LABORATORIO 9 - Verifica delle ipotesi

STATISTICA E LABORATORIO (CDL in INTERNET OF THINGS, BIG DATA, MACHINE LEARNING)

Anno Accademico 2023-2024

Section 1

Test sulla media e sulla varianza

Contenitori

Si desidera valutare la resistenza alla pressione interna di una certa tipologia di contenitori di vetro. In particolare si vuole verificare se la resistenza media supera 175 psi. Per tale fenomeno è ragionevole assumere il modello $N(\mu; \sigma^2)$, dove, sulla base di esperienze precedenti, si pone $\sigma^2 = 100$. Il sistema di ipotesi che si considera è $H_0 : \mu = 175$ vs $H_1 : \mu > 175$.

```
x <- c(184.35,188.69,178.88,168.98,180.29,193.4,177.03,188.9,
       132.52,177.06,180.26,187.93,200.46,195.94,189.98,166.33,
       181.14,178.41,170.06,185.64,185.62,185.47,183.9,185.4,185.27)
mean(x)
```

```
## [1] 181.2764
```

```
# valore osservato per la statistica test z sotto H_0
z_oss=(mean(x)-175)/sqrt(100/25)
z_oss
```

```
## [1] 3.1382
```

```r
# soglia di rifiuto (superiore): si rifiuta H_0
qnorm(0.05,lower.tail=FALSE)
```

```
## [1] 1.644854
```

```r
# p-value
p_value=pnorm(z_oss,lower.tail=FALSE)
p_value
```

```
## [1] 0.0008499442
```

Bilancia (varianza nota)

Si vuole verificare se una bilancia è stata calibrata di modo tale che, in media, fornisca il vero valore del peso dell'oggetto che si considera. Si considera un peso di 25 gr e si assume che una generica pesata sia descritta da una variabile casuale $N(\mu; \sigma^2)$; è nota la precisione della bilancia, che corrisponde ad una varianza $\sigma^2 = 45$. Si effettuano $n = 10$ pesate. Il sistema di ipotesi che si considera è $H_0 : \mu = 25$ vs $H_1 : \mu \neq 25$.

```
x <- c(19,23.2,12.8,18.9,30.9,16.7,31.9,27.2,22.7,16.5)
mean(x)
```

```
## [1] 21.98
```

```
z_oss=(mean(x)-25)/sqrt(45/10)
z_oss
```

```
## [1] -1.423642
```

```r
qnorm(0.025) # soglia inferiore
```

```
## [1] -1.959964
```

```r
qnorm(0.025,lower.tail=FALSE) # soglia superiore
```

```
## [1] 1.959964
```

```r
p_value=2*min(pnorm((mean(x)-25)/sqrt(45/10)),
      pnorm((mean(x)-25)/sqrt(45/10),lower.tail=FALSE)) # p-value
p_value
```

```
## [1] 0.1545502
```

Spese mediche

Sulla base dell'esperienza passata si assume che l'importo in euro delle spese mediche per famiglia in un mese segue una distribuzione $N(66; \sigma^2)$, con σ^2 ignota. Si considerano le spese mensili di un campione casuale semplice di $n = 56$ famiglie. Per valutare se la spesa media sia effettivamente diminuita si definisce il sistema di ipotesi $H_0 : \mu = 66$ vs $H1 : \mu < 66$.

```
x <- c(57,61,56,67,61,56,62,63,63,59,67,62,57,50,65,60,60,64,64,
       63,64,63,60,51,63,60,59,53,58,62,66,60,62,60,54,58,58,60,
       65,63,59,59,63,62,57,57,62,63,59,64,62,57,62,55,66,69)
mean(x)
```

```
## [1] 60.57143
```

```
var(x)
```

```
## [1] 15.23117
```

```
# valore osservato per la statistica test t sotto H_0
t_oss=(mean(x)-66)/sqrt(var(x)/56)
t_oss
```

```
## [1] -10.40909
```

```
# soglia di rifiuto (inferiore) riferita a una t(55):
# si rifiuta H_0
qt(0.05,55)
```

```
## [1] -1.673034
```

```
pt(t_oss,55)
```

```
## [1] 6.506758e-15
```

```r
# in alternativa
t.test(x,alternative="less",mu=66)
```

```
##
##  One Sample t-test
##
## data:  x
## t = -10.409, df = 55, p-value = 6.507e-15
## alternative hypothesis: true mean is less than 66
## 95 percent confidence interval:
##      -Inf 61.44395
## sample estimates:
## mean of x
##  60.57143
```

Bilancia (varianza ignota)

Si considera un problema analogo al precedente, con l'unica differenza che la varianza σ^2 è ignota. Si effettuano $n = 10$ pesate. Il sistema di ipotesi che si considera è $H_0 : \mu = 25$ vs $H_1 : \mu \neq 25$.

```
x <- c(19,23.2,12.8,18.9,30.9,16.7,31.9,27.2,22.7,16.5)
mean(x)
```

```
## [1] 21.98
```

```
mean(x^2)
```

```
## [1] 519.898
```

```
s2 <- mean(x^2)-mean(x)^2
s2
```

```
## [1] 36.7776
```

```
sc2 <- 10*s2/9 # in alterativa, var(x)
sc2
```

```
## [1] 40.864
```

```
t_oss=(mean(x)-25)/sqrt(var(x)/10)
t_oss
```

```
## [1] -1.493952
```

```
qt(0.025,9) # soglia inferiore
```

```
## [1] -2.262157
```

```
qt(0.025,9,lower.tail=FALSE) # soglia superiore: non si rifiuta H_0
```

```
## [1] 2.262157
```

```
p_value=2*min(pt(t_oss,9),
      pt(t_oss,9,lower.tail=FALSE)) #p-value
```

```
p_value
```

```
## [1] 0.169395
```

```
# in alternativa
t.test(x,alternative="two.sided",mu=25)
```

```
##
##  One Sample t-test
##
## data:  x
## t = -1.494, df = 9, p-value = 0.1694
## alternative hypothesis: true mean is not equal to 25
## 95 percent confidence interval:
##  17.40708 26.55292
## sample estimates:
## mean of x
##     21.98
```

Carta di credito

È noto che il 20% dei possessori di carta di credito la utilizza abitualmente per gli acquisti. Dopo una campagna mirata a pubblicizzare l'uso della carta di credito, in un campione di 120 individui, 36 dichiarano di utilizzarla per gli acquisti. Ci si chiede se la campagna pubblicitaria ha modificato la proporzione iniziale di utilizzatori. Il campione osservato, di dimensione $n = 120$, proviene da una popolazione $Ber(p)$, con p ignoto. Si considerano le ipotesi $H_0 : p = 0.2$ vs $H_1 : p \neq 0.2$

```
x <- 36
n <- 120
p <- x/n # stima per p
p
```

```
## [1] 0.3
```

```
se <- sqrt(p*(1-p)/n) # standard error stimato
se
```

```
## [1] 0.041833
```

```
p0 <- 0.2 # valor di p sotto H_0
se0 <- sqrt(p0*(1-p0)/n) # standard error sotto H_0
se0
```

```
## [1] 0.03651484
```

```
# valore osservato per la statistica test z sotto H_0
z_oss=(p-p0)/se0
z_oss
```

```
## [1] 2.738613
```

```
qnorm(0.025) # soglia inferiore
```

```
## [1] -1.959964
```

```
qnorm(0.025,lower.tail=FALSE) # soglia superiore: si rifiuta H_0
```

```
## [1] 1.959964
```

```r
p_value=2*min(pnorm(z_oss),
      pnorm(z_oss,lower.tail=FALSE))
p_value # p-value (approssimato)
```

```
## [1] 0.006169899
```

```r
# in alternativa
prop.test(x,n,p=0.2,correct=FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  x out of n, null probability 0.2
## X-squared = 7.5, df = 1, p-value = 0.00617
## alternative hypothesis: true p is not equal to 0.2
## 95 percent confidence interval:
##  0.2252562 0.3871514
## sample estimates:
##   p
## 0.3
```

```r
# la statistica test e' Z^2 che, sotto H_0,
# ha distribuzione approssimata chi^2(1)
(z_oss)^2
```

```
## [1] 7.5
```

Chip

La direzione di un reparto produttivo afferma che il numero medio di circuiti integrati difettosi prodotti ogni giorno è pari a 25, ma l'osservazione di un campione casuale semplice, relativo a $n = 75$ giorni, ha registrato un totale di 2110 chip difettosi. Si utilizza il modello $P(\lambda)$ per descrivere il numero giornaliero di chip difettosi e si considerano le ipotesi $H_0 : \mu = 25$ vs $H_1 : \mu > 25$.

```
x <- c(26,17,22,23,37,33,29,24,32,27,28,23,27,34,26,21,29,27,17,30,
       28,25,34,31,33,31,25,25,28,25,26,25,30,34,25,25,18,26,31,34,
       28,25,33,24,29,32,38,30,25,24,19,31,32,30,29,33,43,30,32,30,
       27,30,26,32,25,29,31,26,20,26,34,31,34,28,23)
mean(x) # stima per lambda
```

```
## [1] 28.13333
```

```
se <- sqrt(mean(x)/length(x)) # standard error stimato
se
```

```
## [1] 0.6124632
```

```r
se0 <- sqrt(25/length(x)) # standard error sotto H_0
se0
```

```
## [1] 0.5773503
```

```r
# valore osservato per la statistica test z sotto H_0
z_oss=(mean(x)-25)/se0
z_oss
```

```
## [1] 5.427093
```

```r
qnorm(0.01,lower.tail=FALSE) # soglia superiore: si rifiuta H_0
```

```
## [1] 2.326348
```

```r
pnorm(z_oss,lower.tail=FALSE) # p-value (approssimato)
```

```
## [1] 2.863972e-08
```

Lampadina

La durata in ore di un certo tipo di lampadine può essere descritta dal modello *Esp*(λ). L'esperienza passata indica che $\lambda = 0.005$, cioè che la durata media è $\mu = 200$. Considerando un campione casuale semplice di dimensione $n = 70$. Si desidera verificare le ipotesi $H_0 : \mu = 200$ vs $H_1 : \mu > 200$.

```
x <- c(490.8,642.4,808.3,79.3,108,14.5,97.7,434.8,226.3,766.5,107.3,
       280.1,227.4,16.5,634.4,223.4,439.9,342.7,373.8, 583.4,280.5,
       124.3,289.9,72.5,189.2,73.3,124.1,58.5,36.8,96.9,527.7,
       143.4,124.7,105.4,346.9,748.2,54.5,638.8, 157.1,212,
       436.9,409.9,10.3,34.9,394.3,199.6,121.1,19.3,59.4,83.1,
       65.5,126.4,465,222.7,5.6,49.6,905.4,379,25.3, 330.9,73,
       348.2,145.3,105.2,30.5,251.5,43.5,27.5,576,218.1)
mean(x) # stima della durata media
```

```
## [1] 249.5029
```

```
se <- sqrt(mean(x)^2/length(x)) # standard error stimato
```

```r
se
```

```
## [1] 29.8213
```

```r
se0 <- sqrt(200^2/length(x)) # standard error sotto H_0
se0
```

```
## [1] 23.90457
```

```r
# valore osservato per la statistica test z sotto H_0
z_oss=(mean(x)-200)/se0
z_oss
```

```
## [1] 2.070853
```

```r
qnorm(0.01,lower.tail=FALSE) #soglia superiore: si rifiuta H_0
```

```
## [1] 2.326348
```

```r
p_value=pnorm(z_oss,lower.tail=FALSE) # p-value (approssimato)
p_value
```

```
## [1] 0.01918626
```

Catalizzatore

Nell'ambito di un processo produttivo, si vuole studiare la quantità, in Kg, di sottoprodotti di una certa reazione chimica in presenza di un catalizzatore. Se il catalizzatore è attivo, la quantità di sottoprodotti è descritta da una $N(\mu; 3.83)$, con μ ignoto. Si osservano le quantità prodotte in $n = 13$ reazioni chimiche, considerando un catalizzatore esaurito. Si vuole verificare $H_0 : \sigma^2 = 3.83$ vs $H_1 : \sigma^2 \neq 3.83$.

```
x <- c(5.19,10.33,13.67,15.65,11.87,11.58,12.58,9.13,11.86,13.15,
       13.02,11.33,12.75)
mean(x)
```

```
## [1] 11.70077
```

```
mean(x^2)
```

```
## [1] 142.7751
```

```
s2 <- 12*var(x)/13
s2
```

```
## [1] 5.867146
```

```
#valore osservato per la statistica test sotto H_0
chi_oss=length(x)*s2/3.83
chi_oss
```

```
## [1] 19.91459
```

```
qchisq(0.025,12) # soglia inferiore
```

```
## [1] 4.403789
```

```
qchisq(0.025,12,lower.tail=FALSE)#soglia superiore: si rifiuta H_0
```

```
## [1] 23.33666
```

```
p_value=2*min(pchisq(chi_oss,12),
      pchisq(chi_oss,12,lower.tail=FALSE)) # p-value
p_value
```

```
## [1] 0.1374378
```

Section 2

Test sulla differenza delle medie

Inquinamento

Per confrontare, in media, l'efficacia di due diverse metodiche, A e B, per contenere l'inquinamento atmosferico si sono analizzati i fumi prodotti da una certa industria. Si è misurata la quantità di pulviscolo inquinante in g/min, con riferimento a due campioni indipendenti di fumi, ottenuti utilizzando, rispettivamente, il dispositivo anti-inquinante A e B. Siano date due popolazioni normali indipendenti con distribuzione di probabilità $N(\mu_1; \sigma_1^2)$ e $N(\mu_2; \sigma_2^2)$, con varianze note. Si estraggono due campioni casuali semplici di dimensione n_1 e n_2 dalla prima e dalla seconda popolazione, rispettivamente. Si vuole verificare l'ipotesi che le due popolazioni abbiano la stessa media e quindi si considera l'ipotesi nulla $H_0 : \mu_1 = \mu_2 \quad (\mu_1 - \mu_2 = 0)$.

```
summary(dispA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.44   14.87   15.00   15.00   15.14   15.64
```

```
summary(dispB)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.83   14.83   15.02   15.02   15.22   15.80
```

```
mean(dispA)
```

```
## [1] 15.00289
```

```
mean(dispB)
```

```
## [1] 15.02171
```

```r
var(dispA)
```

```
## [1] 0.04587104
```

```r
var(dispB)
```

```
## [1] 0.09956351
```

```r
sp2 <- ((length(dispA)-1)*var(dispA)+(length(dispB)-1)*var(dispB))/
    (length(dispA)+length(dispB)-2)
sp2
```

```
## [1] 0.07271727
```

```r
#valore osservato per la statistica test t a due campioni sotto H_0
t_oss <- (mean(dispA)-mean(dispB))/
    sqrt(sp2*(1/length(dispA)+1/length(dispB)))
t_oss
```

```
## [1] -0.6620645
```

```
qt(0.025,length(dispA)+length(dispB)-2) # soglia inferiore
```

```
## [1] -1.966613
```

```
# soglia superiore: non si rifiuta H_0
qt(0.025,length(dispA)+length(dispB)-2,lower.tail=FALSE)
```

```
## [1] 1.966613
```

```
# p-value
2*min(pt(t_oss,length(dispA)+length(dispB)-2),
      pt(t_oss,length(dispA)+length(dispB)-2,lower.tail=FALSE))
```

```
## [1] 0.5083558
```

```r
# in alternativa
# ipotizzando varianze uguali
t.test(dispA,dispB,alternative="two.sided",var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  dispA and dispB
## t = -0.66206, df = 358, p-value = 0.5084
## alternative hypothesis: true difference in means is not equal to
## 95 percent confidence interval:
##  -0.07471964  0.03708153
## sample estimates:
## mean of x mean of y
##  15.00289  15.02171
```

```r
# ipotizzando varianze diverse (test di Welch)
t.test(dispA,dispB,alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  dispA and dispB
## t = -0.66206, df = 315.06, p-value = 0.5084
## alternative hypothesis: true difference in means is not equal to
## 95 percent confidence interval:
##  -0.07474549  0.03710738
## sample estimates:
## mean of x mean of y
##  15.00289  15.02171
```

```r
# valore osservato per la statistica test F per il confronto tra
# le varianze sotto H_0
f_oss <- var(dispA)/var(dispB)
f_oss
```

```
## [1] 0.4607214
```

```r
qf(0.005,length(dispA)-1,length(dispB)-1) # soglia inferiore
```

```
## [1] 0.6792371
```

```r
# soglia superiore: si rifiuta H_0
qf(0.005,length(dispA)-1,length(dispB)-1,lower.tail=FALSE)
```

```
## [1] 1.47224
```

```
# p-value
2*min(pf(f_oss,length(dispA)-1,length(dispB)-1),
      pf(f_oss,length(dispA)-1,length(dispB)-1,lower.tail=FALSE))
```

```
## [1] 3.148703e-07
```

```
# in alternativa
var.test(dispA,dispB,alternative="two.sided")
```

```
##
##  F test to compare two variances
##
## data:  dispA and dispB
## F = 0.46072, num df = 179, denom df = 179, p-value = 3.149e-07
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3433908 0.6181418
## sample estimates:
## ratio of variances
##          0.4607214
```

Vernici

Si desidera confrontare il tempo di asciugatura in ore di due diverse vernici. Si considerano due campioni casuali semplici di $n_1 = 17$ e $n_2 = 18$ verniciature, rispettivamente. Per verificare le ipotesi $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$.

```
x1 <- c(7.29,6.64,8.31,4.51,5.23,5.71,6.37,0.96,7.03,9.16,6.98,
        8.4,4.01,5.31,3.93,8.04,7.5)
x2 <- c(9.03,6.72,12.21,8.43,8.71,6.99,7.71,9.52,8.04,9.32,5.53,
        12.92,10.47,5.62,5.56,9.24,9.47,8.36)
mean(x1)
```

```
## [1] 6.198824
```

```
mean(x2)
```

```
## [1] 8.547222
```

```
var(x1)
```

```
## [1] 4.254561
```

```
var(x2)
```

```
## [1] 4.278539
```

```
# test F per il confronto tra varianze; si accetta H_0
var.test(x1,x2,alternative="two.sided")
```

```
##
##  F test to compare two variances
##
## data:  x1 and x2
## F = 0.9944, num df = 16, denom df = 17, p-value = 0.995
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3687364 2.7226537
## sample estimates:
## ratio of variances
```

```
sp2 <- ((length(x1)-1)*var(x1)+(length(x2)-1)*var(x2))/
  (length(x1)+length(x2)-2)
sp2
```

```
## [1] 4.266913
```

```
#valore osservato per la statistica test t a due campioni sotto H_0
t_oss <- (mean(x1)-mean(x2))/sqrt(sp2*(1/length(x1)+1/length(x2)))
t_oss
```

```
## [1] -3.361566
```

```
qt(0.005,length(x1)+length(x2)-2) # soglia inferiore
```

```
## [1] -2.733277
```

```
# soglia superiore: si rifiuta H_0
qt(0.005,length(x1)+length(x2)-2,lower.tail=FALSE)
```

```
## [1] 2.733277
```

```r
# p-value
2*min(pt(t_oss,length(x1)+length(x2)-2),
      pt(t_oss,length(x1)+length(x2)-2,lower.tail=FALSE))
```

```
## [1] 0.001972087
```

```
# in alternativa
# ipotizzando varianze uguali
t.test(x1,x2,alternative="two.sided",var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  x1 and x2
## t = -3.3616, df = 33, p-value = 0.001972
## alternative hypothesis: true difference in means is not equal to
## 95 percent confidence interval:
##  -3.769716 -0.927081
## sample estimates:
## mean of x mean of y
##  6.198824  8.547222
```

```r
# ipotizzando varianze diverse (test di Welch)
t.test(x1,x2,alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  x1 and x2
## t = -3.3618, df = 32.896, p-value = 0.001975
## alternative hypothesis: true difference in means is not equal to
## 95 percent confidence interval:
##  -3.7697688 -0.9270286
## sample estimates:
## mean of x mean of y
##  6.198824  8.547222
```

Pressione

Si desidera valutare l'efficacia di un farmaco per abbassare la pressione sistolica. A tale scopo, si misura le pressione ad un campione casuale semplice di $n = 15$ pazienti, prima e dopo il trattamento con il farmaco in esame. Si vuole verificare l'ipotesi $H_0 : \mu_d = 0$ vs $H_1 : \mu_d > 0$.

```
prima <- c(146,145,136,145,146,147,144,147,146,145,145,143,133,
           146,146)
dopo <- c(130,118,120,121,119,141,118,120,119,120,123,120,119,
          129,121)
mean(prima)
```

```
## [1] 144
```

```
mean(dopo)
```

```
## [1] 122.5333
```

```
var(prima)
```

```
## [1] 16.28571
```

```
var(dopo)
```

```
## [1] 39.12381
```

```
diff <- prima-dopo
diff
```

```
##  [1] 16 27 16 24 27  6 26 27 27 25 22 23 14 17 25
```

```
mean(diff)
```

```
## [1] 21.46667
```

```
var(diff)
```

```
## [1] 39.40952
```

```r
# statistica test t per dati appaiati sotto H_0
t_oss=mean(diff)/sqrt(var(diff)/length(diff))
t_oss
```

```
## [1] 13.24371
```

```r
# soglia superiore: si rifiuta H_0
qt(0.01,length(diff)-1,lower.tail=FALSE)
```

```
## [1] 2.624494
```

```r
pt(t_oss,length(diff)-1,
   lower.tail=FALSE) # p-value
```

```
## [1] 1.305968e-09
```

```r
# in alternativa
# test t per dati appaiati
t.test(prima,dopo,alternative="greater",paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  prima and dopo
## t = 13.244, df = 14, p-value = 1.306e-09
## alternative hypothesis: true difference in means is greater than
## 95 percent confidence interval:
##  18.61177      Inf
## sample estimates:
## mean of the differences
##                21.46667
```

Section 3

Test su proporzioni

Cerotti

Ad un gruppo di $n_1 = 224$ fumatori che desiderano smettere di fumare viene prescritto il cerotto alla nicotina, mentre ad altri $n_2 = 245$ fumatori oltre al cerotto si prescrive anche un leggero antidepressivo. Dopo sei mesi, 40 soggetti del primo gruppo e 87 del secondo avevano smesso di fumare. Ci si chiede se è utile associare cerotto ad antidepressivo. Si considera il test per il confronto tra proporzioni per verificare l'ipotesi $H_0 : p_1 = p_2$ vs $H_1 : p_1 \neq p_2$.

```
x1 <- 40
n1 <- 224
x2 <- 87
n2 <- 245
p1 <- x1/n1
p1
```

```
## [1] 0.1785714
```

```
p2 <- x2/n2
p2
```

```
## [1] 0.355102
```

```
p <- (x1+x2)/(n1+n2)
p
```

```
## [1] 0.2707889
```

```
# valore osservato per la statistica test z sotto H_0
z_oss=(p1-p2)/sqrt(p*(1-p)*(1/n1+1/n2))
z_oss
```

```
## [1] -4.297329
```

```
qnorm(0.025) # soglia inferiore
```

```
## [1] -1.959964
```

```
qnorm(0.025,lower.tail=FALSE) # soglia superiore: si rifiuta H_0
```

```
## [1] 1.959964
```

```
# p-value (approssimato)
2*min(pnorm((p1-p2)/sqrt(p*(1-p)*(1/n1+1/n2))),
      pnorm((p1-p2)/sqrt(p*(1-p)*(1/n1+1/n2)),lower.tail=FALSE))
```

```
## [1] 1.728687e-05
```

```r
# in alternativa
prop.test(c(x1,x2),c(n1,n2),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(x1, x2) out of c(n1, n2)
## X-squared = 18.467, df = 1, p-value = 1.729e-05
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.25467274 -0.09838848
## sample estimates:
##    prop 1    prop 2
## 0.1785714 0.3551020
```

```
# la statistica test e' Z^2 che, sotto H_0,
# ha distribuzione approssimata chi^2(1)
(z_oss)^2
```

```
## [1] 18.46704
```

Aspirina

Alcuni anni or sono venne condotto uno studio per valutare l'effetto dell'aspirina per prevenire l'infarto. Si costituì un gruppo di trattamento (somministrazione giornaliera di aspirina) con $n_1 = 11037$ individui e un gruppo di controllo (somministrazione giornaliera di placebo) con $n_2 = 11034$ individui. Dopo un certo periodo si osservarono 139 infartuati nel primo gruppo e 239 nel secondo. Sulla base di questi dati preliminari si interruppe lo studio che doveva durare cinque anni. Si vuole verificare $H_0 : p_1 = p_2$ vs $H_1 : p_1 < p_2$.

```r
x1 <- 139
n1 <- 11037
x2 <- 239
n2 <- 11034
p1 <- x1/n1
p1
```

```
## [1] 0.012594
```

```r
p2 <- x2/n2
p2
```

```
## [1] 0.02166032
```

```r
p <- (x1+x2)/(n1+n2)
p
```

```
## [1] 0.01712655
```

```r
# valore osservato per la statistica test z sotto H_0
z_oss=(p1-p2)/sqrt(p*(1-p)*(1/n1+1/n2))
z_oss
```

```
## [1] -5.190729
```

```
p_value=pnorm(z_oss) # p-value (approssimato)
p_value
```

```
## [1] 1.04736e-07
```

```r
# in alternativa
prop.test(c(x1,x2),c(n1,n2),alternative="less",correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(x1, x2) out of c(n1, n2)
## X-squared = 26.944, df = 1, p-value = 1.047e-07
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.000000000 -0.006195011
## sample estimates:
##     prop 1     prop 2
## 0.01259400 0.02166032
```

Farmaco

Per sperimentare l'efficacia di un principio attivo si considerano tre farmaci: A (senza principio attivo), B (con una certa quantità di principio attivo) e C (con una quantità doppia di principio attivo). Si effettua la sperimentazione su $r = 3$ campioni indipendenti con $n_1 = 126$, $n_2 = 85$ e $n_3 = 119$ pazienti ai quali si somministrano, rispettivamente, i farmaci A, B e C. Con riferimento ad ogni paziente si osserva la variabile multinomiale "Esito finale", con $k = 3$ possibili categorie: y_1 =peggiorato, y_2 =stazionario, y_3 =migliorato. Ci si chiede i tre campioni provengono dalla stessa popolazione multinomiale, cioè se le proporzioni delle tre tipologie di esiti sono ragionevolmente le stesse nei tre casi.

```
x <- matrix(c(10,104,12,5,70,10,10,65,44),3,3,byrow=TRUE)
x
```

```
##      [,1] [,2] [,3]
## [1,]   10  104   12
## [2,]    5   70   10
## [3,]   10   65   44
```

```r
chisq.test(x,correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 35.537, df = 4, p-value = 3.604e-07
```

```r
qchisq(0.01,4,lower.tail=FALSE) # valore critico (soglia superiore).
```

```
## [1] 13.2767
```

```r
pchisq(35.537,4,lower.tail=FALSE) # P-value (approssimato)
```

```
## [1] 3.603032e-07
```

Fisioterapia

Fisioterapia. Si vuole valutare il livello di soddisfazione dei clienti di un centro di fisioterapia, considerando separatamente coloro che svolgono attività ginnica e terapie. Si considerano $r = 2$ campioni indipendenti con $n_1 = 227$ clienti che svolgono attività ginnica e $n_2 = 262$ che seguono terapie, rispettivamente. Ci si chiede se la proporzione di soddisfatti p_1 tra coloro che svolgono attività ginnica e la proporzione di soddisfatti p_2 tra coloro che seguono terapie è la stessa.

```
x <- matrix(c(163,64,154,108),2,2,byrow=TRUE)
x
```

```
##      [,1] [,2]
## [1,]  163   64
## [2,]  154  108
```

```
p1 <- x[1,1]/sum(x[1,])
p1
```

```
## [1] 0.7180617
```

```r
p2 <- x[2,1]/sum(x[2,])
p2
```

```
## [1] 0.5877863
```

```r
p <- sum(x[,1])/sum(x)
p
```

```
## [1] 0.6482618
```

```r
# valore osservato per la statistica test z sotto H_0
z_oss <- (p1-p2)/sqrt(p*(1-p)*(1/sum(x[1,])+1/sum(x[2,])))
z_oss
```

```
## [1] 3.008754
```

```r
qnorm(0.025) # soglia inferiore
```

```
## [1] -1.959964
```

```
qnorm(0.025,lower.tail=FALSE) # soglia superiore: si rifiuta H_0
```

```
## [1] 1.959964
```

```
2*min(pnorm(z_oss),pnorm(z_oss,lower.tail=FALSE)) # p-value (appros
```

```
## [1] 0.002623218
```

```
# in alternativa
prop.test(x,correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  x
## X-squared = 9.0526, df = 1, p-value = 0.002623
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.04673792 0.21381291
## sample estimates:
##    prop 1    prop 2
## 0.7180617 0.5877863
```

```
# stesso risultato con prop.test(c(x[1,1],x[2,1]),c(sum(x[1,]),
# sum(x[2,])),correct=FALSE)
```

```
# la statistica test e' Z^2 che, sotto H_0, ha distribuzione
# approssimata chi^2(1)
(z_oss)^2
```

```
## [1] 9.052598
```

```
# la funzione prop.test() si utilizza quando si hanno due campioni
# multinomiali (in particolare bernoulliani/binomiali)
# con due o piu' campioni bisogna utilizzare la funzione
# chisq.test()
# chisq.test(x,correct=FALSE) fornisce lo stesso risultato
```

Perni

In uno stabilimento industriale ci sono tre macchinari per la produzione di perni di acciaio, che devono rispettare le specifiche di diametro. Per valutarne l'efficacia del procedimento produttivo si considera un campione di $n = 400$ perni per i quali si osservano i valori della variabile X =“Macchinario”, con $r = 3$ categorie (macchinario1, macchinario2, macchinario3), e della variabile Y =“Diametro”, con $k = 3$ categorie (fine, ok, spesso). Si vuole verificare se c'è dipendenza tra il tipo di macchinario utilizzato e la qualità dei perni: H_0 : X e Y indipendenti vs H_1 : X e Y dipendenti.

```
perni <- rbind(cbind(rep("M1",10), rep("Fine",10)),
               cbind(rep("M1",102), rep("Ok",102)),
               cbind(rep("M1",8), rep("Spesso",8)),
               cbind(rep("M2",34), rep("Fine",34)),
               cbind(rep("M2",161), rep("Ok",161)),
               cbind(rep("M2",5), rep("Spesso",5)),
               cbind(rep("M3",10), rep("Fine",10)),
               cbind(rep("M3",60), rep("Ok",60)),
               cbind(rep("M3",10), rep("Spesso",10)))
```

```r
perni <- as.data.frame(perni) # matrice trasformata in data frame
colnames(perni) <- c("Macchinario","Diametro") # nomi delle colonne
str(perni)
```

```
## 'data.frame':    400 obs. of  2 variables:
##  $ Macchinario: chr  "M1" "M1" "M1" "M1" ...
##  $ Diametro   : chr  "Fine" "Fine" "Fine" "Fine" ...
```

```r
# tabella frequenze assolute
tab <- table(perni$Macchinario, perni$Diametro)
tab
```

```
##
##        Fine  Ok Spesso
##     M1   10 102      8
##     M2   34 161      5
##     M3   10  60     10
```

```r
xtot <- margin.table(tab,1) # distribuzione marginale di X
xtot
```

```
##
##  M1  M2  M3
## 120 200  80
```

```r
ytot <- margin.table(tab,2) # distribuzione marginale di Y
ytot
```

```
##
##   Fine     Ok Spesso
##     54    323     23
```

```r
xtot <- as.matrix(xtot) # vettore interpretato come matrice
ytot <- as.matrix(ytot) # vettore interpretato come matrice
```

```r
# tabella di contingenza in caso di indipendenza
tab_ind <- xtot%*%t(ytot)/sum(xtot)
tab_ind
```

```
##    Fine    Ok Spesso
## M1 16.2  96.9    6.9
## M2 27.0 161.5   11.5
## M3 10.8  64.6    4.6
```

```r
# statistica test chi-quadrato
chisq_oss <- sum((tab-tab_ind)^2/tab_ind)
chisq_oss
```

```
## [1] 15.03284
```

```r
# valore critico (soglia superiore): si rifiuta H_0
qchisq(0.01,4,lower.tail=FALSE)
```

```
## [1] 13.2767
```

```
# P-value (approssimato)
pchisq(15.03,4,lower.tail=FALSE)
```

```
## [1] 0.004639398
```

```
# in alternativa si puo' usare il comando
chisq.test(tab,correct=FALSE)
```

```
## Warning in chisq.test(tab, correct = FALSE): Chi-squared approxin
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 15.033, df = 4, p-value = 0.004634
```

Nascite

Con riferimento alla variabile casuale categoriale "Giorno settimanale di nascita", si dispone di $n = 140$ osservazioni campionarie. Si vuole valutare se le nascite siamo distribuite in modo uniforme tra i giorni della settimana.

```
xo <- c(13,23,24,20,27,18,15) # frequenze osservate
xo
```

```
## [1] 13 23 24 20 27 18 15
```

```
prob <- rep(1/7,7) # probabilita' sotto H_0
xa <- 140*prob # frequenze attese sotto H_0
xa
```

```
## [1] 20 20 20 20 20 20 20
```

```r
chisq_oss <- sum((xo-xa)^2/xa) # statistica test chi-quadrato
chisq_oss
```

```
## [1] 7.6
```

```r
# valore critico (soglia superiore): si accetta H_0
qchisq(0.05,6,lower.tail=FALSE)
```

```
## [1] 12.59159
```

```r
pchisq(chisq_oss,6,lower.tail=FALSE) # P-value (approssimato)
```

```
## [1] 0.2688967
```

```r
# in alternativa si puo' usare il comando
chisq.test(xo,p=prob,correct=FALSE)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  xo
## X-squared = 7.6, df = 6, p-value = 0.2689
```

Section 4

Dataset reali

Esercizio 7.2 (Ieva et al., 2016)

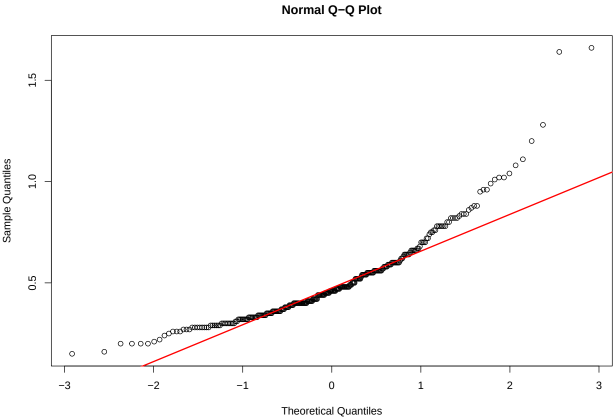
Si consideri il file “siero.txt”. I dati consistono nella misura della concentrazione di trigliceridi nel siero [mmol/l] contenuto nel sangue del cordone ombelicale di un campione di 282 soggetti. Si vuole effettuare un test d’ipotesi sulla media della variabile Siero, in cui l’ipotesi nulla è mu $= 0.4$ contro l’ipotesi alternativa mu > 0.4.

```
setwd("C:/Users/mamel/Desktop/IBML/LAB_STAT9")
siero = read.table(file="siero.txt", header=T)
head(siero)
```

```
##   Siero
## 1  0.15
## 2  0.16
## 3  0.20
## 4  0.20
## 5  0.20
## 6  0.20
```

```
attach(siero)
```

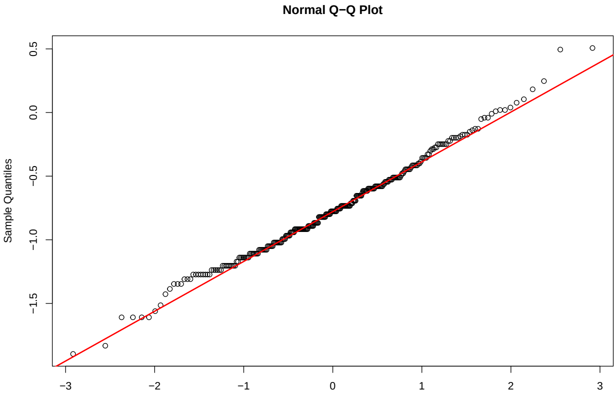
```
qqnorm(Siero)
qqline(Siero,col='red',lwd=2)
```



```
mean(Siero)
```

```
## [1] 0.5058865
```

```
dati = log(Siero)
qqnorm(dati)
qqline(dati,col='red',lwd=2)
```



```
alpha= 0.01
mu0 = log(0.4)
mu0
```

```
## [1] -0.9162907
```

```
t.test(dati, alternative ='greater', mu=mu0, conf.level= 1-alpha)
```

```
##
##  One Sample t-test
##
## data:  dati
## t = 6.6336, df = 281, p-value = 8.39e-11
## alternative hypothesis: true mean is greater than -0.9162907
## 99 percent confidence interval:
##  -0.8158291        Inf
## sample estimates:
##  mean of x
## -0.7610893
```

Esercizio 7.3 (Ieva et al., 2016)

Carichiamo i dati contenuti nel file penicillina.txt: rappresentano la temperatura corporea (espressa in gradi Fahrenheit) di 30 pazienti ricoverati per meningite, che sono stati trattati con un'ampia dose di penicillina. Se dopo 3 giorni è stata osservata una diminuzione di temperatura, il trattamento è stato considerato un successo (S).

Vogliamo stabilire se i dati ci permettono di affermare, con una confidenza del 95%, che il trattamento ha successo in più del 60% dei casi.

In particolare, in base alla richiesta dell'esercizio vorremmo verificare:

```
H_0: p = 0.6 vs H_1: p > 0.6
```

```r
dati <- read.table('penicillina.txt', header=T)
head(dati)
```

```
##   Paziente  Temp Rid_Temp
## 1        1 104.0        S
## 2        2 104.8        S
## 3        3 101.6        F
## 4        4 108.0        S
## 5        5 103.8        S
## 6        6 100.8        S
```

```r
dim(dati)
```

```
## [1] 30  3
```

```r
n <- dim(dati)[1]
```

```
names(dati)
```

```
## [1] "Paziente" "Temp"     "Rid_Temp"
```

```
attach(dati)
p.0 <- 0.6
alpha <- 0.05
z.alpha <- qnorm(1-alpha)
p.camp <- length(which(Rid_Temp=='S'))/n
p.camp
```

```
## [1] 0.5666667
```

```
z_oss <- (p.camp - p.0)/ sqrt(p.0*(1 - p.0)/n)
z_oss
```

```
## [1] -0.372678
```

```
z_oss > z.alpha
```

```
## [1] FALSE
```

```
p_value <- pnorm(z_oss,lower.tail=FALSE)
p_value
```

```
## [1] 0.6453059
```

```
detach(dati)
```