Keep Calm and EnGioI Statistics - N°2.3

Andrea Sansone & Angelo Cignarelli

Giugno 2017

> "Dimmi, quel cervello che mi hai portato era di Hans Delbrück?" "No." "Ah! Be'… Ehm, ti dispiacerebbe dirmi di chi era il cervello che gli ho messo dentro?" "Non si arrabbierà, eh?" "No, io non mi arrabbierò!" "A.B. qualcosa…" "A.B. qualcosa? A.B chi?" "A.B… Norme"
>
> — Frankenstein Junior

Introduzione al capitolo 2.3

Bentornati a *Keep Calm and EnGioI Statistics*, la prima newsletter di statistica a ridotto contenuto calorico. Nelle scorse puntate avete (in teoria) installato R dopo aver capito che la statistica non è il male assoluto; inoltre, avete imparato le differenze fra le diverse tipologie di variabili e alcuni importanti parametri statistici come, ad esempio, la media, la mediana ed il range interquartile.

Ora manca un ultimo piccolo sforzo per completare questo capitolo fondamentale, ovvero, il capitolo col quale cercheremo di porre le fondamenta su cui si baserà tutto il resto. Dopo aver cercato di spiegare come si possono sintetizzare i nostri campioni con i vari indici di stima e dispersione, ci manca sapere se la distribuzione dei nostri valori è normale oppure no perché da questa analisi dipenderà la scelta dei test statistici idonei per verificare le nostre ipotesi. Quindi, alla fine di questo capitolo, speriamo di lasciarvi un bisogno, quello di chiedervi di analizzare la distribuzione di una serie di numeri (i.e. i soliti valori di testosterone dei capitoli precedenti) e chiedervi:

MA TI SEMBRA NORMALE?

Max Gazzè *docet*!

Distribuzione normale

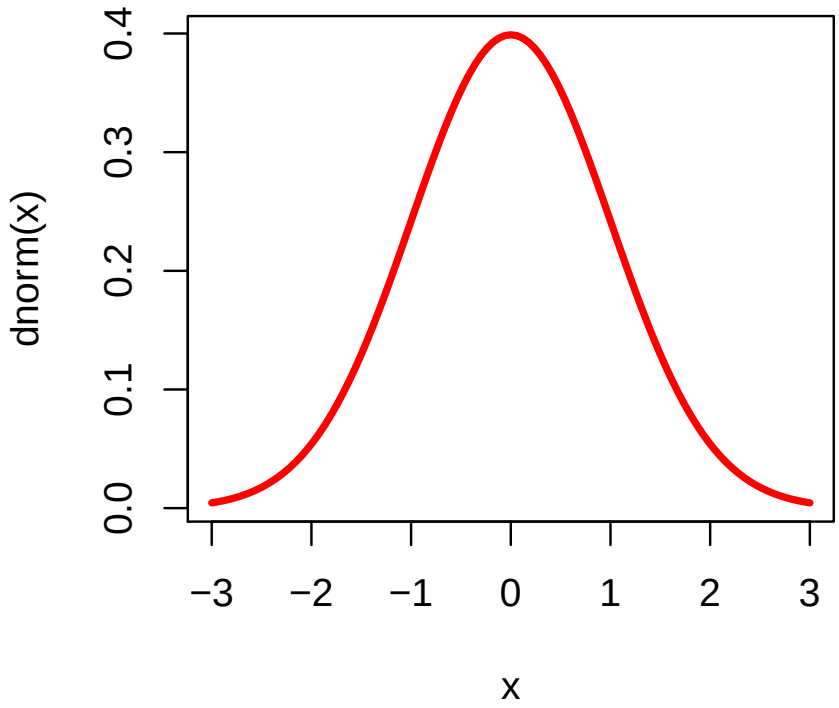
Ovviamente in questa newsletter non parleremo di teoremi centrali del limite, di funzioni gaussiane, di skewness o di kurtosis[1], ma spiegheremo sinteticamente cos'è una distribuzione normale e fornire alcuni piccoli e validi strumenti per verificare se i numeri in nostro possesso sono distribuiti normalmente. Ricordate, ogni volta che usarete un t-test senza sapere se i valori siano distribuiti normalmente, la tomba di Gauss sprofonda di un metro sotto terra.

[1] per questi concetti possiamo consigliarvi un'infinità di testi di statistica: ad esempio i libri della serie *Discovering Statistics* del prof Andy Field sono molto chiari, pieni di *british humor*, gatti e di citazioni heavy metal

Cos'è la distribuzione normale

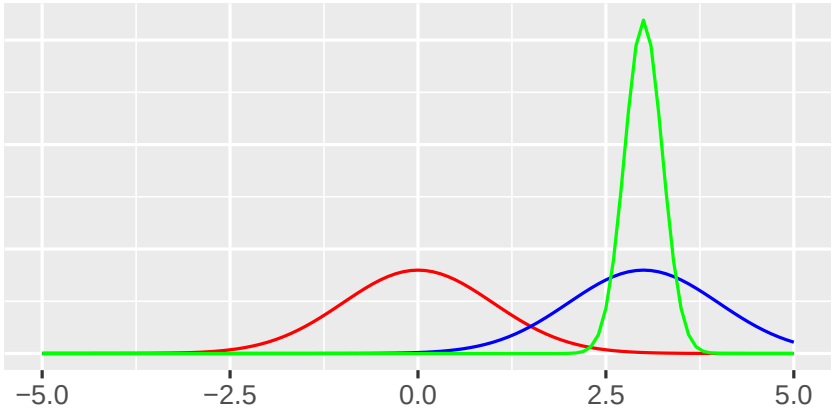
Innanzitutto la distribuzione normale è un concetto che ha che fare con la probabilità. La distribuzione normale è fondamentalmente una *distribuzione di probabilità* che risponde ad alcune specifiche caratteristiche:

- le misure di tendenza centrale (media e mediana) coincidono
- ha una forma a campana e simmetrica



Esistono infinite curve di distribuzione normale, ognuna con una sua media e una sua deviazione standard (rispettivamente $\bar{x}$ e σ). Sapendo a quanto corrispondono questi due parametri, è possibile disegnare la curva: la media indica la posizione sull'asse delle X, mentre la deviazione standard ne indica la forma.

Ricorriamo ad un esempio "pratico" di cosa significhi quest'ultima frase. Il codice ve lo risparmiamo per il momento, limitatevi a guardare l'immagine…



Bene, quelle qui sopra sono tre distribuzioni normali. Iniziamo ad osservare la curva rossa: ha media $\bar{x} = 0$ e sd $\sigma = 1$. La media potete intuirla dall'immagine, mentre la sd date per assodato che sia così. Guardate ora la curva blu: notate come ha la stessa "forma" della curva rossa? Questo accade perchè sd è sovrapponibile: quindi il cambiamento della media (che in questo caso è $\bar{x} = 3$) ha comportato uno spostamento lungo l'asse delle X. Ora concentriamoci sulla curva verde: la media è sempre 3, ma questa volta la deviazione standard ? $\sigma = 0.25$. In questo caso potete intuire come essendo minore la deviazione standard la curva risulterà più "a punta" rispetto alle precedenti osservazioni.

Vi domanderete: ma perchè la curva normale ha tutto questo peso in medicina[2]? La grande importanza della distribuzione normale deriva dal fatto che essendo tutte le curve derivate da due soli parametri è possibile calcolare esattamente le probabilità di un dato evento in relazione alla sua distanza dalla media. Difatti, in **tutte** le curve di normalità, il 95% delle osservazioni cade nello "spazio" compreso fra $\bar{x} - 1.96 * \sigma$ e $\bar{x} + 1.96 * \sigma$, il 99% $\bar{x} - 2.57 * \sigma$ e $\bar{x} + 2.57 * \sigma$. Giusto per completezza, aggiungiamo che nello spazio compreso fra $\bar{x} - \sigma$ e $\bar{x} + \sigma$ cade il 68% delle osservazioni.

[2] O più generalmente in tutta la statistica?

Che significa tutto questo in termini utili? Per fare un esempio, sapendo che la media delle altezze di una determinata popolazione è di 175 cm e che la deviazione standard è 10 cm, possiamo calcolare che il 95% delle persone avrà un'altezza compresa fra 155.4 cm (175-19.6) e 194.6 cm (175 + 19.6). Questi dati saranno poi fondamentali per l'inferenza statistica, cioè per quella parte della statistica che a partire da un campione cerca di ottenere dei dati che siano generalizzabili all'intera popolazione.

Il senso della statistica inferenziale non è particolarmente difficile da capire se prendete come esempio l'emocromo: per sapere quant'è l'ematocrito di una persona potete prendere un campione di sangue e tramite un calcolo inferenziale generalizzarlo all'intero soggetto… oppure dissanguare completamente il povero paziente e fare un calcolo sicuramente più preciso, ma che vi servirà a poco dinanzi all'accusa di omicidio…

La cosa comoda da considerare è anche che in teoria, sapendo che il 95% della popolazione cade entro una certa distanza dalla media (distanza che, come appena scritto sopra, deriva dalla deviazione standard), possiamo ottenere dei dati importanti su quel 5% di soggetti che cade oltre questo limite… Ossia i soggetti *anormali*. Ed ecco

che anche da un punto di vista grammaticale questa parola inizia ad assumere un senso compiuto.

Benissimo, a questo punto vi starete (probabilmente) chiedendo: "Ma se ho una distribuzione e voglio sapere se è normale, come posso fare? Posso chiamare Max Gazzè e chiedere se gli sembra normale?"[3] - e chiaramente c'è una risposta molto più semplice.

[3] Pensate, una battuta orrenda come questa ripetuta due volte nello stesso capitolo!

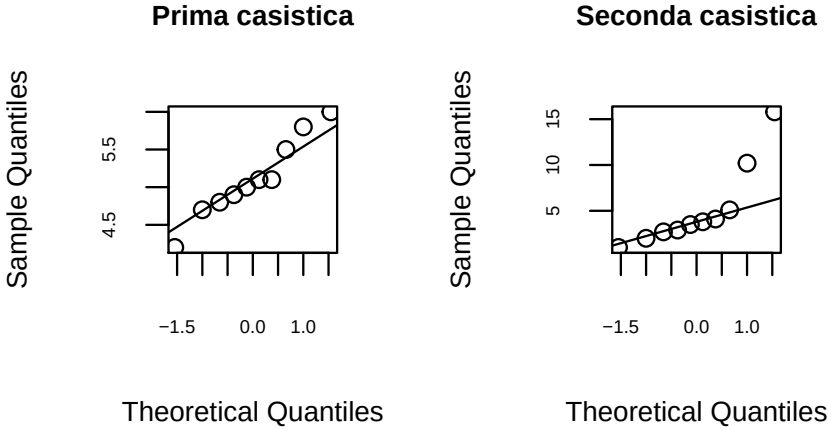
Vediamo come valutare se i nostri valori presentano una distribuzione normale. Abbiamo due possibilità:

- una modalità "grafica"
- una modalità "statistica"

La modalità grafica prevede l'impiego di uno strumento, il QQ plot[4], che confronta la distribuzione cumulata della variabile osservata con la distribuzione cumulata della normale: in parole povere, se i nostri dati presentano una distribuzione normale, i punti di questa distribuzione si addensano sulla diagonale che va dal basso verso l'alto e da sinistra verso destra. Più vicini saranno i punti alla linea, più la nostra distribuzione sarà normale. Questa metodica, sebbene molto "visual" e di impatto, offre il fianco alla soggettività dell'interpretazione non fornendo un valore numerico oggettivo.

[4] Dove QQ sta per quantile-quantile, oltre che a rappresentare due occhi che piangono.

Prendiamo ad esempio le due casistiche di testosterone del numero precedente.

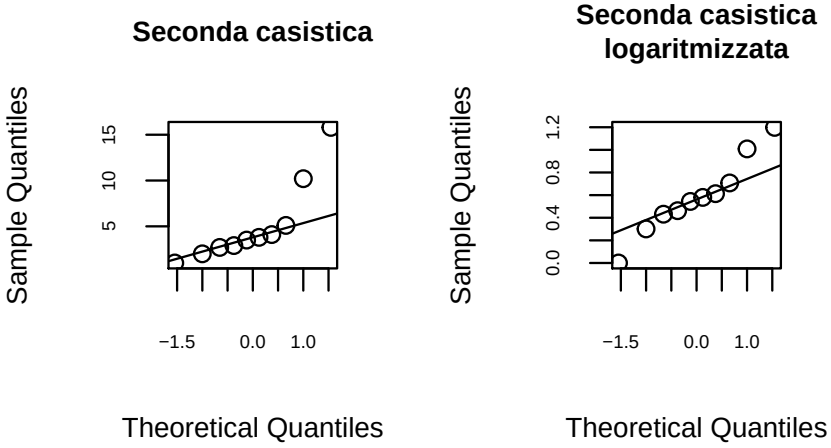


Come noterete, la seconda casistica (che peraltro mostrava anche una deviazione standarda particolarmente ampia) presenta dei valori che si discostano molto dalla "normalità", molto più che nella distribuzione della prima casistica, dove i valori sono molto più vicini alla diagonale.Tuttavia, come affermato in precedenza, con questa modalità non abbiamo elementi solidi per affermare che la distribuzione non/sia normale e, pertanto, abbiamo bisogno di un test. La modalità statistica infati prevede invece un vero e proprio test per cimentare la nostra ipotesi nulla, ovvero che tra la nostra distribuzione e quella normale

non ci sia una differenza. Uno di questi test è lo Shapiro-Wilk[5]; come ogni buon test che si rispetti, anche in questo caso ci aspettiamo un p-value che, in questo caso, se viene < 0.05 vuol dire che l'ipotesi nulla viene rifiutata è che la distribuzione non è normale. Vediamo le due casistiche precedenti, già analizzate graficamente con il Q-Q plot. Nella prima casistica, il p-value è 0.8232963 mentre nella seconda casistica è 0.0062376[6]. Quindi, possiamo affermare che la prima casistica (come intuito dal grafico) presenta una distribuzione normale (visto che la p è > 0.05); al contrario, la seconda ha una distribuzione non normale. In questi casi, una trasformazione logaritmica dei nostri dati può determinare un "normalizzazione" della distribuzione. Proviamo con la nostra seconda casistica.

[5] Per la verità ce ne sono tantissimi: D'agostino's K-squared test, Jarque-Bera test, Anderson-Darling test, Cramer-von Mises normality test, Lilliefors (Kolmogorov-Smirnov) test, Shapiro-Francia, Pearson chi-square, Kolmogorov-Smirnov test

[6] Per calcolarli, abbiamo usato la funzione `shapiro.test()` che, come dice il nome, serve ad applicare la funzione di normalità di Shapiro-Wilk: un p-value <.05 indica che la distribuzione ***non*** è normale



Come si osserva dai grafici, si può notare come i dati trasformati siano molto più vicini alla retta. Infatti il test di Shapiro ci dice che la p non è più significativa:

```
[1] 0.8444329
```

e perciò la distribuzione è diventata normale. In alcuni casi, però, la trasformazione non è possibile (pensate agli zero che non consentono di ottenere un logaritmo) o non determina una normalizzazione. Non disperiamo. In fondo non è fondamentale che la distribuzione sia per forza normale, ma è importante saperlo per non usare test statistici impropriamente. L'ultimo punto per chiudere questo capitolo è sapere che in caso di distribuzioni non-normali è più appropriato indicare la mediana e l'indice interquartile, mentre la media e la deviazione standard o intervalli di confidenza[7] per le distribuzioni normali.

[7] tutta roba spiegata nel numero 2.2

Sporchiamoci le mani…

Nella puntate precedenti vi è stato spiegato come installare R e RStudio e abbiamo cominciato a muovere i primi passi di bimbo. Imparare R è come imparare una lingua, serve coraggio di sbagliare, pratica

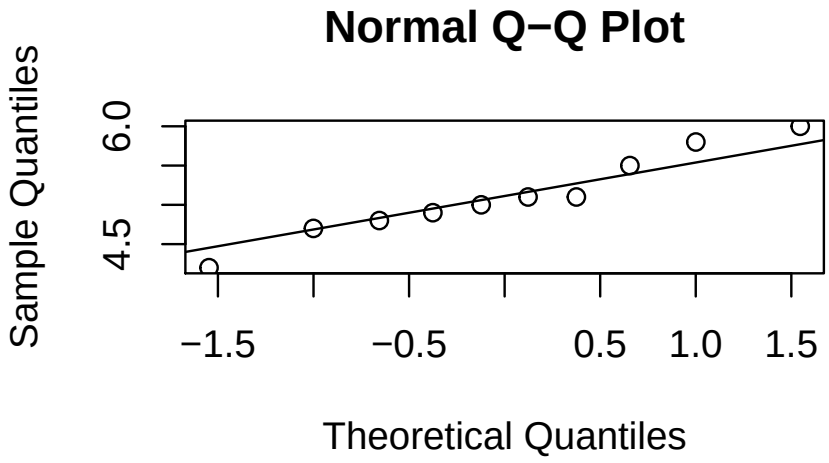
costante e curiosità. All'inizio sembrerà ostico, così come sarebbe ostico sentir parlare per la prima volta una lingua sconosciuta. All'inizio servirà spesso servirsi di tutorial, di Google e di forum, ma una volta superato lo scoglio iniziale (possono volerci mesi…non disperate), si potrà ottenere tutto quello che si può realizzare con software costosi e blasonati, con il vantaggio di analizzare i dati con molta più consapevolezza e velocità. Ora come fare tutte queste belle cose con R. Innanzitutto, se non lo avete già fatto, installare R e RStudio. Fatelo ora altrimenti diventerà sempre più difficile seguire queste newsletter ed in generale vivere[8]. Riprendiamo le nostre casistiche utilizzate in precedenza e rendiamole utilizzabili per alcuni test. A ciascuna di esse, attribuiamo un nome di fantasia che può davvero essere qualunque ("x", "y" o "Dart Fener") con il trucchetto della freccetta seguita dalla c e dai numeri inclusi nelle parentesi tonde.

[8] *R è amore, R è vita.*

```
primaCasistica<-c( 4.2, 5.0, 4.8, 4.7, 4.9, 5.1, 6.0, 5.5, 5.1, 5.8)
secondaCasistica<-c(10.2, 2.0, 3.8, 2.7, 2.9, 5.1, 1.0, 3.5, 4.1,15.8)
```

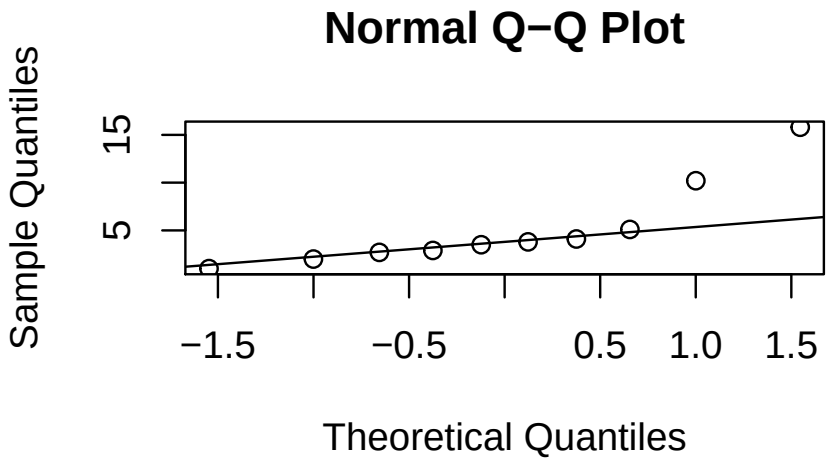
Proviamo quindi ad eseguire i Q-Q plot con il comando `qqnorm` seguito da `qqline` sia per la prima casistica

```
qqnorm(primaCasistica)
qqline(primaCasistica)
```



che per la seconda

```
qqnorm(secondaCasistica)
qqline(secondaCasistica)
```



Dopo aver sbirciato i grafici, non ci resta che effettuare i test di Shapiro-Wilk con il semplice comando shapiro.test

```
shapiro.test(primaCasistica)
```

```
    Shapiro-Wilk normality test

data:  primaCasistica
W = 0.96335, p-value = 0.8233
```

```
shapiro.test(secondaCasistica)
```

```
    Shapiro-Wilk normality test

data:  secondaCasistica
W = 0.76991, p-value = 0.006238
```

Epilogo

SI CONCLUDE COSÌ il terzo ed ultimo appuntamento del secondo capitolo di "*Keep Calm and EnGioI Statistics*" In questo "episodio" abbiamo passato in rassegna un passaggio fondamentale per conoscere la normalità delle distribuzioni dei nostri campioni da cui dipenderanno le scelte dei test statistici. Nel prossimo incontro parleremo appunto di come usare R per ulteriori nobili scopi: ad esempio, come riuscire a non rimanere single nonostante non si conosca la serie TV "Games of Thrones" e come riuscire a non addormentarsi dopo il primo giro della formula 1. Le risposte a questi ed altri interessanti quesiti vi aspettano nel prossimo episodio di "*Keep Calm and EnGioI Statistics*".