# PRIMO APPELLO

Prova a risolvere i seguenti problemi, giustificando il ragionamento seguito.

*La motivazione del processo è molto più importante della risposta numerica.*

Puoi usare una calcolatrice o un regolo per i conti, così come R sul calcolatore dell'aula. PUOI CONTROLLARE I TUOI APPUNTI, LE NOTE DEL CORSO O UN LIBRO DI TESTO TRA QUELLI CONSIGLIATI. Lavora per tuo conto, senza aiuto esterno, ma discuti pure i problemi e le tue soluzioni **finita la prova**. I problemi **non** sono in ordine di difficoltà. Indica nome e cognome (e numero di matricola) sui fogli, così come il numero del problema o della domanda. **Lascia un po' di spazio per i commenti.** Le parti in R possono essere copiate sul foglio, oppure salvate come file (`.R`, `.txt`, `.Rhistory`) indicando con un commento (introdotto dal carattere `#`) a quale esercizio e domanda il codice si riferisce. Se parte di un problema è svolta in R, indicalo sul foglio in corrispondenza del punto dell'esercizio corrispondente. Non dimenticare di caricare il file nella risorsa esamionline al termine dell'esame.

Buon lavoro!

**Problema 1.** L'ufficio personale di una grande azienda vuole studiare il rapporto statistico tra le gratifiche e l'anzianità dei e delle dipendenti. Ogni volta che una gratifica viene assegnata a qualche dipendente, sia $X$ la frazione di dipendenti con anzianità maggiore rispetto alla persona premiata. Si pensa di modellizzare $X$ con una funzione di densità lineare, su un opportuno dominio, $f_X(x) = a\,x + b$.

Siano inoltre $\mu$ e $\sigma^2$ rispettivamente il valore atteso e la varianza di $X$.

1. Qual è il supporto di $X$?
2. Determinare i coeffcienti $a$ e $b$ in funzione di $\mu$.
3. Determinare $\sigma^2$ in funzione di $\mu$.
4. Determinare eventuali restrizioni ai possibili valori che possono assumere la speranza e la varianza di questa classe di variabili aleatorie.

**Soluzione.** Procediamo per punti.

1. Siccome $X$ per definizione è una frazione dei dipendenti può assumere valori solamente nell'intervallo $[0,1]$.
2. La funzione di densità di $X$ è quindi definita da

$$f_X(x) = \begin{cases} a\,x + b & 0 \leqslant x \leqslant 1 \\ 0 & \text{altrimenti} \end{cases}.$$

La condizione di normalizzazione ci dice che, dovendo $f_X$ essere una densità di probabilità, allora deve valere

$$1 = \int_0^1 f_X(x)\,\mathrm{d}x = \int_0^1 a\,x + b\,\mathrm{d}x = \frac{a}{2} + b.$$

Questo stabilisce un legame tra i due coefficienti, ma non risponde ancora alla domanda. Per far entrare in gioco il valore atteso usiamo la definizione

$$\mu = \mathrm{E}[X] = \int_0^1 x f_X(x)\,\mathrm{d}x = a\int_0^1 x^2\,\mathrm{d}x + b\int_0^1 x\,\mathrm{d}x = \frac{a}{3} + \frac{b}{2}.$$

Mettendo assieme questa identità e la precedente otteniamo

$$\begin{cases} \frac{a}{2} + b = 1 \\ \frac{a}{3} + \frac{b}{2} = \mu \end{cases}$$

da cui segue $a = 12\mu - 6$ e $b = 4 - 6\mu$.

3. Per la varianza vale un discorso simile a quello nel punto precedente: partiamo dalla definizione di varianza

$$\sigma^2 = \mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$$

e osserviamo che l'ultimo termine è $\mu^2$, quindi non lo vogliamo toccare. Sviluppiamo invece il primo termine

$$\mathrm{E}[X^2] = \int_0^1 x^2 f_X(x)\,\mathrm{d}x = a\int_0^1 x^3\,\mathrm{d}x + b\int_0^1 x^2\,\mathrm{d}x = \frac{a}{4} + \frac{b}{3}.$$

Dovendo scrivere la varianza in termini di $\mu$ sostituiamo la rappresentazione dei coefficienti $a$ e $b$ in termini di $\mu$

$$\mathrm{E}[X^2] = \frac{12\mu - 6}{4} + \frac{4 - 6\mu}{3} = \mu - \frac{1}{6},$$

da cui otteniamo

$$\mathrm{Var}[X] = \sigma^2 = \mu - \frac{1}{6} - \mu^2.$$

4. La forma stessa di $\sigma^2$ in termini di $\mu$ dovrebbe suggerirci che solo alcuni valori sono possibili (quelli che rendono la varianza non negativa), ma consideriamo la domanda più in generale, imponendo tutte le condizioni che abbiamo raccolto (implicitamente) finora.

   Come prima cosa usiamo il fatto che $f_X$ sia una densità, quindi sia in particolare non negativa. Ricordiamo che è una funzione lineare, quindi ci basta controllarlo agli estremi del supporto:

$$\begin{cases} f_X(0) = b \geqslant 0 \\ f_X(1) = a + b \geqslant 0 \end{cases}$$

   che possiamo rendere come condizioni su $\mu$, grazie alle rappresentazioni dei coefficienti ottenute in precedenza,

$$\begin{cases} 4 - 6\mu \geqslant 0 \\ 12\mu - 6 + 4 - 6\mu \geqslant 0 \end{cases} \longrightarrow \begin{cases} \mu \leqslant \frac{2}{3} \\ \mu \geqslant \frac{1}{3} \end{cases}.$$

   Queste condizioni sono compatibili con la condizione di non negatività della varianza, come si può verificare facilmente. Se partiamo dalla varianza dobbiamo risolvere una disequazione di secondo grado:

$$-\mu^2 + \mu - \frac{1}{6} \geqslant 0 \longrightarrow \mu^2 - \mu + \frac{1}{6} \leqslant 0$$

   da cui $\frac{1}{2} - \frac{1}{2\sqrt{3}} \leqslant \mu \leqslant \frac{1}{2} + \frac{1}{2\sqrt{3}}$, in cui gli estremi sono circa 0.21, minore di $\frac{1}{3}$ e 0.79, maggiore di $\frac{2}{3}$. Quindi affinché $f_X$ sia una densità e la varianza sia non negativa devono valere entrambe le condizioni, ossia $\frac{1}{3} \leqslant \mu \leqslant \frac{2}{3}$.

   Per quanto riguarda la varianza $\sigma^2$, è sicuramente non negativa e varia in termini di $\mu$ come una parabola a concavità verso il basso, il suo massimo sarà in corrispondenza del punto medio dell'intervallo, ossia per $\mu = \frac{1}{2}$, punto in cui $\sigma^2 = -\frac{1}{4} + \frac{1}{2} - \frac{1}{6} = \frac{1}{12}$. Il minimo di $\sigma^2$ sarà agli estremi dell'intervallo di definizione di $\mu$: $\sigma^2\left(\frac{1}{3}\right) = -\frac{1}{9} + \frac{1}{3} - \frac{1}{6} = \frac{1}{18} = \sigma^2\left(\frac{2}{3}\right)$.

**Problema 2.** Un gruppo di ricercatori ha condotto uno studio per confrontare il tempo che le persone dedicano quotidianamente a interazioni affettive con il proprio animale domestico. I partecipanti sono stati divisi in due gruppi: chi possiede un cane e chi possiede un gatto. Ognuno ha riportato il tempo (in minuti) dedicato in un giorno a interazioni affettive (carezze, giochi, ecc.) e il tempo (in anni) da cui convive con quell'animale.

I dati sono riportati nel file `interazioni_animali.csv`, che include le variabili:

- `specie`: ``cane'' o ``gatto''
- `affetto_minuti`: tempo totale in minuti per le interazioni affettive
- `anni_insieme`: anni di convivenza con l'animale

1. Esplora, anche con strumenti grafici, la distribuzione delle variabili per i due gruppi. Che differenze noti? Che somiglianze?
2. Come tratteresti i valori mancanti nei dati? Motiva la tua scelta.
3. Formula un test di ipotesi per verificare se, in media, le persone passano più tempo con il proprio cane rispetto al proprio gatto.
4. Applica il test al dataset. Riporta la statistica test, il p-value e un intervallo di confidenza al 98% per la differenza tra le medie.
5. Ripeti il test considerando solo le persone che convivono con l'animale da almeno 5 anni. Cosa cambia?
6. Quali conclusioni puoi trarre dai risultati? Cosa diresti ai ricercatori?

**Soluzione.** Come prima cosa carichiamo i dati in R e facciamo una prima analisi

```
dati <- read.csv("interazioni_animali.csv", stringsAsFactors = TRUE)
summary(dati)
```

```
      id           specie    affetto_minuti    anni_insieme
 Min.   :  1.00   cane :120   Min.   : 69.65   Min.   : 1.000
 1st Qu.: 55.75   gatto:100   1st Qu.:146.81   1st Qu.: 4.000
 Median :110.50               Median :170.96   Median : 5.000
 Mean   :110.50               Mean   :172.58   Mean   : 5.595
 3rd Qu.:165.25               3rd Qu.:201.24   3rd Qu.: 7.000
 Max.   :220.00               Max.   :294.79   Max.   :13.000
                              NA's   :22
```

La prima colonna è del tutto irrilevante, contiene un numero d'ordine. La seconda colonna contiene la variabile specie, che è categorica (con il parametro `stringsAsFactors = TRUE` abbiamo proprio chiesto che le variabili di tipo stringa, in lettura, siano considerate fattori ossia variabili categoriche). La terza colonna contiene la variabile `affetto_minuti`, che vediamo avere 22 NA. La quarta e ultima colonna contiene il numero di anni trascorsi assieme e non presenta NA.

Possiamo analizzare alcuni indicatori delle variabili per l'intera popolazione e per le due popolazioni identificate dalle due specie. Siccome abbiamo dei valori NA dobbiamo tenerne conto nel calcolo degli indicatori (vedremo poi come procedere per trattarli più in generale).

```
mean(dati$affetto_minuti, na.rm = TRUE)
mean(dati[dati$specie=="gatto",]$affetto_minuti, na.rm = TRUE)
mean(dati[dati$specie=="cane",]$affetto_minuti, na.rm = TRUE)
# 172.5817, 158.6723, 184.173

sd(dati$affetto_minuti, na.rm = TRUE)
```

```r
sd(dati[dati$specie=="gatto",]$affetto_minuti, na.rm = TRUE)
sd(dati[dati$specie=="cane",]$affetto_minuti, na.rm = TRUE)
# 40.47818, 36.65704, 40.00763

mean(dati$anni_insieme)
mean(dati[dati$specie=="gatto",]$anni_insieme)
mean(dati[dati$specie=="cane",]$anni_insieme)
# 5.595455, 5.06, 6.041667

sd(dati$anni_insieme)
sd(dati[dati$specie=="gatto",]$anni_insieme)
sd(dati[dati$specie=="cane",]$anni_insieme)
# 2.140067, 2.078364, 2.095998

boxplot(affetto_minuti ~ specie, data = dati)
boxplot(anni_insieme ~ specie, data = dati)

hist(dati[dati$specie=="gatto",]$affetto_minuti)
hist(dati[dati$specie=="cane",]$affetto_minuti)

hist(dati[dati$specie=="gatto",]$anni_insieme)
hist(dati[dati$specie=="cane",]$anni_insieme)
```

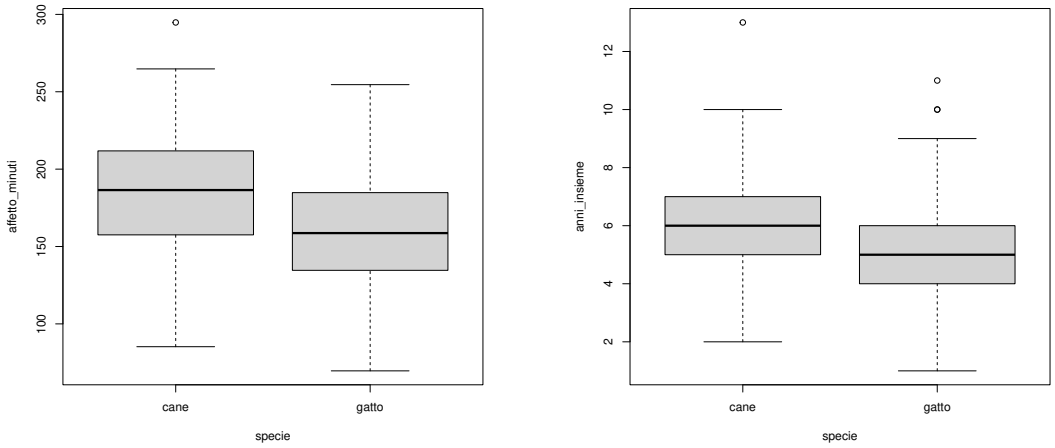


**Figura 1.** Boxplot delle variabili affetto (sinistra) e anni (destra) divisi per popolazione

Da queste primissime analisi osserviamo che le medie delle due popolazioni sembrano leggermente diverse sia per quanto riguarda il tempo affettivo, sia per quanto riguarda gli anni assieme. Viceversa le deviazioni standard sembrano simili. In particolare possiamo osservare che le deviazioni standard nelle singole sottopopolazioni sono minori di quelle della popolazione complessivamente, un possibile indizio che le due sottopopolazioni siano più concentrate, ma in posti diversi (i.e. con media diversa).

Nei boxplot possiamo osservare graficamente le stesse cose appena citate. Negli istogrammi in Figura 2 possiamo notare che la variabile `affetto_minuti` è approssimativamente normale nelle due sottopopolazioni, mentre lo stesso non si può dire della variabile `anni_insieme`. Un altro indizio sulla natura discreta di questa distribuzione potrebbe venirci dal fatto che i valori che assume sono tutti interi.

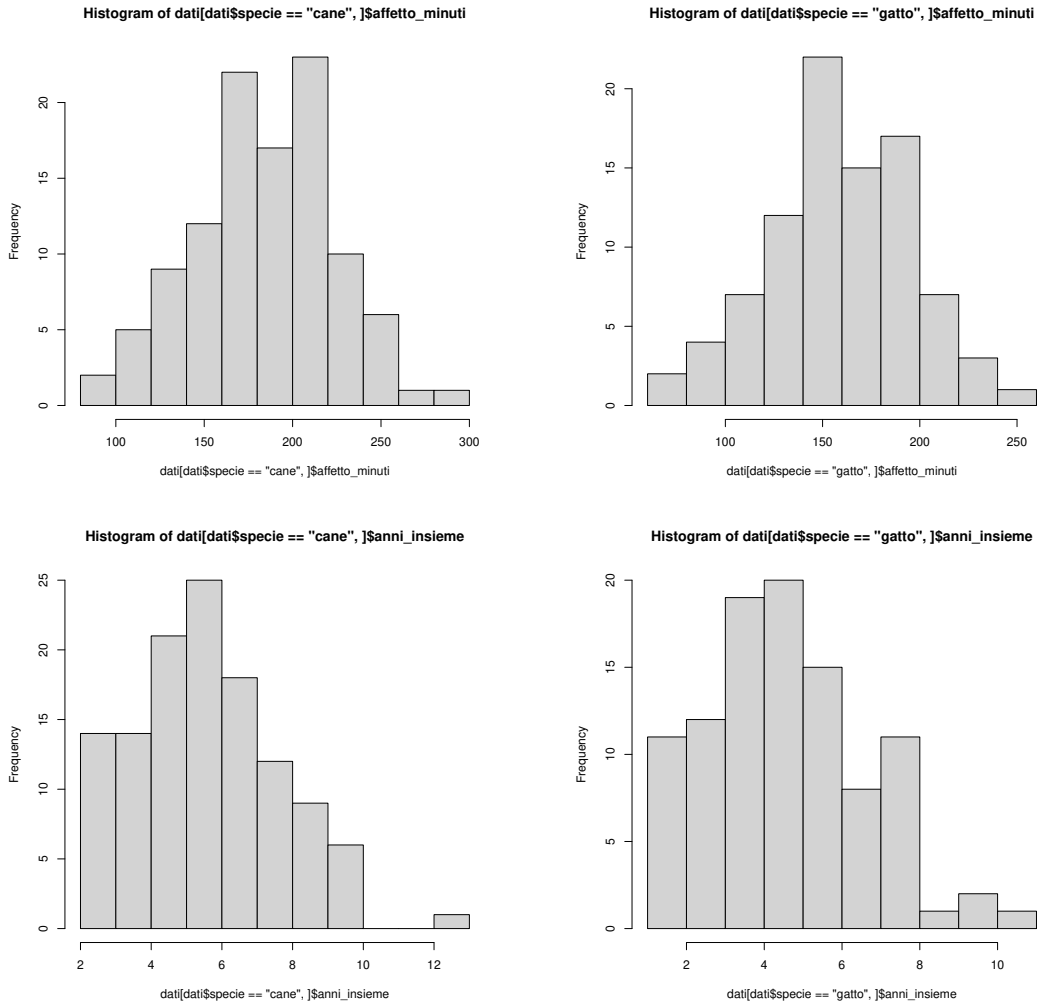


**Figura 2.** Istogrammi delle variabili affetto (sopra) e anni (sotto) per le due specie di animali domestici cani (a sinistra) e gatti (a destra)

Quello che ci aspettiamo nel seguito è quindi osservare una differenza tra le medie della variabile `affetto_minuti` nelle due sottopopolazioni, con i cani a media maggiore. Non possiamo ancora fare congetture sull'impatto che potrà avere la maggior durata in anni del rapporto.

Prima di proseguire con analisi più approfondite, pensiamo a cosa fare con i dati NA. Siccome questi compaiono solamente per la variabile `affetto_minuti` che è il nostro oggetto principale di indagine, possiamo rimuovere senza troppi problemi tutte le righe (i.e. le osservazioni) in cui il valore di quella variabile non compare. Possiamo farlo con `na.omit`:

```
dati_clean <- na.omit(dati)
mean(dati_clean$anni_insieme)
mean(dati_clean[dati_clean$specie=="gatto",]$anni_insieme)
mean(dati_clean[dati_clean$specie=="cane",]$anni_insieme)
# 5.676768, 5.155556, 6.111111

sd(dati_clean$anni_insieme)
sd(dati_clean[dati_clean$specie=="gatto",]$anni_insieme)
sd(dati_clean[dati_clean$specie=="cane",]$anni_insieme)
# 2.098238, 2.076682, 2.024769
```

Vale la pena osservare che ricalcolando ora media e deviazione standard di `anni_insieme` otteniamo dei valori diversi da prima (sia per la popolazione intera sia per le sottopopolazioni): infatti abbiamo tolto osservazioni. Non c'è un cambiamento analogo in `affetto_minuti` perché lì i valori non erano presenti nemmeno prima.

Per quanto riguarda l'impostazione del test statistico vogliamo valutare se una media sia maggiore di un'altra. Abbiamo quindi

$$H_0 : \mu_1 - \mu_2 \geqslant 0$$
$$H_1 : \mu_1 - \mu_2 < 0$$

in cui $\mu_1$ è la media per la popolazione dei gatti e $\mu_2$ la media per la popolazione dei cani. Abbiamo che $H_1$ prende questa forma perché quello che vorremmo mostrare è che ci sia evidenza statistica che la media per la popolazione cani sia *maggiore* (e quindi la differenza sia minore di zero).

Sotto ipotesi che le due popolazioni da confrontare siano normali e a varianza comune (omoschedasticità), dopo aver osservato che i dati non possono essere appaiati (la numerosità è diversa e a ogni riga corrisponde una sola specie), il test che useremo sarà un test t. In un test di questo tipo la statistica ancillare di riferimento è una t di Student a $n+m-2$ gradi di libertà, dove $n$ è il numero di osservazioni relative a gatti e $m$ il numero di quelle relative a cani. Abbiamo

$$T = \frac{(\bar{X}_n - \bar{Y}_m) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n} + \frac{1}{m}\right)}} \sim t\,(n-m-2)$$

dove:

- $\bar{X}_n$, $\bar{Y}_m$ sono le medie campionarie della variabile per gatti e cani rispettivamente
- $n = 90$, $m = 108$ sono le numerosità dei due gruppi;
- $S_p^2 = \frac{(n-1)\,S_1^2 + (m-1)\,S_2^2}{n+m-2}$.

Osserviamo che gli NA erano 22, quindi $n+m-2 = 220-22-2 = 196$, senza bisogno di guardare (dal summary) quanto valessero $n$ ed $m$ distintamente, anche se poi i valori ci occorrono per la parte relativa a $S_p$.

Per poter implementare questo test in R, come prima cosa controlliamo che le ipotesi di normalità e omoschedasticità siano verificate. Possiamo farlo graficamente, anche se esistono metodi più quantitativi:

```
par(mfrow = c(1, 2))
qqnorm(dati_clean$affetto_minuti[dati_clean$specie == "cane"],
       main = "QQ Plot - Cani")
qqline(dati_clean$affetto_minuti[dati_clean$specie == "cane"])
qqnorm(dati_clean$affetto_minuti[dati_clean$specie == "gatto"],
       main = "QQ Plot - Gatti")
qqline(dati_clean$affetto_minuti[dati_clean$specie == "gatto"])
par(mfrow = c(1, 1))

# Test statistici
shapiro.test(dati_clean$affetto_minuti[dati_clean$specie == ``cane"])

Shapiro-Wilk normality test
data:  dati_clean$affetto_minuti[dati_clean$specie == ``cane"]
W = 0.99427, p-value = 0.9365
```

```
shapiro.test(dati_clean$affetto_minuti[dati_clean$specie == ``gatto"])

Shapiro-Wilk normality test
data:  dati_clean$affetto_minuti[dati_clean$specie == ``gatto"]
W = 0.99477, p-value = 0.9797

var.test(affetto_minuti ~ specie, data = dati_clean)
F test to compare two variances

data:  affetto_minuti by specie
F = 1.1912, num df = 107, denom df = 89, p-value = 0.3948
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.795489 1.770428
sample estimates:
ratio of variances
          1.191162
```

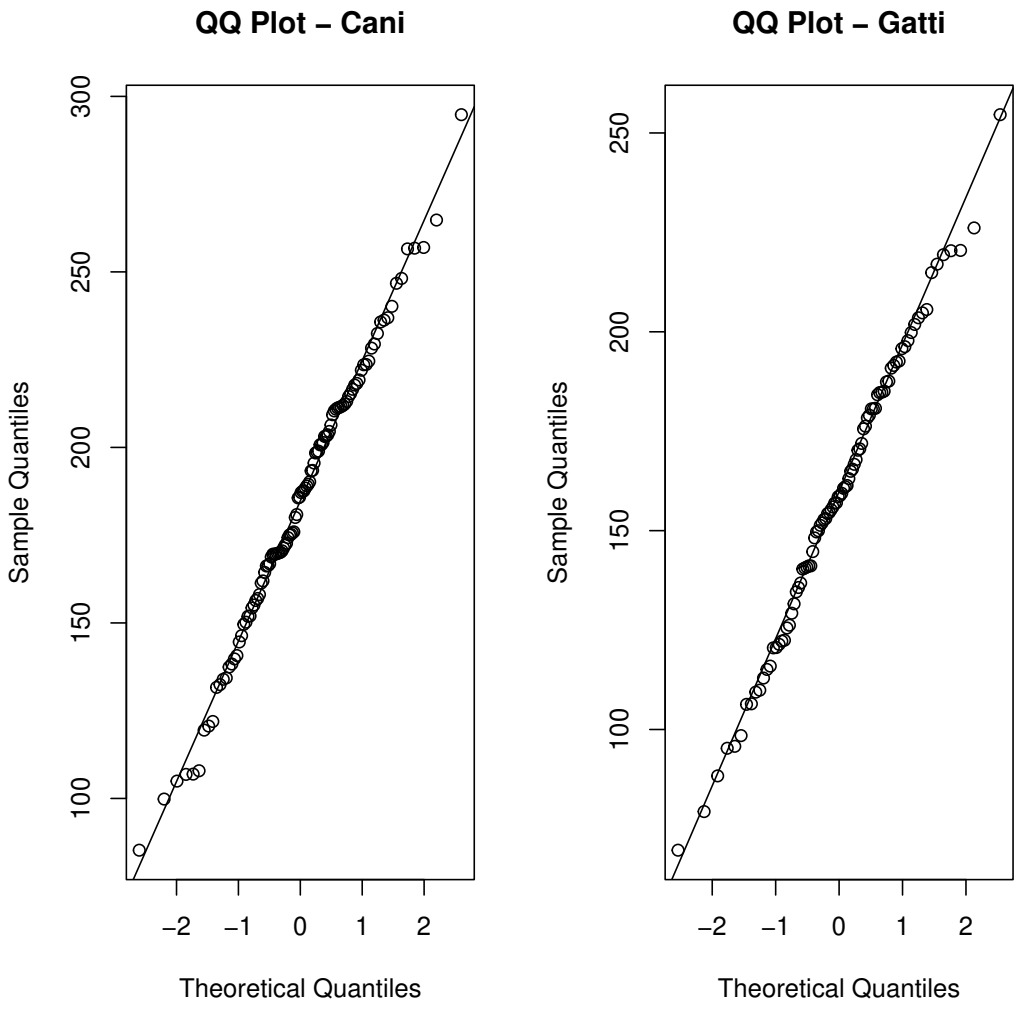


**Figura 3.** I due qqplot

Le rappresentazioni grafiche (e i test fatti, se ne guardiamo i p-value) confermano che si tratta di popolazioni approssimativamente Gaussiane e di uguale varianza.

Possiamo quindi procedere con il test t. A margine possiamo osservare che il test t è robusto rispetto all'ipotesi di non normalità, a patto che il campione sia sufficientemente grande, cosa verificata in questo esempio.

Usiamo qui una notazione leggermente diversa di quella vista altre volte, ossia la notazione funzionale. Più sotto è riportato l'equivalente codice in notazione standard, che, dando lo stesso risultato, appare più pesante.

```
t.test(affetto_minuti ~ specie,
       data = dati_clean,
       var.equal = TRUE, alternative = ``greater", conf.level = 0.98)
```

```
Two Sample t-test

data:  affetto_minuti by specie
t = 4.6381, df = 196, p-value = 3.208e-06
alternative hypothesis: true difference in means between group cane and group
gatto is greater than 0
98 percent confidence interval:
 14.13334      Inf
sample estimates:
 mean in group cane mean in group gatto
           184.1730            158.6723
```

```
t.test(dati_clean[dati_clean$specie=="cane"]$affetto_minuti,
       dati_clean[dati_clean$specie=="gatto"]$affetto_minuti,
       var.equal = TRUE, alternative = "greater", conf.level = .98)
```

Abbiamo quindi evidenza statistica che i proprietari e le proprietarie di cani dedichino loro (in media) più tempo "affettivo".

Per quanto riguarda il caso con almeno 5 anni di convivenza le cose non sono troppo diverse (anche qui sarebbe opportuno controllare che le distribuzioni siano approssimativamente normali e che le varianze siano uguali). Possiamo notare che le medie sono più vicine e che l'estremo inferiore dell'intervallo di fiducia si è avvicinato a 0.

```
t.test(affetto_minuti ~ specie, data = dati_clean[dati_clean$anni_insieme>=5,],
var.equal = TRUE, alternative = ``greater", conf.level = 0.98)
```

```
Two Sample t-test

data:  affetto_minuti by specie
t = 3.2749, df = 138, p-value = 0.0006683
alternative hypothesis: true difference in means between group cane and group
gatto is greater than 0
98 percent confidence interval:
 8.09756     Inf
sample estimates:
 mean in group cane mean in group gatto
           181.4970            159.4264
```

**Problema 3.** Alex lancia un normale dado a 6 facce, ottenendo un punteggio $X$. Tira poi $X$ monete bilanciate.

1. Qual è la probabilità, sapendo che non è uscita alcuna testa, che Alex abbia ottenuto 3 sul dado?

2. Se invece $X$ fosse il numero di messaggi di spam che Alex ha ricevuto il 22 giugno (che in media sono $\alpha$ al giorno) e lanciasse una moneta per ciascun messaggio spam ricevuto, quale sarebbe, al variare di $k$, la probabilità che $X=k$ sapendo che non è uscita alcuna testa?

3. Di che distribuzione si tratta? Qual è il suo valore atteso?

**Soluzione.** Procediamo per punti.

1. Indichiamo con $N$ l'evento "non è uscita alcuna testa tra le monete lanciate". La probabilità richiesta è, riscrivendola con il teorema di Bayes e la formula di fattorizzazione,

$$\begin{aligned} P(X=3\,|\,N) &= \frac{P(N\,|\,X=3)\,P(X=3)}{P(N)} \\ &= \frac{P(N\,|\,X=3)\,P(X=3)}{\sum_{k=1}^{6} P(N\,|\,X=k)\,P(X=k)} \\ &= \frac{\frac{1}{8}\cdot\frac{1}{6}}{\sum_{k=1}^{6}\frac{1}{2^k}\cdot\frac{1}{6}} \\ &= \frac{1}{48}\cdot\frac{6}{1}\cdot\frac{64}{63}=\frac{8}{63}\approx 12.7\% \end{aligned}$$

2. In questo caso $X$ ha una distribuzione di Poisson di parametro $\alpha$, quindi la sua densità discreta (o funzione di massa di probabilità) è

$$p_X(k)=P(X=k)=\frac{\alpha^k}{k!}\mathrm{e}^{-\alpha}.$$

Vogliamo calcolare le probabilità

$$\begin{aligned} P(X=k\,|\,N) &= \frac{P(N\,|\,X=k)\,P(X=k)}{P(N)} \\ &= \frac{\left(\frac{1}{2}\right)^k\frac{\alpha^k}{k!}\mathrm{e}^{-\alpha}}{\sum_{i=0}^{+\infty}\left(\frac{1}{2}\right)^i\frac{\alpha^i}{i!}\mathrm{e}^{-\alpha}}. \end{aligned}$$

Ora lavoriamo separatamente su numeratore e denominatore. Per il numeratore abbiamo

$$\left(\frac{1}{2}\right)^k\frac{\alpha^k}{k!}\mathrm{e}^{-\alpha}=\frac{\left(\frac{\alpha}{2}\right)^k}{k!}\mathrm{e}^{-\alpha},$$

mentre per il denominatore

$$\sum_{i=0}^{+\infty}\left(\frac{1}{2}\right)^i\frac{\alpha^i}{i!}\mathrm{e}^{-\alpha}=\mathrm{e}^{-\alpha}\sum_{i=0}^{+\infty}\frac{\left(\frac{\alpha}{2}\right)^i}{i!}=\mathrm{e}^{-\alpha}\cdot\mathrm{e}^{\frac{\alpha}{2}}=\mathrm{e}^{-\frac{\alpha}{2}}.$$

Mettendo assieme il tutto

$$P(X=k\,|\,N)=\frac{\left(\frac{\alpha}{2}\right)^k}{k!}\mathrm{e}^{-\alpha}\cdot\frac{\left(\frac{\alpha}{2}\right)^k}{k!}\mathrm{e}^{\frac{\alpha}{2}}=\frac{\left(\frac{\alpha}{2}\right)^k}{k!}\mathrm{e}^{-\frac{\alpha}{2}}$$

3. La densità discreta così ottenuta è quella di una distribuzione di Poisson di parametro $\frac{\alpha}{2}$. Tale distribuzione ha valore atteso uguale a $\frac{\alpha}{2}$, in particolare possiamo scrivere

$$\mathrm{E}[X|N] = \frac{\alpha}{2}.$$