Keep Calm and EnGioI Statistics - N°2.2

Andrea Sansone & Angelo Cignarelli

aprile 2017

> Il succo della storia fin qui.
>
> Al principio fu creato l'Universo. Questo fatto ha sconcertato non poche persone ed è stato considerato dai più come una cattiva mossa.
>
> — Douglas Adams, "Ristorante al Termine dell'Universo"

Introduzione

Ci siamo lasciati col precedente numero parlando di indici per la stima delle variabili, in particolare di media, mediana e frequenza; se volete ripassare, o semplicemente se avete perso il PDF, potete ritrovare Keep Calm and EnGioI Statistics episodio 2.1 **sul sito della SIE**. Oggi, invece, vi parleremo di altri indici che ci consentiranno, insieme ai precedenti, di dare maggior dettaglio alla descrizione delle nostre popolazioni: gli indici di dispersione.

Indici di dispersione

La dispersione, al contrario, indica quanto i dati siano distanti tra loro e, di conseguenza, dalla media e/o mediana delle nostre popolazioni; infatti rappresentano la misura della (appunto) dispersione dei valori rispetto ad un parametro centrale. Faremo una rapida rassegna sui principali, alcuni dei quali particolarmente utili nella statistica di tutti i giorni.

La Varianza è una stima dell'accuratezza della media. Nell'esempio di prima sui livelli di testosterone nella nostra popolazione di soggetti maschi avevamo indicato che la media è 5.11 ng/dl. La stessa media si potrà ottenere, tuttavia, anche valori completamente diversi da quelli registrati nella newsletter precedente e magari anche più eterogenei, ad esempio: 10.2, 2.0, 3.8, 2.7, 2.9, 5.1, 1.0, 3.5, 4.1, 15.8[1]. Avrete sicuramente notato che pur avendo la stessa media, i valori sono molto più distanti tra loro. Non è molto difficile intuire che ci siano delle differenze fra le popolazioni, e la varianza ci offre una buona stima di quanto i valori si discosti tra di loro[2].

Come si calcola la varianza? Si parte dalla somma dei quadrati delle differenze rispetto alla media e la si divide per il numero dei soggetti in esame. Perchè dei quadrati, vi chiederete? Perché la somma delle differenze rispetto alla media è… zero! Provateci da soli se non ci credete. Torniamo al nostro esempio: sappiamo che la media calcolata in

[1] questi sono i valori della popolazione del numero precedente: 4.2, 5.0, 4.8, 4.7, 4.9, 5.1, 6.0, 5.5, 5.1, 5.8

[2] Attenzione attenzione, il valore della varianza di per sè non ci dirà granchè, ma servirà per calcolare la deviazione standard che, invece, dà informazioni molto più interpretabili.

entrambi i gruppi è 5.11, quindi per la prima popolazione la varianza si calcolerà così:

$$\frac{(-0.91)^2+(-0.11)^2+(-0.31)^2+(-0.41)^2+(-0.21)^2+(-0.01)^2+(-0.89)^2+(-0.39)^2+(-0.01)^2+(-0.69)^2}{10}$$

dove -0.91, -0.11, -0.31 eccetera sono le differenze fra ciascun valore e la media. Quindi la varianza della prima popolazione sarà: $\frac{2.569}{10} = 0.2$. Mentre per la seconda popolazione sarà calcolata così:

$$\frac{(5.09)^2+(-3.89)^2+(-1.31)^2+(-2.41)^2+(-2.61)^2+(-0.01)^2+(-4.11)^2+(-1.61)^2+(-0.99)^2+(-10.69)^2}{10}$$

cioè $\frac{190.0349}{10} = 19.1$. Avrete notato immediatamente come le 2 varianze siano molto differenti tra loro e, probabilmente, avrete notato che tanto più i valori del campione sono diversi e distanti tra loro tanto maggiore sarà la varianza. Visto e considerato che abbiamo calcolato una somma di quadrati abbiamo un errore espresso in ng/ml quadrati, che non è proprio intellegibile. A questo punto, indovina indovinello, cosa succede se calcoliamo la radice quadrata della varianza?

La deviazione standard, ecco che succede! La famosissima deviazione standard, spesso rappresentata dal simbolo σ o dalla sigla DS, altro non è che la radice quadrata della varianza, appunto. La DS è un indicatore molto più intellegibile della dispersione dei valori e, in parole povere, fornisce un'indicazione numerica di quanto i dati siano vicini o lontani dalla media, ovvero il grado di incertezza da associare alla media che abbiamo calcolato. Nel nostro caso particolare, facendo la radice quadrata delle due varianze otteniamo, quindi: $\sqrt{(0.2)} = 0.44$ e $\sqrt{(19.1)} = 4.3$.

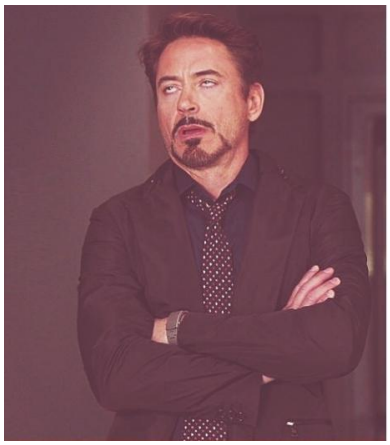
Quindi possiamo affermare che le due popolazioni hanno una media di testosterone di rispettivamente 5.11 ± 0.44 e 5.11 ± 4.3 ng/ml, ovvero, il numero dopo il $\pm$ indica di quanto si scostano mediamente i valori del campione dalla media; nel nostro caso possiamo dire che nella prima popolazione i valori si scostano molto di meno dalla media rispetto alla seconda popolazione. A questo punto, indovina indovinello, cosa succede se dividiamo la deviazione standard con la radice quadrata del numero dei campioni?

Errore standard Ecco che succede! Si ottiene l'errore standard (ES). In sintesi è un indice di precisione delle media. Per spiegarlo in parole povere abbiamo bisogno di una piccola noiosa ma fondamentale digressione, prendendo l'esempio del livello di testosterone dei 10 studenti del numero precedente. Vi ricordate che il livello medio era 5.11 ng/ml? Se misurassi il testosterone in 10 studenti di un'altra classe, troverò quasi sicuramente un altro valore, magari molto vicino, ma comunque diverso...è la Natura, raga, niente di misterioso. Così

via, misurando la nostra variabile in un numero n di campioni, troverò n medie diverse. L'errore standard ci viene in soccorso perché ci stima proprio questa entità di variazione, ovvero, più l'ES è piccolo minore sarà la variazione della media in campioni differenti della nostra popolazione. Pertanto, l'errore standard della prima popolazione risulta $ES = \frac{5.37}{\sqrt[2]{10}} = 1.7$ mentre per la seconda popolazione $ES = \frac{34.63}{\sqrt[2]{10}} = 10.9$. Va da sè che maggiore sarà la numerosità del campione[3], minore sarà l'ES, ovvero più precisa sarà la media. Quindi, di fronte ad un ES molto alto abbiamo la ragionevole necessità di aumentare il numero di misurazioni o, almeno, cercare di investigare il motivo della eccessiva eterogeneità del campione.

[3] essendo la n al denominatore

INTERVALLO DI CONFIDENZA Nella pratica, l'errore standard serve per calcolare l'intervallo di confidenza che rappresenta l'intervallo di valori entro i quali si stima possa cadere, con un livello di probabilità scelto a piacere, il valore vero della popolazione. Nella pratica, si sceglie quasi sempre un livello di probabilità di 0.95 o, più raramente, 0.99, ottenendo rispettivamente l'intervallo di confidenza al 95% o al 99%. Tecnicamente l'intervallo di confidenza al 95% si calcola sommando la media al prodotto di un coefficiente t per l'errore standard. La t si ottiene da tabelle dei valori t per la distribuzione di Student che si trovano facilmente (ne trovi **qui** una semplificata) o si ottiene tramite calcolatori **on line**; l'unica cosa che occore conoscere sono i gradi libertà, che in questo caso sono rappresentati dal numero del campione -1, ovvero 10-1=9. Se vi state chiedendo perché bisogna sottrarre 1, vi meritate la foto a fianco…



Scherzi a parte, i gradi di libertà rappresentano un concetto matematico abbastanza complesso che esula sia dalle nostre competenze che, soprattutto, dallo scopo della newsletter. In ogni caso se volesta approfondire, potete trovare qualcosa dedicato all'uomo della strada **qui** o **qui**.

Tutti gli indici che abbiamo cercato di spiegare fin qui sono applicabili a variabili che presentano una distribuzione normale[4] e, quindi, seguono sempre la media; mentre, quando le variabili non presentano una distribuzione normale, per consuetudine si utilizza un indice più robusto come la mediana a cui si fa seguire, come indice di dispersione, il range interquartile.

[4] ancora un altro numero di pazienza e finalmente parleremo di normalità

IL RANGE INTERQUARTILE è l'ultima misura di dispersione di cui ci occupiamo. Abbiamo imparato che la mediana rappresenta quel valore che si pone esattamente a metà delle nostre osservazioni, giusto? Ci sono quindi 50% di valori al di sotto e 50% al di sopra della mediana. Se dividiamo a metà queste due metà, la conseguenza è che otteniamo una ripartizione in 4 parti uguali della nostra popolazione. Ci sarà quindi un valore corrispondente al 25% delle osservazioni, uno al 50% (la mediana), ed uno al 75%. Si definisce range interquartile (aka IQR) la distanza tra il valore al 25% ed il valore al 75% (la "metà centrale" delle osservazioni) - quindi il valore al 75% (definito il 75° percentile) meno quello al 25% (il 25° percentile). Torniamo al nostro

campione di valori testosterone per un esempio pratico: i valori registrati sono 10.2, 2.0, 3.8, 2.7, 2.9, 5.1, 1.0, 3.5, 4.1, 15.8. Mettiamoli in ordine crescente in modo da rendere più comodo calcolare la mediana[5] che, in questo caso in cui la numerosità del campione è pari sarà la media dei 2 valori centrali 1.0, 2.0, 2.7, 2.9, 3.5, 3.8, 4.1, 5.1, 10.2, 15.8, ovvero $\frac{3.5+3.8}{2} = 3.65$. Il 25% casca fra il primo numero (1.0) e la mediana (3.6), ed è 2.75; il 75% è a metà fra la mediana e l'ultimo valore 15.8, ed è quindi pari a 4.85. Possiamo concludere che il range interquartile delle osservazioni è quindi compreso fra 4.85 e 2.75; l'IQR è dato dalla differenza fra 4.85 e 2.75 ed è quindi 2.1. Se vi state chiedendo come abbiamo ottenuto questi valori, è il caso di procedere alla sezione "**Sporchiamoci le mani**" e proviamo a calcolarli con R.

[5] Piccolo sotto-esercizio: se l'ultimo valore fosse 32.4 invece di 15.8, quanto sarebbe la mediana? Se avete pensato che la mediana non cambia, siete davvero dei super super cool ed in più avrete notato quanto la mediana oscilli poco rispetto alle variazioni dei valori estremi…questo concetto tornerà utile quando affronteremo campioni con distribuzioni non-normali!

Sporchiamoci le mani…

Vediamo ora come calcolare con RStudio tutti gli indici spiegati fino ad ora. Apriamo il nostro affascinante programmino RStudio (o R, se volete essere coraggiosi), e inseriamo i nostri valori di testosterone. Come? facile.it, facile.it, facile.it… Chiamiamo la serie dei nostri numeri con un nome di fantasia, diciamo "testosterone"[6], poi inseriamo una freccina "<-", poi una "c" seguita da una coppia di parentesi in cui incollerete i nostri valori:

[6] Ma potrebbe essere tranquillamente "GiorgioMastrota" o "PeregrinoTuc"

```
testosterone<-c(10.2, 2.0, 3.8, 2.7, 2.9, 5.1, 1.0, 3.5, 4.1, 15.8)
```

Da questo momento, possiamo conoscere tutto quanto della nostra serie di valori. Media e mediana:

```
mean(testosterone)
```

```
## [1] 5.11
```

```
median(testosterone)
```

```
## [1] 3.65
```

Ma con il comando summary possiamo avere **at glance** media, mediana, minimo, massimo, 1° e 3° quartile (rispettivamente il valore al 25% e 75% che abbiamo calcolato precedentemente)

```
summary(testosterone)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.
##    1.00    2.75    3.65    5.11    4.85
##    Max.
##   15.80
```

Deviazione standard e varianza:

```
sd(testosterone)
```

```
## [1] 4.506403
```

```
var(testosterone)
```

```
## [1] 20.30767
```

Range interquartile:

```
IQR(testosterone)
```

```
## [1] 2.1
```

Per l'intervallo di confidenza, la storia si complica un pochino (ma neanche tanto) poiché non esiste nessuna formula one-shot, ma dobbiamo calcolarlo a mano come fatto in precedenza. Diamo un nome alla media con lo stesso procedimento di prima (freccina):

```
media<-mean(testosterone)
```

poi diamo un nome all'intervallo, cioè all'ES (ovvero deviazione standard/radice quadrata di n) moltiplicato per il coefficiente t al 95% (calcolabile con la formula qt) per gradi di libertà pari a n-1 (ovvero 9):

```
intervallo<- qt(0.95,df=9)*sd(testosterone)/sqrt(10)
intervallo
```

```
## [1] 2.612277
```

A questo punto possiamo ricavere il limite inf e quello sup:

```
inf<-media-intervallo
inf
```

```
## [1] 2.497723
```

```
sup<-media+intervallo
sup
```

```
## [1] 7.722277
```

Ultimo esempio, quello del calcolo della frequenza quando stiamo lavorando con variabili categoriche. Poniamo l'esempio del numero precedente in cui abbiamo osservato il numero di cellule vive ("V") vs morte ("M") in un campo di immunofluorescenza

```
cellule<-c("V","V","M","V","V","M","V","V","M","V","V","M","M","V","V","V")
```

Se volessimo sapere il numero assoluto, basterà il comando “table”

```
table(cellule)
```

```
## cellule
##  M  V
##  5 11
```

Altrimenti, per la frequenza e la frequenza percentuale, basterà procedere così

```
prop.table(table(cellule))
```

```
## cellule
##      M      V
## 0.3125 0.6875
```

```
prop.table(table(cellule))*100
```

```
## cellule
##     M     V
## 31.25 68.75
```

Epilogo

SI CONCLUDE COSÌ il secondo appuntamento del secondo appuntamento di “Keep Calm and EnGioI Statistics” In questo “episodio” abbiamo passato in rassegna i principali elementi che serviranno a descrivere i nostri dati e come fare ad ottenerli con R. Nel prossimo incontro parleremo appunto di come usare R per ulteriori nobili scopi: ad esempio, la tecnica per mettersi il mascara senza aprire la bocca e la creazione di un timbro per applicare una macchia di pupù sulle mutande per un simpatico regalo carnevalesco. Le risposte a questi ed altri interessanti quesiti vi aspettano nel prossimo episodio di “Keep Calm and EnGioI Statistics”.