https://datascience-project.eu/

Modello di scheda di formazione

Titolo	Modelli lineari generalizzati: ANOVA
Parole chiave (meta tag)	Analisi multivariata, variabilità interna e tra gruppi, test di ipotesi, modelli lineari
Lingua	Italiano
Obiettivi / Finalità / Risultati di apprendimento	**Lo scopo di questo modulo è presentare i concetti di base dell'analisi della varianza a uno e due fattori (ANOVA), che può essere intesa come un modello lineare di base** **Alla fine di questo modulo sarai capace di:** - **Capire come ANOVA può essere utile per testare se ci sono differenze tra il valore medio di una variabile continua tra diversi livelli di una o più variabili categoriali.** - **Comprendere e identificare le condizioni necessarie per applicare queste tecniche.** - **Condurre analisi della varianza a una o più vie e interpretare i risultati ottenuti.**

Corso di formazione:	
Alfabetizzazione alla scienza dei dati	
Modulo di visualizzazione dei dati e analisi visiva	X
Introduzione alla scienza dei dati per le scienze umane e sociali	
Scienza dei dati per il sociale	
Giornalismo dei dati e storytelling	

Descrizione	In questo modulo formativo verrai introdotto all'uso della modellazione lineare di base per capire come le differenze medie possono essere attribuite o meno all'effetto di variabili categoriali. L'analisi qui presentata è alla base della regressione lineare, che considera anche l'effetto delle variabili continue. Le tecniche descritte in questo modulo formativo si limitano al caso di variabili categoriali (qualitative). A questo proposito, puoi affrontare i contenuti di questo modulo come un'introduzione al Modello lineare generalizzato (MLG)

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

	che utilizza solo fattori categoriali per spiegare la variabilità in una variabile (continua) di interesse. La procedura qui presentata si basa sulla scomposizione della variabilità totale misurata nel campione in diverse fonti: alcune sono residuali (o non spiegate dai fattori considerati) mentre altre provengono da una parte sistematica riconducibile alle diverse categorie dei fattori categoriali.
Contenuti organizzati in 3 livelli	**1. INTRODUZIONE** Le tecniche MLG qui presentate sotto forma di Analisi della varianza (ANOVA) consentono di rispondere a domande potenzialmente interessanti. Qualche esempio: a. I lavoratori e le lavoratrici di una regione percepiscono lo stesso salario annuo medio? b. Gli studenti di un corso che seguono metodi di insegnamento diversi ottengono la stessa media? c. Il consumo medio settimanale di determinati farmaci è diverso a seconda dei gruppi di età e/o sesso? L'ANOVA a un fattore va bene per le domande 1 e 2, mentre la domanda 3 richiede l'ANOVA a due fattori. Il nostro obiettivo è testare l'effetto di una variabile indipendente (fattore) classificata in k diverse categorie (*livelli*) su una variabile dipendente numerica (variabile di risposta), e si basa sulla scomposizione della variabilità totale del campione. Possiamo affrontare questo problema come un test di ipotesi statistica di un'ipotesi nulla (H0; il nostro default) rispetto a un'alternativa (H1; una visione del mondo alternativa). Il test è formulato in termini di medie della popolazione della variabile di risposta attraverso i livelli del/i nostro/i fattore/i. $H_0: \mu_1 = \mu_2 = \cdots = \mu_K$ H_1: At least two different μ_i Le ipotesi richieste per condurre il test ANOVA sono: - Popolazioni normali: la distribuzione della variabile di risposta su ogni livello dovrebbe essere normale - Uguaglianza delle varianze: le varianze della variabile di risposta tra i livelli devono essere le stesse

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

- Semplici indipendenti: i dati campionari su ciascun livello del fattore non sono correlati con gli altri dati campionari (raccolti dagli altri livelli)

2. ANOVA A UN FATTORE

2.1. La procedura

La procedura ANOVA con un fattore si basa sulla seguente equazione:

$$X_{ir} = \mu + \alpha_i + u_{ir}$$

dove Xir è il valore della nostra variabile di risposta per l'individuo r alla categoria (livello) i. Supponiamo che questo valore sia la somma di tre effetti:

- Un valore medio generale (μ), comune a tutti gli individui e a tutti i livelli
- Uno spostamento (αi) che coglie l'influenza media dell'appartenenza al livello i
- Un residuo (uir) , che tiene conto delle variazioni casuali e incontrollate. Si presume che questo residuo si distribuisca normalmente con media nulla.

Il test ANOVA equivale a verificare se gli spostamenti αi sono identici attraverso i livelli k. In caso contrario, ci saranno differenze significative nelle medie.

Prendiamo i dati del campione su X e scomponiamo la sua variabilità (dispersione attorno alle medie del campione) in due parti:

a. La parte interna al gruppo (SSW) rappresenta la variabilità interna.
b. La variabilità tra gruppi (SSB) tiene conto delle differenze tra la media campionaria di ciascun gruppo e la media generale.

La variabilità totale (SST) è solo la somma di SSW+SSB. Se SSB è molto più grande di SSW, questo indica che ci sono differenze significative nelle

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

medie di gruppo. Quindi, ci saranno differenze significative nelle medie tra i livelli del fattore.

Per confrontare il peso relativo di SSB e SSW sulla variabilità totale, li scaleremo dividendoli per il numero di gradi di libertà, ottenendo rispettivamente i valori MSB e MSW.

$$d = \frac{MSB}{MSW} = \frac{\frac{1}{k-1}\chi^2_{k-1}}{\frac{1}{n-k}\chi^2_{n-k}} \sim F^{k-1}_{n-k}$$

Se i presupposti richiesti sono validi, la statistica (d) calcolata come MSB⁄MSW si distribuisce come un modello F. Questa statistica permette di prendere una decisione sul test: più alto è il suo valore, più grande (relativamente) è la variabilità tra gruppi rispetto alla variabilità interna.

Ma come possiamo sapere se d è alto o no? Calcolando il valore p associato a questo test: calcoliamo il valore p (la probabilità alla coda destra della relativa distribuzione F) e se questo valore p è basso rifiutiamo l'ipotesi nulla (cioè ci sono differenze significative nella media tra i livelli)

2.2. Un esempio

A titolo di esempio illustrativo, si supponga di voler testare se il design delle confezioni in cui viene venduta una specifica marca di latte ha qualche influenza sulle vendite. Con questo obiettivo, prendiamo un campione di 12 punti vendita con caratteristiche simili e, fissando lo stesso prezzo per il latte, assegniamo a caso un tipo di confezione (1, 2 o 3). Quindi otteniamo i dati di esempio della nostra variabile di risposta "Vendite", che misura quante migliaia di bottiglie di latte sono state vendute in un mese, come illustrato di seguito:

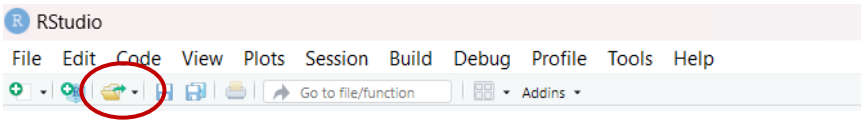
Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Filter

	Sales	Package
1	2.2	1
2	2.5	1
3	2.4	1
4	2.6	1
5	3.1	2
6	2.8	2
7	3.2	2
8	3.3	2
9	2.5	3
10	2.8	3
11	3.2	3
12	2.5	3

I nostri dati di esempio mostrati sopra sono contenuti in un file R, che possiamo aprire andando qui (chiamiamo questo file di dati "Latte"):



Vogliamo verificare se ci sono differenze statisticamente significative nelle vendite medie, a seconda del design della confezione. Stiamo applicando ANOVA con R, che richiede l'installazione di pacchetti specifici:

```
#install and load the relevant packages
install.packages("car")
install.packages("dplyr")
library(car)
library(dplyr)
```

Per applicare ANOVA, dobbiamo prima assicurarci che le ipotesi richieste siano effettivamente valide, quindi eseguiamo le seguenti parti di codice:

```
# test normality (by group)
Milk %>%
  group_by(Package) %>%
  summarise(statistic = shapiro.test(Sales)$statistic,
            p.value = shapiro.test(Sales)$p.value)
```

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Queste righe indicano prima il set di dati considerato ("Latte"), quindi raggruppano i dati in base ai livelli del fattore ("Confezione") e infine eseguono un test di normalità Spahiro sulla nostra variabile di risposta ("Vendite") attraverso i gruppi:

```
  Package statistic p.value
    <dbl>     <dbl>   <dbl>
1       1     0.971   0.850
2       2     0.927   0.577
3       3     0.854   0.241
```

Gli alti valori di p di questo test di normalità per tutti i livelli ci permettono di lavorare sotto l'ipotesi di normalità richiesta. Inoltre, si ipotizza che le varianze siano uguali, il che ci porta a eseguire un test di Barlett sulle varianze omogenee, come illustrato di seguito:

```
# test for homogeneous variances (by group)
bartlett.test(Milk$Sales, Milk$Package)
```

Il valore p visualizzato di seguito suggerisce che questa ipotesi è decisamente realistica:

```
        Bartlett test of homogeneity of variances

data:  Milk$Sales and Milk$Package
Bartlett's K-squared = 1.2076, df = 2, p-value = 0.5467
```

Dato che i presupposti necessari sembrano essere soddisfatti, conduciamo la metodologia ANOVA eseguendo le seguenti linee di codice:

```
# run the ANOVA
anova(lm(Sales ~ Package, Milk))
```

Il che produce il seguente risultato:

```
Analysis of Variance Table

Response: Sales
          Df  Sum Sq Mean Sq F value Pr(>F)
Package    1 0.21125 0.21125  1.6794 0.2241
Residuals 10 1.25792 0.12579
> 
```

I risultati del test ANOVA indicano che i diversi design delle confezioni non sembrano avere un impatto sulle vendite medie: la parte di variabilità spiegata dai diversi livelli del fattore "Confezione" (variabilità tra i gruppi) non è significativamente maggiore della parte residua (variabilità interna). Di conseguenza, il valore p associato a questo test è

Co-funded by the Erasmus+ Programme of the European Union
With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

alto e ci dice che non ci sono motivi per rifiutare l'ipotesi nulla di vendite medie uguali tra i vari design.

3. ANOVA a due fattori

3.1 La procedura

Le idee spiegate per il caso dell'ANOVA a un solo fattore possono essere estese per adattarsi a problemi in cui più di un fattore può influenzare la mia variabile di risposta. Il test ANOVA viene ora esteso per tenere conto di un secondo fattore e di una possibile interazione:

$$X_{ijr} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + u_{ijr}$$

Dove X_"ijr" è il valore della nostra variabile di risposta per l'individuo r alla categoria (livello) i del fattore α e al livello j del fattore β. Assumiamo che questi valori si discostino dalla media generale ("μ"), come somma di quattro effetti:

a. uno spostamento ("α" _i) che cattura l'influenza media dell'appartenenza al livello i del fattore α
b. Un secondo spostamento (β_j) che cattura l'influenza media dell'appartenenza al livello j del fattore β
c. un termine di interazione tra questi due fattori (αβ)_"ij"
d. Un residuo "u" _"ir", che tiene conto delle variazioni casuali e non controllate. Si ipotizza che questo residuo si distribuisca normalmente con media zero.

Ora i confronti tra le diverse parti della variabilità sono più complessi. Ogni fonte di variazione viene confrontata (opportunamente scalata dal numero di gradi di libertà) con la varianza residua. L'intuizione è la stessa dell'ANOVA a un fattore, ma ci sono tre diversi test, come riassunto nella tabella seguente:

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

SOURCE OF VARIATION	SUM OF SQUARES	d.f.	MEAN OF SQUARES	F
Factor α	SS_α	k-1	MS_α	MS_α/MSR
Factor β	SS_β	h-1	MS_β	MS_β/MSR
Interaction ($\alpha\beta$)	$SS_{\alpha\beta}$	(k-1) (h-1)	$MS_{\alpha\beta}$	$MS_{\alpha\beta}/MSR$
Residual	SSR	n-hk	MSR	
Total	SST	n-1		

3.2. Un esempio

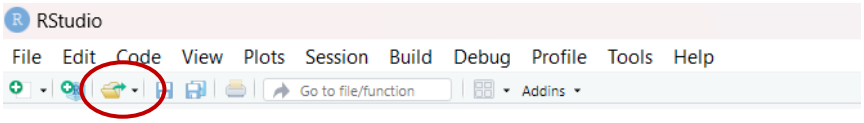
Illustreremo empiricamente il funzionamento dell'ANOVA a due fattori, ipotizzando il seguente problema: un centro sanitario vuole analizzare la potenziale influenza dell'età e del sesso sull'uso di un farmaco. A questo scopo viene condotta un'indagine a campione e gli utenti sono stati raggruppati per età in quattro categorie (bambini, adolescenti, adulti, anziani) e per sesso. È stato creato un campione di 24 persone, selezionando in modo indipendente 3 individui per sesso e gruppo di età. La variabile di risposta è il consumo mensile di questo farmaco (in €) e abbiamo il seguente set di dati:

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

	consumption	sex	age
1	3.0	Male	Child
2	4.0	Male	Child
3	2.8	Male	Child
4	3.2	Female	Child
5	3.0	Female	Child
6	4.1	Female	Child
7	1.8	Male	Teenager
8	1.0	Male	Teenager
9	1.5	Male	Teenager
10	2.1	Female	Teenager
11	1.2	Female	Teenager
12	1.7	Female	Teenager
13	2.5	Male	Adult
14	2.8	Male	Adult
15	3.0	Male	Adult
16	3.0	Female	Adult
17	4.0	Female	Adult
18	2.9	Female	Adult
19	5.0	Male	Senior
20	5.2	Male	Senior
21	6.0	Male	Senior
22	4.9	Female	Senior
23	5.1	Female	Senior
24	6.2	Female	Senior

Anche in questo caso, i dati di esempio mostrati sopra (contenuti in un file R chiamato "medicina") possono essere caricati in Rstudio andando qui:



Ora stiamo applicando un'ANOVA a due fattori (età e sesso) con R, che richiede l'installazione e il caricamento di pacchetti specifici:

```
#install and load the relevant packages
install.packages("car")
install.packages("dplyr")
library(car)
library(dplyr)
```

Per applicare l'ANOVA, dobbiamo innanzitutto verificare se le assunzioni richieste sono effettivamente valide, eseguendo i test di normalità e di

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

uguaglianza delle varianze. I test di normalità (per tutte le fasce d'età e per i due generi) vengono condotti eseguendo:

```
# we test normality by group first
Medicine %>%
  group_by(age,sex) %>%
  summarise(statistic = shapiro.test(consumption)$statistic,
            p.value = shapiro.test(consumption)$p.value)
```

Indichiamo innanzitutto il dataset considerato ("Medicina"), poi raggruppiamo i dati in base ai livelli dei fattori di traino considerati nella nostra analisi ("età" e "sesso") e infine eseguiamo un test di normalità Spahiro sulla variabile "consumo" attraverso tutti i gruppi:

```
  age      sex    statistic p.value
  <fct>    <fct>      <dbl>   <dbl>
1 Child    Male       0.871   0.298
2 Child    Female     0.881   0.328
3 Teenager Male       0.980   0.726
4 Teenager Female     0.996   0.878
5 Adult    Male       0.987   0.780
6 Adult    Female     0.818   0.157
7 Senior   Male       0.893   0.363
8 Senior   Female     0.862   0.274
```

Si noti che ora, quando ci si riferisce ai livelli dei due fattori, occorre considerare tutte le coppie di possibili categorie tra di essi. In tutti i casi troviamo valori p elevati per il test di normalità, che ci permettono di lavorare sotto l'ipotesi di normalità richiesta. Inoltre, è richiesta anche l'omogeneità delle varianze, che in questo caso viene verificata con il test di Levene:

```
#testing for equal variances
leveneTest(consumption ~ age*sex, data=Medicine, center="mean")
```

Il valore p trovato indica che non abbiamo prove empiriche nel campione contro questa ipotesi:

```
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value Pr(>F)
group  7  0.9575 0.4926
```

Poiché le ipotesi necessarie per condurre un processo di ANOVA a due fattori sembrano essere valide, lo facciamo eseguendo le seguenti linee di codice:

```
# two factor ANOVA analysis
anova(lm(consumption ~ age*sex, Medicine))
```

Co-funded by the Erasmus+ Programme of the European Union

	Il risultato dell'analisi si presenta sotto forma della seguente tsabella ANOVA multipla:
	
	`Analysis of Variance Table`
	
	`Response: consumption`
	`Df Sum Sq Mean Sq F value Pr(>F)`
	`age 3 45.250 15.0833 51.8625 1.827e-08 ***`
	`sex 1 0.327 0.3267 1.1232 0.305`
	`age:sex 3 0.223 0.0744 0.2560 0.856`
	`Residuals 16 4.653 0.2908`
	`---`
	`Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1`
	
	I risultati di questa ANOVA a due fattori forniscono informazioni molto utili che permettono di dare una risposta basata sui dati alla nostra domanda di ricerca. I test condotti indicano che i valori medi del consumo del farmaco sono significativamente diversi tra i quattro livelli del fattore "età" (si noti che è l'unico caso in cui abbiamo un basso valore p, che porta a rifiutare l'ipotesi nulla di medie uguali). Tuttavia, non troviamo differenze significative nel consumo medio in base al sesso o alle interazioni tra gruppo di età e sesso.

Autovalutazione (domande e risposte a scelta multipla)	Nell'ANOVA a un fattore, i residui: a) si assume che siano correlati b) Si assume che siano normali c) Non è necessaria alcuna ipotesi sui residui. L'ipotesi nulla in un'ANOVA a un fattore afferma che: a) tutte le medie sono uguali tra i livelli b) ci sono solo due medie uguali c) Tutte le medie sono diverse La statistica ANOVA a due fattori per verificare la significatività del fattore α ha una distribuzione: a) Chi-quadrato b) T di Student c) F di Snedecor
Risorse (video, link di riferimento)	
Materiale correlato	

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

PPT correlato	
Bibliografia	NEWBOLD, P. et al. (2008): Statistics for Management and Economics, (6th edition) Ed. Prentice Hall. Chapter 17, pp. 635-661.
Fornito da	[Uniovi]

Co-funded by the Erasmus+ Programme of the European Union

With the support of the Erasmus+ programme of the European Union. This document and its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.