Design e Analisi di Esperimenti (per l'Interaction Design)*

Davide Rocchesso

This work is produced by OpenStax-CNX and licensed under the
Creative Commons Attribution License 2.0†

Abstract

Una guida rapida alla sperimentazione con soggetti e all'analisi statistica dei dati, nel contesto del design di oggetti interattivi.

Per la valutazione di interfacce e sistemi interattivi esiste una pletora di paradigmi e tecniche, le quali forniscono per lo più informazioni di tipo qualitativo, oppure informazioni quantitative ma di incerta affidabilità statistica. Invece, se l'elemento di interfaccia da analizzare non è troppo complesso e se si possiedono sufficienti risorse, è possibile adattare alla valutazione di interfacce le tecniche usate nella psicologia sperimentale. In altre parole, si può cercare di fare sperimentazione scientifica con soggetti umani.

Definition 1: Scopo
verificare una **ipotesi** che predice una relazione tra due o più **variabili**.

Example 1: Ipotesi
La velocità di lettura di un testo su un display è diversa per il font Helvetica rispetto al font Times.

Il ricercatore manipola le **variabili indipendenti** (es. font) e misura le **variabili dipendenti** (es. tempo di lettura).

Definition 2: Condizioni

- Condizioni sperimentali: Per esempio, la condizione 1 può essere "leggi in Helvetica" e la condizione 2 può essere "leggi in Times"
- Condizioni di controllo (o gruppi di controllo): sono condizioni molto vicine a quelle sperimentali ma non esposte alle variabili indipendenti oggetto di indagine. Ad esempio, un gruppo di controllo può leggere un testo uguale a quello della condizione 1 ma stampato su carta e sovrapposto al display.

Definition 3: Disegni sperimentali
I disegni sperimentali sono caratterizzati dallo spazio $x \times y \times ... \times z$ delle condizioni sperimentali.

Example 2: Disegno 2 x 2
Effetto sull'efficienza di svolgimento di un task dell'introduzione di una nuova versione di software su soggetti esperti e soggetti inesperti.

Allocazione dei soggetti

Definition 4: Between-subjects design
Differenti gruppi di soggetti sono usati per le diverse condizioni.

*Version 1.10: Feb 5, 2007 2:08 am -0600
†http://creativecommons.org/licenses/by/2.0/

- Vantaggi: non ci sono effetti di ordinamento
- Svantaggi: servono molti soggetti

Definition 5: Within-subjects design
Ogni soggetto si presta a tutte le condizioni

- Vantaggi: bastano pochi soggetti e c'è un minore impatto delle differenze individuali
- Svantaggi: possono facilmente emergere effetti dovuti all'ordinamento delle prove. Bisogna quindi applicare **counterbalancing**.

Definition 6: Matched-group procedure
Si fa il match dei gruppi di soggetti sulla base di variabili che non sono direttamente sotto osservazione. Ad esempio, soggetti con grado paragonabile di esperienza sono assegnati ciascuno ad una condizione della variabile indipendente oggetto di indagine.

Example 3: Sperimentazione in Interazione Uomo-Macchina
Esperimento teso a trovare il compromesso ottimo breadth vs. depth negli hyperlink del web. (Larson and Czerwinski, Microsoft, 1998) [1] .

- Condizioni (rami ad ogni livello): (1) 8 x 8 x 8; (2) 16 x 32; (3) 32 x 16
- Soggetti: 16 utenti esperti del web.
- Task: ogni soggetto fa 8 ricerche per ogni condizione, per un totale di 24 ricerche.
- Variabile dipendente: Tempo di ricerca

	Condizioni		
Risultati	**1**	**2**	**3**
Media	58	36	46
SD	23	16	26

Table 1

Conclusione: Breadth è meglio di Depth, ma troppi link possono peggiorare la performance.

1 Statistica per esperimenti

NOTE: Queste note di statistica e gli esempi riportati sono basati sui libri *Statistica per psicologi*[1], di Caudek e Luccio, e *Statistics: an introduction using R*[2] di Crawley.

Definition 7: Ipotesi Nulla H_0
La manipolazione della variabile indipendente non ha alcun effetto sulla variabile dipendente.

Definition 8: Ipotesi Sostantiva H_1
La manipolazione della variabile indipendente ha effetto sulla variabile dipendente. In medicina, si afferma l'efficacia di un trattamento.

Se, per via sperimentale, si accerta che la probabilità associata agli eventi è troppo bassa, l'ipotesi nulla viene respinta e si dimostra l'efficacia del trattamento, o l'esistenza di un effetto della variabile dipendente su quella indipendente.

[1] http://research.microsoft.com/∼marycz/p25-larson.pdf

Definition 9: Livello di significatività α
E' la soglia di probabilità al di sopra della quale accetto l'ipotesi nulla. E' la probabilità di commettere un errore del primo tipo.

Definition 10: Errore del primo tipo
Dice che il trattamento è efficace (o è presente l'effetto) quando in realtà non lo è.

Definition 11: Livello di significatività β
E' la probabilità di accettare l'ipotesi nulla, laddove essa sia effettivamente falsa. E' la probabilità di commettere un errore del secondo tipo.

Definition 12: Errore del secondo tipo
Dice che il trattamento è inefficace (o non è presente l'effetto) quando in realtà l'efficacia c'è.

Definition 13: valore p
E' una stima della probabilità che un certo risultato, o uno ancora più estremo, si sia verificato per caso, in caso di validità dell'ipotesi nulla. Un valore piccolo di p significa che l'ipotesi nulla è difficile da sostenere e che l'effetto misurato è significativo. Piccolo significa usualmente $<$ (0.05)

	Situazione Effettiva	
Ipotesi Nulla	**vera**	**falsa**
accetta	corretto	tipo II
respingi	tipo I	corretto

Table 2

Definition 14: Potenza di un test statistico
E' la probabilità $1-\beta$ di rigettare l'ipotesi nulla quando essa è falsa. Se si vuole rilevare una variazione, dovuta a un trattamento, pari a δ e ci si può ragionevolmente aspettare una variabilità nei dati misurata da una varianza s^2, allora si può dimensionare la numerosità del campione secondo la regola $n = \frac{\text{potenza}s^2 10}{\delta^2}$. Un valore considerato adeguato per la potenza statistica è 0.8. Ad esempio, se la media è circa 20 e la varianza è circa 10, allora per rilevare una variazione del 10 % con potenza 0.8 bisogna usare $n = 20$.

Example 4

- Ipotesi nulla: nascere maschio o nascere femmina non ha influenza sul fatto di diventare programmatori. H_0: $p_m = p_f = 0.5$
- Ipotesi sostantiva: H_1: $p_m > p_f$
- Livello di significatività: $\alpha = 0.05$

Se prendiamo un campione di 10 programmatori e, di questi, 2 sono donne, cosa possiamo concludere? La probabilità di ottenere il caso osservato e tutti quelli altrettanto o più sfavorevoli all'ipotesi nulla è pari a `probabilità di tutte le sequenze con 2 femmine + probabilità di tutte le sequenze con 1 femmina + probabilità di tutte le sequenze con 0 femmine`. La prima di queste tre probabilità, ad esempio, si trova come ${p_m}^8{p_f}^2$ moltiplicata per il numero di possibili sequenze con due femmine, cioè per le combinazioni di 10 su 8 elementi. La somma delle tre probabilità risulta essere complessivamente pari a 0.0547 e quindi, con il livello di significatività adottato, si accetta l'ipotesi nulla. In altri termini l'evento osservato e quelli altrettanto o più sfavorevoli sono troppo poco improbabili per poter accettare l'ipotesi sostantiva.

1.1 Verifica di ipotesi sulla differenza tra due medie (between-subjects design)

Si supponga di avere un campione di numerosità n estratto da una popolazione normale di media μ e varianza σ^2. La media di tale campione è anch'essa una variabile aleatoria normale, con media μ e varianza $\frac{\sigma^2}{n}$. Presi invece due campioni da due popolazioni, il valore atteso della differenza tra le medie di due campioni è

$$\begin{aligned} E\left(\bar{Y}_1 - \bar{Y}_2\right) &= E\left(\bar{Y}_1\right) - E\left(\bar{Y}_2\right) \\ &= \mu_1 - \mu_2 \end{aligned} \tag{1}$$

Se le medie dei due campioni sono indipendenti, allora la varianza della differenza delle medie campionarie è

$$\begin{aligned} \sigma^2\left(\bar{Y}_1 - \bar{Y}_2\right) &= \frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2} \\ &= \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \end{aligned} \tag{2}$$

dove l'ultima uguaglianza assume che la varianza nei due campioni sia la stessa. A partire da un campione, lo stimatore privo di bias della varianza della popolazione è $s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}$. Questo perché, una volta conosciuta la media, i gradi di libertà associati ad un campione di numerosità n sono $n-1$ e nel calcolo della varianza campionaria bisogna dividere per il numero di gradi di libertà.

Dati due campioni di n_1 e n_2 elementi uno stimatore della varianza della popolazione è $\hat{\sigma}^2 = \frac{n_1 {s_1}^2 - n_2 {s_2}^2}{n_1 + n_2 + -2}$. Le ipotesi sulla differenza tra le medie si verificano con la statistica

$$\begin{aligned} t &= \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{\text{differenza tra le medie}}{\text{deviazione standard della differenza tra le medie}} \end{aligned} \tag{3}$$

la quale è distribuita come una t di Student con $n_1 + n_2 + -2$ gradi di libertà. Maggiore è la differenza tra le medie e più siamo convinti dell'efficacia di un trattamento. Maggiore è la varianza dei campioni, e meno ne siamo convinti.

Example 5

Due software didattici (A e B) vengono confrontati esaminando i voti dei test di due classi di studenti. Ci si chiede se i due software siano diversamente efficaci.

Usiamo il software libero R[2] per analizzare i risultati dei test e rispondere alla domanda. La sequenza di istruzioni

```
> classeA
[1] 3 4 4 3 2 3 1 3 5 2
> classeB
[1] 5 5 6 7 4 4 3 5 6 5
> profitto <- c(classeA, classeB)
> label <- factor(c(rep("A", 10), rep("B", 10)))
> boxplot(profitto~label, notch=T, xlab="classi", ylab="profitto")
```

mostra i risultati dei test nelle due classi e disegna il boxplot di Figure 1 (Boxplot). Poiché i solchi dei due plot non si sovrappongono, si può concludere che le medie sono significativamente diverse, con un livello di significatività al 5%. La variabilità è simile in entrambe le classi, e pertanto ha senso procedere con un t-test.

[2] http://www.r-project.org/

Boxplot

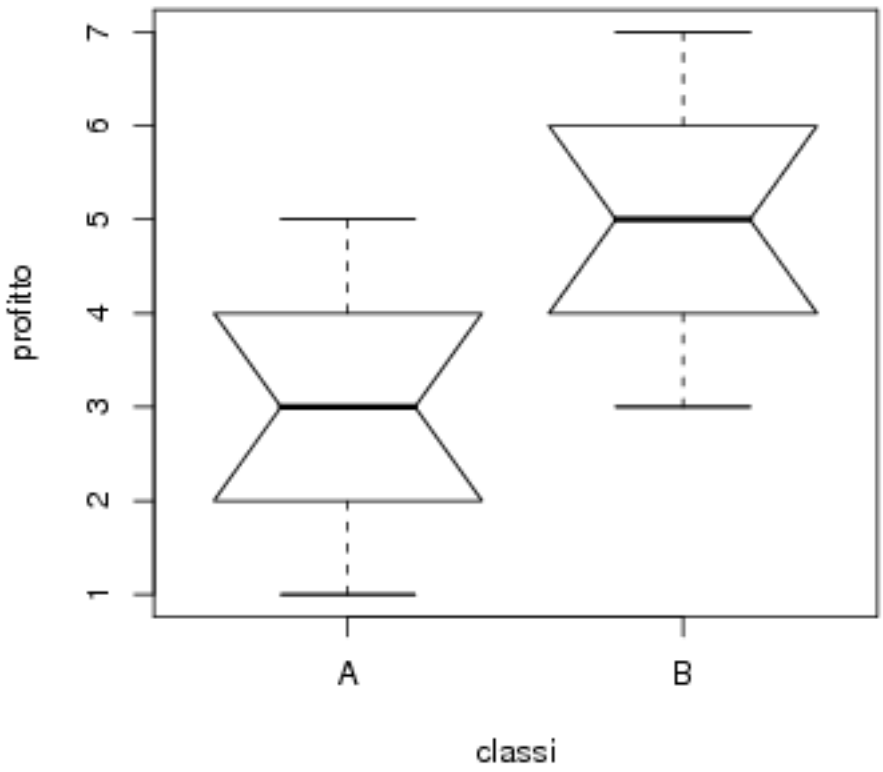

Figure 1: Boxplot.

Possiamo calcolare esplicitamente le varianze campionarie per le due classi e, da queste, il valore della variabile t di Student:

```
> s2A <-  var(classeA)
> s2B <-  var(classeB)
> (mean(classeA) - mean(classeB))/sqrt(s2A/10 + s2B/10)
[1] -3.872983
```

Il valore ottenuto va considerato in modulo, e confrontato con il valore critico per la significatività scelta, il quale è

```
> qt(0.975, 18)
[1] 2.100922
```

Nel primo parametro della chiamata a `qt()`, si noti come il test vada condotto a **due code**, e quindi α divisa per due, perché non c'è una direzione preferenziale di confronto: ci interessa solo la diversità tra le due classi e non una relazione d'ordine. Il secondo parametro è il numero complessivo di gradi di libertà. Poiché il valore calcolato è più grande del valore critico l'ipotesi nulla può essere respinta. Rovesciando il ragionamento, possiamo attribuire un valore di probabilità all'osservazione, supponendo che valga l'ipotesi nulla. Questa probabilità è

```
> 2*pt(-3.872983, 18)
[1] 0.001114540
```

ed è molto inferiore al 5% che ci eravamo posti come soglia. In `R`, il t-test si può condurre con una sola istruzione

```
> t.test(classeA, classeB)

Welch Two Sample t-test

data:  classeA and classeB
t = -3.873, df = 18, p-value = 0.001115
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-3.0849115 -0.9150885
sample estimates:
mean of x mean of y
3         5
```

In un rapporto di ricerca, il risultato dell'analisi si può presentare come "il profitto della classe B (media = 5) è significativamente diverso da quello della classe A (media = 3; t = 3.873, p = 0.0011 (two-tailed), d.f. = 18)".

Con il software libero Octave[3] , la stessa analisi si svolgerebbe con il comando

```
octave> classeA = [3 4 4 3 2 3 1 3 5 2];
octave> classeB = [5 5 6 7 4 4 3 5 6 5];
octave> [p, t, df] = t_test_2(classeA, classeB, "<>")
p = 0.0011145
t = -3.8730
```

[3] http://www.octave.org

```
df = 18
```

L'ipotesi fondante per la verifica di ipotesi sulle medie è che la popolazione sia normale, e di varianza omogenea tra i campioni. Il metodo è robusto a variazioni dalla normalità. Però deve valere l'ipotesi di indipendenza tra le medie, e questo non si può avere con test di tipo "within subjects". Per ovviare a questo problema si può però riformulare il test considerando i dati come provenienti da un unico campione di differenze di punteggi D. L'ipotesi nulla diventa H_0: $E\left(\bar{D}\right) = 0$, e si verifica con

$$t = \frac{\bar{D} - E\left(\bar{D}\right)}{\frac{s_D}{\sqrt{n}}} \tag{4}$$

con $n - 1$ gradi di libertà.

1.2 Regressione multipla

Si abbiano k variabili indipendenti quantitative. Il modello di regressione lineare si esprime come

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_k x_{k,i} + \varepsilon_i \tag{5}$$

Con n osservazioni il modello si può esprimere in forma matriciale

$$y = Xb + \varepsilon \tag{6}$$

Tipicamente il numero di osservazioni è superiore al numero di variabili indipendenti, e pertanto il sistema (6) risulta sovradeterminato. Esso si può risolvere ai minimi quadrati[4] ottenendo

$$b = \left(X^{'}X\right)^{-1} X^{'}y \tag{7}$$

La (7) è la soluzione che minimizza la **Somma quadratica degli errori**SSE $= \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$, dove $\hat{y}_i$ è la stima dell'i-esimo campione fornita dal modello lineare.

Se si hanno k variabili quantitative e una variabile qualitativa il modello di regressione è

$$y_i = \beta_0 + \gamma D_i + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_k x_{k,i} + \varepsilon_i \tag{8}$$

dove D_i è una variabile dicotomica, cioè assume solo i valori 0 o 1.

Example 6

Si studi la relazione tra la prestazione in un test di apprendimento di un software (variabile y) e il numero di volte in cui i soggetti hanno fatto ricorso all'help durante l'apprendimento dello stesso (variabile x). La variabile qualitativa D discrimina tra utenti esperti e novizi.

Ecco una sessione in `R` che produce la regressione ai minimi quadrati (funzione `lm()`, per linear model) e traccia il grafico di Figure 2 (Regressione multipla).

```
> risp_corrette
[1] 80 88 89 62 67 37 62 78 62 77 81 41 26 41 37 57 35 58 22 25 46 18 56 26 38 21
> help
[1] 16 25 21 13 17  2 11 14  8 12 17  7  6 16 17 21 14 25  7 12 19  9 17 14 16  6
```

[4] "Least Squares" <http://cnx.org/content/m10371/latest/>

```
> esperienza
[1] 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
> rce <- split(risp_corrette, esperienza)
> he <- split(help, esperienza)
> plot(help, risp_corrette, type="n", ylab="Risposte corrette",
+ xlab="invocazioni help")
> points(he[[1]], rce[[1]], pch=16)
> points(he[[2]], rce[[2]])
> lm(risp_corrette ~ help + esperienza)

Call:
lm(formula = risp_corrette ~ help + esperienza)

Coefficients:
(Intercept)         help   esperienza
3.738        2.280       33.962

> abline(3.738, 2.28)
> abline(3.738 + 33.962, 2.28, lty=2)
```

Regressione multipla

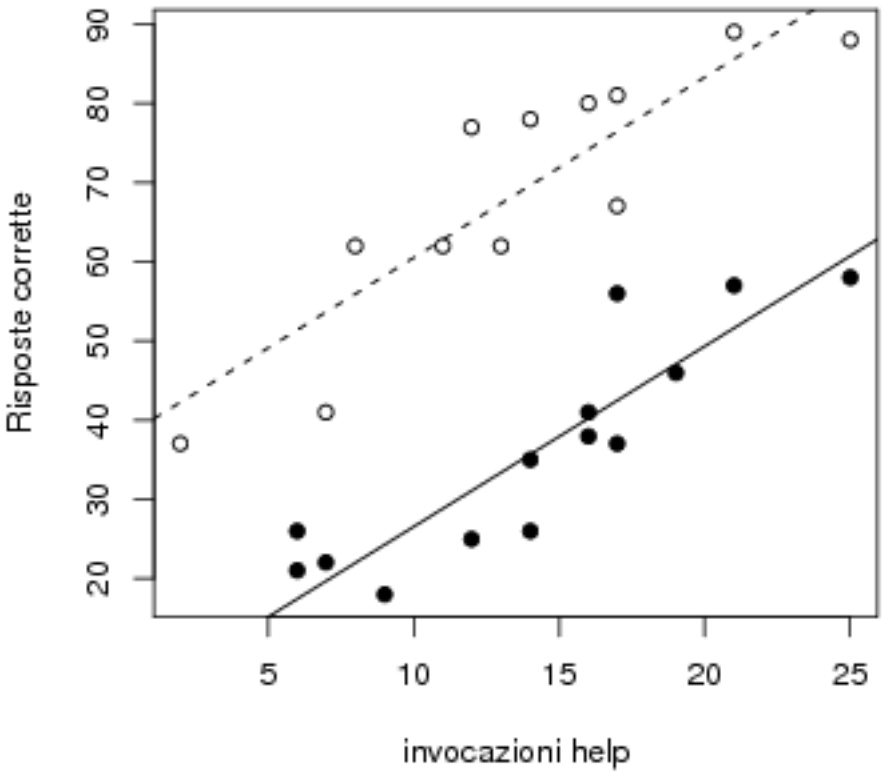

Figure 2: Regressione con una variabile qualitativa e una variabile quantitativa.

Una domanda che è naturale porsi è se e quanto sia significativo il livello di esperienza.

1.3 Analisi della Regressione

Indicata con $\overline{y}$ la media delle osservazioni, e definite la **Somma totale dei quadrati**$\mathrm{SSY} = \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2$ e la **Somma dei quadrati di regressione**$\mathrm{SSR} = \sum_{i=1}^{n} \left(\hat{y}_i - \overline{y}\right)^2$ si dimostra che vale l'ortogonalità tra la variabilità dei dati rispetto al modello lineare e la variabilità del modello rispetto alla media dei dati. Cioè

$$\mathrm{SSY} = \mathrm{SSR} + \mathrm{SSE} \tag{9}$$

Si usa anche dire che la varianza totale è data dalla somma della varianza entro i gruppi (SSE) e tra i gruppi (SSR). Il **Coefficiente di determinazione** è definito come

$$r_{xy}{}^2 = \frac{\text{SSR}}{\text{SSY}} \tag{10}$$

Questo è un numero compreso tra 0 e 1, e vale 1 se non c'è errore residuo, cioè se la regressione fa un fit perfetto. Si dimostra che $r_{xy}{}^2$ è il quadrato del **Coefficiente di correlazione**, definito come il rapporto tra la covarianza e le deviazioni standard campionarie di x e y.

$$\begin{aligned} r_{xy} &= \frac{\sum_{i=1}^{n} x_i - \overline{x} \sum_{i=1}^{n} y_i - \overline{y}}{n S_x S_y} \\ &= \frac{S_{\text{xy}}}{S_x (S_y)} \end{aligned} \tag{11}$$

Proprietà dei coefficienti di regressione multipla

1. Il valore atteso del j-esimo elemento di b è β_j
2. La varianza del j-esimo elemento di b è $\nu_{\text{jj}}\sigma^2$, dove ν_{jj} è il j-esimo elemento diagonale di $\left(X^{'}X\right)^{-1}$ e σ^2 è la varianza del rumore additivo ε
3. $S_E{}^2 = \frac{\text{SSE}}{n-k+1}$ è uno stimatore unbiased di σ.
4. Le variabili di residuo ε_i sono normalmente distribuite.
5. I coefficienti parziali di regressione (elementi di b) sono normalmente distribuiti.

Ulteriori proprietà
sono associabili alla 3 di Proprietà dei coefficienti di regressione multipla, p. 10:

1. $\frac{S_E{}^2 \nu}{\sigma^2}$ è distribuita come una variabile aleatoria $\chi_\nu{}^2$ con ν gradi di libertà.
2. La differenza di due variabili aleatorie χ^2 con ν_1 e ν_2 gradi di libertà è ancora una variabile aleatoria χ^2 con $\nu_1 - \nu_2$ gradi di libertà.
3. Il rapporto di due variabili aleatorie χ^2 con ν_1 e ν_2 gradi di libertà è una variabile aleatoria di tipo F, secondo la relazione $\frac{\frac{\chi_{\nu_1}{}^2}{\nu_1}}{\frac{\chi_{\nu_2}{}^2}{\nu_2}} = F_{\nu_1 \nu_2}$

1.3.1 Verifica di ipotesi sui coefficienti di regressione

Se si deve verificare l'ipotesi nulla H_0: $b_j = \beta_{\text{j, 0}}$ si può applicare, in virtù delle 2 e 3 delle Proprietà dei coefficienti di regressione multipla, p. 10, il t-test con la statistica $t = \frac{b_j - \beta_{\text{j, 0}}}{\sqrt{\frac{\text{SSE}}{n-k+1} \nu_{\text{jj}}}}$

1.3.2 Verifica di minimalità di un modello

Supponiamo di avere un modello con k regressori e di volere verificare l'ipotesi nulla H_0: $\beta_1 = ... = \beta_p = 0$, cioè che esiste un sottomodello ugualmente rappresentativo. Indichiamo con $\text{SSE}_{k\ -\ p}$ lo scarto per la regressione con le $k - p$ variabili rimaste. In virtù delle proprietà (Ulteriori proprietà, p. 10) possiamo scrivere l'equivalenza nel senso della distribuzione statistica

$$\frac{\frac{\text{SSE}_{k\ -\ p} - \text{SSE}}{p}}{\frac{\text{SSE}}{n+-k+-1}} \approx \frac{\frac{\chi_p{}^2}{p}}{\frac{\chi_{n\ -\ k\ -\ 1}{}^2}{n+-k+-1}} \approx F_{\text{p},n-k-1} \tag{12}$$

F-test
Maggiore è l'incremento di errore dovuto alla esclusione di p regressori, minore è la probabilità che valga l'ipotesi nulla. Dati i gradi di libertà di numeratore e denominatore, si legge il valore tabulato per un dato

livello di significatività F_α e si rifiuta l'ipotesi nulla se $F > F_\alpha$. Oppure, il programma calcola la probabilità di errore del primo tipo associata al valore di F calcolato. Se p è piccolo e $F > 1$ si respinge l'ipotesi nulla.

La (12) può anche essere scritta in termini di coefficiente di determinazione:

$$\frac{\frac{\mathrm{SSR}-\mathrm{SSR}_{k\ -\ p}}{p}}{\frac{\mathrm{SSE}}{n+-k+-1}} = \frac{\frac{r^2-r^2{}_{k\ -\ p}}{p}}{\frac{1-r^2}{n+-k+-1}} \tag{13}$$

Un importante caso particolare si ha per $p = k$, per il quale l'ipotesi nulla è: nessuna variabile ha effetto su y. In questo caso il test si fa con

$$\begin{aligned} F &= \frac{\frac{\mathrm{SSR}}{k}}{\frac{\mathrm{SSE}}{n+-k+-1}} \\ &= \frac{\mathrm{between-groups}\quad \mathrm{mean-square}\quad \mathrm{variance}}{\mathrm{within-groups}\quad \mathrm{mean-square}\quad \mathrm{variance}} \end{aligned} \tag{14}$$

Example 7

Si studi la relazione tra lo sviluppo di capacità logiche nella adolescenza (misurate con un test `logictest`) e il tempo speso settimanalmente ai videogame (`vg`) o alla televisione (`tv`). Lo studio è affrontato molto semplicemente in `R`.

```
> vg
[1] 12  7 23  8 19 24  5 31  9 21 24 17 19  6 16 12 12 11 33 14  4  2 21 18 21
> tv
[1] 32  4 12 12 22 21 14  6  9 26 19 18 12 21 18 24 23 23 12 21 26  8 23 12 26
> logictest
[1] 34 12 67 46 43 78 56 69 23 67 89 73 26 41 52 19 83 38 56 43 11 21 54 56 87

> summary(lm(logictest~tv+vg))

Call:
lm(formula = logictest ~ tv + vg)

Residuals:
    Min      1Q  Median      3Q     Max
-28.192 -14.089   1.308  11.183  36.416

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.1431    12.4141   0.898 0.379108
tv            0.6083     0.5203   1.169 0.254910
vg            1.7875     0.4592   3.893 0.000783 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.45 on 22 degrees of freedom
Multiple R-Squared: 0.4249,Adjusted R-squared: 0.3727
F-statistic: 8.129 on 2 and 22 DF,  p-value: 0.002274
```

Il prospetto fornito da `summary(lm())` si può interpretare come segue

- I residui non presentano una distribuzione evidentemente asimmetrica, visti i valori minimo, massimo e mediano, e pertanto si può ragionevolmente assumere la gaussianità;
- Il valore di F è alto ed accompagnato da un p piccolo, e pertanto si può falsificare l'ipotesi nulla secondo cui né i videogiochi né la televisione hanno effetto sullo sviluppo delle capacità logiche;
- Il coefficiente legato alle ore di videogiochi è significativamente (secondo il t-test) positivo, indicando un effetto.

In `octave` lo stesso f-test si svolgerebbe con il comando

```
octave> [p, f, df_num, df_den] = f_test_regression(logictest',
> [ones(1,25);tv;vg]',[[0 1 0];[0 0 1]])
p = 0.0022740
f = 8.1286
df_num = 2
df_den = 22
```

Example 8

Si riprenda lo Example 6 e ci si chieda se è significativo il grado di esperienza. In `R`, da una

```
summary(lm(risp_corrette~help+esperienza))
```

si ricava il coefficiente di determinazione come primo numero riportato nella linea

```
Multiple R-Squared: 0.9051, Adjusted R-squared: 0.8969
```

Il secondo numero, se moltiplicato per 100, si interpreta come riduzione percentuale della varianza apportata dal modello lineare. Trascurando la variabile dicotomica relativa al grado di esperienza, il coefficiente di determinazione risulta da

```
> cor(risp_corrette, help) ^ 2
[1] 0.3099675
```

o da

```
> summary(lm(risp_corrette~help-esperienza))
...
Multiple R-Squared:  0.31,Adjusted R-squared: 0.2812
```

e facendo il calcolo della (13) si trova $F_{1,23} = 144.27$. Il valore di p associato a tale F è molto piccolo, e quindi si rigetta l'ipotesi nulla. In altri termini: il grado di esperienza conta.

Inoltre, possiamo effettuare il t-test per verificare se l'help-on-line è efficace per entrambi i gruppi. Anche tale informazione si può estrarre dal prospetto di

```
summary(lm(risp_corrette~help+esperienza))
```

:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.7382     3.9736   0.941    0.357
help          2.2797     0.2448   9.312 2.89e-09 ***
esperienza   33.9622     2.8275  12.011 2.17e-11 ***
```

Poiché il valore $t = 9.312$ è elevato e la corrispondente $p = 2.89e - 09$ è piccolissima, possiamo considerare provata l'efficacia dell'help.

In `octave`, f-test e t-test si conducono con le due chiamate

```
octave> [p, f, dfn, dfd] = f_test_regression(risp_corrette,
> [ones(1,26); esperienza; help]', [0, 1, 0])
p =  2.1692e-11
f = 144.27
dfn = 1
dfd = 23
octave> [p, t, dfn] = t_test_regression(risp_corrette,
> [ones(1,26); esperienza; help]', [0, 0, 1], 0, ">")
p =  1.4428e-09
t = 9.3115
dfn = 23
```

L'analisi andrebbe completata verificando la assenza di interazioni tra `esperienza` e `help`. Ciò si fa con un altro f-test, aggiungendo il prodotto `esperienza.*help`, e usando i coefficienti di determinazione con e senza il prodotto delle due variabili. In `Octave`:

```
octave>  [p, f, dfn, dfd] = f_test_regression(risp_corrette,
> [ones(1,26); esperienza; help; esperienza.*help]', [0, 0, 0, 1])
p = 0.66901
f = 0.18776
dfn = 1
dfd = 22
```

In `R`, si fa

```
> summary(lm(risp_corrette ~ help + esperienza + help:esperienza))
```

da cui si ottiene

```
Multiple R-Squared: 0.9059,Adjusted R-squared: 0.8931
```

che si può usare insieme al valore del coefficiente di determinazione senza interazioni (0.9051) all'interno della (13), trovando così $F_{1,22} = 0.19$, la cui probabilità associata è

```
> 2*pf(0.19, 1, 22)
[1] 0.6656723
```

In questo caso non c'è evidenza di interazioni tra `esperienza` e `help` e quindi i loro effetti si possono studiare separatamente. Viceversa, la presenza di interazione significherebbe che gli effetti di ciascuna variabile mutano a seconda dei valori assunti dall'altra.

1.3.3 ANOVA

Con il termine ANOVA si indica usualmente una analisi della regressione multipla in cui tutte le variabili indipendenti sono qualitative. Lo scopo è quello di stabilire se esistono differenze statisticamente significative tra i gruppi corrispondenti alle variabili indipendenti.

Definition 15: One-way ANOVA
E' una ANOVA con una sola variabile indipendente qualitativa (o fattore) con m categorie (o livelli). Ci siano n elementi misurati per ogni livello.

$$y_i = \alpha + \gamma_1 D_{1,i} + \gamma_2 D_{2,i} + ... + \gamma_{m-1} D_{m-1,i} + \epsilon_i \tag{15}$$

	D_1	D_2	...	D_{m-1}
Group 1	1	0	...	0
Group 2	0	1	...	0
...	...	...	...	...
Group m-1	0	0	...	1
Group m	0	0	...	0

Table 3

Con questa codifica la matrice $X^{'}X$ risulta invertibile. Ci sono $m-1$ gradi di libertà.
Il test si effettua sulla statistica

$$\begin{aligned} F &= \frac{\text{between-groups mean-square variance}}{\text{within-groups mean-square variance}} \\ &= \frac{\frac{\text{SSA}}{m-1}}{\frac{\text{SSE}}{m(n-1)}} \\ &= \frac{\frac{n}{m-1}\sum_{i=1}^{m}\left(\overline{y}_i - \overline{y}\right)^2}{\frac{1}{m(n-1)}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{\text{i,j}} - \overline{y}_i\right)^2} \end{aligned} \tag{16}$$

il cui valore si confronta con il valore tabulato di $F_{m-1,n-m}$ per un dato valore di significatività α. Se F è circa 1, ciò significa che le differenze tra le medie dei gruppi sono casuali. Invece, se F è significativamente

maggiore di 1 significa che c'è più variazione tra che entro i gruppi, e quindi il raggruppamento gioca un ruolo significativo. Se la one-way ANOVA ha solo due livelli, essa equivale ad un t-test. SSA è la cosiddetta somma dei quadrati di trattamento, e corrisponde alla SSR già definita per la regressione con variabili quantitative.

Example 9

Si misuri mediante un test il profitto di tre gruppi di studenti precedentemente sottoposti a tre programmi di addestramento. Si intende verificare se il tipo di addestramento è ininfluente sul risultato del test.

In R l'ANOVA si esegue nel seguente modo:

```
> test <- c(9, 16, 12, 10,  8, 14,  8, 11,  7,  9, 20, 19, 14, 17, 12)
> addestramento <- c("A", "A", "A", "A", "A", "B", "B", "B", "B", "B",
+ "C", "C", "C", "C", "C")
> summary(aov(test~factor(addestramento)))
                      Df  Sum Sq Mean Sq F value  Pr(>F)
factor(addestramento)  2 123.600  61.800  6.3931 0.01288 *
Residuals             12 116.000   9.667
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

mentre in Octave la stessa analisi si svolge con

```
octave> y1 = [9 16 12 10  8 14  8 11  7  9 20 19 14 17 12];
octave> g = [1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3];
octave> anova(y1, g)

One-way ANOVA Table:

Source of Variation   Sum of Squares    df  Empirical Var
*******************************************************
Between Groups              123.6000     2        61.8000
Within Groups               116.0000    12         9.6667
-------------------------------------------------------
Total                       239.6000    14

Test Statistic f              6.3931
p-value                       0.0129

ans = 0.012877
```

Per un valore di significatività a 0.05 è ragionevole rigettare l'ipotesi nulla e quindi stabilire un effetto del tipo di addestramento.

Nell'analisi ANOVA l'ipotesi nulla H_0 è che non esistano differenze statisticamente significative tra i gruppi corrispondenti alle variabili indipendenti. Se H_0 viene rifiutata, tuttavia, si può procedere con una comparazione a coppie per vedere quali coppie di gruppi sono significativamente differenti. Bisogna fare attenzione,

però, che il livello di significatività α scelto in queste comparazioni deve diminuire con il numero delle stesse. Infatti, supponendo di fare un t-test con $\alpha = 0.05$ su 100 coppie, si ottiene che mediamente 5 di questi confronti sono affetti da errore del I tipo. La probabilità di commettere almeno un errore del I tipo nei 100 confronti è $1 - (1 - \alpha)^{100} = 0.994$ La correzione di Bonferroni prevede che il livello α per n test multipli debba essere diviso per n.

References

[1] C. Caudek and R. Luccio. *Statistica per Psicologi.* Editori Laterza, 2001.

[2] M. J. Crawley. *Statistics: an Introduction using R.* Wiley, 2005.