Lamento del biometrista scalzo (rel. 4)

www.mariocigada.com

A volte mi capita di collaborare all'analisi statistica di dati biomedici. In genere in questi casi chi ha raccolto i dati mi usa la cortesia di trasporli in un qualche formato elettronico (in genere un foglio elettronico) e mi invia il file.
Quando apro il file per la prima volta ho un attimo di trepidazione, poi quasi sempre appena guardo i dati vengo preso da una sensazione strana: mi prende il desiderio di sedermi per terra, col viso rivolto verso il muro, abbracciarmi le ginocchia e cominciare a dondolare recitando mentalmente le potenze di due...

E' possibile che a qualcuno il senso di quest'ultima frase possa sembrare oscuro, se ne avete voglia vi consiglio caldamente di leggere "Lo strano caso del cane ucciso a mezzanotte" di Mark Haddon (ed Einaudi) se non ne avete voglia non fa nulla: vi chiedo solo di avere un po' di pietà per un povero aspirante biometrista.

Ma lamentarsi non serve a niente, così, anche per far contenti la mia mamma ed il mio papà che hanno speso tanti soldi per farmi studiare la psicologia e la teoria della comunicazione, ho deciso di scrivere un piccolo manualetto su come si preparano i dati per un'analisi statistica. Infatti gli psicologi strategico- sistemici dicono che un buon metodo per cercare di evitare i malintesi è quello di commentare le regole per comunicare (in gergo usano il termine metacomunicare); ecco allora una breve (e probabilmente incompleta) comunicazione su come comunicare i dati destinati ad un'analisi statistica.

A) La prima cosa da fare dovrebbe essere aver ben chiaro in testa qual è lo scopo dello studio formulandolo in termini di domanda- risposta. Ovvero a quale domanda la ricerca cerca di dare una risposta?

B) Subito dopo è importante chiarirsi sulla natura della o delle misure che state per fare: sono di tipo quantitativo o qualitativo? Saranno poche o tante? Che cosa si ipotizza influenzi che cos'altro? In altre parole in questa fase è importante distinguere, per ogni dato, di che tipo è e se debba esser considerato un effetto (in gergo in genere gli statistici dicono risposta o anche variabile dipendente) oppure qualcosa che assomiglia ad una causa (potrebbe essere che lo statistico nel parlare si riferisca a questo come ad un fattore, una variabile indipendente o una covariata) .

C) A questo punto sarebbe utile una chiacchierata con lo statistico per cercare di ipotizzare quali saranno i test da utilizzare per l'analisi.

D) Solo ora inizia la raccolta dei dati e di conseguenza la loro introduzione in una qualche forma di foglio elettronico

E) Una buona idea potrebbe essere quella di inviare un file "di prova" con un piccolo campione di dati (diciamo una decina di osservazioni) per verificare se tutto fila liscio, se i file e i dati risultino leggibili, eccetera.

Forse qualcuno, a questo punto, avrà intuito che la scelta di quali analisi statistiche effettuare debba essere fatta prima di iniziare a raccogliere i dati, ma vi dirò di più, sono numerose le cose che vanno definite prima di raccogliere i dati e/o di effettuare gli esperimenti e in genere tutte queste cose vengono raccolte in un documento che prende il nome di protocollo. Le istruzioni per la realizzazione di un buon protocollo esulano dagli scopi di questo scritto, peraltro esistono numerosi manuali a questo scopo; nel mio piccolo ho provato a scrivere un semplice schema riassuntivo degli aspetti principali di un protocollo, potete trovarlo sullo stesso sito web da cui avete scaricato il documento che state leggendo. In particolare i punti A) e B) riportati sopra corrispondono a 2 punti fondamentali della realizzazione del protocollo: rispettivamente la definizione dello scopo della ricerca e la scelta dei cosiddetti "outcome", l'outcome principale e gli outcome secondari.

Inoltre, tra le righe, nei 5 semplici punti A-E son contenute 2 "regole" di carattere generale, mi son reso conto che farò riferimento spesso a queste 2 regole, per cui, per evitare noiose ripetizioni ho pensato di farvi cosa gradita spiegandole qui una volta per tutte.

I regola il gergo
Tutti hanno un loro gergo, sicuramente anche voi l'avete: anche gli statistici ne hanno uno; sarebbe una bella

cosa se, nel comunicare con questo marziano, faceste entrambi lo sforzo di apprendere almeno una parte del gergo altrui; ecco un breve elenco di termini che è probabile vi serviranno: media, varianza, deviazione standard, posizione, dispersione, effetto, variabile, variabile-risposta, fattore, parametro.
Ma come si fa, direte voi, non c'è il tempo di fare un corso di statistica! La soluzione è relativamente semplice: chiedere. E in effetti questa è un'applicazione della II regola (oltre che essere una meta-meta-comunicazione)

II regola Il feedback
Mettetevi d'accordo in anticipo, ma nel dubbio chiedete, quando non siete sicuri di aver chiara una cosa restituite un esempio, come nel punto "E" qui sopra. Cercate di immaginare in anticipo i possibili malintesi e comunicateli prima del patatrac.

Stabilite queste regole generali vediamo in dettaglio a cosa fare attenzione nel caricare i dati, sono 10 semplici punti, come un decalogo, ho scritto anche qualche esempio, spero non vi annoino troppo.

1. Righe o colonne?

A volte è enormemente più comodo avere i dati ordinati riga per riga, a volte colonna per colonna: parlatene prima con chi dovrà analizzare i dati: (regola II). La maggior parte dei programmi per l'analisi statistica lavora meglio con dati organizzati per colonne: come nell'esempio 1a: lungo le righe ci sono le osservazioni (i soggetti in studio) e ogni variabile (cioè le caratteristiche cliniche, antropometriche e quant'altro) occupa una colonna

Es.1a

Sogg	*Peso (Kg)*	*Altezza (cm)*
1	78	180
2	75	173
3	70	165
4	80	178
5	67	170

Es.1b

	Sogg1	Sogg2	Sogg3	Sogg4	Sogg5
Peso	78	75	70	80	67
Altezza	180	173	165	178	170

Ecco, supponiamo di voler correlare peso ed altezza; avendo i dati organizzati come in es.1b è un pasticcio, con molti programmi in circolazione è impossibile, mentre con dati come in es.1a è un gioco da ragazzi.

2. Tante colonne o una colonna e codici di livello? Anche qui regola II

In genere se si prevedono analisi complesse come l'ANOVA multifattoriale, magari quando si vuole stimare anche l'interazione tra più fattori è più comodo avere i dati in una colonna con codici di livello come nell'es.2b. Mentre dati organizzati come nell'Es2a, al contrario, sono più comodi per analisi più semplici come il t di Student (che però prevede 2 soli gruppi, non 3 come in questo caso) o l'ANOVA a una via. Generalmente, comunque, i dati come nell'Es.2a non sono quasi mai utili, a meno che lo statistico non sia anche un programmatore, per cui, nel dubbio è molto meglio usare l'Es.2b come modello cui ispirarsi.

Es.2a

Gruppo1	*Gruppo2*	*Gruppo3*
78	74	68
76	75	71
65	66	67

Es.2b

Peso	*Gruppo*
78	1
76	1
65	1
74	2
75	2
66	2
68	3
71	3
67	3

3. Colori.

Non conosco software (SW) per l'analisi dei dati che siano in grado di operare in funzione del colore con cui un dato è rappresentato: se dovete distinguere tra di loro 2 tipi di dati non ha senso colorarli in modo differente, aggiungete una colonna con un codice come nell'Es.2b.

4. Codici.

E' possibile usare codici alfanumerici anche complessi, ma ricordatevi che durante il "data entry" dovrete ogni volta porre la massima attenzione nel digitare correttamente il codice. Il computer in questo caso si comporta da "cretino d'acciaio", per lui le righe dell'Es.4 qui a fianco contengono 4 codici tutti e 4 differenti tra di loro.

Es.4

Osservazione	*Aspetto della caratteristica x*
1	Continuo
2	Interrotto
3	Interr.
4	Interrottto

5. Numeri e caratteri alfanumerici insieme.

Permettetemi a questo punto di annoiarvi con qualche briciola di informatica, ma se non c'intendiamo bene su questo punto i problemi ci perseguiteranno come l'ombra di Banquo.

Contrariamente a quel che molti pensano, i computer non funzionano con la magia, ma sono soggetti anche loro alle leggi della fisica; per esempio se scrivete qualcosa in una cella di un foglio elettronico il computer deve riservare uno spazio *fisico* nella sua memoria per metterci quello che avete scritto, supponiamo un numero. Fin qui è semplice, ma il computer sa che voi potreste decidere di sostituire quel numero con un numero più grosso e in teoria il numero più grosso che c'è è = a infinito. Ma allora quanto spazio deve riservare il PC per quella cella? Infinito? Ma non si può: la memoria costa e una memoria infinita ha un costo infinito (fuori budget). Allora quei volponi degli informatici hanno inventato un trucco, dividere i dati per tipo: ci sono i numeri, e le cose che non sono numeri e a loro volta ci sono i numeri interi, i numeri con la virgola, i numeri in notazione esponenziale, le date, ecc. Le cose cambiano un po' a seconda del programma che utilizzate, ma in sostanza per ciascun tipo di dato viene riservata una certa area di memoria che dovrebbe andar bene per tutti i dati di quel medesimo tipo. Certo tutto ciò si basa sull'assunto che chi carica i dati non cambi idea e decida di scrivere caratteri dove fino ad ora ha scritto dei numeri e viceversa, ma se uno è ordinato le cose dovrebbero andare così e se non vanno così pazienza: il SW ri-allocherà lo spazio di tutte le celle. Ma attenzione questo modo di lavorare del PC ha una conseguenza molto importante: noi potremo effettuare dei calcoli solo sui dati che per il computer sono "numeri" mentre con dati rappresentati a caratteri (in gergo vengono anche detti "alfanumerici") potremo fare ben poco. Ma non è tutto, forse non sono stato abbastanza chiaro, forse vi è già chiaro, ma in una singola cella di un foglio elettronico non possono stare simultaneamente dati di tipo diverso, per cui, nel dubbio, il programma trasforma tutto in alfanumerico.

Per esempio guardate la tabella nella prossima pagina:

Es.5

Sogg	*Peso*	*Altezza*
1	78 kg	180 cm
2	75 kg	173 cm
3	70 kg	165 cm
4	80 kg	178 cm
5	67 kg	170 cm

Confrontatela con quella dell Es.1a. E' uguale? Non è affatto uguale: con la tabella dell'Es.1a è facile calcolare il peso e la statura media dei 5 soggetti, con la tabella dell'Es.5 e' assolutamente impossibile, se non cancellando a mano tutti i "kg" e tutti i "cm" ed accertandosi che il programma abbia trasformato le 2 colonne in numeri, qualche volta questo avviene automaticamente, qualche volta è necessario ordinare esplicitamente al programma di effettuare la trasformazione. Per questa stessa ragione dovrebbe essere evidente che le unità di misura vanno riportate nell'intestazione e non vanno mai mescolate ai dati numerici. Cfr Es.1a e Es.5

6. Punto decimale e cifre significative

Per identificare le cifre decimali alcuni SW usano il punto, altri la virgola. In genere passare da una notazione all'altra è abbastanza semplice, ma è possibile che lungo l'arduo cammino che porta alla perfezione, vogliate far sì che chi riceva i dati li abbia già nel formato migliore per il suo SW; in ogni caso penso sia ovvio che mescolare, in uno stesso file, dati rappresentati con il punto e dati con la virgola non è mai una buona idea. Purtroppo alcuni fogli elettronici si ritengono molto più intelligenti di chi li usa (magari hanno ragione, chissà...) e scelgono "di testa loro" come rappresentare i caratteri decimali. Mi spiego meglio: nei vecchi fogli elettronici era uso che, in ciascuna cella, i dati alfanumerici venissero allineati a sinistra, mentre i dati numerici a destra. In questo modo, a colpo d'occhio, era possibile individuare un errore. Per esempio, nella tabella qui sotto, i numeri decimali sono rappresentati con la virgola decimale, ma il secondo numero della colonna Bla è stato erroneamente scritto con il punto e allora il nostro cretino d'acciaio lo considera un codice alfanumerico, ma ci usa la cortesia di farcelo notare, allineandolo a sinistra

Bli	Bla
1	12,3
2	13.1
3	11

Alcuni fogli elettronici, come scrivevo sopra, si ritengono particolarmente intelligenti e fanno 2 cose che possono diventare disastrose: innanzitutto può darsi che cambino arbitrariamente il modo di rappresentare i numeri passando dal punto alla virgola in base all'installazione del programma: "se hai installato il SW in italiano vuol dire che sei italiano, quindi i decimali li rappresenti con la virgola, per cui ti cambio tutto con la virgola anche dove tu scrivi il punto" il che potrebbe anche funzionare se non fosse che quello che cambia è solo l'aspetto del dato, mentre la sua rappresentazione interna resta com'era per cui, quando il file verrà letto da un altro SW (il programma per l'analisi statistica), i nodi verranno al pettine rendendo impossibili i calcoli. Il tutto viene poi aggravato dalla seconda, brutta abitudine di questi fogli elettronici "intelligenti" e cioè quella di allineare al centro di ciascuna cella i dati, indipendentemente dal loro tipo. Il che rende il foglio indubbiamente più carino, ma rende anche impossibile accorgersi di eventuali errori di input!!

Sempre a proposito di decimali, vale la pena di porre cura anche al numero di cifre decimali significative; è questo un discorso lungo ma molto importante: se da un lato è sciocco misurare l'età di un gruppo di neonati in mesi, pure è assurdo aggiungere decimali inutili o fittizi. Per esempio se misuro 3 finestre con il metro a stecca:
74 71 75 (cm)
non è giusto che dica che la larghezza media è di 73,3333 cm come se avessi impiegato uno strumento in grado di misurare i micron!

7. Titoli e Abbellimenti vari

Ho un po' paura di essere noioso: ma questo genere di problema mi capita così spesso che vorrei proprio esser sicuro di essermi spiegato bene. Guardate l'esempio nella prossima pagina

Es.7a

Soggetto	*Peso*	*Altezza*	*Sesso*
1	65	168	F
2	70	180	F
3	58	160	F
4	82	177	M
5	75	170	M

In questo caso chi ha caricato i dati ha pensato che sarebbe stato esteticamente gradevole separare in qualche modo i maschi dalle femmine. A parte il riprovevole atteggiamento da bacchettone, ha creato un bel pasticcio: se consideriamo ciascuna colonna come una variabile il PC riterrà che le osservazioni in ciascuna variabile siano 6, di cui una mancante, per cui, volendo calcolare la media dei pesi farà (65+70+58+82+75)/6 anziché diviso 5.
Se poi il mio ineffabile amico avesse messo nella riga bianca dei caratteri di separazione, come una bella fila di asterischi o di "=" , allora la variabile peso avrebbe avuto dei caratteri alfanumerici nella quarta posizione: visto che il PC non sa sommare numeri e caratteri, cercando di calcolare il peso medio tutto quel che otterrei sarebbe un bel messaggio di errore.

8. Valori mancanti (Missing Values)

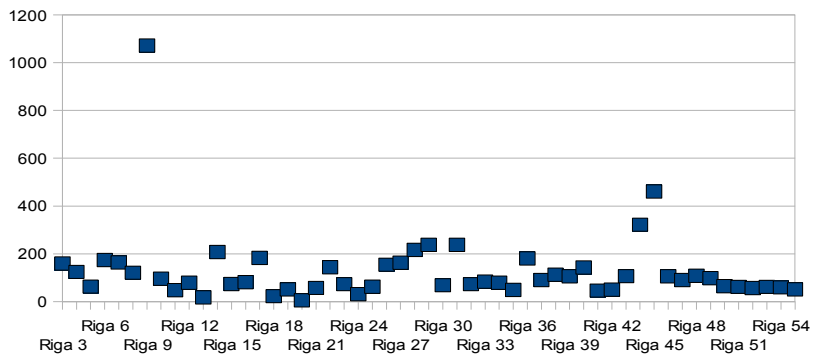
Come forse avete intuito dall'esempio precedente il SW per l'analisi statistica, nella sua diabolica furbizia, prevede la possibilità che qualche osservazione sia andata persa o non sia disponibile. In effetti ciò è indispensabile ma solleva la necessità di accordarsi su come codificare questi valori "missing".
La maggior parte dei SW prevede in questo caso di lasciare semplicemente lo spazio vuoto (blank), ovviamente mai come qui è fondamentale la regola II, comunque vi prego di non mettere uno 0 (zero); zero è un numero, indica che una misura vale zero, non un dato mancante. Provate ad immaginare di aver ricevuto una lettera da un notaio che vi comunica che siete nominati nel testamento di un vostro zio d'America che nemmeno sapevate di avere, è diverso sapere di aver ereditato zero dollari o sapere che per ora non si sa bene cosa avete ereditato, o no?

9. Chi fa i calcoli

Regola II (ancora!) accordarsi su chi effettuerà i calcoli. Se stiamo facendo un confronto tra 2 medie utilizzando il t di Student non ha alcun senso che chi carica i dati si dia la pena di calcolare le 2 medie. Certo esistono delle formule semplificate per calcolare il t di Student ma sono necessarie, oltre alle 2 medie, le 2 deviazioni standard (e le numerosità dei 2 gruppi). Per cui le 2 medie non mi bastano: dovrò comunque ripartire dai dati originari e ricalcolare le medie per calcolare le deviazioni standard eccetera, ma tutto questo lo fa il programma: la fatica per me è uguale, mentre magari per stare a calcolare quelle 2 medie chi mi fornisce i dati ha impegnato tempo e fatica preziosi che forse potevano esser meglio spesi, per esempio studiando la statistica :-)

10. Errori di battitura

Diciamocelo francamente: solo chi non lavora non sbaglia mai, allora è possibile che durante l'introduzione dei dati sia stato fatto qualche errore di battitura; se volete fare una cosa proprio carina potreste provare a fare un semplice grafico dei vostri dati (con un foglio elettronico non è difficile), questo rende subito evidenti eventuali valori "anomali" starà poi a voi andare a controllare se sono veramente valori anomali (ma vanno lasciati così) o se sembrano anomali solo perché corrispondono ad errori di battitura. Per esempio, è facile notare che all'incirca intorno alla riga 9-10 (e probabilmente anche dalle parti della riga 40- 45) dell'insieme di dati rappresentato nel grafico qui sotto ci sono dei valori sospetti.



Ecco, tutto qui, grazie allora per la pazienza e l'attenzione, buon lavoro.