(x - \mu_1)$ e $(y - \mu_2)$, rispettivamente. ∀ punto si scelga, le deviazioni hanno lo stesso segno: sono entrambe ⊕ se il punto prescelto si trova, rispettivamente, a destra e più in alto della media; sono entrambe ⊖ se il punto prescelto si trova, rispettivamente, a sinistra e più in basso della media. Quindi, non solo il loro prodotto è sempre ⊕, ma anche il valore medio – la *covarianza* di X e Y –

$$(1/N)\sum_{i=1}^{N}(x_i - \mu_1)(y_i - \mu_2) = \mathrm{Cov}(X, Y)$$

dove N = numero dei punti, sarà ⊕ e grande (= **21.25** – FIGURA 67).

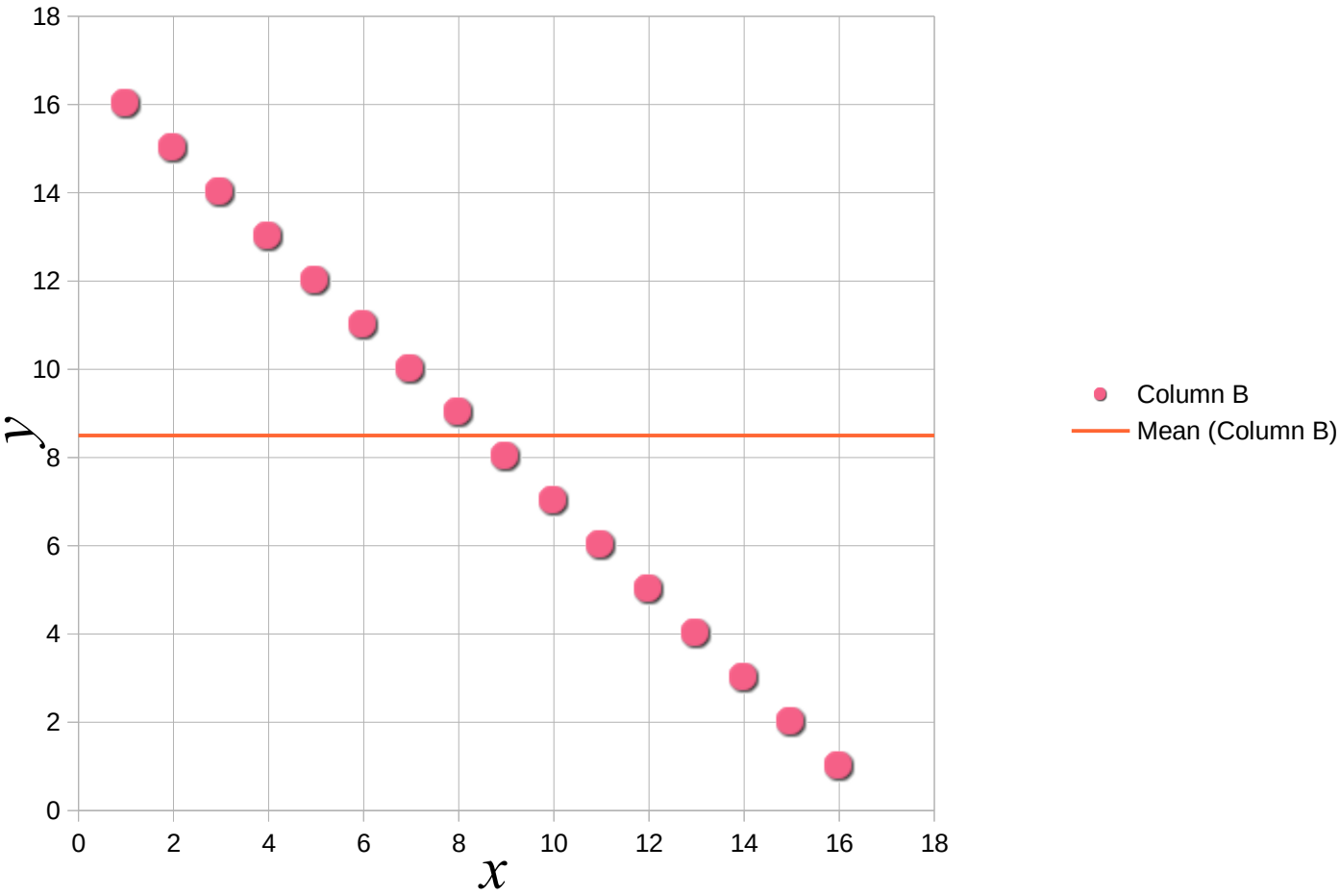


FIGURA 67. Esempio di dipendenza negativa fra variabili.

Anche le variabili X = *larghezza della groppa* di un vitello e Y = *facilità di nascita* dell'animale, sono legate fra loro da una relazione di dipendenza ma, in questo caso, all'aumentare di X si assiste ad una diminuzione di Y (FIGURA 68): tutte le coppie di deviazioni sono di segno opposto ed il valore medio del loro prodotto è uguale a **-21.25**.
Viceversa, nel caso illustrato dalla FIGURA 67(*b*) la dipendenza fra X ed Y è molto debole: Cov(X, Y) = **0.00084128**. Le deviazioni di ciascuna variabile dalla propria media,

$(x - \mu_1)$ e $(y - \mu_2)$, infatti, hanno lo stesso segno per alcuni punti e segno opposto per altri: allora, il risultato del prodotto $(x - \mu_1)\,(y - \mu_2)$ avrà segno $\oplus$ per alcuni punti e $\ominus$ per altri cosicché, sommando i vari prodotti per calcolarne la media, si otterrà un valore prossimo a zero (FIGURA 68).

	x	***y***		
	0.36309162	0.80544619		
	0.78809883	0.56734641		
	0.72466333	0.46373912		
	0.79706468	0.14674177		
	0.04164632	0.37277484		
	0.85607088	0.87911249		
	0.24314566	0.66884471		
	0.25803433	0.46695411		
	0.56499905	0.43346838		
	0.73696219	0.5084493		
MEDIA	**0.53737769**	**0.53128773**		
x - (*x medio*), *y* – (*y medio*)	-0.17428607	0.27415846	[*x* - (*x medio*)]*[*y* – (*y medio*)]	-0.047782
	0.25072114	0.03605868		0.00904067
	0.18728564	-0.0675486		-0.0126509
	0.25968699	-0.384546		-0.0998616
	-0.49573137	-0.1585129		0.07857981
	0.31869319	0.34782476		0.11084938
	-0.29423203	0.13755698		-0.0404737
	-0.27934336	-0.0643336		0.01797117
	0.02762136	-0.0978194		-0.0027019
	0.1995845	-0.0228384		-0.0045582

$$\sum[\boldsymbol{x} - (\boldsymbol{x}\ \textit{medio})]*[\boldsymbol{y} - (\boldsymbol{y}\ \textit{medio})]/10 \qquad \mathbf{0.00084128}$$

FIGURA 68. Il calcolo della covarianza di due variabili X e Y fra loro indipendenti. Gli scostamenti di ciascuna variabile dalla propria media sono indicati in verde se, essendo dello stesso segno, originano un prodotto positivo e in rosso se, essendo di segno opposto, originano un prodotto negativo. Sommando i vari prodotti per calcolarne la media si ottiene, così, un valore della covarianza prossimo a zero.

Il valore medio di $(x_i - \mu_1)(y_i - \mu_2)$, ossia la *covarianza* di X ed Y, fornisce una misura della dipendenza lineare fra le due variabili; valori positivi di Cov(X, Y) indicano che Y aumenta all'aumentare di X; valori negativi indicano che Y diminuisce all'aumentare di X; infine, un valore nullo o prossimo a zero di Cov(X, Y) indica che Y non dipende da X.

Purtroppo, è difficile utilizzare la covarianza come misura della dipendenza fra due variabili poiché il suo valore dipende dalla scala di misura; è necessario, quindi, standardizzare il valore della covarianza allo scopo di renderlo indipendente dalla particolare scala di misura utilizzata di volta in volta rendendo, così, possibile un confronto fra coppie diverse di variabili, ad esempio per stabilire quali di tali coppie di

grandezze sono legate da relazioni di dipendenza più forti e in quali casi, invece, le relazioni di dipendenza sono più deboli.
Dividendo la Cov(X, Y) per il *prodotto delle deviazioni standard delle due variabili* $\sigma_x \sigma_y$, si ottiene il *coefficiente di correlazione* r fra X ed Y:

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_x \sigma_y}.$$

La deviazione standard di X è la *radice quadrata del valore atteso del quadrato della deviazione di X dalla sua media*:

$$\sigma = \{E[(X - \mu)^2]\}^{1/2}.$$

Cosa significa «*valore atteso*»? Facciamo un esempio. Se si lancia una moneta, non si può sapere in anticipo se si otterrà testa o croce; se si effettuano, per es., 10 lanci, non è detto che si ottenga testa e, rispettivamente, croce lo stesso numero di volte (5): infatti, sono possibili molte combinazioni di risultati: 6 volte testa e 4 volte croce e viceversa, 7 volte testa e 3 volte croce e viceversa, 8 volte testa e due volte croce e viceversa e così via. Si può, inoltre, calcolare molto facilmente la probabilità che si verifichi ciascuna di tali combinazioni ma a noi interessa un altro aspetto del problema. Se si lancia la moneta un gran numero di volte, per es. 1000, ci si aspetta di ottenere testa circa 500 volte e croce altrettante volte; in pratica, se il numero di lanci (ripetizioni dell'esperimento casuale «*lancio della moneta*» fosse enormemente grande (tendesse ad ∞), si otterrebbe testa un numero di volte esattamente uguale alla metà dei lanci effettuati. Allora, si dice che il «*valore atteso*» della probabilità di ottenere testa (o croce) è 0,5.
La *deviazione standard* è un indice di dispersione: se il suo valore è grande i dati, ossia i valori della variabile X, sono molto dispersi poiché mediamente si discostano molto dalla media; in altri termini, X presenta una grande variabilità (curva blu della FIGURA 69, in cui $\sigma = 3$). Viceversa, una deviazione standard piccola indica che le osservazioni sono molto concentrate intorno al valore atteso, dal quale si discostano di poco: X è caratterizzata, in questo caso, da una scarsa variabilità (curva rossa della FIGURA 69, in cui $\sigma = 1$).

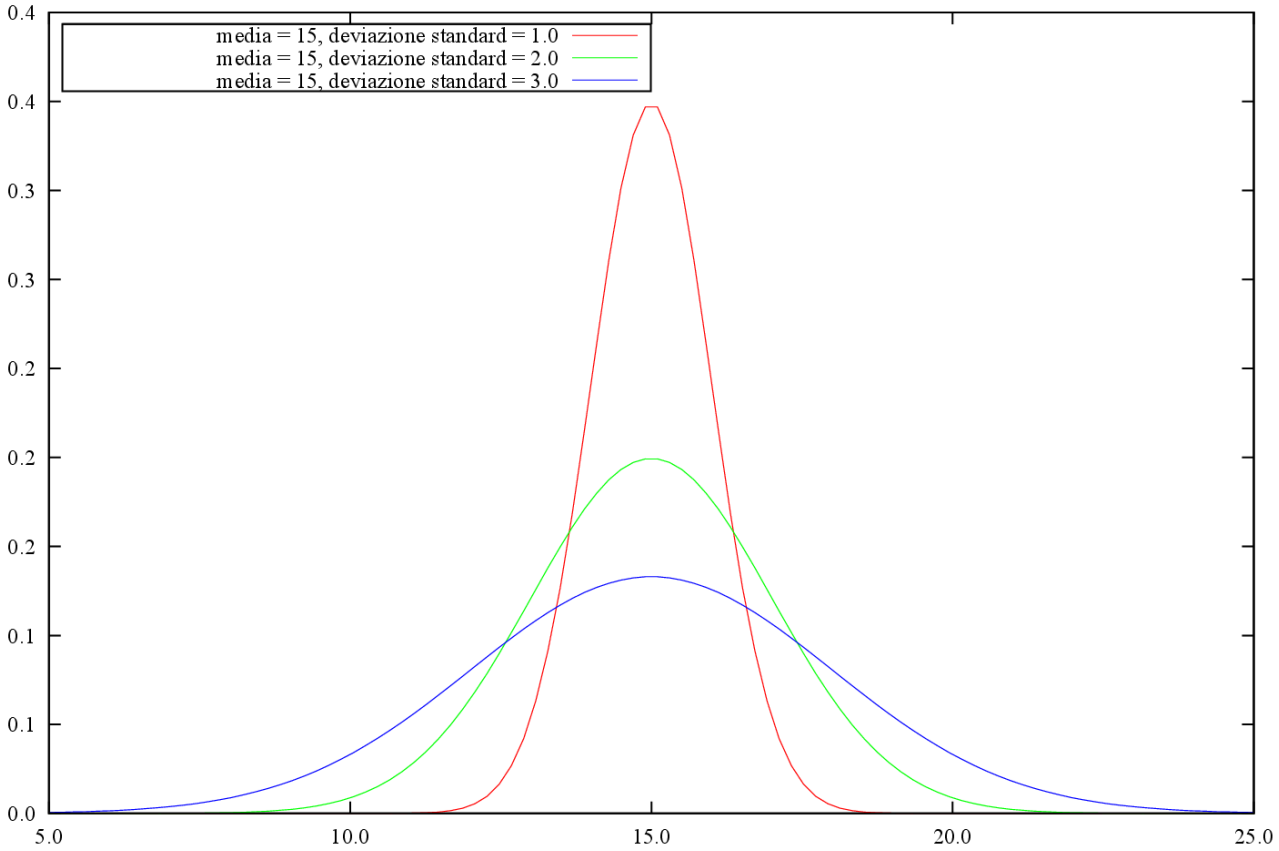


FIGURA 69. Grafico di tre distribuzioni normali con media = 15 e deviazione standard 1, 2 e 3, rispettivamente, costruito con i seguenti comandi:

```
gnuplot> load "NORMAL_FUNC_per_lezione_DESCRIZIONE_LINEARE"
gnuplot> set t postscript enhanced color solid font "TimesNewRoman, 12"
gnuplot> set o "NORMAL_FUNC_per_lezione_DESCRIZIONE_LINEARE.ps"
gnuplot> replot
```

VARIANZA ED EREDITABILITÀ DEI CARATTERI

Individui diversi manifestano stati differenti dei vari caratteri: tale variabilità fenotipica si misura con la **varianza**, il *valore atteso del quadrato della deviazione di X dalla sua media*

$$\sigma^2 = E[(X - \mu)^2],$$

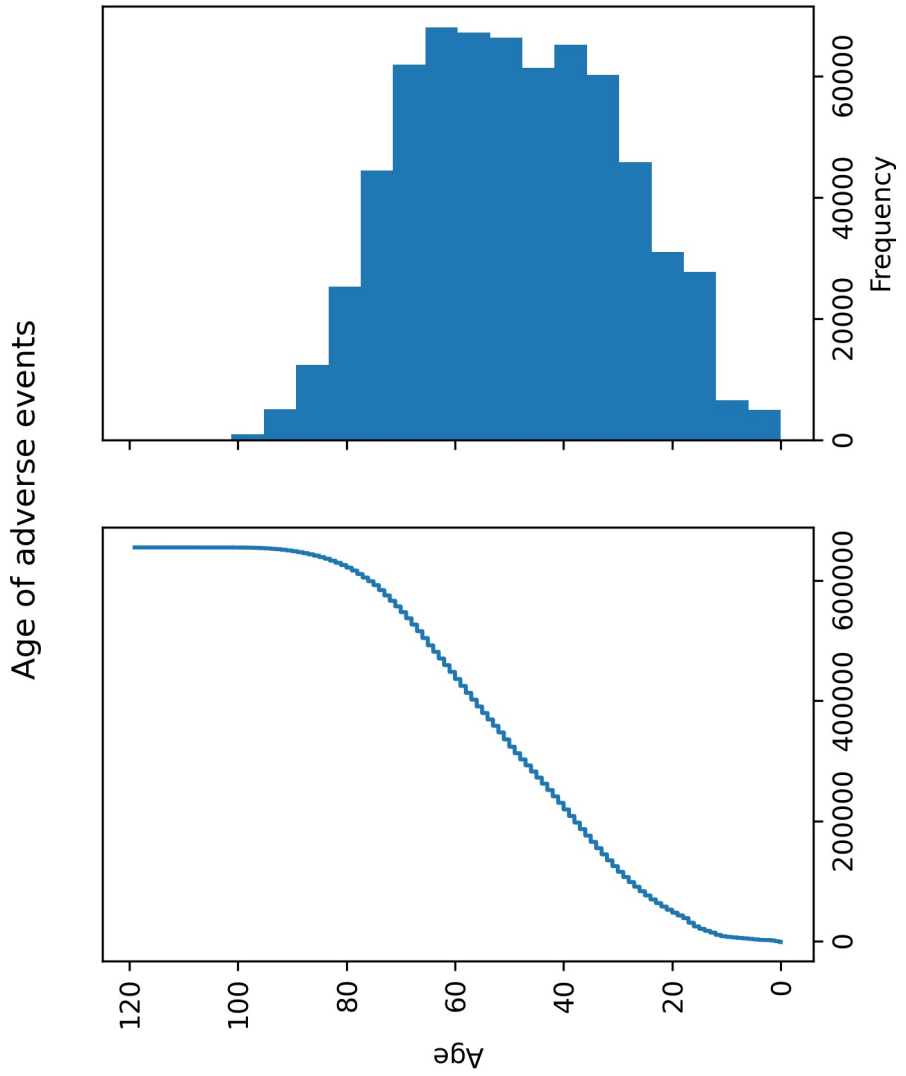
in cui X è lo stato del carattere (valore fenotipico) preso in considerazione.
La **varianza fenotipica** è la somma di varie componenti:

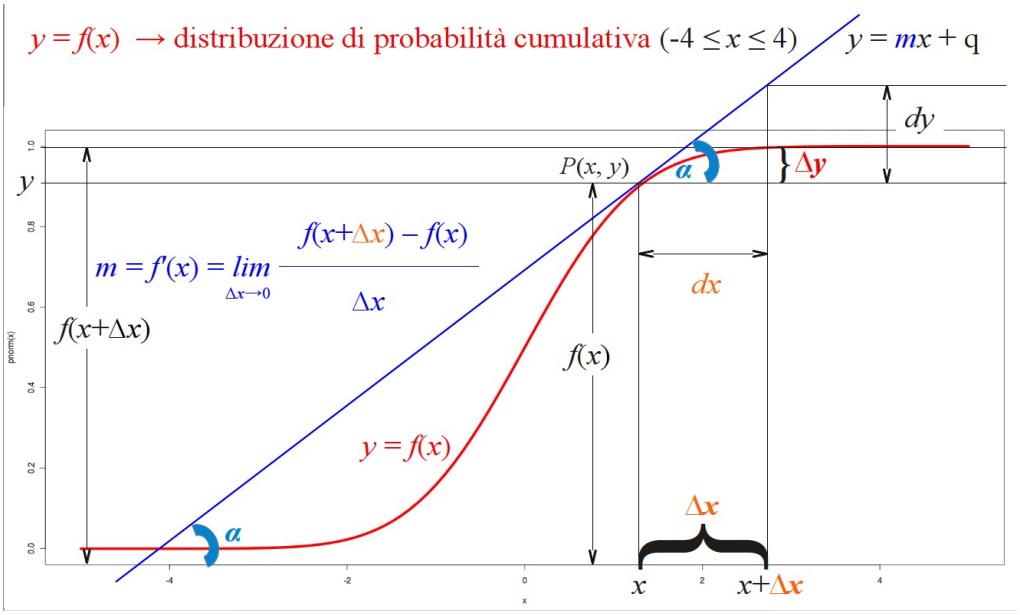
$$V_P = V_A + V_D + V_{IE} + V_{Ep} + V_{Et};$$

dove:
V_P = varianza fenotipica (varianza totale);
V_A = varianza additiva (dovuta ad ogni *locus* in cui i due alleli ivi presenti sommano i rispettivi effetti fenotipici);
V_D = varianza ascrivibile a ciascun *locus* in cui i due alleli che ad esso competono interagiscono fra loro instaurando rapporti di dominanza-recessività;
V_{Ep} = varianza dovuta alla variabilità dell'ambiente materno;
V_{Et} = varianza dovuta alla variabilità dei altri fattori ambientali «estrinseci» (alimentazione, sistema di allevamento, microclima).
L'**ereditabilità** ($\boldsymbol{h^2}$) di un carattere è la *quota di variabilità fenotipica totale ascrivibile all'effetto additivo dei geni*; in altri termini, è il **rapporto fra la varianza additiva e la varianza fenotipica**:

$$\boldsymbol{h^2} = V_A \,/\, V_P.$$

Ciò significa che un determinato carattere d'interesse economico può essere ereditato soltanto nella misura in cui la sua variabilità dipende dall'effetto additivo dei geni. Infatti, nel caso dei caratteri quantitativi le interazioni epistatiche e di dominanza sono trascurabili, a differenza di quanto si osserva per i caratteri che influenzano l'efficienza riproduttiva.


Age of adverse events
Frequency
0
20000
40000
60000
0
200000
400000
600000
120
100
80
60
40
20
0
Age

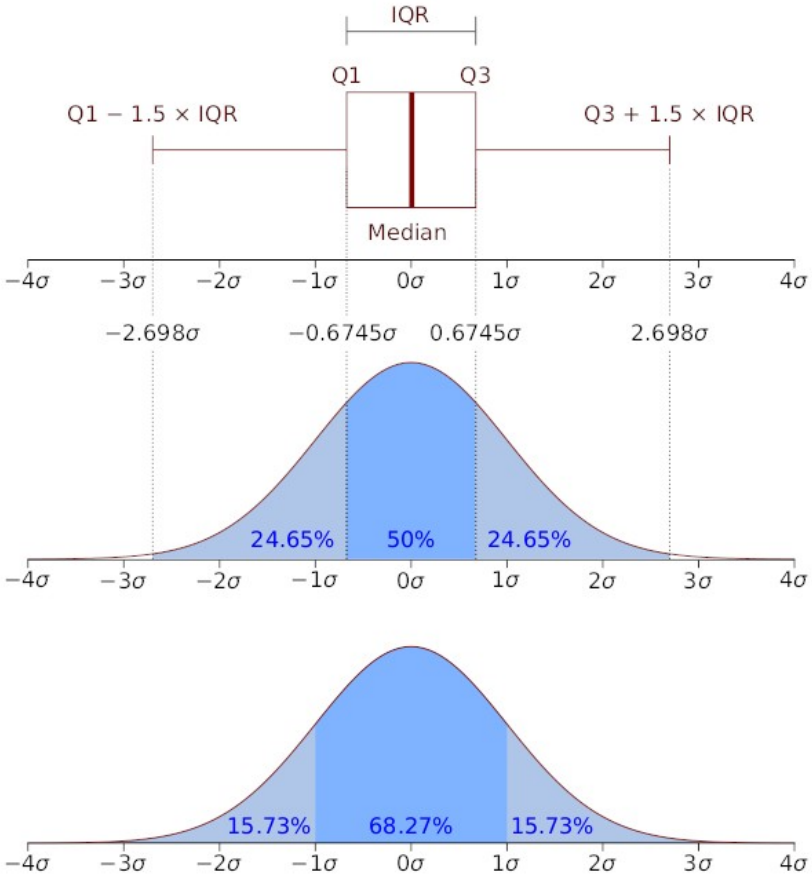


Il significato geometrico dei concetti di differenziale e densità di probabilità. $y = f(x)$ è la distribuzione di probabilità cumulativa della variabile casuale x, $y' = f'(x)$ è la sua derivata (la funzione di densità di x), $dy = f'(x)\,dx$ è il differenziale di y. **Δy** è la probabilità che la variabile sia compresa fra x e $x+\Delta x$, quindi **Δy** è l'area sottesa alla densità $f'(x)$ fra x e $x+\Delta x$: allora $f'(x) = dy/dx$ = AREA/dx = AREA/BASE = ALTEZZA.
Il grafico è stato disegnato con **R**, comando:

```
> curve(pnorm(x),-5,5,col="red",lwd="5").
```

In descriptive statistics, a **box plot** or **boxplot** is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their quartiles. [1] In addition to the box on a box plot, there can be lines (which are called *whiskers*) extending from the box indicating variability outside the upper and lower quartiles, thus, the plot is also termed as the **box-and-whisker plot** and the **box-and-whisker diagram**. Outliers that differ significantly from the rest of the dataset[2] may be plotted as individual points beyond the whiskers on the box-plot. Box plots are non-parametric: they display variation in samples of a statistical population without making any assumptions of the underlying statistical distribution[3] (though Tukey's boxplot assumes symmetry for the whiskers and normality for their length). The spacings in each subsection of the box-plot indicate the degree of dispersion (spread) and skewness of the data, which are usually described using the five-number summary. In addition, the box-plot allows one to visually estimate various L-estimators, notably the interquartile range, midhinge,

range, mid-range, and trimean. Box plots can be drawn either horizontally or vertically.


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
15.73%
68.27%
15.73%