Verità ed esperienza: come la natura genera le osservazioni sperimentali

Andrea Onofri
Dipartimento di Scienze Agrarie ed Ambientali
Universitá degli Studi di Perugia

10 gennaio 2012

Indice

Sommario

Questa esercitazione ha l'obiettivo di introdurre lo studente ai meccanismi di 'confusione' che la natura mette in atto, impedendoci di conoscere direttamente la realtà delle cose. Infatti, a causa di quello che è genericamente definito 'errore sperimentale', i risultati di un esperimento non coincidono mai con la verità 'vera', a meno che l'esperimento non è ripetuto infinite volte. Questo giustifica da parte nostra un atteggiamento prudenziale, che ci impone di presentare i risultati associati sempre ad un'adeguata banda di incertezza (errore standard e/o intervalli di confidenza).

1 Presupposti

Verità 'vera' e verità sperimentale

1. PRESUPPOSTO: I fenomeni biologici seguono una legge di natura (verità 'vera'), che ne costituisce il meccanismo fondamentale.
2. Organizzando un esperimento, le osservazioni sperimentali seguono questo meccanismo di fondo, al quale si sovrappongono tutti i meccanismi di 'confusione', altamente incontrollabili, che vanno sotto il nome di errore sperimentale (verità 'sperimentale').
3. La verità sperimentale è quindi un immagine confusa della verità vera
4. Compito del ricercatore è quello di separare l'informazione (che rappresenta la verità 'vera') dal 'rumore di fondo'

Questo dualismo tra verità 'vera' (inconoscibile) e verità sperimentale (esplorabile tramite un esperimento opportunamente pianificato) è l'aspetto centrale di tutta la biometria e, per essere ben compreso richiede qualche esempio. Gli esempi che seguono sono volutamente irreali, in quanto sono basati su una verità 'vera' creata artificialmente, tramite simulazione. Nella realtà, la verità vera rimane assolutamente ignota.

2 La media algebrica

Generazione dei dati sperimentali

- VERITA' VERA: abbiamo una soluzione erbicida a concentrazione pari a 120 mg/l, che viene misurata tramite un gascromatografo.
- MECCANISMI DI CONFUSIONE: Lo strumento di misura ha un coefficiente di variabilità del 10% (corrispondete ad una deviazione standard pari a 12 mg/l).
- VERITA' SPERIMENTALE: i risultati di analisi chimiche ripetute saranno diversi tra di loro e probabilmente diversi dal valore vero di 120 mg/l.

Come si può riprodurre questo processo?

- Immaginiamo che la natura operi secondo un meccanismo perfettamente gaussiano, dove gli errori positivi e negativi sono equiprobabili, con probabilità decrescente al crescere della distanza dal valore 'vero'
- Utilizziamo il generatore di numeri casuali di Excel, che è disponibile tra gli strumenti aggiuntivi di analisi dei dati.

- Otteniamo i tre valori 109.28, 132.29 e 130.85 (ovviamente, ripetendo l'estrazione i valori cambiano!!!)
- Questo è il risultato dell'esperimento, per creare il quale ci siamo sostituiti alla natura, riproducendo i suoi meccanismi di confusione.

In Excel, assicurarsi di avere installato ed abilitato gli strumenti di analisi dei dati (in Excel 2003 Strumenti/Componenti aggiuntivi/Strumenti di analisi; in Excel 2010: File/Opzioni/Componenti aggiuntivi/Strumenti di analisi/Vai..). Scegliere 'analisi dei dati' dal menù dati e selezionare lo strumento di generazione dei numeri casuali normali. Immettere le informazioni richieste.

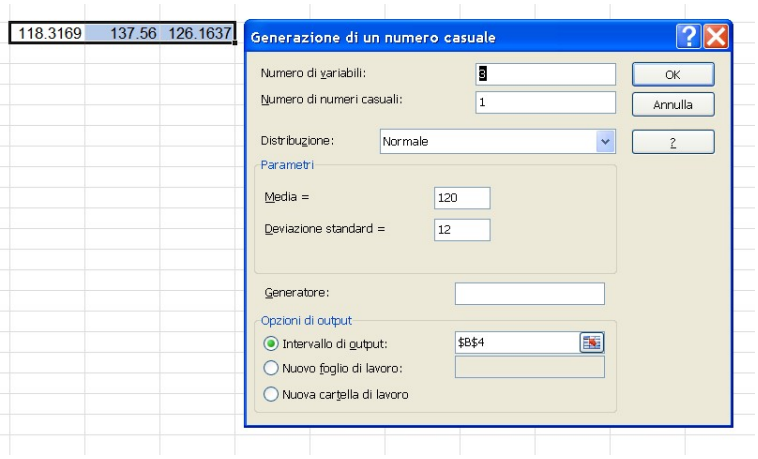


Figura 1: Impiego dello strumenti di generazione dei dati in Excel

Cosa concludiamo?

- In base alle osservazioni in nostro possesso, concludiamo che la cconcentrazione erbicida è pari a 124.14 mg/l, con una deviazione standard pari a 12.89 mg/l.
- La verità sperimentale non coincide con la verità 'vera', ma non siamo molto distanti, compatibilmente con il 10% di variabilità dello strumento di analisi.
- Lo scopo dell'esperimento però non è fornire informazioni sulla verità sperimentale, ma su quella 'vera'
- E' quindi giustificato un atteggiamento prudenziale da parte nostra.
- Che cosa succederebbe se ripetessimo l'esperimento molte altre volte?

La statistica 'frequentista' (così chiamata per distinguerla da quella bayesiana) assume che la verità vera è fissa e la variabilità di osservazione è misurabile attraverso un ipotetico meccanismo di ripetizione degli esperimenti. In questo caso l'esperimento è solo 'elettronico' e possiamo quindi ripeterlo un numero anche molto elevato di volte, attraverso un metodo che prende il nome di 'Monte Carlo', dalla città nella quale si trova il famoso 'casinò'.

Simulazione dei risultati

- Ripetiamo l'estrazione precedente per 10000 volte (ripetiamo l'analisi chimica per 10000 volte, sempre con tre repliche)
- Otteniamo 10000 medie, la cui media è pari a 120.03 e la cui deviazione standard è pari 6.95
- La media delle medie è ora praticamente coincidente con la verità 'vera'
- La deviazione standard delle medie rappresenta la banda di oscillazione della verità 'sperimentale' (come cambiano i risultati dei diversi esperimenti)
- Questa è pari a circa $12.89/\sqrt{(3)}$, cioè l'errore standard della media (SEM)
- Il 95% delle medie sono comprese tra 106.38 e 133.33. Cioè sono all'incirca compresi nell'intervallo tra la media osservata $\pm$ 2 volte l'errore standard (Intervallo di confidenza).

Se ci fossimo fermati al primo esperimento e avessimo concluso che la concentrazione erbicida è pari a 120.03 mg/l $\pm$ 13.9 mg/l, avremmo dato informazioni attendibili sulla verità biologica 'vera' e non solo sui tre dati effettivamente osservati.

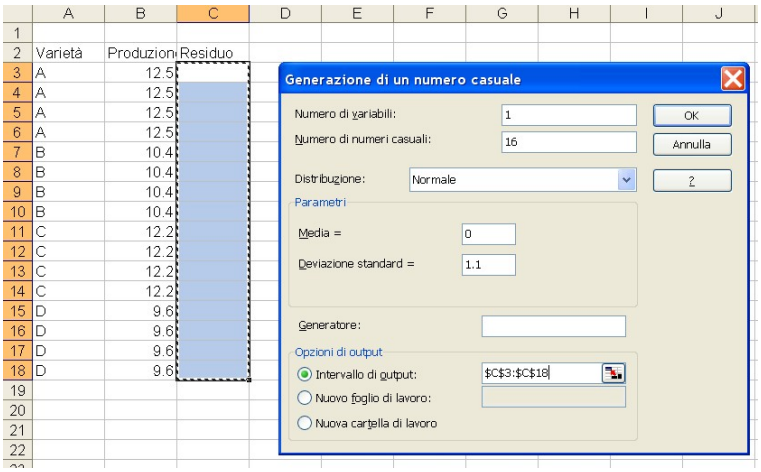
3 Una prova varietale

Generazione dei dati sperimentali

- VERITA' VERA: abbiamo quattro varietà, la cui produttività potenziale in una data situazione ambientale è rispettivamente di 12.5, 10.4, 12.2 e 9.6 t/ha.
- MECCANISMI DI CONFUSIONE: la variabilità ambientale interannuale, l'eterogeneità del suolo, la variabilità individuale e l'errore di misura producono delle oscillazioni produttive normali (gaussiane), con media 0 e deviazioni standard pari ad 1.1 t/ha.
- VERITA' SPERIMENTALE: la prova varietale produce risultati diversi nelle diverse repliche e le medie produttive sono pertanto diverse da quelle vere.

- In un caso (A e C), la graduatoria varietale viene invertita, rispetto alla realtà
- Si veda il foglio 'GenerazioneDati.xls'

Anche in questo caso utilizziamo lo strumenti di generazione dei dati di Excel.



Ed arriviamo a questo risultato

Varietà	Produzione	Residuo	Osservato	Medie
A	12.5	-0.965406	11.53459	**12.02799**
A	12.5	-1.00755	11.49245	
A	12.5	-0.249563	12.25044	
A	12.5	0.334484	12.83448	
B	10.4	-0.602497	9.797503	**10.5777**
B	10.4	1.238192	11.63819	
B	10.4	-0.384551	10.01545	
B	10.4	0.459652	10.85965	
C	12.2	0.130864	12.33086	**12.77947**
C	12.2	0.549699	12.7497	
C	12.2	1.597974	13.79797	
C	12.2	0.039347	12.23935	
D	9.6	-0.051897	9.548103	**8.893183**
D	9.6	-1.708008	7.891992	
D	9.6	-0.333249	9.266751	
D	9.6	-0.734113	8.865887	

4 Il caso della regressione lineare semplice

4.1 Generazione dei dati

L'estensione al caso della regressione lineare è estremamente semplice: si assume che la variabile dipendente è funzione di quella indipendente secondo la retta di equazione generale $y = a + bx$. In questo caso immaginiamo (semplificando di molto la realtà) che la produzione di biomassa (DW) di un prato foraggero incrementi con il tempo (t) secondo una funzione lineare:

$$DW = 12 + 0.5 \cdot t + \epsilon$$

dove ϵ, che assumiamo normalmente distribuito, ha media pari a 0 e deviazione standard pari a 3.5, che si mantiene costante nel tempo:

$$\epsilon \; N(0, \sigma = 3.5)$$

Di conseguenza, i dati che verranno effettivamente osservati nel corso dell'esperimento sono quelli riportati nella colonna a destra.

t (days)	DW vero	epsilon	DW oss
10	17	-0.64175	16.35825
20	22	-2.70908	19.29092
30	27	-7.15393	19.84607
40	32	1.389218	33.38922
50	37	-0.59171	36.40829
60	42	-3.13002	38.86998
70	47	2.664142	49.66414

4.2 Stima dei parametri

Utilizzando le funzioni interne ad Excel per la regressione lineare, otteniamo un modello stimato:

$$y = 0.5559x + 8.3127$$

Anche in questo caso non siamo molto lontani dalla verità 'vera'.

Residui e varianze

Le previsioni effettuate con questo modello, confrontate con i valori osservati portano agli scarti illustrati nella seguente tabella:

DW oss	DW previsto	epsilon
16.35825258	13.8717	2.48655258
19.29091883	19.4307	-0.139781173
19.84606929	24.9897	-5.14363071
33.38921791	30.5487	2.840517914
36.408289	36.1077	0.300589
38.86998422	41.6667	-2.796715784
49.6641419	47.2257	2.438441903

Gli scarti hanno media pari a 0 e deviazione standard pari a 3.02, con 5 gradi di libertà (7 dati meno due parametri stimati).

4.3 Intervalli di confidenza

Nel caso della regressione lineare gli errori standard dei parametri stimati sono pari a:

$$ES(b_0) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}}{SQ_x}}$$

ed:

$$ES(b_1) = \frac{\sigma}{\sqrt{SQ_x}}$$

dove $\bar{x}$ è la media delle X osservate e SQ_x è la loro devianza. Gli errori standard sono quindi 0.062 e 2.792, rispettivamente per b0 e b1. Anche in questo caso possiamo ottenere gli intervalli di confidenza moltiplicando per 2 gli errori standard.

4.4 Conclusione

Con i dati in nostro possesso, possiamo concludere che la pianta cresce secondo una funzione lineare, con b0 = 8.313 ± 5.58 e b1 = 0.556 ± 0.125. Il modello rispecchia bene la verità 'vera' permettendo anche di apprezzare l'incertezza che abbiamo nel trarre conclusioni su ciò che non conosciamo con esattezza.

ATTENZIONE: abbiamo costruito un modello nel quale abbiamo assunto la normalità degli errori, con una deviazione standard costante nel tempo. I dati sono stati generati secondo questo modello, ma non è assolutamente detto che le cose, in natura, vadano in questo modo!

5 Da ricordare

1. La natura genera i dati
2. Noi scegliamo un modello deterministico che simula il meccanismo di generazione dei dati attuato dalla natura.
3. Stimiamo i parametri.
4. Confrontiamo le previsioni con i dati osservati. Determiniamo ϵ e la sua deviazione standard (σ)
5. Assumiamo un modello stocastico ragionevole per spiegare ϵ, quasi sempre di tipo gaussiano, con media 0 e deviazione standard pari a σ, indipendente dalla X (omoscedasticità)

6. Attenzione alle assunzioni di normalità e omoscedasticità, perchè non sono sempre rispettate in natura (ne parleremo in seguito)!

7. Qualunque stima sperimentale deve essere associata ad un indicatore di variabilità (errore standard o intervallo di confidenza).

6 Per approfondimenti

CAMUSSI Alessandro , MOELLER Frank , OTTAVIANO Ercole , SARI GORLA Mirella (1995). Metodi Statistici per la sperimentazione biologica. Zanichelli Editore, 496 pp.