

METODI STATISTICI PER LA RICERCA SOCIALE

CONFRONTO TRA DUE GRUPPI IN

Alessandra Mattei

Dipartimento di Statistica, Informatica, Applicazioni (DiSIA)
Università degli Studi di Firenze
mattei@disia.unifi.it

LM-88 SOCIOLOGIA E RICERCA SOCIALE

Inferenza per la differenza tra le medie

Popolazioni Normali indipendenti con varianze note

- Il volontariato e l'amicizia: Esiste una differenza fra gli studenti universitari che svolgono attività di volontariato e quelli che si dedicano solo allo studio?
- Studenti che hanno partecipato ad alcune attività di volontariato (popolazione 2) *versus* studenti che non hanno mai svolto attività di volontariato (popolazione 1)
- Variabile di interesse: Misura dell'importanza data all'amicizia (su scala da 0 a 100)
- Varianze: $\sigma_1^2 = 36$ e $\sigma_2^2 = 25$
- Sintesi dei dati campionari

Gruppo	n_j	Media
Nessuna attività	27	66.8
Volontariato	35	72.2

Inferenza per la differenza tra le medie

Popolazioni Normali indipendenti con varianze note

```
> ##Dati
> sigma2.1 <- 36; sigma2.2 <- 25
> n1 <- 27
> bary1 <- 66.8
>
> n2 <- 35
> bary2 <- 72.2
>
> ##Stima della differenza tra le medie
> dm <- bary2-bary1
> dm
[1] 5.4
>
> ##Errore standard dello stimatore della differenza
> es.dm <- sqrt(sigma2.1/n1 + sigma2.2/n2)
> es.dm
[1] 1.43095
```

Inferenza per la differenza tra le medie

Popolazioni Normali indipendenti con varianze note

```
> ##Intervallo di confidenza  
> ##Livello di confidenza 1-alpha = 90%  
> alpha <- 0.10  
> zz <- qnorm(1-alpha/2)  
> zz  
[1] 1.644854  
>  
> c(dm - zz*es.dm, dm + zz*es.dm)  
[1] 3.046296 7.753704
```

Inferenza per la differenza tra le medie

Popolazioni Normali indipendenti con varianze note

```
> ##Test per la differenza tra le medie
> ##H0:  $\mu_2 - \mu_1 = 0$  versus  $H_a: \mu_2 - \mu_1 \neq 0$ 
>
> ##Regione critica al livello del 10%
> alpha <- 0.10
> zz <- qnorm(1-alpha/2)
> zz
[1] 1.644854
> ##Regione critica alpha=0.1:  $Z \leq -1.645$  o  $Z \geq 1.645$ 
>
> ##Valore osservato della statistica test
> zoss <- dm/es.dm
> zoss
[1] 3.773716
>
> ##P-valore
> 2*{1-pnorm(abs(zoss))}
[1] 0.000160834
```

Inferenza per la differenza tra le medie

Il miele per curare la tosse

- L'uso del miele per curare la tosse notturna e migliorare la qualità del sonno in bambini con infezioni del tratto respiratorio superiore
- Bambini non trattati (Popolazione 1) versus bambini a cui viene somministrata una singola dose di miele di grano saraceno la sera prima di dormire (Popolazione 2)
- Variabili
 - Gender = Genere (0 = Femmina; 1 = Maschio)
 - Age = Età (in anni)
 - csp = Frequenza della tosse il giorno prima del trattamento
 - treatment = Trattamento (0 = No trattamento; 1 = Miele)
 - csa = Frequenza della tosse dopo il trattamento

Inferenza per la differenza tra le medie

Popolazioni Normali indipendenti con varianze ignote ma uguali

```
> rm(list=ls())
>
> ##Dati
> load("Dati/Honey.RData")
> ls()
[1] "honey"
>
> str(honey)
'data.frame': 72 obs. of 5 variables:
'data.frame': 72 obs. of 5 variables:
$ Gender : num 1 0 0 1 1 1 1 1 0 0 ...
$ Age : num 3.14 2.66 5.18 3.02 5.19 ...
$ cfp : num 4 4 4 4 3 5 4 4 2 4 ...
$ treatment: num 0 1 1 0 1 0 0 0 1 1 ...
$ cfa : num 4 4 2 4 1 3 2 3 0 2 ...
$ cba.cat : num 3 4 2 3 1 2 1 3 1 2 ...
> attach(honey)
```

Inferenza per la differenza tra le medie

Popolazioni Normali indipendenti con varianze ignote ma uguali

```
> ##Stima della differenza tra le medie
> dm<-mean(cfa[treatment==1])- mean(cfa[treatment==0])
> dm
[1] -0.6965251
>
> ##Varianze entro i gruppi
> s2.1 <- var(cfa[treatment==1])
> s2.1
[1] 2.398319
>
> s2.0 <- var(cfa[treatment==0])
> s2.0
[1] 2.602102
```


Inferenza per la differenza tra le medie

Popolazioni Normali indipendenti con varianze ignote ma uguali

```
> ##Varianza pooled
> n0<- sum(treatment==0)
> n0
[1] 37
>
> n1<- sum(treatment==1)
> n1
[1] 35
>
> s2.p <- ({n0-1}*s2.0 + {n1-1}*s2.1)/(n0+n1-2)
> s2.p
[1] 2.503122
>
> ##Errore standard dello stimatore della differenza tra medie
> es.dm <- sqrt(s2.p*{1/n0 + 1/n1})
> es.dm
[1] 0.3730546
```

Inferenza per la differenza tra le medie

Popolazioni Normali indipendenti con varianze ignote ma uguali

```
> ##Test e intervalli di confidenza
> ##IC al livello 1-alpha=0.95
> ##H0:  $\mu_1 - \mu_0 = 0$  versus  $H_a: \mu_1 - \mu_0 \neq 0$ 
>
> treat<- 1-treatment
> t.test(cfa~treat, data=honey, var.equal = TRUE)
```

Two Sample t-test

```
data:  cfa by treat
t = -1.8671, df = 70, p-value = 0.06608
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.44055897  0.04750878
sample estimates:
mean in group 0 mean in group 1
2.114286         2.810811
```

Inferenza per la differenza tra le medie

Popolazioni Normali indipendenti con varianze ignote ma uguali

```
> ##IC al livello 1-alpha=0.90  
> ##H0:  $\mu_1 - \mu_0 = 0$  versus  $H_a: \mu_1 - \mu_0 \neq 0$   
> t.test(cfa~treat, data=honey, conf.level = 0.90, var.equal = TRUE)
```

Two Sample t-test

```
data:  cfa by treat  
t = -1.8671, df = 70, p-value = 0.06608  
alternative hypothesis: true difference in means is not equal to 0  
90 percent confidence interval:  
-1.31837516 -0.07467503  
sample estimates:  
mean in group 0 mean in group 1  
2.114286          2.810811
```

Inferenza per la differenza tra le medie

Popolazioni Normali indipendenti con varianze ignote ma uguali

```
> ##H0:  $\mu_1 - \mu_0 = 0$  versus  $H_a: \mu_1 - \mu_0 < 0$   
> t.test(cfa~treat, data=honey, var.equal = TRUE, alternative = "less")
```

Two Sample t-test

```
data:  cfa by treat  
t = -1.8671, df = 70, p-value = 0.03304  
alternative hypothesis: true difference in means is less than 0  
95 percent confidence interval:  
-Inf -0.07467503  
sample estimates:  
mean in group 0 mean in group 1  
2.114286          2.810811
```

Inferenza per la differenza tra le medie

Popolazioni Normali indipendenti con varianze non note

```
> ###Popolazioni Normali indipendenti con varianze non note
>
> ##IC al livello 1-alpha=0.90
> ##H0:  $\mu_1 - \mu_0 = 0$  versus  $H_a: \mu_1 - \mu_0 \neq 0$ 
> t.test(cfa~treat, data=honey, conf.level = 0.90, var.equal = FALSE)
```

Welch Two Sample t-test

```
data:  cfa by treat
t = -1.8692, df = 69.983, p-value = 0.06578
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
-1.3176636 -0.0753866
sample estimates:
mean in group 0 mean in group 1
2.114286        2.810811
```

Inferenza per la differenza tra le medie

Popolazioni Normali indipendenti con varianze non note

```
> ##H0: mu.1-mu.0 = 0 versus Ha: mu.1-mu.0 < 0  
> t.test(cfa~treat, data=honey, var.equal = FALSE, alternative = "less")
```

Welch Two Sample t-test

```
data:  cfa by treat  
t = -1.8692, df = 69.983, p-value = 0.03289  
alternative hypothesis: true difference in means is less than 0  
95 percent confidence interval:  
-Inf -0.0753866  
sample estimates:  
mean in group 0 mean in group 1  
2.114286        2.810811
```

Inferenza per la differenza tra le medie

Popolazioni qualsiasi indipendenti – Campioni di grandi dimensioni

```
> ###Popolazioni qualsiasi indipendenti - Campioni di grandi dimensioni
> ##Stima della differenza tra le medie
> dm<-mean(cfa[treatment==1])- mean(cfa[treatment==0])
> dm
[1] -0.6965251
>
> ##Varianze entro i gruppi
> s2.1 <- var(cfa[treatment==1])
> s2.1
[1] 2.398319
>
> s2.0 <- var(cfa[treatment==0])
> s2.0
[1] 2.602102
>
> ##Errore standard dello stimatore della differenza tra medie
> es.dm <- sqrt(s2.0/n0 + s2.1/n1)
> es.dm
[1] 0.3726265
```

Inferenza per la differenza tra le medie

Popolazioni qualsiasi indipendenti – Campioni di grandi dimensioni

```
> ##IC al livello 1-alpha=0.90
> alpha<-0.1
> dm-qnrm(1-alpha/2)*es.dm
[1] -1.309441
> dm+qnrm(1-alpha/2)*es.dm
[1] -0.08360908
>
> ##H0:  $\mu_1 - \mu_0 = 0$  versus  $H_a: \mu_1 - \mu_0 > 0$ 
> ##Regione critica al livello alpha=0.1
> qnorm(1-alpha)   ###RC_0.1 =  $Z < -1.28$ 
[1] 1.281552
>
> zoss<- dm/es.dm
> zoss
[1] -1.869231
>
> ##p-valore
> pnorm(zoss)
[1] 0.03079532
```


Inferenza per la differenza tra le medie

Popolazioni Normali dipendenti (dati appaiati)

```
> ## Test per la differenza tra le medie:
> ## Popolazioni Normali dipendenti (dati appaiati)
> ## Focus sui Bambini trattati con il miele
> sum(treatment==1)
[1] 35
> mean(cfa[treatment==1])
[1] 2.114286
> mean(cfp[treatment==1])
[1] 4
>
> ##Stima della differenza tra le medie (post-pre)
> cf.diff <- cfa[treatment==1]-cfp[treatment==1]
> mean(cf.diff)
[1] -1.885714
>
> var(cf.diff)
[1] 1.339496
>
> ##Errore standard della differenza tra le medie (post-pre)
> sd(cf.diff)/length(cf.diff)
[1] 0.0330676
```

Test per la differenza tra le medie

Popolazioni Normali dipendenti (dati appaiati)

```
> ##IC al livello 1-alpha=0.95  
> ##H0:  $\mu_1 - \mu_0 = 0$  versus  $H_a: \mu_1 - \mu_0 \neq 0$   
>  
> t.test(cfa[treatment==1], cfp[treatment==1], paired = TRUE)
```

Paired t-test

```
data: cfa[treatment == 1] and cfp[treatment == 1]  
t = -9.6392, df = 34, p-value = 2.964e-11  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
-2.283283 -1.488145  
sample estimates:  
mean of the differences  
-1.885714
```

Test per la differenza tra le medie

Popolazioni Normali dipendenti (dati appaiati)

```
> ##H0: mu.1-mu.0 = 0 versus Ha: mu.1-mu.0 < 0
>
> t.test(cfa[treatment==1], cfp[treatment==1], paired = TRUE,
+  alternative="less")
```

Paired t-test

```
data:  cfa[treatment == 1] and cfp[treatment == 1]
t = -9.6392, df = 34, p-value = 1.482e-11
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
-Inf -1.554918
sample estimates:
mean of the differences
-1.885714
>
>detach(honey)
```

Inferenza per la differenza tra proporzioni

Corso sull'auto-palpazione del seno e pratica dell'auto-palpazione del seno

- Relazione tra la frequenza di un corso di formazione sull'auto-palpazione del seno e la pratica dell'auto-palpazione del seno
- Donne che frequentano il corso (Popolazione 2) versus donne che non frequentano il corso (Popolazione 2)
- Variabili
 - course = Frequenza del corso (0 = No; 1 = Si)
 - bse = Pratica dell'auto-palpazione del seno a un anno dal corso

Inferenza per la differenza tra proporzioni

Popolazioni bernoulliane indipendenti - Campioni di grandi dimensioni

```
> rm(list=ls())
>
> ##Dati
> Bse <- read.table("Dati/BSE.txt", header=T)
>
> str(Bse)
'data.frame': 800 obs. of 2 variables:
 $ course: int  1 0 0 1 1 1 1 1 0 1 ...
 $ bse    : int  1 1 0 1 1 1 1 1 1 1 ...
>
> attach(Bse)
>
> table(course,bse)

      bse
course 0  1
0      87 335
1      47 331
```

Inferenza per la differenza tra proporzioni

Popolazioni bernoulliane indipendenti - Campioni di grandi dimensioni

```
> ##Sintesi campionarie
> n0<- sum(1-course)
> n0
[1] 422
>
> bse0 <-sum(bse[course==0])
> bse0
[1] 335
>
> n1<- sum(course)
> n1
[1] 378
>
> bse1<-sum(bse[course==1])
> bse1
[1] 331
```

Inferenza per la differenza tra proporzioni

Popolazioni bernoulliane indipendenti - Campioni di grandi dimensioni

```
> ##Proporzioni campionarie
> p0<- bse0/n0
> p0
[1] 0.7938389
>
> p1<- bse1/n1
> p1
[1] 0.8756614
>
> ##Stima della differenza tra proporzioni
> dp <- p1-p0
> dp
[1] 0.08182251
```

Inferenza per la differenza tra proporzioni

Popolazioni bernoulliane indipendenti - Campioni di grandi dimensioni

```
> ##ES dello stimatore della differenza tra proporzioni
> es.dp<-sqrt({p0*(1-p0)/n0}+{p1*(1-p1)/n1})
> es.dp
[1] 0.02599722
>
> ##ES dello stimatore della differenza tra proporzioni
> ##sotto H0: p1=p0
> p.pooled<- (bse0+bse1)/{n1+n0}
> p.pooled
[1] 0.8325
> es.dp.h0<-sqrt(p.pooled*(1-p.pooled)*{1/n0 + 1/n1})
> es.dp.h0
[1] 0.02644493
```


Inferenza per la differenza tra proporzioni

Popolazioni bernoulliane indipendenti - Campioni di grandi dimensioni

```
> ##IC al livello 1-alpha=0.97  
> ##H0:  $p_1 - p_0 = 0$  versus  $H_a: p_1 - p_0 \neq 0$   
> prop.test(c(bse1, bse0), c(n1,n0), conf.level = 0.97, correct=FALSE)
```

2-sample test for equality of proportions without continuity
correction

```
data:  c(bse1, bse0) out of c(n1, n0)  
X-squared = 9.5733, df = 1, p-value = 0.001974  
alternative hypothesis: two.sided  
97 percent confidence interval:  
0.0254062 0.1382388  
sample estimates:  
prop 1      prop 2  
0.8756614 0.7938389
```

Inferenza per la differenza tra proporzioni

Popolazioni bernoulliane indipendenti - Campioni di grandi dimensioni

```
> ##H0:  $p_1 - p_0 = 0$  versus  $H_a: p_1 - p_0 > 0$   
> prop.test(c(bse1, bse0), c(n1,n0), alternative="greater",correct=FALSE)
```

2-sample test for equality of proportions without continuity
correction

```
data:  c(bse1, bse0) out of c(n1, n0)  
X-squared = 9.5733, df = 1, p-value = 0.0009871  
alternative hypothesis: greater  
95 percent confidence interval:  
0.0390609 1.0000000  
sample estimates:  
prop 1      prop 2  
0.8756614 0.7938389  
>  
>detach(Bse)
```