Metodologia di sperimentazione in agricoltura

Andrea Onofri

2019-01-11

Indice

Capitolo 1

Introduzione

1.1 Organizzazione del corso

Questo corso si compone di due parti, che sono trattate in parziale sovrapposizione durante durante il semestre. Nella prima parte, sono fornite le basi teoriche e gli strumenti pratici per pianificare, organizzare e condurre esperimenti scientifici nel settore agrario.

Nella seconda parte, dopo un'introduzione agli argomenti generali relativi alla biostatistica, ci occupiamo dell'analisi e dell'interpretazione dei dati ottenuti nelle prove scientifiche, nonché della presentazione dei risultati tramite tesi, report e/o pubblicazione scientifica. In questa seconda parte, daremo ampio spazio all'impiego di modelli matematici interpretativi, utilizzati sia con finalità descrittive, sia con finalità predittive.

1.2 Esercitazioni pratiche

Questo corso non ha finalità puramente teoriche, ma fondamentalmente pratiche, improntate al 'saper fare'. In sostanza, all fine del corso, gli studenti dovranno essere in grado di pianificare esperimenti ed analizzarne i risultati. E'importante quindi lavorare con alcuni casi studio relativi ad esperimenti molto comuni in ambito agrario, come le prove varietali, le prove di concimazione, le prove di diserbo, i dosaggi biologici e i saggi germinativi.

Per poter risolvere gli esempi proposti, lo studente dovrà quindi avvalersi di un opportuno software statistiche. In questo caso, abbiamo due proposte: Excel (molto comune) e R (freeware e molto avanzato). A lezione tratteremo prevalentemente R, ma gli studenti, soprattutto quelli che non seguono le lezioni, potranno anche affidarsi ad Excel. In questa dispensa proporremo soluzioni per entrambi i software.

Obiettivi specifici del Corso

Come dicevamo, il Corso ha un contenuto prettamente tecnico e gli obiettivi specifici sono fondamentalmente pratici, con una forte attenzione al 'saper fare'.

Gli studenti dovranno:

1. Rispondere a domande generali sulla metodologia sperimentale e sull'organizzazione degli esperimenti
2. Saper disegnare correttamente l'esperimento proposto (uno), scegliendo adeguatamente il layout sperimentale (disegno, numero di repliche) e disegnando la mappa in modo corretto. Gli esperimenti proposto apparterranno alle seguenti categorie: prove di confronto varietale, prove di confronto tra erbicidi, prove di degradazione dei fitofarmaci, dosaggi biologici con erbicidi, prove di concimazione, prove di epoca di semina;
3. Saper eseguire l'ANOVA per il dataset proposto (uno), scelto tra quelli riportati nell'apposito elenco;
4. Saper adattare il modello opportuno (lineare o nonlineare) ad un dataset assegnato, scelto tra quelli riportati nell'apposito elenco. Saper valutare la bontà di adattamento del modello.

Per il punto 3 e il punto 4 gli studenti si dovranno avvalere di un supporto informatico, ad esempio costituito dal software R (citation) o da Excel, con la macro DSAASTAT (per l'ANOVA).

Capitolo 2

Il metodo sperimentale: quando la scienza è scienza

2.1 Introduzione

In una società caratterizzata dal sovraccarico cognitivo immagino sia giusto chiedersi (e chiedere) che cosa sia la scienza, cosa distingua le informazioni scientifiche da tutto quello che invece non è altro che pura opinione, magari autorevole, ma senza il sigillo dell'oggettività.

Per quanto affascinante possa sembrare l'idea del ricercatore che con un'improvviso colpo di genio elabora una stupefacente teoria, dovrebbe essere chiaro che l'intuizione è solo un possibile punto di partenza, che non necessariamente prelude al progresso scientifico, per quanto geniale ed innovativa possa essere. In generale, almeno in ambito biologico, nessuna teoria acquisisce automaticamente valenza scientifica, ma rimane solo nell'ambito delle opinioni, indipendentemente dal fatto che nasca da un colpo di genio, oppure grazie ad un paziente e meticoloso lavoro di analisi intellettuale, che magari si concretizza in un modello matematico altamente elegante e complesso.

Da un punto di vista puramente intuitivo, è ovvio aspettarsi che una prova scientifica debba uscire dall'ambito delle opinioni legate a divergenze di cultura, percezione e/o credenze individuali, per divenire, al contrario, oggettiva e universalmente valida, distinguendosi quindi da altre verità di natura metafisica, religiosa o pseudoscientifica. Che cosa è che permette alla scienza di divenire tale?

A questo proposito, riporto alcuni aforismi significativi:

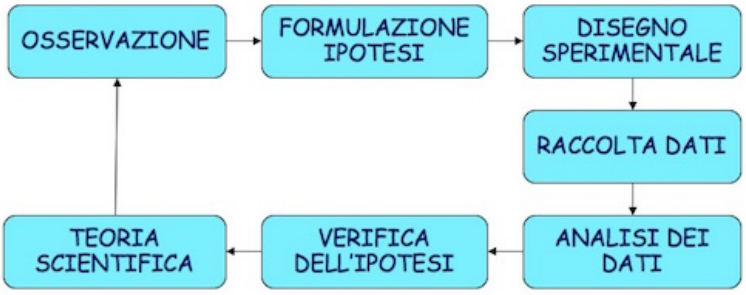


Figura 2.1: Il metodo scientifico Galileiano

1. Proof is a justified true belief (Platone, Dialoghi)
2. The interest I have in believing a thing is not a proof of the existence of that thing (Voltaire)
3. A witty saying proves nothing (Voltaire)

2.1.1 Cosa è quindi una prova scientifica?

La base di tutta la scienza risiede nel cosiddetto 'metodo scientifico', che si fa comunemente risalire a Galileo Galilei (1564-1642) e che è riassunto nella figura seguente.

Senza andare troppo in profondità, è importante notare due aspetti:

1. il ruolo fondamentale dell'esperimento scientifico, che produce dati a supporto di ipotesi pre-esistenti;
2. lo sviluppo di teorie basate sui dati, che rimangono valide fino a che non si raccolgono altri dati che le confutano, facendo nascere nuove ipotesi che possono portare allo sviluppo di nuove teorie, più affidabili o più semplici.

Insomma, l'ingrediente fondamentale di una prova scientifica è che è supportata da un insieme dei dati sperimentali; di fatto, non esiste scienza senza dati! Resta famoso l'aforisma "In God we trust, all the others bring data", attribuito all'ingegnere e statistico americano W. Edwards Deming (1999-1923), anche se pare che egli in realtà non l'abbia mai pronunciato.

2.2 Esperimenti buoni e cattivi!

Non tutti gli esperimenti sono buoni e, di conseguenza, non tutti i dati sono buoni. In particolare, due sono gli elementi che possono portare a dati di diversa affidabilità:

1. Errore sperimentale
2. Campionamento

Vediamo qualche dettaglio in più a proposito di questi due elementi.

2.2.1 L'errore sperimentale

Alla base della raccolta di dati sperimentali vi è un processo di **misurazione**, attraverso la quale il fenomeno oggetto di studio viene caratterizzato con appositi strumenti scientifici, più o meno complessi. Il problema è che nessuna misura può essere considerata precisa in senso assoluto, cioè perfettamente coincidente col valore reale della grandezza misurata, che è destinato a rimanere un'entità incognita e indeterminabile.

In particolare, in ogni esperimento scientifico esiste un potenziale elemento di confusione che gli scienziati conoscono con il termine di **errore sperimentale**, con la cui presenza è necessario confrontarsi sempre e comunque.

Nel misurare una determinata grandezza fisica, indipendentemente dal metodo scelto per la misura, possiamo sempre commettere due tipi di errore: **sistematico** ed **accidentale (casuale)**.

L'errore sistematico è provocato da difetti intrinseci dello strumento o incapacità peculiari dell'operatore e tende a ripetersi costantemente in misure successive. Un esempio tipico è quello di una bilancia non tarata, che tende ad aggiungere 20 grammi ad ogni misura che effettuiamo. Per queste sue peculiarità, l'errore sistematico non è quantificabile e deve essere contenuto al minimo livello possibile tramite la perfetta taratura degli strumenti e l'adozione di metodi di misura rigidamente standardizzati e accettati dalla comunità scientifica mondiale.

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

1. *malfunzionamenti accidentali dello strumento.* Si pensi ad esempio al rumore elettrico di uno strumento, che fa fluttuare i risultati delle misure effettuate;

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

Dato che queste imprecisioni sono assolutamente casuali è chiaro che le fluttuazioni positive (misura maggiore di quella vera) sono altrettanto probabili di quelle negative e tendono a presentarsi con la stessa frequenza quando si ripetano le misure più volte. Di conseguenza, l'errore sperimentale casuale può essere gestito attraverso la **replicazione delle misure**: infatti, se ripeto una misura soggetta a questo tipo di errore, nel lungo periodo gli errori positivi e negativi tendono ad annullarsi reciprocamente e la media delle misure effettuate tende quindi a coincidero con il valore reale della grandezza da misurare.

2.2.2 Il campionamento

Se è vero, e la pratica sperimentale lo conferma, che ripetere le misure porta ad ottenere molti risultati diversi, nasce il problema di capire quante repliche sono necessarie. Se si ripensa a quanto detto finora, dovrebbe risultare evidente che, per ottenere una misura pari all'effettivo (reale) valore della grandezza da misurare, bisognerebbe effettuarne infinite. Tuttavia è altrettanto evidente che questo procedimento è totalmente improponibile!!!

Qual è la strada seguita dagli scenziati, quindi? E' quella di raccogliere un numero finito di misure, sufficientemente basso da essere compatibile con le umane risorse di tempo e denaro, ma sufficientemente alto da essere giudicato attendibile. Qualunque sia questo valore finito, è evidente che ci troviamo difronte solo ad un campione delle infinite misure che avremmo dovuto fare,

Figura 2.2: Conseguenze di un esperimento sbagliato

ma che non abbiamo fatto. La domanda è: questo campione è rappresentativo o no? E' in grado di descrivere adeguatamente la realtà? E' possibile che gli errori sperimentali positivi e negativi non si siano annullati tra loro, confondendosi con l'effetto biologico in studio? In altre parole: possiamo fidarci dei dati che abbiamo raccolto?

La possibilità di raccogliere dati sbagliati è tutt'altro che remota. Gli scienziati american Pons e Fleischmann il 23 Marzo del 1989 diffusero pubblicamente la notizia di essere riusciti a riprodurre la fusione nucleare fredda, causando elevatissimo interesse nella comunità scientifica. Purtroppo le loro misure erano vizite da una serie di problemi e il loro risultato fu clamorosamente smentito da esperimenti successivi.

2.3 Scienza = metodo

Insomma, la scienza deve essere basata sui dati, ma i dati contengono inevitabili fonti di incertezza, legate all'errore sperimentale e al processo di campionamento. Come si può procedere in queste condizioni? Il punto fondamentale

è quello di adottare un metodo sperimentale che consenta di ottenere dati **il più affidabili possibile**. Insomma, questa semplice affermazione significa che bisogna fare eseperimenti ben condotti, precisi, seguendo procedure standardizzate e/o largamente condivise dalla comunità scientifica.

Certo è che, per quanto detto in precedenza, il fatto che i dati provengano da un processo di campionamento impedisce, di fatto, di ottenere un'affidabilità totale. Cosa succederebbe se ripetessimo l'esperimento?

Insomma, bisogna fare alcune considerazioni, che elenco di seguito:

1. in primo luogo si dovrà accettare il fatto che, contrariamente a quanto si potrebbe o vorrebbe credere, non esistono prove scientifiche totalmente certe, ma l'incertezza è un elemento intrinseco della scienza.
2. In secondo luogo si dovranno utilizzare gli strumenti della statistica necessari per quantificare l'incertezza residua, che dovrà essere sempre riportata a corredo dei risultati di ogni esperimento scientifico.
3. Ogni risultato sarà quindi valutato dalla comunità scientifica sullo sfondo della sua incertezza, seguendo alcune regole di natura probabilistica che consentono di stabilire se la prova scientifica è sufficientemente forte per essere considerata tale.

Un elemento fondamentale di valutazione della bontà di un esperimento e dei dati da esso ottenuti sta nella cosiddetta **replicabilità**, cioè nella probabilità di ottenere risultati molto simili (se non uguali) replicando l'esperimento in condizioni analoghe. Per valutare se un esperimento è replicabile è necessario che questo sia descritto con un grado di dettaglio tale da permettere a chinque di ripeterlo, ottenendo risultati comparabili e non contraddittori. Nessun risultato di cui non sia provata la riproducibilità è da considerarsi valido.

E'chiaro comunque che ogni esperimento può essere smentito. Questo non è un problema: la scienza è pronta a considerare una prova scientifica valida fino a che non si raccolgono dati altrettanto affidabili che la confutino. In questo caso, si abbandona la teoria confutata e si abbraccia la nuova. L'abbandono può anche non essere totale: ad esempio la teoria gravitazionale di Newton è ancora oggi valida per molto situazioni pratiche, anche se è stata abbandonoata in favore della teoria della relatività, che spiega meglio il moto dei corpi ad altissime velocità.

In effetti, la scienza considere sempre con attenzione il principio del rasoio di Occam, per il quale si accetta sempre la teoria più semplice per interpretare una dato fenomeno, riservando le teorie più complesse alle situazioni più difficili, che giustificano tale livello di complessità.

2.4 Chi valuta se un esperimento è attendibile?

Quanto detto finora vorrebbe chiarire come il punto centrale della scienza non è la certezza delle teorie, bensì il metodo che viene utilizzato per definirle. Ognuno di noi è quindi responsabile di verificare che le informazioni in suo possesso siano 'scientificamente' attendibili, cioè ottenute con un metodo sperimentale adeguato. Il fatto è che non sempre siamo in grado di compiere questa verifica, perché non abbiamo strumenti 'culturali' adeguati, se non nel ristretto ambito delle nostre competenze professionali. Come fare allora?

L'unica risposta accettabile è quella di controllare l'attendibilità delle fonti di informazione. In ambito biologico, le riviste autorevoli sono caratterizzate dal procedimento di '*peer review*', nel quale i manoscritti scientifici, prima della pubblicazione, sono sottoposti ad un comitato editoriale ed assegnati ad un 'editor', il quale legge il lavoro e contemporaneamente lo invia a due o tre scienziati anonimi e particolarmente competenti in quello specifico settore scientifico (*reviewers* o revisori).

I revisori, insieme all'*editor*, compiono un attento lavoro di esame e stabiliscono se l'evidenza scientifica presentata è sufficientemente 'forte'. Le eventuali critiche vengono presentate all'autore, che è tenuto a rispondere in modo convincente, anche ripetendo gli esperimenti se necessario. Il processo richiede spesso interi mesi ed è abbastanza impegnativo per uno scenziato. E' piuttosto significativa l'immagine presentata in scienceBlog.com, che allego qui.

In sostanza il meccanismo di *peer review* è l'analogo scientifico di un processo, nel quale l'inputato (lavoro scientifico) viene assolto (rilasciato, leggi: rigettato) in presenza di qualunque ragionevole dubbio metodologico. Attenzione: il dubbio che non deve esistere è quello metodologico, dato che il dubbio sul risultato non può essere allontanato completamente e i reviewer controlleranno solo che esso si trovi al disotto della soglia massima, stabilita con metodiche statistiche.

Questo procedimento, se effettuato con competenza, dovrebbe aiutare a separare la scienza dalla pseudo-scienza e, comunque, ad eliminare la gran parte degli errori metodologici dai lavori scientifici.

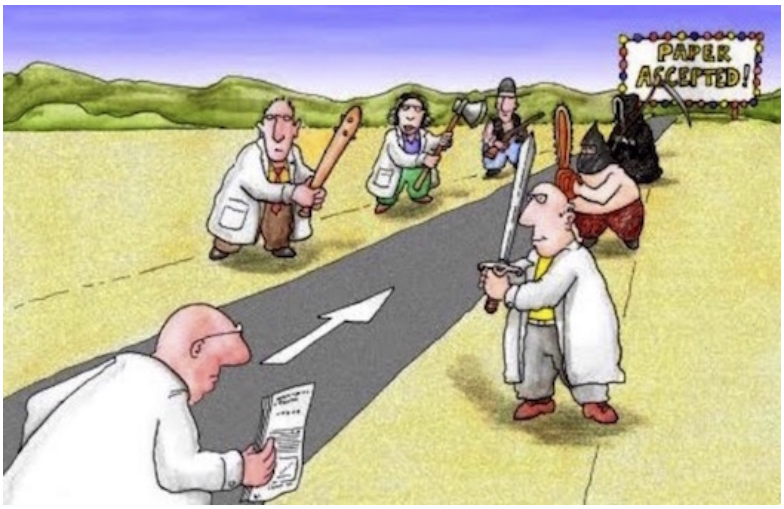


Figura 2.3: Il processo di *peer review*

2.5 Il metodo sperimentale

Almeno in ambito biologico, la definizione del metodo sperimentale è fondamentalmente attribuita allo scienziato inglese Ronald Fisher (1890-1962), che l'ha esplicitata nel suo famoso testo del 1935 (The design of experiments). Mi sembra opportuno riassumerla nelle tre espressioni 'chiave': controllo locale degli errori, replicazione e randomizzazione. Si tratta di:

1. contenere al massimo possibile l'errore sperimentale, con l'adozione di tecniche opportune, in modo da separare le fonti di variabilità, isolando qualla oggetto di studio (controllo locale degli errori);
2. replicare le misure più volte (replicazione)
3. Scegliere le unità sperimentali da misurare in modo totalmente casuale, così da avere un campione rappresentativo ed evitare di confondere gli effetti prodotti dall'errore sperimentale con quelli prodotti dal fenomeno biologico oggetto di studio (randomizzazione)

Vediamo ora un'esempio banale di come procedere.

2.6 Metodi sperimentali validi ed invalidi

Immaginiamo un ricercatore che abbia un'idea brillante: egli ha inventato un nuovo fertilizzante 'prodigioso'. E' evidente che non può presentarsi

alla comunità scientifica declamando le doti di questo fertilizzante, in quanto egli verrebbe immediatamente esposto al pubblico ludibrio, perchè sta presentando delle opinioni, non delle evidenze scientifiche (almeno, così dovrebbe essere, in una società sana... purtroppo in un'era di pseudo-scienza siamo sempre pronti a dar credito a chiunque, senza un'adeguata dose di scetticismo...)

2.6.1 Primo esperimento

Come ogni scenziato, egli deve raccogliere dati. E lo fa, organizzando un esperimento, nel quale prende un campo di mais e lo fertilizza con il suo nuovo composto, ritraendo una produzione del 20% superiore a quella usuale. Ovvoamente, se prova a pubblicare questa notizia, il suo lavoro verrà certamente (si spera...) rigettato, in quanto rimane il dubbio su chi sia la causa dell'effetto riscontrato: il fertilizzante? il clima dell'anno di prova? il suolo? la varietà di mais impiegata? E' chiaro che questo non è un esperimento controllato: il campo trattato e quello non trattato (riferimento) non sono totalmente uguale, eccetto che per il fertilizzante impiegato.

2.6.2 Secondo esperimento

A questo punto il ricercatore pianifica un esperimento comparativo controllato: prende due campio di mais, vicini, con lo stesso terreno, semina la stessa varietà di mais e coltiva i due campi esattamente nello stesso modo, con l'unica differenza che in uno di essi somministra il fertilizzante in studio (campo trattato) e nell'altro no (testimone o controllo). Alla fine osserva che il campo trattato produce 130 tonnellate per ettaro, mentre quello non trattato ne produce 115 e conclude che il nuovo fertilizzante è efficace (+ 12% circa). Infatti egli ritiene che, dato che i due campi sono totalmente uguali, l'incremento di produzione non possa che essere attribuito al fertilizzante. Scrive un report, che, purtroppo, viene rigettato.

Anche se questo secondo esperimento è meglio del primo, permane tuttavia il dubbio che l'effetto si sia prodotto per caso. Potrebbe infatti esserci stata una qualche situazione non osservata che ha avvantaggiato uno dei due campi. Ad esempio un attacco di insetti, una carenza idrica, o qualsivoglia altra situazione. Questo vi sembra improbabile? Non importa, con una sola osservazione il ricercatore non è in grado di provare che il risultato è replicabile.

2.6.3 Terzo esperimento

Avendo imparato la lezione, il ricercatore fa un nuovo esperimento, utilizzando stavolta otto campi: quattro trattati e quattro non trattati. Anche in questo caso osserva un incremento produttivo medio del 12% circa ed è sicuro che l'effetto è replicabile, perché lo ha osservato più volte. Purtroppo, anche questo esperimento non viene considerato affidabile e, di conseguenza, il lavoro non è pubblicabile. Stavolta il problema è che il ricercatore ha scelto i campi trattati con un criterio sistematico (un campo trattato ed uno 'tradizionale' contiguo), cosìcchè i campi trattati sono tutti a sinistra di quelli 'tradizionali'. Ciò crea un ragionevole dubbio: e se vi fosse un gradiente di fertilità da destra verso sinistra? Questo potrebbe dare origine ad una produttività maggiore dei campi a destra, rispetto a quelli a sinistra. In presenza di questo 'ragionevole' dubbio, la prova non può avere valenza scientifica.

2.6.4 Quarto esperimento: quello buono

Il ricercatore prende allora otto campi ed assegna il trattamento a quattro di essi, scelti in modo totalmente casuale. In questo caso è sicuro che, anche se vi fosse un qualche elemento estraneo di confusione (gradiente di fertilità, attacco di insetti...), esso dovrebbe colpire le unità sperimentali casualmente disposte senza creare vantaggi particolari all'uno o all'altro dei due trattamenti. Ovviamente egli non è certo (e non può esserlo) che l'esperimento sia del tutto attendibile; infatti potrebbe essere stato così sfortunato che un qualche elemento estraneo ignoto si è accanito proprio sulle parcelle non trattate, danneggiandone la produttività. Solo che, grazie alla scelta casuale, questa evenienza diviene altamente improbabile, così da rendere i dubbi irragionevoli. In questo caso l'esperimento è controllato, replicato e randomizzato e il risultato ottenuto, in quanto ragionevolmente attendibile, può essere pubblicato.

2.7 Incertezza residua

Insomma, un esperimento valido è controllato, replicato e randomizzato. Tuttavia le misure raccolte sono poche e sono solo un campione di tutte quelle possibili. Infatti il nostro ricercatore ha usato otto campi, ma ne avrebbe potuti usare 16, 32 e così via. Rimane quindi il dubbio, che, se facessimo

altre misure (cioè ampliassimo il campione), queste potrebbero invalidare i risultati ottenuti fino a quel momento.

Se mi è concesso un paragone calcistico, è un po' come chiedersi come finirà una partita di calcio dopo aver assistito solo al primo tempo: in alcune circostanze, quando una delle due squadre ha mostrato una chiara superiorità, la previsione è abbastanza facile, mentre in altre circostanze l'equivalenza dei valori in campo la rende alquanto difficile. In tutti i casi, si tratta solo di una previsione, che può essere sempre smentita alla prova dei fatti.

Anche la scienza funziona così. Noi osserviamo solo il primo tempo, che, nel caso del nostro ricercatore, consiste di otto misure. Osserviamo che le quattro misure del fertilizzato sono tutte in modo consistente molto più alte di quelle del non trattato e quindi possiamo concludere, con ragionevole certezza, che il fertilizzante è efficace. Altrimenti, se la produzione media del trattato è solo lievemente più alta, il nostro esperimento potrà essere inconclusivo, cioè incapace di fugare i dubbi sull'effettiva efficacia del nostro fertilizzante. Avremo bisogno di fare altre prove di conferma.

2.8 Il ruolo della statistica

Abbiamo visto che un esperimento scientifico, anche se ben fatto (controllato, replicato e randomizzato), può portare a evidenze scientifiche più o meno forti. In quest'ottica, la statistica ci fornisce gli strumenti per riassumere le misure effettuate, calcolarne l'incertezza e rappresentare la forza dell'evidenza scientifica, in modo da poter prendere decisioni sull'efficacia dei trattamenti e sull'esigenza di ulteriori verifiche. Imparare a conoscere e comprendere questi strumenti statistici è l'obiettivo di questo corso.

2.9 Conclusioni

In conclusione, possiamo ripartire dalla domanda iniziale: "Che cosa è la scienza?", per rispondere che è scienza tutto ciò che è supportato da dati che abbiano passato il vaglio della *peer review*, dimostrando di essere stati ottenuti con un procedimento sperimentale privo di vizi metodologici e di essere sufficientemente affidabili in confronto alle fonti di incertezza cui sono associati.

Qual è il *take-home message* di questo articolo? Fidatevi solo delle riviste scientifiche attendibili, cioè quelle che adottano un serio processo di *peer review* prima della pubblicazione.

Capitolo 3

Introduzione al disegno sperimentale

3.1 Definizioni

La ricerca scientifica trova la sua unità elementare nell'esperimento, cioè un *processo investigativo, con il quale, seguendo un adeguato protocollo, si osserva e si misura la risposta prodotta da uno o più fattori sperimentali nei soggetti coinvolti nello studio.* Raramente gli esperimenti sono isolati, più spesso fanno parte di uno sforzo collettivo organizzato, generalmente identificato come progetto di ricerca.

Ogni esperimento deve essere attentamente pianificato. Infatti, sappiamo che la variabilità esistente tra soggetti sperimentali, il campionamento, le irregolarità di misura e molti altri fattori perturbativi ci impediscono di osservare la realtà con assoluta precisione. E' come se osservassimo un fenomeno attraverso una sorta di lente deformante, che ci impone di adottare un metodo sperimentale rigoroso, per evitare di attribuire al fenomeno in studio effetti che sono invece puramente casuali.

In particolare, gli esperimenti debbono essere:

1. Precisi
2. Accurati
3. Replicabili/Riproducibili

In mancanza di questi requisiti, al termine dell'esperimento possono rimanere dubbi sui risultati, tali da inficiare la validità delle conclusioni raggiun-

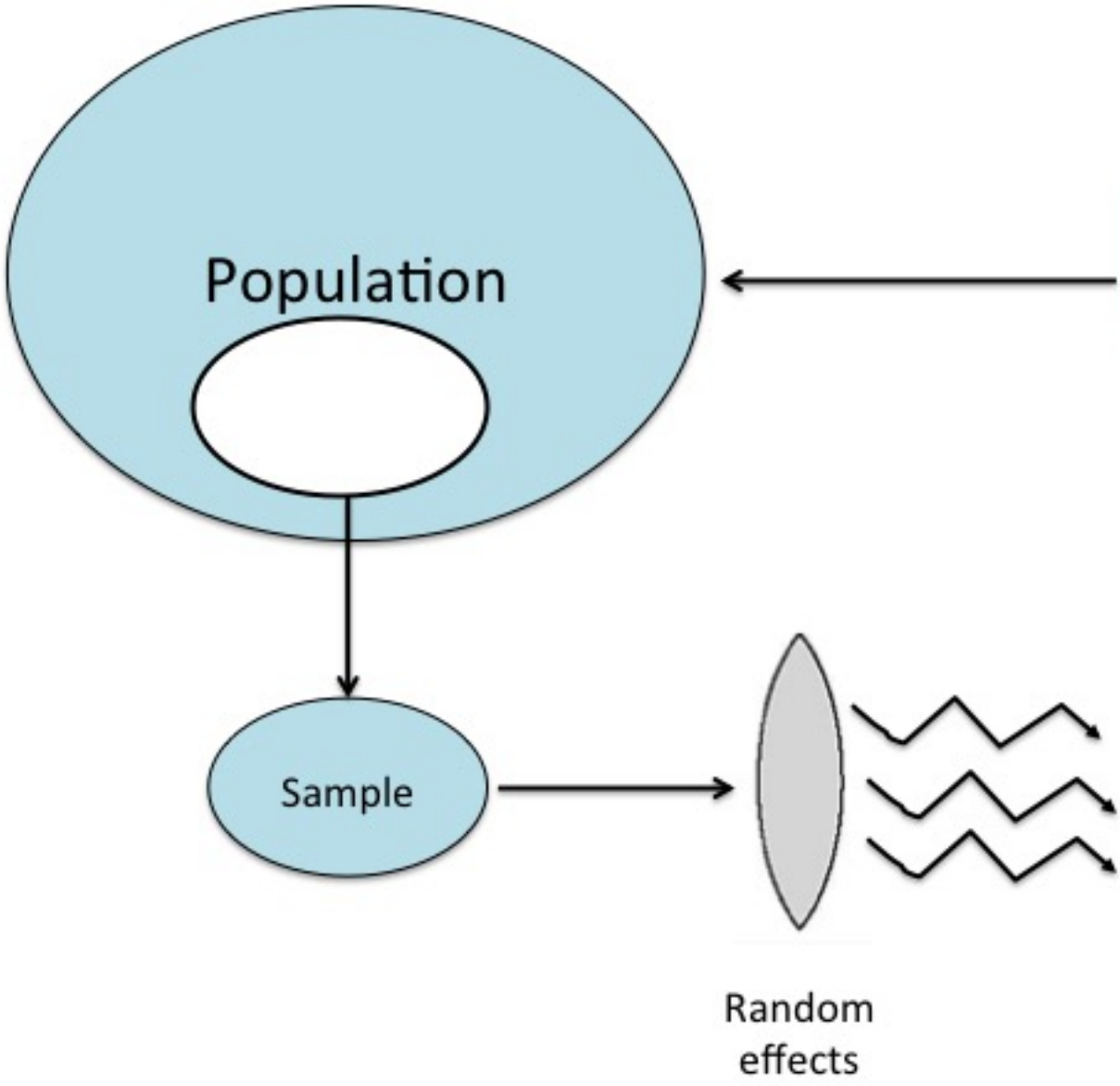


Figura 3.1: Schematizzazione del processo sperimentale

te.

Forse vale la pena di chiarire cosa si intende con precisione, accuratezza e replicabilità/riproducibilità. Abbiamo già visto che la presenza dell'errore sperimentale ci impone di ripetere le misure più volte. La precisione di un esperimento non è altro che la variabilità dei risultati tra una replica e l'altra.

La precisione, da sola, non garantisce che l'esperimento sia affidabile. Abbiamo menzionato nel capitolo precedente che l'errore sperimentale può essere casuale o sistematico. Quest'ultimo può essere dovuto, per esempio, ad uno strumento non accurato che sovrastima tutte le misure. In questo caso, posso ripetere cento volte la misura, ottenendo sempre lo stesso risultato, molto preciso, ma totalmente inaffidabile, nel senso che non riflette la misura reale del soggetto. E'questa l'accuratezza, cioè la capacità di una misura (o di un esperimento) di restituire il valore reale, anche se dopo come media di numero molto elevato di replicazioni.

L'accuratezza è molto più importante della precisione: infatti una misura accurata, anche se imprecisa, riflette bene la realtà, anche se in modo vago. Al contrario, una misura precisa ma inaccurata ci porta completamente fuori strada, perchè non riflette la realtà! Un esperimento/risultato non accurato si dice 'distorto' (*biased*).

Oltre a precisione ed accuratezza, siamo anche interessati alla replicabilità di un esperimento, cioè alla possibilità che questo, se ripetuto in condizioni assolutamente analoghe (stessi soggetti, ambiente, strumenti...) restituisca risultati equivalenti. Alcuni biostatistici distinguono la replicabilità dalla riproducibilità, in quanto considerano quest'ultima come la possibilità di ottenere risultati equivalenti ripetendo una misura in condizioni diverse (diversi soggetti, diverso ambiente...). E'evidente che un esperimento può essere totalmente accurato e replicabile, ma non riproducibile con soggetti e condizioni ambientali diverse. Se è così, le conclusioni raggiunte, anche se accurate, non possono essere generalizzate.

3.2 Elementi fondamentali del disegno sperimentale

La metodica di organizzazione di un esperimento prende il nome di *disegno sperimentale* e deve essere sempre adeguatamente formalizzata trami-

te la redazione di un *protocollo sperimentale* sufficientemente dettagliato da consentire a chiunque la replicazione dell'esperimento e la verifica dei risultati.

Le basi del disegno sperimentale si fanno in genere risalire a Sir Ronald A. Fisher, vissuto in Inghilterra dal 7 Febbraio 1890 al 29 luglio 1962. Laureatosi nel 1912, lavora come statistico per il comune di Londra, fino a quando diviene socio della prestigiosa Eugenics Education Society di Cambridge, fondata nel 1909 da Francis Galton, cugino di Charles Darwin. Dopo la fine della guerra, Karl Pearson gli propone un lavoro presso il rinomato Galton Laboratory, ma egli non accetta a causa della profonda rivalità esistente tra lui e Pearson stesso. Nel 1919 viene assunto presso la Rothamsted Experimental Station, dove si occupa dell'elaborazione dei dati sperimentali e, nel corso dei successivi 7 anni, definisce le basi del disegno sperimentale ed elabora la sua teoria della "analysis of variance". Il suo libro più importante è "The design of experiment", del 1935. E' sua la definizione delle tre componenti fondamentali del disegno sperimentale:

1. controllo degli errori;
2. replicazione;
3. randomizzazione.

3.2.1 Controllo degli errori

Controllare gli errori, o, analogamente, eseguire un esperimento controllato signfica fondamentalmente due cose:

1. adottare provvedimenti idonei ad evitare le fonti di errore, mantenendole al livello più basso possibile (alta precisione);
2. agire in modo da isolare l'effetto in studio (accuratezza), evitando che si confonda con effetti casuali e di altra natura.

Declinare questi principi richiederebbe una vita di esperienza! Vogliamo solo ricordare alcuni aspetti fondamentali.

Campionamento corretto

E'evidente che il primo requisito di un esperimento è una corretta scelta delle unità sperimentali, cioè le più piccole unità che ricevono lo 'stimolo' rappresentato dal trattamento, in modo indipendente da tutte le altre.

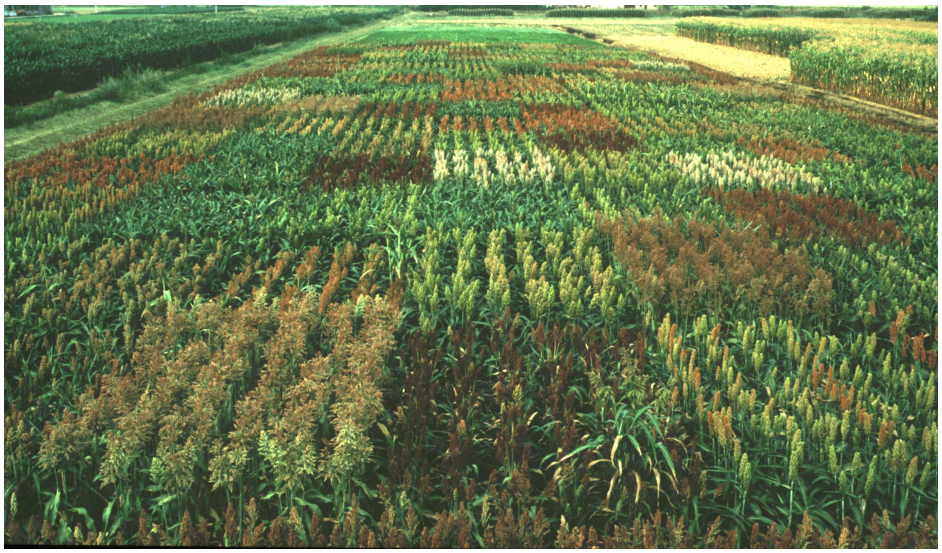


Figura 3.2: Una prova sperimentale in campo (Foto D. Alberati)

E' bene subito comprendere una fondamentale distinsione tra unità sperimentali e unità osservazionali. Le prime sono appena state definite; le seconde sono quelle che costituiscono l'oggetto della misura e possono anche non coincidere con le prime. Ad esempio: immaginiamo di trattare con un diserbante due vasetti, in modo indipendente l'uno dall'altro. Immaginiamo poi di pesare singolarmente le quattro piante di ciascun vasetto; in questa situazione, il vasetto è l'unità sperimentale, le piante sono invece le unità osservazionali. L'elemento discriminante di questo esempio è l'indipendenza: mentre le unità sperimentali hanno ricevuto il trattamento in modo indipendente l'una dall'altra, le unità osservazionali no. Questa differenza è fondamentale, per motivi che vedremo più avanti.

Le unità sperimentali possono essere di varia natura; nel caso degli esperimenti di campo, le unità sperimentali sono dette **parcelle** e sono un pezzetto di terreno, di varia forma e dimensione.

Le unità sperimentali sono scelte per campionamento, che è un elemento fondamentale dell'esperimento. Infatti, il censimento, che riguarda tutti i soggetti di un certo ambito, non è, in se', un esperimento. Ovviamente, il campione deve essere rappresentativo, altrimenti l'esperimento è invalido.

Non è possibile dare indicazioni specifiche di campionamento, perché queste dipendono dalla tipologia di esperimento. Illustriamo quindi solo alcuni criteri generali.

Prima di campionare, dobbiamo avere una chiara visione della cornice di campionamento, cioè della popolazione da cui io devo campionare. Devo effettuare un esperimento valido per l'Italia centrale, per una località particolare, per tutta Italia. Devo fare un esperimento che riguarda una stalla in particolare o tutte le stalle dove si allevano bovini? Di quale razza? E' un passaggio fondamentale, in quanto poi le conclusioni non possono che riferirsi alla cornice della popolazione da cui il campione è stato estratto, non altre. Per esperimenti nell'ambito delle scienze sociali, diviene fondamentale che la cornice di campionamento abbia le seguenti caratteristiche:

1. le unità sono tutte identificabili e reperibili
2. le unità sono tutte caratterizzate (es. Id)
3. è aggiornata e organizzata logicamente
4. non mancano soggetti (che potrebbero quindi sfuggire al campionamento)
5. non ci sono soggetti duplicati
6. non ci sono elementi estranei

Una volta che la popolazione è nota ed organizzata, dobbiamo trovare un criterio di selezione. Fondamentalmente ci sono tre possibilità:

1. campionamento randomizzato (casuale)
2. campionamento stratificato
3. campionamento sistematico

Il campionamento randomizzato è tale che ogni soggetto ha la stessa possibilità di ogni altro di essere incluso nel campione. Tipicamente, questo campionamento è basato su un generatore di numeri casuali, con distribuzione uniforme delle frequenze. Il codice sottostante serve per ottenere cinque elementi casuali da un lotto di 48 identificati con numeri progressivi.

```
sample(1:48, 4)
```

```
## [1] 20 27 42 43
```

Il campionamento casuale può non dare garanzie sufficienti di rappresentatività. Per questo motivo, a volte, si utilizza il **campionamento stratificato**, con il quale si divide la cornice di campionamento in gruppi omogenei e si prelevano un certo numero di soggetti da ogni gruppo. In questo caso è bene ricordare che potrebbe essere auspicabile mantenere nel campione la stessa relazione tra gruppi che esiste nella popolazione. Ad esempio, se in una popolazione di insetti c'è il 10% di maschi e il 90% di femmine, io devo prelevare n maschi e m femmine, tale che $n/m = 0.1$, altrimenti il campione che ottengo potrebbe non essere rappresentativo.

A volte, il campionamento può essere sistematico, nel senso che utilizzo un criterio non casuale, ma in grado di assicurare una certa rappresentatività. Ad esempio, per campionare gli edifici di una via, potrei decidere di prendere il primo a caso e poi procedere prendendone uno si e tre no. Questo campionamento è molto veloce e di facile esecuzione, ma può dare origine a distorsioni.

Una forma di campionamento è quella a cluster: in questo caso suddivido gli elementi in gruppi, scelgo a caso un certo numero di gruppi e poi prendo tutti gli elementi di un gruppo. Ad esempio, devo selezionare i bambini delle scuole elementari del comune di Perugia. In questo caso, invece che selezionare i bambini, posso più velocemente selezionare le scuole e prendere tutti i bambini delle scuole selezionate. Evidentemente il metodo si basa sull'ipotesi che la selezione rappresentativa delle scuole crea anche una selezione rappresentativa di bambini.

A volte si esegue anche un campionamento a quota, cioè si prendono tutti i soggetti che si incontrano in una certa situazione, fino a che non se ne raccolgono un numero prefissato, per alcune classi specificate (Es. 30 donne, 25 uomini, 15 adolescenti e 20 bambini). Questo tipo di campionamento è talvolta utilizzato negli esperimenti medici.

Rigore

Direi che questo aspetto è ovvio e non richiede commenti particolari: una ricerca deve essere condotta 'a regola d'arte'. E' evidente che, ad esempio, se vogliamo sapere la cinetica di degradazione di un erbicida a 20 °C dovremo realizzare una prova esattamente a quella temperatura, con un erbicida uniformemente distribuito nel terreno, dentro una camera climatica capace di un controllo perfetto della temperatura. Gli strumenti dovranno essere ben tarati e sarà necessario attenersi scrupolosamente a metodi validati e largamente condivisi.

Tuttavia, a proporsito di rigore, non bisogna scordare quanto diceva C.F. Gauss a proposito della precisione nei calcoli, e che puà essere anche riferito al rigore nella ricerca : "*Manca di mentalità matematica tanto chi non sa riconoscere rapidamente ciò che è evidente, quanto chi si attarda nei calcoli con una precisione superiore alla necessità*"

Omogeneità

Anche in questo caso, l'importanza di scegliere soggetti uniformi e posti in un ambiente uniforme (nello spazio e nel tempo) è evidente. Bisogna comunque tener presente che i risultati di un esperimento si estendono alla popolazione da cui il campione è estratto e della quale esso rappresenta le caratteristiche. Esperimenti nei quali si restringe il campo di variabilità dei soggetti e dell'ambiente sono certamente più precisi, ma forniscono anche risultati meno generalizzabili. L'importante è avere ben chiaro su quale è il campo di validità che si vuole dare ai risultati. Ad esempio, se si vuole ottenere un risultati riferito alla collina umbra, bisognerà scegliere soggetti che rappresentano bene la variabilità pedo-climatica della collina Umbra; né più, né meno.

Evitare le 'intrusioni demoniache'

Secondo Hurlbert (1984), le intrusioni sono eventi totalmente casuali che impattano negativamente con un esperimento in corso. E'evidente che, ad esempio, un'alluvione, l'attacco di insetti o patogeni, la carenza idrica hanno una pesante ricaduta sulla precisione di un esperimento e sulla sua riuscita. Nello stesso lavoro, Hurlbert usa il termine 'intrusione demoniaca' per indicare quelle intrusioni che, pur casuali, avrebbero potuto essere previste con un disegno più accurato, sottolineando in questo caso la responsabilità dello sperimentatore.

Un esempio è questo: uno sperimentatore vuole studiare l'entità della predazione dovuta alle volpi e quindi usa campi senza steccionate (dove le volpi possono entrare) e campi protetti da steccionate (e quindi liberi da volpi). Se le steccionate, essendo utilizzate dai falchi come punto d'appoggio, finiscono per incrementare l'attività predatoria di questi ultimi, si viene a creare un'intrusione demoniaca, che rende l'esperimento distorto. Il demonio, in questo caso, non è il falco, che danneggia l'esperimento, ma il ricercatore stesso, che non ha saputo prevedere una possibile intrusione.

3.2.2 Replicazione

In ogni esperimento, i trattamenti dovrebbe essere replicati su 2 o più unità sperimentali. Ciò permette di:

1. dimostrare che i risultati sono replicabili (ma non è detto che siano riproducibili!)
2. rassicurare che eventuali circostanze aberranti casuali non abbiano provocati risultati distorti
3. misurare l'errore sperimentale, come variabilità di risposta tra repliche trattate nello stesso modo (precisione dell'esperimento)
4. incrementare la precisione dell'esperimento (più sono le repliche più l'esperimento è preciso, perchè si migliora la stima della caratteristica misurata, diminuendo l'incertezza)

Per poter essere utili, le repliche debbono essere indipendenti, cioè debbono **aver subito tutte le manipolazioni necessarie per l'allocazione del trattamento in modo totalmente indipendente l'una dall'altra**. Le manipolazioni comprendono tutte le pratiche necessarie, come ad esempio la preparazione delle soluzioni, la diluizione dei prodotti, ecc..

La manipolazione indipendente è fondamentale, perchè in ogni parte del processo di trattamento possono nascondersi errori più o meno grandi, che possono essere riconosciuti solo se colpiscono in modo casuale le unità sperimentali. Se la manipolazione è, anche solo in parte, comune, questi errori colpiscono tutte le repliche allo stesso modo, diventano sistematici e quindi non più riconoscibili. Di conseguenza, si inficia l'accuratezza dell'esperimento. Quando le repliche non sono indipendenti, si parla di **pseudorepliche**, contrapposte alle **repliche vere**.

Il numero di repliche dipende dal tipo di esperimento: più sono e meglio è, anche se è necessario trovare un equilibrio accettabile tra precisione e costo dell'esperimento. Nella sperimentazione di campo, 2 repliche sono poche, 3 appena sufficienti, 4 costituiscono la situazione più comune, mentre un numero maggiore di repliche è abbastanza raro, non solo per la difficoltà di seguire l'esperimento, ma anche perché aumentano la dimensione della prova e, di conseguenza, la variabilità del terreno.

3.2.3 Randomizzazione

L'indipendenza di manipolazione non garantisce da sola un esperimento corretto. Infatti potrebbe accadere che le caratteristiche innate dei soggetti, o una qualche 'intrusione' influenzino in modo sistematico tutte le unità sperimentali trattate nello stesso modo, così da confondersi con l'effetto del trattamento. Un esempio banale è che potremmo somministrare un farmaco a quattro soggetti in modo totalmente indipendente, ma se i quattro sog-

getti fossero sistematicamente più alti di quelli non trattati finiremmo per confondere una caratteristica innata con l'effetto del farmaco. Oppure, se le repliche di un certo trattamento si trovassero tutte vicine alla scolina, potrebbero essere più danneggiate delle altre unità sperimentali dal ristagno idrico, il cui effetto si confonderebbe con quello del trattamento stesso.

Questi problemi sono particolarmente insidiosi e si nascondono anche dietro ai particolari apparentemente più insignificanti. La randomizzazione è l'unico sistema per evitare, o almeno rendere molto improbabile, la confusione dell'effetto del trattamento con fattori casuali e/o comunque diversi dal trattamento stesso. La randomizzazione si declina in vari modi:

1. allocazione casuale del trattamento alle unità sperimentali. Gli esperimenti che prevedono l'allocazione del trattamento sono detti 'manipolativi' o 'disegnati'.
2. A volte l'allocazione del trattamento non è possibile o non è etica. Se volessimo studiare l'effetto delle cinture di sicurezza nell'evitare infortuni gravi, non potremmo certamente provocare incidenti deliberati. In questo caso la randomizzazione è legata alla scelta casuale di soggetti che sono 'naturalmente' trattati. Esperimenti di questi tipo, si dicono **osservazionali**. Un esempio è la valutazione dell'effetto dell'inquinamento con metalli pesanti nella salute degli animali: ovviamente non è possibile, se non su piccola scala, realizzare il livello di inquinamento desiderato e, pertanto, dovremo scegliere soggetti che sono naturalmente sottoposti a questo genere di inquinamento, magari perché vivono vicino a zone industriali.
3. Se i soggetti sono immobili, la randomizzazione ha anche una connotazione legata alla disposizione spaziale e/o temporale casuale.

L'assegnazione casuale del trattamento, o la selezione casuale dei soggetti trattati, fanno si che tutti i soggetti abbiano la stessa probabilità di ricevere qualunque trattamento oppure qualunque intrusione casuale. In questo modo, la probabilità che tutte le repliche di un trattamento abbiano qualche caratteristica innata o qualche intrusione comune che li penalizzi/avvantaggi viene minimizzata. Di conseguenza, confondere l'effetto del trattamento con variabilità casuale ('confounding'), anche se teoricamente possibile, diviene altamente improbabile.

Gradienti e blocking

Un esperimento in cui l'allocazione del trattamento, o la scelta dei soggetti trattati, o la disposizione spaziale dei soggetti sono totalmente casuali si dice 'completamente randomizzato'. E' perfettamente valido, perchè non pone dubbi fondati di inaccuratezza. Tuttavia, in alcune circostanze è possibile porre restrizioni (vincoli) alla randomizzazione, perchè ciò porta ad un esperimento più preciso.

In particolare, le unità sperimentali possono presentare delle differenze, ad esempio di fertilità, oppure di sesso. Ad esempio, randomizzare completamente l'allocazione dei trattamenti potrebbe far si che tra le repliche di un trattamento vi siano più maschi che femmine, il che crea un certo livello di 'confounding'. Pertanto potrebbe essere utile divider i soggetti in due gruppi (maschi e femmine), oppure in più gruppi (molto fertile, mediamente fertile, poco fertile...) e randomizzare i trattamenti all'interno di ogni gruppo.

In generale, il *blocking* consiste nel suddividere i soggetti in gruppi uniformi e ripetere lo stesso esperimento (o parte di esso) all'interno di ciascun gruppo, cioè in una situazione di maggiore omogeneità.

Il raggruppamento delle unità sperimentali può tener conto di:

1. vicinanza spaziale (campi, parcelle, stalle ...)
2. caratteristiche fisiche (età, peso, sesso ...)
3. vicinanza temporale
4. gestione dei compiti (tecnico, valutatore, giudice ...)

Chiaramente, randomizzare all'interno del gruppo invece che randomizzare completamente crea un vincolo. A volte i vincoli sono più di uno. Vediamo un esempio. Una certa operazione industriale richiede un solo operatore per essere portata a termine, ma può essere eseguita in quattro modi diversi. Pianificate un esperimento per stabilire qual è il metodo più veloce, avendo a disposizione solo quattro operatori.

L'unità sperimentale è il lavoratore. I metodi sono quattro e, volendo lavorare con quattro repliche, avremmo bisogno di sedici operatori per disegnare un esperimento completamente randomizzato. Possiamo tuttavia considerare che un operatore, in quattro turni successivi, può operare con tutti e quattro i metodi. Quindi possiamo disegnare un esperimento in cui il turno fa da unità sperimentale e l'operatore fa da blocco (blocchi randomizzati). Tuttavia, in ogni blocco (operatore) vi è un gradiente, nel senso che i turni successivi al primo sono via via meno efficienti, perché l'operatore accumula stanchezza.

Per tener conto di questo potremmo allora introdurre un vincolo ulteriore, per ogni operatore, randomizzando i quattro metodi tra i turni, in modo che ogni metodo, in operatori diversi, capiti in tutti i turni. In sostanza, l'operatore fa da blocco, perché in esso sono contenuti tutti i metodi. Ma anche il turno (per tutti gli operatori) fa da blocco, in quanto in esso sono ancora contenuti tutti i metodi. Se non vi è chiaro, ci torneremo sopra più tardi.

Posto che non si deve violare l'indipendenza delle repliche, l'inclusione di vincoli alla randomizzazione è consentita, **ma questa deve sempre essere tenuta presente in fase di analisi dei dati**.

Ronald Fisher diceva "*Analyse them as you have randomised them*". Meglio seguire il consiglio.

E se ricercatori/soggetti sono influenzabili?

Per concludere questa parte, è opportuno menzionare il fatto che, in un esperimento scientifico, il fatto che lo sperimentatore e il soggetto siano coscienti del trattamento somministrato può portare a risultati distorti. Per esempio, nell'eseguire un rilievo, lo sperimentatore può essere influenzato dal sapere con quale diserbante è stata trattata una parcella, cercando inconsciamente conferme alle sue conoscenze pregresse. D'altro canto, nei soggetti sperimentali dotati di coscienza (uomo) sapere di essere stati trattati può influenzare l'esito del trattamento (effetto placebo).

Per evitare questi problemi, soprattutto in ambito medico, un esperimento può essere pianificato come:

1. cieco: l'unità sperimentale o lo sperimentatore non sono coscienti dei dettagli del trattamento;
2. doppio cieco: né l'unità sperimentale né lo sperimentatore sono a coscienza dei dettagli del trattamento

Un esperimento cieco e/o doppio cieco possono non essere eticamente corretti oppure inutili, nel qual caso si torna ad un esperimento tradizionale 'aperto' (*open experiment*: Tutti sanno tutto')

3.2.4 Esperimenti non validi

A questo punto dovrebbero essere chiare le caratteristiche di un esperimento valido. A completamento, cerchiamo di elencare le caratteristiche di un

esperimento non valido.

1. Cattivo controllo degli errori
2. Fondati sospetti di confounding
3. Mancanza di repliche vere
4. Confusione tra repliche vere e pseudo-repliche
5. Mancanza di randomizzazione
6. Presenza di vincoli alla randomizzazione, trascurati in fase di analisi.

Le conseguenze di queste problematiche sono abbastanza diverse.

Cattivo controllo degli errori

Bisogna verificare se il problema è relativo a questioni come la mancanza di scrupolosità, l'uso di soggetti poco omogenei o di un ambiente poco omogeneo, o altri aspetti che inficiano solo la precisione, ma non l'accuratezza dell'esperimento. In questo caso, l'esperimento è ancora valido (accurato), ma la bassa precisione probabilmente impedirà di trarre conclusioni forti. Quindi, un esperimento impreciso si 'elimina' da solo, perché sarà inconclusivo. Di questi esperimenti bisogna comunque diffidare, soprattutto quando siano pianificati per mostrare l'assenza di differenze tra due trattamenti alternativi. Mostrare l'assenza di differenze è facile: basta fare male un esperimento, in modo che vi sia un alto livello di incertezza e quindi l'evidenza scientifica sia molto debole.

Diversa è la situazione in cui un cattivo controllo degli errori, ad esempio l'adozione di metodi sbagliati, porta a mancanza di accuratezza, cioè a risultati che non riflettono la realtà (campionamento sbagliato, ad esempio; oppure strumenti non tarati; impiego di metodi non validati e/o non accettabili). In questo caso venendo a mancare l'accuratezza, l'esperimento deve essere rigettato, in quanto non fornisce informazioni realistiche.

Confounding e correlazione spuria

Abbiamo appena menzionato il problema fondamentale della ricerca, cioè il **confounding**, vale a dire la confusione tra l'effetto del trattamento e un qualche altro effetto casuale, legato alle caratteristiche innate del soggetto o a qualche intrusione più o meno 'demoniaca'. Abbiamo detto che non possiamo mai avere la certezza dell'assenza di confounding, ma abbiamo anche detto che l'adozione di una pratica sperimentale corretta ne minimizza la probabilità.

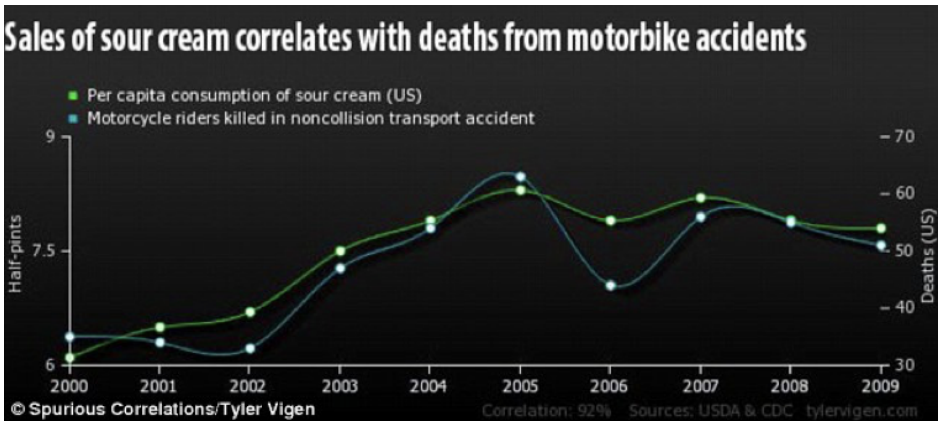


Figura 3.3: Esempio di correlazione spuria

Chiaramente, rimangono dei rischi che sono tipici di situazioni nelle quali il controllo adottato non è perfetto, come capita, ad esempio, negli esperimenti osservazionali. In questo ambito è piuttosto temuta la cosiddetta 'correlazione spuria', una forma di confounding casuale per cui due variabili variano congiuntamente (sono direttamente o inversamente proporzionali), ma in modo del tutto casuale. Esistono, ad esempio, dati che mostrano una chiara correlazione tra le vendite di panna acida e le morti per incidenti in motocicletta. Chiaramente, non esistono spiegazioni scientifiche per questo effetto, che è, ovviamente, del tutto casuale. Il problema è che questa correlazione spuria non è sempre così semplice da rintracciare.

A volte il confounding non è casuale, ma è legato ad una variabile esterna che si agisce all'insaputa dello sperimentatore. Ad esempio, è stato osservato che il tasso di crimini è più alto nelle città che hanno più chiese. La spiegazione di questo paradosso sta nel fatto che esiste un 'confounder', cioè l'ampiezza della popolazione. Nelle grandi città si riscontrano sia una maggiore incidenza criminale, sia un grande numero di chiese. In sostanza, la popolazione determina sia l'elevato numero di chiese che l'elevato numero di crimini, ma queste ultime due variabili non sono legate tra loro da una relazione causa-effetto (A implica B e A implica C, ma B non implica C).

Il confounding non casuale è spesso difficile da evidenziare, soprattutto se le correlazioni misurate sono spiegabili. Inoltre, non è eliminabile con un'accurata randomizzazione, ma solo con l'esecuzione di un esperimento totalmente controllato, nel quale ci si preoccupa di rilevare tutte le variabili necessarie per spiegare gli effetti riscontrati. Di questo è importante tener conto soprattutto negli esperimenti osservazionali, dove il controllo è sempre più difficile e meno completo.

Pseudo-repliche e randomizzazione poco attenta

Per evidenziare questi problemi e comprendere meglio la differenza tra un esperimento corretto e uno non corretto, è utilissima la classificazione fatta da Hurlbert (1984), che riportiamo di seguito.

Vengono mostrati 8 soggetti, sottoposti a due trattamenti (bianco e nero), con 8 disegni sperimentali diversi.

Il disegno A1 è corretto, in quanto si tratta di un esperimento completamente randomizzato. Ugualmente, è valido il disegno A2, nel quale le unità sperimentali sono state divise in quattro gruppi omogenei e sono state trattate in modo randomizzato all'interno di ogni gruppo.

Il disegno A3 è quantomeno 'sospetto': vi sono repliche vere, ma l'allocazione dei trattamenti non è randomizzata ed avviene con un processo sistematico per il quale 'nero' e 'bianco' si alternano. Cosa succederebbe se vi fosse un gradiente di fertilità decrescente da destra verso sinistra? Le unità nere sarebbero avvantaggiate rispetto alle bianche! Insomma, rimangono sospetti di confounding, a meno che non si sia assolutamente certi dell'assenza di gradienti, come capita ad esempio se all'interno dei blocchi, dobbiamo creare una sequenza spazio-temporale. Vediamo tre esempi:

1. ho quattro piante e, per ogni pianta, voglio confrontare un ramo basso con uno alto: è evidente che i due trattamenti sono sempre ordinati in modo sistematico (basso prima di alto).
2. Dobbiamo valutare l'effetto di fitofarmaci somministrati in due epoche diverse (accestimento e inizio-levata); anche qui non possiamo randomizzare, giacchè un'epoca precede sempre l'altra.
3. Dobbiamo confrontare la presenza di residui di un fitofarmaco a due profondità e non possiamo randomizzare, perché una profondità precede sempre l'altra nello spazio.

In queste situazioni l'esperimento rimane valido, anche se la randomizzazione segue un processo sistematico e non casuale.

Il disegno B1 è usualmente invalido: non vi è randomizzazione e ciò massimizza i problemi del disegno A3: la separazione delle unità sperimentali 'bianche' e 'nere' non consente una valutazione adeguata dell'effetto del trattamento, che è confuso con ogni potenziale differenza tra la parte destra e la sinistra dell'ambiente in cui la sperimentazione viene eseguita. Ovviamente, la separazione può essere non solo spaziale, ma anche temporale. Anche in

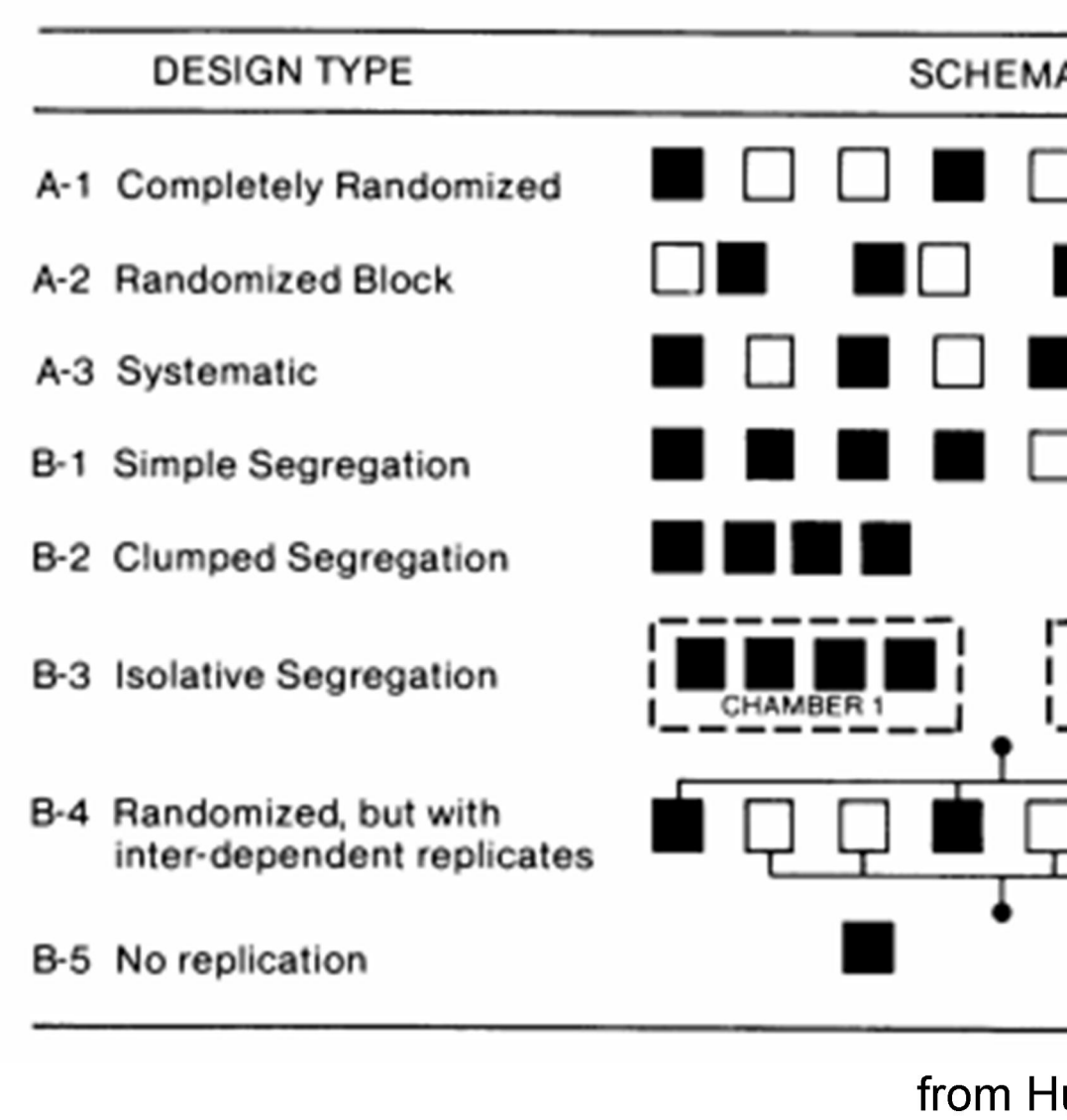


Figura 3.4: Indicazioni per una corretta randomizzazione (Hurlbert, 1984)

questo caso diamo alcuni esempi in cui una situazione come quella descritta in B1 è valida:

1. Vogliamo confrontare la produzione in pianura e in collina. Ovviamente dobbiamo scegliere campioni in due situazioni fisicamente separate
2. Vogliamo confrontare la pescosità di due laghetti
3. Vogliamo confrontare la produttività di due campi contigui.

Queste situazioni sono valide, anche se con una restrizione: non siamo in grado di stabilire a chi debba essere attribuito l'effetto. Ad esempio, per la prima situazione, pianura e collina possono dare produzioni diverse per il suolo diverso, il clima diverso, la precessione colturale diversa o un qualunque altro elemento che differenzi le due località.

Il disegno B2 è analogo al disegno B1, ma il problema è più grave, perché la separazione fisica è più evidente. Questo disegno è totalmente sbagliato, a meno che non siamo specificatamente interessati all'effetto località (vedi sopra).

Il disegno B3 è analogo al disegno B2, ma costituisce una situazione molto frequente nella pratica scientifica. Immaginiamo infatti di voler confrontare la germinazione dei semi a due temperature diverse, utilizzando due camere climatiche e mettendo, in ognuna di esse, quattro capsule Petri identiche. In questa situazione, l'effetto temperatura è totalmente confuso con l'effetto 'camera climatica (località)' e risente di ogni malfunzionamente relativo ad una sola delle due camere. Inoltre, le unità sperimentali con lo stesso trattamento di temperature non sono manipolate in modo indipendente, dato che condividono la stessa camera climatica. Di conseguenza, non si può parlare di repliche vere, bensì di **pseudorepliche**.

Altri esempi di **pseudorepliche** sono schematizzati con il codice B4. Ad esempio:

1. trattare piante in vaso ed analizzare in modo indipendente i singoli individui invece che tutto il vaso;
2. trattare una parcella di terreno e prelevare da essa più campioni, analizzandoli separatamente;
3. trattare una capsula Petri ed analizzare separatamente i semi germinati al suo interno.

Questi disegni, in assenza di repliche vere aggiuntive non sono da considerarsi validi. Ad esempio, se io ho due vasetti trattati in modo totalmente indipendente e da ciascuno di essi prelevo due piante e le analizzo separatamente, il

disegno è caratterizzato da due repliche vere e due pseudorepliche per ogni replica ed è, pertanto, valido.

Il disegno B5 è invece evidentemente invalido, per totale mancanza di repliche.

3.3 Progettazione di un esperimento (protocollo)

Qualunque sia l'ambito scientifico, in ogni esperimento possiamo individuare alcune fasi fondamentali, che proviamo ad elencare:

1. Individuazione del background (ricerca bibliografica)
2. ipotesi scientifica;
3. definizione dell'obiettivo;
4. identificazione dei fattore/i sperimentale/i;
5. identificazione dei soggetti sperimentali e delle repliche;
6. identificazione delle variabili da rilevare;
7. allocazione randomizzata dei trattamenti (mappa dell'esperimento)
8. Esecuzione dell'esperimento

Nell'analizzare questi aspetti, faremo riferimento ad alcuni esempi pratici, che verranno indicati tra breve.

3.3.1 Ipotesi scientifica → obiettivo dell'esperimento

Trascurando la parte di ricerca bibliografica, che è pur fondamentale, nel metodo scientifico galileiano, il punto di partenza di un esperimento è l'**ipotesi scientifica**, che determina l'obiettivo dell'esperimento. Si tratta del passaggio fondamentale dal quale dipende in modo logico tutto il lavoro successivo. Gli obiettivi debbono essere:

1. rilevanti
2. chiaramente definiti;
3. specifici;
4. misurabili;
5. raggiungibili/realistici;
6. temporalmente organizzati.

Il rischio che si corre con obiettivi mal posti è quello di eseguire una ricerca dispersiva, con raccolta di dati non necessari e/o mancanza di dati fondamentali, con costi più elevati del necessario e un uso poco efficiente delle risorse. In genere, prima si definisce un obiettivo generale, seguito da uno o più obiettivi specifici, in genere proiettati su un più breve spazio temporale e che possono essere visti anche come le fasi necessarie per raggiungere l'obiettivo generale.

Poniamo quattro esempi pratici.

3.3.2 Casi di studio - 1

Esempio 1 - Diserbo chimico

Si suppone che gli erbicidi A, B e C siano più efficaci di D, E ed F verso *Solanum nigrum*, una comune pianta infestante delle colture di pomodoro. L'obiettivo generale della ricerca sarà quello di trovare un'efficace soluzione per l'eliminazione di *Solanum nigrum* dal pomodoro. Gli obiettivi specifici saranno:

1. valutare l'efficacia erbicida di A, B e C, confrontandola con quella di D, E ed F;
2. valutare la selettività degli anzidetti erbicidi verso il pomodoro;

Esempio 2 - Valutazione varietale

L'ipotesi è che le varietà di girasole A, B e C non hanno la stessa base genetica e quindi non sono tutte ugualmente produttive. L'obiettivo generale è quello di capire quale tra A, B e C sia più adatta alle condizioni pedoclimatiche della collina Umbra.

Gli obiettivi specifici sono quelli di valutare:

1. produttività di A, B e C
2. stabilità produttiva di A, B e C

Esempio 3 - Diserbo parziale

Nella barbabietola da zucchero, il diserbo localizzato lungo la fila consente di diminuire l'impiego di erbicidi. Tuttavia, se la coltura precedente ha prodotto

semi e se non abbiamo effttuato una lavorazione profonda per interrarli, la coltura sarà più infestata e quindi sarà più difficile ottenere una buona produttività con il diserbo parziale.

Su questa ipotesi costruiamo un esperimento volto a valutare l'interazione tra lavorazione del terreno e diserbo chimico. Per raggiungere questo obiettivo generale, proveremo a valutare se:

1. il diserbo parziale consente di ottenere produzioni comparabili a quelle del diserbo totale; 2.l'effetto è indipendente dalla lavorazione effettuata.

Esempio 4 - Colture poliennali

L'ipotesi scientifica è affine a quella dell'esempio 2, ma, in questo caso, vogliamo porre a confronto tre varietà di erba medica (A, B e C). La differenza sta nel fatto che l'erba medica è una coltura poliennale e quindi vogliamo capire se il giudizio di merito è indipendente dall'anno di coltivazione.

I nostri obiettivi specifici saranno quindi:

1. valutare la produttività media delle varietà in prova
2. valutare le oscillazione nei quattro anni di durata del cotico erboso

Esempio 5 - Inquinamento da micotossine

Secondo le notizie in bibliografia, i datteri confezionati in vendita nei supermercati contengono elevate quantità di micotossine. L'obiettivo generale è quello di verificare il livello di infestazione e vedere se questo cambia con il metodo di confezionamento.

3.3.3 Identificazione dei fattori sperimentali

Dopo aver definito l'obiettivo di un esperimento, è necessario chiarire esattamente gli stimoli a cui saranno sottoposte le unità sperimentali. Uno 'stimolo' sperimentale prende il nome di **fattore sperimentale**, che può avere più **livelli**. I livelli del fattore sperimentale prendono il nome di **trattamenti (o tesi) sperimentali**.

3.3.4 Esperimenti (multi)fattoriali

In alcuni casi è necessario inserire in prova più di un fattore sperimentale. In questo caso si parla di esperimenti **fattoriali**, che possono essere **incrociati (crossed)** quando sono presenti in prova tutte le possibili combinazioni dei livelli di ogni fattore, oppure di esperimenti **innestati (nested)** quando i livelli di un fattore cambiano al cambiare dei livelli dell'altro.

Ad esempio:

1. Immaginiamo di voler studiare due fattori sperimentali: la varietà di girasole (tre livelli: A, B e C) e la concimazione (2 livelli: pollino e urea). Abbiamo quindi 6 possibili trattamenti (combinazioni): A-pollina, A-urea, B-pollina, B-urea, C-pollina e C-urea. Il disegno è completamente incrociato.
2. Immaginiamo di voler confrontare due specie in agricoltura biologica (orzo e triticale), con tre varietà ciascuna (A, B e C per orzo, D, E e F per triticale). Anche in questo caso abbiamo sei trattamenti: orzo-A, orzo-B, orzo-C, triticale-D, triticale-E e triticale-F, ma il disegno è innestato, perché per il fattore sperimentale 'varietà' i livelli cambiano a seconda dei livelli del fattore 'specie'.

3.3.5 Aggiungere un controllo?

In alcuni casi si pone il problema di inserire in prova un trattamento che funga da riferimento per tutti gli altri. In questi casi si parla comunemente di **controllo** o **testimone**, che può essere

- non sottosposto a trattamento
- trattato con placebo
- trattato secondo le modalità usuali di riferimento

3.3.6 Fattori sperimentali di trattamento e di blocco

Finora abbiamo menzionato quelli che, in lingua inglese, vengono definiti *treatment factor* (trattamenti sperimentali). Tuttavia, possono esserci altri fattori sperimentali non allocati, ma 'innati' e legati alla collocazione spazio-temporale o alle caratteristiche dei soggetti. Questi fattori vengono definiti, sempre in inglese, *blocking factors*. Di questi fanno parte, ad esempio, il

blocco, la località ed ogni altro elemento che permette di raggruppare i soggetti. Anche questi *blocking factors* devono essere chiaramente identificati ed elencati.

Su questa base identifichiamo i fattori sperimentali negli esempi precedenti.

3.3.7 Casi di studio - 2

Esempio 1

Il fattore sperimentale oggetto di studio sarà il diserbo del pomodoro, con 5 livelli inseriti in prova (6 trattamenti sperimentali): A, B, C, D, E ed F. Inoltre, si ritiene opportuno inserire in prova un testimone non trattato (NT), che ci permetterà di quantificare la percentuale di malerbe controllate. Inoltre, sarà anche necessario inserire in prova un testimone scerbato manualmente (ST), che ci permetterà di quantificare eventuali perdite produttive dovute alla competizione residua o alla fitotossicità del trattamento. In totale, avremo quindi 8 tesi sperimentali. Come usuale in pieno campo, l'esperimento verrà disegnato a blocchi randomizzati e sarà pertanto necessario inserire un fattore di blocco.

Esempio 2

Il fattore sperimentale in studio sarà la varietà di girasole con 3 livelli inclusi in prova (varietà A, B e C). Come testimone, inseriremo la varietà di riferimento per la zona (D). Dato che eseguiremo questa prova su un terreno nel quale vi sono due chiari gradienti di fertilità, disegneremo l'esperimento considerando due fattori di blocco: trasversale e longitudinale (spiego meglio tra poco...). Poichè dobbiamo valutare la stabilità produttiva, dovremo ripetere l'esperimento più volte (es. in tre anni diversi) e quindi avremo un secondo fattore sperimentale, incrociato con il primo.

Esempio 3

In questo caso avremo due fattori sperimentali incrociati: il diserbo con due livelli (totale o parziale, localizzato sulla fila) e la lavorazione con tre livelli (aratura profonda, aratura superficiale e *minimum tillage*). Non vi è

la necessità di un testimone, ma avremo la necessità di un fattore di blocco. In totale, avremo sei tesi sperimentali.

Esempio 4

Il fattore sperimentale in studio sarà la varietà di erba medica con 3 livelli inclusi in prova (varietà A, B e C) ai quali aggiungiamo il riferimento di zona (D) come testimone. Come nel caso del girasole, dovremo valutare la stabilità produttiva negli anni, ma, dato che abbiamo una coltura poliennale, non avremo bisogno di ripetere la prova, ma potremo ripetere le osservazioni per quattro anni sulla stessa prova.

Esempio 5

Per questo esperimento vengono considerate tre diverse modalità di confezionamento (carta, busta di plastica, scatola di plastica perforata). Non vi è necessità di un testimone, ma, dato che le diverse confezioni verranno acquistate in diversi supermercati e dato che sospettiamo differenze nella conservazione tra un supermercato e l'altro, utilizzeremo il supermercato come fattore di raggruppamento.

3.3.8 Identificazione delle unità sperimentali e delle repliche

Cornice di campionamento e numero di repliche

Per i primi quattro esempi verranno eseguite prove di pieno campo, nella Media Valle del Tevere, che rappresenta la cornice di campionamento adeguata per l'obiettivo previsto. Sappiamo di dover selezionare appezzamenti di terreno

1. rappresentativi della Media Valle del Tevere,
2. omogenei.

L'omogeneità dell'ambiente è fondamentale per aumentare la precisione dell'esperimento. La scelta dell'appezzamento è chiaramente fondamentale ed è guidata dall'esperienza, tenendo conto anche di aspetti come la facilità di

accesso e la vicinanza di strutture (laboratori, capannoni...) che consentano un'accurata esecuzione degli eventuali prelievi.

Oltre alla scelta dell'appezzamento, si possono anche utilizzare alcune strategie per favorire una buona omogeneità delle parcelle. Spesso si usa far precedere la prova da una coltura di 'omogeneizzazione', ad esempio avena, che è molto avida di azoto e lascia nel terreno poca fertilità residua. Oppure un prato di erba medica, che, grazie agli sfalci periodici, lascia il terreno libero da piante infestanti.

Trattandosi di esperimenti di campo, il numero di repliche sarà di quattro, per ogni trattamento e l'unità sperimentale sarà una parcella, della quale dovremo valutare forma e dimensioni.

Per il quinto esempio, la cornice di campionamento sarà data dal territorio del comune di Perugia. L'unità sperimentale sarà la confezione e la scelta del numero di repliche dovrà essere compatibile con la capacità di analisi per la determinazione dell'inquinamento da micotossine. E' ragionevole pensare che 30 repliche (90 confezioni totali) possano essere adeguate per rappresentatività e facilità di gestione.

Campionamento delle unità sperimentali

Per le quattro prove di pieno campo, una volta scelto l'appezzamento, dovremo campionare le parcelle di terreno. Questa operazione viene usualmente eseguita su carta, redigendo la **mappa dell'esperimento**. In primo luogo, si decide la **dimensione e la forma della parcella**.

L'aspetto fondamentale è che ogni parcella deve contenere un numero di piante sufficientemente alto da essere rappresentativo. Per questo motivo le colture a bassa fittezza hanno sempre bisogno di parcelle più grandi che non quelle ad alta fittezza. La dimensione non deve tuttavia eccedere una certa soglia, in quanto con essa aumenta anche la variabilità del terreno e, di conseguenza, diminuisce l'omogeneità dell'esperimento. Per questo motivo, talvolta si preferisce diminuire la dimensione delle parcelle ed, avendo lo spazio sufficiente, aumentare il numero delle repliche.

Nello stabilire la dimensione delle parcelle, dovremo tener conto del fatto che la parte più delicata è il bordo, in quanto le piante che si trovano lungo il bordo esterno risentono di condizioni diverse dalle altre piante situate al centro della parcella (**effetto bordo**). Questo determina variabilità all'interno della parcella, che possiamo minimizzare raccogliendo solo la parte centrale.

Si viene così a distinguere la superficie totale della parcella dalla superficie di raccolta (**superficie utile**), che può essere anche molto minore di quella totale.

In generale si ritiene che le colture ad elevata fittezza (frumento, cereali, erba medica...) dovrebbero avere parcelle di almeno 10-20 m^2, mentre a bassa fittezza (mis, girasole...) dovrebbero avere parcelle di almeno 20-40 m^2. Queste dimensioni sono riferite alla superficie utile di raccolta, non alla dimensione totale.

Per quanto riguarda la forma, le parcelle quadrate minimizzano l'effetto bordo, perché, a parità di superficie, hanno un perimetro più basso. Tuttavia esse sono di più difficile gestione, in quanto, considerando il fronte di lavoro di una seminatrice o una mietitrebbiatrice parcellare, possono richiedere la semina o la raccolta in più passate, il che finisce per essere una fonte di errore. Per questo motivo le parcelle sono usualmente rettangolari, con una larghezza pari a quella della macchina impiegata per la semina.

Per i quattro esempi in studio potremmo utilizzare una dimensione delle parcelle di 20 m^2 per l'erba medica (2 m di larghezza per 10 m di lunghezza) e di 22.5 m^2 per pomodoro, mais e barbabietola da zucchero (2.25 m di larghezza per 10 metri di lunghezza).

A questo punto possiamo redigere la mappa per il primo esempio. Dato che il campo di prova è largo 30 metri e lungo 400 metri, potremmo immaginare di disegnare otto file di percelle in senso trasversale (8 x 2.25m = 18m), di modo che l'esperimento non sia troppo lungo (il che ne aumenterebbe la variabilità), ma rimanga spazio sufficiente ai lati, per evitare di avvicinarsi troppo alle scoline, dove possono manifestarsi ristagni idrici.

La mappa dell'esperimento è un elemento fondamentale e deve riportare tutte le informazioni relative al disegno sperimentale. E'anche importante indicare la direzione del Nord, in modo da facilitare l'orientamento della mappa stessa. Notare inoltre che, intorno alla prova, abbiamo sistemato altre parcelle fuori esperimento con funzione di 'bordi'. In questo modo si evita che i bordi esterni delle parcelle esterne siano esposti a condizioni molto diverse dagli altri, cosa che potrebbe accentuare l'effetto 'bordo', di cui abbiamo parlato in precedenza. Queste parcelle verranno trattate in modo ordinario (semina e diserbo tradizionale del pomodoro).

Per l'esperimento relativo all'esempio 5, l'unità sperimentale è una singola confezione di datteri, con le tipologie previste dal piano.

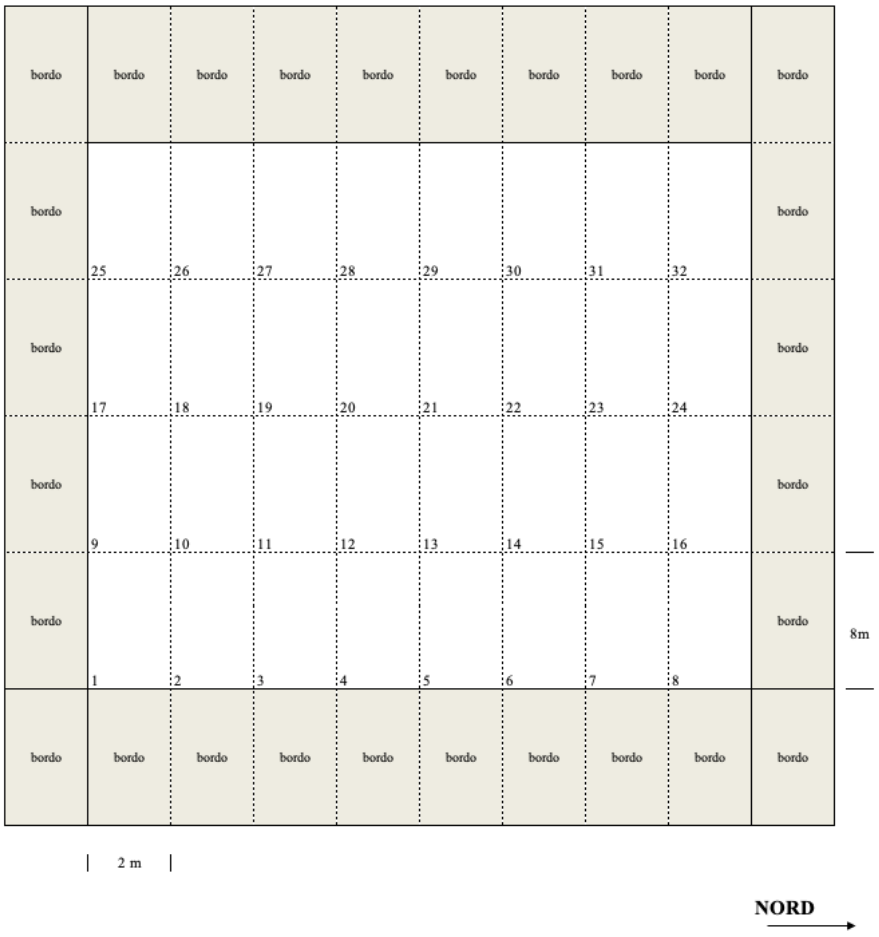


Figura 3.5: Mappa dell'esperimento relativo all esempio 1

Figura 3.6: Tipologie delle confezioni di datteri

Siccome è abbastanza scomodo campionare confezioni di datteri casualmente all'interno del Comune di Perugia, si preferisce un campionamento stratificato, selezionando 10 supermercati rappresentativi, nelle zone più densamente popolate della città. All'interno di ogni supermercato, si selezioneranno casualmente tre repliche per ogni tipo di confezione.

3.3.9 Scelta delle variabili da rilevare

Durate e al termine dell'esperimento, sarà necessario rilevare le più importanti caratteristiche dei soggetti sperimentali, sia quelle innate, sia quelle indotte dai trattamenti sperimentali. Per ogni singolo carattere, l'insieme delle modalità/valori che ognuno dei soggetti presenta prende il nome di **variabile** (proprio perché varia, cioè assume diversi valori, a seconda del soggetto). Ad esempio, quando stiamo studiando l'effetto di due diserbanti su piante infestanti appartenenti ad una certa specie, posto che l'unità sperimentale è costituita da una singola pianta, possiamo avere le seguenti variabili: il prodotto diserbante con cui ogni pianta è stata trattata, il peso di ogni pianta prima del trattamento, il peso di ogni pianta dopo il trattamento.

Le variabili sperimentali possono essere molto diverse tra di loro ed è piuttosto importante saperle riconoscere, perché questo condiziona il tipo di analisi statistica da eseguire.

Variabili nominali (categoriche)

Le variabili nominali esprimono, per ciascun soggetto, l'appartenenza ad una determinata categoria o raggruppamento. L'unica caratteristica delle categorie è l'esclusività, cioè un soggetto che appartiene ad una di esse non può appartenere a nessuna delle altre. Variabili nominali sono, ad esempio, il sesso, la varietà, il tipo di diserbante impiegato, la modalità di lavorazione e così via. Le variabili categoriche permettono di raggruppare i soggetti, ma non possono essere utilizzate per fare calcoli, se non per definire le proporzioni dei soggetti in ciascun gruppo.

Variabili ordinali

Anche le variabili ordinali esprimono, per ciascun soggetto, l'appartenenza ad una determinata categoria o raggruppamento. Tuttavia, le diverse categorie sono caratterizzate, oltre che dall'esclusività, anche da una relazione di ordine, nel senso che è possibile stabilire una naturale graduatoria tra esse. Ad esempio, la risposta degli agricoltori a domande relative alla loro percezione sull'utilità di una pratica agronomicapuò essere espressa utilizzando una scala con sei categorie (0, 1, 2, 3, 4 e 5), in ordine crescente da 0 a 5. Di conseguenza possiamo confrontare categorie diverse ed esprimere un giudizio di ordine (2 è maggiore di 1, 3 è minore di 5), ma non possiamo eseguire operazioni matematiche, tipo sottrarre dalla categoria 3 la categoria 2 e così via, dato che la distanza tra le categorie non è necessariamente la stessa.

Variabili quantitative discrete

Le variabili discrete sono caratterizzate dal fatto che possiedono, oltre alle proprietà dell'esclusività e dell'ordine, anche quella dell'equidistanza tra gli attributi (es., in una scala a 5 punti, la distanza – o la differenza – fra 1 e 3 è uguale a quella fra 2 e 4 e doppia di quella tra 1 e 2). Le variabili discrete consentono la gran parte delle operazioni matematiche e, spesso, possono essere analizzate con metodiche parametriche, facendo riferimento alla distribuzione normale, che, pur essendo continua, in alcune condizioni può essere assunta come buona approssimazione di molte distribuzioni discrete.

Variabili quantitative continue

Le variabili quantitative continue possiedono tutte le proprietà precedentemente esposte (esclusività delle categorie, ordine, distanza) oltre alla continuità, almeno in un certo intervallo. Tipiche variabili continue sono l'altezza, la produzione, il tempo, la fittezza...

Dato che gli strumenti di misura nella realtà sono caratterizzati da una certa risoluzione, si potrebbe arguire che misure su scala continua effettivamente non esistono. Tuttavia questo argomento è più teorico che pratico e, nella ricerca biologica, consideriamo continue tutte le misure nelle quali la risoluzione dello strumento è sufficientemente piccola rispetto alla grandezza da misurare. Viceversa, le variabili continue sono piuttosto rare nelle scienze economiche e sociali in genere.

La quantità di informazione fornita dagli strumenti di valutazione cresce passando dalle scale nominali, di più basso livello, a quelle quantitative continue, di livello più elevato. Variabili esprimibili con scale quantitative continue o discrete possono essere espresse anche con scale qualitative, adottando un'opportuna operazione di classamento. Il contrario, cioè traformare in quantitativa una variabile qualitativa, non è invece possibile.

Rilievi visivi e sensoriali

Nella pratica sperimentale è molto frequente l'adozione di metodi di rilievo basati sull'osservazione di un fenomeno attraverso uno dei sensi (più spesso, la vista, ma anche gusto e olfatto) e l'assegnazione di una valutazione su scala categorica, ordinale o, con un po' di prudenza, quantitativa discreta o continua. Ad esempio, il ricoprimento delle piante infestanti, la percentuale di controllo di un erbicida e la sua fitotossicità vengono spesso rilevati visivamente, su scale da 0 a 100 o simili.

I vantaggi di questa tecnica sono molteplici:

1. Basso costo ed elevata velocità
2. Possibilità di tener conto di alcuni fattori perturbativi esterni, che sono esclusi dalla valutazione, contrariamente a quello che succede con metodi oggettivi di misura
3. non richiede strumentazione costosa

A questi vantaggi fanno da contraltare alcuni svantaggi, cioè:

1. Minor precisione (in generale)

2. Soggettività
3. L'osservatore può essere prevenuto
4. Difficoltà di mantenere uniformità di giudizio
5. Richiede esperienza specifica e allenamento

I rilievi sensoriali sono ben accettati nella pratica scientifica in alcuni ambiti ben definiti, anche se richiedono attenzione nell'analisi dei dati non potendo essere assimilati *tout court* con le misure oggettive su scala continua.

Variabili di confondimento

Quando si pianificano i rilievi da eseguire, oppure anche nel corso dell'esecuzione di un esperimento, bisogna tener presente non soltanto la variabile che esprime l'effetto del trattamento, ma anche tutte le variabili che misurano possibili fattori di confondimento.

Ad esempio, immaginiamo di voler valutare la produttività di una specie arborea in funzione della varietà. Immaginiamo anche di sapere che, per questa specie, la produttività dipende anche dall'età. Se facciamo un esperimento possiamo utilizzare alberi della stessa età per minimizzare la variabilità dei soggetti. Tuttavia, se questo non fosse possibile, per ogni albero dobbiamo rilevare non solo la produttività, ma anche l'età, in modo da poter valutare anche l'effetto di questo fattore aggiuntivo e separarlo dall'effetto della varietà. In questo modo l'esperimento diviene molto più preciso.

3.3.10 Casi di studio - 3

Per gli esempi in studio, immaginiamo per semplicità di dover rilevare la produzione per gli esempi da 1 a 4 e il contenuto di micotossine per l'esempio 5. Inoltre, per l'esempio 2, immaginiamo di dover rilevare anche il peso di mille semi. Per questo, prenderemo dalla produzione di granella di ogni parcella, quattro subcampioni da mille semi, da sottoporre a successive misure.

3.3.11 Allocazione dei trattamenti

Il problema dell'allocazione dei trattamenti non si pone con l'esempio 5, in quanto, trattandosi di un esperimento osservazionale, le confezione sono già

'naturalmente' trattate, cioè appartengono già, all'atto del campionamento, alla tipologia di confezionamento prescelta.

Per quanto riguarda gli altri esempi, abbiamo già redatto la mappa secondo le necessità. A questo punto si pone il problema di decidere quali parcelle trattare con cosa, nel rispetto dei trattamenti e delle repliche prescelte. Per questo fine, semplici esperimentipossono anche essere disegnati a mano; per esperimenti più complessi potremo utilizzare il package *agricolae* in R (de Mendiburu, 2017).

3.3.12 Casi di studio - 4

Esempio 1

Questo esempio va disegnato a blocchi randomizzati; tuttavia, a titolo di esempio, esamineremo anche la possibilità che venga disegnato a randomizzazione completa. Quest'ultimo disegno è il più semplice e consiste nell'assegnare ogni trattamento a quattro parcelle casualmente scelte. Con R bisognerà prima creare il vettore dei nomi delle tesi e il numero di repliche per tesi

```
library(agricolae)
trt <- c("A", "B", "C", "D", "E", "F", "NT", "TS")
reps <- rep(4, 8)
design <- design.crd(trt, r=reps, seed=777, serie=0)
design$book
```

```
##    plots r trt
## 1      1 1   E
## 2      2 1   C
## 3      3 1   B
## 4      4 2   C
## 5      5 1   F
## 6      6 1  TS
## 7      7 1  NT
## 8      8 1   D
## 9      9 2  NT
## 10    10 1   A
## 11    11 2  TS
## 12    12 2   F
## 13    13 3  NT
```

```
## 14    14 3   C
## 15    15 3  TS
## 16    16 2   D
## 17    17 3   D
## 18    18 2   E
## 19    19 3   E
## 20    20 4  NT
## 21    21 2   A
## 22    22 4   D
## 23    23 4   E
## 24    24 3   A
## 25    25 4   C
## 26    26 2   B
## 27    27 4   A
## 28    28 3   F
## 29    29 3   B
## 30    30 4  TS
## 31    31 4   F
## 32    32 4   B
```

Possiamo ora riportare la randomizzazione sulla mappa disegnata in precedenza.

Questo schema è eccellente se l'ambiente è molto uniforme. Tuttavia, nel caso di un esperimento di camp è lecito aspettarsi un gradiente trasversale, dato che il campo sarà certamente meno fertile vicino alle scoline. Per questo motivo disegneremo l'esperimento a blocchi randomizzati, dividendo prima l'appezzamento in quattro blocchi perpendicolari al gradiente di fertilità. Ad esempio il blocco 1 conterrà le parcelle 1, 9, 17, 25 2, 10, 18 e 26, cioè le prime due colonne della mappa, con un numero di parcellè esattamente uguali al numero delle tesi. Il blocco 2 conterrà le colonne 3 e 4 e così via. Dato che il gradiente è trasversale, le parcelle di un stesso blocco saranno più omogenee che non parcelle su blocchi diversi. Dopo aver diviso la mappa in quattro blocchi di otto parcelle, possiamo allocare gli otto trattamenti a random all'interno di ogni blocco. Con R è possibile utilizzare il codice seguente (notare che la numerazione assegnata da R è diversa dalla nostra, anche se possiamo far riferimento ai valori crescenti all'interno di ogni blocco).

```
reps <- 4
designRCBD <- design.rcbd(trt, r=reps, seed=777, serie=2)
book2 <- designRCBD$book
book2
```

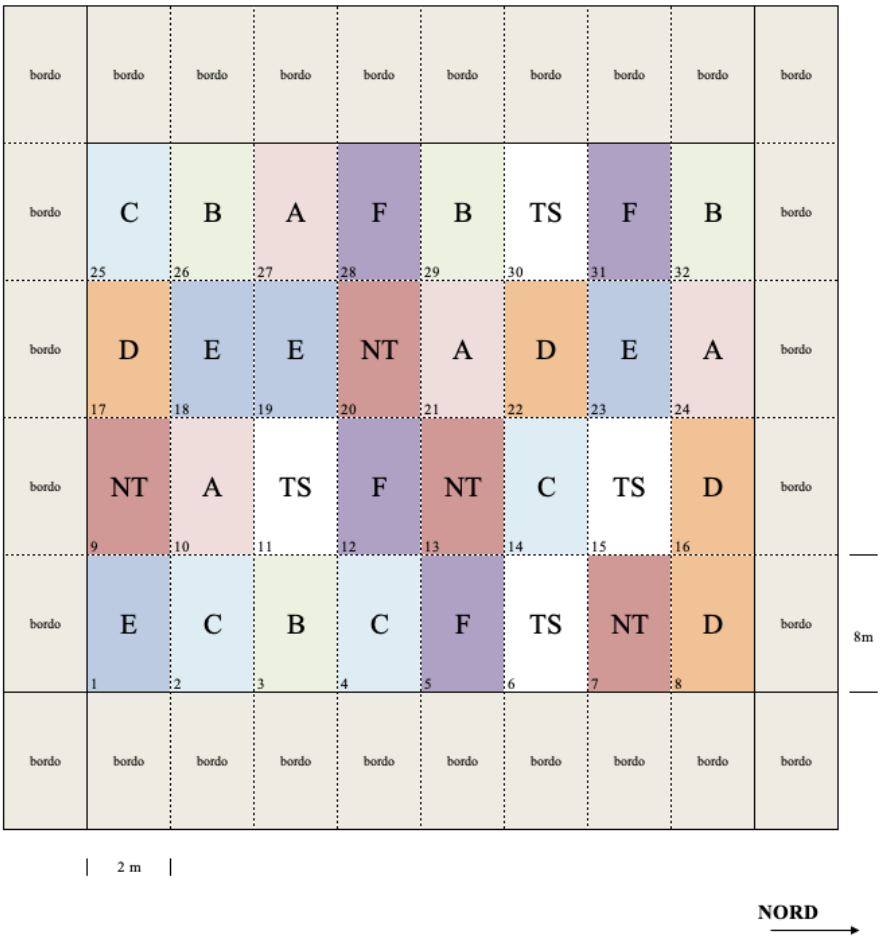


Figura 3.7: Schema sperimentale a randomizzazione completa per l'Esempio 1

```
##    plots block trt
## 1    101     1   E
## 2    102     1   B
## 3    103     1  NT
## 4    104     1   F
## 5    105     1   D
## 6    106     1  TS
## 7    107     1   C
## 8    108     1   A
## 9    201     2   F
## 10   202     2   A
## 11   203     2   C
## 12   204     2   E
## 13   205     2  TS
## 14   206     2   B
## 15   207     2  NT
## 16   208     2   D
## 17   301     3  TS
## 18   302     3  NT
## 19   303     3   F
## 20   304     3   A
## 21   305     3   B
## 22   306     3   E
## 23   307     3   C
## 24   308     3   D
## 25   401     4   D
## 26   402     4  TS
## 27   403     4   A
## 28   404     4   F
## 29   405     4   E
## 30   406     4   C
## 31   407     4   B
## 32   408     4  NT
```

```
zigzag(designRCBD) # zigzag numeration
```

```
##    plots block trt
## 1    101     1   E
## 2    102     1   B
## 3    103     1  NT
## 4    104     1   F
```

```
## 5    105     1   D
## 6    106     1  TS
## 7    107     1   C
## 8    108     1   A
## 9    208     2   F
## 10   207     2   A
## 11   206     2   C
## 12   205     2   E
## 13   204     2  TS
## 14   203     2   B
## 15   202     2  NT
## 16   201     2   D
## 17   301     3  TS
## 18   302     3  NT
## 19   303     3   F
## 20   304     3   A
## 21   305     3   B
## 22   306     3   E
## 23   307     3   C
## 24   308     3   D
## 25   408     4   D
## 26   407     4  TS
## 27   406     4   A
## 28   405     4   F
## 29   404     4   E
## 30   403     4   C
## 31   402     4   B
## 32   401     4  NT
```

```
print(designRCBD$sketch)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,] "E"  "B"  "NT" "F"  "D"  "TS" "C"  "A"
## [2,] "F"  "A"  "C"  "E"  "TS" "B"  "NT" "D"
## [3,] "TS" "NT" "F"  "A"  "B"  "E"  "C"  "D"
## [4,] "D"  "TS" "A"  "F"  "E"  "C"  "B"  "NT"
```

Anche in questo caso, riportiamo tutto sulla mappa.

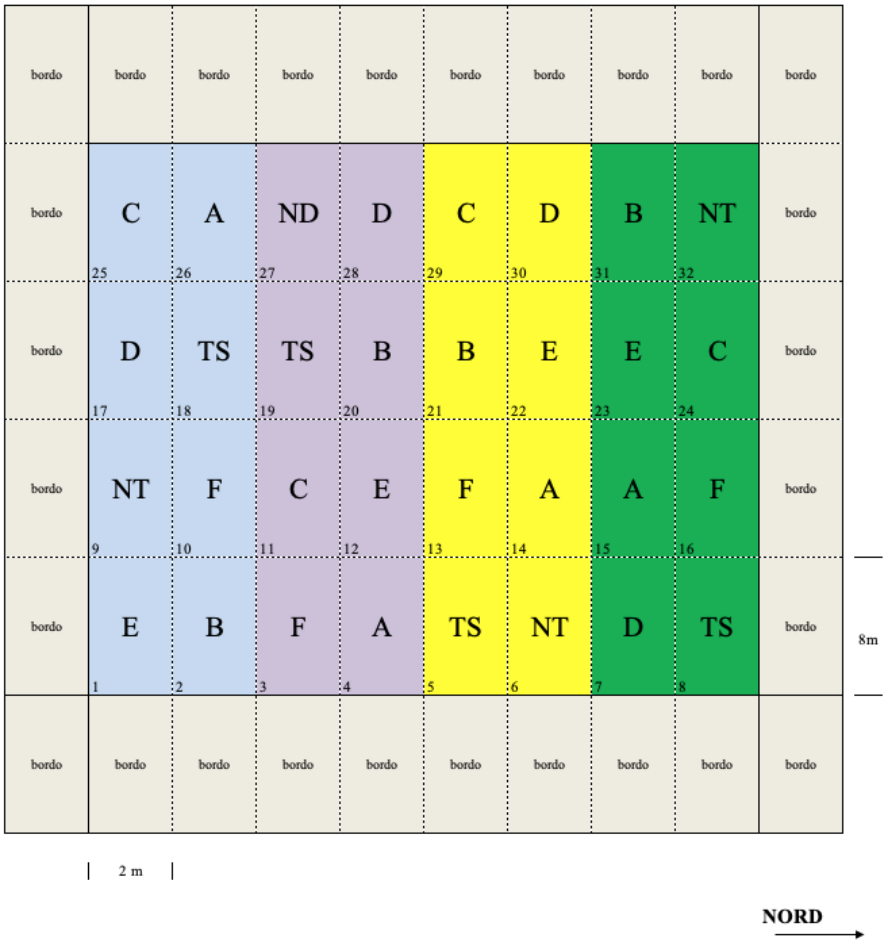


Figura 3.8: Schema sperimentale a blocchi randomizzati per l'Esempio 1

Esempio 2

In questo caso, per ognuno dei tre anni di prova, la mappa contiene una griglia 4 x 4, analoga a quella dell'esperimento precedente, ma più piccola. Nella mappa potremo quindi identificare, esclusi i bordi, quattro colonne e quattro righe. Dato che abbiamo presupposto l'esistenza di un gradiente trasversale e lungitudinale (tra righe e tra colonne), l'allocazione dei trattamenti dovrà esser fatta in modo che ognuno di essi si trovi su ogni riga e ogni colonna. Questo tipo di disegno prende il nome di **Quadrato latino**.

Anche in questo caso potremo chiedere ad R di aiutarci a trovare la combinazione corretta (anche se questo potrebbe essere comodamente fatto a mano).

```
trt <- c("A", "B", "C", "D")
designLS <- design.lsd(trt, seed=543, serie=2)
designLS$book
```

```
##    plots row col trt
## 1    101   1   1   C
## 2    102   1   2   A
## 3    103   1   3   B
## 4    104   1   4   D
## 5    201   2   1   D
## 6    202   2   2   B
## 7    203   2   3   C
## 8    204   2   4   A
## 9    301   3   1   B
## 10   302   3   2   D
## 11   303   3   3   A
## 12   304   3   4   C
## 13   401   4   1   A
## 14   402   4   2   C
## 15   403   4   3   D
## 16   404   4   4   B
```

A questo punto dobbiamo considerare che questa prova deve essere ripetuta in tre anni. La ripetizione di una prova è sempre fondamentale, in quanto consente di verificare non solo la replicabilità dell'esperimento (che è dimostrata dalle repliche), ma anche la sua riproducibilità (riguardare le definizioni di replicabilità e riproducibilità). In questo caso poi la ripetizione dell'esperi-

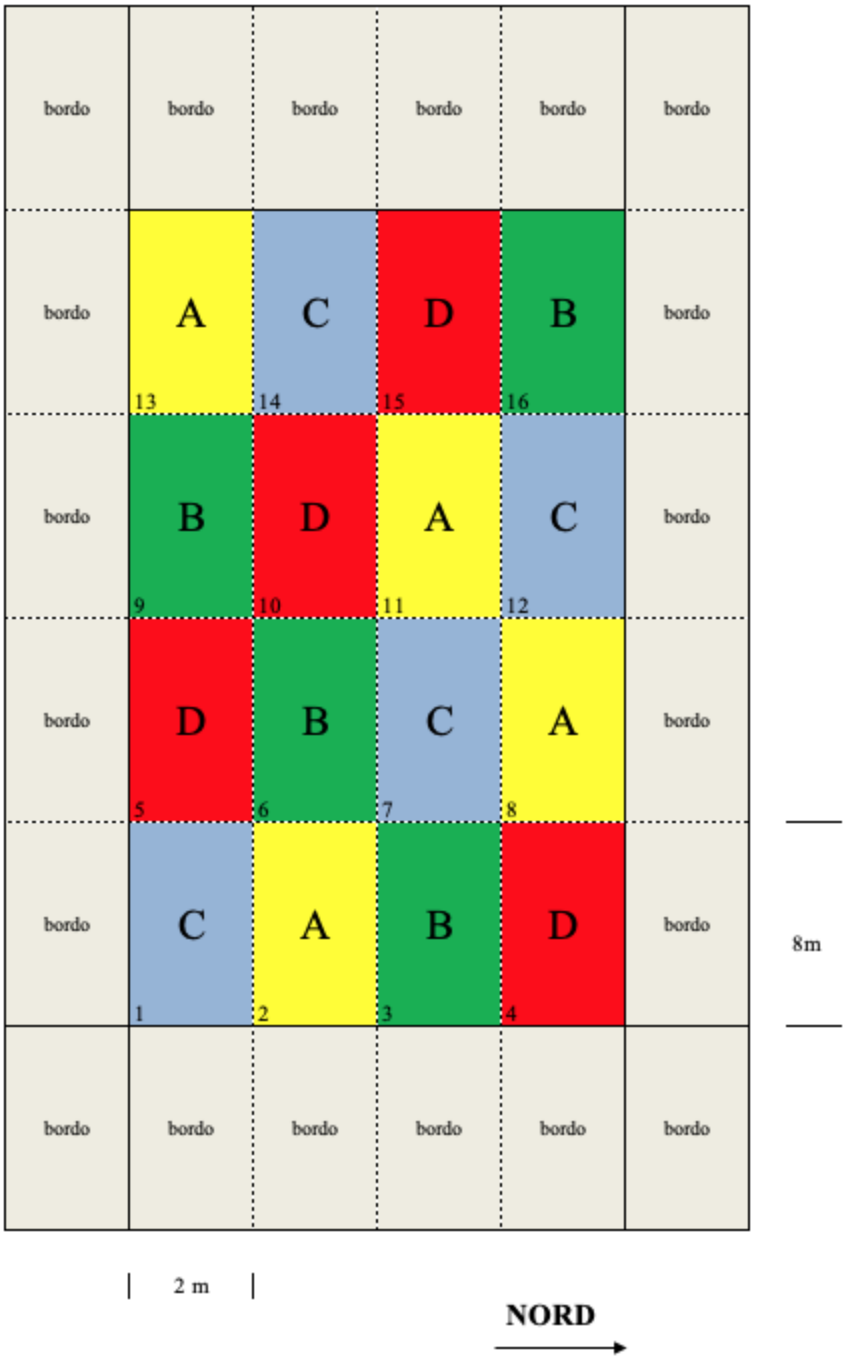


Figura 3.9: Schema sperimentale a quadrato latino per l'Esempio 2 (un anno)

mento è indispensabile per misurare la stabilità produttiva, cioè l'oscillazione delle produzioni da un anno all'altro.

Ovviamente è anche importante verificare la stabilità produttiva da una località all'altra, che consente di valutare l'esistenza di macro-areali, nei quali è possibile consigliare le stesse varietà. Un'aspetto fondamentale è comunque quello di **definire una diversa randomizzazione in ogni anno/località**, per evitare che le stesse varietà siano sempre nelle stesse posizioni, che potrebbe dare origine a dubbi di confounding. La definizione delle randomizzazioni per il secondo e terzo anno è lasciata per esercizio.

Un'altro aspetto da considerare è la metodica impiegata per la determinazione del peso di 1000 semi. Abbiamo già visto che, per aumentare la precisione e la rappresentatività, da tutta la granella raccolta da una parcella preleviamo quattro lotti da 1000 semi, di cui determinare il peso. In questo modo, per ogni trattamento avremo 16 valori (quattro repliche x quattro lotti per replica). Ovviamente non possiamo affermare di avere 16 repliche, in quanto solo le parcelle sono da considerare repliche, in quanto ricevono il trattamento (varietà) in modo indipendente. I quattro lotti raccolti da ogni parcella sono unità osservazionali (perché ne viene rilevato il peso), ma non unità sperimentali, perché appartengono alla stessa parcella e non sono indipendenti. I quattro lotti si dicono **sub-repliche**, quindi il disegno ha quattro repliche e quattro sub-repliche per replica (**disegno a quadrato latino con sottocampionamento**). I due strati di errore (variabilità tra repliche e variabilità tra sub-repliche entro replica), devono essere mantenuti separati in fase di analisi, altrimenti l'analisi è invalida, perché è condotta come se avessimo un più alto grado di precisione (16 repliche) rispetto a quello che abbiamo effettivamente (una sorta di millantato credito!).

Esempio 3

In questo caso abbiamo un disegno fattoriale con due livelli a blocchi randomizzati. Nel principio, questo disegno non ha nulla di diverso da quello relativo all'esempio 1, fatto salvo un minor numero di trattamenti (solo 6). Anche in questo caso, ci facciamo aiutare da R.

```
trt <- c(3,2) # factorial 3x2
design2way <-design.ab(trt, r=4, serie=2, design="rcbd", seed=777)
book <- design2way$book
levels(book$A) <- c("PROF", "SUP", "MIN")
```

```
levels(book$B) <- c("TOT", "PARZ")
book
```

```
##    plots block    A    B
## 1    101     1  SUP PARZ
## 2    102     1 PROF PARZ
## 3    103     1 PROF  TOT
## 4    104     1  MIN  TOT
## 5    105     1  SUP  TOT
## 6    106     1  MIN PARZ
## 7    107     2  MIN  TOT
## 8    108     2  SUP  TOT
## 9    109     2  MIN PARZ
## 10   110     2 PROF  TOT
## 11   111     2  SUP PARZ
## 12   112     2 PROF PARZ
## 13   113     3  MIN  TOT
## 14   114     3  SUP  TOT
## 15   115     3 PROF PARZ
## 16   116     3  MIN PARZ
## 17   117     3  SUP PARZ
## 18   118     3 PROF  TOT
## 19   119     4  MIN PARZ
## 20   120     4 PROF  TOT
## 21   121     4 PROF PARZ
## 22   122     4  MIN  TOT
## 23   123     4  SUP  TOT
## 24   124     4  SUP PARZ
```

La mappa risultante è visibile più sotto.

Questo disegno è totalmente appropriato, ma ci costringe a lasciare parecchio spazio tra una parcella e l'altra, per poter manovrare con la macchina per la lavorazione del terreno. Sarebbe utile raggruppare le parcelle caratterizzate dalla stessa lavorazione, in modo da poter lavorare su superfici più ampie. Ne guadagnerebbe l'uniformità dell'esperimento e l'accuratezza dei risultati. Possiamo quindi immaginare un disegno a un fattore, con parcelle di dimensione doppia (**main-plots**), sulle quali eseguire, in modo randomizzato le lavorazioni del terreno. Succesivamente, ogni main-plot può essere suddivisa in due e, su ognuna delle due metà, possono essere allocati in modo random i due trattamenti di diserbo. In questo modo ci troviamo ad operare

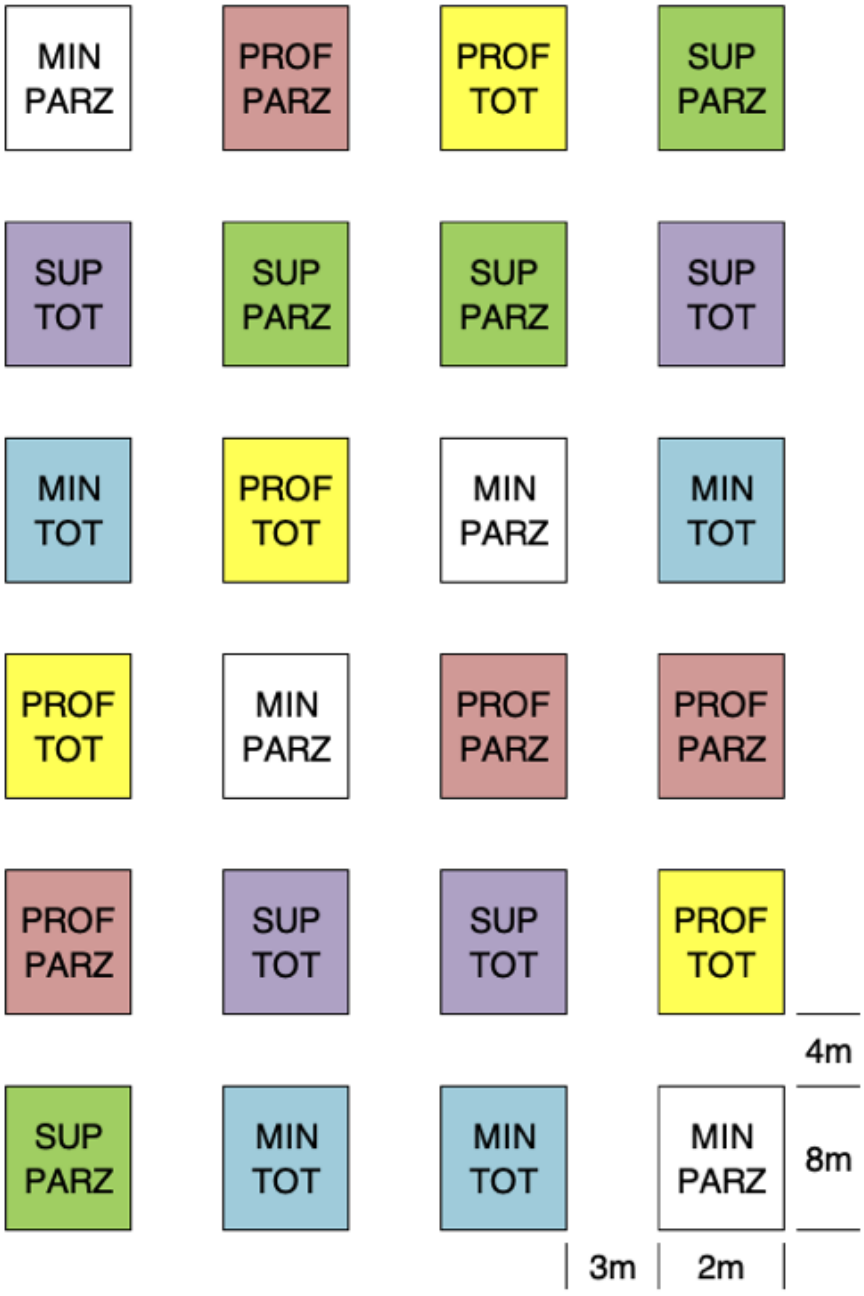


Figura 3.10: Schema sperimentale fattoriale a blocchi randomizzati per l'Esempio 3

con parcelle di due dimensioni diverse: le main-plots per le lavorazioni e le sub-plots per il diserbo. Questo tipo di schema prende il nome di **parcella suddivisa** (**split-plot**), ed è piuttosto comune nella sperimentazione di pieno campo.

Proviamo ad utilizzare R per redigere il piano sperimentale.

```
lavorazione <- c("PROF", "SUP", "MIN")
diserbo <- c("TOT", "PARZ")
designSPLIT <- design.split(lavorazione, diserbo, r=4, serie=2, seed=777)
book <- designSPLIT$book
book
```

```
##    plots splots block lavorazione diserbo
## 1    101      1     1          SUP    PARZ
## 2    101      2     1          SUP     TOT
## 3    102      1     1         PROF     TOT
## 4    102      2     1         PROF    PARZ
## 5    103      1     1          MIN    PARZ
## 6    103      2     1          MIN     TOT
## 7    104      1     2          SUP    PARZ
## 8    104      2     2          SUP     TOT
## 9    105      1     2          MIN     TOT
## 10   105      2     2          MIN    PARZ
## 11   106      1     2         PROF     TOT
## 12   106      2     2         PROF    PARZ
## 13   107      1     3          MIN     TOT
## 14   107      2     3          MIN    PARZ
## 15   108      1     3          SUP     TOT
## 16   108      2     3          SUP    PARZ
## 17   109      1     3         PROF     TOT
## 18   109      2     3         PROF    PARZ
## 19   110      1     4         PROF    PARZ
## 20   110      2     4         PROF     TOT
## 21   111      1     4          MIN     TOT
## 22   111      2     4          MIN    PARZ
## 23   112      1     4          SUP    PARZ
## 24   112      2     4          SUP     TOT
```

In alcune circostanze, soprattutto nelle prove di diserbo chimico, potrebbe trovare applicazione un altro tipo di schema sperimentale, nel quale, in ogni blocco, un trattamento viene applicato a tutte le parcelle di una riga e l'altro

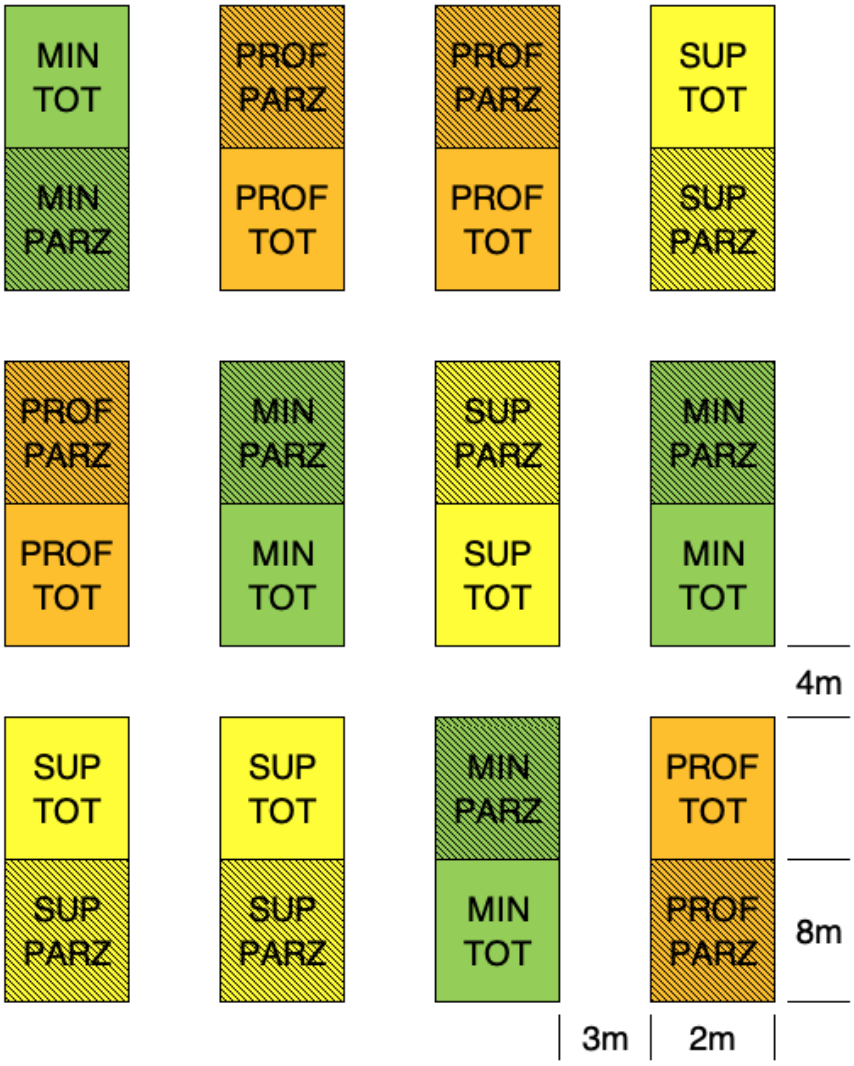


Figura 3.11: Schema sperimentale split-plot a blocchi randomizzati per l'Esempio 3

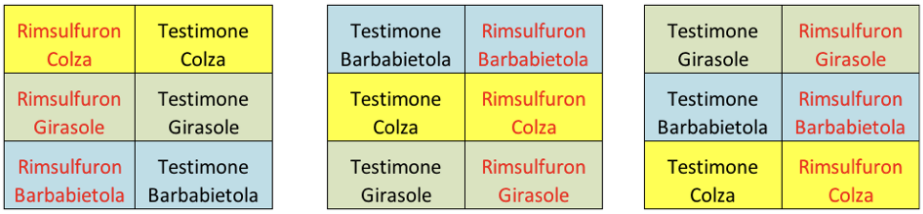


Figura 3.12: Schema sperimentale strip-plot

trattamento a tutte le parcelle di una colonna. Ad esempio, il disegno sottostante mostra una prova nella quale il terreno è stato diserbato in una striscia nel senso della lunghezza e, dopo il diserbo, le colture sono state seminate in striscia, nel senso della larghezza. Questo disegno è detto **strip-plot** ed è molto comodo perché consente di lavorare velocemente.

Esempio 4

La prova di erba medica è fondamentalmente un esperimento a blocchi randomizzati, il cui piano è riportato più sotto. Tuttavia, si tratta di una coltura poiliennale nella quale ripeteremo le misurazioni ogni anno sulle stesse parcelle. le misure ripetute non sono randomizzate (non possono esserlo), ma seguono una metrica temporale. Proprio per questo sviluppo lungo la scala del tempo, i dati che si raccolgono in questi esperimenti a misure ripetute sono detti **dati longitudinali**. Guardando bene il disegno si capisce anche per si parla di **split-plot nel tempo**. Esempi affini sono relativi all'analisi di accrescimento con misure non distruttive (esempio l'altezza) oppure i prelievi di terreno a profondità diverse, anche se, in quest'ultimo caso, la metrica delle misure ripetute è spaziale, non temporale.

Si può notare una certa analogia con il sottocampionamento illsutrato più sopra, nel senso che vengono prese più misure per parcella. Tuttavia, bisogna tener presente che nel sottocampionamento le diverse misure sono solo repliche e non vi è nessuna esigenza di distinguere tra quelle prese nella stessa parcella. Invece, nel caso delle misure ripetute ognuna di esse ha interesse individuale, in quanto espressione di un'anno particolare.

Esempio 5

Per questo disegno osservazionale, la mappa non è necessaria. Tuttavia, si può notare che, in ogni supermercato, abbiamo un disegno a randomizzazione

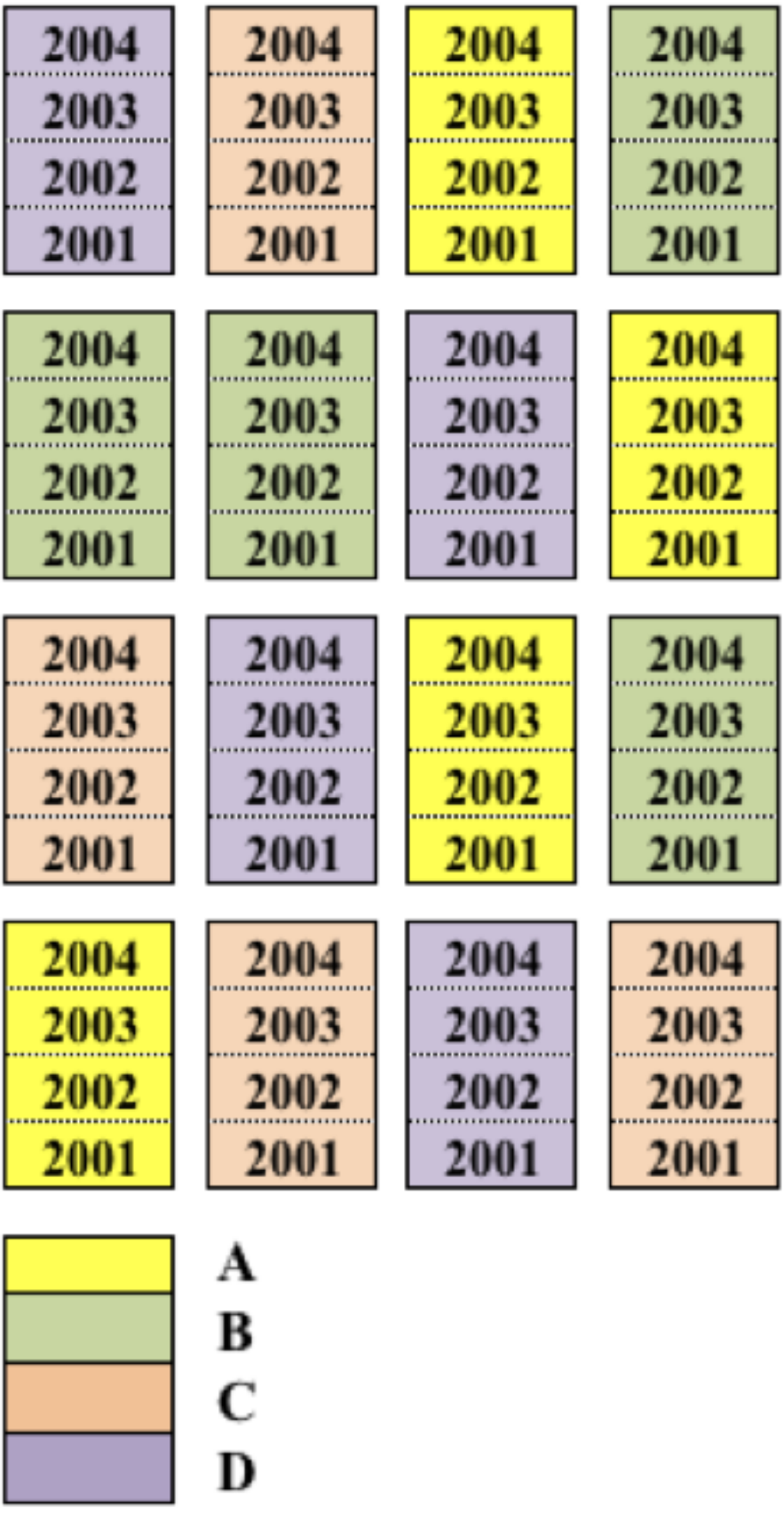


Figura 3.13: Schema sperimentale a blocchi randomizzati con misure ripetute

completa, con tre tipi di confezioni e tre repliche, cioè nove confezioni scelte a random da un lotto più grande. Insomma, si tratta di un esperimento ripetuto 9 volte che, pertanto, ha una certa affinità con l'esperimento ripetuto dell'Esempio 2.

3.3.13 Impianto delle prove

Da questo punto in poi, subentrano le competenze agronomiche e fitopatologiche necessarie per codurre gli esperimenti, Mi piace solo ricordare alcune pratiche usuali nella sperimentazione di pieno campo, destinate a migliorare l'efficienza della prova.

1. Seminare a densità più alte e poi diradare, per assicurare una migliore uniformità di impianto
2. Prelevare da ogni parcella più campioni ed, eventualmente, omogeneizzarli o mediare i risultati ottenuti (vedi il caso dei 1000 semi)
3. Considerare le caratteristiche naturalmente meno variabili (es. la produzione areica e non la produzione per pianta)

Voglio inoltre ricordare che gli esperimenti parcellari configurano una situazione nella quale, per l'elevata cura che si pone nelle tecniche agronomiche, si riesce ad ottenere una produttività almeno del 20% superiore rispetto a quanto avviene nella normale pratica agricola.

3.4 Come scrivere un progetto di ricerca o un report di ricerca

Quanto abbiamo finora esposto costituisce uno schema generale che può essere adottato per redigere un progetto di ricerca o un report sui risultati ottenuti (tesi, pubblicazione). Bisogna provare che la ricerca che si è eseguita è precisa, accurata e replicabile/riproducibile e, di conseguenza, i risultati sono validi.

Nella redazione di un progetto di ricerca o di un report, è fondamentale tratteggiare bene i seguenti elementi:

1. Titolo della ricerca
2. Descrizione del problema e background scientifico
3. Ipotesi scientifica, motivazioni e obiettivi
4. Tipo di esperimento e durata

5. Disegno sperimentale: trattamenti sperimentali (tesi) a confronto con dettagli relativi all'applicazione
6. Unità sperimentali e criteri per la loro selezione. Dettagli su repliche e randomizzazione
7. Dettagli su eventuali tecniche di 'blocking'
8. Variabili da rilevare/rilevate
9. Dettagli su come le variabili saranno/sono state rilevate
10. Esposizione dei risultati (solo report)
11. Discussione (solo report)
12. Conclusioni (solo report)

Alcuni aspetti che divengono elemento di valutazione del progetto e/o del report sono i seguenti:

1. La selezione dei metodi deve essere coerente con gli obiettivi
2. Descrizione dettagliata dei materiali e metodi (bisogna che chiunque sia in grado di replicare l'esperimento)
3. Esposizione dei risultati chiara e convincente
4. Discussione approfondita e con molti riferimenti alla letteratura.

3.5 Per approfondimenti

1. Hurlbert, S., 1984. Pseudoreplication and the design of ecological experiments. Ecological Monographs, 54, 187-211
2. Kuehl, R. O., 2000. Design of experiments: statistical principles of research design and analysis. Duxbury Press (CHAPTER 1)
3. LeClerg, E.; Leonard, W. & Clark, A., 1962. Field Plot Technique. Burgess Publishing Company, (CHAPTER 3)
4. Felipe de Mendiburu (2017). agricolae: Statistical Procedures for Agricultural Research. R package, version 1.2-8. `https://CRAN.R-project.org/package=agricolae`

Capitolo 4

Per iniziare: introduzione ad R

4.1 Cosa è R?

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

E'uno strumento molto potente, anche da un punto di vista grafico, ma necessita di una certa pratica, in quanto manca di una vera e propria interfaccia grafica (Graphical User Interface: GUI) e, di conseguenza, è spesso necessario scrivere codice.

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

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

4.2 Oggetti e assegnazioni

4.3 Costanti e vettori

R lavora con valori, stringhe di caratteri, vettori e matrici, che vengono assegnati alle variabili con opportuni comandi. Ad esempio, il comando:

```
y  <-  3
y
```

```
## [1] 3
```

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

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

```
## [1] 1 2 3
```

crea un vettore x contenente i numeri 1,2 e 3. Bisogna precisare che con il termine 'vettore' in R non ci si riferisce al vettore algebrico, ma più semplicemente ad una serie di numeri (o strighe) consecutivi, rappresentati convenzionalmente da R in una riga.

4.4 Matrici

Oltre ai vettori, in R possiamo definire le matrici. Ad esempio il comando:

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

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

Come già mostrato, per visualizzare il contenuto di una variabile basta digitare il nome della variabile. Ad esempio:

```
z
```

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

Gli elementi di una matrice possono essere richiamati con un opportuno utilizzo delle parentesi quadre:

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

```
## [1] 3
```

4.5 Operazioni ed operatori

Le variabili possono essere create anche con opportune operazioni algebriche, che si eseguono utilizzando i normali operatori (+, -, *, /). Ad esempio:

```
f  <-  2 * y
f
```

```
## [1] 6
```

4.6 Funzioni ed argomenti

Per eseguire operazioni particolari si utilizzano, in genere, le funzioni. Una funzione è richiamata con un nome ed uno o più argomenti. Ad esempio, il comando:

```
log(5)
```

```
## [1] 1.609438
```

Calcola il logaritmo naturale di 5 e richiede un solo argomento, cioè il numero di cui calcolare il logaritmo. Al contrario, il comando:

```
log(100, 2)
```

```
## [1] 6.643856
```

Calcola il logaritmo in base 2 di 100 e richiede due argomenti, cioè il numero di cui calcolare il logaritmo e la base del logaritmo. Quando sono necessari due o più argomenti essi debbono essere messi nell'ordine esatto (in questo caso prima il numero poi la base) oppure debbono essere utilizzati i riferimenti corretti. Ad esempio, i due comandi:

```
log(100, base=2)
```

```
## [1] 6.643856
```

```
log(base=2, 100)
```

```
## [1] 6.643856
```

restituiscono lo stesso risultato, al contrario dei due comandi seguenti:

```
log(100, 2)
```

```
## [1] 6.643856
```

```
log(2, 100)
```

```
## [1] 0.150515
```

4.7 Dataframe

Oltre a vettori e matrici, in R esiste un altro importante oggetto, cioè il *dataframe*, costituito da una tabella di dati con una o più colonne di variabili e una o più righe di dati. A differenza della matrice, il dataframe può essere utilizzato per memorizzare variabili di diverso tipo (numeri e caratteri). Un dataframe può essere creato unendo più vettori, come nell'esempio seguente.

```
parcelle  <-  c(1, 2, 3, 4, 5, 6)
tesi  <-  factor(c("A", "A", "B", "B", "C", "C"))
dati  <-  c(12, 15, 16, 13, 11, 19)
tabella  <-  data.frame("Parc"=parcelle,"Tesi"=tesi,"Produzioni"=dati)
tabella
```

```
##   Parc Tesi Produzioni
## 1    1    A         12
## 2    2    A         15
## 3    3    B         16
## 4    4    B         13
## 5    5    C         11
## 6    6    C         19
```

Per utilizzare i dati in un dataframe, bisognerà accedere ai singoli vettori colonna che lo costituiscono. Per far questo possiamo utilizzare l'estrattore *$*:

```
tabella$Parc
```

```
## [1] 1 2 3 4 5 6
```

oppure possiamo utilizzare gli indici, che nel caso del dataframe, cioè una struttura dati bidimensionale, sono due, uno per le righe e uno per le colonne, separati da virgole:

```
tabella[,1]
```

```
## [1] 1 2 3 4 5 6
```

oppure si può usare il comando *attach()*, che crea immediatamente tre vettori (Pianta, Varietà e Altezza), disponibili per le successive elaborazioni.Possiamo osservare infatti che, dopo aver creato la matrice 'tabella', digitando quanto segue R ci mette a disposizione il vettore 'Produzioni'.

```
attach(tabella)
```

```
## The following objects are masked from tabella (pos = 5):
##
##     Parc, Produzioni, Tesi
```

```
Produzioni
```

```
## [1] 12 15 16 13 11 19
```

I dataframe possono essere editati velocemente utilizzando il comando *fix*, che fa apparire una finestra di editing tipo 'foglio elettronico'.

4.8 Quale oggetto sto utilizzando?

Per avere informazioni sulla natura di un oggetto creato in R, posso usare la funzione `str()`, come nell'esempio seguente:

```
str(tabella)
```

```
## 'data.frame':    6 obs. of  3 variables:
##  $ Parc      : num  1 2 3 4 5 6
##  $ Tesi      : Factor w/ 3 levels "A","B","C": 1 1 2 2 3 3
##  $ Produzioni: num  12 15 16 13 11 19
```

Vediamo infatti che R ci informa che l'oggetto 'tabella' è in realtà un dataframe composto da tre colonne, di cui la prima e la terza sono numeriche, mentre la seconda è una variabile qualitativa (fattore).

4.9 Consigli per l'immissione di dati sperimentali

I dati delle prove sperimentali si possono o importare in R da altri software (ad esempio Excel) oppue si possono digitare direttamente in R. In quest'ultimo caso, in genere, si crea un vettore per ogni colonna di dati e, successivamente, si riuniscono i vettori in un dataframe, che viene poi salvato nel workspace, come vedremo in seguito.

4.9.1 Immissione manuale di dati

L'immissione dei dati in R (e quindi la creazione di vettori) può essere velocizzata utilizzando la funzione `scan()`, separando i dati con INVIO (questo è comodo perchè ci permette di lavorare senza abbandonare il tastierino numerico!). L'immissione termina quando si digita un INVIO a vuoto.

```
dati <- scan()
1: 12
2: 14
3: 16
4: 18
5: 20
6:
Read 5 items
dati
[1] 12 14 16 18 20
```

La stessa funzione può essere anche utilizzata per immettere comodamente stringhe di caratteri, con un opportuno impiego dell'argomento `what`. In questo caso è possibile omettere le virgolette.

```
tesi  <-  scan(what = "character")
1: aurelio
2: aurelio
3: aurelio
```

```
4: claudio
5: claudio
6: claudio
7: latino
8: latino
9: latino
10:
Read 9 items
tesi
[1] "aurelio" "aurelio" "aurelio" "claudio"
    "claudio" "claudio" "latino"  "latino"  "latino"
>
```

4.9.2 Immissione di numeri progressivi

Per creare una serie progressiva, si può utilizzare il comando `seq(n,m,by=step)` che genera una sequenza da n a m con passo pari a *step*.

```
parcelle  <-  seq(1,50,1)
parcelle
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
## [24] 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
## [47] 47 48 49 50
```

4.9.3 Immissione dei codici delle tesi e dei blocchi

A volte i codici delle tesi sono sequenze ripetute di stringhe. Ad esempio, i primi quattro dati potrebbero essere riferiti alla varietà BAIO, i secondi quattro alla varietà DUILIO, i successivi quattro alla varietà PLINIO. Per creare velocemente questo vettore, possiamo utilizzare la funzione `rep()`, in questo modo.

```
tesi  <-  factor(c("BAIO", "DUILIO", "PLINIO"))
tesi
```

```
## [1] BAIO   DUILIO PLINIO
## Levels: BAIO DUILIO PLINIO
```

```
tesi  <-  rep(tesi,each=4)
tesi
```

```
##  [1] BAIO   BAIO   BAIO   BAIO   DUILIO DUILIO DUILIO DUILIO PLINIO PLINIO
## [11] PLINIO PLINIO
## Levels: BAIO DUILIO PLINIO
```

Notare l'uso della funzione `factor()` per creare un vettore di dati qualitativi (fattore). Allo stesso modo, per immettere i codici dei blocchi possiamo utilizzare la stessa funzione in un modo diverso. Ammettiamo infatti che i quattro valori di ogni tesi appartengano rispettivamente ai quattro blocchi; si opera quindi in questo modo.

```
tesi  <-  (c (1, 2, 3, 4))
tesi <- rep(tesi, times=3)
tesi
```

```
##  [1] 1 2 3 4 1 2 3 4 1 2 3 4
```

4.9.4 Leggere e salvare dati esterni

Oltre che immessi da tastiera, i dati possono essere importati in R da files esterni, Inoltre, gli oggetti di R creati nel corso di una sessione possono essere memorizzati su files esterni. Partiamo dal presupposto di aver creato (come frequentemente avviene) il nostro database con EXCEL e di volerlo importare in R nel DATAFRAME *dati*.

Creiamo in EXCEL la tabella riportata di seguito, che si riferisce a 20 piante di mais.

Pianta	Var	Altezza
1	N	172
2	S	154
3	V	150
4	V	188
5	C	162
6	N	145
7	C	157
8	C	178
9	V	175
10	N	158
11	N	153
12	N	191
13	S	174

Pianta	Var	Altezza
14	C	141
15	N	165
16	C	163
17	V	148
18	S	152
19	C	169
20	C	185

La procedura è la seguente:

1. salviamo questa tabella nel file di testo: *comma delineated* 'import.csv'. Per far questo scegliere 'Menù - File - Salva con nome'. Scegliere un nome per il file ed indicare: 'Tipo file = CSV (delimitato dal separatore di elenco) (*.csv). Salvare quindi il file in una directory prescelta.
2. Avviare una sessione R, cambiare la directory predefinita del sistema, scegliendo con il menu File, Change Directory, la cartella nella quale abbiamo memorizzato il file di importazione.
3. Leggere il file di testo in un dataframe, con il seguente comando:

```
setwd("myWorkingDir")
dati  <-  read.csv("import.csv", header=TRUE)
```

Il comando appena descritto ha successo per file CSV creati con la versione inglese di Windows, caratterizzati dal punto come separatore decimale e dalla virgola come separatore di elenco. Se invece il computer fosse settato all'italiana, con la virgola come separatore decimale e il punto e virgola come separatore di elenco, allora si potrebbe utilizzare la funzione `read.csv2()` (stessa sintassi). Con questi due comandi, in R viene creato un dataframe di nome dati, contenente le tre colonne della tabella 'import.csv' appena creata, comprese le intestazioni di colonna.

I dati contenuti in un dataframe o in qualunque altro oggetto possono essere salvati in un file esterno (in formato R binario):

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

ed eventualmente ricaricati:

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

Per scrivere in un file di testo (in questo caso *comma delineated*, ma il separatore di elenco può essere modificato secondo le nostre esigenze con l'argomento `sep`) si utilizza il seguente comando:

```
write.table(dati, "residui.csv", row.names=FALSE,
              col.names=TRUE, sep=",")
```

4.10 Alcune operazioni comuni sul dataset

4.10.1 Selezionare un subset di dati

E' possibile estrarre da un dataframe un subset di dati utilizzando la funzione:

```
subset(dataframe, condizione)
```

Ad esempio, se consideriamo il dataframe tabella creato in precedenza, è possibile selezionare tutte le righe relative alle Tesi A e C come segue:

```
tabella2  <-  subset(tabella, Tesi == "A" | Tesi == "C")
tabella2
```

```
##   Parc Tesi Produzioni
## 1    1    A         12
## 2    2    A         15
## 5    5    C         11
## 6    6    C         19
```

Notare il carattere "|" che esprime la condizione logica OR. La condizione logica AND si esprime con il carattere "&". L'esempio seguente isola i record in cui le varietà sono A o C e, contemporaneamente, la produzione è minore di 19.

```
tabella3  <-  subset(tabella, Tesi == "A" | Tesi == "C" &
                       Produzioni < 19)
tabella3
```

```
##   Parc Tesi Produzioni
## 1    1    A         12
## 2    2    A         15
## 5    5    C         11
```

4.10.2 Ordinare un vettore o un dataframe

Un vettore (numerico o carattere) può essere ordinato con il comando `sort`:

```
y  <-  c(12, 15, 11, 17, 12, 8, 7, 15)
sort(y, decreasing = FALSE)
```

```
## [1]  7  8 11 12 12 15 15 17
```

```
z  <-  c("A", "C", "D", "B", "F", "L", "M", "E")
sort(z, decreasing = TRUE)
```

```
## [1] "M" "L" "F" "E" "D" "C" "B" "A"
```

Un dataframe può essere invece ordinato con il comando `order()`, facendo attenzione al segno meno utilizzabile per l'ordinamento decrescente.

```
dataset[order(dataset$z, dataset$y), ]
dataset[order(dataset$z, -dataset$y), ]
```

4.11 Workspace

Gli oggetti creati durante una sessione di lavoro vengono memorizzati nel cosiddetto workspace. Per il salvataggio del workspace nella directory corrente si usa il menu (File/Save Workspace) oppure il comando:

```
save.image('nomefile.RData')
```

Il contenuto del workspace viene visualizzato con:

```
ls()
```

Il workspace viene richiamato da menu (File/Open Workspace) oppure con il comando:

```
load('nomefile.RData')
```

Per un lavoro efficiente in R è bene tenere il workspace molto pulito, eliminando gli oggetti non necessari. La completa eliminazione degli oggetti nel workspace si esegue con:

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

Uno o più oggetti specifici possono essere eliminati con:

```
rm(oggetto1, oggetto2, .....)
```

Gli oggetti possono anche essere richiamati in base alla loro posizione; ad esempio il comando:

```
rm(list=ls()[3:4])
```

elimina il terzo e il quarto oggetto dal workspace.

Un comando particolarmente utile è il seguente:

```
rm(list=ls()[ls()!="oggetto1"])
```

che permette di eliminare dal workspace ogni oggetto meno "oggetto1". Si possono utilizzare anche clausole logiche più articolate come la seguente:

```
rm(list=ls()[ls()!="oggetto1" & ls()!="oggetto2"])
```

che elimina tutto meno "oggetto1" e "oggetto2".

4.12 Script o programmi

Come è possibile memorizzare dati e workspace, è anche possibile creare uno script (procedura, funzione...) da memorizzare e richiamare in seguito. Nel caso più semplice è possibile scrivere comandi in un semplice editor di testo e salvarli in un file con estensione '.r'. I comandi possono poi essere riutilizzati per semplice copia ed incolla sulla console, opppure, nel caso in cui si utilizzi Rstudio (FILE -> APRI SCRIPT o NUOVO SCRIPT) selezionando il comando (o i comandi) da inviare alla console e premendo la combinazione CTRL + INVIO.

Lavorare con scripts è molto comodo e consigliabile perchè non si deve partire da zero ad ogni sessione, ma è sufficiente correggere i comandi digitati in sessioni precedenti.

Oltre agli script, è possibile creare funzioni personalizzate fino ad arrivare a veri e propri programmi (packages). Immaginiamo ad esempio di voler scrivere una funzione che, dato il valore della produzione rilevata in una parcella di orzo di 20 $ m^2 $ (in kg) e la sua umidità percentuale, calcoli automaticamente il valore della produzione secca in kg/ha. La funzione che dobbiamo implementare è:

$$PS = PU \cdot \frac{100 - U}{100} \cdot \frac{10000}{20}$$

ove PS è la produzione secca in kg/ha e PU è la produzione all'umidità U in kg per 20 $ m^2 $.

Scriveremo un file di testo (ad esempio con il *Block notes* o con l'editor interno ad R):

```
PS  <-  function(PU, U) {
  PU*((100-U)/100)*(10000/20)
}
```

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

Aprendo una nuova sessione in R, possiamo ricaricare in memoria il file di programma (FILE - SORGENTE CODICE R, oppure da console, con il comando:

```
source('prova.r')
```

A differenza di quanto avviene con uno script, i comandi memorizzati nella funzione non vengono eseguiti, ma la funzione 'PS' diviene disponibile nel workspace e può essere utilizzata nel modo seguente:

```
PS(20,85)
```

4.13 Interrogazione di oggetti

A differenza di altri linguaggi statistici come SAS o SPSS, R immagazzina i risultati delle analisi negli oggetti, mostrando un output video piuttosto minimale. Per ottenere informazioni è necessario interrogare opportunamente gli oggetti che al loro interno possono contenere altri oggetti da cui recuperare le informazioni interessanti. Gli oggetti che contengono altri oggetti sono detti **liste**.

Ad esempio, se vogliamo calcolare autovettori ed autovalori di una matrice, utilizziamo la funzione 'eigen'. Questa funzione restituisce una lista di oggetti, che al suo interno contiene i due oggetti values (autovalori) e vectors (autovettori). Per recuperare l'uno o l'altro dei due risultati (autovettori o autovalori) si usa l'operatore di concatenamento (detto anche estrattore) $.

```
matrice  <-  matrix(c(2,1,3,4),2,2)
matrice
```

```
##      [,1] [,2]
## [1,]    2    3
## [2,]    1    4
```

```
ev  <-  eigen(matrice)
ev
```

```
## eigen() decomposition
## $values
## [1] 5 1
##
## $vectors
##            [,1]       [,2]
## [1,] -0.7071068 -0.9486833
## [2,] -0.7071068  0.3162278
```

```
ev$values
```

```
## [1] 5 1
```

```
ev$vectors
```

```
##            [,1]       [,2]
## [1,] -0.7071068 -0.9486833
## [2,] -0.7071068  0.3162278
```

4.14 Altre funzioni matriciali

Oltre che autovettori ed autovalori di una matrice, R ci permette di gestire altre funzioni di matrice. Se ad esempio abbiamo le matrici:

$$Z = \begin{pmatrix} 1 & 22 & 3 \end{pmatrix} \quad Y = \begin{pmatrix} 3 & 2 \end{pmatrix}$$

queste possono essere caricate in R con i seguenti comandi:

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

Possiamo poi ottenere la trasposta di Z con il comando:

```
t(Z)
```

```
##      [,1] [,2]
```

```
## [1,]    1    2
## [2,]    2    3
```

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

```
Y%*%Z
```

```
##      [,1] [,2]
## [1,]    7   12
```

Possiamo calcolare l'inversa di Z con:

```
solve(Z)
```

```
##      [,1] [,2]
## [1,]   -3    2
## [2,]    2   -1
```

4.15 Cenni sulle funzionalità grafiche in R

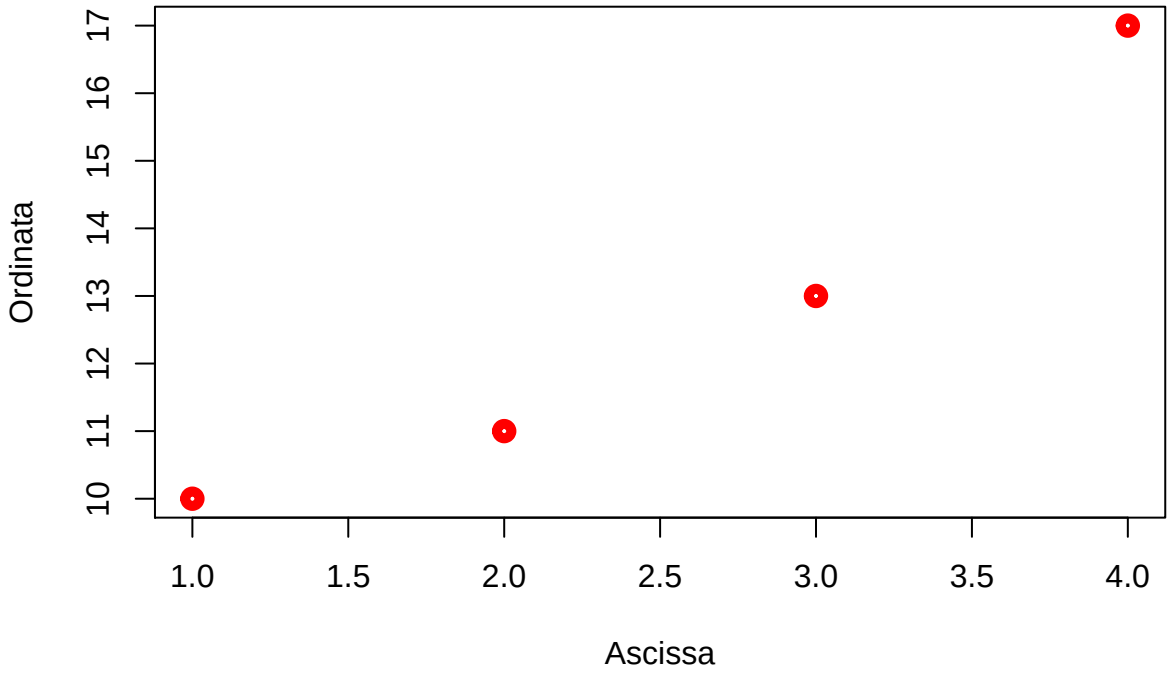
R è un linguaggio abbastanza potente e permette di creare grafici interessanti. Ovviamente un trattazione esauriente esula dagli scopi di questo testo, anche se è opportuno dare alcune indicazioni che potrebbero essere utili in seguito. La funzione più utilizzata per produrre grafici è:

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

ovex ed y sono i vettori con le coordinate dei punti da disegnare. `Type` rappresenta il tipo di grafico (''p''produce un grafico a punti,''l''un grafico a linee,''b''disegna punti uniti da linee,''h'' disegna istogrammi), 'Title disegna il titolo del grafico, sub il sottotitolo,xlab e ylab le etichette degli assi, col è il colore dell'oggetto, lwd il suo spessore, lty il tipo di linea e cos'i via.

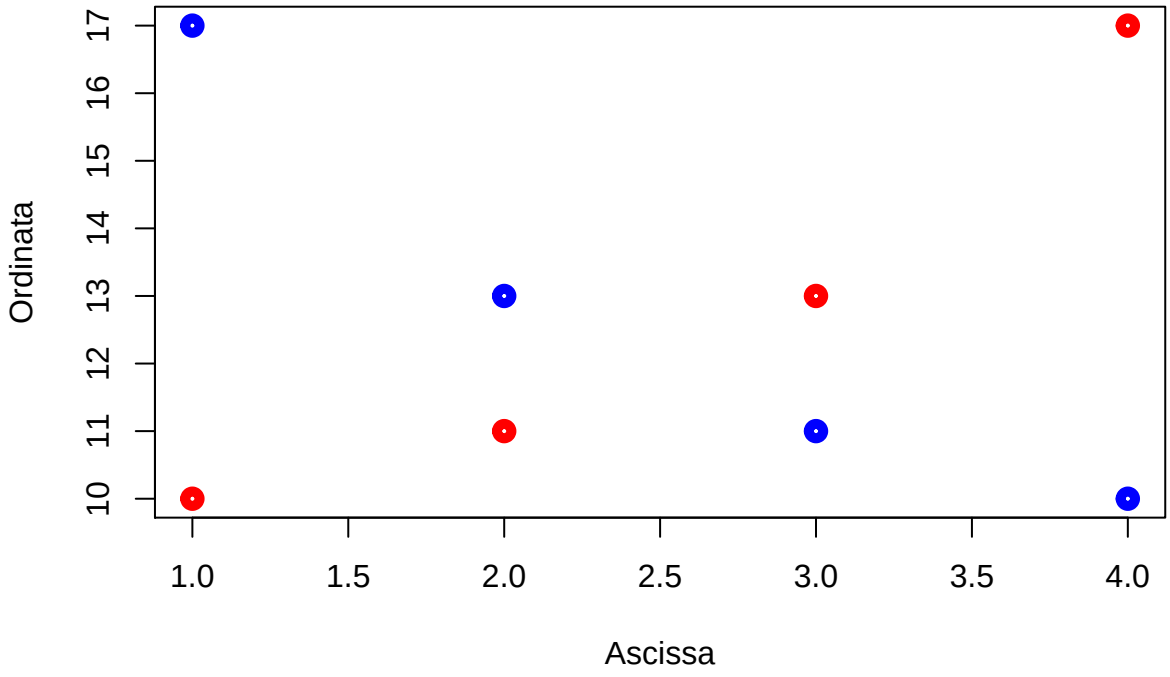
Per una descrizione più dettagliata si consiglia di consultare la documentazione on line. A titolo di esempio mostriamo l'output dei comandi:

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



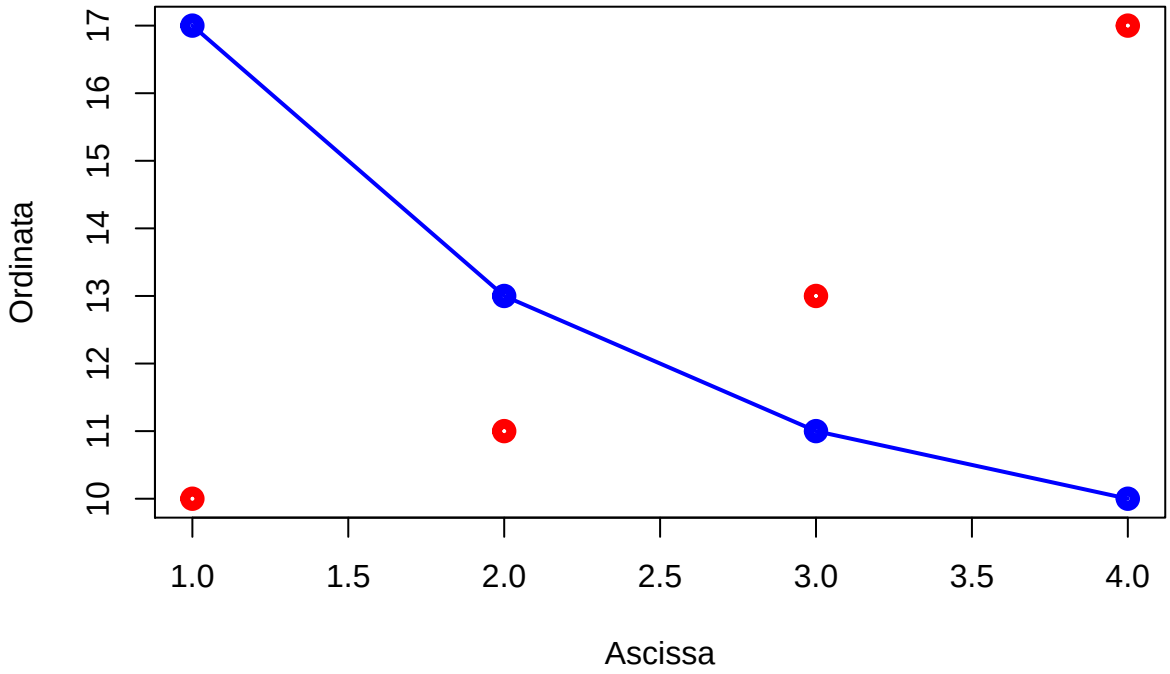
Per sovrapporre un'altra serie di punti alla precedente possiamo utilizzare il comando 'points':

```
y2  <-  c(17,13,11,10)
plot(x, y, "p", col="red", lwd=5,xlab="Ascissa", ylab="Ordinata")
points(x, y2, col="blue", lwd=5)
```



Se avessimo voluto sovrapporre una grafico a linee avremmo invece utilizzato la funzione 'lines':

```
plot(x, y, "p", col="red", lwd=5,xlab="Ascissa", ylab="Ordinata")
points(x, y2, col="blue", lwd=5)
lines(x, y2, col="blue", lwd=2)
```



Per disegnare una curva si può utilizzare la funzione:

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

dove l'argomento 'add' serve per specificare se la funzione deve essere aggiunta ad un grafico preesistente.

Per aggiungere un titolo ad un grafico possiamo utilizzare la funzione:

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

mentre per aggiungere una legenda utilizziamo la funzione:

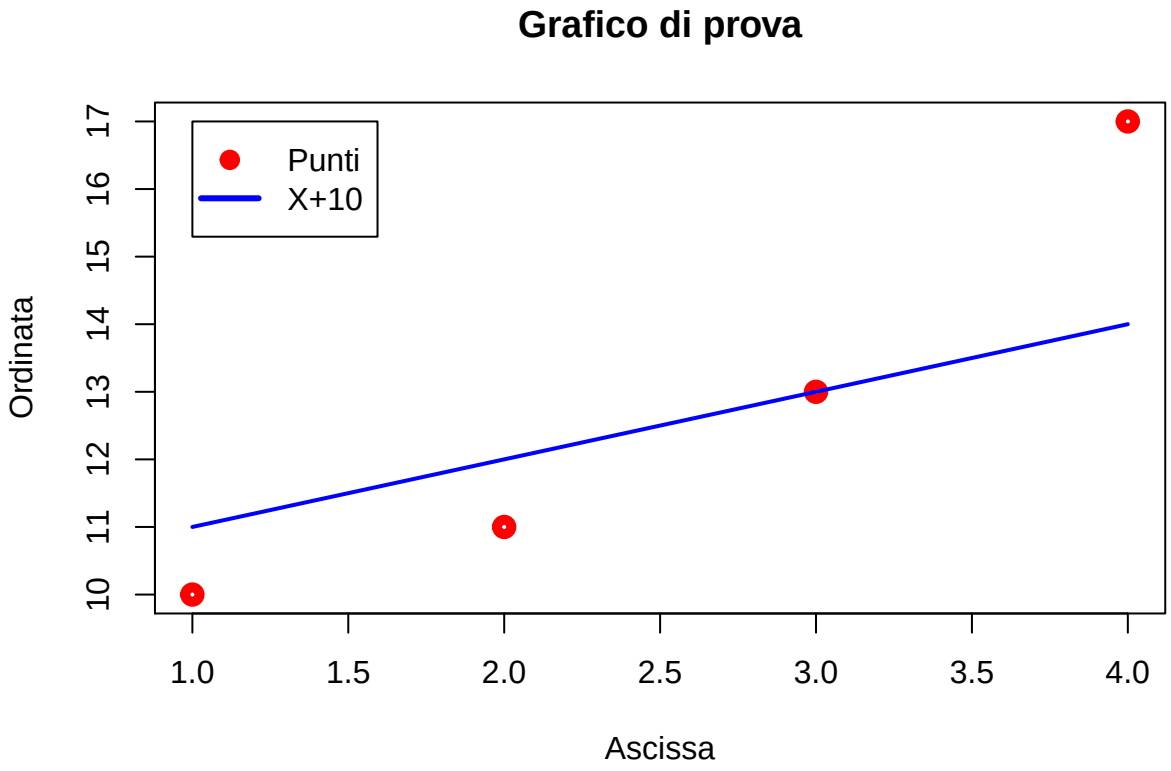
```
legend(Xcoord, YCoord , legend=c("Punti","X+10"), pch=c(19,-1),
       col=c("Red","Blue"),
       lwd=c(3,3), lty=c(0,3))
```

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

Ad esempio:

```
plot(x, y, "p", col="red", lwd=5, xlab="Ascissa",
       ylab = "Ordinata")
curve(10+x, add=TRUE, lty=1, lwd=2, col="blue")
```

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

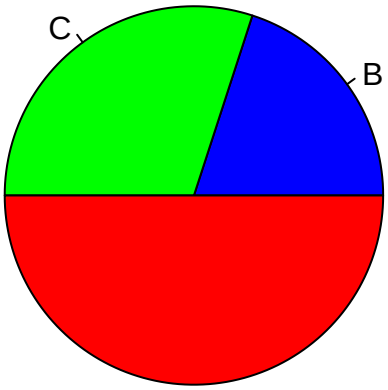


L'ultima cosa che desideriamo menzionare è la possibilità di disegnare grafici a torta, utilizzando il comando:

```
pie(vettoreNumeri, vettoreEtichette, vettoreColori)
```

Ad esempio il comando:

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



4.16 Per approfondimenti

Per approfondimenti si consiglia la consultazione di:

- Maindonald J. Using R for Data Analysis and Graphics - Introduction, Examples and Commentary. (PDF, data sets and scripts are available at JM's homepage.

Per conoscere più a fondo l'ambiente di svilupp RStudio, consiglio la lettura di:

- Oscar Torres Reina, 2013. Introductio to RStudio (v. 1.3). This homepage

Capitolo 5

Primo passo: la descrizione dei dati raccolti

5.1 Le variabili quantitative: analisi chimiche e altre misurazioni fondamentali

In un precedente capitolo abbiamo visto che la replicazione è uno degli elementi fondamentali di un sperimento. Abbiamo anche visto che le n repliche effettuate sono solo un campione delle infinite misure che avremmo potuto fare, ma che non abbiamo fatto per insufficienza di risorse (usualmente tempo, spazio e lavoro). Di conseguenza, dopo aver terminato un esperimento ci troviamo con un collettivo di misure, che debbono essere riassunte e descritte.

Se i dati sono quantitativi (su scala continua o discreta), è possibile e necessario descrive almeno due caratteristiche del dataset, vale a dire:

1. tendenza centrale (location)
2. dispersione (shape)

Vediamo ora quali sono le statistiche più utilizzate per descrivere un campione.

5.1.1 Indicatori di tendenza centrale

La media aritmetica è un concetto molto intuitivo che non necessita di particolari spiegazioni: si calcola con R mediante la funzione `mean(vettore)`. In Excel, si calcola con la funzione "=MEDIA(intervallo)".

Per esempio, carichiamo il dataset 'heights' contenuto nel package 'aomisc' e calcoliamo la media delle altezze.

```
library(aomisc)
data(heights)
mean(heights$height)
```

```
## [1] 164
```

Un altro indicatore di tendenza centrale è la *mediana*, data dal valore che bipartisce la distribuzione di frequenza in modo da lasciare lo stesso numero di termini a sinistra e a destra. Se abbiamo una serie di individui ordinati in graduatoria, la mediana è data dall' individuo che occupa il posto (n + 1)/2 o, se gli individui sono in numero pari, dalla media delle due osservazioni centrali. Il comando per calcolare la mediana in R è `median(vettore)`. In Excel, si utilizza la funzione "=MEDIANA(intervallo)".

La mediana è un indicatore più robusto della media: infatti, supponiamo di avere cinque valori:

1 - 4 - 7 - 9 - 10

La media è pari a 6.2, mentre la mediana è pari a 7 (valore centrale). Se cambiano il numero più alto in questo modo:

1 - 4 - 7 - 9 - 100

la media di questi cinque valori sarà 24.2, mentre la mediana sarà sempre pari a 7. Insomma, la mediana non è influenzata da valori estremi (*outliers*), in senso positivo o negativo.

```
median(heights$height)
```

```
## [1] 162.5
```

5.1.2 Indicatori di variabilità

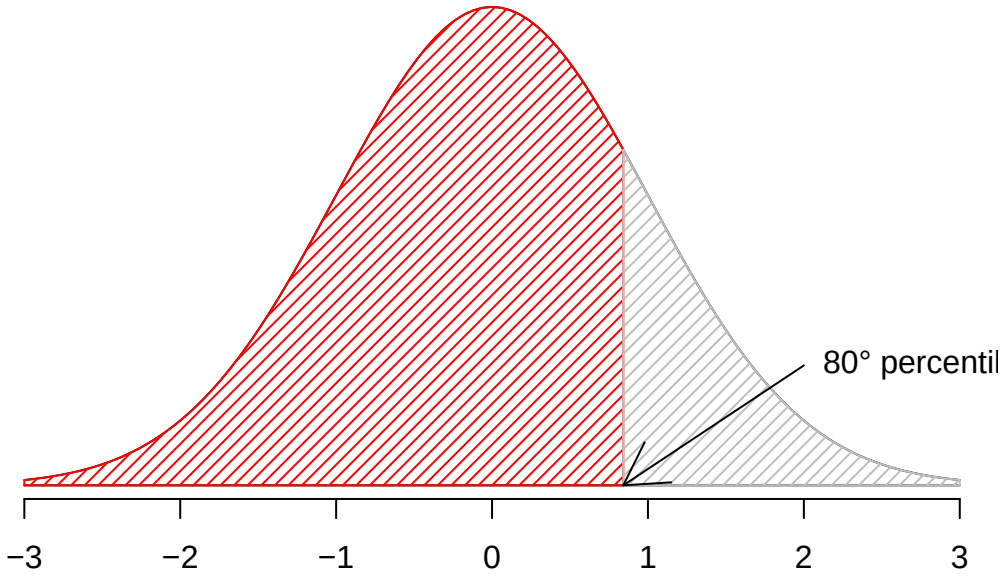
Gli indicatori di tendenza centrale, da soli, non ci informano su come le unità sperimentali tendono a differire l'una dall'altra: ad esempio una media

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.

Pertanto, i risultati di un processo di misurazione non possono essere descritti solo con la media, ma è necessario anche calcolare un indice di variabilità. Tra essi, il più semplice 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.

Invece del campo di variazione, possiamo utilizzare i cosiddetti *percentili*, che bipartiscono la popolazione di partenza in modo da lasciare una certa quantità di termini alla sua sinistra e la restante quantità alla sua destra. 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%.

```
percentile <- 0.8
curve(dnorm(x),from=-3,to=3,axes=FALSE, ylab="", xlab="")
axis(1)
lines(c(-3,3),c(0,0))
valori.rosso<-seq(-3,qnorm(percentile),length=100)
x.rosso<-c(-3,valori.rosso,qnorm(percentile),-3)
y.rosso<-c(0,dnorm(valori.rosso),0,0)
polygon(x.rosso,y.rosso,density=20,angle=45, col="red")
valori.grigio<-seq(qnorm(percentile),3,length=100)
x.grigio<-c(qnorm(percentile),valori.grigio,3,qnorm(percentile))
y.grigio<-c(0,dnorm(valori.grigio),0,0)
polygon(x.grigio,y.grigio,density=20,angle=45, col="grey")
text(x=2,y=0.1,"80° percentile", pos=4)
arrows(2,0.1,qnorm(percentile),0)
```



Per descrivere la variabilità di un collettivo è possibile utilizzare, ad esempio, il 25esimo e il 75esimo percentile: se questi sono molto vicini, significa che il 50 % dei soggetti è compreso in un intervallo piccolo e quindi la variabilità della popolazione è bassa. Per calcolare questi due valori con Excel possiamo utilizzare le funzioni "=PERCENTILE.INC(intervallo, 0.25)" e "=PERCENTILE.INC(intervallo, 0.75)". Per quanto riguarda R, i comandi sono dati più sotto.

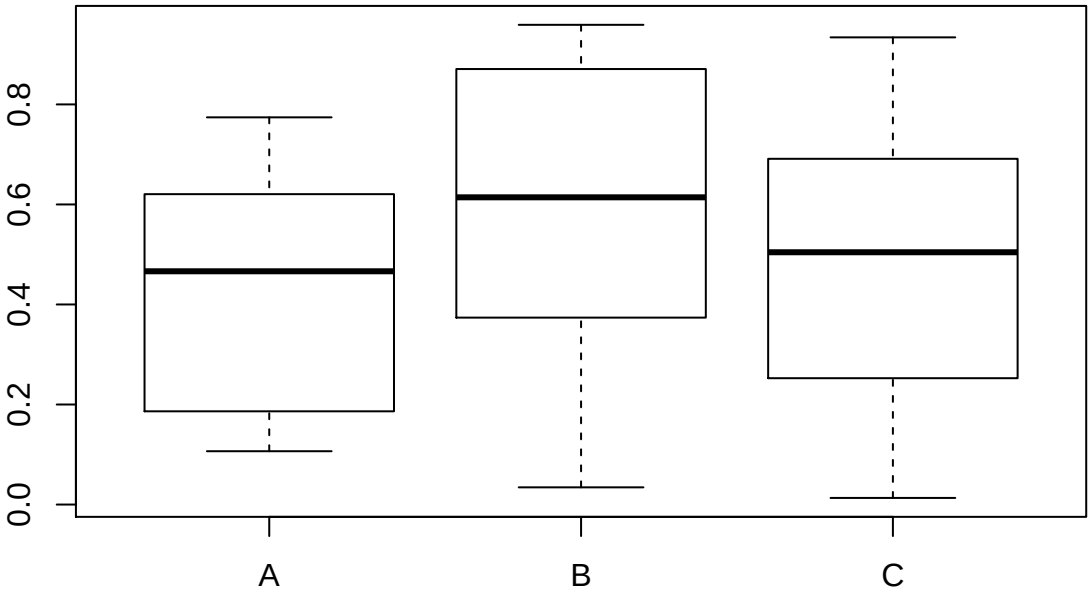
```
quantile(heights$height, probs = c(0.25, 0.75))
```

```
##     25%     75%
## 152.75 174.25
```

A questo proposito, possiamo introdurre il concetto di *boxplot* (grafico Box-Whisker). Si tratta di una scatola che ha per estremi il 25esimo e il 75esimo percentile ed è tagliata da una linea centrale in corrispondenza della mediana. Dalla scatola partono due linee verticali che identificano il valore massimo e il minimo. Se il massimo (o il minimo) distano dalla mediana più di 1.5 volte la differenza tra la mediana stessa e il 75esimo (o 25esimo) percentile, allora le linee verticali si fermano ad un valore pari ad 1.5 volte il 75esimo (o il 25esimo) percentile rispettivamente ed i dati esterni vengono raffigurati come outliers. I boxplot sono solitamente usati per descrivere campioni numerosi nei quali esista un qualche criterio di raggruppamento. In basso abbiamo create tre gruppi con una funzione di estrazione di numeri casuali.

```
set.seed(1234)
A <- runif(20)
```

```r
B <- runif(20)
C <- runif(20)
series <- rep(c("A", "B", "C"), each = 20)
values <- c(A, B, C)
boxplot(values ~ series)
```



Con Excel, non esiste una funzione in grado di disegnare un boxplot automaticamente e si possono seguire le indicazioni a questo link.

Oltre ad esprimere la variabilità di una popolazione con un intervallo (campo di variazione o coppia di percentili) è possibile utilizzare diversi indici sintetici di variabilità, tra cui i più diffusi sono la devianza, la varianza, la deviazione standard ed il coefficiente di variabilità.

La **devianza** (generalmente nota come SS, cioè somma dei quadrati) è data da:

$$SS = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Si tratta di un indicatore caratterizzato da significato geometrico molto preciso, collegabile alla somma dei quadrati delle distanze euclidee di ogni osservazione rispetto alla media. In R, non vi è una funzione per il calcolo della devianza (o meglio, esiste una possibilità nell'ambito dei modelli linea-

ri, ma è troppo presto per introdurla...). Possiamo allora un'espressione del tipo:

```
sum( (heights$height - mean(heights$height))^2 )
```

```
## [1] 4050
```

Come misura di 'distanza', la devianza ha alcune importanti proprietà (che vedremo meglio in seguito), ma essendo una somma, il valore finale dipende dal numero di scarti da sommare e quindi non è possibile operare confronti tra collettivi formati da un diverso numero di individui. Si può quindi definire un altro indice, detto *varianza* (nei software di uso più corrente si parla di varianza campionaria, e definito come segue:

$$\sigma^2 = \frac{SS}{n-1}$$

La varianza permette di confrontare la variabilità di collettivi formati da un numero diverso di individui, anche se permane il problema che questo indicatore è espresso in un'unità di misura al quadrato, rispetto a quella delle osservazioni originali: ad esempio se le osservazioni sono espresse in metri, la varianza è espressa in metri quadrati.

Per eliminare questo problema si ricorre alla radice quadrata della varianza, cioè la *deviazione standard*, che si indica con s. La deviazione standard è espressa nella stessa unità di misura dei dati originari ed è quindi molto informativa sulla banda di oscillazione dei dati rispetto alla media.

Spesso la variabilità dei dati è in qualche modo proporzionale alla media: collettivi con una media alta hanno anche una variabilità alta e viceversa. Per questo motivo viene utilizzato spesso il *coefficiente di variabilità*:

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

che è un numero puro e non dipende dall'unità di misura e dall'ampiezza del collettivo, sicché è molto adatto ad esprimere ad esempio l'errore degli strumenti di misura e delle apparecchiature di analisi.

Varianza e deviazione standard sono molto facili da calcolare in R, grazie alle funzioni `var()`, `sd()`. In Excel, abbiamo le funzioni "=DEV.Q(intervallo)", "=VAR(intervallo)" e "=DEV.ST(intervallo)", rispettivamentep er devianza, varianza e deviazione standard.

In genere, la deviazione standard, per le sue caratteristiche, viene utilizzata come indicatore dell'incertezza assoluta associata ad una determinata misurazione, mentre il coefficiente di variabilità (incertezza relativa percentuale; CV), è molto adatto ad esprimere l'errore degli strumenti di misura e delle apparecchiature di analisi.

```
var(heights$height)
```

```
## [1] 213.1579
```

```
sd(heights$height)
```

```
## [1] 14.59993
```

```
sd(heights$height)/mean(heights$height) * 100
```

```
## [1] 8.902395
```

5.1.3 Arrotondamenti

Il calcolo della media e della deviazione standard (sia a mano che con il computer) porta all'ottenimento di un numero elevato di cifre decimali. E' quindi lecito chiedersi quante cifre riportare nel riferire i risultati della misura. L'indicazione generale, da prendere con le dovute cautele è che nel caso della media si riportano un numero di cifre decimali pari a quello rilevato nella misura, mentre per gli indicatori di variabilità si dovrebbe utilizzare un decimale in più.

5.2 Descrizione dei sottogruppi

In biometria è molto comune che il gruppo di unità sperimentali sia divisibile in più sottogruppi, dei quali vogliamo conoscere alcune statistiche descrittive. Abbiamo già visto il boxplot; ora potremmo voler calcolare le medie per gruppo. Per questo, possiamo utilizzare la funzione 'tapply()':

```
with(heights, tapply(height, var, mean) )
```

```
##      C      N      S      V
## 165.00 164.00 160.00 165.25
```

dove `height` è la variabile che contiene i valori da mediare, `var` è la variabile che contiene la codifica di gruppo, `mean` è la funzione che dobbiamo calco-

lare. Ovviamente `mean` può essere sostituito da qualunque altra funzione ammissibile in R, come ad esempio la deviazione standard.

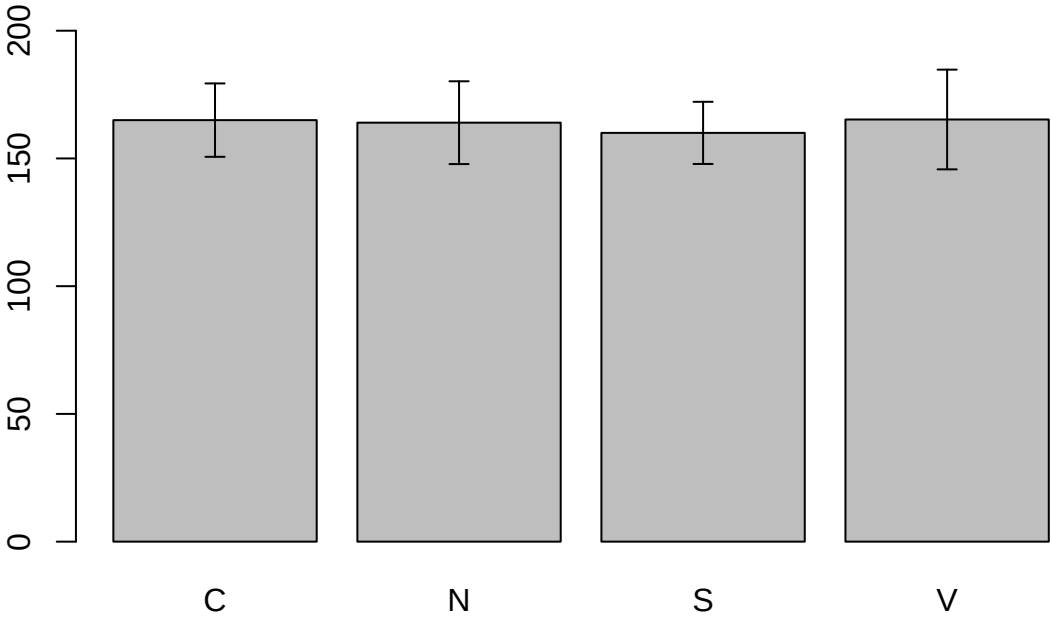
Spesso vogliamo calcolare più di una funzione (ad esempio, la media e la deviazione standard). Per questo possiamo tilizzare il package 'plyr' e la funzione 'ddplyr()'.

```
library(plyr)
descript <- ddply(heights, ~var, summarise,
                  Media = mean(height),
                  SD = sd(height))
descript
```

```
##   var  Media       SD
## 1   C 165.00 14.36431
## 2   N 164.00 16.19877
## 3   S 160.00 12.16553
## 4   V 165.25 19.51709
```

Con la funzione soprastante abbia creato un nuovo dataset (descript), che può essere utilizzato per il plotting, ad esempio per creare un grafico a dispersione, con l'indicazione della dispersione dei dati. Per sapere le coordinate relative al centro di ogni barra, dobbiamo creare un oggetto con la funzione 'barplot()'. Questa funzione, oltre che disegnare il grafico, restituisce appunto le coordinate necessarie.

```
coord <- barplot(descript$Media, names.arg = descript$var,
                 ylim = c(0, 200))
arrows(coord, descript$Media - descript$SD,
       coord, descript$Media + descript$SD,
       length = 0.05, angle = 90, code = 3)
```



Il grafico non è bellissimo; per ora ci accontenteremo, ma può essere migliorate con un po' di esercizio.

In Excel, è molto conveniente utilizzare la funzione "tabella pivot".

5.3 Distribuzioni di frequenza e classamento

Avendo a che fare con variabili qualitative, possiamo considerare la *frequenza assoluta*, cioè il numero degli individui che presentano una certa modalità. 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.

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. Nel caso prima accennato, la frequenza relativa degli eterotteri è pari a 100/500, cioè 0.2.

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

Le distribuzioni di frequenza possono essere costruite anche per le variabili quantitative, tramite un'operazione di classamento, che consiste nel creare classi con intervalli opportuni. Su queste distribuzioni di frequenza possiamo quindi calcolare frequenze assolute, relative e cumulate. In genere, se abbiamo un collettivo molto numeroso è conveniente aggregare i *dati* in forma di distribuzioni di frequenza, perché la lettura delle informazioni è molto più facile. Qui facciamo un esempio, anche se il dataset che utilizzeremo ('heights') non è così numeroso.

Vogliamo:

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

La soluzione al punto 1 con R è facile, attraverso l'impiego della funzione `table()`. La funzione `length()` restituisce il numero di elementi in un vettore.

```
#Frequenze assolute
table(heights$var)
```

```
##
## C N S V
## 7 6 3 4
```

```
#Frequenze relative
with(heights, table(var)/length(var) )
```

```
## var
##    C    N    S    V
## 0.35 0.30 0.15 0.20
```

```
#Frequenze percentuali
with(heights, table(var)/length(var) * 100 )
```

```
## var
##  C  N  S  V
## 35 30 15 20
```

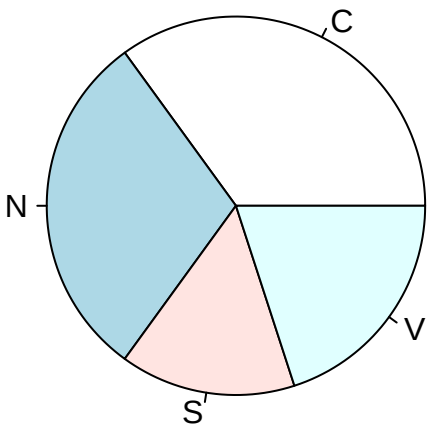
Per la variabile altezza, che è di tipo quantitativo, si utilizza lo stesso comando `table(vettore)`, ma occorre specificare l'ampiezza delle classi di frequenza con la funzione `cut()` e l'argomento `breaks()`, con il quale vengono specificati gli estremi superiori della classe (inclusi per default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

```
freq <- table(cut (heights$height,
           breaks = c(140,150,160,170,190,200)))
freq
```

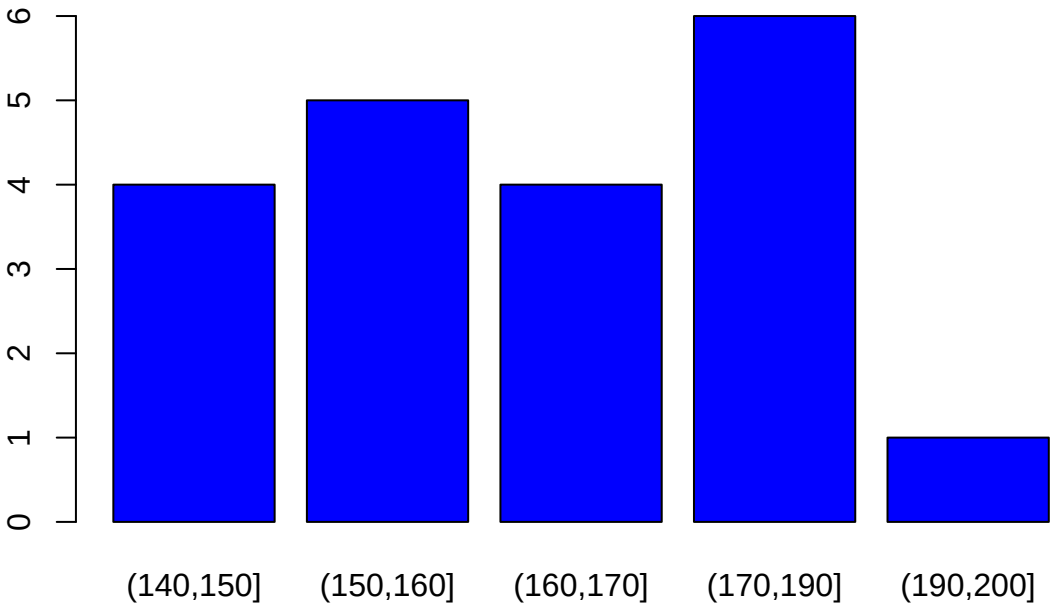
```
##
## (140,150] (150,160] (160,170] (170,190] (190,200]
##         4         5         4         6         1
```

Per disegnare i grafici si utilizzano le funzioni `pie()` e `barplot()`.

```
pie(table(heights$var))
```




```
barplot(freq, col="blue")
```



In Excel, l'operazione di classamento può essere effettuata utilizzando la formula "=FREQUENZA(matriceDati, matriceClassi)", che tuttavia è una formula di matrice e quindi deve essere immessa in un intervallo e consolidata con la combinazione tasti "SHIFT+CTRL+INVIO".

5.4 Statistiche descrittive per le distribuzioni di frequenza

Il più semplice indicatore di tendenza centrale, utilizzabile con qualunque tipo di dati è la *moda*, cioè il valore della classe che presenta la maggior frequenza. Ovviamente, se la variabile è quantitativa, si assume come moda il punto centrale della classe con maggior frequenza. L'individuazione della moda è banale e non richiede calcoli di sorta.

Nel caso di distribuzioni di frequenza per caratteri ordinabili (qualitativi e quantitativi), oltre alla moda possiamo calcolare la *mediana* e gli altri percentili.

Oltre a questi, per le distribuzioni di frequenza dicaratteri quantitativi è anche possibile calcolare la media, come illustrato in precedenza, insiema tutti gli indicatori di variabilità già citati.

5.5 Distribuzioni di frequenza bivariate: le tabelle di contingenza

In alcuni casi in ciascuna unità sperimentale del collettivo vengono studiati due (o più) caratteri e, di conseguenza, si ha a che fare con distribuzioni di frequenza bivariate (o multivariate). In questo caso si possono costruire delle tabelle di contingenza, cioè delle tabelle a due entrate nelle quali ogni numero rappresenta la frequenza congiunta (in genere assoluta) per una particolare combinazione delle due variabili.

Ad esempio consideriamo le variabili Varietà (con i valori SANREMO e FANO) e 'Forma delle bacche' (con i valori LUNGO, TONDO, OVALE), riportati nella tabella di contingenza che creeremo come matrice.

```
tabCon <- matrix(c(37, 45, 32, 74, 61, 59), nrow = 2, ncol = 3,
                 byrow = F)
row.names(tabCon) <- c("SANREMO", "FANO")
colnames(tabCon) <- c("LUNGO", "TONDO", "OVALE")
tabCon
```

```
##          LUNGO TONDO OVALE
## SANREMO     37    32    61
## FANO        45    74    59
```

Ogni riga della tabella sovrastante costituisce una distribuzione condizionata della forma del frutto, dato un certo valore della Varietà, mentre ogni colonna costituisce una distribuzione condizionata della Varietà, data una certa forma del frutto.

Capitolo 6

Connessione

Se guardiamo le due distribuzioni condizionate per SANREMO e FANO possiamo notare che esiste una certa differenza. Potremmo chiederci quindi se il presentarsi di una data modalità del carattere Varietà (SANREMO o FANO) influenza il presentarsi di una particolare modalità del carattere Forma del frutto. Se ciò non è vero si parla di indipendenza delle variabili (allora le distribuzioni condizionate sono uguali) altrimenti si parla di dipendenza o connessione. In caso di indipendenza, le distribuzioni condizionate delle due variabili dovrebbero essere uguali tra loro, cioè la frequenza relativa condizionale di X per una data modalità di Y deve essere uguale alla frequenza relativa condizionale di X per l'altra modalità di Y e quindi alla frequenza marginale di X.

Ad esempio, per il carattere LUNGO la frequenza relativa marginale è pari ad 82/308=0.266 (82 è la somma dei pomodori di forma allungata, mentre 308 è il numero totale dei pomodori); in caso di indipendenza, questa frequenza dovrebbe essere la stessa, indipendentemente dal fatto che il pomodoro sia di varietà SANREMO oppure Fano. In cifre, la frequenza assoluta condizionata per LUNGO|SANREMO dovrebbe essere pari a 0.266x130=34.6. mentre LUNGO|FANO dovrebbe essere pari a 0.266x178=47.4. Con questi principi, possiamo costruire la tabella delle frequenze assolute attese, in caso di indipendenza completa tra i due caratteri.

```
expF <- matrix(c(34.6, 47.4, 44.7, 61.3, 50.6, 69.4),
                 nrow = 2, ncol = 3,
                 byrow = F)
row.names(expF) <- c("SANREMO", "FANO")
colnames(expF) <- c("LUNGO", "TONDO", "OVALE")
```

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

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

dove f_o sta per frequenza osservata ed f_a sta per frequenza attesa nel caso indipendenza. Questo indice assume valore pari a zero nel caso di indipendenza completa (le frequenze osservate sono uguali a quelle attese) ed assume un valore positivo tanto più alto quanto maggiore è la connessione tra i due caratteri, fino ad un valore massimo dato dal prodotto del numero degli individui per il valore minimo tra il numero di righe - 1 e il numero di colonne - 1:

$$\max \chi^2 = n \cdot \min(r - 1, c - 1)$$

Nel nostro caso, potremmo calcolare il chi quadro in questo modo:

```
sum( ((tabCon - expF) ^ 2) /expF )
```

```
## [1] 10.22348
```

Esiste anche un comando più semplice, che consiste nell'utilizzare la funzione `as.table()` per forzare la matrice `dati` in una tabella di contingenza ed applicare la funzione 'summary()'.

```
summary( as.table (tabCon))
```

```
## Number of cases in table: 308
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 10.223, df = 2, p-value = 0.006027
```

Il valore massimo di chi quadro è pari a 308 e di conseguenza il valore osservato, espresso in relazione al valore massimo è pari a 10.22/308=0.033. Si può quindi concludere che la connessione tra i due caratteri è piuttosto debole.

Capitolo 7

Correlazione

Se abbiamo a che fare con variabili quantitative, possiamo calcolare l'indice di connessione previa opportuna divisione in classi di frequenza delle due variabili in studio. Oltre a ciò, con variabili quantitative è possibile esplorare l'esistenza della cosidetta relazione di variazione congiunta, che si ha quando al variare di una variabile cambia anche il valore dell'altra.

La variazione congiunta si quantifica tramite il **coefficiente di correlazione** costituito dal rapporto tra la codevianza (o somma dei prodotti) delle due variabili e il prodotto delle loro devianze. Il coefficiente di correlazione varia tra -1 e +1: un valore pari a +1 indica concordanza perfetta (tanto aumenta una variabile, tanto aumenta l'altra), mentre un valore pari a -1 indica discordanza perfetta (tanto aumenta una variabile tanto diminuisce l'altra). Un valore pari a 0 indica assenza di qualunque grado di variazione congiunta tra le due variabili (assenza di correlazione). Valori intermedi tra quelli anzidetti indicano correlazione positiva (se positivi) e negativa (se negativi).

In R, per calcolare la correlazione tra due variabili si usa la funzione `cor()`. In Excel, abbiamo la funzione "=CORRELAZIONE(intervallo1, intervallo2)"

Proviamo a considerare questo esempio: il contenuto di olio di 9 lotti di acheni di girasole è stato misurato con due metodi diversi ed è riportato più sotto.

```
a <- c(45, 47, 49, 51, 44, 37, 48, 44, 53)
b <- c(44, 44, 49, 53, 48, 34, 47, 46, 51)
```

Valutare la correlazione tra i risultati dei due metodi di analisi.

```
cor(a, b)
```

```
## [1] 0.8960795
```

Possiamo osservare che il coefficiente di correlazione è abbastanza vicino ad 1 e quindi possiamo concludere che esiste un buon grado di concordanza tra i due metodi di analisi.

Capitolo 8

Final Words

We have finished a nice book.

References