Laboratorio 5

Test t di Student

5.1 Analisi del dataset FRUITFLY.DAT

I dati `fruitfly.dat` si riferiscono alla fecondità dei moschini della frutta, valutata come numero medio giornaliero di uova prodotte nei primi 14 giorni di vita, come rilevate da ciascuna di 25 femmine appartenenti a tre linee genetiche: RS, SS e NS. Si vuole verificare:

1. se c'è differenza fra le prime due linee e la terza;
2. se c'è differenza fra le prime due.

```
> fruit <- read.table('fruitfly.dat', col.names=c('RS','SS','NS'))
> attach(fruit)
> fruit
     RS   SS   NS
 1 12.8 38.4 35.4
 ......
25 23.6 10.8 47.4
```

Scegliamo come strumento per l'analisi il test t di Student. Prima di utilizzarlo dobbiamo verificare che siano soddisfatte le ipotesi di base: normalità e omoschedasticità dei dati.

5.1.1 Confronto fra (RS, SS) e NS

Definiamo la prima quantità d'interesse:

```
> RSS <- c(RS,SS)
```

Per avere un'idea di come sono distribuiti i dati (simmetria, dispersione ...):

```
> boxplot(RSS, NS)
> par(mfrow=c(2,1))     # per vedere piu' grafici nella stessa finestra
> hist(RSS, nclass=8, freq=F)
> hist(NS, nclass=8, freq=F)
```

Per la normalità:

```
> par(mfrow=c(1,2), pty='s')
> qqnorm(RSS)
> qqline(RSS)
> qqnorm(NS)
> qqline(NS)
```

Per verificare l'omoschedasticità possiamo dare un'occhiata alle varianze campionarie.

```
> var(RSS)
[1] 77.00249
> var(NS)
[1] 79.9596
```

Queste paiono abbastanza simili. Inoltre, anche dal confronto dei boxplot, pareva che le due distribuzioni empiriche avessero variabilità comparabile.
Assumendo che i dati in `RSS` e in `NS` siano realizzazioni indipendenti, rispettivamente da una $N(\mu_{RSS}, \sigma^2_{RSS})$ e una $N(\mu_{NS}, \sigma^2_{NS})$, si può calcolare il test del rapporto di verosimiglianza (W_P) per verificare $H_0 : \sigma^2_{RSS} = \sigma^2_{NS}$ contro $H_0 : \sigma^2_{RSS} \neq \sigma^2_{NS}$.
Questo test viene condotto tramite il comando

```
> var.test(RSS,NS)

        F test to compare two variances

data:  RSS and NS
F = 0.963, num df = 49, denom df = 24, p-value = 0.8833
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4564086 1.8658070
sample estimates:
ratio of variances
         0.9630174
```

Il risultato indica che si possono considerare le due varianze uguali.
Adesso possiamo utilizzare la funzione `t.test` di `R` per un test t di Student a due campioni bilaterale, per saggiare l'ipotesi $\mathrm{H}_0 : \mu_{RSS} = \mu_{NS}$.

```
> t.test(NS, RSS, var.equal=T)

        Two Sample t-test

data:  NS and RSS
t = 4.1286, df = 73, p-value = 9.587e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  4.619188 13.240812
```

```
sample estimates:
mean of x mean of y
   33.372    24.442
```

La quantità `t` è il valore osservato della statistica test; `df` sono i gradi di libertà della distribuzione della statistica test sotto H_0, `p-value` è il livello di significatività osservato, cioè $2 * \Pr(t_{73} > |t|)$. Verifichiamolo:

```
> 2*( 1 - pt(4.1286, 73) )
[1] 9.585708e-05
```

Poiché il valore-p è molto basso (< 0.01), si rifiuta l'ipotesi di uguaglianza delle medie fra i due gruppi RSS e NS. Se si fissa il livello del test a 0.05, allora si può verificare che il valore osservato t della statistica test si trova nella regione di rifiuto. Infatti la soglia destra della regione di rifiuto, data dal quartile di livello 0.975 di una distribuzione t con 73 gradi di libertà, cade nel punto:

```
> qt(0.975, 73)
[1] 1.992997
```

e la regione di rifiuto è $R = (|t| > 1.992997)$.

Passiamo ora al confronto fra RS e NS. Cominciamo col verificare la normalità:

```
> par(mfrow=c(1,3))
> boxplot(RS, NS)
> hist(RS, freq=F)
> qqnorm(RS)
> qqline(RS)
```

L'adattamento ad una normale pare buono. Circa l'omoschedasticità:

```
> var(RS)
[1] 60.41007
> var(NS)
[1] 79.9596
```

Le varianze campionarie sembrano diverse. Il confronto dei boxplot lasciava però intuire una variabilità comparabile. Proviamo a considerare il test per la verifica dell'ipotesi di uguaglianza delle varianze

```
> var.test(RS, NS)

        F test to compare two variances

data:  RS and NS
F = 0.7555, num df = 24, denom df = 24, p-value = 0.4974
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
```

```
 0.3329286 1.7144557
sample estimates:
ratio of variances
         0.7555074
```

Ancora una volta, è adeguato supporre l'omoschedasticità. Procediamo con il test:

```
> t.test(NS, RS, var.equal=T)

         Standard Two-Sample t-Test

data:  NS and RS
t = 3.4251, df = 48, p-value = 0.001268
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  3.351692 12.880308
sample estimates:
 mean of x mean of y
    33.372    25.256
```

Concludiamo anche in questo caso con il rifiuto di H_0.

NB: Nel caso le varianze delle due popolazioni non possano essere assunte uguali, R permette di utilizzare un test approssimato per confrontare le medie dei due campioni. In questo caso, la chiamata è del tipo:

```
> t.test(campione1, campione2, var.equal=F)
```

Esercizio: Svolgere l'analisi sui dati trasformati mediante la trasformata logaritmica, specificando bene l'ipotesi posta sotto verifica.
Esercizio: Fare il confronto fra SS e NS.

5.1.2 Confronto fra RS e SS

Svolgiamo le solite analisi preliminari.

```
> par(mfrow=c(1,2))
> boxplot(RS,SS)
> qqnorm(SS)
> qqline(SS)
```

I dati del secondo gruppo sono asimmetrici e le variabilità paiono meno confrontabili dei casi precedenti.

```
> var(RS)
[1] 60.41007
> var(SS)
[1] 95.42293
```

I due gruppi hanno varianze campionarie molto diverse. Proviamo a considerare il test per la verifica dell'ipotesi di uguaglianza delle varianze.

```
> var.test(RS,SS)

        F test to compare two variances

data:  RS and SS
F = 0.6331, num df = 24, denom df = 24, p-value = 0.2698
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2789774 1.4366273
sample estimates:
ratio of variances
          0.633077
```

Nonostante l'apparente diversità, il test indica una differenza non significativa. Passiamo quindi al test t.

```
> t.test(RS, SS, var.equal=F)

         Standard Two-Sample t-Test

data:  RS and SS
t = 0.6521, df = 48, p-value = 0.5176
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.39843  6.65443
sample estimates:
 mean of x mean of y
    25.256    23.628
```

In questo caso si accetta l'ipotesi H_0. Le soglie della regione di accettazione possono essere trovate tramite:

```
> qt(0.975, 48)
[1] 2.010635
```

Quindi la regione di accettazione è costituita da tutti i valori t tali che $|t| < 2.010635$, che comprende anche il nostro t osservato.

Esercizio: Provare a ripetere l'esercizio trasformando il dati mediante trasformazione logaritmica. In particolare, commentare l'effetto della trasformazione su asimmetria e normalità.

```
> detach()
```

5.2 Analisi del dataset CAPTOPRIL.DAT

I dati sono relativi a misurazioni della pressione sistolica e diastolica del sangue di un gruppo di 15 pazienti, prima e dopo la somministrazione del farmaco *captopril*. Si vuole verificare l'efficacia del farmaco nell'abbassare le due pressioni.

```
> capto <- read.table('capto.dat')
> attach(capto)
> capto
    Sp  Sd  Dp  Dd
 1 210 201 130 125
......
15 154 131 100  82
```

In questo caso non è possibile condurre un test t a due campioni per verificare l'uguaglianza delle medie delle distribuzioni di, ad esempio, `Sp` e di `Sd`. Infatti, se anche l'ipotesi di normalità fosse adeguata, l'ipotesi di indipendenza non può certamente esserlo, essendo `Sp` e `Sd` misurazioni in tempi diversi ma sugli stessi soggetti. Il problema non si pone se si considera la differenza `SD= Sd -Sp`. Infatti, assumendo che le coppie (`Sd`,`Sp`) siano realizzazioni indipendenti da una normale bivariata con componenti marginali $N(\mu_{Sd}, \sigma^2_{Sd})$ e $N(\mu_{Sp}, \sigma^2_{Sp})$, rispettivamente, e con coefficiente di correlazione ρ, allora l'insieme delle differenze `SD= Sd -Sp` è un campione casuale da una $N(\mu_{SD}, \sigma^2_{SD})$, con $\mu = \mu_{Sd} - \mu_{Sp}$ e $\sigma^2 = \sigma^2_{Sd} + \sigma^2_{Sp} - 2\rho\sigma^2_{Sd}\,\sigma^2_{Sp}$. Quindi, si può verificare l'ipotesi nulla $H_0 : \mu_{Sd} = \mu_{Sp}$, verificando l'ipotesi $H_0 : \mu_{SD} = 0$, attraverso un test t ad un campione.

Costruzione delle differenze e verifica della normalità.

```
> SD <- Sd-Sp
> DD <- Dd-Dp
> par(mfrow=c(2,1))
> boxplot(SD)
> qqnorm(SD)
> qqline(SD)
```

La numerosità campionaria è bassa ...

```
> boxplot(DD)
> qqnorm(DD)
> qqline(DD)
```

Procediamo con l'analisi delle differenze di pressione sistolica. Si usa il test t ad un campione per verificare l'ipotesi $\mathrm{H}_0 : \mu_{SD} = 0$.

```
> t.test(SD)

        One-sample t-Test

data:  SD
```

```
t = -8.1228, df = 14, p-value = 1.146e-06
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -23.93258 -13.93409
sample estimates:
 mean of x
 -18.93333
```

Si rifiuta l'ipotesi che SD abbia media nulla. Il test t ad un campione sulle differenze corrisponde al test t per dati appaiati.

```
> t.test(Sd, Sp, paired=T)

         Paired t-Test

data:  Sd and Sp
t = -8.1228, df = 14, p-value = 1.146e-06
alternative hypothesis: true mean of differences is
 not equal to 0
95 percent confidence interval:
 -23.93258 -13.93409
sample estimates:
 mean of x - y
     -18.93333
```

Il risultato è identico a quello ottenuto dal test t ad un campione sulle differenze. In realtà si richiede di verificare se c'è stato un miglioramento, cioè se la pressione si è abbassata. E' percio' piu' adeguato considerare un problema di verifica d'ipotesi unilaterale:

$$\mathrm{H}_0 : \mu_{SD} \geq 0$$
$$\mathrm{H}_1 : \mu_{SD} < 0.$$

```
> t.test(Sd, Sp, paired=T, alternative='l')

         Paired t-Test

data:  Sd and Sp
t = -8.1228, df = 14, p-value = 5.732e-07
alternative hypothesis: true mean of differences is
 less than 0
95 percent confidence interval:
        -Inf -14.82793
sample estimates:
 mean of x - y
     -18.93333
```

Si rifiuta l'ipotesi nulla. Quindi, è ragionevole pensare che il farmaco porti all'abbassamento della pressione. Il quantile di riferimento a livello 0.05 è

```
> qt(0.05, 14)
[1] -1.76131
```

$t < -1.76131$ indica che siamo nella regione di rifiuto del test che è la coda sinistra della t di Student con 14 gradi di libertà.

Esercizi: Ripetere le analisi per la pressione diastolica. E se si volesse confrontare il rapporto delle due pressioni sistolica e diastolica prima e dopo la somministrazione del farmaco?