ELEMENTI DI STATISTICA DESCRITTIVA ED INFERENZIALE

Per gli studenti del 1° Anno della Facoltà di Agraria

APPUNTI DALLE LEZIONI (A.A. 2002/2003)

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

Pubblicato in proprio (2002)

ELEMENTI DI STATISTICA DESCRITTIVA

Corso di Matematica con principi di statistica ed informatica
1° Anno – 1° Semestre

Definizione

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...).

Statistica descrittiva e inferenziale

Quando si raccolgono informazioni in riferimento ad un certo fenomeno, ci si trova ad aver a che fare con una mole 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 studiare l'altezza media delle piante di mais di un determinato appezzamento, possiamo anche pensare di entrare nell'appezzamento in studio e misurare, una per una, le altezze di tutte le piante. Se invece vogliamo sapere l'altezza media delle piante di mais di una certa varietà, coltivata su tutto il comprensorio della Valle del Tevere, non saremo mai in grado di misurare le altezze di tutte le piante allevate in quel comprensorio, se non a costi troppo elevati. Pertanto effettueremo le nostre misure su un numero ridotto di piante, scelte a caso tra tutte quelle presenti nel comprensorio in studio.

Nella situazione anzidetta, chi effettua l'indagine non è interessato solo agli individui effettivamente misurati e quindi non può utilizzare (se non inizialmente) tecniche di statistica descrittiva. Infatti l'interesse è rivolto a tutti gli individui, compresi quelli che non sono stati direttamente misurati. In questo senso, le piante misurate costituiscono solo un campione di tutte quelle presenti nel comprensorio della Valle del Tevere. Il procedimento per cui dalle caratteristiche di un sottogruppo di individui, estratto a caso da un gruppo più grande, si cerca di risalire alle caratteristiche del gruppo più grande prende il nome di *inferenza statistica.* La disciplina relativa si chiama *statistica inferenziale.*

Il procedimento scientifico

Come già accennato, i campi di applicazione della statistica sono numerosi e spaziano dalla meteorologia alle scienze sociale, alle ricerche di marketing ecc.. Inoltre, la statistica trova applicazione in tutte le scienze sperimentali, come, tra le altre, le scienze agronomiche, le tecnologie alimentari e le discipline relative allo sviluppo rurale.

In tutti i casi, il ruolo della metodologia statistica é essenziale nell'applicazione del metodo scientifico, che è basato sulla formulazione di un ipotesi induttiva, che deve essere poi verificata deduttivamente mediante un esperimento appositamente pianificato. L'analisi dei dati ottenuti consente di provare l'ipotesi e formularne eventualmente una seconda.

La metodologia statistica consente di seguire questo cammino logico intervenendo in tutte le tappe: nella definizione del problema e nella formulazione di un'ipotesi precisa, nell'organizzazione dell'esperimento adatto a verificarla e nell'analisi dei dati ottenuti. Infatti il rilevamento deve essere fatto sulla base di criteri precisi, che consentano di ottenere informazioni pertinenti circa il problema in studio. Inoltre i dati grezzi non sono di solito suscettibili di un'interpretazione diretta, ma debbono essere ridotti e sintetizzati con metodiche di statistica descrittiva. Il procedimento inferenziale consente poi di prendere una decisione quanto più possibile obiettiva circa l'ipotesi formulata.

L'utilizzazione di un appropriato trattamento dei dati è inoltre particolarmente importante alfine di superare le principali difficoltà della sperimentazione biologica legate alla presenza di quello che viene definito *errore sperimentale,* cioè l'insieme delle variazioni indotte da fattori non controllati, i cui effetti si sovrappongono a quello del fattore in studio. Ad esempio, se siamo interessati a saggiare un nuovo fitofarmaco capace di eliminare gli insetti dannosi, dobbiamo organizzare un apposito esperimento, tenendo però presente che l'effetto insetticida non dipende solo dalle caratteristiche del fitofarmaco, ma anche, ad esempio, dalla suscettibilità dell'individuo trattato. Il problema consiste quindi nel valutare l'efficacia dell'insetticida, indipendentemente dalla suscettibilità dell'insetto trattato , il che può essere fatto adottando un adeguato disegno sperimentale. Analoga situazione può essere riscontrata nella chimica analitica: ad esempio quando misuriamo il contenuto di alcool nel vino, dobbiamo tenere presente che questo può essere influenzato dall'imprecisione dello strumento di misura, in modo che ogni analisi che facciamo può dare un risultato lievemente diverso dall'analisi precedente. E' chiaro quindi che un risultato assolutamente preciso potrebbe essere ottenuto solo con un numero di analisi infinito, il che non è tecnicamente fattibile. Allora procederemo eseguendo le analisi "in doppio" o in "triplo" ed adotteremo procedimento di inferenza statistica che ci consentano di risalire dai risultati delle due o tre analisi eseguite, ai risultati che si sarebbero ottenuti eseguendo un numero infinito di analisi.

Collettivo e unità sperimentale

In sostanza, in statistica si ha sempre a che fare con un *collettivo,* cioè con un insieme di individui (animali, piante, terreni, foglie ...) sui quali è stata studiata una certa caratteristica (peso, altezza, contenuto in fosforo, larghezza), in grado di assumere diversi valori e, pertanto, detta *variabile*. Il singolo individuo prende 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...).

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 conveniente 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 precedenti a 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/e classe/i precedente/i.

Rappresentazione grafica delle distribuzioni di frequenza

Oltre che in tabella, le frequenze possono essere anche riportate in grafico. Per variabili qualitative si usano in genere grafici ad istogramma o a torta, come quello in figura 1, relativo al collettivo di insetti prima indicato.

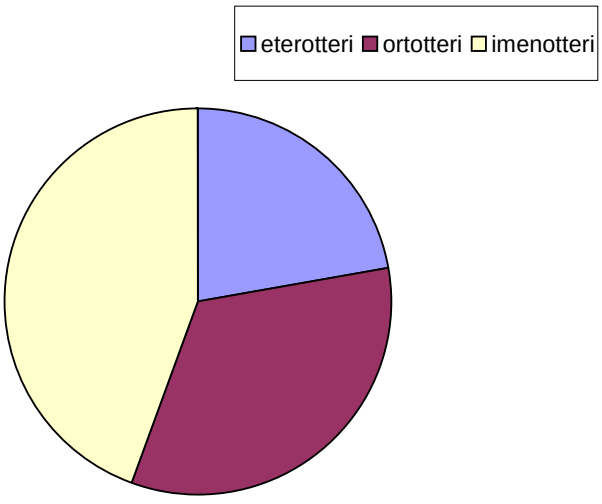


Figura 1. Esempio di un grafico a torta.

Nel caso di variabili quantitative, si usano spesso istogrammi di frequenza, dove la variabile rilevata si pone sull'asse delle ascisse, mentre la frequenza sull'asse delle ordinate, come nel caso dell'esercizio successivo.

Esercizio 1

Sono state rilevate le altezze di 3000 piante di mais. I dati sono i seguenti
145 – 147 –188 175 –176 (seguono altre 2992 misure).....182 – 147 148.
Valutare la distribuzione delle frequenze assolute, relative e cumulate.
Per una migliore comprensione dei dati questi vengono suddivisi in classi di frequenza, considerando intervalli di dieci centimetri.
La distribuzione delle frequenze assolute, relative e cumulate è quella riportata in tabella 1.

Tabella 1. Distribuzione delle frequenze assolute, relative e cumulate delle altezze di 3000 piante di mais

Classi	*Frequenze assolute*	*Frequenze relative*	*Frequenze cumulate*
150 – 155	115	0,038	115
155 – 160	224	0,075	339
160 – 165	399	0,133	738
165 – 170	547	0,182	1285
170 – 175	594	0,198	1879
175 – 180	494	0,165	2373
180 – 185	374	0,125	2747
185 – 190	176	0,059	2923
190 - 195	77	0,026	3000

Le frequenze assolute possono essere riportate in grafico come in Figura 2.

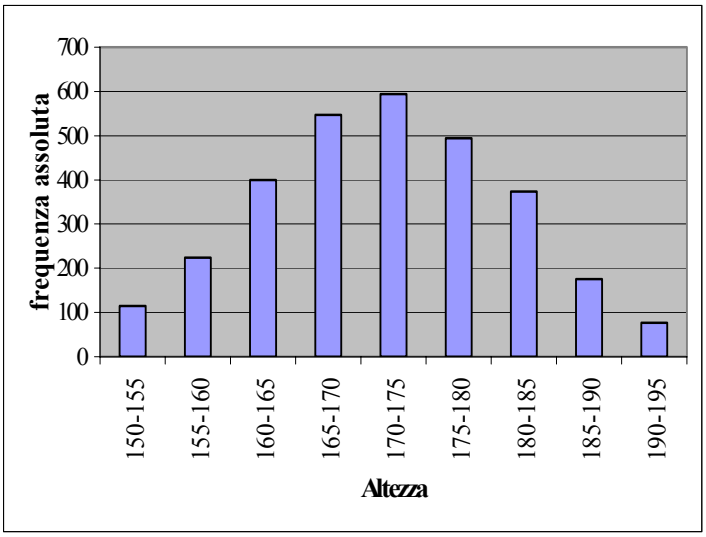


Figura 2. Rappresentazione grafiche delle frequenze assolute in tabella 1.

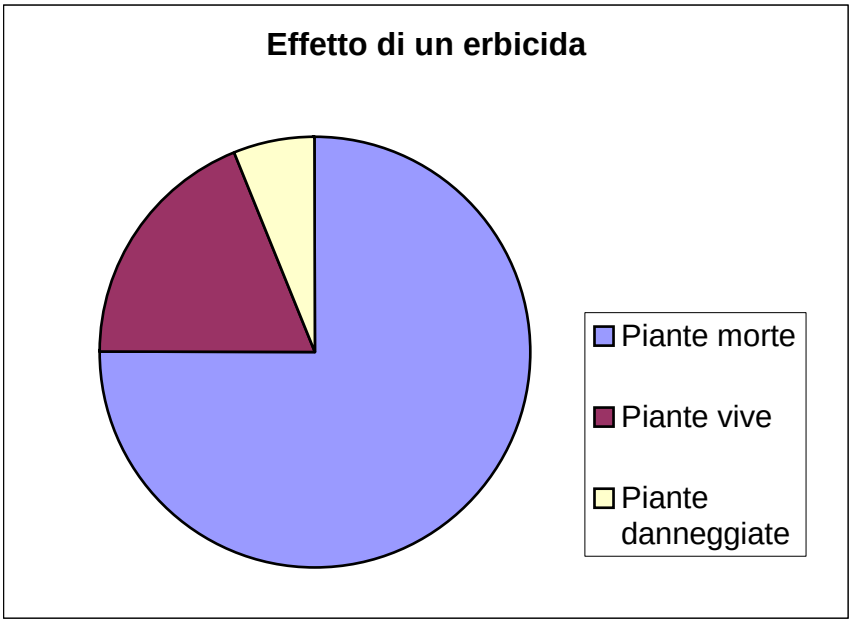
Esercizio 2

Dopo un trattamento con un erbicida, le 400 piante trattate possono essere classificate come segue: morte (300), vive (75) o danneggiate, ma non morte (25). Valutare la distribuzione delle frequenze assolute e relative.

La distribuzione di frequenze assolute e relativa è

Classi	*Frequenze assolute*	*Frequenze relative*
Morte	300	0,75
Vive	75	0,19
Danneggiate	25	0,06

In questo caso, siccome le classi di frequenza non possono essere logicamente ordinate, non ha senso calcolare le frequenze cumulate. I dati possono essere rappresentati in una torta, come segue.



Indici di tendenza centrale: media, moda e mediana

Nel caso di variabili statistiche qualitative, le informazioni relative alle frequenze nelle classi (modalità) costituiscono un'informazione sufficiente per un'analisi adeguata dei dati.

Nel caso di variabili quantitative, dato un insieme di dati o una distribuzione di frequenza, è possibile calcolare degli indici aggiuntivi, che rispecchino il più possibile le informazioni contenute nell'insieme dei dati.

Un'informazione fondamentale è quella relativa alla tendenza centrale della popolazione, espressa, tra gli altri, da tre indicatori, cioè la media, la moda e la mediana.

La *media aritmetica* 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.

La *moda* è invece la classe che presenta la maggior frequenza.

La *mediana* è data dal termine 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.

Percentili

I percentili costituiscono una famiglia di indicatori analoghi alla mediana. Hanno questo nome in quanto 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).

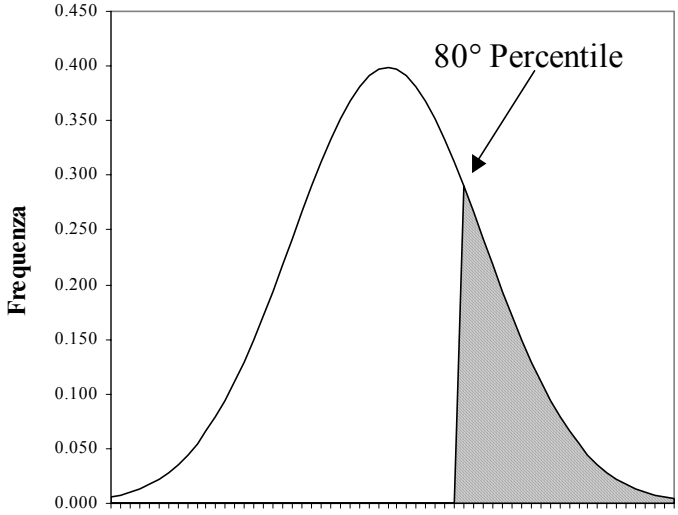


Figura 3. Esempio dell'80° percentile, in una distribuzione di frequenza per una variabile quantitativa su scala continua e con classi di frequenza di ampiezza estremamente ridotta.

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:

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

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

Dividendo la devianza per il numero dei gradi di libertà (numero degli individui del collettivo meno uno) si ottiene la *varianza* (generalmente indicata con σ^2):

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

La radice quadrata della varianza costituisce la *deviazione standard*, che si indica con σ.

Il coefficiente di variabilità è un indice percentuale, dato dal rapporto fra la deviazione standard e la media, moltiplicato per 100. E' interessante per confrontare tra di loro le variabilità di due o più collettivi e/o variabili.

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

Esercizio 3

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 media, devianza, varianza, deviazione standard e coefficiente di variabilità.

In questo caso si tratta dei dati relativi a sei singole unità sperimentali. I conti procedono come segue:

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

Non trattandosi di una distribuzione di frequenza la moda non è calcolabile, mentre la mediana è pari a:

$$Mediana = \frac{6.3 + 6.2}{2} = 6.25$$

$$SS = (6.5 - 6.15)^2 + (5.7 - 6.15)^2 + (6.4 - 6.15)^2 + (6.3 - 6.15)^2 +$$
$$+ (6.2 - 6.15)^2 + (5.8 - 6.15)^2 = 0.535$$

$$\sigma^2 = \frac{0.535}{5} = 0.107$$

$$\sigma = \sqrt{0.107} = 0.327$$

$$CV = \frac{0.327}{6.15} \times 100 = 5.32$$

Esercizio 4

Calcolare le statistiche descrittive per i dati relativi all'esercizio 1. In quale percentile si trovano due individui alti rispettivamente 160 e 190 cm?

Trattandosi di una distribuzione di frequenza, la media viene calcolata come segue:

Classi	*Valore centrale*	*Frequenze assolute*	*Somma della classe*
150 – 155	152.5	115	17537.50
155 – 160	157.5	224	35280.00
160 – 165	162.5	399	64837.50
165 – 170	167.5	547	91622.50
170 – 175	172.5	594	102465.00
175 – 180	177.5	494	87685.00
180 – 185	182.5	374	68255.00
185 – 190	187.5	176	33000.00
190 - 195	192.5	77	14822.50
		Somma =	*515'505.00*

La media è pari a 515'505 / 3000 = 171.84
La moda è pari a 172.5, così come la mediana.

La devianza è calcolata come segue:

Classi	*Valore centrale*	*Frequenze assolute*	*Scostamenti delle classi*	*Scostamenti al quadrato*	*Somma della classe*
150 – 155	152.5	115	-19.34	373.84	42991.86
155 – 160	157.5	224	-14.34	205.49	46030.26
160 – 165	162.5	399	-9.34	87.14	34769.75
165 – 170	167.5	547	-4.34	18.79	10279.35
170 – 175	172.5	594	0.66	0.44	262.6816
175 – 180	177.5	494	5.66	32.09	15853.56
180 – 185	182.5	374	10.67	113.74	42539.59
185 – 190	187.5	176	15.67	245.39	43189.03
190 - 195	192.5	77	20.67	427.04	32882.25
				Devianza =	268798.30

Gli altri indici di variabilità si calcolano analogamente all'esercizio precedente.
Un individuo alto 160 cm si lascia a sinistra 115 + 224 = 339 individui, cioè l'11% della popolazione. Si trova pertanto nel 12° percentile. Allo stesso modo, un individuo alto 180 cm si trova nel 80° percentile.

Distribuzioni bivariate

In alcuni casi in ciascuna unità sperimentale del collettivo vengono studiati due 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 esistenti tra le due variabili; 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. Principalmente, esistono due tipi di relazioni:

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).

Esercizio 5

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.

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):

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

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

La dimostrazione delle due formule non è richiesta. Si noterà comunque che, mentre la formula per il calcolo di b_0 è banale, la formula per il calcolo di b_1 porta al numeratore la codevianza di X e Y ed al denominatore la devianza di 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.

La figura sottostante mostra due esempi di regressione con diversi valori di R^2. E' comprensibile visivamente come la regressione in (A) si più attendibile di quella in (B)

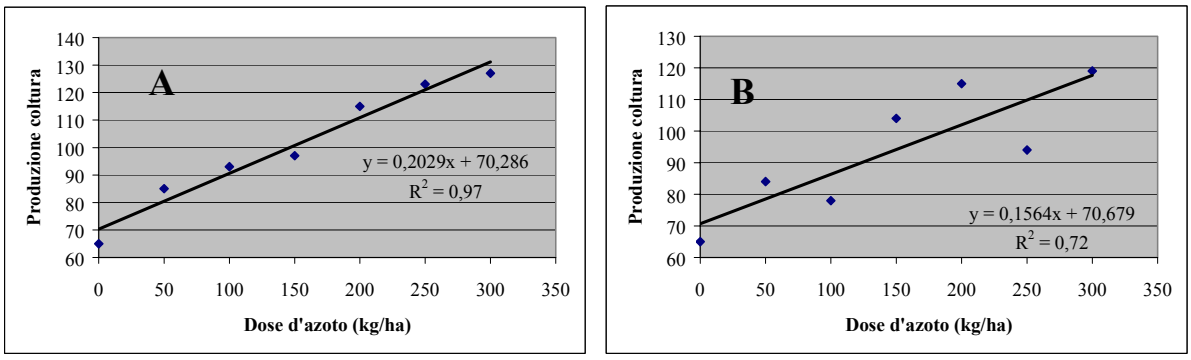


Figura 4. Due esempi di regressioni con diversi coefficienti di determinazione

Esercizio 6

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 la dose richiesta per inibire del 50% lo sviluppo della pianta trattata (ED50).

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:

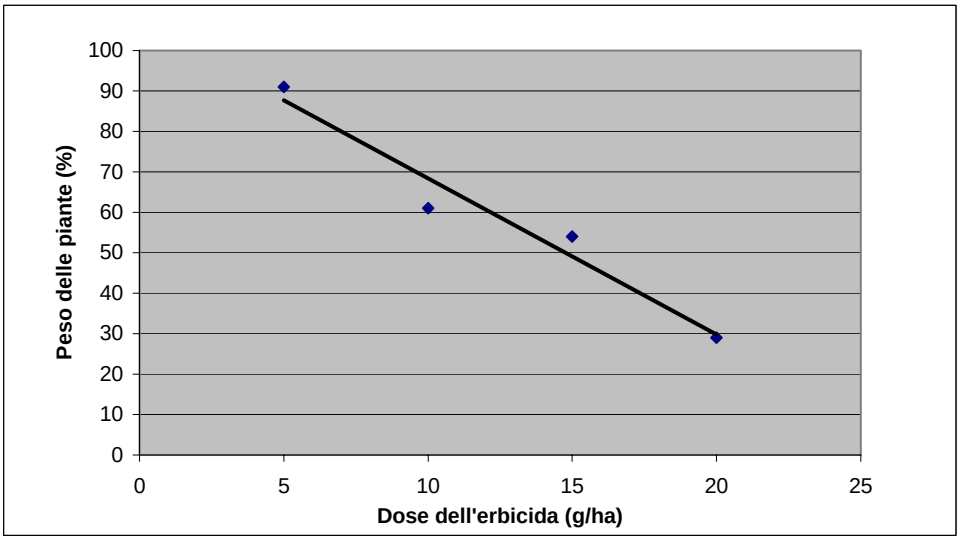


Figura 5. Analisi di regressione con i dati dell'esercizio 6

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.

Descrizione della distribuzione di frequenza di variabili quantitative: la 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 altezze 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*.

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 di un gruppo di *n* individui conosciamo (o riusciamo a stimare) la media e la deviazione standard, è come se conoscessimo ogni singolo individuo del gruppo: infatti con la media e la deviazione standard possiamo ricostruire l'intera distribuzione di frequenza dei dati.
2) la curva ha due asintoti e tende a 0 quando x tende a ± ∞. Questo ci dice che dato un certo fenomeno, tutte le misure sono possibili, ma 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 è pari al 99%.

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.

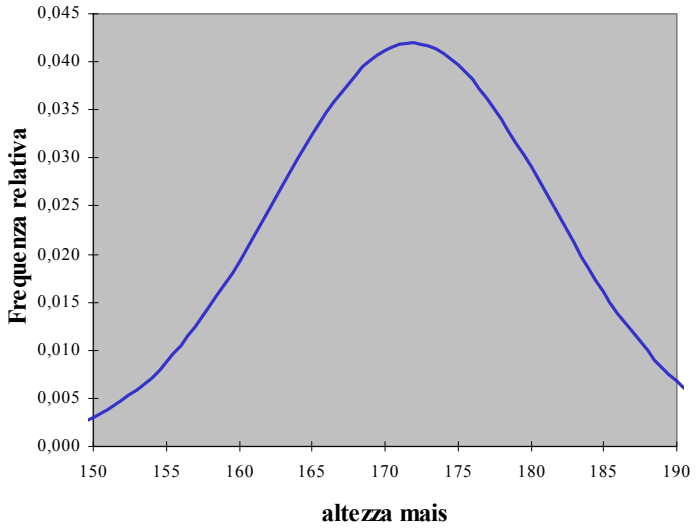


Figura 6. Distribuzione di frequenza Gaussiana per una popolazione normale con la stessa media e la stessa deviazione standard dei dati relativi all'esercizio 1.

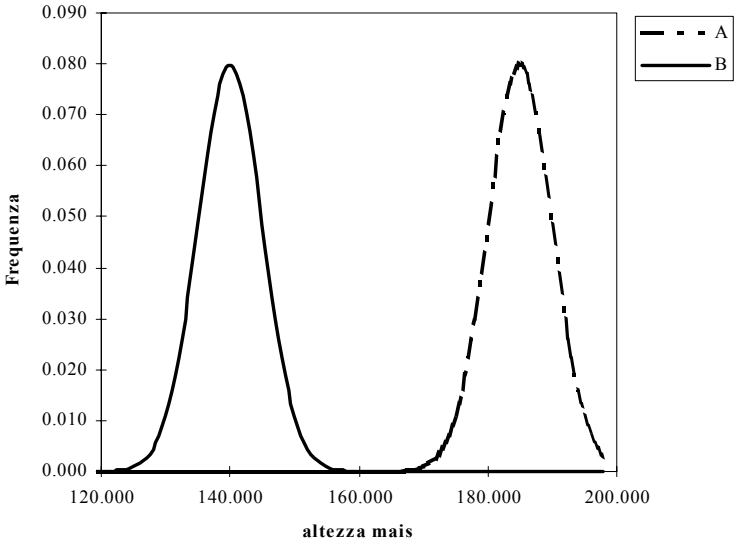


Figura 7. Esemplificazione grafica di due popolazioni normali A e B. A è

caratterizzata da $\mu = 185$ e $\sigma = 5$, mentre B è caratterizzata da $\mu = 140$ e $\sigma = 5$.

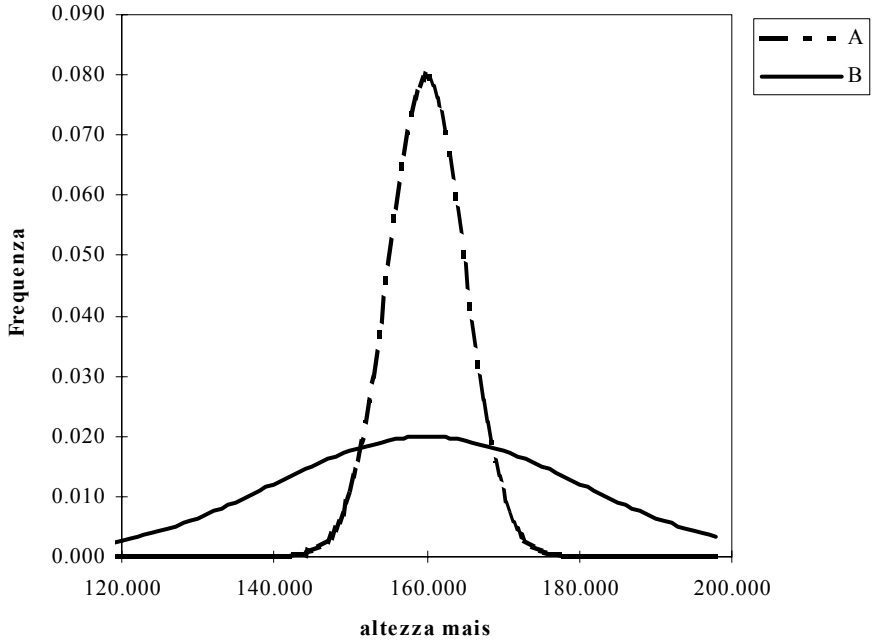


Figura 7. Esemplificazione grafica di due popolazioni normali A e B. A è caratterizzata da $\mu = 160$ e $\sigma = 5$, mentre B è caratterizzata da $\mu = 160$ e $\sigma = 20$. Notare che B è più bassa di A.

Trasformazione delle variabili

Per comodità, gli integrali della funzione di densità di frequenza, sono riportati in apposite tavole. Ovviamente le popolazioni normali sono infinite (perché infiniti sono i valori possibili per μ e σ). Siccome non è pensabili tabulare gli integrali della funzione di frequenza per ogni distribuzione normale, è stato tabulato solo l'integrale della funzione di distribuzione di una popolazione di riferimento con $\mu = 0$ e $\sigma = 1$ (dette tavole di z, si veda Tabella 2). Questa popolazione è detta *popolazione normale standardizzata* e qualunque altra popolazione, con opportuna trasformazione (detta standardizzazione) può essere ricondotta a questa.

Standardizzazione della variabili

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.

Questo modo di procedere è quello che viene comunemente adottato in statistica: si assume che la popolazione in studio si comporti secondo una distribuzione di riferimento (ed es. la normale di Gauss), si studia la distribuzione di riferimento e si estrapolano le conclusioni alla popolazione in studio.

Il procedimento è abbastanza complicato, tuttavia sarà sufficiente che lo studente

acquisisca almeno un esempio di questo modo di procedere, come di seguito indicato.

Tabella 2. Tavole di z: distribuzione normale standardizzata (integrale da z a + ∞).

Z	0	1	2	3	4	5	6	7	8	9
0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
3	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
3,5	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002
3,6	0,0002	0,0002	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
3,7	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
3,8	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001

(1) Si puntualizza che nella tabella z la prima colonna a sinistra indica l'unità e il primo decimale della cifra da cercare, mentre la prima riga indica il secondo decimale della cifra da cercare. In sostanza, se si dovesse cercare la probabilità di un valore pari a 3.27, la si dovrebbe cercare all'intersezione tra la riga che comincia con 3.2 e la colonna che comincia con 7 (0.0005 in tabella 2).

Esercizio 7

Abbiamo analizzato un campione di acqua proveniente da un pozzo con un contenuto medio di cloro pari a 1 meq l^{-1}; abbiamo eseguito l'analisi con uno strumento caratterizzato da un coefficiente di variabilità pari al 4%. La misura ottenuta per quel campione è di 1.1 meq l^{-1}. E' possibile che questa misura così alta 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 a 1.1 meq l^{-1}? La risposta può essere trovata ricorrendo alle tavole dell'integrale di probabilità della popolazione normale standardizzata, sapendo che 1.1 meq l^{-1} corrispondono a ad un valore standardizzato pari a 2.5. Infatti:

$$z = \frac{1.1 - 1}{0.04} = 2.5$$

La probabilità di ottenere questo valore o uno più alto di questo da una popolazione normale standardizzata è pari al 0.62% circa. Questo si può desumere dalla tabella 2, all'intersezione della riga che porta un valore iniziale pari a 2.5 e della colonna che porta un valore pari a 0 [(1)]

Si tratta quindi di un evento abbastanza raro. Siamo quindi autorizzati a sospettare l'inquinamento del pozzo o un errore di misura, dato che un' oscillazione casuale di tale entità è molto poco probabile.

STATISTICA DESCRITTIVA – 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.

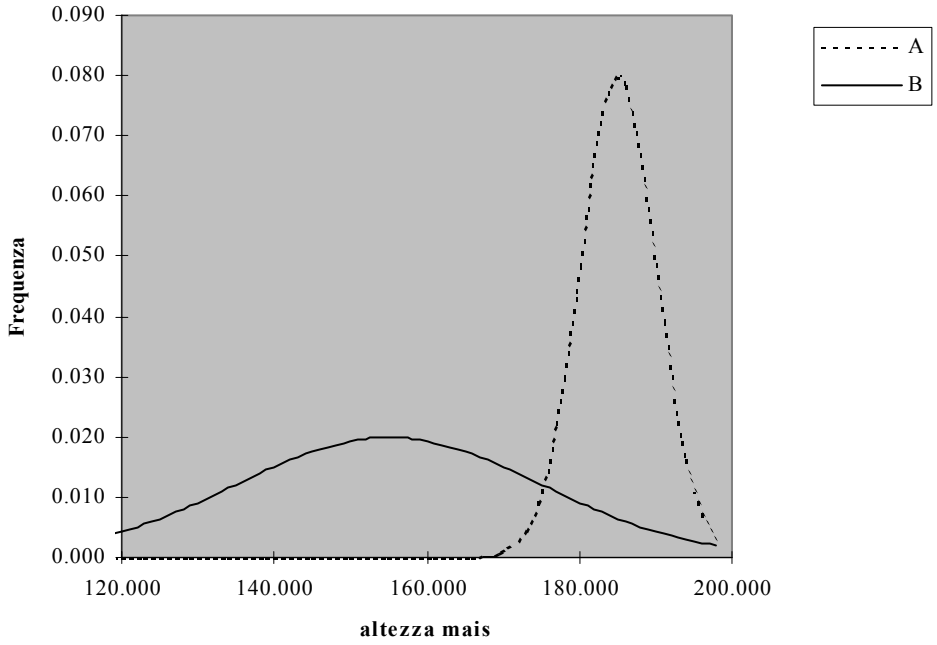
Le due popolazioni hanno

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) Su un campione di 500 olive si riscontra che 225 sono state attaccate da *Dacus oleae* (mosca dell'olivo). Stabilire:

a) frequenza assoluta delle piante attaccate_____

b) frequenza relativa delle piante attaccate _____

3) 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.

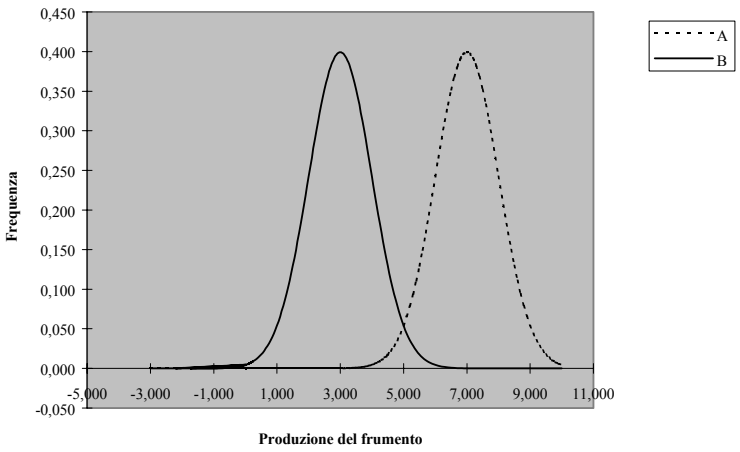
La concimazione ha incrementato:

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 ☐



4) Su un campione di 150 grappoli d'uva, si riscontra che 25 sono stati attaccati da *Botrytis cinerea* (muffa grigia). Stabilire:

a) frequenza assoluta dell'attacco _____

b) frequenza relativa dell'attacco_____

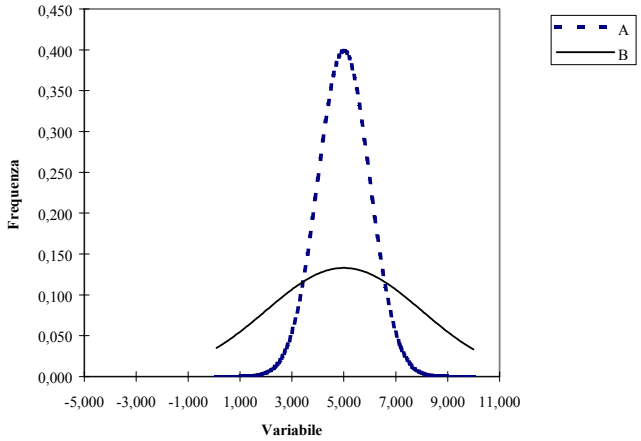
5) Cosa rappresenta il grafico sottostante?

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



6) Dalla popolazione delle altezze delle piante di mais alla fioritura si è estratto un campione di 10 individui. Le misure ottenute sono:

150, 160, 155, 154, 172, 137, 136, 148, 155, 157

Determinare:

a) media,

b) devianza, varianza e deviazione standard

7) L'analisi di correlazione si usa per:

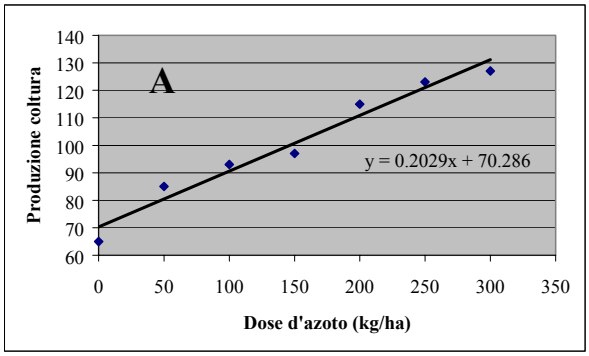
a) verificare se due medie sono significativamente diverse tra loro

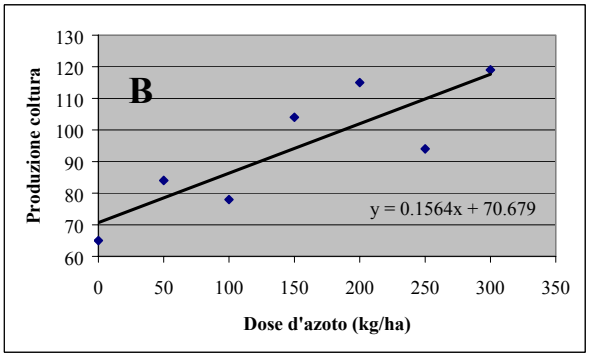
b) verificare se una variabile detta dipendente è funzione di un altra variabile detta indipendente

c) Verificare se due variabili variano in modo congiunto

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

9) 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.





10) La retta nel grafico A 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

11) La retta nel grafico B 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

12) Osservare la figura seguente. Le curve di regressione riportate hanno:

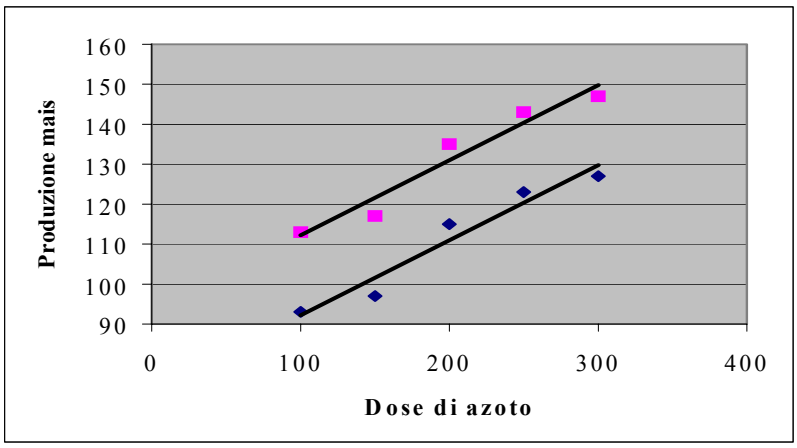
a) lo stesso coeff. di regressione, diversi B_0 e valori di R^2 simili.

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.



13) 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?

14) Un ricercatore ha eseguito sei repliche della stessa analisi chimica ed ha ottenuto i seguenti risultati:
101 – 126 – 97 – 117 – 121 – 94 nanogrammi per grammo
Calcolare media, devianza, varianza, deviazione standard e coefficiente di variabilità delle misure effettuate. Secondo voi, quanto è l'errore di misura dell'apparecchio utilizzato per l'analisi? Motivare la risposta.

15) 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.

ELEMENTI DI STATISTICA INFERENZIALE

Corso di Complementi di Matematica statistica ed informatica
1° Anno – 2° Semestre

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.

A questo proposito è necessario puntualizzare che l'inferenza 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.*

Il problema delle misure

Il problema dell'inferenza statistica è molto comune nell'attività di chiunque si occupi della misura di fenomeni biologici. 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 q uale 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.

Le unità sperimentali

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**. Bisogna 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.

Esercizio 8

Da un terreno agrario è stato estratto casualmente un campione di 4 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 l'errore che abbiamo commesso nel rilievo (errore legato ad un contenuto nel terreno non uniforme, alla tecnica di raccolta e di misura)?
Questo problema può essere risolto pensando che i risultati delle infinite analisi che potrebbero essere eseguite su un terreno agrario dovrebbero distribuirsi normalmente, con una media pari al contenuto medio di fosforo del terreno e una deviazione standard proporzionale all'errore che commettiamo nel prelievo.
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 valore più probabile del contenuto medio di fosforo nel terreno è pari a 12.4 ppm, mentre il valore più probabile per l'errore di determinazione è del 20.6%
Si capisce inoltre come 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.

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

Abbiamo visto in precedenza che facendo ricorso ad un numero elevato di ripetizioni otteniamo delle stime più affidabili. A questo punto dobbiamo fare una distinzione tra variabilità della misura ed errore di stima. La variabilità della misura tiene conto di tutte le possibili fonti di variabilità che alterano il valore della misura e quindi trasformano la misura stessa in una distribuzione (spesso normale) di frequenza delle misure possibili. Ci si richiama quindi al concetto di popolazione normale: la misura più probabile (cioè quella con la frequenza più alta se facessimo un infinito numero di determinazioni) corrisponde con la media delle misure effettuate, mentre la deviazione standard è una misura della variabilità stessa.

Se avessimo fatto un infinito numero di misure avremmo ottenuto una stima perfetta della misura effettuata e della sua variabilità (che è ineliminabile). Quindi, in sostanza, la stima può essere perfetta anche se la misura è viziata da un errore (per esempio perché l'apparecchio non è perfettamente funzionante). Di conseguenza, il concetto di variabilità della misura è diverso dal concetto di precisione della stima.

Per quanto detto sopra, la precisione della stima dipende sia dalla variabilità della misura,

sia dal numero di repliche che effettuiamo.

Possiamo a questo punto definire un indice che misura la precisione della stima che è detto **errore standard** ed è definito:

$$errore\ s\tan dard = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

dove σ è la deviazione standard e n è il numero di repliche. Dall'espressione soprascritta possiamo dedurre che l'errore standard aumenta all'aumentare della deviazione standard e diminuisce all'aumentare del numero delle ripetizioni, annullandosi quando questo tende ad infinito.

Pertanto, l'errore standard può essere assunto come un errore di stima associato alla determinazione della media.

Esercizio 9
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:

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

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 $\overline{X}$ 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:

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

Intervalli di confidenza di una media

L'errore standard può essere utilizzato per la costruzione degli intervalli di confidenza della media, con la formula seguente:

$$\mu = \overline{X} \pm t_{0,05;n-1} \times s_{\overline{X}}$$

La formula è piuttosto chiara: si tratta di ipotizzare che la media incognita dell'intero collettivo di dati (μ) è pari alla media stimata ($\overline{X}$) a cui viene aggiunta e tolta una quantità proporzionale all'errore standard. La costante di proporzionalità è data dal valore $t_{\alpha;n-1}$ che può essere desunto dalla tabella sottostante, per il livello di probabilità α prescelto e per il numero di gradi di libertà relativi (numerosità del campione meno una unità).

Tabella 3. Valori critici della distribuzione di t.

Gradi di libertà	Probabilità di errore (α)				Gradi di libertà	Probabilità di errore (α)			
	0.1	**0.05**	**0.025**	**0.01**		**0.1**	**0.05**	**0.025**	**0.01**
1	6.314	12.706	25.452	63.656	**31**	1.696	2.040	2.356	2.744
2	2.920	4.303	6.205	9.925	**32**	1.694	2.037	2.352	2.738
3	2.353	3.182	4.177	5.841	**33**	1.692	2.035	2.348	2.733
4	2.132	2.776	3.495	4.604	**34**	1.691	2.032	2.345	2.728
5	2.015	2.571	3.163	4.032	**35**	1.690	2.030	2.342	2.724
6	1.943	2.447	2.969	3.707	**36**	1.688	2.028	2.339	2.719
7	1.895	2.365	2.841	3.499	**37**	1.687	2.026	2.336	2.715
8	1.860	2.306	2.752	3.355	**38**	1.686	2.024	2.334	2.712
9	1.833	2.262	2.685	3.250	**39**	1.685	2.023	2.331	2.708
10	1.812	2.228	2.634	3.169	**40**	1.684	2.021	2.329	2.704
11	1.796	2.201	2.593	3.106	**41**	1.683	2.020	2.327	2.701
12	1.782	2.179	2.560	3.055	**42**	1.682	2.018	2.325	2.698
13	1.771	2.160	2.533	3.012	**43**	1.681	2.017	2.323	2.695
14	1.761	2.145	2.510	2.977	**44**	1.680	2.015	2.321	2.692
15	1.753	2.131	2.490	2.947	**45**	1.679	2.014	2.319	2.690
16	1.746	2.120	2.473	2.921	**46**	1.679	2.013	2.317	2.687
17	1.740	2.110	2.458	2.898	**47**	1.678	2.012	2.315	2.685
18	1.734	2.101	2.445	2.878	**48**	1.677	2.011	2.314	2.682
19	1.729	2.093	2.433	2.861	**49**	1.677	2.010	2.312	2.680
20	1.725	2.086	2.423	2.845	**50**	1.676	2.009	2.311	2.678
21	1.721	2.080	2.414	2.831	**55**	1.673	2.004	2.304	2.668
22	1.717	2.074	2.405	2.819	**60**	1.671	2.000	2.299	2.660
23	1.714	2.069	2.398	2.807	**65**	1.669	1.997	2.295	2.654
24	1.711	2.064	2.391	2.797	**70**	1.667	1.994	2.291	2.648
25	1.708	2.060	2.385	2.787	**75**	1.665	1.992	2.287	2.643
26	1.706	2.056	2.379	2.779	**80**	1.664	1.990	2.284	2.639
27	1.703	2.052	2.373	2.771	**85**	1.663	1.988	2.282	2.635
28	1.701	2.048	2.368	2.763	**90**	1.662	1.987	2.280	2.632
29	1.699	2.045	2.364	2.756	**95**	1.661	1.985	2.277	2.629
30	1.697	2.042	2.360	2.750	**100**	1.660	1.984	2.276	2.626

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.

Esercizio 10
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:

$$\mu = 4.1 \pm 2.262 \times 0.167 = 4.1 \pm 0.378$$

Questo ci permette di affermare che la produzione media per pianta del vigneto (quella vera, che rimane ignota) è compresa tra 3.722 e 4.478. 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% (α = 0.01), ottenendo:

$$\mu = 4.1 \pm 3.250 \times 0.167 = 4.1 \pm 0.543$$

In questo caso possiamo affermare che la produzione media per pianta del vigneto è compresa tra 3.557 e 4.643, con una probabilità d'errore dell'1%. Come si vede, per diminuire la probabilità d'errore abbiamo dovuto allargare l'intervallo di confidenza.

Esercizio 11
Immaginiamo una popolazione di dati composta da 84 individui (una popolazione piccola, ma motivata da esigenze di brevità!). Questi 84 individui sono in realtà 84 appezzamenti di mais; in ognuno di questi 84 appezzamenti sono state rilevate due variabili: (1) la presenza di piante infestanti (in % di ricoprimento) (2) la produzione della coltura (in t /ha).
E' quindi evidente che si tratta di una popolazione bivariata, i cui dati sono riportati in Tabella 4; inoltre è chiaro che tra le due variabili esiste una relazione di dipendenza, nel senso che la produzione del mais (variabile dipendente) dipende direttamente dalla presenza di piante infestanti (variabile indipendente).
Applicando a questa popolazione i nostri indici descrittivi potremmo concludere quanto segue:

RICOPRIMENTO FLORA INFESTANTE
La media è 38.07, mentre la deviazione standard è 41.53

PRODUZIONE MAIS
La media è 11.85, mentre la deviazione standard è 1.085

REGRESSIONE LINEARE
La relazione di regressione tra le due variabili è

$$Y = 12.682 - 0.0218\,X$$

La produzione decresce al crescere del ricoprimento delle piante infestanti, secondo la funzione sopra indicata.
Ora, immaginiamo che la popolazione di dati appena descritta sia in

realtà assolutamente ignota e che per qualche motivo sia necessario compiere lo studio anzidetto. Immaginiamo di non avere il tempo o le possibilità di studiare tutti gli 84 individui, ma di potere effettuare lo studio solo su quattro di essi.
Immaginiamo quindi di utilizzare un algoritmo di estrazione casuale per scegliere gli individui da studiare ed immaginiamo che questo algoritmo ci abbia indicato gli individui numero 17, 33, 35 e 71.
Ovviamente noi non abbiamo nessun interesse specifico nei confronti dei quattro individui campionati, ma abbiamo interesse a stimare le caratteristiche dell'intera popolazione.
Per gli individui considerati possiamo determinare che il ricoprimento medio di piante infestanti è pari a 25.80 con deviazione standard pari a 27.33; inoltre, la produzione media di mais è pari a 11.69 con deviazione standard pari a 0.98.
Si può osservare che in realtà queste stime puntuali, pur essendo abbastanza vicine ai valori della popolazione intera, tuttavia non sono esattamente uguali.

Tabella 4 . Esempio di una popolazione bivariata di dati, relativi ad un esperimento agronomico.

Num. dato	Ricoprimento piante infestanti (X)	Produzione mais (Y)	Num. dato	Ricoprimento piante infestanti (X)	Produzione mais (Y)	Num. dato	Ricoprimento piante infestanti (X)	Produzione mais (Y)
1	0.00	12.80	29	10.00	12.75	57	46.25	12.11
2	0.10	12.59	30	10.20	12.48	58	48.80	13.74
3	0.13	12.75	31	11.35	12.13	59	52.60	11.54
4	0.15	12.94	32	11.51	11.09	60	52.60	12.28
5	0.20	12.57	33	12.41	11.18	61	53.78	11.75
6	0.30	12.75	34	17.60	11.19	62	55.20	12.01
7	0.30	12.39	35	20.05	12.25	63	56.25	10.80
8	0.40	12.95	36	21.45	12.37	64	57.60	12.56
9	1.28	13.01	37	22.50	12.10	65	57.75	11.65
10	1.30	13.02	38	22.50	12.09	66	60.00	11.05
11	2.06	12.25	39	22.60	11.49	67	60.10	12.15
12	2.65	12.38	40	22.60	12.96	68	62.60	11.54
13	2.85	12.64	41	25.15	12.17	69	62.69	11.01
14	3.93	12.65	42	25.25	12.78	70	65.11	10.40
15	5.00	12.54	43	25.35	11.74	71	65.75	10.59
16	5.00	12.53	44	27.80	11.98	72	67.33	11.23
17	5.00	12.75	45	28.80	12.33	73	76.35	11.06
18	5.10	12.69	46	28.85	12.39	74	80.20	10.54
19	5.19	12.58	47	30.15	12.00	75	82.60	11.70
20	5.20	12.58	48	31.91	12.23	76	85.00	9.94
21	5.20	12.98	49	33.75	10.16	77	97.70	9.96
22	5.40	12.99	50	35.45	10.15	78	100.15	11.49
23	5.50	12.45	51	37.50	12.11	79	110.00	10.57
24	5.71	12.12	52	40.10	11.99	80	125.00	9.88
25	7.55	12.54	53	40.10	12.89	81	145.10	10.12
26	7.65	12.58	54	40.34	11.89	82	170.00	8.54
27	7.65	13.01	55	42.60	11.20	83	172.70	8.06
28	10.00	12.49	56	45.00	12.01	84	185.10	8.80

Se però calcoliamo gli intervalli di confidenza ($p = 0.05$) delle stime otteniamo che la produzione media di mais dell'intera popolazione è pari a 11.69 ± 1.57, mentre il livello medio di infestazione è pari a 25.80 ± 43.49. In entrambi i casi le nostre stime includono la vera media (in realtà ignota) della popolazione da cui si è estratto un campione, anche se la stima che abbiamo, soprattutto nel caso del

livello di infestazione è molto imprecisa (intervallo di confidenza molto ampio). Ciò è ovviamente dovuto all'elevata variabilità della popolazione di origine e al basso numero di individui nel campione, che hanno fatto innalzare notevolmente il valore dell'errore standard.

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 parte dal calcolo della deviazione standard del residuo, che si esegue come di seguito illustrato.

Esercizio 11 (segue)
Consideriamo i quattro individui campionati dall'intera popolazione:

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 le metodiche note. Arriviamo alla seguente conclusione:
$b_0 = 12.411$
$b_1 = - 0.028$
Possiamo notare che si tratta di valori vicini a quelli ottenuti lavorando sull'intera popolazione, ma non uguali.
Dovendo calcolare l'errore di stima della regressione possiamo notare che, se la equazione di regressione appena calcolata ($Y = 12.411 – 0.028 X$) fosse vera, allora ai valori di X dei quattro individui campionati dovrebbero corrispondere i valori attesi riportati in tabella.

Num. dato	Ricoprimento piante infestanti (X)	Produzione mais (osservata)	Produzione mais (attesa)	(Osservato – Atteso)^2
17	5.00	12.75	Y = 12.411 – 0.028 x 5 = 12.271	0.229
33	12.41	11.18	Y = 12.411 – 0.028 x 12.41 = 12.065	0.784
35	20.05	12.25	Y = 12.411 – 0.028 x 20.05 = 11.853	0.158
71	65.75	10.59	Y = 12.411 – 0.028 x 65.75 = 10.59	0.000

La somma dei quadrati è pari a 1.17, che corrisponde alla cosiddetta devianza residua (residua, perché non spiegata dalla regressione). Con i procedimenti usuali, possiamo dire che la varianza residua è pari alla devianza diviso il numero di gradi di libertà, che nel caso della regressione è pari ad n-2. La deviazione standard residua è pari alla radice quadrata della varianza residua. Nel caso specifico la deviazione standard residua è pari a 0.7650.

Dopo aver calcolato la deviazione standard del residuo, dobbiamo calcolare l'errore standard della pendenza e dell'intercetta. La prima quantità si calcola:

$$s_{b1} = \frac{s_{residuo}}{\sqrt{SQ_X}}$$

mentre l'errore standard dell'intercetta si calcola:

$$s_{b0} = s_{residuo}\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{SQ_X}}$$

Esercizio 11 (segue)
Nel caso specifico dell'esercizio 11, l'errore standard della pendenza è pari a:

$$s_{b1} = \frac{0.7650}{\sqrt{2240.93}} = \frac{0.7650}{47.33} = 0.016$$

mentre l'errore standard dell'intercetta è pari a:

$$s_{b0} = 0.7650 \times \sqrt{\frac{1}{4} + \frac{(25.80)^2}{2240.93}} = 0.5658$$

Ora che sono noti gli errori standard di ciascun parametro, gli intervalli di confidenza si calcolano nel modo già spiegato per la media, considerando che i gradi di libertà di una regressione sono sempre (n-2).

Esercizio 11 (segue)
Per quanto riguarda l'intercetta, i limiti di confidenza per una probabilità d'errore del 5% sono pari a:
$\beta_1 = b_1 \pm t_{0.05,\,2} \cdot s_{b1} = -0.028 \pm 4.3027 \text{ x } 0.016 = -0.028 \pm 0.0688$

Per quanto riguarda l'intercetta, i limiti di confidenza per una

probabilità d'errore del 5% sono pari a:
$\beta_0 = b_0 \pm t_{0.05,\,2} \cdot s_{b0} = 12.411 \pm 4.3027 \text{ x } 0.5658 = 12.411 \pm 2.4345$

Ancora una volta possiamo notare che i parametri della popolazione di 84 individui sono effettivamente contenuti all'interno degli intervalli di confidenza.

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

Confronto tra due medie: il test t di Student

Fino ad ora ci siamo occupati di una sola popolazione e di un eventuale campione rappresentativo di essa. Nella sperimentazione agraria, tuttavia, 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.

Esercizio 12
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 superiore al limite di confidenza inferiore 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:

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

Dove $\overline{s}^2$ 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 il test di t assume un valore tanto più alto, quanto più è significativa la differenza tra le due medie. O meglio: tanto più è elevato il valore del test di t, quanto più bassa è la probabilità di sbagliare affermando che la differenza tra le due medie è significativa.

La probabilità d'errore che riteniamo accettabile va fissata all'inizio dell'esperimento ed è in genere pari al 5% o all' 1% ($p = 0.05$ o $p = 0.01$).

Ora, fissata la probabilità d'errore che riteniamo accettabile, per rispondere alla domanda fatta all'inizio dobbiamo sapere qual è il valore minimo del test che ci consente di concludere che la differenza tra le medie è significativa. Questo valore può essere desunto dalle tavole di t riportate in precedenza, considerando un numero di gradi di libertà pari ad $[(n_1-1)+(n_2-2)]$, con n_1 ed n_2 pari alle numerosità dei due campioni a confronto.

Esercizio 13
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 sono riportati di seguito. Si può notare come il test di t porta ad un valore di t calcolato pari a 2.944, che è più alto del valore di t tabulare per 6 gradi di libertà (2.447). 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 inferiore al 5%, cioè al livello di protezione prefissato.

	T	**NT**
	85	101
	78	95
	91	89
	81	94
Media	**83.75**	**94.75**

Dev.	**94.75**	**72.75**
n	**4**	**4**
Media generale		89.25
Varianza media		27.92
err.stand. diff.		3.73
Test t		2.944

Confronto tra due frequenze: il test di χ^2

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 frequenza degli individui che presentano una certa modalità. Dovendo confrontare tra loro una serie di frequenze, non possiamo utilizzare il test di t, ma dobbiamo ricorrere ad un altro test, detto di χ^2.

La forma del test è la seguente:

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

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

Come nel caso del test di t, il valore ottenuto per il χ^2 viene confrontato con i valori tabulati, per l'opportuno livello di probabilità e per un numero di gradi di libertà pari al numero dei confronti indipendenti (Tabella 5). Se il valore calcolato è superiore a quello tabulato significa che l'ipotesi nulla non può essere accettata.

Anche in questo caso un esempio chiarirà l'uso del test.

Esercizio 14
Un erbicida viene comunemente utilizzato in quanto, salvo oscillazioni di efficacia casuali, è in grado di controllare l'80 % delle piante infestanti presenti in un campo di mais (questo valore è noto e non è soggetto ad errore). Lo stesso erbicida viene utilizzato in miscela con un coadiuvante, che dovrebbe migliorarne l'efficacia erbicida. In questo caso, su 50 piante osservate ne sono morte 48. Si tratta di un'oscillazione di efficacia casuale, o invece la presenza del coadiuvante ha migliorato l'azione dell'erbicida?
E' evidente che la variabile rilevata è di tipo qualitativo (pianta viva o morta) e che le unità sperimentali sono costituite dalle singole piante infestanti. In questo caso l'ipotesi nulla è che il coadiuvante non ha influenzato l'efficacia dell'erbicida. Se questo è vero, ci si aspetta che dopo il trattamento effettuato con l'erbicida in miscela con il

coadiuvante, su 50 piante osservate, se ne dovrebbero trovare 40 morte (80%) che è appunto la frequenza attesa (f_a). In realtà se ne sono contate 48 morte (frequenza osservata). Possiamo impostare il test di χ^2, come segue:

$$\chi^2 = \frac{(48-40)^2}{40} = 1.6$$

Il risultato ottenuto viene confrontato con il valore riportato nella Tabella 5, che è pari a 3.841. Siccome il valore calcolato è più basso di quello tabulare, dobbiamo concludere che l'ipotesi nulla può essere accettata: non abbiamo elementi per ritenere che il coadiuvante abbia incrementato l'efficacia dell'erbicida.

Esercizio 15

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.

Tabella 5. Valori critici per la distribuzione di χ^2.

GL	Probabilità (α)				GL	Probabilità (α)			
	0.1	0.05	0.025	0.01		0.1	0.05	0.025	0.01
1	2.706	3.841	5.024	6.635	**31**	41.422	44.985	48.232	52.191
2	4.605	5.991	7.378	9.210	**32**	42.585	46.194	49.480	53.486
3	6.251	7.815	9.348	11.345	**33**	43.745	47.400	50.725	54.775
4	7.779	9.488	11.143	13.277	**34**	44.903	48.602	51.966	56.061
5	9.236	11.070	12.832	15.086	**35**	46.059	49.802	53.203	57.342
6	10.645	12.592	14.449	16.812	**36**	47.212	50.998	54.437	58.619
7	12.017	14.067	16.013	18.475	**37**	48.363	52.192	55.668	59.893
8	13.362	15.507	17.535	20.090	**38**	49.513	53.384	56.895	61.162
9	14.684	16.919	19.023	21.666	**39**	50.660	54.572	58.120	62.428
10	15.987	18.307	20.483	23.209	**40**	51.805	55.758	59.342	63.691
11	17.275	19.675	21.920	24.725	**41**	52.949	56.942	60.561	64.950
12	18.549	21.026	23.337	26.217	**42**	54.090	58.124	61.777	66.206
13	19.812	22.362	24.736	27.688	**43**	55.230	59.304	62.990	67.459
14	21.064	23.685	26.119	29.141	**44**	56.369	60.481	64.201	68.710
15	22.307	24.996	27.488	30.578	**45**	57.505	61.656	65.410	69.957
16	23.542	26.296	28.845	32.000	**46**	58.641	62.830	66.616	71.201
17	24.769	27.587	30.191	33.409	**47**	59.774	64.001	67.821	72.443
18	25.989	28.869	31.526	34.805	**48**	60.907	65.171	69.023	73.683
19	27.204	30.144	32.852	36.191	**49**	62.038	66.339	70.222	74.919
20	28.412	31.410	34.170	37.566	**50**	63.167	67.505	71.420	76.154
21	29.615	32.671	35.479	38.932	**55**	68.796	73.311	77.380	82.292
22	30.813	33.924	36.781	40.289	**60**	74.397	79.082	83.298	88.379
23	32.007	35.172	38.076	41.638	**65**	79.973	84.821	89.177	94.422
24	33.196	36.415	39.364	42.980	**70**	85.527	90.531	95.023	100.425
25	34.382	37.652	40.646	44.314	**75**	91.061	96.217	100.839	106.393
26	35.563	38.885	41.923	45.642	**80**	96.578	101.879	106.629	112.329
27	36.741	40.113	43.195	46.963	**85**	102.079	107.522	112.393	118.236
28	37.916	41.337	44.461	48.278	**90**	107.565	113.145	118.136	124.116
29	39.087	42.557	45.722	49.588	**95**	113.038	118.752	123.858	129.973
30	40.256	43.773	46.979	50.892	**100**	118.498	124.342	129.561	135.807

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, che in genere richiede l'uso del computer per poter essere applicato a problemi appena più complessi di questo.

Esercizio 16
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**

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. Quindi salvo oscillazioni casuali, se facciamo il rapporto tra la varianza dei trattamenti e quella dell'errore (test F), questo rapporto dovrebbe oscillare intorno ad 1. In realtà, esiste un valore limite per questo rapporto riportato nella tabella seguente. Per l'uso della tabella 7, si deve considerare che essa è a doppia entrata e porta sulle colonne i gradi di libertà dei trattamenti e sulle righe i gradi di libertà dell'errore. Possiamo osservare che il valore limite tabulare è pari a 4.26 (per 2 e 9 gradi di libertà). Il valore di F da noi calcolato è invece pari a 57.52, molto superiore a quello tabulato. Pertanto possiamo rifiutare l'ipotesi nulla, affermando 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.

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 (nel nostro caso è pari a 4).*

Nel nostro caso quindi:

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

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 C 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 = 9,25 che è superiore alla MDS). Allo stesso modo B è stato superiore ad A (A – B = 26.75 che è superiore alla MDS).

Possiamo quindi concludere che tutti e tre i trattamenti si sono comportati diversamente.

Figura 7. Valori critici della distribuzione di F ($\alpha = 0.05$). Sulle colonne i gradi di libertà dei trattamenti, sulle righe i gradi di libertà dell'errore.

G.L.	1	2	3	4	5	6	7	8	9	10	11	12	15	20	25	30	35	40	50	100
1	161.4	199.5	215.7	224.6	230	234	237	239	241	242	243	244	246	248	249	250	250	251	252	253
2	18.51	19.00	19.16	19.25	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5	19.5
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.70	8.66	8.63	8.62	8.60	8.59	8.58	8.55
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.86	5.80	5.77	5.75	5.73	5.72	5.70	5.66
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.62	4.56	4.52	4.50	4.48	4.46	4.44	4.41
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.94	3.87	3.83	3.81	3.79	3.77	3.75	3.71
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.51	3.44	3.40	3.38	3.36	3.34	3.32	3.27
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.22	3.15	3.11	3.08	3.06	3.04	3.02	2.97
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.01	2.94	2.89	2.86	2.84	2.83	2.80	2.76
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.85	2.77	2.73	2.70	2.68	2.66	2.64	2.59
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.72	2.65	2.60	2.57	2.55	2.53	2.51	2.46
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.62	2.54	2.50	2.47	2.44	2.43	2.40	2.35
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.53	2.46	2.41	2.38	2.36	2.34	2.31	2.26
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.46	2.39	2.34	2.31	2.28	2.27	2.24	2.19
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.40	2.33	2.28	2.25	2.22	2.20	2.18	2.12
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.35	2.28	2.23	2.19	2.17	2.15	2.12	2.07
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.31	2.23	2.18	2.15	2.12	2.10	2.08	2.02
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.27	2.19	2.14	2.11	2.08	2.06	2.04	1.98
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.23	2.16	2.11	2.07	2.05	2.03	2.00	1.94
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.20	2.12	2.07	2.04	2.01	1.99	1.97	1.91
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.18	2.10	2.05	2.01	1.98	1.96	1.94	1.88
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.15	2.07	2.02	1.98	1.96	1.94	1.91	1.85
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24	2.20	2.13	2.05	2.00	1.96	1.93	1.91	1.88	1.82
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.11	2.03	1.97	1.94	1.91	1.89	1.86	1.80
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16	2.09	2.01	1.96	1.92	1.89	1.87	1.84	1.78
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.07	1.99	1.94	1.90	1.87	1.85	1.82	1.76
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.17	2.13	2.06	1.97	1.92	1.88	1.86	1.84	1.81	1.74
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12	2.04	1.96	1.91	1.87	1.84	1.82	1.79	1.73
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.03	1.94	1.89	1.85	1.83	1.81	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.01	1.93	1.88	1.84	1.81	1.79	1.76	1.70
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.07	2.04	1.96	1.88	1.82	1.79	1.76	1.74	1.70	1.63
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.92	1.84	1.78	1.74	1.72	1.69	1.66	1.59
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05	2.01	1.97	1.89	1.81	1.75	1.71	1.68	1.66	1.63	1.55
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.99	1.95	1.87	1.78	1.73	1.69	1.66	1.63	1.60	1.52
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.84	1.75	1.69	1.65	1.62	1.59	1.56	1.48
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.93	1.89	1.81	1.72	1.66	1.62	1.59	1.57	1.53	1.45
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.91	1.88	1.79	1.70	1.64	1.60	1.57	1.54	1.51	1.43
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.90	1.86	1.78	1.69	1.63	1.59	1.55	1.53	1.49	1.41
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.89	1.85	1.77	1.68	1.62	1.57	1.54	1.52	1.48	1.39

STATISTICA INFERENZIALE – 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 7 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?