Introduzione alla statistica Inferenziale con R

Stefano Bussolon

18 marzo 2019

Introduzione alla statistica Inferenziale con R

Il fine dell'analisi inferenziale è di fare delle inferenze su di una popolazione a partire dalle osservazioni di un campione.

Il fine dell'analisi inferenziale univariata è di stimare il valore di un parametro della popolazione a partire da una statistica calcolata sul campione.
Il fine dell'analisi inferenziale bivariata è quello di stimare la significatività di una relazione fra due variabili.
Le analisi multivariate sono concettualmente un'estensione dell'analisi bivariata.

Le analisi bivariate

Tipi di variabili e statistica

La tipologia di statistica inferenziale da applicare si basa sulla tipologia di variabili. Possiamo distinguere fra variabili categoriali, ordinali, ad intervalli e a rapporti.

Queste quattro tipologie possono essere raggruppate in variabili nominali (categoriali e, generalmente, ordinali) e variabili numeriche (a intervalli, a rapporti).

La tipologia di statistica che può essere applicata si basa sulla tipologia delle variabili indipendenti e dipendenti.

	dipendente nominale	dipendente numerica
indipendente nominale	chi quadro	t-test, ANOVA
indipendente numerica	analisi discriminante, regressione logistica	correlazione, regressione

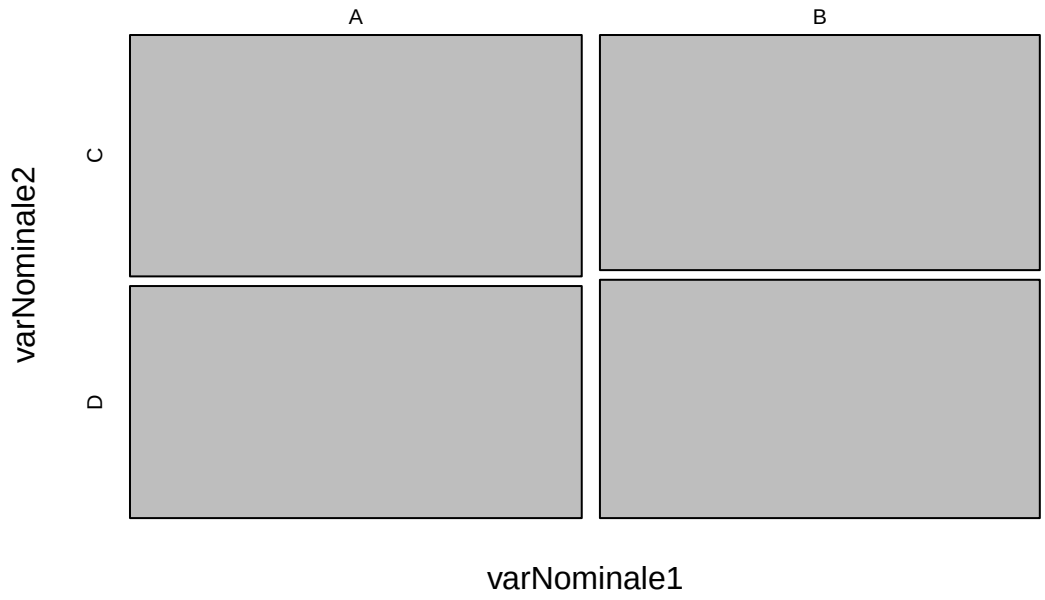
La distinzione fra variabile indipendente e dipendente è centrale in caso di indipendente nominale e dipendente numerica, è più sfumata nel caso di due variabili nominali o due variabili numeriche.

Due variabili nominali

Per valutare la relazione fra due variabili nominali, viene utilizzata la statistica χ^2. In `R`, per calcolare il test di χ^2 si usa la funzione `chisq.test`.

Per mostrarne l'utilizzo, generiamo due variabili categoriali indipendenti, visualizziamo il grafico delle variabili, e calcoliamo la statistica.

```
# stabilisco il numero di osservazioni
osservazioni <- 600
varNominale1 <- sample(c("A","B"),osservazioni, replace = TRUE)
varNominale2 <- sample(c("C","D"),osservazioni, replace = TRUE)
plot(table(varNominale1,varNominale2))
```



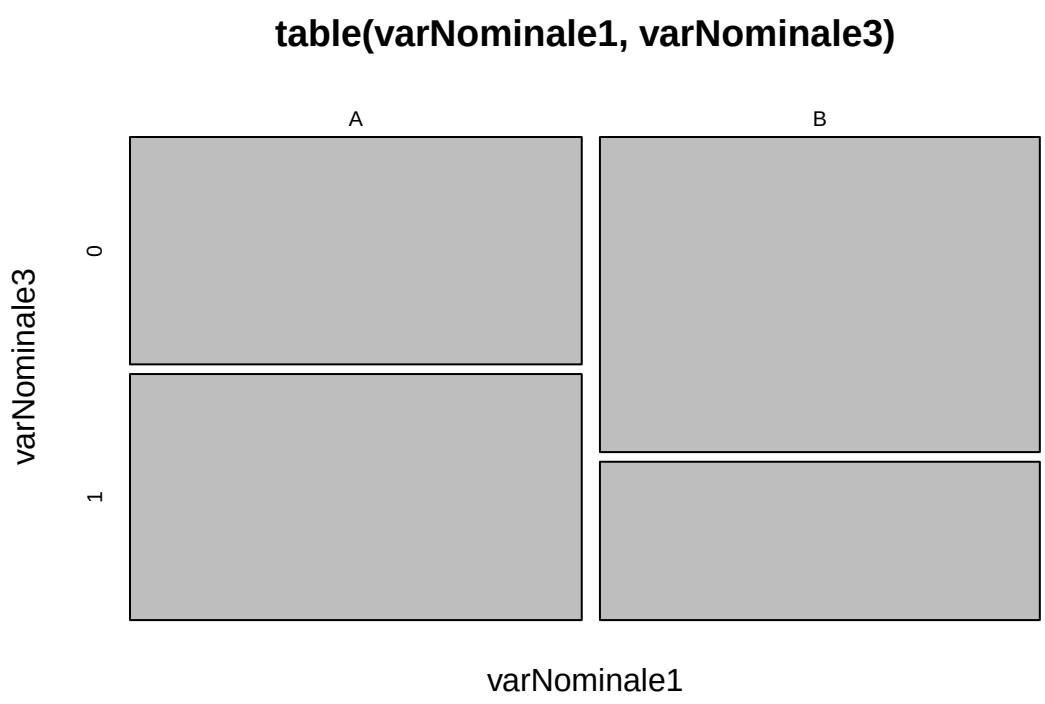

```
chi2_a <- chisq.test(varNominale1,varNominale2)
chi2_a
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  varNominale1 and varNominale2
## X-squared = 0.058951, df = 1, p-value = 0.8082
```

La funzione ci dice che ha applicato la statistica Pearson's Chi-squared test with Yates' continuity correction. Che i gradi di libertà sono 1 ((2-1)*(2-1)). Il valore della statistica è 0.0589512 ed il p-value è 0.808. Come prevedibile il p-value è superiore a `0.05`, e dunque non si può rifiutare l'ipotesi nulla di indipendenza fra le due variabili.

Nell'esempio sucessivo, creiamo `varNominale3`, che invece non è indipendente da `varNominale1`. Disegnamo il grafico e calcoliamo la statistica.

```
isNominaleA <- as.integer(varNominale1=="A")
# genero una variabile nominale che "dipende" da varNominale1
varNominale3 <- rbinom(length(varNominale1),1,.35+isNominaleA*.2)
plot(table(varNominale1,varNominale3))
```




```
chi2_b <- chisq.test(varNominale1,varNominale3)
chi2_b
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  varNominale1 and varNominale3
## X-squared = 20.275, df = 1, p-value = 6.707e-06
```

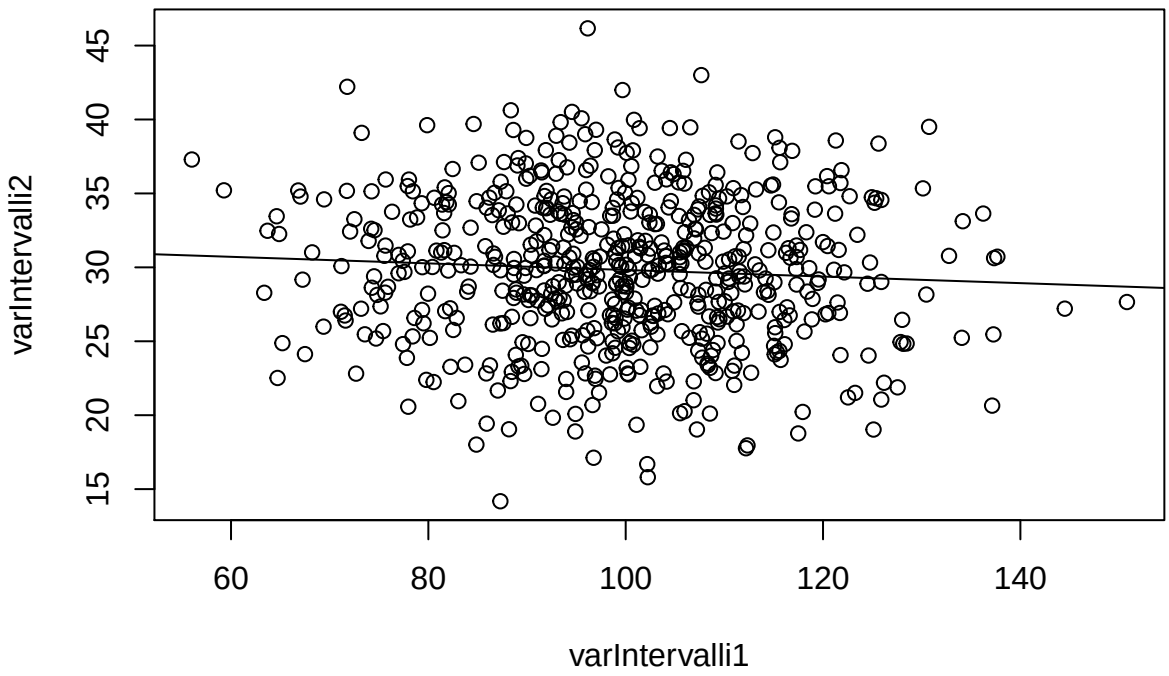
Anche in questo caso la statistica applicata è la Pearson's Chi-squared test with Yates' continuity correction, ed anche in questo caso i gradi di libertà sono 1 ((2-1)*(2-1)). Il valore della statistica in questo caso è 20.2750617 ed il p-value è 6.71e-06. In questo caso il p-value è inferiore a `0.05`, e dunque dobbiamo rifiutare l'ipotesi nulla di indipendenza fra le due variabili.

Due variabili numeriche

In caso di due variabili numeriche, si utilizza la funzione `cor.test(x,y)`.

Anche in questo caso, generiamo due variabili (numeriche) indipendenti e creiamo lo scatterplot. Valutiamo se le variabili superano il test di normalità, e calcoliamo il test di correlazione

```
varIntervalli1 <- rnorm(osservazioni, mean=100, sd=15)
varIntervalli2 <- rnorm(osservazioni, mean=30, sd=5)
plot(varIntervalli1,varIntervalli2)
lineare <- lm(varIntervalli2 ~ varIntervalli1)
abline(lineare)
```



Normalità delle variabili numeriche

Poiché le statistiche inferenziali parametriche assumono una distribuzione delle osservazioni di tipo normale, è generalmente opportuno valutare la distribuzione osservata di una variabile non soltanto attraverso metodi grafici e descrittivi, ma anche attraverso dei test di normalità.
Utilizzeremo due di questi test:

- Il test di *Kolmogorov-Smirnov* permette di confrontare due distribuzioni arbitrarie, e può essere usato per il confronto fra la distribuzione osservata e la distribuzione normale;
- Il test di normalità *Shapiro-Wilk* è finalizzato a valutare la normalità della distribuzione osservata.

Le due misure possono dare risultati differenti. Risulta pertanto necessario un processo di valutazione che tenga conto sia dei risultati dei test che dell'analisi grafica della distribuzione.

Questa regola pratica vale in ogni ambito della ricerca e dell'analisi dei dati: la metodologia ci indica delle procedure che è opportuno seguire, per minimizzare il rischio di errori che mettano a repentaglio affidabilità e validità della ricerca.

Le procedure, però, non vanno seguite pedissequamente. Conoscere i princîpi e gli assunti dell'analisi dei dati ci permette di fare delle inferenze ragionevolmente robuste anche nei casi, e sono molti, in cui non è possibile una applicazione meccanica della procedura.

Il processo

In genere si segue un processo che prevede:

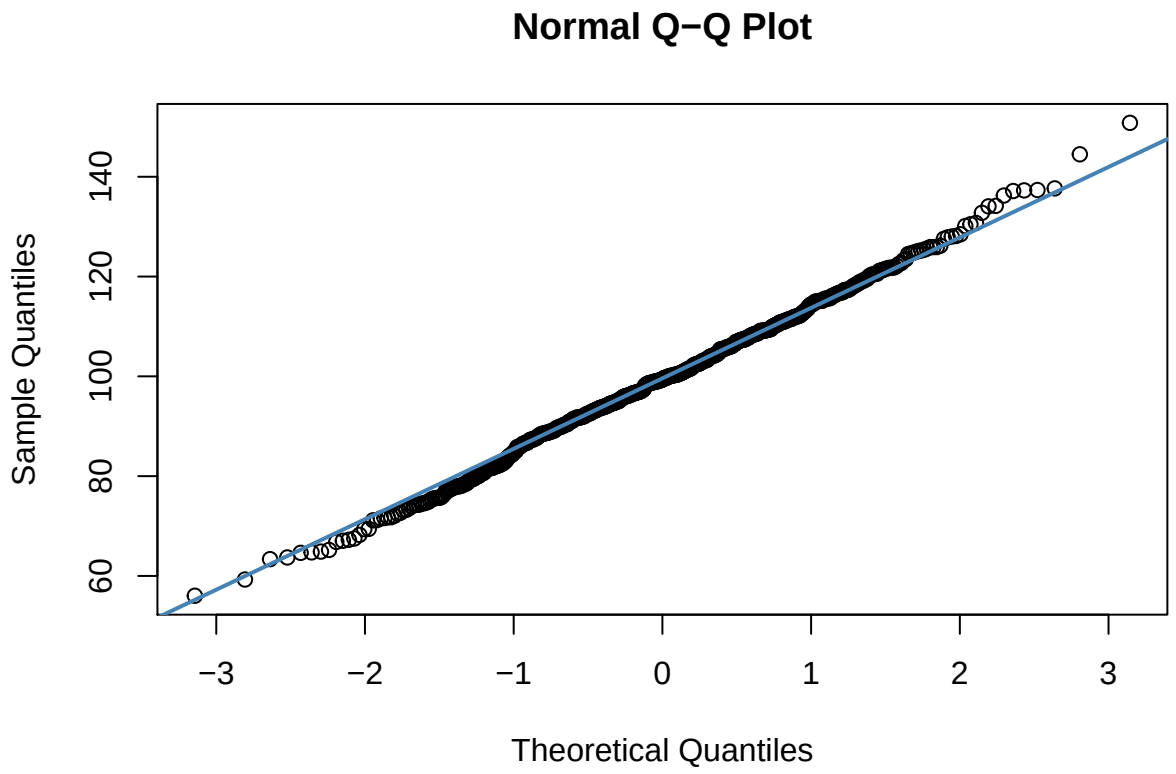
- la visualizzazione del grafico di dispersione e la linea di regressione

- la valutazione della normalità, sia graficamente che attraverso gli opportuni test
- la valutazione grafica della linearità dell'eventuale correlazione
- il test di linearità
- se gli assunti di normalità o di linearità sono violati, l'utilizzo del test inferenziale non parametrico; altrimenti, del test inferenziale parametrico

QQnorm e qqline

le funzioni `qqnorm()` e `qqline()` ci permettono di visualizzare graficamente la normalità di una distribuzione.

```
qqnorm(varIntervalli1)
# disegno una linea blu, di spessore 2
qqline(varIntervalli1, col = "steelblue", lwd = 2)
```



Se la distribuzione delle osservazioni nel grafico è lineare e sovrapponibile alla linea, si può assumere che la distribuzione sia normale.

Test di Shapiro-Wilk

Applichiamo il test di Shapiro-Wilk per valutare la normalità della distribuzione `varIntervalli1`.

```
stest_1 <- shapiro.test(varIntervalli1)
stest_1
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  varIntervalli1
## W = 0.99797, p-value = 0.6994
```

La funzione ci dice che ha applicato la statistica Shapiro-Wilk normality test. Il valore della statistica è 0.9979677 ed il p-value è 0.699. Il p-value è superiore a `0.05`, e dunque non si rifiuta l'ipotesi nulla di normalità della variabile.

Test di Kolmogorov-Smirnov

Confrontiamo la distribuzione di `varIntervalli1` con la distribuzione normale teorica, attraverso la funzione `ks.test`.

```
kstest_1 <- ks.test(varIntervalli1, "pnorm", mean = mean(varIntervalli1), sd=sd(varIntervalli1))
kstest_1
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  varIntervalli1
## D = 0.023824, p-value = 0.8853
## alternative hypothesis: two-sided
```

La funzione ci dice che ha applicato la statistica One-sample Kolmogorov-Smirnov test. Il valore della statistica è 0.023824 ed il p-value è 0.885. Il p-value è superiore a `0.05`, e dunque non si può rifiutare l'ipotesi nulla di normalità della variabile.

Esercizio

testare la normalità della variabile `varIntervalli2`.

Correlazione: test parametrico

Per valutare la correlazione fra due variabili numeriche si utilizza la funzione `cor.test()`. La funzione permette di applicare tre metodi diversi: Pearson (di default), Kendall e Spearman. In caso di non violazione degli assunti (normalità, linearità) si usa il metodo parametrico, di Pearson.

```
# non è necessario specificare method="pearson"
corTest_1 <- cor.test(varIntervalli1,varIntervalli2)
corTest_1
```

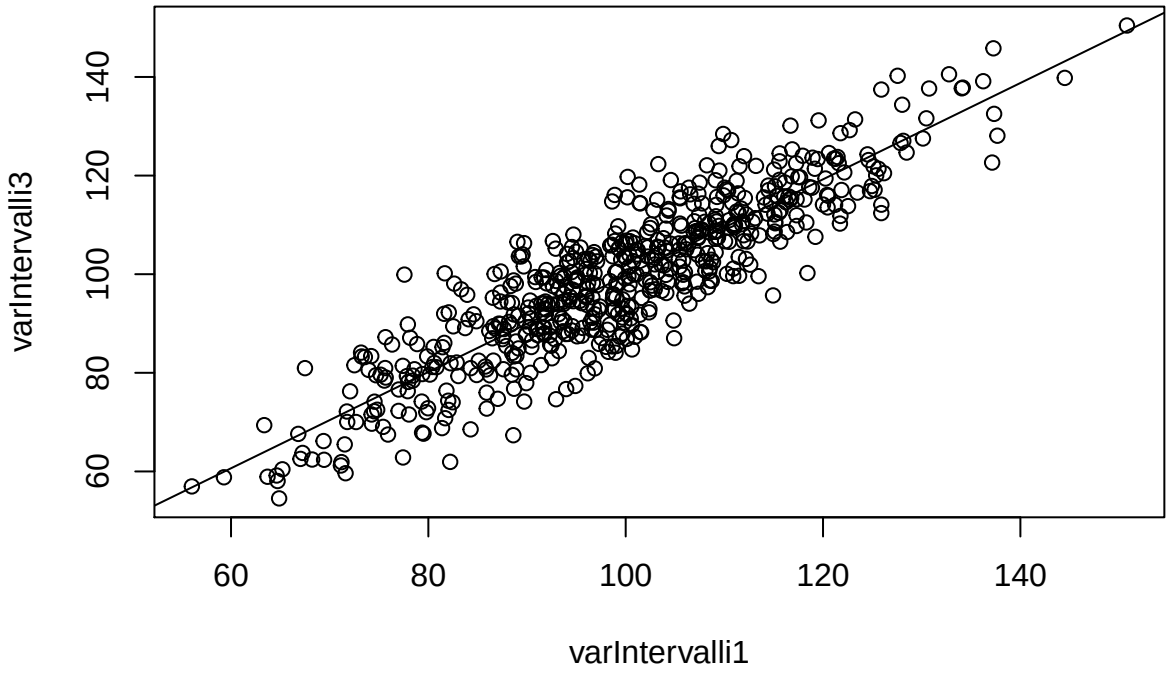
```
##
##  Pearson's product-moment correlation
##
## data:  varIntervalli1 and varIntervalli2
## t = -1.6146, df = 598, p-value = 0.1069
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.14516100  0.01423733
## sample estimates:
##         cor
## -0.06588213
```

La funzione ci dice che ha applicato la statistica Pearson's product-moment correlation. Che i gradi di libertà sono 598 (n-2). Il valore della statistica è -1.6145919 ed il p-value è 0.107. La correlazione è pari a -0.0658821.

Come prevedibile il p-value è superiore a `0.05`, e dunque non si può rifiutare l'ipotesi nulla di indipendenza fra le due variabili.

Nel prossimo esempio, generiamo `varIntervalli3`, una variabile ad intervalli *dipendente* da `varIntervalli1`

```
varIntervalli3 <- varIntervalli1 + rnorm(osservazioni, mean=0, sd=8)
plot(varIntervalli1,varIntervalli3)
corTest_2 <- cor.test(varIntervalli1,varIntervalli3)
lineare_1 <- lm(varIntervalli3 ~ varIntervalli1)
abline(lineare_1)
```




```
corTest_2
```

```
##
##  Pearson's product-moment correlation
##
## data:  varIntervalli1 and varIntervalli3
## t = 47.276, df = 598, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8700223 0.9039856
## sample estimates:
##       cor
## 0.8882113
```

In questo caso il valore della statistica è 47.2762134 ed il p-value è <2e-16. La correlazione è pari a 0.8882113. Come prevedibile il p-value è inferiore a `0.05`, e dunque non si può rifiutare l'ipotesi nulla di indipendenza fra le due variabili.

Testare la linearità

Il test parametrico assume che la correlazione sia lineare.

La non linearità può essere diagnosticata attraverso la visualizzazione del grafico di dispersione delle due variabili, il grafico di dispersione dei residui sui valori attesi, o sulla variabile X.

Da un punto di vista inferenziale, è possibile applicare la statistica *Harvey-Collier*, che valuta la linearità della correlazione: `harvtest` (va caricata la libreria `lmtest`).

```
# install.packages("lmtest")
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

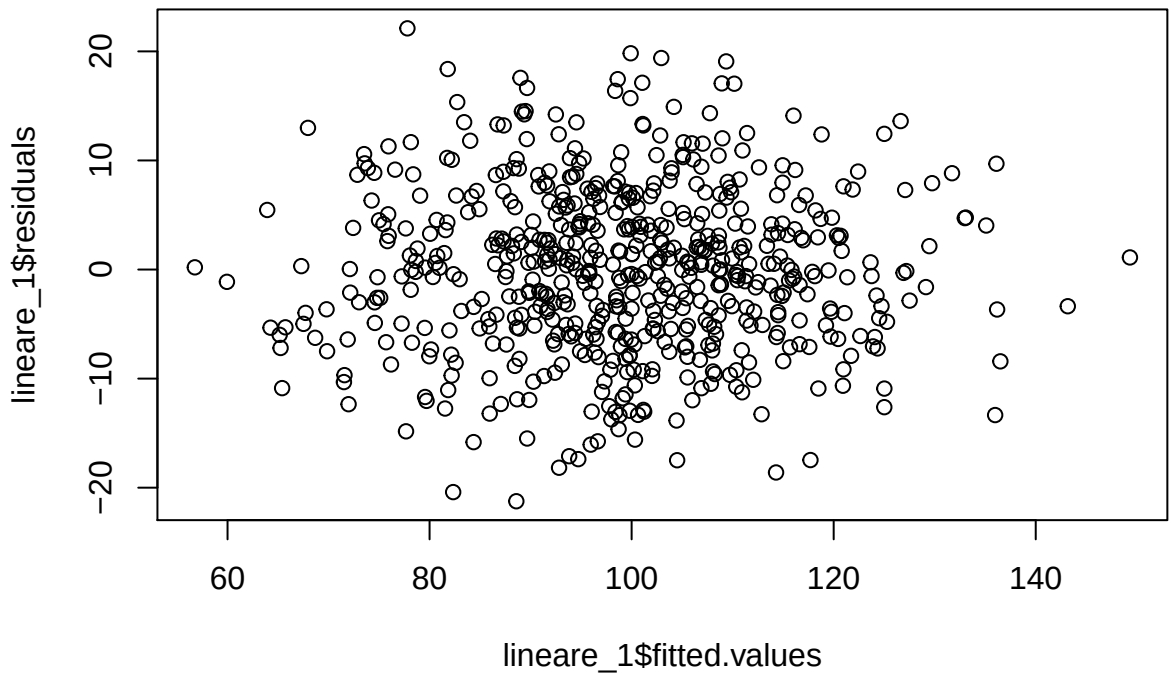
```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
htest_1 <- harvtest(varIntervalli3 ~ varIntervalli1, order.by = ~ varIntervalli1)
htest_1
```

```
##
##  Harvey-Collier test
##
## data:  varIntervalli3 ~ varIntervalli1
## HC = 1.1214, df = 597, p-value = 0.2625
```

La funzione ci dice che ha applicato la statistica Harvey-Collier test. Il valore della statistica è 1.1214448 ed il p-value è 0.263. Il p-value è superiore a `0.05`, e dunque non si può rifiutare l'ipotesi nulla di linearità della regressione.

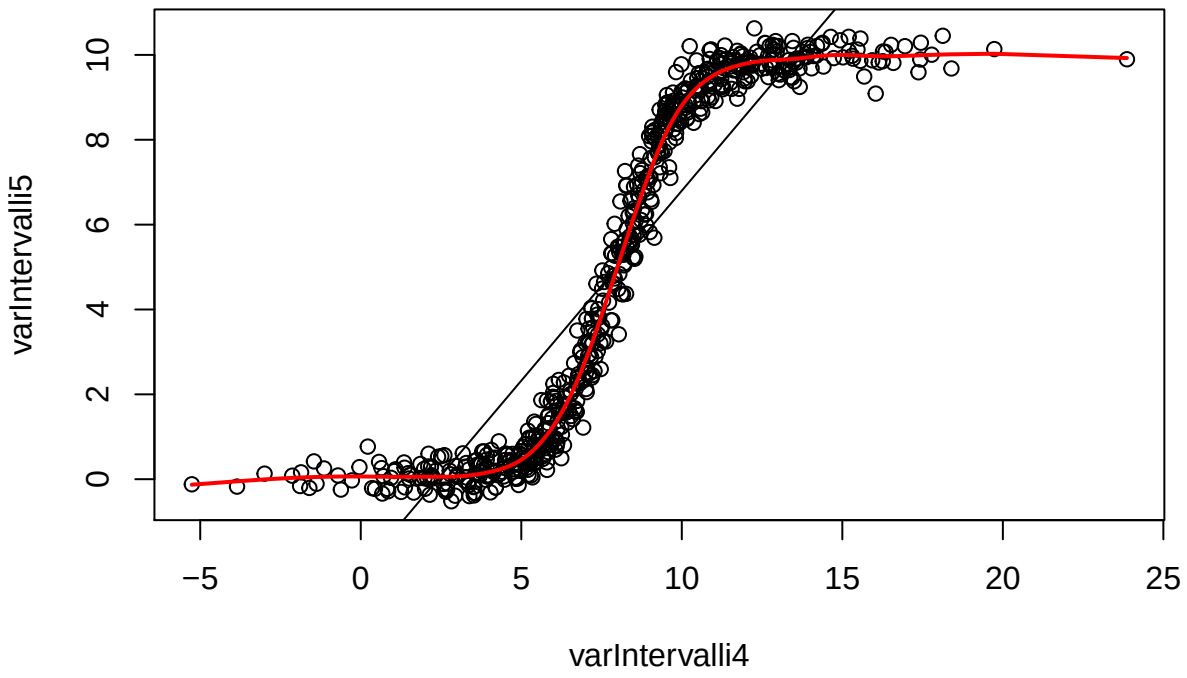
```
plot(lineare_1$fitted.values,lineare_1$residuals)
```



Esempio di correlazione non lineare

```
varIntervalli4 <- rnorm (osservazioni,0,4)
varIntervalli5 <- (1 / (1 + exp(-varIntervalli4)))*10+rnorm(osservazioni,0,0.3)
varIntervalli4 <- varIntervalli4+8+rnorm(osservazioni,0,0.3)


plot(varIntervalli4,varIntervalli5)
lineare_2 <- lm(varIntervalli5 ~ varIntervalli4)
abline(lineare_2)
# disegno una curva che segue i punti
lines(smooth.spline(varIntervalli4,varIntervalli5), col='red', lwd=2)
```



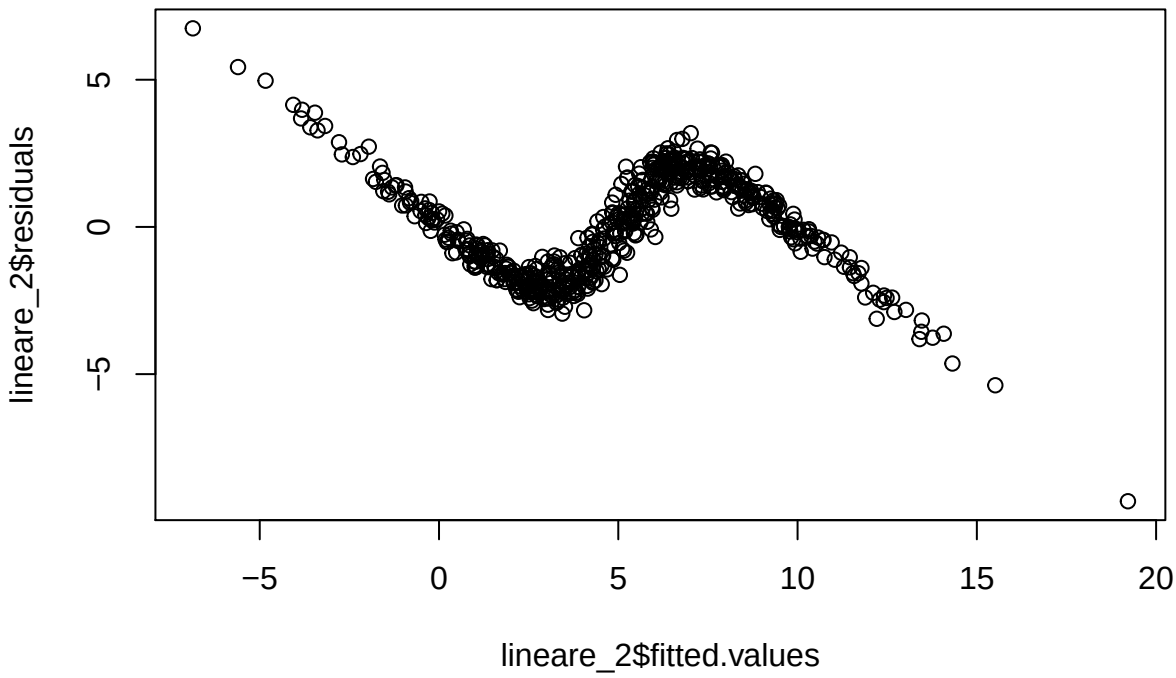

```
htest_2 <- harvtest(varIntervalli5 ~ varIntervalli4, order.by = ~ varIntervalli4)
htest_2
```

```
##
##  Harvey-Collier test
##
## data:  varIntervalli5 ~ varIntervalli4
## HC = 12.424, df = 597, p-value < 2.2e-16
```

Il valore della statistica è 12.4244172 ed il p-value è <2e-16; essendo inferiore a `0.05`, dobbiamo rifiutare l'ipotesi nulla di linearità della regressione.

La non linearità appare evidente anche dalla curva `smooth.spline` e dal grafico di dispersione dei residui.

```
plot(lineare_2$fitted.values,lineare_2$residuals)
```



Coefficiente di Spearman

Nelle circostanze in cui la relazione fra le due variabili non sia lineare, ma tenda ad essere comunque monotona, è possibile utilizzare il modello non-parametrico della correlazione: ρ di Spearman.

In questo modello, il calcolo della relazione si effettua non sui valori delle due variabili, ma sulla loro posizione ordinale.

Questa statistica, pertanto, può essere applicata anche nella circostanza in cui una o entrambe le variabili siano di tipo ordinale.

```
# specifico il metodo: Spearman
corTest_3 <- cor.test (varIntervalli4,varIntervalli5, method='spearman')
corTest_3
```

```
##
##  Spearman's rank correlation rho
##
## data:  varIntervalli4 and varIntervalli5
## S = 869980, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9758338
```

La funzione ci dice che ha applicato la statistica Spearman's rank correlation rho. Il valore della statistica è 8.69982×10^5 ed il p-value è $< 2e-16$. La correlazione è pari a 0.9758338.

Indipendente categoriale, dipendente numerica

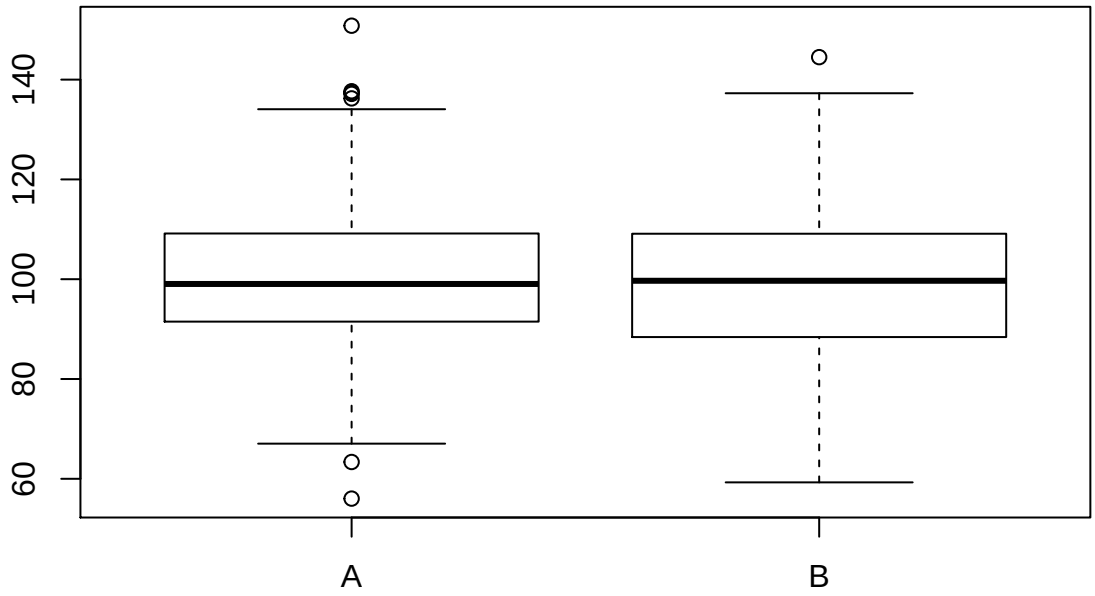
In caso di indipendente categoriale (a 2 livelli) e dipendente numerica si usa la funzione `t.test`.

Anche in questo caso, prima di optare per il test parametrico, è necessario valutare la normalità delle variabili dipendenti. In secondo luogo, è necessario valutare l'omogeneità delle varianze delle due variabili.

Nel primo esempio, applichiamo il test alla variabile dipendente `varNominale1` e alla indipendente `varIntervalli1`.

`t.test` si aspetta i due gruppi di osservazione, che possono essere estratti attraverso un filtro (`varIntervalli1[varNominale1=="A"]' e 'varIntervalli1[varNominale1=="B"]`). In alternativa si può usare la sintassi `t.test(dipendente ~ indipendente)`

```
boxplot(varIntervalli1~varNominale1)
```




```
# per comodità, creiamo le due variabili distinte
intervalli1A <- varIntervalli1[varNominale1=="A"]
intervalli1B <- varIntervalli1[varNominale1=="B"]
# calcoliamo la varianza
var_1A <- var(intervalli1A)
var_1B <- var(intervalli1B)
# calcolo il rapporto fra la maggiore e la minore
max(var_1A,var_1B)/min(var_1A,var_1B)
```

```
## [1] 1.224456
```

Testare l'omogeneità della varianza (omoschedasticità)

Uno dei test per valutare che le varianze non siano significativamente diverse è il test di Bartlett, attraverso la funzione `bartlett.test`. Se il test risulta non significativo (p-value > 0.05) non si rifiuta l'ipotesi nulla di omoschedasticità.

```
bartlett_1 <- bartlett.test(varIntervalli1 ~ varNominale1)
bartlett_1
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  varIntervalli1 by varNominale1
## Bartlett's K-squared = 3.057, df = 1, p-value = 0.08039
```

La funzione ci dice che ha applicato la statistica Bartlett test of homogeneity of variances. Il p-value è 0.0804.

Esercizio

Valutare la normalità di intervalli1A e intervalli1B

La funzione `t.test`

Se gli assunti non sono violati, si applica il Test di Student con la funzione `t.test()`.

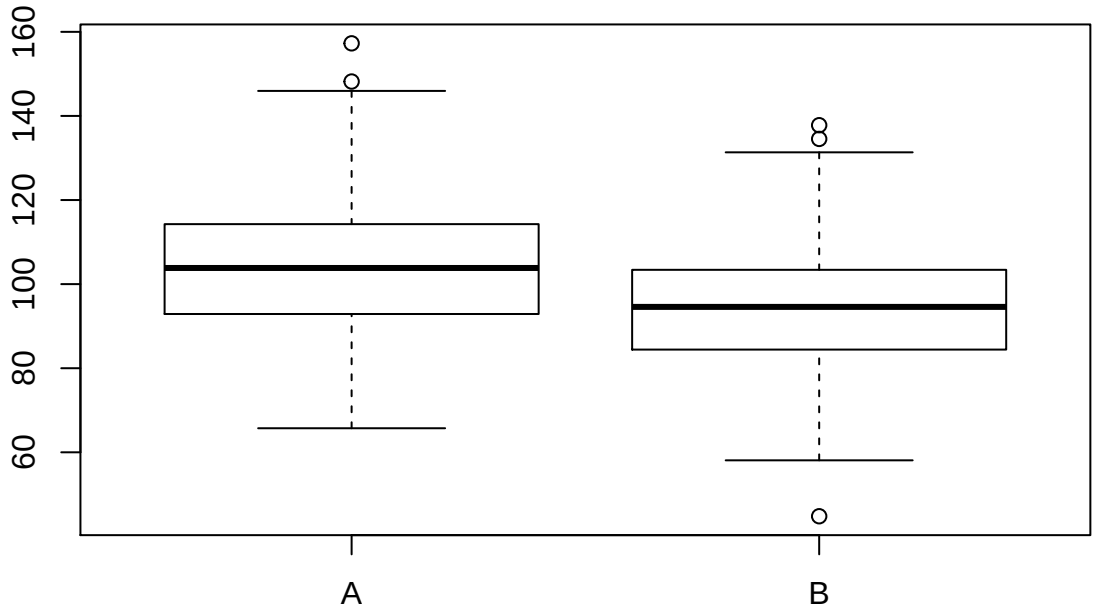
```
ttest_1 <- t.test(intervalli1A,intervalli1B)
# sintassi alternativa: t.test(varIntervalli1 ~ varNominale1)
ttest_1
```

```
##
##  Welch Two Sample t-test
##
## data:  intervalli1A and intervalli1B
## t = 1.134, df = 588.44, p-value = 0.2573
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.007534  3.760594
## sample estimates:
## mean of x mean of y
## 100.03770  98.66117
```

La funzione ci dice che ha applicato la statistica Welch Two Sample t-test. Che i gradi di libertà sono 588.4382385. Il valore della statistica è 1.1339924 ed il p-value è 0.2572595. Come prevedibile il p-value è superiore a `0.05`, e dunque non si può rifiutare l'ipotesi nulla di indipendenza fra le due variabili.

Nel prossimo esempio, generiamo `varIntervalli5`, una variabile ad intervalli *dipendente* da `varNominale1`

```
varIntervalli5 <- rnorm(osservazioni, mean=95, sd=15)+isNominaleA*10
# creiamo il boxplot
boxplot(varIntervalli5~varNominale1)
```

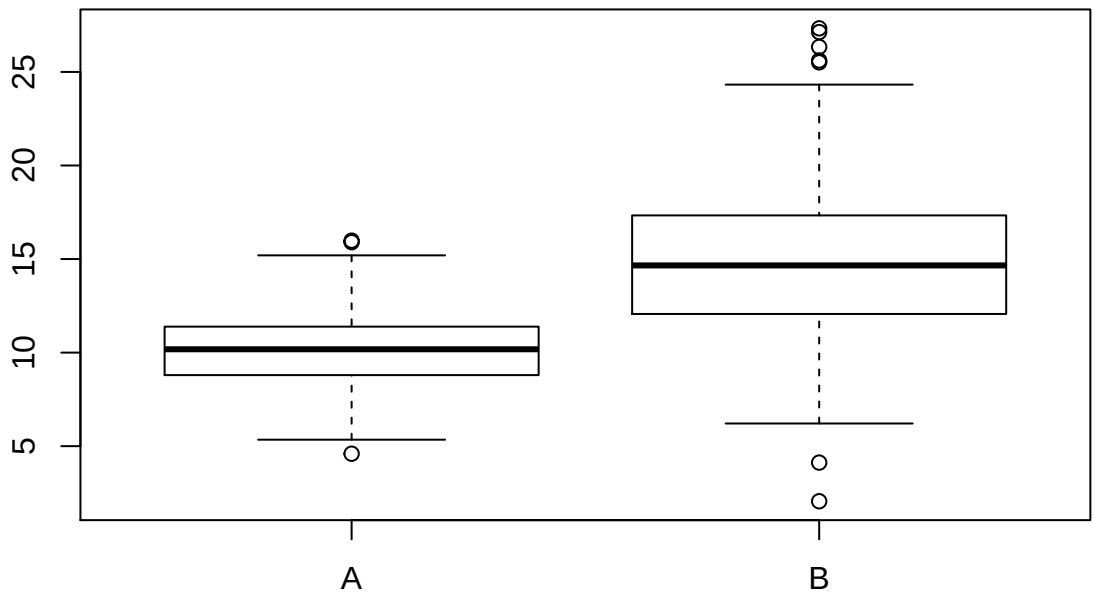



```r
# applichiamo la funzione t.test
ttest_2 <- t.test(varIntervalli5[varNominale1=="A"],varIntervalli5[varNominale1=="B"])
ttest_2
```

```
##
##  Welch Two Sample t-test
##
## data:  varIntervalli5[varNominale1 == "A"] and varIntervalli5[varNominale1 == "B"]
## t = 7.9272, df = 597.79, p-value = 1.103e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   7.554496 12.530497
## sample estimates:
## mean of x mean of y
## 103.93491  93.89242
```

In questo caso il valore della statistica è 7.9271954 ed il p-value è $1.1033242 \times 10^{-14}$, inferiore a `0.05`, e dunque si deve rifiutare l'ipotesi nulla di indipendenza fra le due variabili.

```r
varNominale4 <- c(rep("A",osservazioni/2),rep("B",osservazioni/2))
varIntervalli6 <- c(rnorm(osservazioni/2,10,2),rnorm(osservazioni/2,15,4))
var_6A <- var(varIntervalli6[1:(osservazioni/2)])
var_6B <- var(varIntervalli6[(osservazioni/2+1):osservazioni])
boxplot(varIntervalli6 ~ varNominale4)
```




```
# calcolo il rapporto fra la maggiore e la minore
max(var_6A,var_6B)/min(var_6A,var_6B)
```

```
## [1] 3.960426
```

```
bartlett_2 <- bartlett.test(varIntervalli6 ~ varNominale4)
bartlett_2
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  varIntervalli6 by varNominale4
## Bartlett's K-squared = 131.44, df = 1, p-value < 2.2e-16
```

Test non parametrico: test di Wilcoxon

In caso di violazione degli assunti di normalità o di omoschedasticità, si può usare il test non parametrico di Wilcoxon, con la funzione `wilcox.test`.

```
wilcox.test(varIntervalli6[1:(osservazioni/2)],
                        varIntervalli6[(osservazioni/2+1):osservazioni])
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  varIntervalli6[1:(osservazioni/2)] and varIntervalli6[(osservazioni/2 + 1):osservazioni]
## W = 13012, p-value < 2.2e-16
```

```
## alternative hypothesis: true location shift is not equal to 0
```

L'analisi della varianza

Nelle circostanze in cui la variabile dipendente abbia più di due livelli, oppure nel caso di più di una variabile indipendente, è necessario applicare l'Analisi della Varianza (anova).

Anche in questo caso, prima di optare per il test parametrico, è necessario valutare sia la normalità dei residui della variabile indipendente chee l'omogeneità delle varianze delle due variabili.

La funzione aov

R mette a disposizione, per il calcolo dell'analisi della varianza, la funzione $aov(y\ x)$, dove y è la variabile dipendente, numerica, e x è il fattore.

Per mostrare l'uso di aov creiamo una variabile nominale (`varFattore1`) con 3 livelli (`X`,`Y`,`Z`) e calcoliamo l'analisi della varianza utilizzando la variabile ad intervalli `varIntervalli1`. Poiché le due variabili sono del tutto indipendenti, ci aspettiamo che l'ipotesi nulla non sia falsificata.

```
varFattore1 <- factor (c(rep('X',osservazioni/3),rep('Y',osservazioni/3),rep('Z',osservazioni/3)))

aov_1 <- aov(varIntervalli1~varFattore1)
summary(aov_1)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## varFattore1   2    633   316.7   1.438  0.238
## Residuals   597 131445   220.2
```

```
m_aov_1 <- matrix(unlist(summary(aov_1)),ncol = 5)
```

Utilizzando la funzione `summary` su `aov` è possibile avere il dettaglio dei risultati dell'analisi. Nel caso di una analisi ad una via, avremo una tabella con due righe. La seconda riga calcola i gradi di libertà, la somma dei quadrati, e la media dei quadrati dei residui. La prima riga calcola i gradi di libertà, la somma dei quadrati, e la media dei quadrati del modello; inoltre, calcola la statistica $F = 1.4383669$; infine, calcola il p-value: $p = 0.24$.

Secondo esempio

Creiamo la variabile numerica `varIntervalli7`, che non è indipendente da `varFattore1`. In questo caso ci aspettiamo che l'analisi della varianza risulti significativa.

```
varIntervalli7 <- varIntervalli1
# rendiamo varIntervalli7 non indipendente da varFattore1
varIntervalli7[varFattore1=="X"]<- varIntervalli7[varFattore1=="X"]-3
varIntervalli7[varFattore1=="Y"]<- varIntervalli7[varFattore1=="Y"]+7
aov_2 <- aov(varIntervalli7~varFattore1)
summary(aov_2)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## varFattore1   2   8308    4154   18.87 1.13e-08 ***
## Residuals   597 131445     220
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
m_aov_2 <- matrix(unlist(summary(aov_2)),ncol = 5)
```

In questo caso, poiché `varIntervalli7` è stata costruita su `varFattore1`, la statistica $F = 18.8673415$ ed il p-value: $p =< 0.001$; l'ipotesi nulla va dunque rifiutata.

Verifica degli assunti

Come ogni approccio parametrico, anche l'analisi della varianza fa delle assunti:

- indipendenza delle osservazioni
- distribuzione normale dei *residui*
- omoschedasticità: la varianza dell'errore è costante
- gli errori sono fra loro indipendenti

Distribuzione dei *residui*

Si assume che i residui abbiano una distribuzione normale, con media pari a 0, e varianza costante fra i gruppi. Per testare l'ipotesi di normalità, è possibile usare il test di Shapiro-Wilk sui residui del modello dell'analisi della varianza:

```r
shapiro.test(aov_2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_2$residuals
## W = 0.99776, p-value = 0.6137
```

Per testare l'ipotesi di omoschedasticità, si può usare il test di Bartlett:

```r
bartlett.test(varIntervalli7~varFattore1)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  varIntervalli7 by varFattore1
## Bartlett's K-squared = 5.7113, df = 2, p-value = 0.05752
```

Confronti multipli

Confronti multipli ed errore

L'analisi della varianza ci permette di verificare se le differenze fra le medie di tre o più campioni sono da attribuire all'errore campionario, o se sono significative.

Una volta rifiutata l'ipotesi nulla, però, resta da determinare quali differenze sono significative. L'analisi della varianza, infatti, ci dice se vi è almeno una coppia di gruppi la cui differenza è significativa, ma non ci dice quali differenze lo sono.

Per poter determinare quali differenze sono significative, diventa necessario confrontare i gruppi a due a due.

Si potrebbe decidere di utilizzare, per confrontare a due a due i diversi gruppi, il t-test. Ma applicare ripetutamente il t-test aumenta la probabilità di incorrere in un errore del primo tipo.

Diventa dunque necessario adottare dei test di confronti multipli capaci di mantenere sotto controllo l'errore del I tipo.

Il test di Tukey

Attraverso il test di Tukey è possibile mantenere l'errore di tipo I entro un predeterminato valore di α (generalmente pari a 0.05).

Il test di Tukey permette di *correggere* il p-value in base al numero di confronti che vengono effettuati nel confronto multiplo, senza però penalizzare eccessivamente la statistica.

La funzione R TukeyHSD

La funzione di R per il calcolo del confronto con il metodo Tukey è `TukeyHSD`. La funzione si applica sul risultato della corrispondente analisi della varianza.

```
(confronti1 <- TukeyHSD(aov_2, ordered = TRUE))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##     factor levels have been ordered
##
## Fit: aov(formula = varIntervalli7 ~ varFattore1)
##
## $varFattore1
##          diff       lwr       upr     p adj
## Z-X 0.5881703 -2.898196  4.074537 0.9170580
## Y-X 8.1714236  4.685057 11.657790 0.0000002
## Y-Z 7.5832532  4.096887 11.069620 0.0000013
```

La funzione ritorna una tabella, con una riga per ogni confronto, dove vengono mostrate:

- la coppia confrontata (es, il confronto fra il gruppo Y ed il gruppo X); l'ordine è tale che il gruppo con media più alta è davanti all'altro;
- la differenza (positiva) fra i due gruppi;
- l'intervallo di confidenza della differenza; ad esempio, nel secondo confronto (Y-X), la differenza è di 8.1714236, l'intervallo di confidenza va da un minimo di 4.6850569 ad un massimo di 11.6577903. *p adj* è il p-value aggiustato, che nel confronto `C - A` è pari a <0.001.

Test non parametrico

Vi sono circostanze in cui l'analisi della varianza non può essere applicata, in quanto vengono meno alcuni assunti o condizioni:

- non si può assumere la normalità della distribuzione degli errori
- il numero di osservazioni per ogni gruppo è minore di 10
- la variabile dipendente non è ad intervalli, ma ordinale

In questi casi è possibile applicare il test non parametrico di Kruskal-Wallis

R: la funzione kruskal.test

Applichiamo il test di *Kruskal-Wallis* al nostro secondo esempio.

```
(kt_1 <- kruskal.test(varIntervalli7~varFattore1))
```

```
##
##  Kruskal-Wallis rank sum test
##
```

```
## data:  varIntervalli7 by varFattore1
## Kruskal-Wallis chi-squared = 37.7, df = 2, p-value = 6.508e-09
```

Leggere i risultati

La funzione restituisce il metodo, Kruskal-Wallis rank sum test; la statistica: 37.7004682; i gradi di libertà: 2; il p-value = <0.001.

Anova a due vie

Due variabili indipendenti

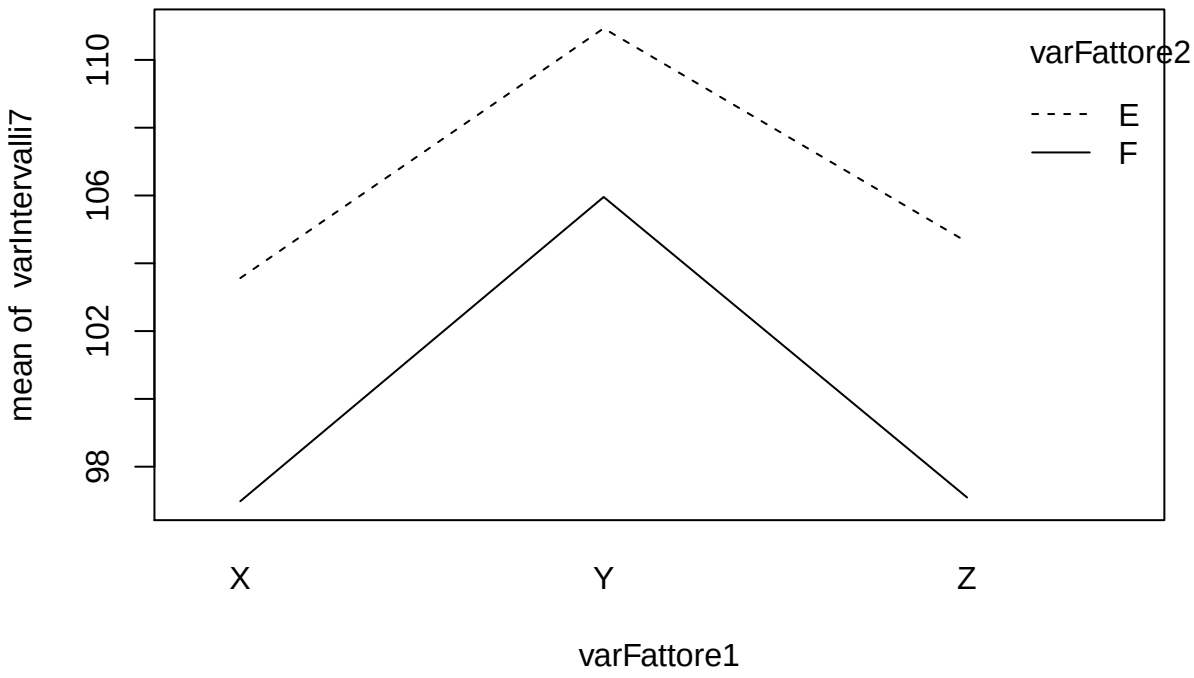
L'analisi della varianza che abbiamo introdotto, può essere estesa anche ai casi in cui le variabili indipendenti sono più di una.

Nell'analisi della varianza a due vie, si indaga la relazione fra due variabili indipendenti, entrambe categoriali, ed una variabile dipendente, quantitativa.

In questa sezione analizziamo la circostanza in cui le variabili indipendenti sono due, ma la logica rimane la stessa anche nelle circostanze in cui le variabili indipendenti sono più di due.

Creiamo la variabile `varFattore2` con due livelli. Rappresentiamo l'effetto delle due variabili indipendenti sulla variabile dipendente usando la funzione `interaction.plot`.

```
varFattore2 <- factor (rep(c(rep('E',osservazioni/6),rep('F',osservazioni/6)),3))
# modifichiamo varIntervalli7 in modo da creare un effetto di varFattore2
varIntervalli7[varFattore2=="E"]<- varIntervalli7[varFattore2=="E"]+5
interaction.plot (varFattore1,varFattore2,varIntervalli7)
```



Calcoliamo l'analisi della varianza, utilizzando la funzione `aov(dipendente~indipendente1+indipendente2+indipendente1:` dove `indipendente1:indipendente2` rappresenta l'interazione delle due variabili indipendenti.

```
aov_3 <- aov(varIntervalli7~varFattore1+varFattore2+varFattore1:varFattore2)
summary(aov_3)
```

```
##                           Df Sum Sq Mean Sq F value   Pr(>F)
## varFattore1                2   8308    4154   18.84 1.17e-08 ***
## varFattore2                1   6076    6076   27.55 2.13e-07 ***
## varFattore1:varFattore2    2    168      84    0.38    0.684
## Residuals                594 130998     221
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m_aov_3 <- matrix(unlist(summary(aov_3)),ncol = 5)
```

`summary(aov_3)` ha 4 righe: i residui (ultima riga), `varFattore1`, `varFattore2` e l'interazione. I p-value sono <0.001 per il primo fattore, <0.001 per il secondo, 0.68 per l'interazione.

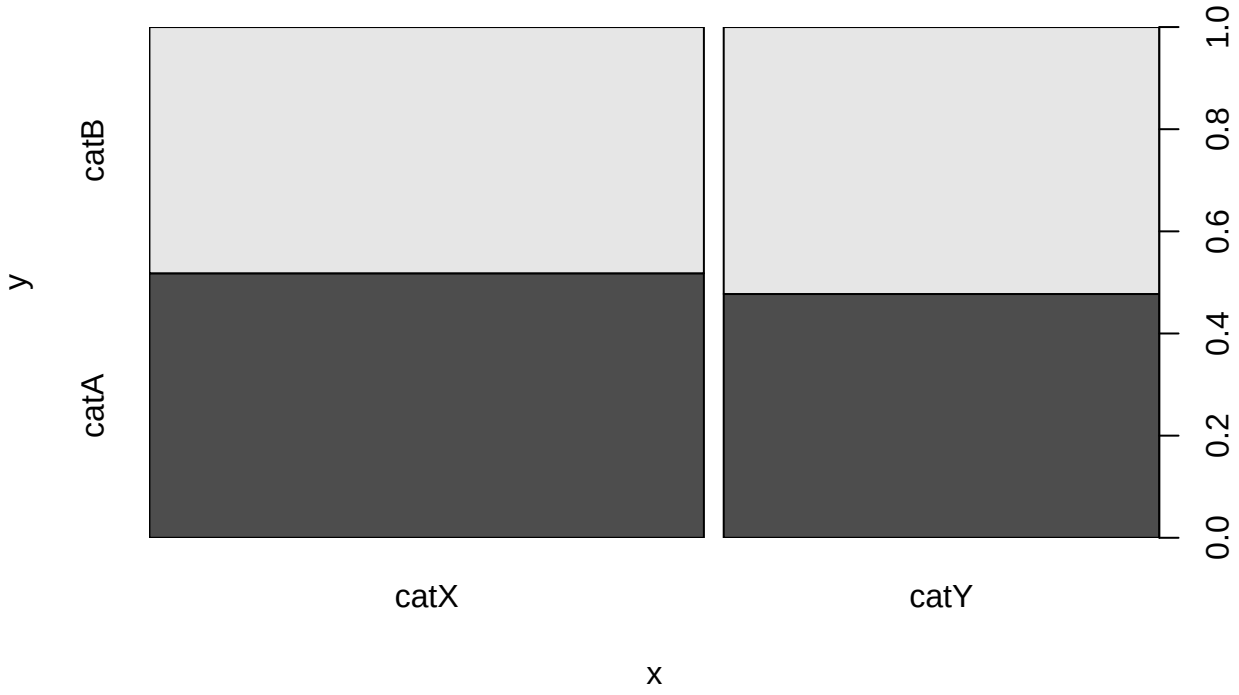
Esercizi

Caricare il file https://s3.eu-central-1.amazonaws.com/bussolon/dati/df_simulato_2.RDS nel dataframe df_simulato_2. Rappresentare graficamente e calcolare le statistiche inferenziali fra le seguenti variabili

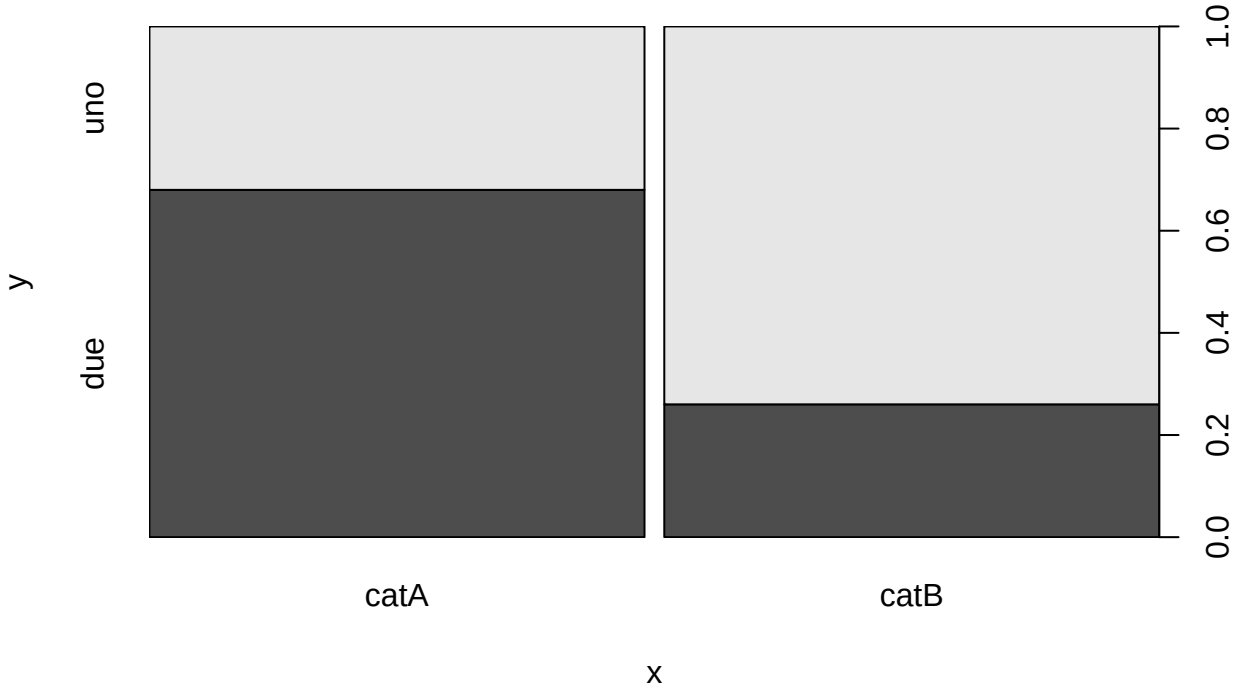
- nominale1 e nominale2
- nominale2 e nominale3
- nominale2 e intervalli1

- nominale1 e intervalli3
- intervalli1 e intervalli2
- intervalli1 e intervalli4

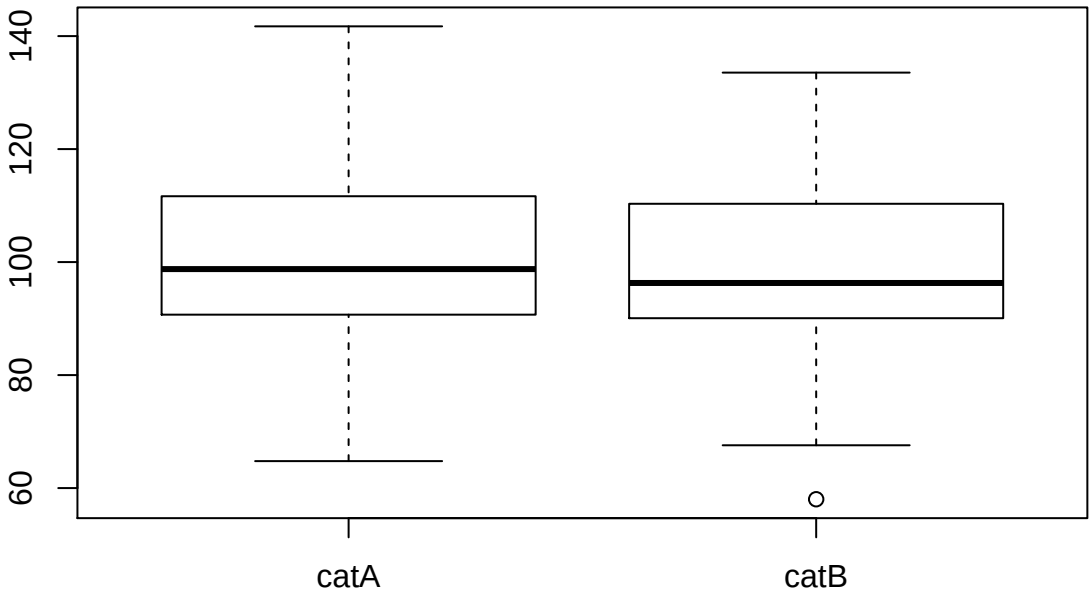
Per ogni coppia, visualizzare il grafico appropriato, e decidere se, in base all'analisi inferenziale, l'ipotesi nulla debba o meno essere rifiutata.

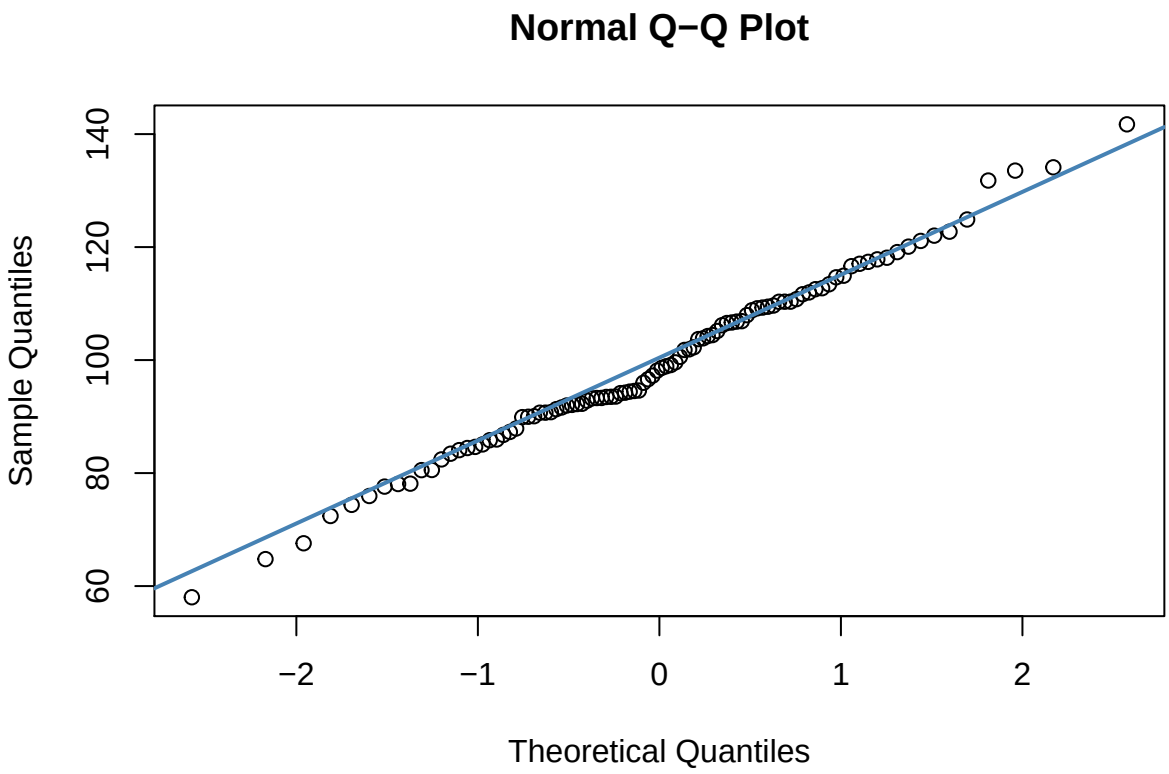



```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df_simulato_2$nominale1 and df_simulato_2$nominale2
## X-squared = 0.040584, df = 1, p-value = 0.8403
```

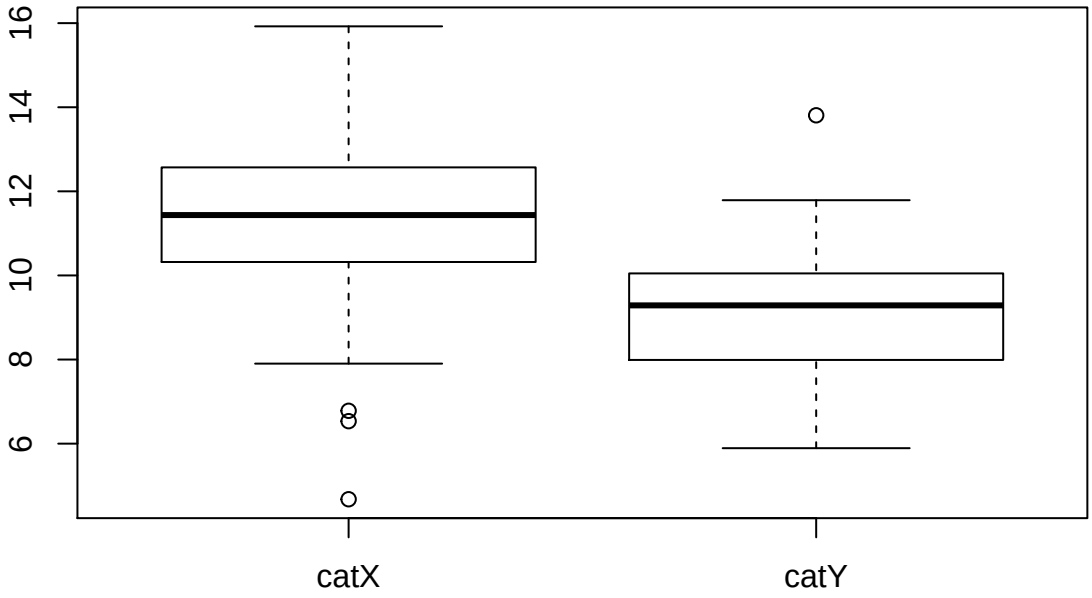


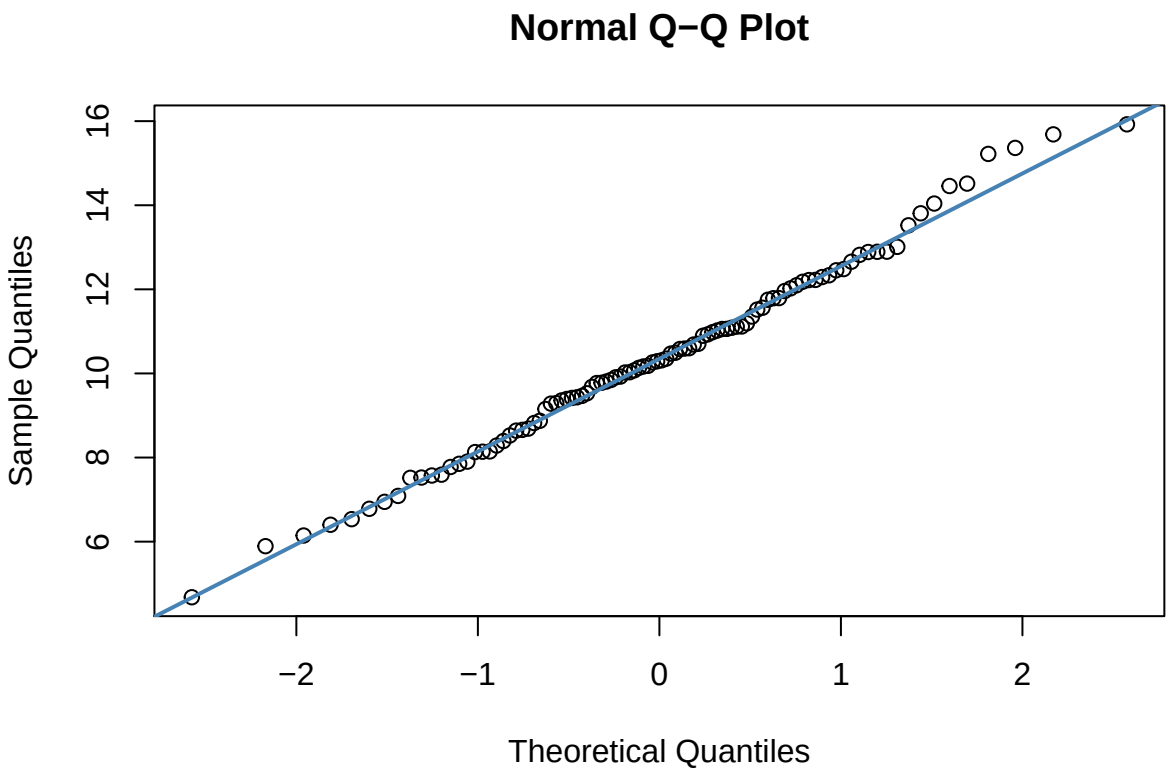

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df_simulato_2$nominale2 and df_simulato_2$nominale3
## X-squared = 16.058, df = 1, p-value = 6.144e-05
```


140
120
100
80
60
catA
catB

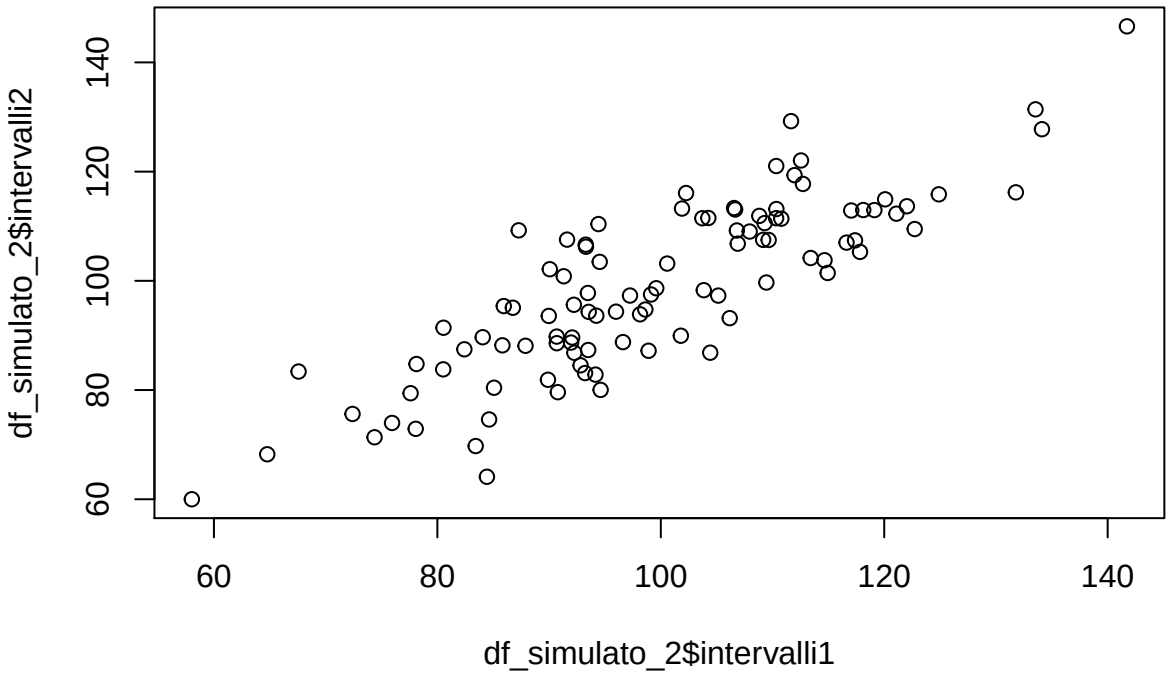



```
##
##  Welch Two Sample t-test
##
## data:  df_simulato_2$intervalli1 by df_simulato_2$nominale2
## t = 0.20222, df = 97.996, p-value = 0.8402
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.577041  6.842638
## sample estimates:
## mean in group catA mean in group catB
##           99.87326           99.24046
```

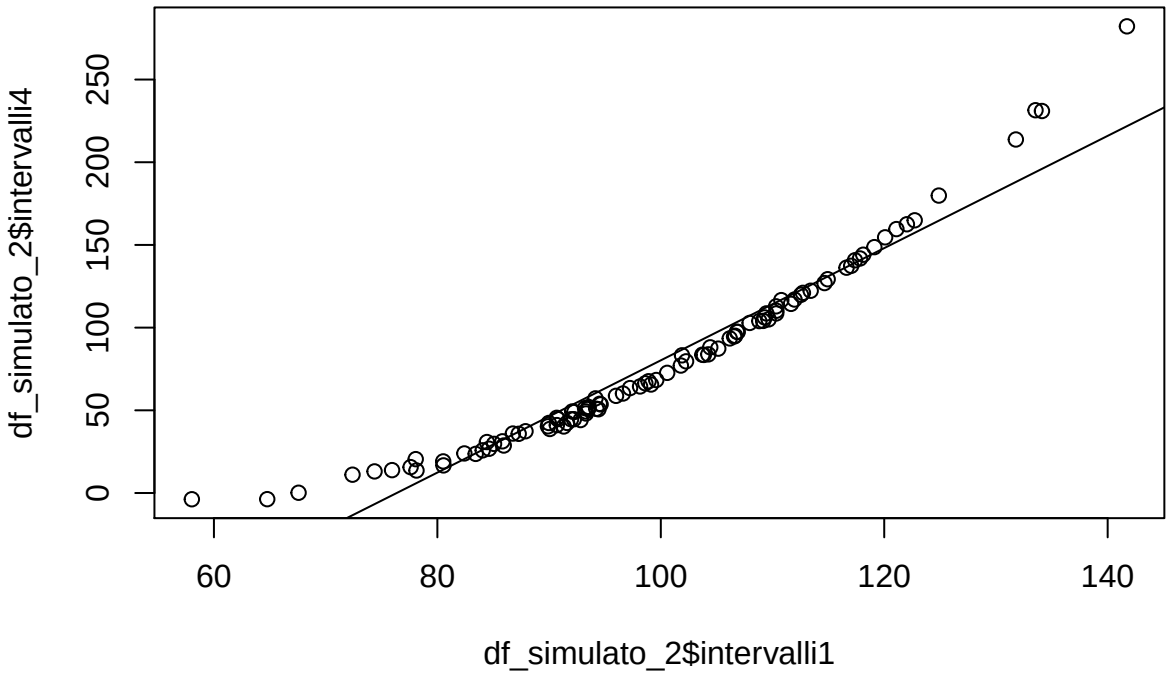

16
14
12
10
8
6
catX
catY




```
##
##  Welch Two Sample t-test
##
## data:  df_simulato_2$intervalli3 by df_simulato_2$nominale1
## t = 6.1944, df = 97.078, p-value = 1.411e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.580704 3.071191
## sample estimates:
## mean in group catX mean in group catY
##          11.410072           9.084124
```




```
##
##  Pearson's product-moment correlation
##
## data:  df_simulato_2$intervalli1 and df_simulato_2$intervalli2
## t = 15.535, df = 98, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7753392 0.8919912
## sample estimates:
##       cor
## 0.8433266
```




```
##
##  Spearman's rank correlation rho
##
## data:  df_simulato_2$intervalli1 and df_simulato_2$intervalli4
## S = 380, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9977198
```