METODOLOGIA SPERIMENTALE IN AGRICOLTURA

LABORATORIO DI BIOMETRIA CON R
(http://www.R-project.org/)

APPUNTI DALLE LEZIONI (bozze – Settembre 2005)

DOCENTE
Andrea Onofri
Dipartimento di Scienze Agroambientali e della Produzione Vegetale
Sezione di Agronomia e Coltivazioni erbacee
Borgo XX Giugno 74
06121 PERUGIA
Tel: 075-5856324
onofri@unipg.it

UNITA' I: IL PROCEDIMENTO SCIENTIFICO SPERIMENTALE

OBIETTIVO

Comprendere l'ambito nel quale si muove la statistica e la sua importanza a supporto della sperimentazione agronomica. Introdurre alcune nozioni in grado di formare una terminologia scientifica di base.

SOMMARIO

1. Definizione di statistica e introduzione alla biometria
2. Il procedimento scientifico sperimentale e la statistica
3. Statistica descrittiva ed inferenziale
4. L'errore sperimentale
5. Le repliche
6. Collettivo, unità sperimentale e variabili statistiche
7. Variabili quantitative e qualitative

SPIEGAZIONE

Definizione di statistica e introduzione alla biometria

In genere, con il termine statistica si intende la disciplina che studia le tecniche per la raccolta dei dati e la loro elaborazione, in modo da ottenere il più elevato numero di informazioni in riferimento al fenomeno in studio (chimico, fisico, biologico, sociologico, psicologico...).

I campi di applicazione della statistica sono numerosi e spaziano dalla meteorologia alle scienze sociali, alle ricerche di marketing ecc.. Inoltre, la statistica trova applicazione in tutte le scienze sperimentali, come, le scienze biologiche, l'agronomia, le tecnologie alimentari e le discipline relative allo sviluppo rurale.

L'esigenza di conoscere la statistica per chiunque si occupi di scienze sperimentali nasce già dal momento in cui si procede alla raccolta dei dati, che si esegue con procedimenti di **misurazione**, utilizzando strumenti ed unità di misura adeguati. Da questo punto di vista, le operazioni sono tutt'altro che banali, in quanto nessuna misura può essere considerata precisa in senso assoluto, cioè perfettamente coincidente col valore reale della grandezza misurata, che rimane un'entità incognita e inconoscibile. Oltre che assisterci nei procedimenti di misurazione, la statistica ci insegna anche a gestire correttamente le misure ottenute, a sintetizzare i dati in modo da far emergere efficacemente l'informazione in essi contenuta, a testare ipotesi scientifiche e a confrontare trattamenti sperimentali. Quella parte di statistica che ci insegna a misurare correttamente i fenomeni biologici e, in genere, a risolvere problemi legati alla sperimentazione biologica si dice **biometria** (o **biostatistica)** e costituisce l'argomento del prosieguo di questo testo.

Il procedimento scientifico sperimentale e la statistica.

La statistica e la biometria in particolare sono parte integrante ed essenziale dell'applicazione

del **metodo scientifico sperimentale**, basato sulla formulazione di un ipotesi induttiva, che deve essere poi verificata deduttivamente mediante un esperimento appositamente pianificato. Alla pianificazione e alla esecuzione dell'esperimento segue la raccolta e l'analisi dei dati, che consente di verificare o rigettare l'ipotesi iniziale ed, eventualmente, di formularne una successiva (fig. I.1).

Ad esempio, potremmo chiederci quale procedimento tecnologico è più opportuno per produrre vino di qualità. E' chiaro che una scelta di merito può essere effettuata solo se siamo in grado di misurare la qualità e quindi di mettere a confronto i diversi procedimenti in uno schema sperimentale adeguato. Allo stesso modo potremmo chiederci quale tra due varietà d frumento è migliore in un dato ambiente pedo-climatico; in questo caso organizzeremo un esperimento di confronto con metodiche adeguate a misurare la produzione di granella delle due varietà e stabilire quale è la più produttiva. Dopo aver stabilito il procedimento tecnologico o la varietà migliore potremmo chiederci quale effetto avrebbe il cambio della temperatura alla quale viene condotto quel procedimenti tecnologico o il cambio del livello di concimazione per la varietà in studio.

In questo senso, il procedimento scientifico è chiaramente iterativo: ad ogni ipotesi ne segue una successiva, che consente di approfondire il livello di conoscenze raggiunto, in una sequenza senza fine (si spera!). La statistica biometrica consente di seguire questo cammino logico intervenendo in tutte le sue tappe: nella definizione del problema e nella formulazione di un'ipotesi precisa, nell'organizzazione dell'esperimento adatto a verificarla, nella raccolta dei dati e nella loro analisi.

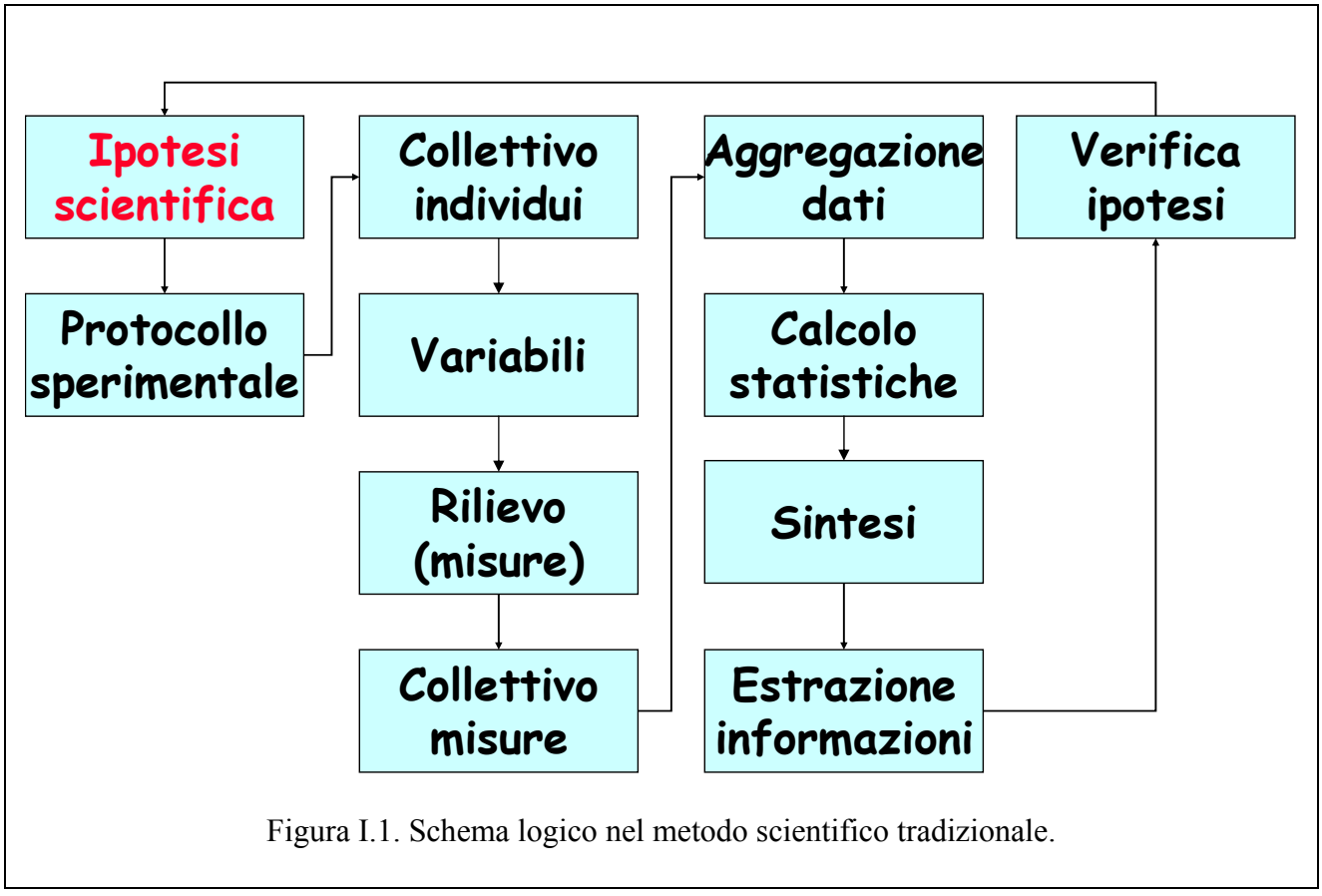


Figura I.1. Schema logico nel metodo scientifico tradizionale.

Statistica descrittiva e inferenziale

Alla fine di un esperimento ci troviamo in mano una mole anche notevole di dati grezzi. Di conseguenza, il primo problema che ci si trova ad affrontare è quello di sintetizzare la massa di dati

grezzi in pochi numeri o indicatori particolarmente informativi, utilizzando metodiche grafiche o numeriche, che siano in grado di descrivere la massa di dati, senza alterarne il senso complessivo. Questa parte della statistica è nota con il nome di **statistica descrittiva**.

Talvolta, la semplice descrizione dei dati grezzi non è il vero scopo dell'indagine statistica. Infatti spesso si studiano fenomeni per i quali non è possibile prendere in considerazione un numero di individui sufficientemente elevato. Ad esempio, se vogliamo sapere la produttività media delle piante di mais di un determinato appezzamento, possiamo anche decidere di raccogliere tutte le piante dell'appezzamento, contarle e determinarne la produzione unitaria. Se invece vogliamo sapere la produttività media delle piante di mais in un intero comprensorio, probabilmente non saremo in grado di valutare tutte le singole piante e i singoli appezzamenti di quel comprensorio, se non con costi e tempi troppo elevati. Pertanto effettueremo le nostre misure su un numero ridotto di appezzamenti (campione rappresentativo; ritorneremo tra breve su questo concetto), scelti a caso tra quelli presenti nel comprensorio in studio e cercheremo poi di risalire dalle caratteristiche degli appezzamenti considerati alle caratteristiche dell'intero comprensorio. Questa operazione prende il nome di **inferenza statistica** e la disciplina relativa si chiama **statistica inferenziale**.

Siccome le ipotesi che si fanno all'inizio dell'esperimento sono di carattere generale, mentre gli esperimenti per loro natura non possono che essere condotti su un numero limitato di individui, è proprio la statistica inferenziale che ci consente di confermare o rigettare l'ipotesi iniziale, completando quindi il procedimento scientifico iniziandone uno nuovo

L'errore sperimentale

Come abbiamo già detto, il problema di fondo che richiede l'adozione della statistica nelle scienze sperimentali è legato all'imprecisione delle misure effettuate. In particolare, nel misurare una determinata grandezza fisica possiamo commettere due tipi di errore: **sistematico** ed **accidentale**.

L'errore sistematico è provocato da difetti intrinseci dello strumento o incapacità peculiari dell'operatore e tende a ripetersi costantemente in misure successive. Un esempio tipico è quello di una bilancia non tarata, che tende ad aggiungere 20 grammi ad ogni misura che effettuiamo. Per queste sue peculiarità, l'errore sistematico non è quantificabile e non costituisce l'oggetto di metodologie statistiche particolari. L'unico modo per contenerlo al minimo livello possibile è quello di ripetere le misure con metodiche diverse, in modo da arrivare ad una perfetta taratura e standardizzazione degli strumenti e/o delle procedure di misura.

L'errore accidentale (o casuale) è invece legato a fattori variabili nel tempo e nello spazio, quali:

1 - *malfunzionamenti accidentali dello strumento*. Si pensi ad esempio al rumore elettrico di uno strumento, che fa fluttuare i risultati delle misure effettuate;
2 - *imprecisioni o disattenzioni casuali dell'operatore*. Si pensi ad esempio ad un banale errore di lettura dello strumento, che può capitare soprattutto ad un operatore che esegua moltissime misure manuali con procedure di routine;
3 - *irregolarità dell'oggetto da misurare* unite ad una precisione relativamente elevata dello strumento di misura. Si pensi alla misurazione del diametro di una biglia apparentemente sferica, con uno strumento molto preciso: è facile che compaiano errori legati all'irregolarità della biglia o al fatto che l'operatore non riesce a misurare la stessa nel punto in cui il suo diametro è massimo. Oppure, più semplicemente si pensi alla misurazione della produzione di granella di una certa varietà di frumento: anche ipotizzando di avere uno strumenti di misura perfetto e quindi esente da errore, la produzione mostrerebbe comunque una fluttuazione naturale da pianta a pianta, in base al patrimonio genetico e, soprattutto, in base alle condizioni di coltivazione che non possono essere standardizzate oltre ad un certo livello (si pensi alla variabilità del terreno agrario).

Dato che queste imprecisioni sono assolutamente casuali è chiaro che le fluttuazioni positive (misura maggiore di quella vera) sono altrettanto probabili di quelle negative e si ripetono con la stessa frequenza. Per questo motivo, l'incertezza di misura che l'errore casuale introduce può essere ridotta e identificata grazie alla **ripetizione della misura.**

Le repliche

Dovrebbe essere intuitivamente chiaro che un risultato assolutamente accurato può essere ottenuto solo ripetendo la misura infinite volte, il che non è tecnicamente fattibile. Per questo motivo le misure vengono ripetute un numero finito di volte (**repliche**) e vengono poi adottate metodiche di inferenza statistica per risalire dai risultati delle misure effettuate a quelli che si sarebbero ottenuti con un numero infinito di repliche.

Nel decidere quante repliche effettuare, bisogna tener presente che la ripetizione delle misure da una parte aumenta la precisione delle stime, ma dall'altra introduce dei costi per l'operatore e va quindi valutata con estrema attenzione, in considerazione della tipologia di misura da effettuare e delle caratteristiche dello strumento da utilizzare.

Collettivo, unità sperimentale e variabili statistiche

In sostanza, in statistica si ha sempre a che fare con un *collettivo,* cioè con un insieme di misure o di individui (animali, piante, terreni, foglie ...) sui quali è stata studiata (o meglio misurata) una certa caratteristica (peso, altezza, contenuto in fosforo, larghezza), in grado di assumere diversi valori e, pertanto, detta *variabile*. Il singolo individuo o la singola misura prendono il nome di *unità sperimentale.*

Variabili qualitative e quantitative

Le variabili statistiche possono essere *qualitative*, se esprimono una qualità dell'individuo, (ad esempio colore e forma delle foglie e dei frutti; si ricordino i "famosi" piselli di Mendel). Una variabile qualitativa non viene misurata, ma classificata in categorie sulla base delle modalità con cui essa si presenta (piselli lisci o rugosi, verdi o gialli).

D'altra parte esistono le *variabili quantitative*, che possono essere misurate su una *scala discreta* (numero di insetti suscettibili ad un certo insetticida, numero di semi germinati in certe condizioni ambientali...) o su una *scala continua* (produzione delle piante o altezza degli alberi...).

VERIFICA

1. Dare una definizione sintetica di statistica e di biometria
2. Cosa si intende per statistica inferenziale?
3. Cosa si intende per statistica descrittiva?
4. Che cosa è un collettivo?
5. Qual è la differenza tra variabili qualitative e quantitative?

UNITA' II: INTRODUZIONE AD R

OBIETTIVO

Imparare le operazioni fondamentali in R, indispensabili per continuare questo corso.

SOMMARIO

SPIEGAZIONE

Introduzione ad R

R è un software cugino di S-PLUS, con il quale condivide la gran parte delle procedure ed una perfetta compatibilità. Rispetto al cugino più famoso, è completamente freeware (sotto la licenza GNU General Public Licence della Free Software Foundation) ed è nato proprio per mettere a disposizione degli utenti un software gratuito, attraverso il quale passare ad S-PLUS, laddove disponibile, mantenendo comunque la capacità di lavorare in proprio senza usare software di frodo.

E' uno strumento molto potente, anche da un punto di vista grafico, ma necessita di una certa pratica, in quanto manca di una interfaccia grafica (Graphical User Interface: GUI) molto avanzata e di conseguenza l'interazione con l'utente avviene a livello di linea di comando (*prompt*).

Inoltre, si tratta di un programma *Open Source*, cioè ognuno può avere accesso al suo codice interno ed, eventualmente, proporne modifiche. Altro vantaggio è che, oltre che un programma, è anche un linguaggio *object oriented*, che può essere facilmente esteso dall'utente.

Per evitare noiosi errori che possono essere molto comuni per chi è abituato a lavorare in ambiente MS-DOS, è bene precisare subito che R, come tutti i linguaggi di derivazione UNIX, è *case sensitive*, cioè distingue tra lettere maiuscole e lettere minuscole.

Per facilitare la lettura, nel prosieguo di questo testo utilizzeremo un carattere rosso per indicare ciò che l'utente deve digitare al prompt di R (indicato con >) e utilizzeremo un carattere blu per indicare l'output di R.

Assegnazioni

R lavora con numeri, vettori e matrici, da assegnare a variabili con opportuni comandi. Ad esempio, il comando:

```
y <- 3
```

assegna il valore 3 alla variabile *y*. Invece il comando:

```
x <- c(1,2,3)
```

crea un vettore *x* contenente i numeri 1,2 e 3. Bisogna comunque precisare che con il termine vettore in R non ci si riferisce alla nozione usuale di vettore algebrico, m più semplicemente ad una stringa di valori consecutivi, rappresentati convenzionalmente da R in una riga.

Oltre a numeri e vettori, in R possiamo definire le matrici. Ad esempio il comando:

```
z <- matrix(c(1,2,3,4,5,6,7,8),2,4,byrow=TRUE)
```

crea una matrice *z* a 2 righe e 4 colonne, contenente i numeri da 1 a 8. La matrice viene riempita per riga.

Per visualizzare il contenuto di una variabile basta digitare il nome della variabile. Ad esempio:

```
> z
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    5    6    7    8
>
```

Le variabili possono essere create anche con opportune operazioni algebriche:

```
> f<-2*y
> f
[1] 6
>
```

Nel caso di una matrice, gli elementi possono essere richiamati con un opportuno utilizzo delle parentesi quadre:

```
> z[1,3]
[1] 3
>
```

Oggetti e Dataframe

Y, x, z ed f sono oggetti di R, che, essendo numerici, possono essere utilizzati per effettuare operazioni di ogni tipo. Tuttavia, l'oggetto più importante per i nostri fini è il DATAFRAME, nel quale possiamo memorizzare il database delle nostre osservazioni sperimentali in attesa di successive analisi. In questa sede partiremo dal presupposto di aver creato (come frequentemente avviene) il nostro database con EXCEL e di volerlo importare in R, registrandolo nel DATAFRAME *dati*.

Creiamo in EXCEL la tabella riportata di seguito (tab II.1), che si riferisce a 20 piante di mais. Salviamo questa tabella in un file di testo "import.dat". Per far questo scegliere Menù File, Salva con nome. Scegliere un nome per il file e indicare Tipo file = testo delimitato da tabulazione (*.txt). Salvare quindi il file in una directory prescelta (immaginiamo di avere scelto il nome import.txt). Avviare quindi una sessione R cambiare la directory predefinita del sistema, scegliendo con il menu File, Change Directory, la cartella nella quale abbiamo memorizzato il file di importazione.

Per leggere il file di testo nel DATAFRAME si usano i seguenti comandi:

```
> dati<-read.table("import.txt",header=TRUE)
```

Con questo comando, in R viene creato un dataframe di nome `dati`, contenente le tre colonne della tabella `import.dat` appena creata, comprese le intestazioni di colonna.

L'oggetto dataframe è disponibile in R, e può essere richiamato molto semplicemente digitandone il nome.

Tabella II.1. Esempio di un database da importare in R

Pianta	Varietà	Altezza
1	N	172
2	S	154
3	V	150
4	V	188
5	C	162
6	N	145
7	C	157
8	C	178
9	V	175
10	N	158
11	N	153
12	N	191
13	S	174
14	C	141
15	N	165
16	C	163
17	V	148
18	S	152
19	C	169
20	C	185

Infatti, si può osservare che digitando:

```
> dati
```

R risponde con il seguente output:

```
   Pianta Varieta Altezza
1       1       N     172
2       2       S     154
3       3       V     150
4       4       V     188
5       5       C     162
6       6       N     145
7       7       C     157
8       8       C     178
9       9       V     175
10     10       N     158
11     11       N     153
12     12       N     191
13     13       S     174
14     14       C     141
15     15       N     165
16     16       C     163
17     17       V     148
18     18       S     152
```

```
19      19        C      169
20      20        C      185
```

Per un veloce accesso al DATAFRAME, si può usare il comando FIX.
I dati nel DATAFRAME possono essere salvati in un file esterno:

```
> save(file="dati1.rda", dati)
```

ed eventualmente ricaricati:

```
> load("dati1.rda")
```

Per utilizzare i dati nel DATAFRAME, bisognerà accedere ai singoli vettori colonna che lo costituiscono. Per far questo si usa il comando *attach*, che crea immediatamente tre vettori (Pianta, Varietà e Altezza), che sono disponibili per le succesive elaborazioni.

Digitando infatti:

```
> attach(dati)
> Varieta
```

R risponde con il seguente output:

```
[1] N S V V C N C C V N N N S C N C V S C C
Levels: C N S V
```

Oppure digitando:

```
> Altezza
```

R risponde con il seguente output:

```
 [1] 172 154 150 188 162 145 157 178 175 158 153 191 174 141 165
163 148 152 169
[20] 185
>
```

Per caricare una parte di DATAFRAME in un altro DATAFRAME, si può usare il seguente metodo:

```
> dati2<-data.frame(X=Varieta, Y=Altezza)
> dati2
   X   Y
1  N 172
2  S 154
3  V 150
4  V 188
5  C 162
6  N 145
7  C 157
8  C 178
9  V 175
```

```
10 N 158
11 N 153
12 N 191
13 S 174
14 C 141
15 N 165
16 C 163
17 V 148
18 S 152
19 C 169
20 C 185
>
```

Workspace

Gli oggetti creati durante una sessione di lavoro vengono memorizzati nel cosiddetto workspace. Il contenuto di quest'ultimo può essere visualizzato:

```
> ls()
```

cancellato

```
> rm(list=ls())
```

salvato nella directory corrente:

```
> save.image("nomefile.RData")
```

e richiamato, per proseguire il lavoro dal punto in cui lo si è interrotto:

```
> load("nomefile.RData")
```

Script o programmi

Come è possibile memorizzare dati e workspace, è anche possibile scrivere programmi (procedure, funzioni...) da memorizzare e richiamare in seguito.

Nel caso più semplice, immaginiamo di voler scrivere un programma che, dato il valore della produzione rilevata in una parcella di orzo di 20 m^2 (in kg) e la sua umidità percentuale, calcoli automaticamente il valore della produzione secca in kg/ha.

La funzione che dobbiamo implementare è:

$$PS = PU \cdot \frac{100 - U}{100} \cdot \frac{10'000}{20}$$

ove PS è la produzione secca in kg/ha e PU è la produzione all'umidità U in kg per 20 m^2.
Scriveremo un file di testo (ad esempio con il *Block notes*):

```
PS <- function(PU, U) {
```

```
PU*((100-U)/100)*(10000/20)

}
```

Notare l'uso delle parentesi graffe. Registreremo il file di testo con il nome (esempio) "prova.r".

Aprendo una nuova sessione in R, possiamo ricaricare in memoria il file di programma ed utilizzarne la funzione somma, nel modo seguente:

```
> source("prova.r")
> ls()
[1] "PS"
> PS(20,85)
[1] 1500
```

Interrogazione di oggetti

A differenza di altri linguaggi statistici come SAS o SPSS, R immagazzina i risultati delle analisi negli oggetti, mostrando un output video piuttosto minimale. Per ottenere informazioni è necessario interrogare opportunamente gli oggetti che al loro interno possono contenere altri oggetti da cui recuperare le informazioni interessanti. Gli oggetti che contengono altri oggetti sono detti **liste**. Ad esempio, se vogliamo calcolare autovettori ed autovalori di una matrice, utilizziamo la funzione `eigen`. Questa funzione restituisce una lista di oggetti, che al suo interno contiene i due oggetti `values` (autovalori) e `vectors` (autovettori). Per recuperare l'uno o l'altro dei due risultati (autovettori o autovalori) si usa l'operatore di concatenamento `$`.

```
> matrice<-matrix(c(2,1,3,4),2,2)
> ev<-eigen(matrice)
> ev
$values
[1] 5 1

$vectors
           [,1]       [,2]
[1,] -0.7071068 -0.9486833
[2,] -0.7071068  0.3162278

> ev$values
[1] 5 1
> ev$vectors
           [,1]       [,2]
[1,] -0.7071068 -0.9486833
[2,] -0.7071068  0.3162278
>
```

Altre funzioni matriciali

Con R abbiamo la possibilità di gestire funzioni di matrice. Se ad esempio abbiamo le matrici:

$$Z = \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix} \qquad Y = \begin{pmatrix} 3 & 2 \end{pmatrix}$$

possiamo caricarle in R con i seguenti comandi:

```
> Z<-matrix(c(1,2,2,3),2,2)
> Y<-matrix(c(3,2)1,2)
```

Possiamo poi ottenere la trasposta di Z con il comando:

```
> t(Z)
     [,1] [,2]
[1,]    1    2
[2,]    2    3
```

Possiamo moltiplicare Y e Z utilizzando l'operatore %*%:

```
> Y%*%Z
     [,1] [,2]
[1,]    7   12
```

Possiamo calcolare l'inversa di Z con:

```
> solve(Z)
     [,1] [,2]
[1,]   -3    2
[2,]    2   -1
```

Cenni sulle funzionalità grafiche in R

R è un linguaggio abbastanza potente e permette di creare grafici piuttosto interessanti. Ovviamente un trattazione esauriente esula dagli scopi di questo testo, anche se è opportuno dare alcune indicazioni che potrebbero essere utili in seguito.

La funzione più utilizzata per produrre grafici è:

```
plot(x,y, type, xlab, ylab, col, lwd, lty...)
```

ove x ed y sono i vettori con le coordinate dei punti da disegnare. `Type` rappresenta il tipo di grafico ("p" produce un grafico a punti, "l" un grafico a linee, "b" disegna punti uniti da linee, "h" disegna istogrammi), `Title` disegna il titolo del grafico, `sub` il sottotitolo, `xlab` e `ylab` le etichette degli assi, `col` è il colore dell'oggetto, `lwd` il suo spessore, `lty` il tipo di linea e così via. Per una descrizione più dettagliata si consiglia di consultare la documentazione on line. A titolo di esempio mostriamo che i comandi

```
> x<-c(1,2,3,4)
> y<-(10,11,13,17)
> plot(x,y,"p",col="red",lwd=5,xlab="Ascissa",ylab="Ordinata")
```

producono come output il grafico in figura II.1:

Per sovrapporre un altro grafico al precedente una funzione possiamo utilizzare la funzione:

```
curve(funzione, Xiniziale, Xfinale, add=TRUE),
```

con il metodo `add`. Per aggiungere un titolo possiamo utilizzare la funzione:

```
title(main="Titolo"),
```

mentre per aggiungere una legenda utilizziamo la funzione:

```
> legend(Xcoord, YCoord , legend=c("Punti","X+10"), pch=c(19,-1),
col=c("Red","Blue"), lwd=c(3,3), lty=c(0,3))
```

ove i vettori indicano, per ogni elemento della legenda, il testo che deve essere riportato (`legend`), il tipo di simbolo (`pch`, con -1 che indica nessun simbolo), il colore (`col`), la larghezza (`lwd`) e il tipo di linea (`lty`, con 0 che indica nessuna linea).

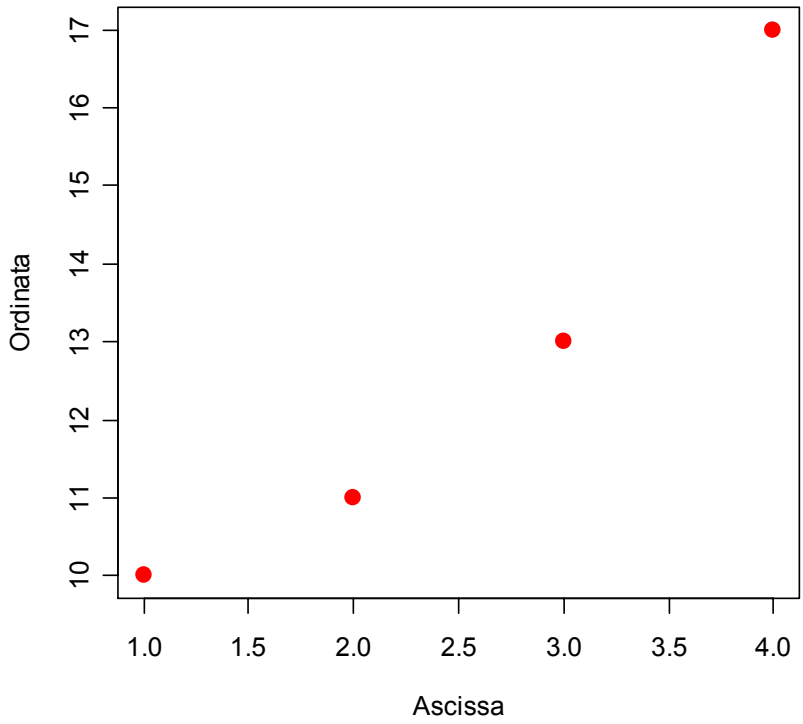


Figura II.1 Esempio di grafico con R

Ad esempio, dopo aver creato il grafico in figura II.1, utilizzando i comandi:

```
> curve(10+x,add=TRUE,lty=1,lwd=2,col="blue")
> title(main="Grafico di prova")
>     legend(1,17,     legend=c("Punti","X+10"),pch=c(19,-1),col=c("Red","Blue"),
lwd=c(3,3), lty=c(0,1))
>
```

possiamo ottenere il seguente grafico riportato in figura II.2

L'ultima cosa che desideriamo menzionare è la possibilità di disegnare grafici a torta, utilizzando il comando `pie(vettoreNumeri,vettoreEtichette,vettoreColori)`. Ad esempio il comando:

```
> pie(c(20,30,50),label=c("A","B","C"),col=c("blue","green","red"))
```

produce l'output riportato in figura II.3.

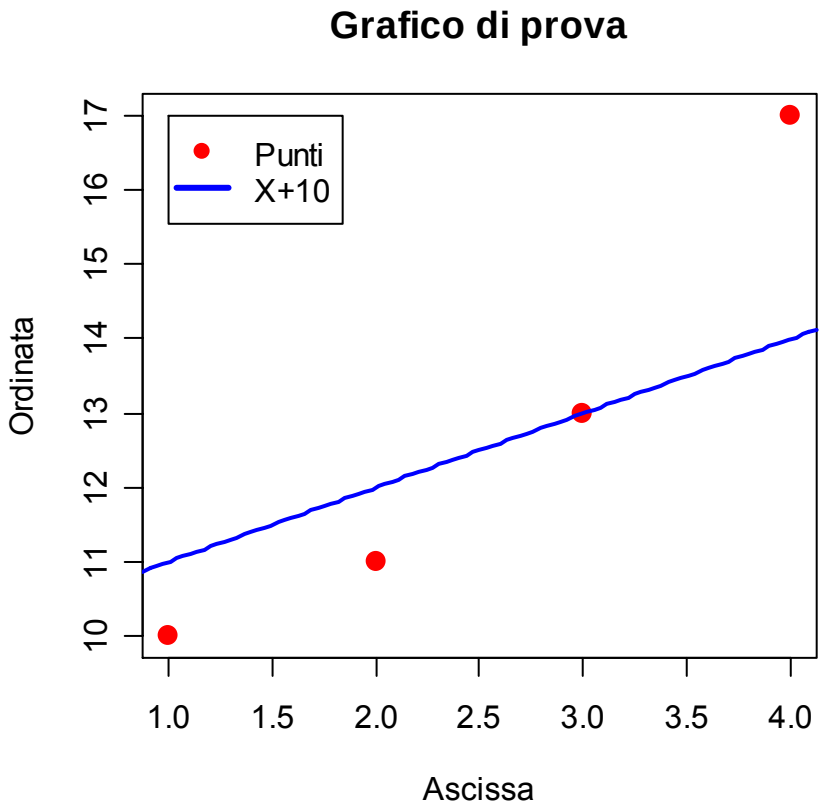


Figura II.2. Esempio di grafico multiplo con legenda in R.

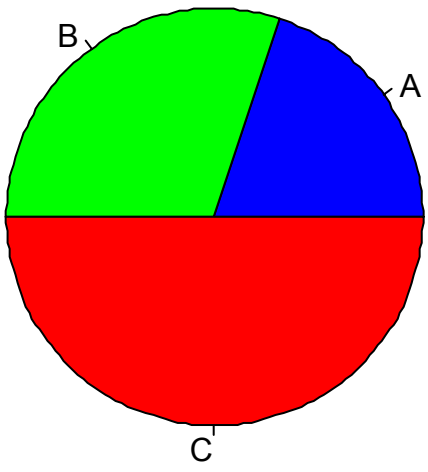


Figura II.3. Esempio di grafico di grafico a torta in R.

UNITA' III: LE STATISTICHE DESCRITTIVE

OBIETTIVO

Imparare le tecniche di base per l'analisi statistica descrittiva dei dati sperimentali. Sviluppare l'abilità di utilizzare R per il calcolo delle statistiche descrittive

SOMMARIO

1- Distribuzioni di frequenza
2 - Densità di frequenza
3 - Indici di tendenza centrale
4 - Calcolo delle medie di sottogruppi
5 - Indici di variabilità

SPIEGAZIONE

Abbiamo già visto che lo scopo della statistica descrittiva è quello di sintetizzare l'informazione contenuta nei dati grezzi, in modo da darne una lettura il più immediata possibile, senza però aver la pretesa di estendere alcuna considerazione a dati sperimentali diversi da quelli analizzati. Gli strumenti fondamentali che abbiamo a disposizione per questo fine sono le distribuzioni di frequenza, le loro rappresentazioni grafiche e i principali indici di sintesi che le descrivono.

Distribuzioni di frequenza

Avendo a che fare con un numero elevato di dati, è conveniente considerare le frequenze delle unità sperimentali: la *frequenza assoluta* non è altro che il numero degli individui che presentano una certa misura (per un carattere quantitativo) o una certa modalità (per un carattere qualitativo).

Ad esempio, se su 500 insetti 100 sono eterotteri, 200 sono imenotteri e 150 sono ortotteri, possiamo concludere che la frequenza assoluta degli eterotteri è pari a 100.

Se abbiamo a che fare con variabili quantitative su scala continua, prima di calcolare le frequenze è necessario suddividere l'intervallo delle misure in una serie di *classi di frequenza*.

Ad esempio, se abbiamo considerato 3000 piante di mais ed abbiamo osservato che 115 hanno altezze comprese tra 150 e 155 cm, possiamo conclude che la frequenza degli individui della classe 150-155 cm è pari a 115.

Oltre alle frequenze assolute, possiamo considerare anche le *frequenze relative*, che si calcolano dividendo le frequenze assolute per il numero totale degli individui del collettivo.

Nei casi prima accennati, la frequenza relativa degli eterotteri è pari a 100/500, cioè 0.2, mentre la frequenza relativa degli individui nella classe 150-155 è pari a 115/3000, cioè 0.038.

Se abbiamo una variabile quantitativa o comunque una variabile nella quale le modalità o le classi di frequenza possono essere logicamente ordinate, oltre alle frequenze assolute e relative possiamo prendere in considerazione le cosiddette *frequenze cumulate*, che si ottengono cumulando i valori di tutte le classi di frequenza che precedono quella considerata.

Ad esempio se tra le 3000 piante di mais anzidette 224 hanno altezze comprese tra 155 e 160 cm, la frequenza cumulata della classe è pari a 224+115 = 339, che si ottiene sommando alla frequenza assoluta di classe la frequenza assoluta della classe precedente.

Aggregare i dati in forma di distribuzioni di frequenza è estremamente conveniente, perché la lettura delle informazioni in essi contenute è molto più facile! Il prezzo da pagare è una lieve

perdita di informazione, come sarà chiaro nell'esercizio seguente.

CASO STUDIO III.1
In un campo di mais sono state rilevate su 20 piante le altezze e la varietà di ciascuna pianta (tab. III.1).

Tabella III. 1. Dati relativi al caso studio III.1

n. pianta	***Varietà***	***Altezza***
1	*N*	*172*
2	*S*	*154*
3	*V*	*150*
4	*V*	*188*
5	*C*	*162*
6	*N*	*145*
7	*C*	*157*
8	*C*	*178*
9	*V*	*175*
10	*N*	*158*
11	*N*	*153*
12	*N*	*191*
13	*S*	*174*
14	*C*	*141*
15	*N*	*165*
16	*C*	*163*
17	*V*	*148*
18	*S*	*152*
19	*C*	*169*
20	*C*	*185*

1 - valutare la distribuzione delle frequenze assolute, relative e percentuali degli individui di ciascuna varietà;
2 - valutare la distribuzione delle frequenze assolute, relative, percentuali assolute cumulate dell'altezza di tutti gli individui, considerando classi di ampiezza pari a 5 cm;
3 - Disegnare la torta delle frequenze relative della varietà e l'istogramma delle frequenze assolute dell'altezza.

La soluzione manuale del problema è banale e quindi viene lasciata al lettore. Si illustrerà invece l'elenco delle procedure R necessarie per la soluzione del problema.

1 -Supponiamo di aver caricato i dati in un dataframe e di aver utilizzato il comando `attach` *per rendere disponibile il vettore varietà (si veda l'unità precedente). Il comando R per ottenere le distribuzioni di frequenza è* `table:`

```
> table(Varieta)
Varieta
C N S V
7 6 3 4
>
```

Per ottenere le frequenze relative e percentuali si deve dividere ciascuna frequenza assoluta per il numero dei dati, che in R coincide con la lunghezza del vettore. Il comando è `lenght`.

```
> table(Varieta)/length(Varieta)
Varieta
   C    N    S    V
0.35 0.30 0.15 0.20
>
> table(Varieta)/length(Varieta)*100
Varieta
 C  N  S  V
35 30 15 20
>
```

2 - Per la variabile altezza, che è di tipo quantitativo, si utilizza lo stesso comando `table`*, ma occorre specificare l'ampiezza delle classi con la funzione* `cut` *e e il comando* `breaks`*, che specifica gli estremi superiori della classe (inclusi per default nella classe stessa). Per le frequenze cumulate si usa il comando* `cumsum`*.*

```
> table(cut(Altezza, c(140,150,160,170,190,200)))

(140,150] (150,160] (160,170] (170,190] (190,200]
        4         5         4         6         1

>table(cut(Altezza, (140,150,160,170,190,200)))/length(Altezza)

(140,150] (150,160] (160,170] (170,190] (190,200]
     0.20      0.25      0.20      0.30      0.05

>table(cut(Altezza, c(140,150,160,170,190,200)))/length(Altezza)*100

(140,150] (150,160] (160,170] (170,190] (190,200]
       20        25        20        30         5
> cumsum(table(cut(Altezza, c(140,150,160,170,190,200))))
[1]  4  9 13 19 20
```

3 - Per disegnare i grafici si usano i comandi `pie` *e* `plot`*.*

```
> pie(table(Varieta)/length(Varieta))
> plot(table(cut(altezza,
c(140,150,160,170,190,200))),lwd=10,col="blue",ylab="Frequenza")
>
```

Che producono l'output riportato in figura III.1

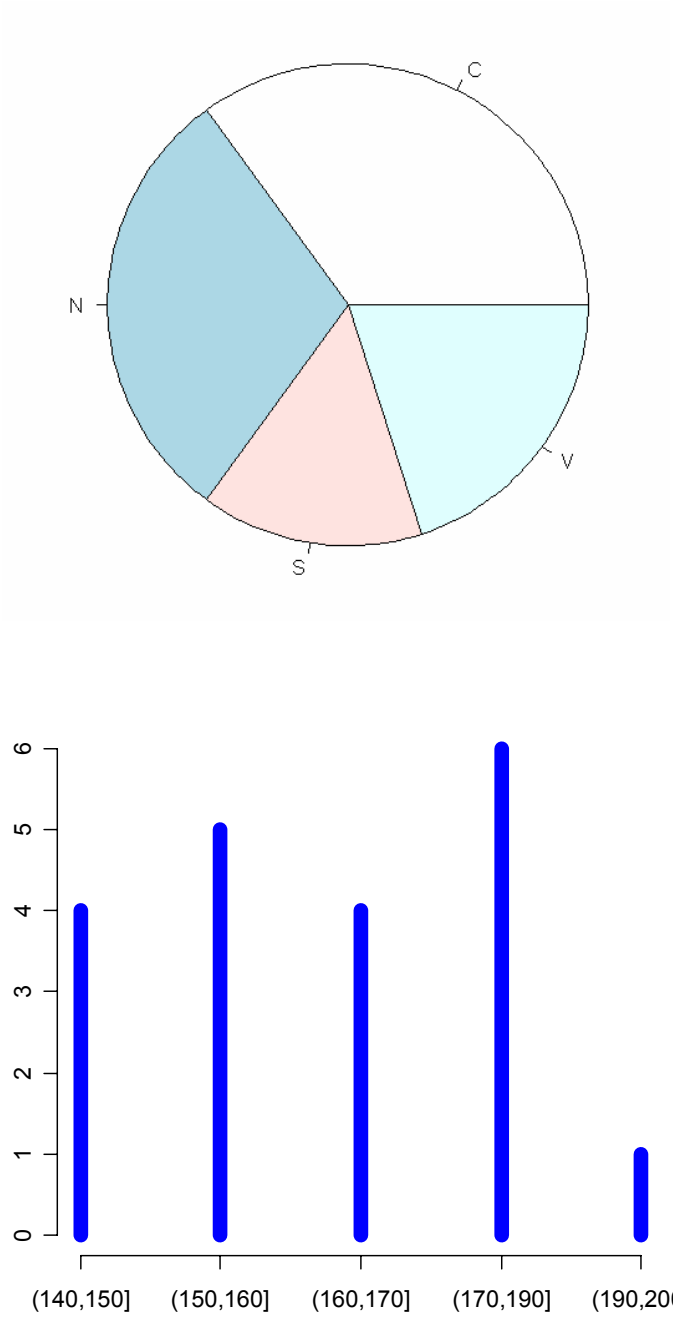


Figura III. 1. Grafici relativi al caso studio III.1

Densità di frequenza

Le densità di frequenza rappresentano la frequenza associata a ciascun punto dell'intervallo della classe. Si cerca in questo modo di evitare che classi molto ampie abbiano frequenze più alte di classi più rappresentative, ma molto strette. Ad esempio, Se ho due classi di altezza, la prima da 160 a 165 cm e la seconda da 165 a 175 cm e ho 5 individui nella prima classe e 5 nella seconda, è chiaro che la seconda classe contiene lo stesso numero di individui della prima, ma è molto più ampia. Se usiamo le sole frequenze non riusciamo ad evidenziare questo fatto, ma se dividiamo la frequenza di classe per l'ampiezza dell'intervallo otteniamo appunto la densità di frequenza:

$$d = \frac{n}{a}$$

che nel primo caso è pari a 1 e nel secondo caso è pari a 0.5, che è proprio l'informazione corretta di come si distribuiscono le unità all'interno degli intervalli.

Nel caso di R, le densità di frequenza vengono calcolate facendo riferimento alle frequenze relative:

$$d = \frac{f}{a}$$

e si ottengono con la funzione `hist`, che può anche essere utilizzata per disegnare gli istogrammi delle densità.

CASO STUDIO III.1 (segue)
4 - Calcolare le densità di frequenza per l'altezza delle 20 piante in tabella 1 e disegnarne i relativi istogrammi. Considerare le seguenti classi: (130-140], (140-160], (160-170], (170-190].

Se non lo abbiamo già fatto, carichiamo il dataframe "dati"e rendiamone disponibili i relativi vettori con il comando `attach`*. Utilizziamo il comando* `table` *per ottenere le distribuzioni di frequenza :*
In particolare, calcoliamo le frequenze assolute

```
> table(cut(Altezza, breaks=c(140,160,170,190)))

(140,160] (160,170] (170,190]
        9         4         6
>
```

Calcoliamo le densità di frequenza:

```
> hist(Altezza, c(130,160,170,200),plot=FALSE)
$breaks
[1] 130.0000 160.0000 170.0000 200.0000

$counts
[1] 9 4 7

$intensities
[1] 0.01499998 0.02000000 0.01166667

$density
[1] 0.01499998 0.02000000 0.01166667

$mids
[1] 145.0000 165.0000 185.0000

$xname
[1] "Altezza"

$equidist
[1] FALSE

attr(,"class")
[1] "histogram"
```

Disegnamo l'istogramma relativo

```
> hist(Altezza, c(130,160,170,200))
```

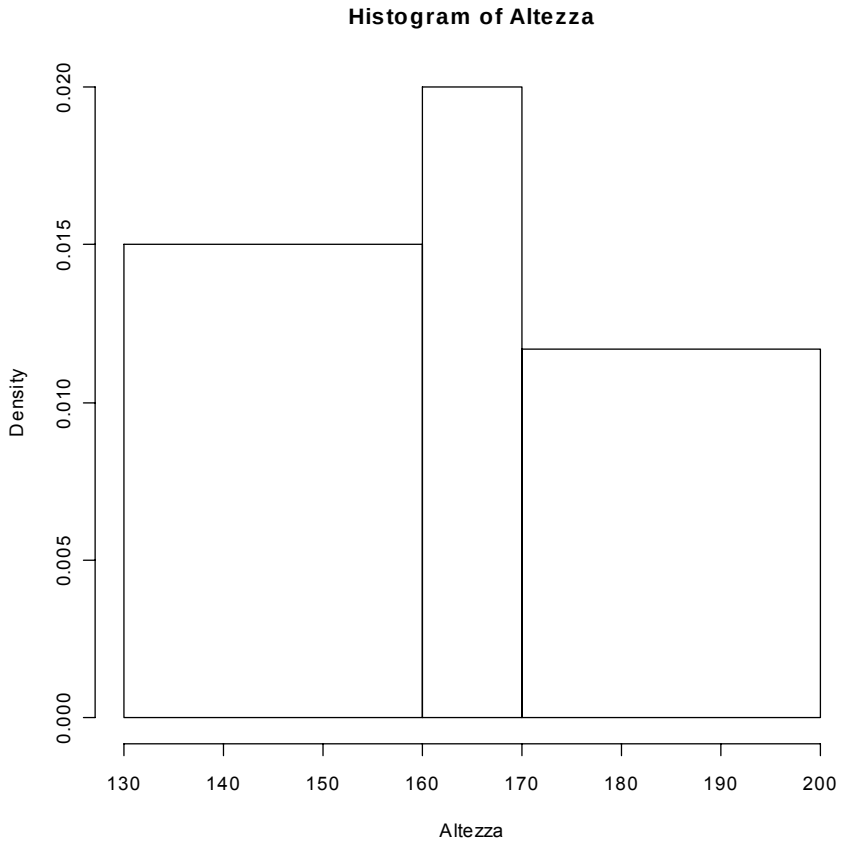


Figura III. 2. Grafico delle densità di frequenza delle altezze relative al caso studio III.1

E' da notare che la chiamata della funzione `hist`, restituisce una serie di informazioni, come le frequenze assolute e le densità di frequenza, che sono contraddistinte da una parola chiave preceduta dal segno $. queste parole chiave sono gli attributi di ogni proprietà dell'oggetto `hist` e ad esse si può accedere con l'operatore di concatenamento $:

```
> freq<-hist(Altezza, c(130,160,170,200),plot=FALSE)
> freq$counts
[1] 9 4 7
```

oppure:

```
> hist(Altezza, c(130,160,170,200),plot=FALSE)$counts
[1] 9 4 7
>
```

Indici di tendenza centrale: moda, mediana e media

E' possibile descrivere alcune caratteristiche dell'insieme dei dati, attraverso alcuni indici di statistica descrittiva. In particolare, è interessante adottare degli indici che misurino in qualche modo la tendenza centrale di una popolazione, cioè se esiste un valore attorno al quale si aggregano i dati.

Il più semplice indicatore di tendenza centrale, utilizzabile con qualunque tipo di dati è la *moda*,

cioè la classe che presenta la maggior frequenza. Ovviamente, se la variabile è quantitativa, si assume come moda il punto centrale della classe con maggior frequenza. Se le classi sono di diversa ampiezza si debbono considerare le densità di frequenza e non le frequenze.

L'individuazione della moda è banale e non richiede calcoli di sorta.

Nel caso di distribuzioni di frequenza per caratteri ordinabili (qualitativi e quantitativi), oltre alla moda possiamo calcolare la mediana, data dal valore che bipartisce la distribuzione di frequenza in modo da lasciare lo stesso numero di termini a sinistra e a destra.

Se abbiamo una serie di individui ordinati in graduatoria, la mediana è data dall'individuo che occupa il posto *(n + 1)/2* o, se gli individui sono in numero pari, dalla media delle due osservazioni centrali.

Il comando per calcolare la mediana in R è `median(vettore)`.

La mediana è legata al concetto di ripartizione ed è il primo di una serie di indicatori detti *quantili*, o, se parliamo di frequenze percentuali, *percentili*. Un percentile bipartisce la popolazione normale in modo da lasciare una certa quantità di termini alla sua sinistra e la restante quantità alla sua destra. I percentili sono 99: ad esempio il primo percentile bipartisce la popolazione in modo da lasciare a sinistra l'1% dei termini e alla destra il restante 99%. Allo stesso modo l'ottantesimo percentile bipartisce la popolazione in modo da lasciare a sinistra l'80% dei termini e alla destra il restante 20% (figura 1).

Per calcolare l'ottantesimo e il novantesimo percentile dell'altezza dei dati nel caso studio III.1, usiamo il comando `quantile()`.

```
> quantile(Altezza,probs=c(0.8,0.9))
  80%   90%
175.6 185.3
>
```

Collegato all'uso dei percentili possiamo introdurre il concetto di **boxplot**, che è una grafico a scatola, i cui estremi sono il 25 e il 75 percentile, tagliata da una linea centrale in corrispondenza della mediana e dotata di due linee verticali tagliate in corrispondenza di una linea orizzontale tracciata ad una distanza pari a 1.5 volte il 25 e il 75 percentile rispettivamente. Se tutti i dati rientrano negli intervalli così costituiti, le linee orizzontali si pongono in corrispondenza del valore inferiore o maggiore rispettivamente.

```
> boxplot(Altezza)
>
```

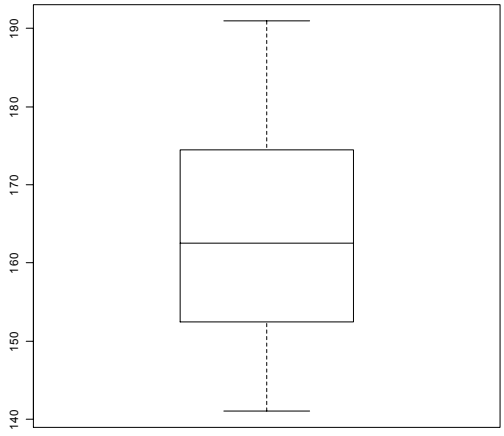


Figura III.2. Esempio di Boxplot

Nel caso di variabili quantitative, è possibile calcolare anche la *media aritmetica,* che è un concetto molto intuitivo ed esprime, in genere, quanta parte dell'intensità totale del fenomeno compete, in media, a ciascuna unità sperimentale. Si indica con μ e si calcola facendo la somma dei valori relativi alla variabile rilevata in tutti gli individui, e dividendola per il numero degli individui del collettivo.

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

Quando si ha a che fare con distribuzioni di frequenze, la media può essere calcolata moltiplicando il valore centrale di una classe per il numero degli individui che appartengono a quella classe, secondo la seguente espressione.

$$\mu = \frac{\sum_{i=1}^{n} f_i \cdot x_i}{n}$$

Il valore centrale di una classe è dato dalla semisomma degli estremi della classe stessa.
In R, la media si calcola con il comando `mean`.

```
> mean(Altezza)
[1] 164
>
```

Scelta dell'indicatore di tendenza centrale

La scelta dell'indice di tendenza centrale dipende dal tipo di dati con cui abbiamo a che fare, secondo la tabella III.2.

Tabella III.2. Scelte possibili per alcune statistiche descrittive

Indice	**Carattere**		
	qualitativo nominale	**qualitativo ordinale**	**quantitativo (continuo o discreto)**
moda	SI	SI	SI
mediana	NO	SI	SI
percentili	NO	SI	SI
media	NO	NO	SI
campo di variazione	NO	NO	SI

Quando possiamo utilizzare più indici (variabili quantitative) dobbiamo tener presente che la mediana è un indicatore più robusto della media. Infatti, supponiamo di avere cinque valori:

1 - 4 - 7 - 9 - **10**

La media è pari a 6.2, mentre la mediana è pari a 7 (valore centrale).

Se cambiano il numero più alto in questo modo:

1 - 4 - 7 - 9 - **100**
la media di questi cinque valori sarà 24.2, mentre la mediana sarà sempre pari a 7.

Calcolo delle medie di sottogruppi

In biometria è molto comune che il gruppo di unità sperimentali sia divisibile in più sottogruppi. Ad esempio potremmo avere un gruppo di nove unità sperimentali (parcelle di terreno) coltivate con tre livelli di concimazione azotata (tre parcelle per ogni livello di concimazione). Si può essere interessati a calcolare la media di ogni livello di concimazione, utilizzando il comando `by`. La sintassi di questo comando è:

```
by(var, indice, mean)
```

dove `var` è la variabile che contiene i valori da mediare, `indice` è la variabile che contiene la codifica di gruppo, `mean` è la funzione che dobbiamo calcolare. Ovviamente `mean` può essere sostituito da qualunque altra funzione che vogliamo calcolare per ciascun sottogruppo.

Esempio III.1

Considerare la seguente tabella di dati produttivi per tre livelli di concimazione azotata (N):

N=30 kg/ha	***N=60 kg/ha***	***N=90 kg/ha***
15	*18*	*21*
16	*22*	*24*
19	*19*	*25*

Calcolare le medie per ogni livello di concimazione.

In R:

```
> Prod<-c(15,16,19,18,22,19,21,24,25)
> Conc<-c(30,30,30,60,60,60,90,90,90)
> by(Prod,Conc,mean)
INDICES: 30
[1] 16.66667
------------------------------------------------------------
INDICES: 60
[1] 19.66667
------------------------------------------------------------
INDICES: 90
[1] 23.33333
>
```

La funzione `by` restituisce un oggetto che può essere considerato come un vettore di medie (nel caso dell'esempio formato da tre elementi) associato ad un vettore che contiene gli indici, considerati come nomi di riga. Questo ci consente di riutilizzare in calcoli successivi sia le medie, sia gli indici.

Esempio III.1 (segue)

Considerate le medie sopra indicate e caricatele in un vettore MedieProd.

Caricate i livelli di concimazione nel vettore MedieConc.

In R:

```
> medie<-by(Prod,Conc,mean)
> MedieProd<-as.vector(medie)
> MedieConc<-as.vector(row.names(medie))
> MedieProd
[1] 16.66667 19.66667 23.33333
> MedieConc
[1] "30" "60" "90"
```

Se vogliamo trasformare il vettore MediConc in un vettore di numeri anzichè di stringhe (come ce lo restituisce R), possiamo operare come segue

```
> MedieConcNum<-as.numeric(MedieConc)
> MedieConcNum
[1] 30 60 90
```

Indici di variabilità dei fenomeni: devianza, varianza, deviazione standard e coefficiente di variabilità

Gli indici di tendenza centrale non ci informano su come le unità sperimentali tendono ad assumere misure che sono diverse l'una dall'altra. In sostanza una media pari a 100 può essere ottenuta con tre individui che misurano 99, 100 e 101 rispettivamente o con tre individui che misurano 1, 100 e 199. E' evidente che in questo secondo gruppo gli individui sono molto più differenti tra loro (dispersi) che nel primo gruppo.

Quindi, quando si vuole descrivere un gruppo di unità sperimentali, è necessario utilizzare non solo un indice della tendenza centrale, ma anche un indice di variabilità, che ci consenta di stabilire come si colloca ogni singolo individuo rispetto alla tendenza centrale dell'insieme.

Il più semplice indice di variabilità è il *campo di variazione*, che è la differenza tra la misura più bassa e la misura più alta. In realtà, non si tratta di un vero e proprio indice di variabilità, in quanto dipende solo dai termini estremi della distribuzione e non necessariamente cresce al crescere della variabilità degli individui.

Esistono diversi indici di variabilità, tra cui i più diffusi sono la devianza, la varianza, la deviazione standard ed il coefficiente di variabilità.

L'indice SS:

$$SS = \sum_i (x_i - \mu)^2$$

costituisce la somma dei quadrati degli scarti (SS) ed è noto con il termine di *devianza*.

La devianza è un indicatore che ha un significato geometrico molto preciso, collegabile alla somma dei quadrati delle distanze euclidee di ogni osservazione rispetto alla media. Come misura di 'distanza', ha alcune importanti proprietà (che vedremo meglio in seguito), ma essendo la somma di scarti, il valore finale dipende dal numero di scarti da sommare e quindi non è possibile operare confronti tra collettivi formati da un diverso numero di individui.

Si può quindi definire un altro indice, detto **varianza** (nei software di uso più corrente si parla di **varianza della popolazione**), e definito come segue:

$$\sigma^2 = \frac{\sum_i (x_i - \mu)^2}{n}$$

In realtà, per motivi che saranno più chiari in seguito, più che la varianza di una popolazione, normalmente si usa un altro indicatore, detto **varianza campionaria** o più semplicemente **varianza** (senza ulteriori specifiche) ottenuto dividendo la devianza per il numero dei gradi di libertà (numero degli individui del collettivo meno uno):

$$s^2 = \frac{SS}{n-1} = \frac{\sum_i (x-\mu)^2}{n-1}$$

Per collettivi molto grandi i due indici pressappoco coincidono ed entrambi esprimono in qualche modo la distanza media al quadrato delle osservazioni rispetto alla media del collettivo. Entrambi consentono di confrontare la variabilità di collettivi formati da un numero diverso di individui, anche se permane il problema che la distanza tra le osservazioni e la media è su una scala al quadrato: ad esempio se le osservazioni sono espresse in metri, la varianza è espressa in metri quadrati.

Per eliminare questo problema si ricorre alla radice quadrata della varianza costituisce la *deviazione standard*, che si indica con σ.

$$\sigma = \sqrt{\sigma^2}$$

La deviazione standard è espressa nella stessa unità di misura dei dati originari ed è quindi molto informativa sulla banda di oscillazione dei dati rispetto alla media.

Un problema di questo indice è che spesso la variabilità dei dati è in qualche modo proporzionale alla media: collettivi con una media alta hanno anche una variabilità alta e viceversa. Per questo motivo viene utilizzato spesso il coefficiente di variabilità, che è un numero puro, che non dipende dall'unità di misura e dall'ampiezza del collettivo, sicché è molto adatto ad esprimere ad esempio l'errore degli strumenti di misura e delle apparecchiature di analisi:

$$CV = \frac{\sigma}{\mu} \times 100$$

CASO STUDIO III.2
Una varietà di frumento è stata saggiata in sei appezzamenti della Media Valle del Tevere, per verificarne la produttività. Le produzioni ottenute (in t ha^{-1}) sono state:
6.5 – 5.7 – 6.4 – 6.3 – 6.2 – 5.8
Valutare mediana, media, devianza, varianza, deviazione standard e coefficiente di variabilità.

La mediana è 6.25 (media aritmetica dei due termini centrali 6.2 e 6.3). La media è:

$$m = \frac{6.5 + 5.7 + 6.4 + 6.3 + 6.2 + 5.8}{6} = 6.15$$

La devianza è pari a:

$$SS = \frac{(6.5-6.15)^2 + (5.7-6.15)^2 + (6.4-6.15)^2 + (6.3-6.15)^2 + (6.2-6.15)^2 + (5.8-6.15)^2}{6} = 0.535$$

Il calcolo della varianza e della deviazione standard è banale.

Con R, il lavoro procede come segue:

1 - Immissione dati:

```
> prod<-c(6.5,5.7,6.4,6.3,6.2,5.8)
```

2 - Calcolo mediana

```
> median(prod)
[1] 6.25
```

3 - Calcolo media

```
> mean(prod)
[1] 6.15
```

4 - Calcolo varianza

```
> var(prod)
[1] 0.107
```

5 - Calcolo devianza

```
> var(prod)*(length(prod)-1)
[1] 0.535
```

6 - Calcolo deviazione standard

```
> sqrt(var(prod))
[1] 0.3271085
NB. Si può utilizzare anche la funzione sd(prod),
disponibile nelle nuove versioni di R
```

7 - Calcolo del coefficiente di variabilità

```
> sqrt(var(prod))/mean(prod)*100
[1] 5.318838
>
```

Alcuni dei risultati anzidetti possono essere facilmente ottenuti ricorrendo alla funzione `summary`

```
> summary(prod)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.700   5.900   6.250   6.150   6.375   6.500
>
```

Un esempio pratico

In un capitolo precedente abbiamo visto come il metodo fondamentale per fronteggiare l'imprecisione degli strumenti di misura è quello di effettuare più repliche della stessa misurazione. In questo modo ci si viene a trovare con un collettivo di misure. In questo caso si prenderà come risultato la media delle misure effettuate, che sarà associata ad un indicatore di variabilità che attesti in qualche modo l'errore della misura, o meglio la sua incertezza.

In genere, la deviazione standard, per le sue caratteristiche, viene utilizzata come indicatore dell'**incertezza assoluta** associata ad una determinata misurazione. Questa incertezza si riferisce soltanto **all'errore casuale**, del quale costituisce una stima affidabile.

L'incertezza può anche essere espressa in forma di coefficiente di variabilità (**incertezza relativa percentuale; CV**), che, come già detto, è molto adatto ad esprimere l'errore degli strumenti di misura e delle apparecchiature di analisi:

CASO STUDIO III.3
Un ricercatore ha eseguito sei repliche della stessa analisi chimica per valutare la concentrazione di un fitofarmaco nelle foglie di lattuga. Ha ottenuto i seguenti risultati:
101 – 126 – 97 – 117 – 121 – 94 nanogrammi per grammo
Stabilire la concentrazione del fitofarmaco e l'incertezza associata con questa determinazione. Qual è l'errore di misura dell'apparecchio utilizzato per l'analisi?

La concentrazione della sostanza nella lattuga è:

$$\frac{101+126+97+117+121+94}{6} = 109 \ \ ng/g$$

L'incertezza di questa stima è pari a:

$$\sqrt{\frac{(101-109)^2+(126-109)^2+(97-109)^2+(117-109)^2+(121-109)^2+(94-109)^2}{5}} = 13.6$$

Globalmente, possiamo concludere che la concentrazione è pari a:

109 ± 13.6 ng/g

L'errore di misura dell'apparecchio è pari al:

$$\frac{13.6}{109} \times 100 = 12.5\%$$

La soluzione con R è ottenuta analogamente all'esempio precedente

Arrotondamenti

Normalmente il calcolo della media e della deviazione standard (sia a mano che con il computer) porta all'ottenimento di un numero elevato di cifre decimali. E' quindi lecito chiedersi quante cifre riportare nel riferire i risultati della misura. Nel caso della media è poco corretto riportare un numero di cifre decimali superiore a quello rilevato nella misura. Ad esempio, se stiamo misurando l'altezza di una pianta con uno strumento capace di misurare i millimetri, non è giusto riportare un risultato come 1.9347 metri, perché la cifra finale non è stata effettivamente misurata, ma è solo frutto dei calcoli.

Si conviene invece che nel riportare le misure di variabilità si dovrebbe utilizzare un decimale in più.

RIEPILOGO

In questo capitolo abbiamo utilizzato le seguenti funzioni R:

table (vettoredati)	calcola le frequenze assolute di fenomeni qualitativi
table (cut(vettoredati, c(estremi di classe)))	frequenze assolute per variabili continue
pie(table(vettoredati))	disegna grafici a torta

plot(table(vettoredati))	disegna istogrammi
hist(vettoredati,vettore estremi di classe)	disegna istogrammi della densità di distribuzione)
median(vettoredati)	calcola la mediana
quantile(vettoredati, probs=c(prob1,prob2,...)	quantili
boxplot(vettoredati)	boxplot
min, max, range(vettoredati)	minimo, massimo e intervallo di variazione
mean(vettoredati)	media
summary(vettoredati)	alcune statistiche descrittive
var(vettoredati)	varianza campionaria
sqrt(var(vettoredati))	deviazione standard

ESERCIZI PROPOSTI

1) Su un campione di 500 olive si riscontra che 225 sono state attaccate da *Dacus oleae* (mosca dell'olivo). Stabilire: a) frequenza assoluta e relativa delle piante attaccate. (R: 225; 0.45)
2) Alla fioritura del mais si sono misurate le altezze di dieci piante. Le misure ottenute sono:
   150, 160, 155, 154, 172, 137, 136, 148, 155, 157
   Determinare: a) media; b) devianza, varianza e deviazione standard. (R: 152.4; 1010.4; 112.27; 10.596)
3) Un ricercatore sta eseguendo uno studio sulle produzioni di un vigneto di Sangiovese. Per questo motivo, ha misurato la produzione unitaria di 500 piante, ottenendo la seguente distribuzione di frequenze assolute.

Classi (kg/pianta)	Frequenze assolute
2, - 2,5	21
2,5 – 3	46
3 – 3,5	78
3,5 - 4	102
4 - 4,5	106
4,5 - 5	69
5 - 5,5	51
5,5 - 6	27

Stabilire:
1) la frequenza relativa della classe di produzione da 4 a 4,5 kg/pianta;
2) la frequenza cumulata della classe da 3 a 3,5 kg/pianta;
3) media, varianza, deviazione standard e coefficiente di variabilità;
4) in quale percentile si trova una pianta che produce 5,5 kg/pianta?

UNITA' IV: RELAZIONI TRA VARIABILI

OBIETTIVO

Sviluppare l'abilità di utilizzare R per l'analisi statistica descrittiva delle relazioni tra variabili sperimentali (dipendenza, dipendenza in media, correlazione e regressione).

SOMMARIO

1- Due variabili qualitative: tabelle di contingenza e dipendenza
2 - Due variabili quantitative: correlazione e regressione

SPIEGAZIONE

In alcuni casi in ciascuna unità sperimentale del collettivo vengono studiati due (o più) caratteri e, di conseguenza, si ha a che fare con distribuzioni di frequenza bivariate. Procedendo seconda quanto detto in precedenza, è possibile calcolare separatamente per ciascuna delle due variabili gli indici di statistica descrittiva finora accennati (media, varianza, deviazione standard ecc…). In questo modo è possibile avere un'ottima descrizione di ognuna delle due variabili, ma non è possibile avere informazioni sulle relazioni o sui legami esistenti tra loro; ad esempio non è possibile sapere come si comporta una variabile man mano che l'altra cambia di valore.

E' quindi utile avere la possibilità di calcolare degli indici statistici che descrivano in qualche modo le relazioni esistenti tra le due variabili.

Due variabili quantitative: tabelle di contingenza ed indici di dipendenza (connessione)

Indipendentemente dal tipo di variabili in studio, quando si ha a che fare con un numero notevole di individui è possibile costruire delle **tabelle di contingenza:** si tratta di tabelle a due entrate nelle quali ogni numero rappresenta la **frequenza congiunta** (in genere assoluta) per una particolare coppia di valori delle due variabili. Ad esempio consideriamo le variabili di fantasia X=Varietà (con i valori SANREMO e FANO) e Y=Forma delle bacche (con i valori LUNGO, TONDO, OVALE), nella tabella sottostante il valore 37 indica il numero di individui che presentano congiuntamente la modalità SANREMO e la modalità LUNGO. I totali mostrano le **frequenze marginali** delle due variabili separatamente.

Tabella IV. 1. Esempio di tabella di contingenza

	LUNGO	**TONDO**	**OVALE**	**Totale**
SANREMO	37	32	61	**130**
FANO	45	74	59	**178**
Totale	**82**	**106**	**120**	**308**

Ogni riga della tabella di cui sopra (esclusi i totali) costituisce una distribuzione condizionata della variabile Y, dato un certo valore della X (Y|SANREMO e Y|FANO). Viceversa ogni colonna (X|LUNGO, X|TONDO e X|OVALE).

In simboli:

	$\mathbf{Y_1}$	...	$\mathbf{Y_j}$	...	$\mathbf{Y_k}$	
X_1	n_{11}	...	n_{1j}	...	n_{1k}	$\mathbf{n_{1.}}$
...	...	...		...	...	
X_i	n_{i1}	...	n_{ij}	...	n_{ik}	$\mathbf{n_{1.}}$
...	...	...		...	...	
$\mathbf{X_h}$	n_{h1}	...	n_{hj}	...	n_{hk}	
Totale	$\mathbf{n_{.1}}$	...	$\mathbf{n_{.j}}$	...	$\mathbf{n_{.k}}$	**308**

Dipendenza

Se guardiamo le due distribuzioni condizionate Y|SANREMO e Y|FANO possiamo notare che esiste una certa differenza. Potremmo chiederci quindi se il presentarsi di una data modalità del carattere X (SANREMO o FANO) influenza il presentarsi di una particolare modalità del fenomeno Y. Se ciò non è vero si parla di **indipendenza** delle variabili (allora le distribuzioni condizionate sono uguali) altrimenti si parla di dipendenza o **connessione**.

In caso di indipendenza, le distribuzioni condizionate di Y dovrebbero essere uguali tra loro e alla distribuzione marginale di X. In simboli:

$$\frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n}$$

il che equivale a dire che la frequenza relativa condizionale di X per una data modalità di Y deve essere uguale alla frequenza relativa marginale di X. In sostanza, se esiste indipendenza, ogni valore dentro alla tabella di contingenza dovrebbe essere pari a:

$$\bar{n}_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}$$

Possiamo quindi ricostruire la tabella, delle frequenze assolute attese, in caso di indipendenza completa

	LUNGO	**TONDO**	**OVALE**	**Totale**
SANREMO	$\frac{82 \times 130}{308} = 34.6$	$\frac{106 \times 130}{308} = 44.7$	50.6	**130**
FANO	$\frac{82 \times 178}{308} = 47.4$	61.3	69.4	**178**
Totale	**82**	**106**	**120**	**308**

A questo punto è logico costruire un indice statistico di connessione, detto χ^2 che misuri lo scostamento tra le frequenze osservate e quelle attese nell'ipotesi di indipendenza perfetta:

$$\chi^2 = \sum \frac{(f_o - f_a)^2}{f_a}$$

dove f_o sta per frequenza osservata ed f_a sta per frequenza attesa nel caso indipendenza. Questo indice assume valore pari a zero nel caso di indipendenza completa (le frequenze osservate sono uguali a quelle attese) ed assume un valore positivo tanto più alto quanto maggiore è la connessione

tra i due caratteri.

Nel caso in esame:

$$\chi^2 = \frac{(37-34.6)^2}{34.6} + \frac{(32-44.7)^2}{44.7} + ... = 10.22$$

Per valutare il significato del valore ottenuto, nel campo della statistica descrittiva si suole dividere l'indice per il suo valore massimo, che è proporzionale al numero di righe e di colonne della tabella:

$$\max \chi^2 = n \cdot \min(h-1, k-1)$$

Nel nostro caso il massimo è pari a:

$$\max \chi^2 = 308 \cdot \min(3-1, 2-1) = 308$$

In sostanza, l'indice ottenuto è pari a 10.22/308=0.033; il valore risultante è piuttosto vicino a zero e ci fa sospettare che la connessione tra i due caratteri non è molto elevata.

Per il calcolo dell'indice di connessione con R, dobbiamo prima creare una tabella di contingenza, con il comando `table` che abbiamo già utilizzato per costruire le distribuzioni di frequenza. A questa tabella di contingenza applicheremo poi il comando `summary`.

CASO STUDIO IV.1
Immaginiamo di aver rilevato su venti piante di frumento appartenenti a quattro varietà (N, S, V e C) la presenza o l'assenza di una determinata malattia. Vogliamo vedere se il carattere varietà è connesso con la presenza della malattia.

Tabella IV.2. Dati relativi al caso studio IV. 1.

Pianta	Varietà	Attacco fungino
1	N	SI
2	S	SI
3	V	NO
4	V	NO
5	C	NO
6	N	SI
7	C	NO
8	C	SI
9	V	SI
10	N	SI
11	N	NO
12	N	NO
13	S	NO
14	C	NO
15	N	NO
16	C	SI
17	V	SI
18	S	NO
19	C	NO
20	C	SI

Se immaginiamo che i dati sono stati già registrati in formato testo (file

"caso2.dat", il lavoro procede come segue:

1 - Cambiamo directory di lavoro, con il menu FILE(CHANGE DIR)

2 - Leggiamo il file nel dataframe connessione:

```
> connessione<-read.table("caso2.dat",header=TRUE)
>
```

3 - rendiamo disponibili le colonne dei dati

```
> attach(connessione)
>
```

4 - creiamo la tabella di contingenza

```
> contingenza<-table(Varieta,Attacco)
> contingenza
        Attacco
Varieta NO SI
      C  4  3
      N  3  3
      S  2  1
      V  2  2
>
```

5 - Calcoliamo il χ2

```
> summary(contingenza)
Number of cases in table: 20
Number of factors: 2
Test for independence of all factors:
        Chisq = 0.27898, df = 3, p-value = 0.964
        Chi-squared approximation may be incorrect>
```

Come si può vedere, il comando `summary` *restituisce, tra l'altro, il valore dell'indice di connessione, da riutilizzare per effettuare un calcolo successivo. Per capire come estrarlo, utilizziamo il comando* `str`.

```
> str(summary(contingenza))
List of 7
 $ n.vars   : int 2
 $ n.cases  : int 20
 $ statistic: num 0.279
 $ parameter: num 3
 $ approx.ok: logi FALSE
 $ p.value  : num 0.964
 $ call     : NULL
 - attr(*, "class")= chr "summary.table"
```

Vediamo che il valore che ci interessa è contenuto in `$ statistic`. *Questo parametro potrà essere utilizzato congiuntamente al comando* `summary`, *come segue.*

```
> chi<-summary(contingenza)$statistic
> chi
```

```
[1] 0.2789803

> chi/(length(Varieta)*(min(dim(contingenza)-1)))
[1] 0.01394901
```

Possiamo concludere che la connessione è molto debole.

Se non abbiamo i dati, ma abbiamo una semplice tabella di contingenza, possiamo caricarla in una matrice ed utilizzare il comando `as.table` per informare il sistema sul fatto che quella matrice è in realtà una tabella di contingenza.

```
> cont<-matrix(c(4,3,2,2,3,3,1,2),4,2)
> cont
     [,1] [,2]
[1,]    4    3
[2,]    3    3
[3,]    2    1
[4,]    2    2
> colnames(cont)<-c("SI","NO")
> rownames(cont)<-c("C","N","S","V")
> cont
  SI NO
C  4  3
N  3  3
S  2  1
V  2  2
> summary(as.table(cont))
Number of cases in table: 20
Number of factors: 2
Test for independence of all factors:
        Chisq = 0.27898, df = 3, p-value = 0.964
        Chi-squared approximation may be incorrect
```

Variabili quantitative

Se abbiamo a che fare con variabili quantitative, possiamo calcolare l'indice di connessione ricorrendo a divisioni in classi di frequenza opportunamente scelte. Oltre a ciò, con variabili quantitative è possibile esplorare l'esistenza di altri tipi di relazioni tra variabili, di cui sono estremamente importanti le seguenti:

1) *variazione congiunta (covariazione*): si ha quando al variare di una variabile cambia il valore dell'altra in modo abbastanza analogo, ma senza che si possa in qualche modo stabilire un nesso causale tra una variabile e l'altra;
2) *dipendenza*: si ha quando una variabile (detta dipendente) è funzione dell'altra (detta indipendente). In questo modo tra le variabili si può stabilire un nesso diretto causa-effetto.

Ad esempio, su una popolazione di piante di mais si potrebbe misurare (a) l'altezza delle piante e la lunghezza delle foglie. Oppure su una popolazione di piante di pomodoro si potrebbe misurare (b) la produzione di bacche e la quantità di concime utilizzata da ogni pianta. Oppure ancora si potrebbe su una serie di vini diversi si potrebbe misurare (c) la gradazione alcolica e il contenuto in zucchero dell'uva prima della pigiatura.

Emerge una differenza fondamentale tra i tre esempi riportati. Nel caso dell'esempio (a) ci può aspettare che piante di mais più alte abbiano anche foglie più lunghe, ma è evidente che non è

possibile stabilire una relazione funzionale di dipendenza tra una variabile e l'altra. In altre parole, è l'altezza delle piante che dipende dalla lunghezza delle foglie o viceversa? Probabilmente ne' l'una ne' l'altra cosa! In questo caso si può solo parlare di variazione congiunta, non di dipendenza. Ciò non è vero per gli esempi (b) e (c): è infatti evidente come la produzione del pomodoro (variabile dipendente) dipende direttamente dalla dose di concime (variabile indipendente) e come la gradazione del vino (variabile dipendente) dipende dal contenuto in zucchero dell'uva (variabile indipendente).

Nel caso dell'esempio (a), il ricercatore è interessato a stabilire l'entità della variazione congiunta delle due variabili rilevate, mentre nei casi (b) e (c) il ricercatore potrebbe essere interessato a definire l'equazione matematica che lega la variabile dipendente alla variabile indipendente. Il primo problema è risolvibile mediante analisi di CORRELAZIONE, mentre il secondo problema è risolvibile mediante analisi di REGRESSIONE.

Coefficiente di correlazione

Un indicatore statistico per descrivere il grado di variazione congiunta di due variabili è il *coefficiente di correlazione*. Il calcolo è abbastanza semplice: dato un collettivo statistico composto da *n* unità sperimentali, sulle quali sono state rilevate due variabili statistiche ($X1_i$ e $X2_i$ con *i* che va da 1 ad *n* e medie rispettivamente pari a μ_{X1} e μ_{X2}), definiamo *coefficiente di correlazione* (r), la misura:

$$r = \frac{\sum_{i=1}^{n}\left[(X1_i - \mu_{x1})(X2_i - \mu_{x2})\right]}{\sqrt{\sum_{i=1}^{n}(X1_i - \mu_{x1})^2 \sum_{i=1}^{n}(X2_i - \mu_{x2})^2}}$$

La quantità al numeratore viene detta *codevianza (*o *somma dei prodotti)*, mentre si può notare che al numeratore, sotto radice, abbiamo il prodotto delle devianze delle due variabili.

Il coefficiente di correlazione varia tra –1 e +1 (la dimostrazione di questa proprietà non è necessaria): un valore pari a +1 indica concordanza perfetta (tanto aumenta una variabile, tanto aumenta l'altra), mentre un valore pari a –1 indica discordanza perfetta (tanto aumenta una variabile tanto diminuisce l'altra). Un valore pari a 0 indica assenza di qualunque grado di variazione congiunta tra le due variabili (assenza di correlazione). Valori intermedi tra quelli anzidetti indicano correlazione positiva (se positivi) e negativa (se negativi).

In R, per calcolare covarianza (pari alla codevianza diviso n-1) e correlazione tra due variabili si usano molto semplicemente le funzioni `cov` e `cor`

CASO STUDIO IV.2
Il contenuto di olio degli acheni di girasole è stato misurato con due metodi diversi; le misurazioni sono stata eseguite su quattro campioni. I risultati (espressi in percentuale) sono come segue:

N° campione	*Metodo 1*	*Metodo 2*
1	*46*	*45*
2	*47*	*49*
3	*49*	*51*
4	*51*	*49*

Verificare se esiste una buona concordanza tra i due tipi di analisi.

Questo tipo di problema può essere risolto mediante analisi di correlazione, in quanti si tratta di descrivere (misurare) il grado di variazione congiunta delle due variabili misurate su ognuna delle unità sperimentali (i campioni analizzati).
Per motivi di comodità, converrà organizzare il calcolo in tre fasi. In primo luogo è conveniente calcolare le statistiche descrittive della variabile X1 (media e devianza).

N° campione	$X1_i$	$X1_i - \mu_{X1}$	$(X1_i - \mu_{X1})^2$
1	46	46-48.25=-2.25	5.0625
2	47	47-48.25=-1.25	1.5625
3	49	49-48.25=0.75	0.5625
4	51	51-48.25=2.75	7.5625
Media =48.25		Devianza =14.75	

In secondo luogo possiamo calcolare le stesse statistiche per la variabile X2.

N° campione	$X2_i$	$X2_i - \mu_{X2}$	$(X2_i - \mu_{X2})^2$
1	45	45-48.5=-3.5	12.25
2	49	49-48.5=0.5	0.25
3	51	51-48.5=2.5	6.25
4	49	49-48.5=0.5	0.25
Media =48.5		Devianza =19.00	

In terzo luogo possiamo calcolare la codevianza, moltiplicando tra loro gli scostamenti dalla media delle due variabili

N° campione	$X1_i - \mu_{X1}$	$X2_i - \mu_{X2}$	Prodotto
1	-2.25	45-48.5=-3.5	7.875
2	-1.25	49-48.5=0.5	-0.625
3	0.75	51-48.5=2.5	1.875
4	2.75	49-48.5=0.5	1.375
		Codevianza =10.5	

A questo punto possiamo calcolare il coefficiente di correlazione semplice:

$$r = \frac{10.5}{\sqrt{14.75 \times 19}} = 0.6272$$

Possiamo osservare che r si trova approssimativamente a metà strada tra 1 (correlazione positiva perfetta) e 0 (assenza di correlazione). In questo senso possiamo concludere che esiste un certo grado di concordanza tra i due metodi di analisi, ma esso non deve essere considerato particolarmente buono.
In R, l'analisi procede nel seguente modo:

```
> correlazione<-read.table("Caso3.dat",header=TRUE)
```

```
> correlazione
  Camp Met1 Met2
1    1   46   45
2    2   47   49
3    3   49   51
4    4   51   49
> attach(correlazione)
> cov(Met1,Met2)
[1] 3.5
> cor(Met1,Met2)
[1] 0.6272151
>
```

Se vogliamo disegnare un grafico di dispersione usiamo la funzione `plot`

```
>plot(Met1,Met2)
```

che disegna in modo molto grezzo il grafico desiderato.

Analisi di regressione

In alcuni casi le due variabili rilevate sulle unità sperimentali sono tali che possiamo ipotizzare che una relazione di dipendenza diretta, sulla base di considerazioni biologiche, sociali, chimiche, fisiche ecc… In sostanza, è possibile individuare una *variabile dipendente* (detta anche *variabile regressa*) e una *variabile indipendente* (detta anche *regressore*).

In questo caso, la conoscenza del semplice grado di correlazione tra le due variabili può non essere sufficiente per i nostri scopi, mentre potrebbe essere necessaria la conoscenza diretta della funzione matematica che lega la variabile dipendente alla variabile indipendente. *In questa sede, per motivi di semplicità, restringiamo il nostro interesse alle funzioni lineari e, in particolare, all'equazione di una retta*.

Nel momento in cui ipotizziamo che tra le due variabili esiste una relazione lineare, rappresentabile con una linea retta di equazione generica:

$$Y = mX + q$$

o meglio (in statistica):

$$Y = b_1 X + b_0$$

il problema è ridotto alla determinazione dei valori di b_1 (detto in statistica *coefficiente di regressione*) e b_0 che sono rispettivamente la pendenza della retta e l'intercetta (intersezione con l'asse delle Y).

L'esigenza di fare una analisi di regressione si presenta, in genere, perché vogliamo essere in grado di prevedere i valori della Y qualunque sia il valore della X (o viceversa).

Il problema sarebbe assolutamente banale se i punti fossero perfettamente allineati, il che non si verifica mai in statistica, almeno per due motivi:

1) le relazioni biologiche non sono quasi mai perfettamente lineari, ma lo sono solo approssimativamente;
2) le variabili osservate sulle unità sperimentali fluttuano a causa del possibile errore sperimentale.

E'quindi necessaria una procedura di interpolazione, che viene eseguita analiticamente ricorrendo alle formule seguenti (n è il numero di unità sperimentali mentre μ_X e μ_Y sono le medie

delle due variabili):

$$b_1 = \frac{\sum_{i=1}^{n}[(X_i - \mu_X)(Y_i - \mu_Y)]}{\sum_{i=1}^{n}(X_i - \mu_X)^2}$$

$$b_0 = \mu_Y - b_1\mu_X$$

Per arrivare alle due formula anzidette bisogna considerare che la retta dei minimi quadrati è quella che rende appunto minimi gli scostamenti tra i valori osservati ed i valori attesi; è cioè necessario minimizzare le quantità:

$$Q = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2 = \sum_{i=1}^{n}(Y_i^2 + \beta_0{}^2 + \beta_1{}^2 X_i^2 - 2Y_i\beta_0 - 2Y_i X_i\beta_1 + 2\beta_0\beta_1 X_i)^2 =$$

$$= \sum_{i=1}^{n} Y_i^2 + n\beta_0{}^2 + \beta_1{}^2\sum_{i=1}^{n} X_i^2 - 2\beta_0\sum_{i=1}^{n} Y_i - 2\beta_1\sum_{i=1}^{n} Y_i X_i + 2\beta_0\beta_1\sum_{i=1}^{n} X_i$$

Calcolando le derivate parziali (rispetto a β_0 e β_1 che al momento sono le nostre incognite), si ottiene:

$$\frac{dQ}{d\beta_0} = 2n\beta_0 - 2\sum Y_i + 2\beta_1\sum X_i$$

$$\frac{dQ}{d\beta_1} = 2\beta_1\sum X_i^2 - 2\sum X_i Y_i + 2\beta_0\sum X_i$$

Ricordando che la funzione Q raggiunge il suo minimo quando la derivata prima è uguale a zero, otteniamo il seguente sistema:

$$\begin{cases} n\beta_0 + \beta_1\sum X_i = \sum Y_i \\ \beta_1\sum X_i^2 + \beta_0\sum X_i = \sum X_i Y_i \end{cases}$$

Risolvendo per β_0 la prima equazione si ottiene:

$$\beta_0 = \frac{\sum Y_i}{n} - \beta_1 \cdot \frac{\sum X_i}{n}$$

sostituendo nella seconda equazione, si ottiene:

$$\beta_1\sum X_i^2 + \left(\frac{\sum Y_i}{n} - \beta_1 \cdot \frac{\sum X_i}{n}\right)\sum X_i = \sum X_i Y_i$$

da cui:

$$\beta_1\sum X_i^2 + \left(\frac{\sum Y_i \cdot \sum X_i}{n} - \beta_1 \cdot \frac{\left(\sum X_i\right)^2}{n}\right) = \sum X_i Y_i$$

$$\beta_1 \sum X_i^2 - \beta_1 \cdot \frac{\left(\sum X_i\right)^2}{n} = \sum X_i Y_i - \frac{\sum Y_i \cdot \sum X_i}{n}$$

$$\beta_1 \left(\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n} \right) = \sum X_i Y_i - \frac{\sum Y_i \cdot \sum X_i}{n}$$

$$\beta_1 = \frac{\sum X_i Y_i - \frac{\sum Y_i \cdot \sum X_i}{n}}{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}}$$

si noti che:

$$SQ(X) = \sum (X_i - \overline{X})^2 = \sum (X_i^2 - 2X_i\overline{X} + \overline{X}^2) = \sum X_i^2 - 2n\overline{X}\sum X_i + n\overline{X}^2$$

essendo:

$$\overline{X} = \frac{\sum X_i}{n}$$

otteniamo:

$$SQ(X) = \sum X_i^2 - 2\frac{\sum X_i}{n}\sum X_i + n\left(\frac{\sum X_i}{n}\right)^2$$

$$= \sum X_i^2 - 2\left(\frac{\sum X_i}{n}\right)^2 + \left(\frac{\sum X_i}{n}\right)^2 = \sum X_i^2 + \left(\frac{\sum X_i}{n}\right)^2$$

Analogamente si dimostra che:

$$SP(XY) = \sum \left[(X_i - \overline{X})(Y_i - \overline{Y}) \right] = \sum X_i Y_i - \frac{\sum Y_i \cdot \sum X_i}{n}$$

Per cui:

$$\beta_1 = \frac{SP(XY)}{SQ(X)}$$

Quando questo calcolo viene eseguito con l'aiuto del computer, l'output dell'analisi comprende in genere un indicatore detto *coefficiente di determinazione* (R^2). Questo indicatore numericamente è il quadrato del coefficiente di correlazione lineare, ma concettualmente indica la quota parte della variabilità della Y che è attribuibile alla dipendenza lineare dalla X; in sostanza si tratta di un indicatore della bontà della regressione: più è vicino ad 1 e più la regressione è buona.

Un altro metodo per valutare la bontà dell'adattamento è quello di guardare un grafico dei residui, che riporta gli scostamenti tra i valori osservati e quelli attesi (residui) in funzione dei valori attesi corrispondenti (vedi esempio sottostante). In una regressione adeguata i residui non debbono essere troppo ampi, debbono essere in parte positivi e in parte negativi, ma distribuiti casualmente nel grafico.

In R, l'analisi di regressione viene facilmente effettuata utilizzando la funzione `lm`. Il valore di R^2 si ottiene applicando il comando `summary` all'output di `lm` e vari tipi di grafico si ottengono aplicando il comando `plot` all'output di `lm`. Per il calcolo dei valori predetti si utilizza la funzione

`predict`, applicata ad un `data.frame`.

CASO STUDIO IV.3
Un diserbante (una sostanza chimica che riduce lo sviluppo delle piante) utilizzata a quattro dosi crescenti ha ridotto lo sviluppo di una pianta infestante come indicato più sotto.

Dose di erbicida (g/ha)	*Peso delle piante infestanti (%)*
5	*91*
10	*61*
15	*54*
20	*29*

Calcolare dose media, peso medio, deviazione standard del peso, errore standard. Fare un grafico a dispersione del peso vs la dose.
Calcolare la dose richiesta per inibire del 50% lo sviluppo della pianta trattata (ED50). Calcolare il peso delle piante trattate con una dose pari a 12.5 g/ha di erbicida. Calcolare R^2.

Le unità sperimentali in questo caso sono le piante infestanti trattate, a proposito delle quali sono state rilevate due variabili: la dose di trattamento ed il peso dopo il trattamento. Si può notare che all'aumentare della dose diminuisce il peso delle piante (a causa dell'effetto diserbante) ed è inoltre lecito ipotizzare che vi sia una relazione diretta tra le due variabili in esame, nel senso che la dose agisce da variabile indipendente (perché fissata dallo sperimentatore) ed il peso agisce da variabile dipendente (perché costituisce la risposta della pianta alla dose applicata). E' anche chiaro che è la dose dell'erbicida a determinare il peso e non mai viceversa.
Si tratta quindi di una classica analisi di regressione, che può essere eseguita come segue.
In primo luogo si può calcolare la devianza di X e la devianza di Y.

N° campione	*Dose (X)*	$(X_i - \mu_X)$	$(X_i - \mu_X)^2$
1	*5*	-7.5	56.25
2	*10*	-2.5	6.25
3	*15*	2.5	6.25
4	*20*	7.5	56.25
Media =12.5		*Devianza =125*	

N° campione	*Peso (Y)*	$Y_i - \mu_Y$	$(Y_i - \mu_Y)^2$
1	*91*	32.25	1040.0630
2	*61*	2.25	5.0625
3	*54*	-4.75	22.5625
4	*29*	-29.75	885.0625
Media =58.75		*Devianza =1952.75*	

La codevianza di X e Y è pari a

N° campione	$X_i - \mu_X$	$Y_i - \mu_Y$	$(X_i - \mu_X)(Y_i - \mu_Y)$
1	-7.5	32.25	-241.875
2	-2.5	2.25	-5.625
3	2.5	-4.75	-11.875
4	7.5	-29.75	-223.125
		Codevianza = - 482.5	

Da questo ricaviamo che:

$$b_1 = \frac{-482.5}{125} = -3.86$$
$$b_0 = 58.75 + 3.86 \times 12.5 = 107$$

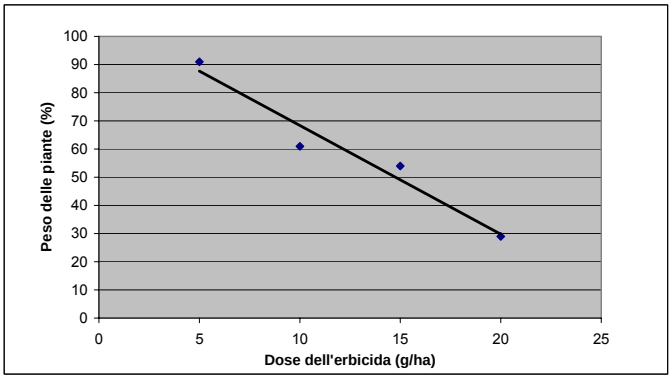
La funzione cercata è quindi:

$Y = 107 - 3.86\,X$

Il coefficiente di correlazione è pari a:

$$r = \frac{-482.5}{\sqrt{125 \times 1952.75}} = -0.97661$$

che ci indica ulteriormente come le due variabili sono negativamente correlate e come questa correlazione è piuttosto buona.
Il coefficiente di determinazione è pari al quadrato del coefficiente di correlazione ed è pari a 0.9538: si può concludere che la regressione è molto buona (valore vicino ad 1).
La funzione trovata è riportata nel grafico sottostante:



Con la funzione di regressione ottenuta possiamo calcolare la dose che ha provocato il 50% di inibizione di sviluppo nella pianta trattata: già graficamente si può notare che la dose è intorno a 15 g/ha. Più precisamente si può calcolare che:

$$Y = 107 - 3.86 \times X$$

$$X = -\frac{Y - 107}{3.86}$$

$$posto\, Y = 50$$

$$X = 14.767$$

Quindi l'ED50 è pari a 14.67 grammi.

Con R si opera nel seguente modo (notare l'impiego della funzione `lm(Y~X)` *e ricordare che il carattere ~ si ottiene tramite una combinazione dei tasti ALT+0126):*

```
> regressione<-read.table("Caso4.dat",header=TRUE)
> regressione
  Dose Peso
1    5   91
2   10   61
3   15   54
4   20   29
> attach(regressione)
> cor(Dose,Peso)
[1] -0.9766051
> model <- lm(Peso~Dose)

Call:
lm(formula = Peso ~ Dose)

Coefficients:
(Intercept)         Dose
     107.00        -3.86

> summary(model)

Call:
lm(formula = Peso ~ Dose)

Residuals:
   1    2    3    4
 3.3 -7.4  4.9 -0.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  107.000      8.229  13.002  0.00586 **
Dose          -3.860      0.601  -6.423  0.02339 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1

Residual standard error: 6.719 on 2 degrees of freedom
Multiple R-Squared: 0.9538,     Adjusted R-squared: 0.9306
F-statistic: 41.25 on 1 and 2 DF,  p-value: 0.02339
```

Per disegnare il grafici si ricorre al metodo usuale `plot`*; per aggiungere la funzione, si può utilizzare il metodo* `abline`*, che sovrappone un grafico ad un altro.*

```
> plot(Dose,Peso)
> abline(model)
```

Per calcolare il peso delle piante trattate con 12.5 g/ha di erbicida si ricorre alla funzione predict:

```
> predict(model, data.frame(Dose=12.5))
[1] 58.75
```

Mentre l''ED50 può essere ricavata grazie ad una serie di calcoli, estrendo dall'oggetto `model` *i valori dei coefficienti di regressione*

```
> (50-model$coefficients[1])/model$coefficients[2]
(Intercept)
   14.76684
```

CASO STUDIO IV.4
Eseguire le regressioni relative alle due serie di dati sottostanti ed osservare che i diversi valori di R^2 *comportano una diversa bontà di adattamento.*

1 - Immissione dei dati

```
> x<-c(25,50,75,100,125,150)
> y1<-c(201,173,144,117,101,75)
> y2<-c(205,153,164,107,111,68)
```

2 - Calcolo delle regressioni

```
> mod1<-lm(y1~x)
> mod2<-lm(y2~x)
```

Per ottenere il coefficiente di determinazione si va a leggere la proprietà `r.squared` *degli oggetti* `mod1` *e* `mod2`

```
> summary(mod1)$r.squared
[1] 0.9933147
> summary(mod2)$r.squared
[1] 0.9034583
```

2 - Esecuzione dei grafici
Per questo scopo si utilizzano le funzioni plot, abline (per sovrapporre la retta di regressione) e text (per sovrapporre del testo al grafico.

```
> plot(x,y1,type="p", col="blue", lwd=5)
> abline(mod1,col="red")
> text(100,180,"R^2=0.99")


> plot(x,y2,type="p", col="blue", lwd=5)
> abline(mod2,col="red")
> text(100,180,"R^2=0.90")
>
```

L'output finale è riportato in figura IV. 1.

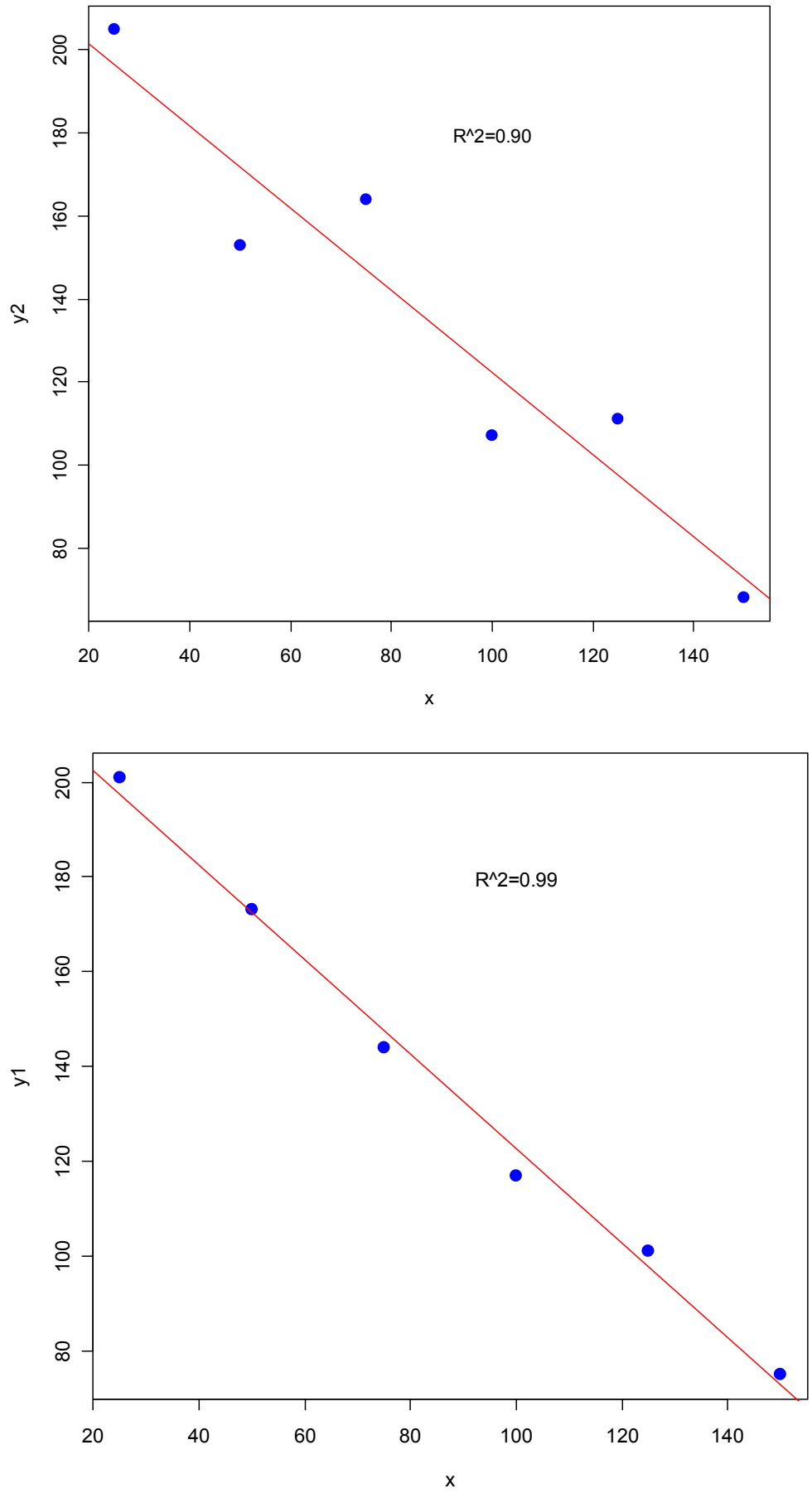


Figura IV. 1. Grafici di due rette di regressione con diverso valore di R^2.

RIEPILOGO

In questo capitolo abbiamo utilizzato le seguenti funzioni R:

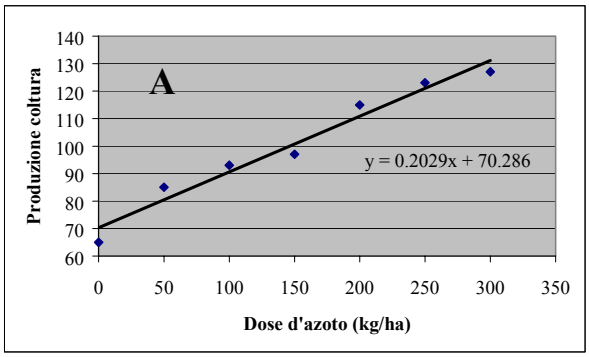
table (vettoredati X, vettoredati Y)	costruisce una tabella di contingenza
as.table(matrice)	considera una matrice a due dimensioni come una tabella di contingenza
summary(tabella di contingenza)	calcola alcune statistiche descrittive sulle tabelle di contingenza
str	elenca gli attributi degli oggetti, che possono poi essere utilizzati nei calcoli
by(vettoreati,vettoregruppi,statistica)	calcola statistiche per gruppi su un vettore
cov(vettoreX,vettoreY)	calcola la covarianza
cor(vettoreX,vettoreY)	calcola la correlazione
plot(vettoreX,vettoreY)	esegue un grafico di dispersione
ALT+0126	~
lm(Y~X)	esegue una analisi di regressione
summary(lm(Y~X))	fornisce alcune statistiche dell'analisi di regressione
abline(lm(Y~X))	sovrappone la retta di regressione al grafico di dispersione

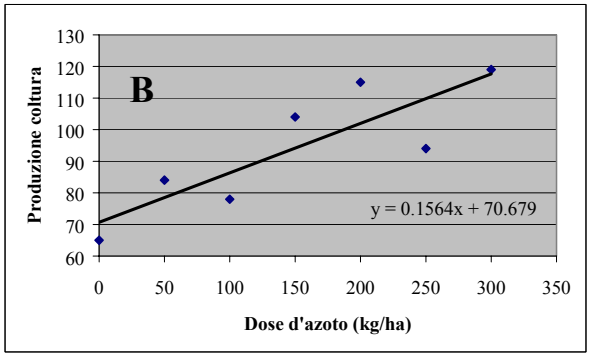
ESERCIZI PROPOSTI

1) L'analisi di correlazione si usa per (cancellare la/le risposta/e sbaglata/e:
   a) verificare se due medie sono significativamente diverse tra loro
   b) verificare se una variabile detta dipendente è funzione di un altra variabile detta indipendente
   c) Verificare se due variabili variano in modo congiunto

2) Qual è il significato del termine R^2, nell'analisi di regressione?

3) Osservate attentamente i grafici A e B, che rappresentano due analisi di regressione. Quale delle due rette è caratterizzato da un R^2 pari a 0.72 e quale da un R^2 pari a 0.97? Spiegare succintamente il perchè.





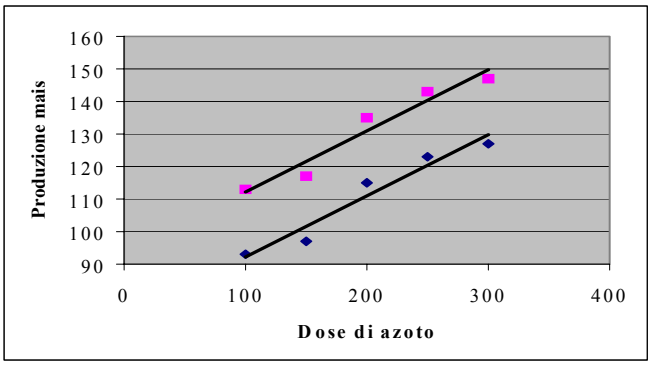
4) La retta nel grafico A (esercizio 5) è caratterizzata da:
   a) un B_0 pari a 96.5
   b) un B_0 pari a 70.28
   c) un B_0 pari a 0.2029

5) La retta nel grafico B (esercizio 5) è caratterizzata da:
   a) un coefficiente di regressione (B_1) pari a 0.1564
   b) un coefficiente di regressione (B_1) pari a 70.679
   c) un coefficiente di regressione (B_1) pari a 0.2029

6) Un campo di mais è concimato con tre dosi crescenti di azoto e pari a 0, 150 e 300 kg/ha. Le produzioni osservato sono rispettivamente pari a 5, 9 e 12 t/ha. Stabilire la relazione esistente tra dose di concimazione e produzione, il coefficiente di correlazione, l'equazione di regressione ed il valore di R^2.

7) Osservare la figura seguente. Le curve di regressione riportate hanno:
   a) lo stesso coeff. di regressione, diversi B_0 e valori di R^2 simili.

b) Lo stesso B_0, diversi coeff. di regressione e valori di R^2 simili.
c) Lo stesso coeff. di regressione, diversi B_0 e valori di R^2 molto diversi.
d) Diversi coeff. di regressione e B_0, valori di R^2 simili.
Motivare succintamente la risposta.



UNITA' V: DALLA POPOLAZIONE AL CAMPIONE: IL CALCOLO DI PROBABILITA'

OBIETTIVO

Introdurre alcuni concetti di base del calcolo di probabilità. Introdurre le variabili casuali. Illustrare le variabili casuali più utilizzate in pratica. Comprendere l'uso delle variabili casuali per l'assegnazione delle probabilità ad alcuni eventi.

SOMMARIO

MAPPA CONCETTUALE

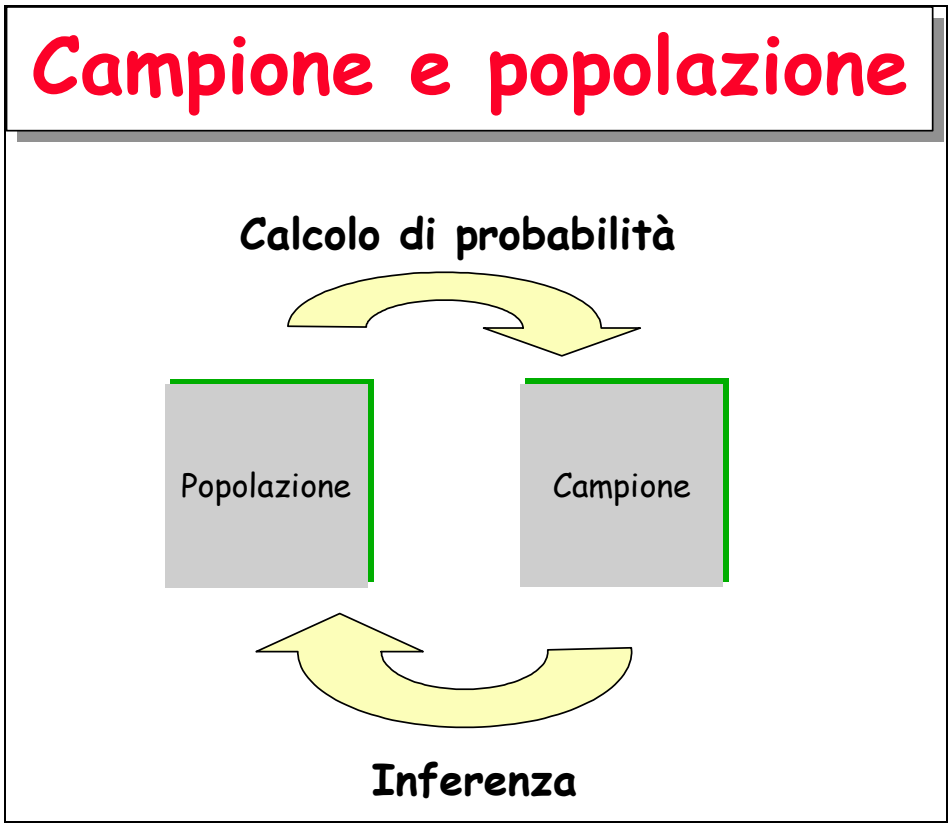


Figura V.1. Schema del processo di inferenza statistica.

SPIEGAZIONE

Alcune volte i collettivi sono così numerosi che non possono essere studiati nella loro interezza. Per questo motivo si estrae un campione casuale sul quale si eseguono le necessarie misure che debbono poi essere utilizzate per comprendere le caratteristiche dell'intera popolazione. E' evidente, comunque, che la popolazione rimane un'entità non conoscibile e qualunque delle sue caratteristiche deve essere dedotta su una base probabilistica: *"dato che il campione è in questo modo è allora probabile che la popolazione abbia queste caratteristiche"*. Per poter fare questo discorso occorre avere informazioni relative al calcolo di probabilità, cioè occorre sapere, data una certa popolazione con certe caratteristiche come sono i campioni che si estraggono da questa popolazione. Usando queste informazioni al contrario possiamo effettuare la cosiddetta inferenza statistica.

Cenni sul calcolo di probabilità

Il calcolo di probabilità insegna ad assegnare la probabilità agli eventi. Se consideriamo la probabilità di un evento singolo (**probabilità semplice**), secondo la **definizione classica** essa è data dal numero di casi favorevoli sul totale dei casi possibili:

$$probabilità\ di\ un\ evento = \frac{numero\ dei\ casi\ favorevoli}{numero\ dei\ casi\ possibili}$$

Questa definizione ci aiuta ad assegnare la probabilità ad un gran numero di eventi, come ad esempio il lancio della moneta (la probabilità di ottenere testa è 0.5, dato che due sono gli eventi possibili, uno dei quali è quello favorevole) o di un dado a sei facce (la probabilità di ottenere 1 è pari a 0.167), ma in alcune situazioni possono presentarsi problemi relativi al peso da assegnare ai diversi casi possibili. Infatti nel caso di una moneta sappiamo che, se essa non è truccata, i due casi possibili (testa o croce) sono equiprobabili. Ma se abbiamo una popolazione di insetti composta da un imprecisato numero di maschi e di femmine e vogliamo calcolare la probabilità di incontrare un maschio, non riusciamo a farlo secondo la definizione classica, che ci porterebbe a concludere che detta probabilità è pari a 0.5, visto che gli eventi possibili sono due, cioè di incontrare un maschio o una femmina.

Un'altra definizione di probabilità è quella *a posteriori* (frequentistica):

$$probabilità\ di\ un\ evento = \frac{numero\ di\ esperimenti\ favorevoli}{numero\ complessivo\ di\ esperimenti}$$

Nel caso precedente, potremmo effettuare 100 estrazioni (cioè ripetere 100 volte l'esperimento di estrazione), osservare che si incontrano 99 maschi ed una femmina e concludere, che la probabilità di estrarre una maschio è pari a 0.99. Queesta definizione è etremamente utile in alcuni casi, ma presenta anch'essa un problema: il numero di esperimenti effettuati influenza la probabilità di un evento. Se dalla popolazione precedente effettuiamo 10 estrazioni ed otteniamo 9 maschi ed una femmina concludiamo erroneamente che la probabilità cercata è pari a 0.90.

Esiste una terza definizione di probabilità, quella *soggettivistica*, legata all'aspettativa che ognuno nutre sul fatto che un evento si realizzi oppure no. In casi estremi anche questa definizione di probabilità è estremamente utile pur nella sua "soggettività". In pratica, le tre definizioni di probabilità sono tutte vere e vengono utilizzate insieme per calcolare la probabilità di un evento singolo.

Probabilità di eventi complessi

Anche se misurare la probabilità di eventi semplici non sempre è cosa agevole, esistono regole definite per ricavare la probabilità di eventi complessi, cioè costituiti da più eventi semplici di cui sia nota la probabilità.

Parliamo di **eventi complessi indipendenti** quando il verificarsi dell'uno non influenza la probabilità che si verifichi l'altro (es. due lanci di dado consecutivi). In questo caso, se A e B sono i due eventi e P(A) e P(B) sono le loro probabilità semplici, risulta che (**prodotto logico**)

$$P(A\,e\,B) = P(A) \cdot P(B)$$

Se i due eventi non sono indipendenti, nel senso che il verificarsi dell'uno influenza la probabilità con cui si verifica l'altro (es. la probabilità di ottenere il numero 1 e il numero 2 consecutivamente a tombola) il loro prodotto logico è pari a

$$P(A\,e\,B) = P(A) \cdot P(B\,/\,A)$$

Nel nostro caso la probabilità sarebbe pari a 1/90 x 1/89, dato che l'aver estratto l'1 influenza la probabilità di estrarre il due, dato che alla seconda estrazione il numero cercato è l'unico su 89 e non più su 90, visto che un numero è stato già estratto.

Se vogliamo sapere la probabilità di un evento complesso risultante dal verificarsi di due eventi semplici in alternativa (o l'uno o l'altro), parliamo di **somma logica**:

$$P(A\,o\,B) = P(A) + P(B) - P(A\,e\,B)$$

L'ultimo termine è inutile se i due eventi sono alternativi, nel senso che non possono verificarsi insieme. La probabilità di avere 1 o 2 nel lancio di un dado è pari a 1/6 + 1/6 (i due eventi sono alternativi) mentre la probabilità di estrarre da un mazzo un asso o una carta di bastoni è pari ad 4/40 + 10/40 – 1/40 = 13/40, ove 1/40 è la probabilità che la carta estratta sia contemporaneamente un asso e una carta di bastoni.

Se gli eventi sono alternativi, possiamo definire l'**evento complementare**: ad esempio ottenere una carta di coppe dal mazzo ha come evento complementare non ottenere una carta di coppe. La probabilità dell'evento complementare è pari a:

$$p(E) = 1 - p(\overline{E})$$

Calcolo combinatorio

Nel caso di eventi molto complessi sono necessarie ulteriori nozioni per poter essere in grado di valutarne la probabilità. In particolare è spesso necessario ricorrere al calcolo combinatorio per sapere il numero totale di eventi possibili e quindi determinarne la probabilità secondo la definizione classica.

Nel calcolo combinatorio distinguiamo permutazioni, disposizioni semplici o con ripetizione e combinazioni.

Le **permutazioni** sono come gli anagrammi ed indicano in quanti modi (ordinamenti) diversi possono essere presi *n* oggetti. Le permutazioni di *n* elementi sono date da:

$$Permutazioni = n!$$

In R il fattoriale si calcola con la funzione Prod(1:n):
Ad esempio, data una gara tra quattro squadre, quante sono le possibili classifiche? In R:

```
> prod(1:4)
[1] 24
>
```

Si ricorda che il calcolo dei fattoriali diviene complesso quando n è maggiore di circa 170. In questo caso, la soluzione può essere approssimata ricorrendo alla funzione gamma o meglio al logaritmo di gamma, sapendo che

$$n! = \gamma(n+1)$$

e che:

$$\log(n!) = \log(\gamma(n+1))$$

Le **disposizioni semplici** sono le possibili scelte di *k* elementi ordinati da un insieme composto da *n* oggetti (disposizioni di *n* elementi di classe *k*). Ad esempio, quali sono i possibili podi in una gara tra 8 atleti? Al primo posto possono esserci 8 persone diverse, al secondo posto ce ne possono essere 7 e al terzo 6; le combinazioni possibili possono essere 8 x 7 x 6 = 336.
In generale:

$$Disposizioni = \frac{n!}{(n-k)!}$$

Le disposizioni coincidono con le permutazioni se n=k.
Le **disposizioni con ripetizione** (disposizioni con ripetizione di *n* elementi di classe *k*) sono come le disposizioni, ma ogni oggetto, dopo essere stato scelto viene rimesso nell'insieme di partenza. Date 10 lettere (da A a L), quante combinazioni ordinate (ABCD è diverso da BACD ad esempio) da quattro lettere posso effettuare? Per la prima lettera ho dieci possibilità, altrettante per la seconda e così via. In generale:

$$Disposizioni\ con\ ripetizione = n^k$$

Quante sono le possibili disposizioni di X 1 2 in una schedina al totocalcio? Esattamente 3^{13}.

Le **combinazioni** (combinazioni di *n* elementi di classe *k*, con k≤n) sono analoghe alle disposizioni con ripetizione, ma senza considerare l'ordine con cui gli oggetti si presentano. Nel caso dell'esempio precedente, ABCD è uguale a BACD, BADC e così via. Nel caso specifico, date quattro lettere vi sono 4! permutazioni possibili che per noi sono assolutamente equivalenti. In sostanza il numero delle combinazioni è dato dal numero delle disposizioni con ripetizione di *k* elementi su *n* diviso per il numero delle permutazioni di *k* elementi (coefficiente binomiale)

$$combinazioni = \frac{n!}{(n-k)!k!} = \binom{n}{k}$$

Il coefficiente binomiale, in R è dato dalla funzione `choose(n,k)`.

Le variabili casuali

Se con le metodiche finora illustrate siamo in grado di calcolare la probabilità degli eventi, possiamo anche costruire delle **variabili casuali**, cioè dei modelli matematici (funzioni) che assegnano la relativa probabilità ad ogni valore X assumibile dal fenomeno in studio. Per variabili discrete:

$$p_i = P(X = x_i) = P(x_i)$$

ove p_i è la probabilità che si presenti ogni modalità x_i della variabile casuale X, con

$$p_i \geq 0 \quad e \quad \sum_{i=1}^{n} p_i = 1$$

Nel caso delle variabili casuali p_i è detta **funzione di frequenza** o **distribuzione di frequenza**.

Immaginiamo un mazzo di carte con i quattro assi, tre due, due tre e un quattro (10 carte); definiamo la variabile casuale

$$P(x) = \begin{cases} 0.4 \ se \ x = 1 \\ 0.3 \ se \ x = 2 \\ 0.2 \ se \ x = 3 \\ 0.1 \ se \ x = 4 \end{cases}$$

Oltre alla funzione di frequenza, si può definire anche la *funzione cumulata di frequenza*, detta anche **funzione di ripartizione**:

$$P(X \leq x) = \sum_{x_i \leq x} P(X = x_i)$$

Possiamo definire la media (valore atteso) di una variabile casuale discreta come:

$$\mu = E(X) = \sum_{i=1}^{k} x_i \cdot P(X = x_i)$$

e la varianza come:

$$\sigma^2 = Var(X) = E(X - EX)^2 = \sum_{i=1}^{k} \left[(x_i - \mu)^2 \cdot P(X = x_i) \right]$$

Se abbiamo variabili casuali continue (**funzione di densità** o **densità di frequenza**):

$$Y = f(x),$$

non cambia nulla, salvo il fatto che la funzione di ripartizione (probabilità cumulata), la media e la devianza sono definite ricorrendo al concetto di integrale:

$$P(X \leq x) = \int_{-\infty}^{x} f(u)du$$

$$\mu = E(X) = \int_{-\infty}^{+\infty} xf(x)dx$$

$$\sigma^2 = Var(X) = E(X - EX)^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)dx$$

Variabili casuali empiriche e teoriche

Molto spesso nello studio dei collettivi o delle popolazioni naturali possiamo far riferimento a variabili casuali che utilizzano funzioni matematiche note. In questo caso possiamo utilizzare la funzione e le sue proprietà matematiche ed algebriche per descrivere e comprendere il fenomeno stesso (modello matematico descrittivo). In particolare, se studiamo un fenomeno naturale che segue una determinata distribuzione o densità di frequenza, possiamo calcolare la probabilità di un qualunque evento che sia realizzato e che si debba realizzare. Vediamo ora alcune variabili casuali che possono essere utilizzate per interpretare fenomeni di interesse agrario e biologico in genere.

Variabili casuali discrete: la distribuzione binomiale

Ogni esperimento che consiste in un insieme di prove indipendenti ripetute, per ognuna delle quali abbiamo solo due esiti possibili (successo ed insuccesso), con una probabilità di successo costante, viene detto **esperimento Bernoulliano**. Nell'ambito di questi esperimenti, spesso siamo interessati a conoscere la probabilità di ottenere *k* successi su *n* prove, che può essere descritta attraverso la **variabile casuale binomiale**.
Poniamo di sapere che in una Facoltà di Agraria con un numero molto elevato di studenti il rapporto tra maschi e femmine sia pari a 0.7 e quindi che la probabilità di incontrare un maschio sia pari a P = 0.7 (evento semplice). Deve essere estratto a sorte un viaggio studio per quattro studenti e, per una questione di pari opportunità, si preferirebbe che fossero premiati in ugual misura maschi e femmine. Qual è la probabilità che un simile evento si realizzi?
La probabilità cercata si può ottenere pensando che abbiamo un evento "estrazione" che può dare due risultati possibili (maschio o femmina) e che deve essere ripetuto quattro volte. In ogni estrazione la probabilità di ottenere un maschio è pari a *P* mentre quella di ottenere una femmina (evento complementare) è pari a 1- *P* = *Q* = 0.3; ATTENZIONE!!!!!! ciò è vero se la popolazione è sufficientemente numerosa da pensare che la singola estrazione non cambia la probabilità degli eventi nelle successive (eventi indipendenti). La probabilità che su quattro estrazioni si verifichi 2 volte l'evento "maschio" e due volte l'evento femmina è data dal prodotto delle probabilità di ciascun evento (*teorema della probabilità composta*), cioè:

$0.7 \times 0.7 \times 0.3 \times 0.3 = 0.7^2 \times 0.3^2$

In generale, data una popolazione molto numerosa, nella quale gli individui si presentano con due modalità possibili (in questo caso maschio e femmina) e posto di sapere che la frequenza con cui si presenta la prima modalità è pari a *p* (in questo caso la frequenza dei maschi è pari a 0.7), mentre la frequenza della seconda modalità è pari a $q = 1 - p$, se vogliamo estrarre da questa popolazione *n* elementi, la probabilità che *x* di questi presentino una delle due modalità è data da:

$p^n \times q^{(n-x)}$

La formula di cui sopra, tuttavia, non risolve il nostro problema, in quanto noi vogliamo che vengano estratte due femmine, indipendentemente dall'ordine con cui esse vengono estratte (prima, seconda, terza o quarta estrazione), mentre la probabilità che abbiamo appena calcolato è quella

relativa all'evento in cui le due femmine sono estratte al terzo e quarto posto. Di conseguenza (*teorema della probabilità totale*) dobbiamo sommare tutte le probabilità relative all'estrazione di due femmine in prima e seconda posizione, oppure in seconda e terza, oppure in seconda e prima e così via. Il numero delle combinazioni possibili per 2 maschi in quattro estrazioni (combinazione di 4 elementi su 2), oppure delle due femmine in quattro estrazioni sono date dal coefficiente binomiale:

$$\binom{n}{k} = \frac{n!}{(n-k)!k!\cdot}$$

Moltiplicando le due equazioni date in precedenza otteniamo l'equazione della distribuzione binomiale:

$$\frac{n!}{k!\cdot(n-k)!} \cdot p^k \cdot q^{(n-k)}$$

che nel caso specifico da il risultato:

$$\frac{4!}{2!(4-2)!} \cdot 0.7^2 \cdot 0.3^2 = \frac{24}{4} \cdot 0.49 \cdot 0.09 = 0.2646$$

che è appunto la probabilità cercata

In R, utilizziamo la funzione `dbinom(successi, prove, probabilità semplice)` per calcolare le probabilità della distribuzione binomiale, ogni volta in cui vogliamo sapere la probabilità di ottenere *k* successi in *n* prove:

```
> dbinom(2,4,0.7)
[1] 0.2646
```

La funzione binomiale costituisce una variabile casuale:

$$P(X = k) = \frac{n!}{k!\cdot(n-k)!} \cdot p^k \cdot q^{(n-k)}, \quad k = 0,1,.....,n$$

ove:

$$\mu = E(X) = np$$

e:

$$\sigma^2 = np(1-p)$$

La funzione di ripartizione (probabilità cumulata) è:

$$P(X = x) = \sum_{x_i \le x} P(X = x_i)$$

ed è ottenibile in R con la funzione `pbinom(successi, prove, probabilità semplice)`. Nell'esempio, se vogliamo sapere la probabilità totale di estrarre meno di tre

femmine (≤ 2 femmine), possiamo operare in questo modo:

```
> pbinom(2,4,0.3)
[1] 0.9163
```

Che risulta dalla somma della probabilità di estrarre 0, 1, 2 femmine:

```
> zero<-dbinom(0,4,0.3)
> uno<-dbinom(1,4,0.3)
> due<-dbinom(2,4,0.3)
> zero+uno+due
[1] 0.9163
```

La funzione di ripartizione può anche essere utilizzata al contrario, per determinare i quantili, cioè il numero di successi che corrispondono ad una probabilità cumulata pari ad α:

```
> qbinom(0.9163,4,0.3)
[1] 2
```

CASO STUDIO V.1
Da una popolazione di insetti che ha un rapporto tra maschi e femmine pari a 0.5, qual è la probabilità di campionare casualmente 2 maschi e 8 femmine?

```
> dbinom(2,10,0.5)
[1] 0.04394531
```

CASO STUDIO V.2
Riportare su un grafico la funzione di ripartizione binomiale, per p=0.5 e n=5. Costruire anche la densità di frequenza, utilizzando le opportune funzioni R.

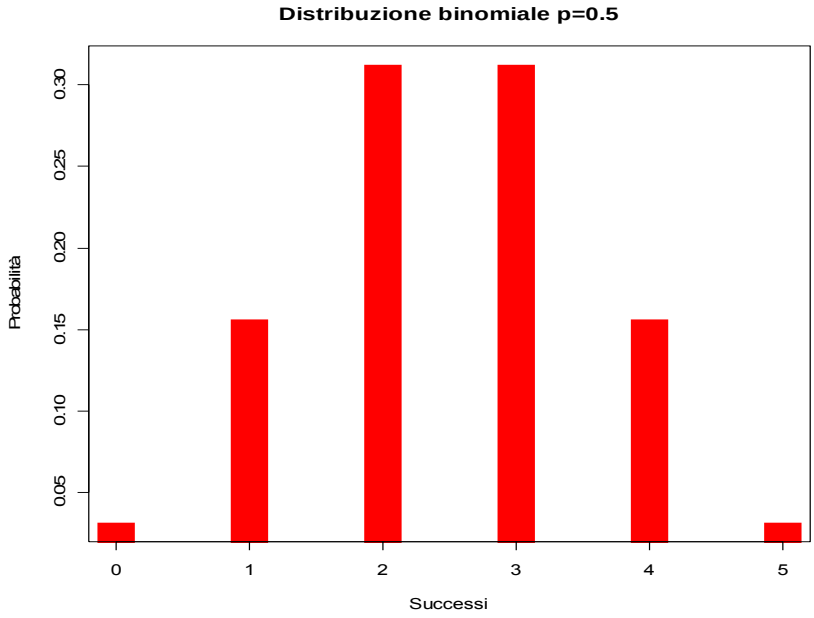
```
> a<-dbinom(0,5,0.5)
> b<-dbinom(1,5,0.5)
> c<-dbinom(2,5,0.5)
> d<-dbinom(3,5,0.5)
> e<-dbinom(4,5,0.5)
> f<-dbinom(5,5,0.5)
> x<-c(a,b,c,d,e,f,)
> plot(c(0,1,2,3,4,5),x,type="h",main="Distribuzione
binomiale p=0.5",xlab="Successi",ylab="Probabilità", lwd=30,
col="red")
```

L'output grafico è in figura V.2.

```
> a<-pbinom(0,5,0.5)
> b<-pbinom(1,5,0.5)
> c<-pbinom(2,5,0.5)
> d<-pbinom(3,5,0.5)
> e<-pbinom(4,5,0.5)
> f<-pbinom(5,5,0.5)
> x<-c(a,b,c,d,e,f,)
> plot(c(0,1,2,3,4,5),x,type="h",main="Distribuzione
binomiale p=0.5",xlab="Successi",ylab="Probabilità", lwd=30,
```

```
col="red")
```

L'output grafico è in figura V.3.




```
>
```

Figura V.2. Istogramma della distribuzione binomiale per p=0.5 e n=5.

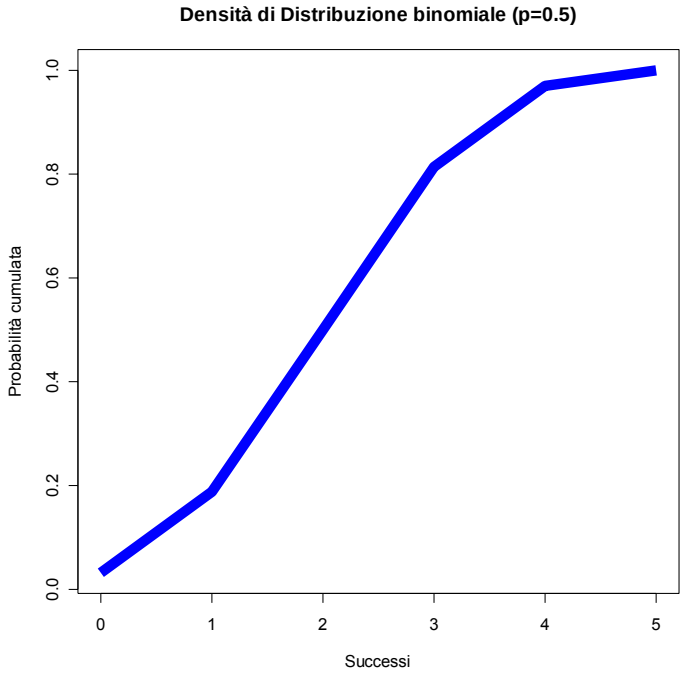


Figura V.3. Istogramma della distribuzione binomiale cumulata per p=0.5 e n=5.

Variabili casuali continue: la distribuzione normale (curva di Gauss)

Anche senza voler entrare molto in dettaglio delle problematiche poste dalla statistica è necessario accennare come in natura esistono un infinito numero di popolazioni possibili: si pensi a

quanti fenomeni biologici si possono studiare e misurare. Tuttavia, da tempo si è notato che le misurazioni fatte in relazione alla gran parte dei fenomeni biologici possono in ultima analisi essere ricondotte ad una sola distribuzione di frequenze, la cosiddetta *distribuzione normale*.

Si richiamiamo alla mente i dati relativi all'esercizio 1: abbiamo visto che le 3000 altezza potevano essere organizzate nella distribuzione di frequenza riportata in Tabella 1 e in Figura 2. Dalla figura si osserva che si tratta di una distribuzione di frequenze ad istogramma, rappresentabile con una funzione discontinua. Tuttavia, se immaginiamo di aumentare infinitamente il numero degli individui, possiamo anche pensare di restringere l'ampiezza delle classi di frequenza, fino a farle divenire infinitamente piccole. In questo modo la nostra distribuzione di frequenza tende ad assumere una forma a campana, che potrebbe essere descritta con una funzione continua detta *curva di Gauss* (figura 6).

La curva è descritta dalla seguente funzione:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}};$$

ove P(x) è la frequenza di una certa misura x, mentre μ e σ sono rispettivamente la media e la deviazione standard della popolazione. Le distribuzioni di frequenza che possono essere descritte con la curva di Gauss, prendono il nome di *distribuzioni normali*.

Si può dimostrare che μ e σ sono rispettivamente la media e la deviazione standard della variabile casuale distribuzione normale.

Studiare le principali proprietà matematiche della curva di Gauss è estremamente utile, perché, se supponiamo che essa possa descrivere la gran parte dei fenomeni biologici naturali, possiamo estendere le caratteristiche della curva e all'andamento del fenomeno in studio. Ad esempio, senza voler entrare troppo in dettaglio, il semplice esame grafico della curva di Gauss consente le seguenti osservazioni:

1) La forma della curva dipende da solo da μ e σ (figure 7 e 8). Ciò significa che, se prendiamo un gruppo di individui e partiamo dal presupposto (assunzione parametrica) che in relazione ad un determinato carattere quantitativo (es. altezza) la distribuzione di frequenza è normale (e quindi può essere descritta con una curva di GAUSS), allora basta conoscere la media e la deviazione standard degli individui e immediatamente conosciamo l'intera distribuzione di frequenza e non abbiamo più bisogno di esporre le frequenze delle diverse classi.
2) la curva ha due asintoti e tende a 0 quando x tende a ± ∞. Questo ci dice che se assumiamo che un fenomeno è descrivibile con una curva di Gauss, allora assumiamo che tutte le misure sono possibili, anche se la loro frequenza decresce man mano che ci si allontana dalla media;
3) Se la curva di Gauss è stata costruita utilizzando le frequenze relative, l'integrale della funzione è uguale ad 1. Infatti la somma delle frequenze relative di tutte le varianti possibili non può che essere uguale ad 1;
4) la curva è simmetrica. Questo indica che la frequenza dei valori superiori alla media è esattamente uguale alla frequenza dei valori inferiori alla media. Non solo; dato un certo valore γ qualunque, la frequenza dei valori superiori a μ+γ è uguale alla frequenza dei valori inferiori a μ+γ
5) Allo stesso modo se γ = σ, possiamo dire che la frequenza dei valori superiori a μ+σ è uguale alla frequenza dei valori inferiori a μ-σ. Questa frequenza è pari a circa il 15.87%. Allo stesso modo la frequenza degli individui superiori a μ+2σ è pari al 2.28% (questi valori si ricavano dall'integrale della curva di Gauss o funzione di distribuzione normale);
6) Considerando la somma degli eventi, possiamo dire che la frequenza degli individui con misure superiori a μ+σ più la frequenza degli individui inferiori a μ-σ è del 31.74%. Allo stesso modo, possiamo dire che la frequenza dei valori compresi tra μ+σ e μ-σ è pari al 68.26%.
7) Così procedendo, ricorrendo all'integrale della funzione di distribuzione normale, possiamo

sapere che la frequenza dei valori compresi tra μ+1.96σ e μ-1.96σ è pari al 95% e che la frequenza dei valori compresi tra μ+2.575σ e μ-2.575.

In sostanza, possiamo concludere che data una popolazione distribuita normalmente, con media μ e deviazione standard σ, ricorrendo all'integrale della funzione di distribuzione, possiamo calcolare quale è la frequenza di ogni possibile individuo. Siccome il concetto di frequenza è strettamente associato a quello di probabilità (nel senso che la frequenza di una particolare variante è uguale alla probabilità che abbiamo di estrarre quella variante dalla popolazione), possiamo anche affermare che la probabilità di estrarre una certa misura o un certo intervallo di misure da una popolazione normale può essere calcolata ricorrendo all'integrale della funzione di densità di frequenza.

CASO STUDIO V.3
Disegnare due curve normali con media pari a10 e deviazione standard pari rispettivamente a 2 e 4, sovrapponendo le due curve sullo stesso grafico. Disegnare inoltre due curve normali con media pari a 4 e 8 e deviazione standard pari a 3.
Si noti l'uso della funzione `dnorm(x,media,deviazione standard)` *combinata con la funzione* `curve` *(cheserve a plottare funzioni algebriche e/o di altra natura) e il metodo* `add` *per sovrapporre i grafici.*

```
> curve(dnorm(x,10,2),2,18, lwd=5,col="red", main="Curve
        normali", xlab="Variabile",ylab="Frequenza")
> curve(dnorm(x,10,4),2,18, lwd=5,col="blue",  add=TRUE)
```

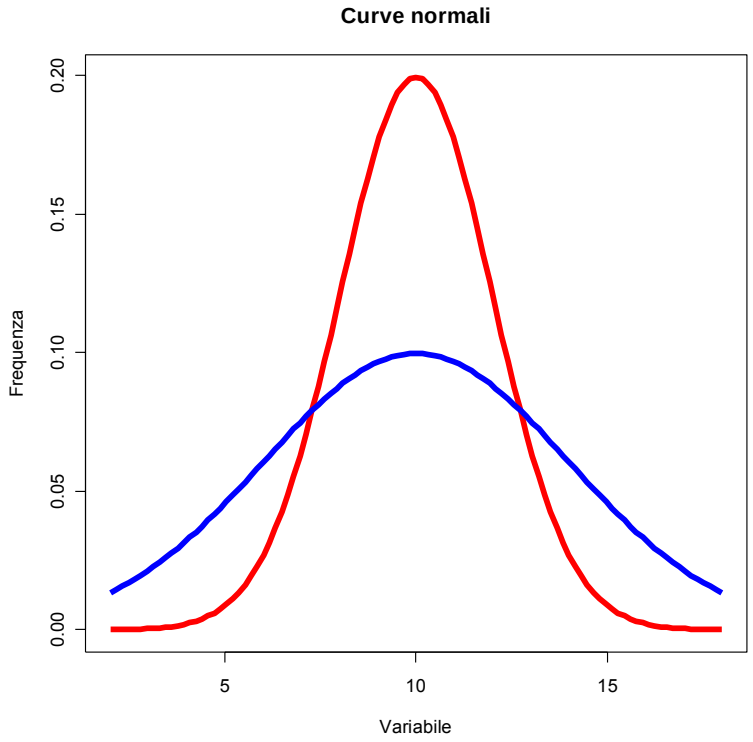


Figura V.4. Grafico di due distribuzioni normali con la stessa media e diversa deviazione standard.

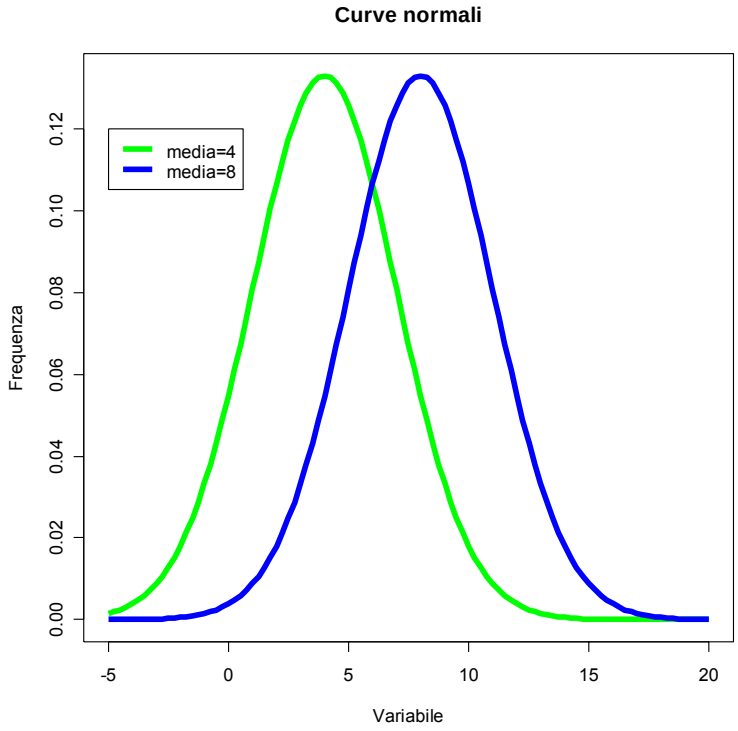


Figura V.5. Grafico di due distribuzioni normali con diversa media e stessa deviazione standard.

Trasformazione e standardizzazione delle variabili

Le popolazioni normali sono infinite (perché infiniti sono i valori possibili per μ e σ), ma con opportune trasformazioni dei dati possono tutte essere ricondotte ad una sola popolazione di riferimento con $\mu = 0$ e $\sigma = 1$, detta *popolazione normale standardizzata*. Gli integrali della funzione di ripartizione di quest'ultima popolazione sono riportati nelle cosiddette tavole di z (Tabella 2)

Trasformare una popolazione (o comunque un insieme) di dati (misure) significa aggiungere ad ognuno di essi una quantità costante e/o moltiplicare ognuno di essi per una quantità costante. La trasformazione si rilette sul valore della media e della deviazione standard dei dati in modo altamente prevedibile.

In particolare, tutti i dati della popolazione possono essere addizionati ad un numero *n*. In questo caso, la media della popolazione trasformata è pari alla media della popolazione non trasformata + *n*. Lo stesso vale se tutti i dati sono moltiplicati per un numero comune *n*. In questo caso anche la media è uguale al prodotto della media della popolazione non trasformata per *n*.

Esempio

Considerate i dati
(a) 12 ; 14 ; 16 ;18 ; 11. La media è pari a: 14.2
Se ad ogni dato aggiungiamo il numero 2, otteniamo:
(b) 14 ; 16 ; 18 ; 20 ; 13. La nuova media è 16.5
Se invece consideriamo la serie:
(c) 24 ; 28 ; 32 ; 36 ; 22. La media è 28.4

Lo stesso vale se tutti i dati sono moltiplicati per un numero comune *n*. In questo caso anche la media è uguale al prodotto della media della popolazione non trasformata per *n*.

Se invece della media consideriamo la deviazione standard, le trasformazioni addittive non hanno alcun effetto, mentre le trasformazioni moltiplicative fanno si che la deviazione standard sia moltiplicata per *n*.

Esempio

Considerate i dati dell'esempio precedente.
(a) 12 ; 14 ; 16 ;18 ; 11. $\sigma = 2.86$
Se ad ogni dato aggiungiamo il numero 2, otteniamo:
(b) 14 ; 16 ; 18 ; 20 ; 13. $\sigma = 2.86$
Se invece consideriamo la serie:
(c) 24 ; 28 ; 32 ; 36 ; 22. $\sigma = 5.72$

Ora se prendiamo un insieme di dati (x) calcoliamo la media e la deviazione standard e poi prendiamo ogni dato ci sottraiamo la media e dividiamo il risultato per la deviazione standard, secondo la funzione

$$z = \frac{x - \mu}{\sigma}$$

otteniamo un insieme di dati trasformati la cui media è zero e la cui deviazione standard è 1.

Esempio

Considerate i dati:
(a) 2 ; 5 ; 8; $\mu = 5;\ \sigma = 3$
Se ad ogni dato sottraiamo 5 e dividiamo il risultato per 3, otteniamo la serie:
(b) –1 ; 0 ; 1; $\mu = 0;\ \sigma = 1$

In questo modo, qualunque sia la popolazione normale di partenza, possiamo trasformarla in una popolazione normale standardizzata; ciò ci permette di risolvere il problema del calcolo di frequenza o di probabilità semplicemente ricorrendo alle tavole degli integrali della popolazione normale standardizzata che sono tabulati nella maggior parte dei testi di statistica. Nel nostro caso, per calcolare la probabilità di una distribuzione normale standardizzata, basterà utilizzare la funzione `dnorm(x)`, senza specificare la media e la deviazione standard.

Densità di frequenza, funzione di distribuzione e quantili

In R, tre sono le funzioni collegate alla distribuzione normale:
1 - `dnorm(x,media, ds)` calcola la probabilità associata ad un certo valore x;
2 - `pnorm(val,media, ds)` calcola la probabilità cumulata per x≤val;
3 - `qnorm(α,meadia,ds)` calcola il valor x che lascia alla sua sinistra una probabilità α.
Utilizzando queste tre funzioni possiamo fare tutti i calcoli di probabilità necessari.

CASO STUDIO V.4
Qual è la probabilità che, da un pozzo con un contenuto medio di cloro pari a 1 meq l^{-1}, eseguendo l'analisi con uno strumento caratterizzato da un

coefficiente di variabilità pari al 4%, si ottenga una misura pari o superiore a 1.1 meq l^{-1}? E' possibile che questa misura sia stata ottenuta casualmente, oppure è successo qualcosa di strano (errore nell'analisi o inquinamento del pozzo)?
Questo problema può essere risolto immaginando che se è vero che il pozzo ha un contenuto medio di 1 meq l^{-1} i contenuti di cloro dei campioni estratti da questo pozzo dovrebbero essere distribuiti normalmente, con media pari ad 1 e deviazione standard pari a 0.04 (si ricordi la definizione di coefficiente di variabilità). Qual è la probabilità di estrarre da questa popolazione una misura pari superiore a 1.1 meq l^{-1}? La risposta può essere trovata ricorrendo ad R:

```
> 1-pnorm(1.1,mean=1,sd=4*1/100)
[1] 0.006209665
```

Si utilizza 1 - `pnorm()` in quanto la funzione `pnorm` restituisce l'integrale della funzione di ripartizione da -inf. a z, evento complementare quello da noi considerato.
Allo stesso risultato si può ricorrere utilizzando l'argomento `lower.tail`

```
> pnorm(1.1,mean=1,sd=4*1/100,lower.tail=FALSE)
[1] 0.006209665
```

CASO STUDIO V.5
Nello stesso strumento sopra indicato e considerando lo stesso tipo di analisi, calcolare:
1 - la probabilità di ottenere una misura inferiore a 0.75
2 - la probabilità di ottenere una misura superiore a 1.5
3 - la probabilità di ottenere una misura compresa tra 0.95 e 1.05

Stabilire inoltre:
1 - la misura che è superiore al 90% di quelle possibili
2 - la misura che è inferiore al 70% di quelle possibili
3 - le misure entro le quali si trova il 95% delle misure possibili

```
> pnorm(0.75,1,4*1/100)
[1] 2.052263e-10
> pnorm(1.5,1,4*1/100,lower.tail=FALSE)
[1] 3.732564e-36
> pnorm(1.05,1,4*1/100)-pnorm(0.95,1,4*1/100)
[1] 0.7887005
> qnorm(0.9,1,0.04)
[1] 1.051262
> qnorm(0.7,1,0.04,lower.tail=FALSE)
[1] 0.979024
```

l'ultima espressione è equivalente alla seguente:

```
> qnorm(0.3,1,0.04)
[1] 0.979024
```

```
> qnorm(0.975,1,0.04)
[1] 1.078399
> qnorm(0.025,1,0.04)
[1] 0.9216014
>
```

La distribuzione delle medie campionarie

Il problema precedente dovrebbe aver chiarito come, dato uno strumento di analisi caratterizzato da un errore pari al 4%, se dobbiamo analizzare una sostanza la cui concentrazione è pari ad 1, le misure ottenute, nel 95% dei casi oscilleranno tra 1.07 e 0.92. In realtà, come abbiamo già avuto modo di ricordare, noi non eseguiremmo mai una singola analisi, ma ripeteremo la misura almeno due o tre volte, calcolando poi la media. Il problema allora è: esiste una variabile casuale che descrive la distribuzione delle medie di tutti gli infiniti campioni estraibili dalla popolazione anzidetta. Si può dimostrare che, data una popolazione normalmente distribuita con media μ e deviazione standard σ, le medie campionarie sono anch'esse normalmente distribuite con media μ e deviazione standard pari a:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

dove n è la dimensione del campione.

CASO STUDIO V.6
Nello stesso strumento indicato al cosa studio 5 (CV=4%), immaginando di eseguire analisi in triplicato di una sostanza la cui concentrazione è pari a 1 ng/l, calcolare:
1 - la probabilità di ottenere una media campionaria < 0.75
2 - la probabilità di ottenere una media campionaria > 1.5
3 - la probabilità di ottenere una media campionaria compresa tra 0.95 e 1.05

Stabilire inoltre:
4 - la media campionaria superiore al 90% di quelle possibili
5 - la media campionaria inferiore al 70% di quelle possibili
6 - le medie campionarie entro le quali si trova il 95% delle misure possibili

```
(punto 1)
> pnorm(0.75,1,0.04/sqrt(3))
[1] 1.305861e-27
(punto 2)
> pnorm(1.5,1,0.04/sqrt(3),lower.tail=FALSE)
[1] 2.997427e-104
punto (3)
> pnorm(1.05,1,0.04/sqrt(3))-pnorm(0.95,1,0.04/sqrt(3))
[1] 0.9696172
(punto 4)
> qnorm(0.9,1,0.04/sqrt(3))
[1] 1.029596
(punto 5)
> qnorm(0.7,1,0.04/sqrt(3),lower.tail=FALSE)
[1] 0.9878895
(punto 6)
> qnorm(0.975,1,0.04/sqrt(3))
```

```
[1] 1.045263
> qnorm(0.025,1,0.04/sqrt(3))
[1] 0.9547366
>
```

Del precedente caso studio è di particolare interesse il punto 6; in altre parole si afferma che, data una sostanza di concentrazione pari ad 1 ng/l, che deve essere dosata con uno strumento che ha un errore di misura del 4%, la concentrazione vera è indeterminabile, giacché a causa dell'anzidetto errore di misura i risultati delle analisi si comporteranno come una variabile casuale normale, con media pari alla concentrazione incognita e deviazione standard pari a 0.04. Per ottenere la vera concentrazione del campione l'unico modo sarebbe ripetere infinite analisi. Tuttavia, possiamo considerare anche che, se preleviamo un campione di n individui dalla popolazione in esame (cioè se ripetiamo l'analisi n volte), abbiamo il 95% di probabilità che la media delle *n* determinazioni effettuate sia compresa tra 0.95 e 1.04. Questi margini di incertezza si restringono se *n* aumenta e si annullano quando *n* diviene infinito.

Questa affermazione, così posta, è contestuale e vale solo per il mio strumento con CV=4% e media da determinare pari a 1 ng/l. Se volessimo fare un discorso di validità più generale, potremmo pensare alla standardizzazione delle misure, in modo da avere, qualunque sia la sostanza da analizzare e qualunque sia l'errore di misura dell'apparecchio, una distribuzione delle misure con media pari a 0 e σ pari ad 1. Di conseguenza la distribuzione delle medie campionarie sarà normale, con media pari a 0 e deviazione standard pari a $1/\sqrt{n}$.

Nella distribuzione normale standardizzata delle medie campionarie, il 95% delle misure è compreso tra:

```
> qnorm(0.975)
[1] 1.959964
> qnorm(0.025)
[1] -1.959964
>
```

cioè:

$$-1.96 < \frac{x-\mu}{\frac{\sigma}{\sqrt{n}}} < 1.96$$

$$\mu - 1.96\frac{\sigma}{\sqrt{n}} < x < \mu + 1.96\frac{\sigma}{\sqrt{n}}$$

cioè possiamo concludere che il 95% delle misure (medie di n determinazioni), in genere, è compreso la media vera più o meno una quantità costante, pari ad un multiplo dell'errore standard.

La distribuzione t di Student

Abbiamo visto che la variabile casuale standardizzata e medie campionarie standardizzate si distribuiscono norrmalmente con media pari a 0 e deviazione standard pari a $1/\sqrt{n}$. Tuttavia, per operare la standardizzazione è necessario conoscere la deviazione standard σ della popolazione originaria. In alcuni casi questo valore non è noto e deve essere stimato a partire dalla deviazione standard del campione (s). In questo caso la quantità:

$$t = \frac{x - \mu}{\frac{s}{\sqrt{n}}}$$

non segue la distribuzione normale, in quanto esiste un margine di incertezza in più, che aumenta la probabilità delle misure lontane dalla media. Si può dimostrare che la quantità anzidetta si distribuisce secondo una distribuzione detta **t di Student** con ν gradi di libertà, pari alla numerosità del campione-1. Più cresce ν, più cresce la precisione di *s*, più la distribuzione di *t* tende a quella normale standardizzata.

Esempio: la forma della distribuzione t di Student
Nell'esempio sottostante si può osservare che la distribuzione t di Student è molto vicina alla normale già con 24 gradi di libertà (gl).

```
> curve(dnorm(x),-3,+3,col="Black",lwd=3,xlab="",ylab="Probabilità")
> curve(dt(x,2),-3,+3,col="Blue",lwd=3, add=TRUE)
> curve(dt(x,6),-3,+3,col="Red",lwd=3, add=TRUE)
> curve(dt(x,24),-3,+3,col="Green",lwd=3, add=TRUE)
> legend(-3,0.35,legend=c("normale","t con 2 gl","t con 6 gl","t con
    24 gl"),col=c("Black","Blue","Red","Green"),lty=c(1,1,1,1),
    lwd=c(3,3,3,3))
>
```

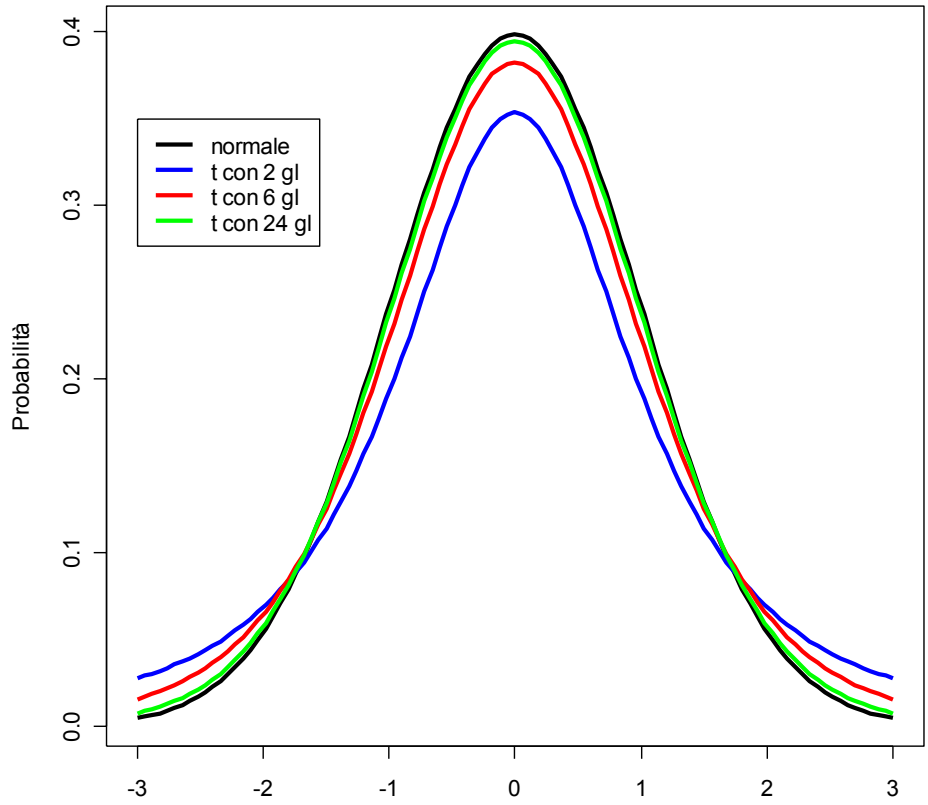


Figura V.6. Distribuzione normale a confronto con alcune distribuzioni t di Student per un diverso numero di gradi di libertà

CASO STUDIO V.7

Con una bilancia abbiamo pesato quattro piante di mais trattate allo stesso modo e provenienti da un appezzamento piuttosto grande. Le misure sono: 125, 128, 136 e 142 g di sostanza secca. Ammesso che si tratti di un campione rappresentativo, valutare la probabilità che questi quattro individui siano estratti da una popolazione con media pari a 150 g. Verificare anche se la media del campione è esterna all'intervallo che contiene il 95% delle misure più probabili.

Il campione ha una media $\bar{x}$ *= 132.75 e deviazione standard s =7.719.*

```
> peso<-c(125, 128, 136 , 142 )
> media<-mean(peso)
> devst<-sqrt(var(peso))
> media
[1] 132.75
> devst
[1] 7.719024
```

Possiamo considerare in prima approssimazione che la popolazione da cui è estratto il campione abbia una deviazione standard pari a quella del campione. Di conseguenza, la quantità:

$$t = \frac{x - 150}{\frac{7.719}{\sqrt{4}}}$$

ove x è la media del campione, si distribuisce secondo la variabile casuale t di Student (σ è stimato). Nel nostro caso, t = - 4.47

```
> errst<-devst/sqrt(4)
> errst
[1] 3.859512
> tosservato<-(media-150)/errst
> tosservato
[1] -4.469477
```

La probabilità corrispondente è:

```
> pt(tosservato,3)
[1] 0.01043776
```

Come dice il valore di porbabilità, questa misura non fa parte del 95% delle misure più probabili; infatti i valori di t situati al 2.5 e al 97.5 mo percentile della distribuzione di t sono:

```
> qt(0.025,3)
[1] -3.182446
> qt(0.975,3)
[1] 3.182446
```

Che corrispondono ad un peso di:

$$t = \frac{x - \mu}{\frac{s}{\sqrt{n}}}$$

$$x = t\frac{s}{\sqrt{n}} + \mu = \pm 3.18 \times 3.86 + 150$$

```
> qt(0.025,3)*errst+150
[1] 137.7173
> qt(0.975,3)*errst+150
[1] 162.2827
```

Distribuzione F di Fisher

Se da una popolazione normale N(μ,σ) estraiamo due campioni indipendenti otteniamo due stime s_1 ed s_2 della deviazione standard σ. Se operiamo infinite volte l'estrazione di coppie di campioni e ogni volta misuriamo la quantità:

$$F = \frac{s_1^2}{s_2^2}$$

otteniamo la variabile casuale F di Fisher, con ν_1 gradi di libertà al numeratore (relativi ad s^2_1) e ν_2 gradi di libertà al denominatore (relativi ad s^2_2). La distribuzione F è fortemente asimmetrica, con mediana pari ad 1.

Esempio: la forma della distribuzione F
Nell'esempio sottostante si può osservare che la distribuzione t di Student è asimmetrica (Figura V.7).

```
> curve(df(x,3,3),0,+3,col="Black",lwd=3,xlab="",
     ylab="Probabilità",ylim=c(0,1.5))
> curve(df(x,10,10),0,+3,col="Blue",lwd=3,add=TRUE)
> curve(df(x,50,50),0,+3,col="Red",lwd=3,add=TRUE)
> curve(df(x,3,50),0,+3,col="Green",lwd=3,add=TRUE)
> legend(2,1.3,legend=c("3,3 gl","10,10 gl","50,50 gl","3,50
     gl"),col=c("Black","Blue","Red","Green"),lty=c(1,1,1,1),
     lwd=c(3,3,3,3))
```

Altre variabili casuali di interesse per lo sperimentatore

Oltre a quelle accennate, esistono molte altre variabili casuali, sia continue che discrete, utilizzate nell'inferenza statistica. Menzioniamo solamente la variabile χ^2, la variabile casuale ipergeometrica ed esponenziale negativa. Le relative densità di distribuzione, funzioni di ripartizioni e quantili sono disponibili in R e possono essere utilizzate nel calcolo, con la sintassi usuale.

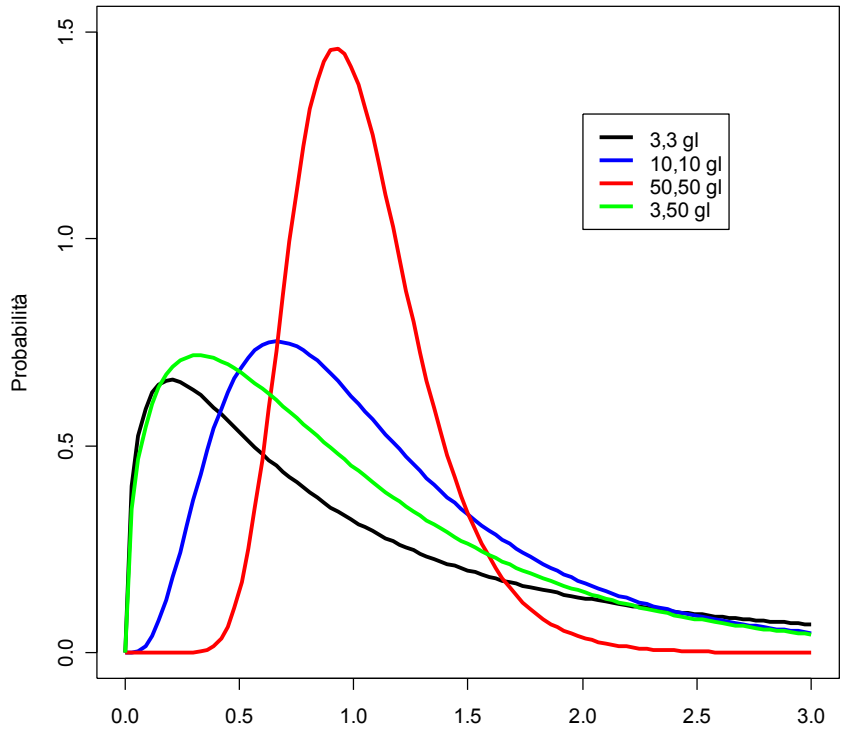


Figura V.7. Distribuzioni di F per diversi gradi di libertà

RIEPILOGO

In questo capitolo abbiamo utilizzato, tra le altre, le seguenti funzioni R:

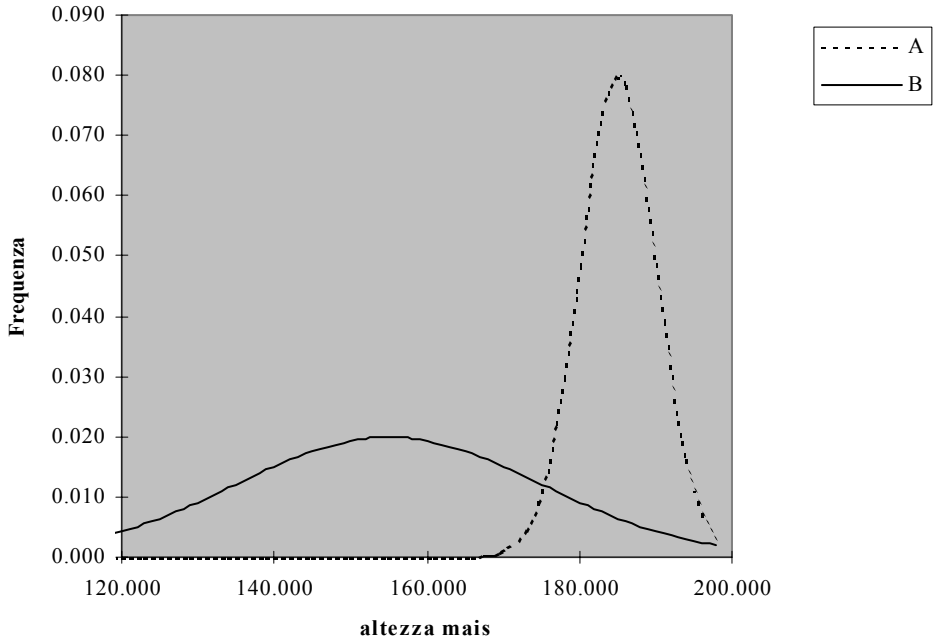
prod(1:n)	calcolo del fattoriale di n
choose(n,k)	coefficiente binomiale
dbinom(k,n,p)	densità di distribuzione binomiale k successi su n prove probabilità p
pbinom(k,n,p)	integrale della distribuzione binomiale (funzione di ripartizione)
qbinom(α,n,p)	quantile α in una distribuzione binomiale per n prove e probabilità p
dnorm(x,mean=μ, sd=σ)	densità di distribuzione normale di x. Default mean=0, sd=1)
pnorm(x,mean=μ, sd=σ)	Funzione di ripartizione distribuzione normale da -inf a 0
qnorm(α,mean=μ, sd=σ)	quantile α

ESERCIZI PROPOSTI

1) Il grafico sottostante rappresenta la distribuzione di frequenza delle produzioni altezze delle piante di mais nel caso di un ibrido (A) e di una linea pura (B). Scegliere la risposta appropriata motivando opportunamente la scelta.

Le due popolazioni hanno

a) la stessa altezza media, ma l'ibrido (A) ha maggior deviazione standard □
b) l'ibrido (A) ha altezza media e deviazione standard maggiore □
c) la linea pura (B) ha altezza media e deviazione standard maggiore □
d) L'ibrido (A) ha maggiore altezza media e minore deviazione standard □



2) Il grafico sottostante rappresenta la distribuzione di frequenza delle produzioni del frumento non concimato (B) e concimato (A). Scegliere l'affermazione esatta, motivando opportunamente la risposta.

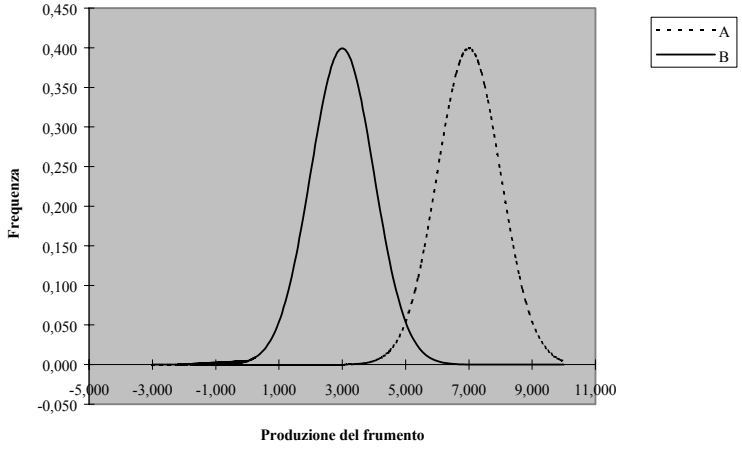
La concimazione ha incrementato:

a) la produttività media e la variabilità dei risultati produttivi ☐

b) solo la variabilità dei risultati produttivi ☐

c) solo la produttività media. ☐

d) Non ha avuto effetti di sorta ☐



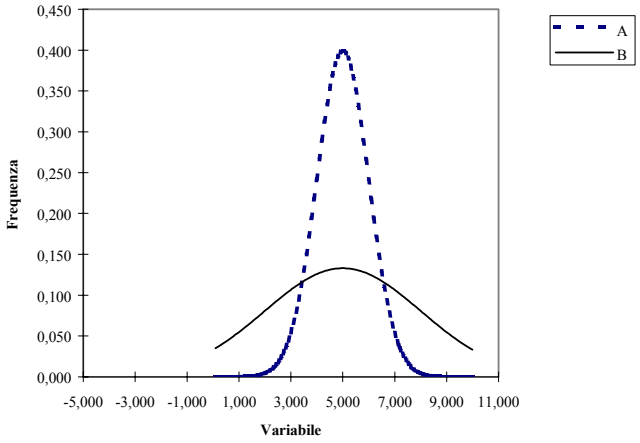
3) Cosa rappresenta il grafico sottostante?

a) Due popolazioni normali con $\mu_a = \mu_b$, $\sigma_a < \sigma_b$

b) Due popolazioni normali con $\mu_a = \mu_b$, $\sigma_a > \sigma_b$

c) Due popolazioni normali con $\mu_a > \mu_b$, $\sigma_a = \sigma_b$

d) Due popolazioni binomiali, con p = 5


0,450
0,400
0,350
0,300
0,250
0,200
0,150
0,100
0,050
0,000
Frequenza
-5,000
-3,000
-1,000
1,000
3,000
5,000
7,000
9,000
11,000
Variabile
A
B

UNITA' VI. DAL CAMPIONE ALLA POPOLAZIONE: LA STIMA DEI PARAMETRI

OBIETTIVO

Introdurre gli studenti al disegno sperimentale e al processo inferenziale di stima dei parametri.

SOMMARIO

SPIEGAZIONE

Introduzione all'inferenza statistica

Finora abbiamo visto che, dato un collettivo di misure o dati ricavati da unità sperimentali in relazione ad una o due variabili, è possibile ricavare una serie di indicatori descrittivi, funzione dei dati e capaci di descrivere alcune delle caratteristiche dell'intero collettivo, come la tendenza centrale, la variabilità, la variazione congiunta e la dipendenza lineare.

Ovviamente ciò soddisfa solo alcune delle esigenze del ricercatore o del tecnico che abbia a che fare con un collettivo di misure o dati sperimentali. Infatti, come abbiamo avuto modo di accennare, spesso lo sperimentatore non è interessato solo ai dati in suo possesso, in quanto li considera un campione rappresentativo di una popolazione più ampia che non si è potuta studiare nel suo complesso, per motivi di tempo, di costo, di opportunità o di fattibilità. E' evidente comunque che l'interesse dello sperimentatore è rivolto alla popolazione, non al campione da questa estratto. In sostanza, noto che sia il campione, è necessario estrapolare (o *inferire*) da questo le caratteristiche della popolazione che lo ha generato.

Questa operazione di inferenza può essere compiuta grazie alle nozioni apprese nel capitolo precedente, riguardo al calcolo di probabilità: infatti l'estrazione di un determinato campione è un evento del quale possiamo calcolare la probabilità se assumiamo che la popolazione in studio si comporti secondo una determinata distribuzione statistica di riferimento (ed es. la normale di Gauss). Vedremo in seguito alcuni esempi di questo modo di operare e ragionare. Anticipiamo però che i problemi di inferenza si dividono in due grandi gruppi: stima dei parametri e test d'ipotesi. Con la stima dei parametri, partendo da alcune statistiche descrittive del campione inferiamo le caratteristiche della popolazione che ha generato il campio, mentre con il test d'ipotesi tentiamo di verificare l'esattezza di ipotesi sperimentali disponibili *a priori*.

Campionamento

L'inferenza statistica è possibile se e solo se il campione è *rappresentativo*; per essere tale, un

campione deve essere composto da un numero sufficiente di unità estratte casualmente dalla popolazione, in modo che ogni singolo individuo della popolazione ha la stessa probabilità di tutti gli altri di essere incluso nel campione medesimo. *Il problema della selezione del campione (campionamento) è un problema centrale di ogni metodologia sperimentale: se il campionamento non è rappresentativo, i dati raccolti non potranno mai permettere nessuna conclusione in relazione al fenomeno in studio.*

Anche le misure che si eseguono in relazione a qualunque fenomeno biologico costituiscono un problema di campionamento. Infatti, dovrebbe essere oramai chiaro che ogni misura che effettuiamo in natura è soggetta ad un errore, più o meno evidente, legato alle cause più disparate, che vanno dall'imprecisione di misura all'effetto di non determinabili cause perturbatrici esterne. Questa semplice osservazione ci obbliga, ogni volta che dobbiamo eseguire una misura, a ripetere la determinazione più volte, in modo da minimizzare l'impatto delle possibili fonti di errore. In questo modo ci troviamo ad avere a che fare con un campione di misure estratto casualmente dall'infinito universo di tutte le misure possibili.

E' evidente che le misure effettuate non ci interessano in se', perché la nostra attenzione è rivolta a tutta la popolazione di misure possibili. Quest'ultima è di solito caratterizzata da una distribuzione normale, con una certa media (il valore più probabile per la misura cercata), ma anche con una certa variabilità che in qualche modo riflette l'entità degli errori possibili. Vediamo quindi che il problema della misura è in realtà un vero e proprio problema di inferenza statistica, con il quale, **cerchiamo di ottenere delle stime più o meno attendibili per alcuni valori, che in realtà sono destinati a rimanere ignoti.**

La sperimentazione agraria

Come si inseriscono le osservazioni finora effettuate nella realtà operativa di un agronomo? La connessione è evidente se si pensa che la gran parte delle informazioni tecniche o scientifiche vengono ottenuta grazie ad un lavoro di ricerca sperimentale, attraverso l'esecuzione di appositi *esperimenti scientifici,* nei quali si realizzano espressamente situazioni controllate, in modo da verificare l'effetto di un *trattamento sperimentale* e confrontarlo con situazioni diverse ed alternative.

L'effetto del trattamento in esame viene valutato attraverso apposite misure, da eseguire sugli individui inclusi nell'esperimento e sottoposti al trattamento in studio. Questi individui non rappresentano in genere l'intero universo degli individui disponibili, bensì un campione da esso estratto e che si considera rappresentativo dell'intero universo.

Ad esempio se vogliamo studiare un farmaco non possiamo somministrare questo farmaco all'intera popolazione mondiale, ma dovremo somministrarlo ad un campione di individui nel quale dovranno essere incluse tutte le età, entrambi i sessi, tutte le razze e così via, in modo che le conclusioni a cui arriviamo alla fine possano essere estese all'intera popolazione mondiale.

Questo modo di procedere comporta sempre un certo grado di incertezza, che rende fondamentale l'adozione di una metodologia sperimentale corretta e supportata da un razionale impiego della statistica.

Organizzazione di un esperimento

La prima cosa da fare nell'organizzare un esperimento, dopo averne deciso l'obiettivo e aver quindi stabilito quali sono i trattamenti da studiare, è quella di individuare le unità sperimentali a cui somministrare il trattamento in studio.

Le unità sperimentali possono essere costituite da individui (un albero, una animale, un vaso, un uomo, un'analisi chimica) oppure, come nel caso della sperimentazione agronomica, da piccoli appezzamenti di terreno, che vengono chiamati **parcelle**.

La scelta delle unità sperimentali è particolarmente critica per un esperimento corretto, proprio

per il concetto di rappresentatività di cui si è parlato finora. Ciò è particolarmente importante per le parcelle di terreno che debbono essere sempre di dimensioni giudiziosamente scelte. Nello stabilire la dimensione delle parcelle va tenuto conto del presumibile **effetto *di bordo*** che si verificherà, cioè del diverso sviluppo che le piante perimetrali assumono sotto l'influenza delle parcelle o dei viottoli contigui. Esempi: un albero o una varietà a taglia bassa sarà ombreggiato e riceverà danno dalla vicinanza di un albero o di una varietà alta, e viceversa; una parcella non concimata (o non irrigata) può risentire della concimazione (o dell'irrigazione) fatte alla parcella vicina; l'allettamento di una parcella può danneggiare quella vicina; l'esistenza di un viottolo dà agio alle piante prospicienti di godere di più spazio, di più acqua, di più nutrimento delle piante situate all'interno della parcella.
Nel caso di esperienze di alimentazione su animali, qualcosa di simile all'effetto di bordo si verifica all'inizio della prova; in questi casi è tassativo di lasciar passare un certo numero di giorni tra l'inizio della somministrazione della razione sperimentale e l'inizio della raccolta dei dati di produzione. Si dà così modo all'organismo dei soggetti di mettersi a regime.

E’evidente che le situazioni “di bordo” debbono comunque essere escluse dai rilievi finali, per evitare un sensibile incremento dell’errore sperimentale.

Per quanto riguarda la dimensione ottimale delle parcelle, varia a seconda della variabilità del terreno, della fittezza di coltivazione, del tipo di trattamenti che si sperimenta. In genere con parcelle piccole diminuisce, entro certi limiti, l'errore dovuto all'eterogeneità del terreno, mentre aumenta quello dovuto alla variabilità delle piante e all’imprecisione delle misure. Con parcelle grandi spesso ci si illude di avvicinarsi di più alle condizioni colturali di pieno campo, ma si incorre in tale variabilità delle condizioni ambientali che l'effetto dei trattamenti rischia di essere mascherato. Inoltre, la dimensione e la forma delle parcelle non può prescindere da considerazioni relative ai macchinari che verranno eventualmente utilizzati per la semina, per la raccolta, o a considerazioni relative alla disponibilità del seme

Il trattamento sperimentale e il concetto di replicazione (replica)

Alle unità sperimentali prescelte viene imposto il trattamento sperimentale da studiare, seguendo le procedure richieste dal trattamento stesso. Dal punto di vista metodologico, dovrebbe essere ormai chiaro che ogni trattamento sperimentale sia applicato non solo su un’unità sperimentale, ma su un numero di unità sperimentali maggiore dell’unità.

Ognuna delle unità sperimentali a cui è stato applicato lo stesso trattamento viene chiamata **replica** o **replicazione**. L’insieme delle replicazioni costituisce il campione su cui verranno fatte le successive analisi statistiche: è evidente che questo campione è estratto dall’infinito numero di individui simili che si sarebbero potuti considerare nel corso dell’esperimento.

Il numero di replicazioni da adottare su un certo esperimento o la dimensione delle parcelle dipende dalla natura dell’esperimento: più questo è alto e maggiore è la precisione dell’esperimento, ma anche i costi ad esso connessi in termini di tempo e denaro. Nella sperimentazione agraria il numero più usuale di ripetizioni oscilla tra 3 e 6; limitazioni nelle disponibilità di superficie, di soggetti, di mezzi finanziari o di lavoro impediscono, generalmente, di fare più numerose ripetizioni, anche se la precisione aumenta con l'aumentare di queste. Peraltro ben poco guadagno in precisione c'è da attendersi quando si superano 8-10 ripetizioni. In genere le ripetizioni devono essere tanto più numerose quanto più il terreno o i soggetti sono disformi e quanto più esigui ci si attende che siano gli effetti dei trattamenti; le ripetizioni possono essere ridotte al minimo in condizioni opposte, cioè di grande uniformità ambientale e con trattamenti sperimentali a effetti molto marcati.

Talora si possono fare solo due ripetizioni perché il numero dei trattamenti è molto elevato, lo spazio a disposizione scarso, le disponibilità di manodopera limitate: anche se una prova con solo due ripetizioni e tutti altro che perfetta, può comunque consentire di trarre conclusioni corrette e non arbitrarie. Solo nel caso che l'effetto di un trattamento sia grandissimo può farsi a meno delle ripetizioni: ad esempio, la grande scoperta del valore fertilizzante delle scorie Thomas sui pascoli

inglesi fu fatta su un'unica grande parcella a Cocile Park. Ma la successiva messa a punto della miglior tecnica di concimazione ha potuto essere fatta solo con metodi di sperimentazione precisi basati sulle ripetizioni.

Comunque si scelgano le unità sperimentali e il numero di replicazioni, la cosa più importante per una opportuna applicazione delle metodiche statistiche in un esperimento di qualunque natura, **la regola fondamentale è che le unità sperimentali sottoposte ai diversi trattamenti differiscano tra loro solo per il trattamento oggetto di studio**.

Ad esempio, se vogliamo confrontare due livelli di concimazione azotata, dobbiamo farlo in modo che le piante trattate con una certa dose di concime differiscano da quelle trattate con un'altra dose solo per quello che riguarda la concimazione e non, ad esempio, per la varietà, l'irrigazione o altri fattori sistematici. E' evidente infatti che se trattiamo un gruppo di piante con un certo concime ed un altro gruppo con un concime diverso, allevando questo secondo gruppo su un terreno più fertile, è evidente che alla fine l'effetto misurato non potrà essere imputato al trattamento in studio (il concime), ma dalla fertilità del terreno. La massima cura deve essere messa nell'organizzazione dell'esperimento, su questo fondamentale aspetto relativo alla metodologia sperimentale.

Il rilievo dei dati

Abbiamo già accennato come ogni esperimento sia basato sull'esecuzione di una serie di misure, da effettuarsi nel momento opportuno per evidenziare l'effetto di un determinato trattamento. Bisogna tener presente che ogni esperimento necessita di una continua attenzione da parte del ricercatore, in modo da poter annotare appena si manifestano tutte le differenze di aspetto che si evidenzino tra le parcelle o i soggetti.

Oltre ai rilievi biometrici (peso, altezza ecc…), sono molto importanti anche i rilievi visivi, soprattutto per quelle variabili che non possono essere misurate facilmente, come lo stadio di sviluppo di una pianta, il vigore, gli attacchi di malattie, l'allettamento, l'infestazione di malerbe, la fitotossicità di certi prodotti, ecc.. Il rilievo visivo consiste nell'individuare una scala percentuale (ad esempio, % di piante attaccate, % di superficie allettata, ecc.) o una scala arbitraria di punti (ad esempio da 1 a 5, da 1 a 9, ecc.) e nell'assegnare ad ogni unità sperimentale il punteggio opportuno in relazione al carattere in studio. In ogni caso, nell'eseguire una misura visiva, lo sperimentatore non deve fare mai riferimento ai trattamenti, ma deve invece valutare ogni soggetto senza sapere di che tesi si tratta, in modo da non commettere errori di giudizio. Tecnica ottima è che più osservatori, dopo essersi ben accordati sui criteri generali prima di iniziare le osservazioni, procedano indipendentemente alle osservazioni stesse.

Stima puntuale dei parametri di una popolazione

Seguendo le indicazioni finora proposte è evidente che quando eseguiamo un esperimento sottoponiamo ad una certo trattamento sperimentale un dato numero di unità sperimentali, che (come già detto) sono solo un campione di quelle possibili. Tuttavia noi col nostro esperimento vogliamo tirare conclusioni generiche valide per l'intera popolazione da cui il campione è stato estratto (*stima dei parametri della popolazione*).

E' intuitivo pensare che, data una popolazione se da questa immaginiamo di estrarre a caso un campione di n individui, è probabile che la media del campione sia pari alla media della popolazione da cui questo è stato estratto. Infatti gli individui intorno alla media nella popolazione di partenza sono i più frequenti e quindi sono quelli che hanno la massima probabilità di essere inclusi nel campione. E' ovvio che questo è vero se il campione è rappresentativo (cioè se è estratto a caso e sufficientemente numeroso). Questa osservazione intuitiva ci consente di affermare che dato un campione estratto casualmente da una popolazione normalmente distribuita, **la media e la deviazione standard del campione sono una stima non distorta della media e della deviazione standard della popolazione di origine**. Per la dimostrazione di questo assunto rimandiamo a

pubblicazioni più specifiche, ma ricordiamo che solo la deviazione standard campionaria (cioè quella ottenuta come:

$$s = \sqrt{\frac{SS}{n-1}}$$

dove SS è la devianza del campione) è una stima corretta della deviazione standard della popolazione. Bisogna comunque notare che i reali valori dei parametri (media e deviazione standard) della popolazione di origine rimangono comunque ignoti, ma si può affermare che con la massima probabilità questi sono uguali a quelli del campione estratto.

Più in generale, dato un campione, le statistiche descrittive calcolate per questo campione (media, varianza, deviazione standard, regressione, correlazione ecc..) possono essere estrapolate alla popolazione che ha generato il campione stesso, senza che questo possa essere in qualche modo oggetto di critica. In fin dei conti è la migliore stima che abbiamo. Questo tipo di stima si definisce **stima puntuale**, perché ad ogni valore ignoto di un certo parametro della popolazione (ad es. la media) associamo una certa stima puntiforme, cioè costituita da un singolo valore.

ESEMPIO VI.1
Da un terreno agrario è stato estratto casualmente un campione di 5 buste da 20 grammi ciascuna di terreno. Il terreno presente in ogni busta viene analizzato per conoscere il contenuto in fosforo assimilabile. I dato ottenuti sono 9 – 10 - 14 – 16 - 13 ppm, rispettivamente per le cinque buste. Qual è il contenuto di fosforo nel terreno e qual è la sua deviazione standard (variabilità naturale del contenuto di fosforo nel terreno, errore di campionamento e di misura)?
Questo problema può essere risolto pensando che il campione da noi estratto (cinque buste) sia rappresentativo dell'intera popolazione e, di conseguenza, le statistiche descrittive del campione possono essere assunte come stime puntuali delle statistiche descrittive della popolazione.
La media delle cinque misure nel campione è pari a 12.4 ppm, mentre la deviazione standard è pari 2.88 ppm. Ne consegue che il coefficiente di variabilità è pari al 20.6%. Come abbiamo visto questi risultati possono essere estrapolati all'intera popolazione di tutte le misure possibili. Possiamo quindi concludere che il campione è estratto da un terreno il cui contenuto medio di fosforo è pari a 12.4 ppm con una deviazione standard pari a 2.88 ppm.
I reali valori di contenuto medio ed errore rimangono ignoti: le nostre conclusioni sono raggiunti solamente su base probabilistica; si tratta delle conclusione più probabili, ma non certe.

Quanto detto vale anche per la proporzione (la proporzione del campione p è una stima non distorta di π), mentre nel caso della varianza lo stimatore non distorto è la varianza campionaria (cioè quelle ottenuta dividendo la devianza per n-1).

La stima puntuale è molto comoda, ma anche molto imprecisa: possibile che la popolazione intera abbia proprio la stessa media o la stessa deviazione standard del campione che noi abbiamo estratto?

La risposta è che questo è altamente improbabile. Perciò dobbiamo associare alla stima puntuale una banda di incertezza, passando quindi alla cosiddetta stima per intervallo.

La precisione della stima e l'errore standard

Nel paragrafo precedente abbiamo affermato che la media incognita della popolazione (μ) è

stimata dalla media del campione ($\bar{x}$). Tuttavia, il calcolo di probabilità illustrato nel capitolo precedente ci ha insegnato che, data una popolazione normale con media μ e deviazione standard σ, se estraiamo infiniti campioni di *n* elementi, le medie campionarie sono distribuite normalmente con media μ e deviazione standard pari alla quantità:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

detta **errore standard di una media**. Questo può essere guardato come la variabilità della media campionaria, cioè una misura dell'incertezza legata alla stima della media. In altre parole, l'errore standard è una misura dell'errore di stima. In questo senso, l'errore standard è un concetto molto diverso dalla deviazione standard σ della popolazione da cui il campione è estratto, che ne rappresenta la variabilità naturale ineliminabile. Infatti, se dalla popolazione avessimo estratto un campione di infinite misure (n = ∞) avremmo ottenuto una stima perfetta di μ (errore standard pari a 0), nonostante la variabilità naturale σ. In altre parole, la stima può essere perfetta anche se la misura è viziata da un errore (per esempio perché l'apparecchio non è perfettamente funzionante).

E' bene sottolineare ancora come l'errore standard (e quindi la precisione della stima di μ) dipende sia dalla variabilità della misura, sia dal numero di repliche che effettuiamo ed aumenta all'aumentare della deviazione standard e diminuisce all'aumentare del numero delle ripetizioni, annullandosi quando questo tende ad infinito.

ESERCIZIO VI.2

In un vigneto, si vuole conoscere la produzione d'uva per pianta. Non avendo tempo e risorse sufficienti per misurare tutte le piante del vigneto, si scelgono dieci piante a caso e si misura la loro produzione, che risulta pari rispettivamente a:

3.6 – 4.2 – 5.2 – 3.4 - 3.9 - 4.1 – 4.7 – 4.2 – 3.9 – 3.8

La stima più probabile della produzione per pianta del vigneto è data dalla media delle misure effettuate:

$$\bar{X} = \frac{3.6+4.2+5.2+3.4+3.9+4.1+4.7+4.2+3.9+3.8}{10} = 4.1$$

La variabilità della misura (che include, tra l'altro, la variabilità individuale delle viti, la variabilità della fertilità del terreno e l'errore di misura dell'operatore) può essere stimata dalla deviazione standard del campione:

$$s = \sqrt{\frac{(3.6-4.1)^2 + (4.2-4.1)^2 + ... + (3.8-4.1)^2}{9}} = 0.527$$

Si ricorda che $\bar{X}$ è il simbolo indicato per la stima di μ, mentre s è il simbolo per indicare la stima di σ. Con le lettere greche si indica invece la vera media e la vera deviazione standard dell'intera popolazione di piante del vigneto (che rimangono ignote).

Come errore di stima della media possiamo prendere l'errore standard:

$$s_{\bar{X}} = \frac{0.527}{\sqrt{10}} = 0.167$$

In R, l'errore standard può essere calcolato con le consuete formule

```
> x<-scan()
1: 3.6  4.2  5.2  3.4  3.9  4.1  4.7  4.2  3.9  3.8
11:
Read 10 items
> mean(x)
[1] 4.1
> se<-sqrt(var(x))/sqrt(length(x))
> se
[1] 0.1666667
>
```

Notare l'uso della funzione `scan`*, che memorizza su una stringa i numeri (solo numeri) digitati nella console di R.*

Intervalli di confidenza di una media

Nel capitolo precedente abbiamo già illustrato come il 95% delle medie campionarie sono comprese nell'intervallo ± 1.96 volte l'errore standard. Di conseguenza, se affermiamo che:

$\mu = \overline{X} \pm 1.96 \times s_{\overline{X}}$

abbiamo una probabilità di essere nel giusto del 95% ed una probabilità d'errore del 5%.

Questo ragionamento, tuttavia, presuppone di conoscere la quantità σ della popolazione. Più frequentemente, σ viene stimato a partire da s, cioè dalla deviazione standard del campione. In questa situazione, abbiamo visto che le medie campionarie sono distribuite tra $\pm t_{\alpha, \nu}$ ove α è il grado di confidenza ricercato (ad esempio il 95%, che taglia una probabilità di errore del 5%, distribuita al 2.5% a destra e al 2.5% a sinistra della media) e ν è il numero di gradi di libertà della deviazione standard del campione (n-1). Gli intervalli di confidenza della media, pertanto, possono essere costruiti in questo modo:

$$\mu = \overline{X} \pm t_{\alpha;n-1} \times \frac{s}{\sqrt{n}}$$

E'bene ribadire che se vogliamo usare R per il calcolo deille bande di confidenza, dobbiamo fare attenzione al valore α; infatti se vogliamo la banda di inferenza del 95%, dobbiamo indicare un valore α = (1 - 0.95)/2 (distribuzione ad una coda).

In sostanza, dato un certo livello di probabilità d'errore (ad esempio α = 0.05, cioè probabilità d'errore pari al 5%), possiamo costruire un intervallo che molto probabilmente contiene il vero ed ignoto valore della media della popolazione da cui il campione è stato estratto. Più esattamente, questa affermazione è tanto probabile da lasciare solo un 5% di margine d'errore.

CASO STUDIO VI.1
Riprendendo i dati dell'Esercizio 9 abbiamo già osservato come, sulla base del campione esaminato, possiamo concludere che il valore più probabile della produzione media per pianta nel vigneto è pari a 4.1 kg. Questa stima ci lascia un po' insoddisfatti: come è possibile che la produzione per pianta di un intero vigneto sia proprio uguale a quella delle dieci piante misurate? Se ci calcoliamo allora l'intervallo di confidenza della media per un livello di probabilità pari al 5% (α = 0.05) otteniamo:

$\mu = 4.1 \pm 2.262 \times 0.167 = 4.1 \pm 0.377$

Questo ci permette di affermare che la produzione media per pianta del vigneto (quella vera, che rimane ignota) è compresa tra 4.447 e 3.723. Se il campione era effettivamente rappresentativo, possiamo avere fiducia che facendo questa affermazione non abbiamo più del 5% di probabilità d'errore.
Se volessimo essere ancora più tranquilli, potremmo calcolare l'intervallo di confidenza della media per un livello di probabilità pari all'1% ($\alpha = 0.01$), ottenendo:

$$\mu = 4.1 \pm 3.250 \times 0.167 = 4.1 \pm 0541$$

In questo caso possiamo affermare che la produzione media per pianta del vigneto è compresa tra 4.641 e 3.559, con una probabilità d'errore dell'1%. Come si vede, per diminuire la probabilità d'errore abbiamo dovuto allargare l'intervallo di confidenza.

In R, anche il calcolo dei limiti di confidenza si ottiene con le formule usuali, ricorrendo alla funzione `qt(α, df=n)` per ottenere il quantile α della distribuzione t. Ricordare che in R qt resituisce la distribuzione ad una coda!

```
> limfid05<-qt(0.975,df=length(x)-1)*se
> limfid05
[1] 0.3770262

> limfid01<-qt(0.995,df=length(x)-1)*se
> limfid01
[1] 0.5416393
```

L'errore standard e gli intervalli di confidenza nell'analisi di regressione

Come avrete intuito, il calcolo dell'errore standard e degli intervalli di confidenza ci consente di aggiungere alle nostre stime una banda d'incertezza; in questo modo possiamo comunque stare al riparo da errori macroscopici, anche se rimane il fatto che non potremo mai conoscere con assoluta precisione una certa caratteristica della nostra popolazione.

Lo stesso problema va affrontato nel caso dell'analisi di regressione. Come si ricorderà, eseguire una analisi di regressione in una popolazione di dati bivariata, consiste nel determinare due parametri: l'intercetta (β_0) e la pendenza (β_1) in modo da caratterizzare la retta che esprime la relazione funzionale tra le due variabili.

Anche in questo caso se non abbiamo a disposizione l'intera popolazione possiamo eseguire l'analisi di regressione su un campione rappresentativo che sia stato estratto da questa. In questo modo otterremo dei valori di intercetta (b_0) e pendenza (b_1) che sono delle stime dei valori reali dell'intera popolazione. Anche queste stime, come nel caso della media, dovranno essere corredate dei relativi intervalli di confidenza.

Il calcolo degli intervalli di confidenza nell'analisi di regressione è un po' più complicato e verrà demandato ad R. Ricordiamo tuttavia che alla base di questo calcolo vi è la determinazione degli errori standard rispettivamente di b_0 e b_1 che hanno esattamente lo stesso significato dell'errore standard di una media e costituiscono una misura dell'errore che si commette nella stima delle due quantità incognite β_0 e β_1. Anche in questo caso gli intervalli di confidenza si calcolano moltiplicando l'errore standard di b_0 e b_1 per il valore di t per il livello α prescelto e per un numero

di gradi di libertà pari ad n-2, cioè esattamente il numero delle coppie di dati, meno 2, cioè il numero di parametri stimati.

CASO STUDIO VI.2
Consideriamo il seguente campione (quattro individui):

Num. dato	Ricoprimento piante infestanti (X)	Produzione mais (Y)
17	5.00	12.75
33	12.41	11.18
35	20.05	12.25
71	65.75	10.59

Eseguiamo su di essi l'analisi di regressione con R, utilizzando il comando `summary` *sull'output dell'analisi di regressione:*

```
> x<-c(5,12.41,20.05,65.75)
> y<-c(12.75,11.18,12.25,10.59)
> model<-lm(y~x)
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
        1         2         3         4
 0.478407 -0.885316  0.397364  0.009545

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.41078    0.56588  21.932  0.00207 **
x           -0.02784    0.01616  -1.722  0.22712
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 0.7651 on 2 degrees of freedom
Multiple R-Squared: 0.5973,     Adjusted R-squared: 0.396
F-statistic: 2.967 on 1 and 2 DF,  p-value: 0.2271

>
```

Per quanto riguarda la pendenza, i limiti di confidenza per una probabilità d'errore del 5% sono pari a:

$\beta_1 = b_1 \pm t_{0.05,2} \cdot s_{b1} = -0.028 \pm 4.3027 \times 0.016 = -0.028 \pm 0.0688$

Per quanto riguarda l'intercetta, i limiti di confidenza per una probabilità d'errore del 5% sono pari a:

$\beta_0 = b_0 \pm t_{0.05,2} \cdot s_{b0} = 12.411 \pm 4.3027 \times 0.5658 = 12.411 \pm 2.4345$

In R (si vedano gli attributi dell'oggetto `summary(model)`*):*

```
>summary(model)$coefficients[4:4]*qt(0.975,df=model
$df.residual)
       <NA>
0.06953662
```

```
>summary(model)$coefficients[3:3]*qt(0.975,df=model
$df.residual)
    <NA>
2.434793
```

Il calcolo degli intervalli di confidenza è abbastanza importante, perché ci ha portato alla fine a fare un'affermazione di tipo probabilistico, che non è necessariamente vera, ma che invece è condizionata da una certa possibilità d'errore, che è comunque nota e fissata a priori, ancor prima di compiere la misurazione.

Questo modo di procedere è tipico della statistica inferenziale, nata appunto per le situazioni nelle quali non si possono avere certezze deterministiche, ma soltanto una stima, affidabile, salvo un certo rischio d'errore.

UNITA' VII. DAL CAMPIONE ALLA POPOLAZIONE: INTRODUZIONE AL TEST D'IPOTESI ED ESEMPI

OBIETTIVO

Introdurre gli studenti al test d'ipotesi.

SOMMARIO

1 - Il test d'ipotesi
2 - Errore di prima e seconda specie
3 - Ipotesi alternative semplici e complesse
4 - Confronto tra due medie: test di t di Student
5 - Confronto tra una proporzione teorica ed una proporzione osservata: metodo esatto
6 - Confronto tra due proporzioni osservate: il test di χ^2
7 - Confronto tra più medie: il test di F nell'ANOVA

SPIEGAZIONE

Il test d'ipotesi

La statistica inferenziale non si pone solo l'obiettivo di comprendere le caratteristiche di una popolazione a partire dai dati raccolti per un campione rappresentativo, ma si pone anche l'obiettivo di verificare ipotesi fatte *a priori* su alcuni aspetti di interesse biologico. Immagiamo ad esempio di avere una popolazione nota $N(\mu, \sigma)$. Immaginiamo di ipotizzare che su questa popolazione abbia agito un certo trattamento sperimentale che ha spostato la media della popolazione sul valor $\nu>\mu$. Per verificare questa ipotesi, immaginiamo di prendere un campione di *n* individui e misurare il valore della media, trovandolo pari a $X_m \pm s > \mu$. Ci chiediamo: il valore X_m è effettivamente superiore a μ perchè il trattamento ha avuto l'effetto ipotizzato, oppure si tratta di una normale oscillazione legata all'errore di campionamento?

Il test d'ipotesi aiuta a verificare ipotesi di questo tipo, su una base probabilistica. Il procedimento è di questo tipo:

1 - Si formula l'ipotesi nulla (corrispondente ad affermare che non vi è effetto) e l'ipotesi alternativa;
2 - Si calcola la probabilità che l'ipotesi nulla sia vera;
3 - Se il livello di probabilità è inferiore ad una certa soglia α prefissata (generalmente 0.05), si rifiuta l'ipotesi nulla e si accetta l'ipotesi alternativa.

> *Esempio VII.1*
> *Poniamo di monitorare con uno strumento di analisi la concentrazione di una determinata sostanza in un pozzo. Poniamo di sapere che la concentrazione sia nota e pari a μ=250 μg/l, con una deviazione standard pari a 180 μg/l. Improvvisamente, estraiamo un campione d'acqua e facendo tre analisi otteniamo un valore medio pari a 550 μg/l. Possiamo sospettare che il pozzo si è inquinato?*
> *L'ipotesi è che la concentrazione di nitrati nel pozzo sia descrivibile con una distribuzione di frequenza normale, con media pari a 250 e deviazione standard pari a 180*
> *La media osservata nel campione analizzato è pari a X_m = 550.*

Poniamo l'ipotesi nulla:
$H_0 : \mu = 250$
che significa che non vi è stato inquinamento e maggior concentrazione del campione rispetto alla popolazione è imputabile solo al caso (errore di campionamento o di analisi). L'ipotesi alternativa è:
$H_1 : v = 550 > \mu$;
cioè il pozzo è inquinato e di conseguenza la concentrazione dell'acqua è cresciuta fino ad un valore v, chepuò essere stimato tramite la media X_m del campione .
Possiamo ora calcolare la probabilità di estrarre da una popolazione normale N(250, 180) un campione con media pari a 550.
Il calcolo di probabilità ci dice che (si rimanda a testi specializzati per la dimostrazione) data una popolazione normale N(μ,σ) la media dei campioni di n *individui estratti da questa popolazione si distribuisce normalmente con media μ e deviazione standard pari all'errore standard:*

$$e.s. = \frac{\sigma}{\sqrt{n}}$$

Pertanto, la probabilità che da N(250, 180) si estragga un campione con media pari o superiore 550 è pari a (in R):

```
> 1-pnorm(550,mean=250,sd=(180/sqrt(3)))
[1] 0.001946209
>
```

Si tratta quindi di una caso che fa parte dello 0.19 % di casi più rari. Quindi la probabilità di osservare un campione come il nostro con l'ipotesi nulla vera è ben al disotto del livello prefissato, pari ad α=0.05. Per questo motivo possiamo rifiutare l'ipotesi nulla e concludere che effettivamente il pozzo è inquinato.

Errore di prima e seconda specie

Il test d'ipotesi ci ha condotto ad una decisione che è quella di rifiutare l'ipotesi nulla. Ma siamo proprio sicuri che la nostra decisione corrisponda alla "verità vera"? No, perchè mentre noi abbiamo deciso che è vera l'ipotesi alternativa, in realtà potrebbe essere vera l'ipotesi nulla. Qual è la probabilità che ci sbagliamo? Noi sbagliamo a rifiutare l'ipotesi nulla se essa in realtà è vera; siccome questa opzione (ipotesi nulla vera) ha lo 0.19% di probabilità di verificarsi, la probabilità d'errore corrisponde appunto allo 0.19%. Quest tipo di errore (rifiutare erroneamente l'ipotesi nulla) si dice **errore di prima specie**, che è appunto ≤ al livello α prefissato.

Esempio VII.2
Nel caso dell'esempio precedente, poniamo di aver rilevato una concentrazione media del campione pari a 400. In questo caso la probabilità è pari a

```
> 1-pnorm(400,mean=250,sd=180/sqrt(3))
[1] 0.07445734
```

ed è quindi inferiore al livello α prefissato. Pertanto, possiamo accettare l'ipotesi nulla.

Nel caso di accettazione dell'ipotesi nulla, potremmo commettere un errore se in realtà è vera l'ipotesi alternativa (cioè che il campione è in realtà estratto da una popolazione con media $\nu > \mu$). Il rischio di commettere questo errore (detto di seconda specie, cioè accettare erroneamente l'ipotesi nulla) si corre fino a che la concentrazione media del campione è pari a:

```
> qnorm(0.95, mean=250, sd=180/sqrt(3))
[1] 420.9382
```

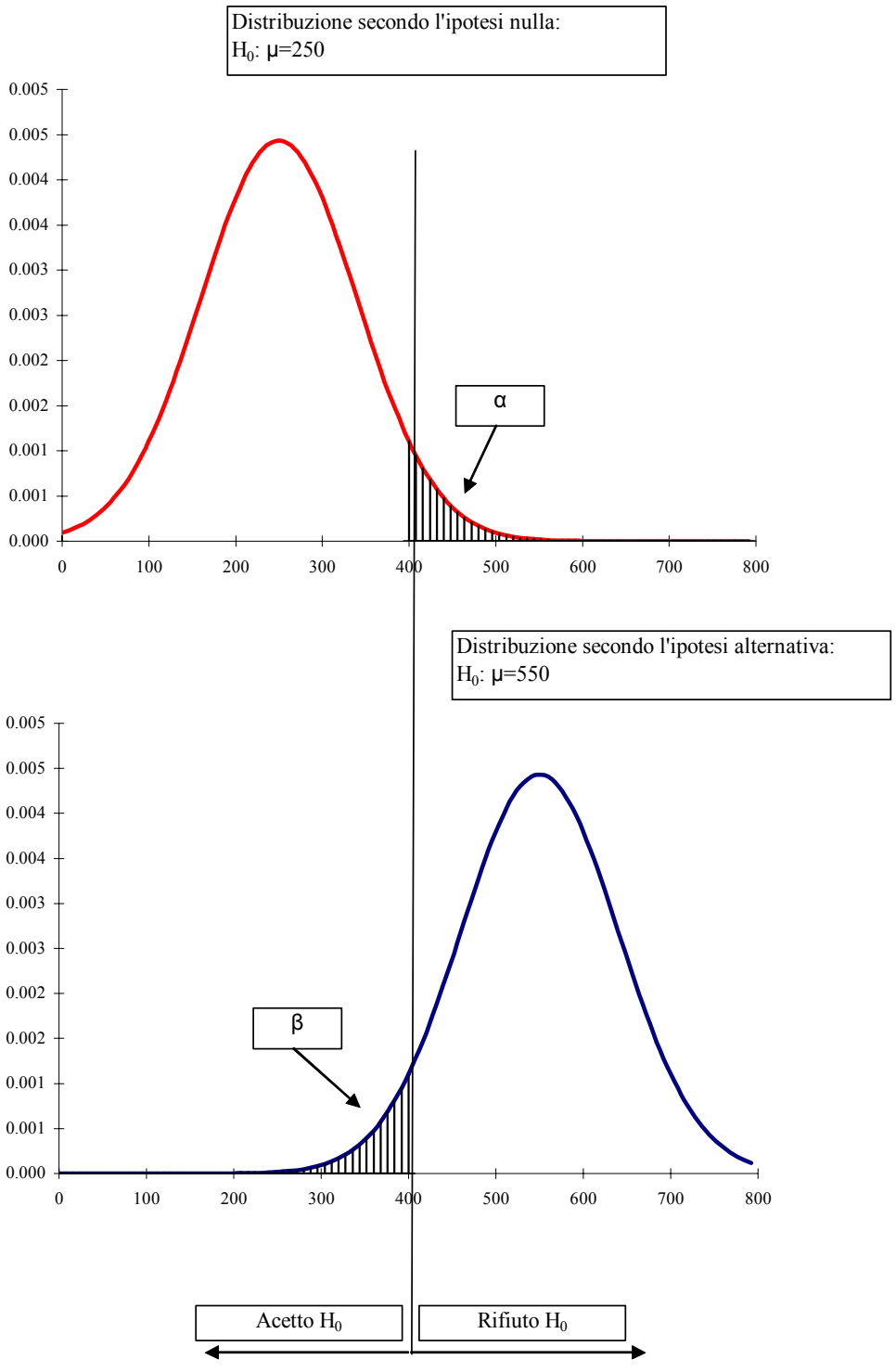


Figura VII.1. Esemplificazione grafica dell'errore di I e II specie

Infatti, con una concentrazione pari a 421, la probabilità di errore di prima specie è appunto pari a 0.05. In sostanza, 421 è il valore critico, oltre il quale scegliamo di rifiutare l'ipotesi nulla. In altre parole, il nostro test non riesce a mettere in luce nessun inquinamento se la concentrazione è inferiore a 421. Ma se la concentrazione è pari a 421 (valore limite) e accettiamo l'ipotesi nulla, la nostra probabilità di errore di seconda specie (ipotizzando ν=550, ad esempio) è pari a:

```
> pnorm(421,mean=550,sd=180/sqrt(3))
[1] 0.1072469
```

cioè al 10.7 % circa. La probabilità d'errore di seconda specie è tanto più bassa quanto più alto è il valore di ν e quanto è più bassa la variabilità delle popolazioni, che può essere ridotta con una sperimentazione accurata o con l'aumento della numerosità del campione.

L'errore di seconda specie è detto β e il suo complemento 1 - β è detto potenza del test (89.3%) in quanto è la probabilità di mettere in luce le differenze effettivamente esistenti, cioè di rifiutare correttamente l'ipotesi nulla.

Ipotesi alternative semplici e complesse

Nel caso precedente abbiamo valutato un'ipotesi alternativa complessa H_1: $\nu > \mu$. Se *a priori* non abbiamo elementi per sostenere che $\nu > \mu$, possiamo testare ipotesi alternative semplici, del tipo $\nu \neq \mu$. In questo caso, la probabilità di errore α deve essere considerata a destra e a sinistra della media, 2.5% per parte (test a due code), in quanto ν potrebbe essere sia maggiore che minore di μ. Di questo deve essere tenuto conto nel calcolo di probabilità con R.

Confronto tra due medie: il test t di Student

Analizziamo un primo esempio pratico di test d'ipotesi: nella sperimentazione agraria si ha spesso interesse a considerare due popolazioni per scoprire se queste sono diverse per il carattere o i caratteri considerati. Più in particolare, siccome ognuna delle popolazioni sarà descritta dalla sua media, saremo interessati a rispondere al quesito se **l'eventuale differenza rilevata tra le due medie e da ritenersi una differenza reale, effettiva e con un preciso significato biologico**. In sostanza, in termini statistici, dovremo stabilire se la differenza tra le medie è *significativa* oppure da attribuire a fattori casuali e quindi *non significativa*.

E' intuitivo comprendere che, anche se il problema può sembrare banale, esso non lo è; basti ripensare al fatto che ogni media stimata si porta dietro un alone di incertezza, definito appunto dall'intervallo di confidenza.

> *Esempio VII.3*
> *Un ricercatore ha coltivato due varietà di grano con diverse caratteristiche delle cariossidi (VICTO e LUCREZIA), per valutare quale delle due ha. Per ciascuna delle due varietà ha coltivato 3'500'000 piante circa. Alla fine dell'esperimento ha determinato il peso ettolitrico della granella. Questa determinazione non può essere eseguita su tutta la massa della granella, ma su un quantitativo di poche decine di grammi di cariossidi; di conseguenza lo sperimentatore, dopo aver accuratamente mescolato la massa di granella di ciascuna varietà, estrae un campione di cinque contenitori di granella da 50 g ed esegue quindi cinque determinazioni per varietà. E' evidente che i cinque contenitori di granella sono un campione casuale, scelto tra tutti quelli che si sarebbero potuti estrarre da ciascuna varietà di grano; si è scelto di eseguire l'analisi su cinque contenitori per migliorare la stima del peso ettolitrico medio delle due varietà di frumento, diminuendo*

l'importanza di eventuali inaccuratezze nell'analisi e nel campionamento. E'anche evidente come il peso ettolitrico del frumento è una caratteristica soggetta ad una certa variabilità naturale, legata al fatto che le cariossidi non sono tutte uguali e alla possibilità di commettere errori nel campionamento e nella misurazione del peso ettolitrico stesso.

I risultati sono i seguenti:

VICTO (peso ettolitrico): 65 – 68 –69 –71 – 78; la media per questa varietà è pari a 70.2, mentre la deviazione standard è pari a 4.87
Possiamo quindi calcolare l'errore standard che è pari a 2.18 e quindi l'intervallo di confidenza della media, che è pari a 70.2 ± 6.04

LUCREZIA (peso ettolitrico): 70 – 71 – 74 – 78 – 84: la media è 75.4, mentre la deviazione standard è pari a 5.73
In questo caso l'errore standard è pari a 2.56, mentre l'intervallo di confidenza per la media è pari a 75.4 ± 7.11

Possiamo affermare che la varietà Lucrezia ha un peso ettolitrico più alto di Victo?
Questa semplice domanda ci mette in difficoltà: è evidente infatti che il peso ettolitrico medio di LUCREZIA è maggiore di quello di VICTO, ma è anche vero che dei cinque dati relativi a LUCREZIA, solo uno è superiore a tutti quelli relativi a VICTO. E' anche vero che esiste un certo margine di variabilità intorno alla media, misurato dal coefficiente di variabilità, che in qualche modo rende incerta la nostra stima. Cosa sarebbe successo se avessimo effettuato un numero superiore di analisi? Inoltre, si può osservare che il limite di confidenza superiore per VICTO (70.2 + 6.04 = 76.24) è inferiore al limite di confidenza superiore per LUCREZIA (75.4 – 7.11 = 68.29).

Nell'approccio dell'esercizio soprastante si manifesta tutto il potenziale della statistica inferenziale, che ci consente di prendere decisioni in casi dubbi come questo.

E' evidente che la decisione dovrà essere basata su due aspetti:

1) l'ampiezza della differenza tra le medie: più la differenza tra le due medie è alta e più è probabile che essa sia significativa;

2) l'ampiezza dell'errore standard. Più è elevata la variabilità dei dati e quindi l'errore di stima è più è bassa la probabilità che le differenze osservate tra le medie siano significative.

Questi due aspetti sono stati utilizzati per definire il cosiddetto **test di t** :

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$$

Si può osservare che il test di t in realtà non è altro che il rapporto tra le quantità indicate in precedenza ai punti 1 e 2: infatti la quantità al numeratore è la differenza tra le medie dei due campioni, mentre la quantità al denominatore è il cosiddetto errore standard della differenza tra due medie, che si calcola a partire dalla media ponderata delle deviazioni standard dei due campioni, secondo la formula seguente:

$$s_{\bar{X}_1-\bar{X}_2} = \sqrt{2\bar{s}^2 \frac{n_1+n_2}{n_1 n_2}}$$

Dove $\bar{s}^2$ è la varianza mediata dei due campioni e si calcola sommando le devianze dei due campioni e dividendole per la somma dei rispettivi gradi di libertà:

$$\bar{s}^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2}$$

Secondo quanto detto in precedenza, è evidente che t assume un valore pari a 0 se le due medie sono uguali e si scosta dallo zero sia in senso positivo che negativo in modo tanto più elevato quanto più elevata è la differenza osservata tra le due medie. L'ipotesi nulla può essere posta in questi termini:

$$H_0 : \mu_1 = \mu_2$$

ove μ_1 e μ_2 sono le medie delle popolazioni da cui sono estratti i due campioni, stimate dalle quantità X_1 e X_2. Se l'ipotesi nulla è vera, valori di t diversi da zero sono tanto meno probabili quanto più grandi in valore assoluto.

Si può dimostrare che nel caso in cui è vera l'ipotesi nulla, il valore di t calcolato si distribuisce seguendo la variabile casuale *t di Student*. Utilizzando la legge definita per questa variabile casuale possiamo calcolare la probabilità che ha di verificarsi il valore di t calcolato nel caso in cui l'ipotesi nulla è vera.

L'ipotesi alternativa semplice può essere definita:

$$H_0 : \mu_1 \neq \mu_2$$

Se abbiamo elementi sufficienti, possiamo anche adottare ipotesi alternative complesse, del tipo

$$H_0 : \mu_1 < \mu_2 \quad oppure \quad H_0 : \mu_1 > \mu_2$$

ma queste ipotesi alternative debbono essere ragionevolmente fatte prima di eseguire l'esperimento, non dopo aver visto i risultati.

CASO STUDIO VII.1
Un prodotto in grado di diminuire la crescita delle piante (brachizzante) viene spruzzato su quattro parcelle di orzo e messo a confronto con quattro parcelle trattate solo con acqua. Al termine del ciclo produttivo, viene rilevata l'altezza dell'orzo; i risultati sono i seguenti:

TRATTATO: 85, 78, 91, 81
NON TRATTATO: 101 95 89 94

Stabilire se il trattamento ha avuto effetto.

Come è evidente, si tratta di effettuare un confronto tra le medie di due campioni composti da quattro unità sperimentali e sottoposti a due diversi trattamenti sperimentali (brachizzato e non brachizzato). L'ipotesi nulla è che il trattamento brachizzante non ha avuto effetto e quindi che le due medie non sono significativamente diverse tra di loro. In altre parola l'ipotesi è che la differenza osservata sia dovuta solo al caso (o meglio all'errore). Per testare questa ipotesi nulla impostiamo un test di t.
I risultati ed i calcoli necessari in R sono riportati di seguito. In particolare,

si utilizza la funzione `t.test`.

```
> x<-c(85,78,91,81)
> y<-c(101,95,89,94)
>  t.test(x,y,alternative="two.sided",paired=FALSE,
var.equal=TRUE)

        Two Sample t-test

data:  x and y
t = -2.9443, df = 6, p-value = 0.02580
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -20.141871  -1.858129
sample estimates:
mean of x mean of y
    83.75     94.75
>
```

Si può notare come il test di t porta ad un valore di t calcolato pari a -2.944, corrispondente ad una probabilità di 0.0258, che è inferiore alla soglia α*=0.05. Possiamo quindi rifiutare l'ipotesi nulla e affermare che il trattamento brachizzante ha avuto un effetto significativo. Nel fare questa affermazione abbiamo una probabilità d'errore di primo tipo pari al 2.58%. Notare ancora che abbiamo richiesto un test di t a due code (two.sided) per esplorare l'ipotesi alternativa semplice, abbiamo supposto l'omogeneità delle varianze (var.equal=TRUE: i due campioni sono estratti da popolazioni con varianze incognite ma uguali) e che le osservazioni non sono appaiate (ad esempio non sono eseguite sulle stesse unità sperimentali in tempi diversi).*
I limiti di confidenza riportati sono quelli relativi alla differenza tra le due medie, che è appunto pari all'errore standard della differenza (SED), moltiplicato per il valore di t corrispondente ad una probabilità pari a 0.05 (0.025 per coda) e 6 gradi di libertà (pari ad $[(n_1-1)+(n_2-2)]$*, dove* n_1 *ed* n_2 *sono le numerosità dei due campioni a confronto). Il SED si calcola:*

$$SED = \sqrt{\bar{s}^2 \frac{n_1 + n_2}{n_1 n_2}}$$

Dove $\bar{s}^2$ *è la varianza mediata dei due campioni e si calcola sommando le devianze dei due campioni e dividendole per la somma dei rispettivi gradi di libertà:*

$$\bar{s}^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2}$$

ossia:

```
> SSX=var(x)*3
> SSY<-var(y)*3
> SXY<-(SSX+SSY)/6
> SED<-sqrt(SXY/2)
> SED*qt(0.025,6)
[1] -9.141871
```

```
> (mean(x)-mean(y))+SED*qt(0.025,6)
[1] -20.14187
> (mean(x)-mean(y))-SED*qt(0.025,6)
[1] -1.858129>
```

Confronto tra una proporzione osservata ed una proporzione attesa: metodo esatto

Il test di t è molto utile, ma soltanto nel caso in cui si abbia a che fare con caratteri quantitativi, cioè con variabili misurate su una scala continua, per le quali sia possibile calcolare delle statistiche descrittive, come appunto la media.

In molti casi, gli sperimentatori sono interessati a rilevare alcune caratteristiche qualitative, come ad esempio lo stato di una pianta in seguito ad un trattamento (morta o viva), il colore dei semi (si ricordino i piselli verdi e gialli di Mendel) ed altre caratteristiche che non sono misurabili facilmente su una scala continua.

Avendo a che fare con variabili qualitative l'unico dato che si può rilevare è la proporzione degli individui che presentano una certa modalità. Dovendo confrontare tra loro due proporzioni osservate, non possiamo utilizzare il test di t, ma dobbiamo ricorrere ad un altro test, detto della soluzione esatta di Fisher.

Illustriamo questo test con un esempio.

> *Esempio VII.4*
> *Immaginiamo di avere un erbicida (composto chimico in grado di controllare le piante infestanti) e di sapere che, in condizioni controllate di temperatura e umidità e salvo oscillazioni di efficacia casuali, questo fitofarmaco è in grado di controllare il 75% (π_0) delle piante di* Solanum nigrum*, in una data fase fenologica. Organizziamo un esperimento per sapere se utilizzando questo erbicida in miscela con un coadiuvante possiamo incrementare la sua fitotossicità. Trattiamo 50 piante di* Solanum nigrum *con una miscela erbicida+coadiuvante ed otteniamo un numero di piante morte pari al 96%. Si tratta di un'oscillazione di efficacia casuale o possiamo concludere che il coadiuvante ha avuto un effetto significativo?*

Nel caso accennato, l'ipotesi nulla può essere definita in questo modo:

$H_0 : \pi_1 = \pi_0$

mentre l'ipotesi alternativa semplice può essere definita come:

$H_0 : \pi_1 \neq \pi_0$

Dobbiamo fare riferimento all'ipotesi alternativa semplice in quanto non sappiamo *a priori* se il coadiuvante è in grado di innalzare l'efficacia erbicida, infatti molte sostanza supposte coadiuvanti possono invece diminuire l'efficacia delle molecole con cui vengono miscelate. Nel caso in cui l'ipotesi alternativa è vera, il valore di π_1 rimane ignoto, mentre la sua stima sarà data da $p_1 = 0.96$.

Se poniamo vera l'ipotesi nulla possiamo utilizzare la variabile casuale binomiale per calcolare la probabilità di estrarre un campione con una proporzione pari a 0.96 da una popolazioni con una proporzione pari a 0.75. Se questa probabilità fosse inferiore ad $\alpha = 0.025$ (test bilaterale e a due code!) rifiutiamo l'ipotesi nulla.

> *Esempio VII.4 (segue)*
> *La probabilità di ottenere p=0.96 a partire da una popolazione in cui π_0 = 0.75 con una numerosità del campione pari a n=50 è data da (in R):*

```
> pbinom(48,50,0.75,lower.tail=FALSE)
```

```
[1] 1.000502e-05
>
```

L'ipotesi nulla può essere rifiutata, con una probabilità d'errore molto bassa.

Nell'utilizzare la funzione di distribuzione binomiale in R, bisogna ricordare che:

```
pbinom(k, n, π)
```

calcola la probabilità cumulata da 0 a k. Se specifichiamo l'opzione `lower.tail=FALSE`, la funzione restituisce la probabilità cumulata da k+1 a n.

Confronto tra due frequenze osservate: il test di χ^2

In questo caso non abbiamo informazioni sulla proporzione della popolazione intera, ma sono disponibili solo due proporzioni osservate. Anche in questo caso, il test verrà illustrato con l'aiuto di un esempio.

Esempio VII.5
Per verificare se un coadiuvante aumenta l'efficaca di un erbicida organizziamo un esperimento in cui utilizziamo l'erbicida da solo e con il coadiuvante. Nel primo caso (erbicida da solo) otteniamo 56 morti su 75 piante trattate, mentre nel secondo caso otteniamo 48 morti su 50 piante trattate.

Si può notare che la differenza con l'esempio precedente risiede nel fatto che in questo caso non sappiamo *a priori* la proporzione di morti per la popolazione trattata, ma dobbiamo stimarla a partire dai dati sperimentali.

In questo caso i risultati del nostro esperimento si riducono ad una tabella di contingenza di questo tipo:

	Piante morte	**Piante vive**	**Totale**
Erbicida	56	19	*75*
Erbicida + coadiuvante	48	2	*50*
Totale	*104*	*21*	*125*

Abbiamo già visto nel capitolo IV come è possibile costruire un indice di indipendenza (detto χ^2) che misura la quantità della dipendenza tra le due variabili rilevate. Ci stiamo infatti chiedendo se il numero dei morti è indipendente dal tipo di trattamento oppure no. Ricordiamo che questo indice è pari a:

$$\chi^2 = \sum \frac{(f_o - f_a)^2}{f_a}$$

dove f_o sta per frequenza osservata ed f_a sta per frequenza attesa nel caso in cui sia verificata l'ipotesi nulla.

Sappiamo già che questo indice è pari a 0 in caso di indipendenza e cresce progressivamente se esiste un certo grado di dipendenza tra le due variabili. Siccome siamo in ambito inferenziale dobbiamo ricordare che a noi non interessa il confronto tra proporzioni osservate nel caso in studio,

ma interessa il confronto delle proporzioni delle popolazione da cui i due campioni (uno di 75 l'altro di 50 individui) sono stati estratti. Poniamo l'ipotesi nulla in questi termini:

$$H_0 : \pi_1 = \pi_2 = \pi$$

ove π_1 è la proporzione della popolazione da cui è estratto il campione trattato solo con erbicida e π_2 è la proporzione della popolazione da cui è estratto il campione trattato con erbicida+coadiuvante. Visto che noi abbiamo solo delle stime di π_i e π_2, queste sono soggette alle normali oscillazioni legate al campionamento sperimentale e quindi potrebbe verificarsi una valore di χ^2 maggiore di 0 per il semplice effetto del caso. Ma quanto è probabile questa evenienza. A questa domanda possiamo rispondere considerando che, se n è sufficientemente grande (n>30) il valore osservato di χ^2 segue appunto la distribuzione della variabile casuale χ^2, con un numero di gradi di libertà ν pari al numero dei confronti (in questo caso 1). Possiamo quindi utilizzare la funzione `pchisq(x, df= ν)` per calcolare la probabilità relativa al caso dell'ipotesi nulla vera.

Esempio VII.5 (segue)
Calcoliamo il valore di χ^2 come segue:

```
> summary(as.table(tabella))
Number of cases in table: 125
Number of factors: 2
Test for independence of all factors:
        Chisq = 9.768, df = 1, p-value = 0.001776
>
```

Il valore è pari a 9.768, che nella distribuzione della variabile casuale χ^2 ha una probabilità pari a (probabilità da 9.768 a +infinito):

```
> 1-pchisq(9.768,1)
[1] 0.001775755
```

Notare che il valore di probabilità è già fornito come output della funzione `summary` *applicata ad una tabella di contingenza.*
Sulla base del valore di probabilità calcolato, posso rifiutare l'ipotesi nulla con una probabilità d'errore pari allo 0.178%.
Allo stesso risultato si può pervenire per altra via utilizzando la funzione `chisq.test`*, come vedremo nel caso studio seguente.*

CASO STUDIO VII.1
Si vuole valutare l'efficacia di due insetticidi (A e B). 257 insetti vengono trattati con A e 244 con B. Nel primo caso muoiono 41 insetti, nel secondo caso ne muoiono 64. Possiamo concludere che gli insetticidi sono ugualmente efficaci?
Anche in questo caso si tratta di una variabile qualitativa (insetto vivo o morto) rilevata su due campioni rispettivamente di 257 e 244 insetti, trattati in due modi diversi (insetticida A ed insetticida B). L'ipotesi nulla è che i due insetticidi non differiscono in modo significativo.
L'esempio è analogo a quello precedente, solo che in questo caso la frequenza attesa non è esplicita e va anche essa stimata a partire dai dati sperimentali: di fatto si tratta di confrontare due frequenze osservata (e non una frequenza osservata con una teorica). Si tratta quindi di un test d'indipendenza: vogliamo sapere se le frequenze sono indipendenti dal tipo di trattamento.

I dati sono stati esemplificati nella tabella sottostante

	Morti	**Vivi**	**Totale**
Insetticida A	41	216	*257*
Insetticida B	64	180	*244*
Totale	*105*	*396*	*501*

Se fosse vera l'ipotesi nulla, la frequenza dei morti per i due insetticidi dovrebbe essere uguale e pari a (41+64)/(257+244), cioè a 105/501=0.210.

Quindi le frequenze attese sono le seguenti:

	Morti	**Vivi**	**Totale**
Insetticida A	53.86	203.15	*257*
Insetticida B	51.138	192.86	*244*
Totale	*105*	*396*	*501*

Possiamo ora calcolare il test in questo modo:

$$\chi^2 = \frac{(41-53.86)^2}{53.86} + \frac{(64-51.138)^2}{51.138} + \frac{(216-203.15)^2}{203.15} + \frac{(180-192.86)^2}{192.86} = 7.979$$

I gradi di libertà sono soltanto uno, perché abbiamo un solo confronto (tra A e B). Il valore tabulato è quindi 3.841; possiamo concludere che l'ipotesi nulla può essere rifiutata e possiamo quindi affermare che l'insetticida B è più efficace di A.

LISTATO R

```
> tab<-matrix(c(41,64,216,180),2,2)
> rownames(tab)<-c("A","B")
> colnames(tab)<-c("Morti","Vivi")
> tab
  Morti Vivi
A    41  216
B    64  180
> chisq.test(tab,correct=FALSE)

        Pearson's Chi-squared test

data:  tab
X-squared = 7.9789, df = 1, p-value = 0.004733
```

OPPURE

```
> a<-c("A","B")
> b<-c("M","V")
> tab<-as.table(matrix(c(41,64,216,180),2,2,dimnames=list(a,b)))
> tab
   M   V
A 41 216
B 64 180
> summary(tab)
Number of cases in table: 501
Number of factors: 2
Test for independence of all factors:
        Chisq = 7.979, df = 1, p-value = 0.004733
>
```

Confronto tra più di due trattamenti: il test F nell'ANOVA

Nella pratica della ricerca sperimentale in genere il ricercatore esegue più di due trattamenti e quindi il suo obiettivo è quello di confrontare tra loro parecchie medie. Compiere una serie di test di t operando sulle medie prese a coppie è un lavoro abbastanza tedioso e per questo esiste una procedura molto più pratica è potente che è detta analisi della varianza (ANOVA) e che è basata su un test di confronto tra varianze, detto test F di Fisher.

L'analisi della varianza è uno strumento molto complesso, ma anche estremamente potente, che consente di risolvere un ampio spettro di problemi statistici. Non è possibile in questa sede una trattazione approfondita della materia, tuttavia si ritiene interessante riportare un semplice esempio dell'uso di questa tecnica.

Esempio VII.6
Un ricercatore vuole valutare l'effetto di tre ceppi di microrganismi (A, B e C) sulla degradazione di un pesticida nel terreno. A questo fine prepara 12 campioni di terreno e li contamina con una concentrazione nota di erbicida, uguale per tutti i campioni. Successivamente aggiunge i tre ceppi di microrganismi, in modo da contaminare, con ciascun ceppo, quattro campioni di terreno, scelti a caso. Successivamente pone i dodici campioni di terreno in cella climatica alle medesime condizioni di temperatura, illuminazione ed umidità. Dopo 21 giorni analizza la concentrazione dell'erbicida nel terreno, riscontrando i seguenti valori (in μg/g):

Replica	*Ceppo A*	*Ceppo B*	*Ceppo C*
1	**150**	**121**	**115**
2	**161**	**125**	**111**
3	**152**	**131**	**120**
4	**149**	**128**	**122**
Media	153.00	126.25	117.00
Devianza	90	54.75	74

L'analisi della varianza (ANOVA) è basata sul presupposto che la variabilità totale (devianza totale) dei dati sperimentali, può essere scomposta in due quote: la devianza dovuta al trattamento (SS_t) e la varianza dovuta all'errore sperimentale (SS_e). Ognuna delle due devianze può essere divisa per i rispettivi gradi di libertà per ottenere le relative varianze (QM_t e QM_e). Il trattamento è significativo se è solo se induce nei dati sperimentali una varianza superiore a quella dell'errore, cioè se il valore:

$$F = \frac{QM_t}{QM_e}$$

è superiore ad 1. Ancora una volta dobbiamo considerare che QM_e e QM_t come tutte le grandezze stimate è soggetta ad oscillazione casuale che segue la distribuzione della variabile casuale F di Fisher. Questa variabile casuale può essere utilizzata per il calcolo di probabilità del valore calcolato di F con l'ipotesi nulla vera, cioè:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_n = \mu$$

Esempio VII.6 (segue)
La variabilità totale dei 12 dati (devianza = 3014.917, con 11 gradi di libertà) può essere scomposta in due quote:

1 – Devianza del trattamento: che fa variare la media di ogni trattamento rispetto alla media generale. Questa quota è legata all'effetto del trattamento in studio ed è pari a 2796,16. Questo valore si ottiene dalla devianza delle tre medie (699,04) moltiplicata per il numero di repliche (4). La devianza dei trattamenti ha 2 gradi di libertà (numero dei trattamenti –1).

2 – Devianza dell'errore: che è la somma delle devianze tra i quattro individui trattati di ogni trattamento. È evidente che le differenze tra gli individui trattati allo stesso modo non possono che essere imputate all'errore sperimentale (devianza dell'errore = 218.75, con 9 gradi di libertà, tre per ogni trattamento).

È possibile osservare che se sommiamo la devianza e i gradi di libertà dell'errore con quelli dei trattamenti, otteniamo la devianza e i gradi di libertà totali.
Dalle rispettive devianze possiamo calcolare le varianze, secondo la seguente tabella dell'analisi della varianza:

Fonte della variazione	*Devianza*	*GL*	*Varianza*	*F*
Trattamenti	2796.167	2	1398.083	57.52114
Errore	218.75	9	24.30556	
Totale	3014.917	11		

Il principio è che, se l'ipotesi nulla è vera i trattamenti non hanno avuto effetto e quindi non possono aver indotto una variabilità dei dati superiore a quella dell'errore.

In R questo calcolo sii compie direttamente con la funzione `lm`*,ricordandosi però di specificare che la variabile esplicativa x è quantitativa, con il comnado* `factor`*.*

1 - Immissione dati

```
> x<-c(1,1,1,1,2,2,2,2,3,3,3,3)
> y<-c(150,161,152,149,121,125,131,128,115,111,120,122)
> x<-factor(x)
```

2 - Calcolo delle medie

```
> by(y,x,mean)
INDICES: 1
```

```
[1] 153
------------------------------------------------------------
INDICES: 2
[1] 126.25
------------------------------------------------------------
INDICES: 3
[1] 117
>
```

3 - Calcolo ANOVA

```
> model<-lm(y~x)
> anova(model)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          2 2796.17 1398.08  57.521 7.465e-06 ***
Residuals  9  218.75   24.31
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

L'ipotesi nulla può quindi essere rifiutata con una probabilità d'errore molto bassa (7.465 x 10^{-6}).

La Minima Differenza Significativa (MDS)

I calcoli precedenti ci portano ad affermare che i tre ceppi di microrganismi influenzano in modo diverso la degradazione dell'erbicida in studio. Questa informazione, tuttavia, non risolve ancora il problema iniziale: sappiamo infatti che i ceppi A, B e C non sono uguali, ma che almeno uno è diverso dagli altri; tuttavia non sappiamo quale dei tre è diverso dagli altri e soprattutto quale è il ceppo che più velocemente degrada l'erbicida in studio.

A questo proposito però la varianza dell'errore ci permette di calcolare la minima differenza significativa (MDS), con la formula seguente

$$\mathrm{MDS}(p < 0.05) = t \times \sqrt{\frac{2 \times \mathrm{Var}_{err}}{n}}$$

ove t è il valore di *t* tabulato per la probabilità desiderata (nel nostro caso 0.05) e per un numero di gradi di libertà pari alla varianza dell'errore, var_{err} è la varianza dell'errore ed *n* è il numero di repliche. Questa minima differenza significativa non è altro che il valore critico da adottare per ognuno dei confronti possibili tra medie e coincide con il calcolo del limite di confidenza medio, sulla base della variabilità d'errore. Se la differenza tra due medie supera la MDS. rifiutiamo l'ipotesi nulla e concludiamo che le due medie sono diverse per una probabilità di errore inferiore al livello α prefissato.

Esempio VII.6
Nel nostro caso:

$$\mathrm{MDS}(p < 0.05) = 2.262 \times \sqrt{\frac{2 \times 24.31}{4}} = 7.87$$

In R:

```
>sqrt(2*anova(model)$Mean[2]/(anova(model)$Df[2]))*qt(0.975,anova(model)$Df[2])
[1] 7.886069
```

Ciò significa che se la differenza tra due medie eccede il valore di 7.87, questa deve essere considerata significativa. In questo modo è possibile fare tutti i confronti a coppie tra le medie ed è quindi possibile appurare che il ceppo B è stato quello che ha consentito la più veloce degradazione dell'erbicida in studio, in quanto la sua differenza con B è significativa (C-B = 19,25 che è superiore alla MDS). Allo stesso modo C è stato superiore ad A (A – C = 16 che è superiore alla MDS).
Possiamo quindi concludere che tutti e tre i trattamenti si sono comportati diversamente.

Il test d'ipotesi nell'analisi di regressione

Quando le unità sperimentali in studio sono un campione di una popolazione più ampia, anche l'analisi di regressione presenta gli stessi margini di incertezza già discussi in precedenza, riassumibili nella banda di inferenza a cui è associata la stima puntuale dei valori di ogni parametro (pendenza ed intercetta nel caso della regressione lineare). Di questo non si può non tenere conto quando si vogliano prendere decisioni statistiche nell'analisi di regressione.

Le ipotesi che più frequentemente un ricercatore vuole saggiare nell'ambito dell'analisi di regressione sono le seguenti:

1 - Si può concludere che il valore di un parametro (ad es. la pendenza della retta) è significativamente diverso da zero?
2 - Si può concludere che la retta di regressione nel suo complesso descrive in modo adeguato i dati sperimentali?

Nel primo caso, l'ipotesi nulla può essere formulata in questo modo:

$$H_0 : \beta_1 = 0$$

mentre l'ipotesi alternativa è:

$$H_0 : \beta_1 \neq 0$$

Per valutare la veridicità dell'ipotesi nulla possiamo far riferimento alla distribuzione *t* di Student, dato che il rapporto tra il valore di un parametro stimato e il suo errore standard si distribuisce appunto coma la variabile casuale citata in precedenza:

$$t_\nu = \frac{b_i}{s_{bi}}$$

dove ν è il numero di gradi di libertà dell'errore (residuo della regressione).

Per quello che riguarda la seconda domanda (la regressione è una buona descrizione dei dati sperimentali?) è possibile effettuare la cosiddetta Analisi della Varianza della Regressione, che consiste nel suddividere la devianza totale dei dati in due componenti: quella dovuta alla regressione e quella dovuta all'errore, analogamente a quanto già visto per l'ANOVA nel capitolo precedente.

In questo caso, la devianza dell'errore della regressione (residuo) è facilmente calcolata come la somma dei quadrati degli scostamenti tra i valori attesi secondo la retta di regressione e i valori

osservati.

Una regressione non è significativa se la devianza che essa spiega non è superiore a quella spiegata dall'errore. La grandezza:

$$F = \frac{QM_{regressione}}{QM_{residuo}}$$

si distribuisce secondo la variabile casuale F di Fischer, con i gradi di libertà relativi alle grandezze al numeratore e al denominatore.

CASO STUDIO VII.2

Consideriamo il seguente campione (quattro individui) già considerato nel capitolo precedente:

Num. dato	Ricoprimento piante infestanti (X)	Produzione mais (Y)
17	5.00	12.75
33	12.41	11.18
35	20.05	12.25
71	65.75	10.59

Eseguiamo su di essi l'analisi di regressione e valutiamo la bontà della regressione, verificando anche che nessuno dei parametri stimati sia uguale a zero.

Analogamente a quanto già fatto in precedenza:

1 - Immettiamo i dati

```
> x<-c(5,12.41,20.05,65.75)
> y<-c(12.75,11.18,12.25,10.59)
```

2 - Eseguiamo la regressione:

```
model<-lm(y~x)
```

3 - Controlliamo l'output in termini di stima dei parametri:

```
Call:
lm(formula = y ~ x)

Residuals:
        1         2         3         4
 0.478407 -0.885316  0.397364  0.009545

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.41078    0.56588  21.932  0.00207 **
x           -0.02784    0.01616  -1.722  0.22712
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` '
1

Residual standard error: 0.7651 on 2 degrees of freedom
Multiple R-Squared: 0.5973,     Adjusted R-squared: 0.396
F-statistic: 2.967 on 1 and 2 DF,  p-value: 0.2271
```

Possiamo osservare come la pendenza in effetti non è significativamente

diversa da 0 con limite critico decisionale α = 0.05. Allo stesso modo la regressione non è significativa, dato che la statistica F è pari a 2.967, che corrisponde ad una probabilità di errore nel rifiutare l'ipotesi nulla pari a 0.2271. Notare che la funzione:

```
> pf(2.967,1,2,lower.tail=FALSE)
[1] 0.2271207
```

restituisce lo stesso risultato. Questo test di F è unilaterale, in quanto si testa l'ipotesi che F osservato sia strettamente maggiore di 1.

4 - Esecuzione dell'ANOVA della regressione
Agli stessi risultati si può pervenire anche applicando il comando `anova` *all'output del comando* `lm`*, grazie ad una più comprensibile tabellina ANOVA.*

```
> anova(model)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value Pr(>F)
x          1 1.73663 1.73663   2.967 0.2271
Residuals  2 1.17065 0.58532
```

ESERCIZI PROPOSTI

1) Otto parcelle di mais perfettamente uguali sono state concimate con due tipi di concimi (quattro parcelle per ogni concime). Al momento della raccolta sono state prelevate 5 piante per appezzamento e, per ciascuna pianta, è stato determinato il peso della granella prodotta. I risultati sono come segue:

Concime 1 (t/ha)	***Concime 2 (t/ha)***
120	*114*
115	*113*
112	*116*
121	*109*
116	*107*

Determinare:

1) Le statistiche descrittive dei due campioni (media, varianza, deviazione standard);
2) Calcolare gli intervalli di confidenza delle due medie;
3) Verificare se il le due concimazioni hanno avuto un diverso effetto sulla produzione della coltura

2) Due varietà di tabacco sono state inoculate con lo stesso virus. Dopo l'inoculazione, da ciascuna popolazione è stato estratto una campione di 100 piante ed è stato contato il numero di individui malati. E' risultato che nel campione tratto dalla popolazione A si sono

riscontrati 25 individui malati, mentre nel campione estratto dalla popolazione B si sono riscontrati 50 individui malati. Verificare se le due popolazioni sono caratterizzate da un diverso grado di sensibilità al virus in studio.

3) Considerando una popolazione normale ed un campione da essa estratto, che cosa si intende con i simboli:

$\sigma, \mu, s, \bar{x}$

4) Si immagini di aver spruzzato un insetticida su una popolazione di insetti, di aver estratto un campione di 20 individui superstiti e di averne registrato il sesso. Si immagini di aver riscontrato 5 maschi e 25 femmine superstiti**.** Ipotizzando che nella popolazione originaria (prima del trattamento) maschi e femmine fossero ugualmente rappresentati, stabilire se è corretto affermare che i maschi sono più sensibili delle femmine all'insetticida in studio.

5) Immaginate di sapere che un insetticida controlla il 60% di individui di *Lobesia botrana*. Organizzate allora un esperimento per vedere se lo stesso insetticida è più efficace quando utilizzato in miscela con un coadiuvante. Dall'esperimento ottenete il seguente risultato: 35 insetti morti su 40 trattati. Cosa concludete in relazione all'effetto del coadiuvante e perchè?

6) Qual è il valore critico della distribuzione di t (test bilaterale o a due code) per una probabilità del 5% e per un campione di numerosità pari a 21?

7) Si immagini di dover confrontare sette varietà di frumento con quattro ripetizioni e si immagini di impiegare lo strumento statistico dell' ANOVA. Impostare l'ipotesi nulla, scomporre i gradi di libertà ed impostare il test F.

8) Immaginiamo che il test di F dell'ANOVA di cui all'esercizio 10 abbia data un risultato pari a 7.5. Tenendo conto dei valori tabulati dell'F, cosa possiamo concludere sulla sette varietà di frumento?

9) Immaginiamo che la varietà IRNERIO abbia prodotto 5.2 t/ha, la varietà AURELIO 5.6 t/ha e la verietà GENIO 4.2 t/ha. La MDS calcolata per $p<0.05$ è risultata pari a 0.7 t/ha. Si calcoli se le tre varietà differiscono tra loro in modo significativo.

10) Da una popolazione di piante di mais è estratto un campione casuale di dieci individui, con le seguenti altezze:

155 - 159 - 160 - 167 - 168 - 169 - 172 - 172 - 178 - 179

Determinare:
1 - Media
2 - Varianza, deviazione standard
4 - Limiti di confidenza della media ($p<0,05$)

11) Un ricercatore ha rilevato la produzione di cinque parcelle di mais non diserbate (nelle quali cioè non sono stati fatti trattamenti per il controllo della flora infestante) ed ha ottenuto le seguenti produzioni:
50 – 55 – 48 – 45 – 39 kg/ha

Altre cinque parcelle sono state diserbate con l'erbicida A ed hanno mostrato le seguenti produzioni:

101 – 107 –109 –110 –99 kg/ha

Un successivo gruppo di cinque parcelle è stato invece diserbato utilizzando l'erbicida B, con le seguenti produzioni

120 – 119 – 115 –121 –108 kg/ha

Quale è stato il trattamento che ha determinato la maggior produzione di mais?

Esame del 14/01/2005

Considerate la seguente prova varietale, nella quale sono state misurate le produzioni e le altezze di 12 varietà di colza, secondo uno schema a randomizzazione completa con tre ripetizioni.

1 - Stabilire le medie di produzione e altezza per ogni varietà;
2 - eseguire l'ANOVA per la produzione;
3 - Calcolare la Minima Differenza Significativa (p=0.05)
4 - Verificare se la produzione è correlata con l'altezza (operando sulle medie di varietà).

Tesi	Varietà	Altezza	Produzione
1	PI469823	143	12.32
1	PI469823	143	10.88
1	PI469823	149	14.29
2	PI509072	131	8.01
2	PI509072	145	7.84
2	PI509072	138	10.43
3	PI469732	130	12.89
3	PI469732	135	9.92
3	PI469732	135	13.62
4	PI469741	132	12.77
4	PI469741	134	13.97
4	PI469741	132	12.77
5	PI469778	125	9.8
5	PI469778	122	9.8
5	PI469778	102	11.8
6	PI469778	140	13.02
6	PI469778	140	10.53
6	PI469778	156	11.15
7	GOLDEN	122	13.05
7	GOLDEN	138	12.08
7	GOLDEN	138	9.7
8	RESTON	160	13.53
8	RESTON	122	11.03
8	RESTON	155	11.02
9	BNI11	145	19.95
9	BNI11	145	18.53
9	BNI11	145	20.46
10	CAROLUS	144	20.97
10	CAROLUS	153	15.96
10	CAROLUS	140	16.91
11	EXTRA	153	16.84
11	EXTRA	160	14.92
11	EXTRA	165	18.89
12	PANTHER	128	12.68
12	PANTHER	150	16.24
12	PANTHER	148	12.03

Svolgimento

1 - Calcolo medie per varietà

```
> dati<-read.table("g:\prova.txt",header=TRUE)
> dati
> attach(dati)
> MeanH<-by(Altezza,Variet,mean)
> MeanYield<-by(Produzione,Variet,mean)
> MeanH
```

```
INDICES: BNI11
[1] 145
------------------------------------------------------------------------------
INDICES: CAROLUS
[1] 145.6667
------------------------------------------------------------------------------
INDICES: EXTRA
[1] 159.3333
------------------------------------------------------------------------------
INDICES: GOLDEN
[1] 132.6667
------------------------------------------------------------------------------
INDICES: PANTHER
[1] 142
------------------------------------------------------------------------------
INDICES: PI469732
[1] 133.3333
------------------------------------------------------------------------------
INDICES: PI469741
[1] 132.6667
------------------------------------------------------------------------------
INDICES: PI469778
[1] 116.3333
------------------------------------------------------------------------------
INDICES: PI469823
[1] 145
------------------------------------------------------------------------------
INDICES: PI509072
[1] 138
------------------------------------------------------------------------------
INDICES: RESTON
[1] 145.6667
------------------------------------------------------------------------------
INDICES: PI469779
[1] 145.3333

> MeanYield
INDICES: BNI11
[1] 19.64667
------------------------------------------------------------------------------
INDICES: CAROLUS
[1] 17.94667
------------------------------------------------------------------------------
INDICES: EXTRA
[1] 16.88333
------------------------------------------------------------------------------
INDICES: GOLDEN
[1] 11.61
------------------------------------------------------------------------------
INDICES: PANTHER
[1] 13.65
------------------------------------------------------------------------------
INDICES: PI469732
[1] 12.14333
------------------------------------------------------------------------------
INDICES: PI469741
[1] 13.17
------------------------------------------------------------------------------
INDICES: PI469778
[1] 10.46667
------------------------------------------------------------------------------
INDICES: PI469823
[1] 12.49667
```

```
------------------------------------------------------------------------------
INDICES: PI509072
[1] 8.76
------------------------------------------------------------------------------
INDICES: RESTON
[1] 11.86
------------------------------------------------------------------------------
INDICES: PI469779
[1] 11.56667
>
```

2 - ANOVA

```
> model<-lm(Produzione~Variet)
> anova(model)

Analysis of Variance Table

Response: Produzione
          Df Sum Sq Mean Sq F value    Pr(>F)
Variet    11 340.13   30.92  10.729 8.656e-07 ***
Residuals 24  69.17    2.88
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
>
```

3 - MDS

```
> sqrt(2*anova(model)$Mean[2]/(anova(model)$Df[2]))*qt(0.975,anova(model)$Df[2])
[1] 1.011482
```

4 - Correlazione

```
> MeanYield<-as.vector(MeanYield)
> MeanH<-as.vector(MeanH)
> cor(MeanYield,MeanH)
[1] 0.5442164
```

Esame del 09/02/2005

E' stata organizzata una prova sperimentale (con tre ripetizioni) per valutare la densità di semina ottimale per il colza. La tabella sottostante riporta i risultati produttivi della coltura. Calcolare:

1 - medie produttive per ciascuna densità di semina;
2 - ANOVA;
3 - Utilizzare la devianza del residuo per calcolare il coefficiente di variabilità della prova;
4 - Usando le medie, calcolare la relazione di regressione tra densità e produzione.

Dati sperimentali

Dens	Prod
20	2.33
40	2.50
60	2.56
80	2.60
100	2.73
20	2.35
40	2.49
60	2.58

80	2.62
100	2.75
20	2.39
40	2.52
60	2.54
80	2.59
100	2.77

Svolgimento

1 - Importazione dati in R

```
> dati<-read.table("h:\import.dat",header=TRUE)
> dati
   Dens Prod
1    20 2.33
2    40 2.50
3    60 2.56
4    80 2.60
5   100 2.73
6    20 2.35
7    40 2.49
8    60 2.58
9    80 2.62
10  100 2.75
11   20 2.39
12   40 2.52
13   60 2.54
14   80 2.59
15  100 2.77
> attach (dati)
```

2 - Calcolo medie per densità;

```
> by(Prod,Dens,mean)
INDICES: 20
[1] 2.356667
------------------------------------------------------------------------------
INDICES: 40
[1] 2.503333
------------------------------------------------------------------------------
INDICES: 60
[1] 2.56
------------------------------------------------------------------------------
INDICES: 80
[1] 2.603333
------------------------------------------------------------------------------
INDICES: 100
[1] 2.75
```

3 - ANOVA ad un livello

```
> densf<-factor(Dens)
> aov<-anova(lm(Prod~densf))
> aov
Analysis of Variance Table

Response: Prod
          Df   Sum Sq  Mean Sq F value    Pr(>F)
densf      4 0.247173 0.061793  140.44 9.676e-09 ***
Residuals 10 0.004400 0.000440
```

```
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

4 - Calcolo coefficiente di variabilità

```
> devstres<-sqrt(aov$Mean[2])
> devstres

0.02097618
> CV=devstres/mean(Prod)*100
> CV

0.8210925
```

4 - Regressione lineare

```
> medie<-by(Prod,Dens,mean)
> Y<-as.vector(medie)
> Y
[1] 2.356667 2.503333 2.560000 2.603333 2.750000
>
> X<-as.numeric(row.names(medie))
> X
[1]  20  40  60  80 100
> > Y<-as.vector(medie)
> Y
[1] 2.356667 2.503333 2.560000 2.603333 2.750000
> X<-as.numeric(row.names(medie))
> X
[1]  20  40  60  80 100
> modello<-lm(Y~X)
> summary(modello)

Call:
lm(formula = Y ~ X)

Residuals:
        1         2         3         4         5
-0.020667  0.037333  0.005333 -0.040000  0.018000

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.2886667  0.0371962  61.530 9.46e-06 ***
X           0.0044333  0.0005608   7.906  0.00422 **
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 0.03547 on 3 degrees of freedom
Multiple R-Squared: 0.9542,     Adjusted R-squared: 0.9389
F-statistic: 62.51 on 1 and 3 DF,  p-value: 0.004218

>
```