Keep Calm and EnGioI Statistics - N°2.1

Andrea Sansone & Angelo Cignarelli

Marzo 2017

> Non mi fido molto delle statistiche, perché un uomo con la testa nel forno acceso e i piedi nel congelatore statisticamente ha una temperatura media.
>
> — Charles Bukowski

Introduzione al capitolo 2

Bentornati a Keep Calm and EnGioI Statistics, la prima newsletter di statistica in accaio inox 18/10[1]. Ci eravamo lasciati a fine febbraio con un sacco di domande sulla statistica, l'universo e tutto il resto. Se vi siete persi il primo capitolo, lo trovate a **questo indirizzo**.

[1] oggi in omaggio anche un set di dodici pentole con fondo fuso alto un centimetro!

Grazie innanzitutto a tutti quelli che ci hanno dato i loro commenti sulla newsletter; in particolare grazie a tutti coloro che *non* ci hanno inviato i loro insulti. Se volete rimediare, potete ancora farlo sulla pagina di Facebook di EnGioI (**link**) o via mail ai nostri indirizzi di posta elettronica (**Andrea** e **Angelo**).

In questo secondo appuntamento cerchiamo di entrare un po' più nel vivo della questione. Dalla prossima volta ci saranno anche esempi pratici - ragion per cui trovate in calce le istruzioni su come reperire ed installare il software di statistica che useremo nei prossimi appuntamenti.

I dati, o per meglio dire, le variabili

Il primo passo per capire i dati è riconoscere di quale tipo siano, a quali categorie appartengano e le sostanziali differenze tra di loro. La scelta dei test statistici appropriati sarà molto più semplice una volta comprese queste differenze. La prima fondamentale domanda da porci è: la variabile esprime una *quantità* o una *qualità*?

Se il dato indica una qualità, ci troviamo di fronte ad una *variabile categorica* o *qualitativa*. Esempi:

- Maschio/Femmina
- Cellula sana/Cellula malata
- Vivo/Morto[2]

[2] La variabile "lunedì mattina" non è considerata in ambito scientifico, purtroppo.

In pratica, le variabili categoriche non sono misurabili in termini numerici. Alcune di queste variabili possono indicare una valore crescente nell'ambito di una scala arbitraria, come ad esempio:

1. insufficiente
2. sufficiente
3. buono
4. ottimo

In questo caso avremo a che fare con una variabile categorica *ordinale*[3], altrimenti la variabile è definita categorica *nominale*.

[3] proprio perché esiste un ordine prestabilito!

Le variabili *quantitative* invece sono numeriche e pertanto misurabili. Possono essere *continue* o *discrete*: le prime sono indicative di una quantità continua (ad esempio, il numero di metri percorsi in una passeggiata, la concentrazione di un ormone, etc…) mentre le altre sono indicative di una quantità numerica ben definita (ad esempio, il numero di ruote che possiede un veicolo - non può possedere tre ruote e mezza!). Le variabili possono essere inoltre quantitative *per scala di rapporti* o *di intervalli*: questa ulteriore distinzione è un po' più astrusa, probabilmente, ma è facilmente spiegata da un paio di esempi. Il peso è una variabile per scala di rapporti: non possiamo cambiare lo 0 perché non avrebbe senso parlare di una cosa che pesa "meno 10 kg". La temperatura espressa in gradi Celsius è invece una variabile quantitativa per scala di intervalli: lo 0 è un'entità numerica fissata arbitrariamente[4]. Ha senso dire che una cosa pesa il doppio di un'altra, mentre ha meno senso dire che una giornata è "il doppio più calda" di un'altra.

[4] La scala Kelvin invece ha uno 0 assoluto, per cui è una variabile per scala di rapporti.

Schema riassuntivo:

- Qualitative (o categoriche)
  - Nominali (maschio/femmina; cellula sana/patologica)
  - Ordinali ("insufficiente", "sufficiente", "buono" e "ottimo")
- Quantitative
  - Continue (peso, altezza, glicemia, temperatura)
  - Discrete (numero di farmaci assunti)

Da un punto di vista statistico, vedremo quando è possibile utilizzare test appropriati per confrontare variabili quantitative o qualitative. Un t test non è utilizzabile per confrontare la prevalenze del genere maschile in una popolazione (abbiamo vista prima che il genere è una variabile qualitativa), così come non è possibile usare un test chi^2 per confrontare il livello di espressione del gene *Van Gogh*[5].

[5] Sì, esiste un gene con questo nome (**provare per credere**).

Statistica descrittiva

Una volta compreso quali tipi di variabili compongono il nostro database, è importante cominciare ad effettuare una analisi descrittiva sia per ragioni di sintesi e sia per avere un'idea su quanto sia omogeneo

il nostro campione. Ricordiamo che i nostri dati provengono sempre da un campione che si spera sia espressione (il più fedele possibile) dell'intera popolazione da esaminare. Tuttavia, per quanto il nostro campione possa avvicinarsi idealmente alla popolazione, va sottolineato come i nostri dati esprimeranno sempre una stima. Alcuni indici che andremo a discutere in questa sezione, hanno proprio la funzione di stimare le varibili del nostro campione[6] e indicarci quanto è omogeneo il nostro campione e con quanta confidenza possiamo traslare le informazioni che otteniamo alla popolazione intera[7]. Uno dei principi fondamentali (quasi tautologico) è che tanto maggiore è la numerosità del campione, tanto più questo somiglierà alla popolazione.

[6] Il livello di testosterone circolante in un campione di 20 soggetti maschi; oppure, il diametro degli adipociti misurato in un campione di tessuto adiposo ottenuto da 10 ratti.

[7] Il testosterone circolante o il diametro degli adipociti misurato in un 10 soggetti o ratti, rispettivamente, ovviamente non sarà lo stesso; ci sono indici che permettono di capire quanto si discostano le misurazioni le une dalle altre

In ciabatte in tangenziale andiamo a stimare, ovvero della media, mediana, et similia

Gli indici per la stima del campione sono tanti e pertanto vorremmo risparmiarvi tutti quelli che abbiamo dovuto (giustamente) imparare a memoria per superare gli esami di statistica, ma che poi non vengono utilizzati di routine per la redazione di tesi, articoli scientifici e simili, concentrandoci sui principali [8]. A un certo punto, dopo aver raccolto tutti i dati che ci interessano (e magari qualcuno che non ci interessa, ma non si sa mai cosa ha in serbo il futuro), siamo pronti ad effettuare una prima analisi. La prima cosa che si fa è abitualmente stimare come i nostri dati sono distribuiti: è possibile che siano tutti concentrati in uno spazio molto ristretto,o al contrario è possibile che siano molto lontani l'uno dall'altro. Gli indici di tendenza centrale servono a capire verso dove tende la nostra distribuzione.

[8] Facciamo riferimento a media armonica, media geometrica ecc.

La media è sicuramente tra i parametri più popolari in ambito scientifico (e non solo) per la stima di una variabile di tipo quantitativo. Dal punto di vista algebrico si tratta, molto semplicemente, della somma dei valori della nostra variabile di interesse diviso il numero delle osservazioni; dal punto di vista statistco esprimre proprio una stima di quale sia il livello di testosterone circolante o del diametro degli adipociti dei ratti dell'esempio di cui sopra. Per farlo, basterà fare la somma dei valori che ho registrato e dividerle per il numero di persone ratti che ho esaminato (ovvero 10). Se ad esempio i soggetti del nostro campione mostrano un testosterone di 4.2, 5.0, 4.8, 4.7, 4.9, 5.1, 6.0, 5.5, 5.1, 5.8 ng/dl, la media sarà

$$\frac{(4.2 + 5.0 + 4.8 + 4.7 + 4.9 + 5.1 + 6.0 + 5.5 + 5.1 + 5.8)}{10}$$

ovvero 5.11. Non vorremmo spaventarvi troppo presto, ma è importante dirvi già da ora che la media non rappresenta sempre il miglior parametro per stimare la variabile di interesse. Per il momento vi ba-

sti sapere che occorrerà verificare che la nostra variabile mostri una distribuzione normale[9] dei dati prima di utilizzare la media a cuor leggero.

[9] ne parleremo diffusamente in un prossimo capitolo

La mediana è un indice che si applica a variabili di tipo quantitativo o qualitativo ordinale e rappresenta quel valore che si trova esattamente a metà dei dati osservati. Tornando all'esempio precedente, la mediana è 5.05[10].

Per calcolarla manualmente, occorre disporre i valori in maniera crescente e identificare il valore centrale, ovvero il valore che occupa la posizione (n + 1)/2. Nel caso di una variabile con un numero dispar di dati è semplicissimo, in quanto se abbiamo 5 valori, la mediana casca sul terzo; se ne abbiamo 7, sul quarto; e così via. Se la nostra popolazione ha un numero pari di soggetti, come ad esempio per il nostro campione di 10 soggetti la mediana sarà la media dei due valori centrali - quindi la media tra 5.0 e 5.1 (rispettivamente il 5° e 6° valore).

[10] Avrete notato come la mediana si avvicini molto alla media in questo caso e quindi vi starete chiedendo "Perchè mi devo calcolare pure la mediana se è quasi uguale alla media?". Se vi siete fatti questa domanda, siete dei super cool, ma non avete ancora raggiunto il livello super Saiyan per ricevere la risposta...ancora qualche newsletter di pazienza

La frequenza Nel caso di una variabile di tipo qualitativo/categorico, il parametro fondamentale nella statistica descrittiva è rappresentato dalla frequenza, ovvero il numero con cui una determinata *qualità* si presenta. La frequenza può essere:

- assoluta
- relativa
- percentuale

Prendiamo ad esempio una scolaresca di 24 studenti o un campo di immunofluorescenza di 20 cellule di cui vogliamo conoscere la frequenza di, rispettivamente, studenti maschi o la frequenza di cellule positive ad un marcatore di apoptosi (i.e. annessina V). La frequenza assoluta[11] non è altro che il numero di studenti maschi (nel caso specifico, 12) o delle cellule apoptotiche (nel caso specifico, 5). La frequenza relativa, invece, rappresenta la frequenza di una determinata categoria rispetto al numero totale del campione, ed è un numero che può andare da 0 a 1. Nel nostro esempio, la frequenza relativa degli studenti maschi è 0.5 ovvero

$$\frac{12}{24} = 0,5$$

mentre la frequenza delle cellule apoptotiche è 0.2 ovvero

$$\frac{5}{20} = 0,2$$

[11] quasi ci vergognamo a spiegarla, ma tant'è se for dummies deve essere, for dummies sia

Moltiplicando semplicemnte la frequenza relativa per 100, abbiamo la frequenza percentuale, un parametro molto intuitivo che ci indica

come il 50% degli studenti sia di genere maschile o come il 20% delle cellue siano apoptotiche.

Chiaramente non tutti gli indici di tendenza centrale sono applicabili a tutte le variabili. La media fra i colori degli occhi non può esistere; la mediana fra variabili nominali non ha senso a meno di non avere variabili ordinali.

Sporchiamoci le mani…

In ogni capitolo, cercheremo di inserire in calce un paragrafo in cui metteremo in pratica ciò che è stato spiegato in precedenza. Per questo obiettivo, abbiamo deciso di diffondere i rudimenti per l'impiego di un software che useremo per il resto della nostra newsletter e si chiama *R*. Si pronuncia *arr*[12] ma la maggior parte di noi italianizza il nome in "Erre". Perché non SPSS, vi chiederete? Ci sono una serie di buone ragioni. La prima, è un software open-source, quindi gratuito e modificabile se vi sentite particolarmente coraggiosi. La seconda, funziona nello stesso modo su ogni sistema operativo (Windows, Mac o Linux). La terza, e più importante, è che R vi costringe a riflettere su cosa state facendo; questo processo mentale, all'inizio potrà sembrare di intralcio, ma, una volta superato l'impatto iniziale (che bisogna ammettere è abbastanza raggelante) consente di ottenere molti più risultati con molta più consapevolezza. Con SPSS, come avevamo scritto nel primo "capitolo", è facile che premendo bottoni a caso alla fine si arrivi ad un risultato con $p < 0.05$; tuttavia, come già detto, **"significativo" non è un sinonimo di "importante"** Ciò posto, nulla vi vieta di leggere la parte teorica di *Keep Calm and EnGioI Statistics* e poi lavorare col vostro software preferito[13].

[12] Come il tipico verso dei pirati. Yohohoho!

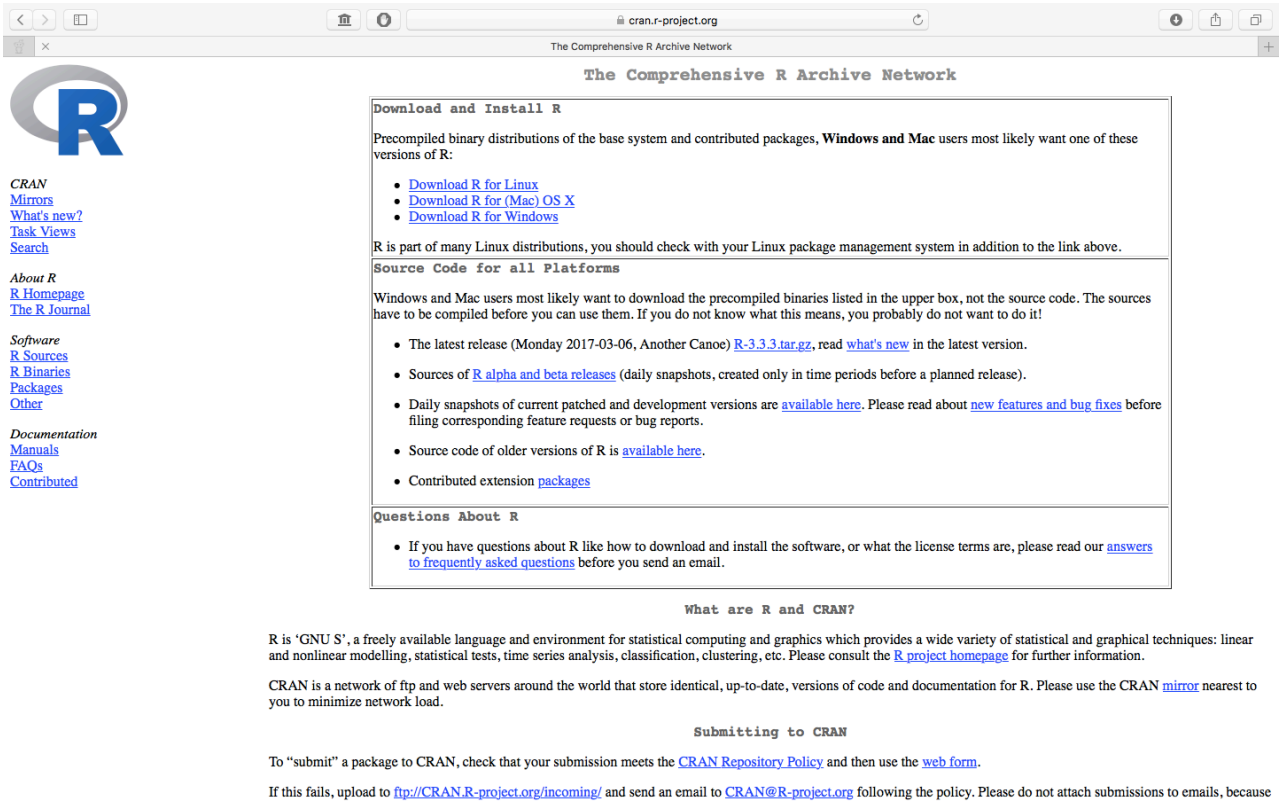
[13] Purchè non sia quel software commerciale che inizia con "ex" e finisce con "cel" :(

Installare R

R è, come detto poco fa, un software gratuito, scaricabile dal sito **The Comprehensive R Archive Network** - in breve, CRAN.

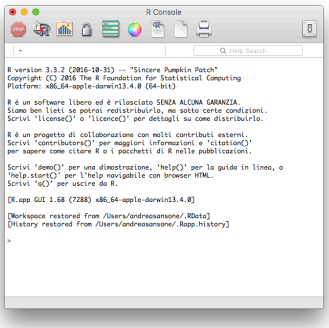
Nella pagina seguente potete vedere la pagina principale di CRAN in tutta la sua bellezza. Sì, è orrenda. L'importante è che vediate bene quelle simpatiche scritte in alto, dove c'è scritto "Download R for"… Cliccate sul vostro sistema operativo, scaricate il file più in alto nella lista e voilà! Avete appena ottenuto, senza alcuna fatica, la versione più recente di R. Al momento in cui scriviamo è disponibile la versione 3.3.3, uscita il 7 marzo.

Chi di voi usa Mac dovrà scaricare anche un'altra app che si chiama XQuartz; il link lo trovate sempre su CRAN nella pagina dei download ma se siete pigri potete anche cliccare **qui**. Se avete problemi con



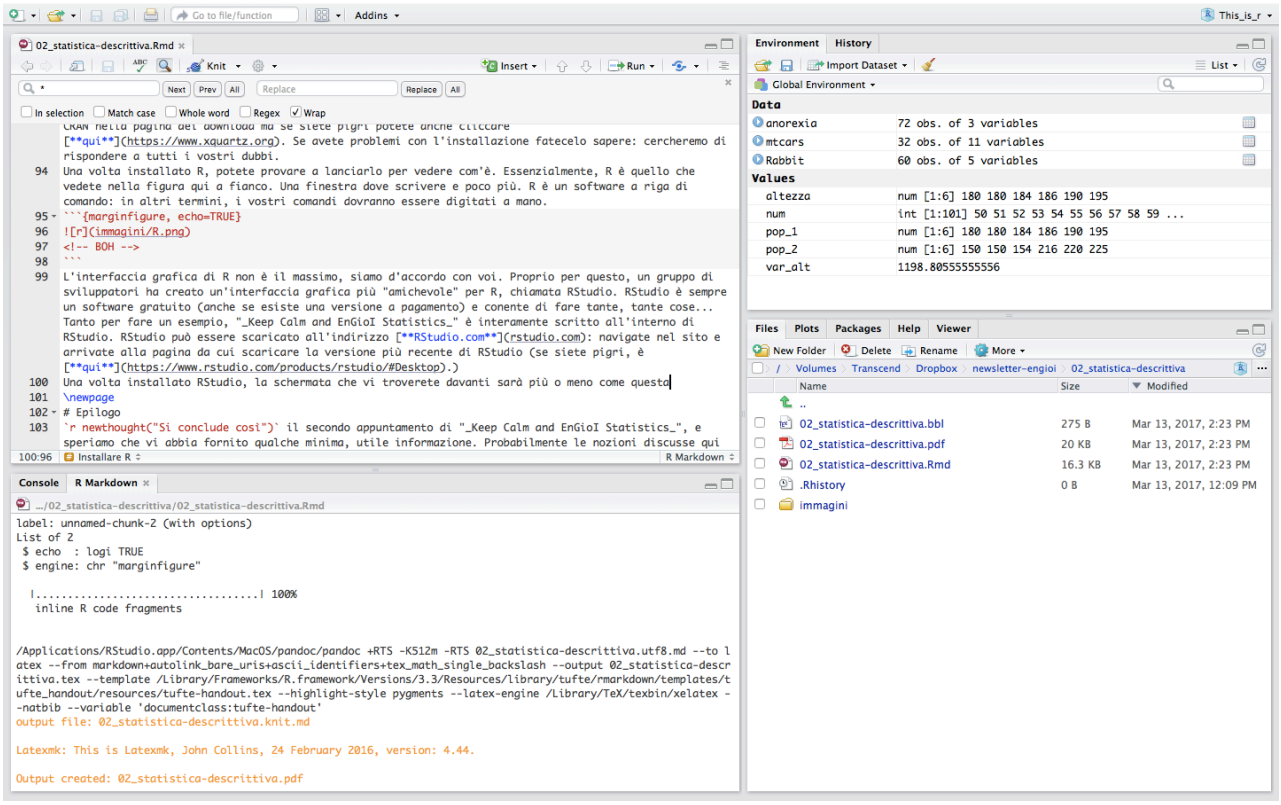
l'installazione fatecelo sapere: cercheremo di rispondere a tutti i vostri dubbi.

Una volta installato R, potete provare a lanciarlo per vedere com'è. Essenzialmente, R è quello che vedete nella figura qui a fianco. Una finestra dove scrivere e poco più. R è un software a riga di comando: in altri termini, i vostri comandi dovranno essere digitati a mano.



L'interfaccia grafica di R non è il massimo, siamo d'accordo con voi. Proprio per questo, un gruppo di sviluppatori ha creato un'interfaccia grafica più "amichevole" per R, chiamata RStudio. RStudio è sempre un software gratuito (anche se esiste una versione a pagamento) e conente di fare tante, tante cose... Tanto per fare un esempio, "*Keep Calm and EnGioI Statistics*" è interamente scritto all'interno di RStudio. RStudio può essere scaricato all'indirizzo **RStudio.com**: navigate nel sito e arrivate alla pagina da cui scaricare la versione più recente di RStudio (se siete pigri, è **qui**.)

Una volta installato RStudio, la schermata che vi troverete davanti sarà più o meno come questa. Non temete: quello che vedete nell'immagine è il frutto di svariate ore di lavoro!



Epilogo

Si conclude così il secondo appuntamento di "*Keep Calm and EnGioI Statistics*". Probabilmente le nozioni discusse qui sopra sono già alla portata di tutti, ma sono per molti versi le basi per costruire tutti i concetti chiave della statistica, fra cui il teorema del limite centrale. Questo teorema rappresenta il fulcro su cui gira tutta (o quasi) la statistica inferenziale, che in fin dei conti è quella che più ci interessa.

Soprattutto, in questo "episodio" avete fatto il primo passo nel mondo serio della statistica con l'installazione di R. Nel prossimo incontro parleremo appunto di come usare R per gli scopi che più ci interessano: ad esempio, la tecnica per grattarsi le orecchie con i piedi e l'uso del grasso di balena come rimedio naturale per l'alitosi. Le risposte a questi ed altri interessanti quesiti vi aspettano nel prossimo episodio di "*Keep Calm and EnGioI Statistics*".