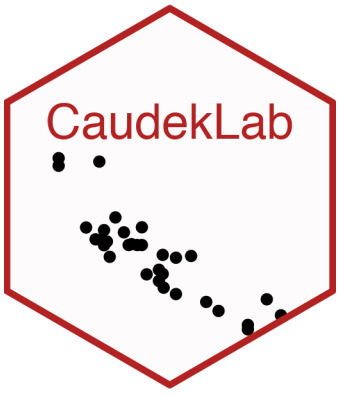
Psicometria

Corrado Caudek

2021-11-25

Questo documento è stato realizzato con:

- LaTeX e la classe memoir (http://www.ctan.org/pkg/memoir);
- R (http://www.r-project.org/) e RStudio (http://www.rstudio.com/);
- bookdown (http://bookdown.org/) e memoiR (https://ericmarcon.github.io/memoiR/).



Nel blog della mia pagina personale sono forniti alcuni approfondimenti degli argomenti qui trattati.

https://ccaudek.github.io/caudeklab/

Indice

Data della versione presente: Novembre 25, 2021.

Prefazione

Data Science per psicologi contiene il materiale delle lezioni dell'insegnamento di *Psicometria B000286* (A.A. 2021/2022) rivolto agli studenti del primo anno del Corso di Laurea in Scienze e Tecniche Psicologiche dell'Università degli Studi di Firenze.

L'insegnamento di Psicometria si propone di fornire agli studenti un'introduzione all'analisi dei dati in psicologia. Le conoscenze/competenze che verranno sviluppate in questo insegnamento sono quelle della *Data science*, ovvero le conoscenze/competenze che si pongono all'intersezione tra statistica (ovvero, richiedono la capacità di comprendere teoremi statistici) e informatica (ovvero, richiedono la capacità di sapere utilizzare un software).

La psicologia e la Data Science

> It's worth noting, before getting started, that this material is hard. If you find yourself confused at any point, you are normal. Any sense of confusion you feel is just your brain correctly calibrating to the subject matter. Over time, confusion is replaced by comprehension [...] — Richard McElreath

Sembra sensato spendere due parole su un tema che è importante per gli studenti: quello indicato dal titolo di questo Capitolo. È ovvio che agli studenti di psicologia la statistica non piace. Se piacesse, forse studierebbero Data Science e non psicologia; ma non lo fanno. Di conseguenza, gli studenti di psicologia si chiedono: "perché dobbiamo perdere tanto tempo a studiare queste cose quando in realtà quello che ci interessa è tutt'altro?'' Questa è una bella domanda.

C'è una ragione molto semplice che dovrebbe farci capire perché la Data Science è così importante per la psicologia. Infatti, a ben pensarci, la psicologia è una disciplina intrinsecamente statistica, se per statistica intendiamo quella disciplina che studia la variazione delle caratteristiche degli individui nella popolazione. La psicologia studia *gli individui* ed è proprio la variabilità inter- e intra-individuale ciò che vogliamo descrivere e, in certi casi, predire. In questo senso, la psicologia è molto diversa dall'ingegneria, per esempio. Le proprietà di un determinato ponte sotto certe condizioni, ad esempio, sono molto simili a quelle di un altro ponte, sotto le medesime condizioni. Quindi, per un ingegnere la statistica è poco importante: le proprietà dei materiali sono unicamente dipendenti dalla loro composizione e restano costanti. Ma lo stesso non può dirsi degli individui: ogni individuo è unico e cambia nel tempo. E le variazioni tra gli individui, e di un individuo nel tempo, sono l'oggetto di studio proprio della psicologia: è dunque chiaro che i problemi che la psicologia si pone sono molto diversi da quelli affrontati, per esempio, dagli ingegneri. Questa è la ragione per cui abbiamo tanto bisogno della *data science* in psicologia: perché la *data science* ci consente di descrivere la variazione e il cambiamento. E queste sono appunto le caratteristiche di base dei fenomeni psicologici.

Sono sicuro che, leggendo queste righe, a molti studenti sarà venuta in mente la seguente domanda: perché non chiediamo a qualche esperto di fare il "lavoro sporco" (ovvero le analisi statistiche) per noi, mentre noi (gli psicologi) ci occupiamo solo di ciò che ci interessa, ovvero dei problemi psicologici slegati dai dettagli "tecnici" della *data science*? La risposta a questa domanda è che non è possibile progettare uno studio psico-

logico sensato senza avere almeno una comprensione rudimentale della *data science*. Le tematiche della *data science* non possono essere ignorate né dai ricercatori in psicologia né da coloro che svolgono la professione di psicologo al di fuori dell'Università. Infatti, anche i professionisti al di fuori dall'università non possono fare a meno di leggere la letteratura psicologica più recente: il continuo aggiornamento delle conoscenze è infatti richiesto dalla deontologia della professione. Ma per potere fare questo è necessario conoscere un bel po' di *data science*! Basta aprire a caso una rivista specialistica di psicologia per rendersi conto di quanto ciò sia vero: gli articoli che riportano i risultati delle ricerche psicologiche sono zeppi di analisi statistiche e di modelli formali. E la comprensione della letteratura psicologica rappresenta un requisito minimo nel bagaglio professionale dello psicologo.

Le considerazioni precedenti cercano di chiarire il seguente punto: la *data science* non è qualcosa da studiare a malincuore, in un singolo insegnamento universitario, per poi poterla tranquillamente dimenticare. Nel bene e nel male, gli psicologi usano gli strumenti della *data science* in tantissimi ambiti della loro attività professionale: in particolare quando costruiscono, somministrano e interpretano i test psicometrici. È dunque chiaro che possedere delle solide basi di *data science* è un tassello imprescindibile del bagaglio professionale dello psicologo. In questo insegnamento verrano trattati i temi base della *data science* e verrà adottato un punto di vista bayesiano, che corrisponde all'approccio più recente e sempre più diffuso in psicologia.

Come studiare

> I know quite certainly that I myself have no special talent. Curiosity, obsession and dogged endurance, combined with self-criticism, have brought me to my ideas. — Albert Einstein

Il giusto metodo di studio per prepararsi all'esame di Psicometria è quello di seguire attivamente le lezioni, assimilare i concetti via via che essi vengono presentati e verificare in autonomia le procedure presentate a lezione. Incoraggio gli studenti a farmi domande per chiarire ciò che non è stato capito appieno. Incoraggio gli studenti a utilizzare i forum attivi su Moodle e, soprattutto, a svolgere gli esercizi proposti su Moodle. I problemi forniti su Moodle rappresentano il livello di difficoltà richiesto per superare l'esame e consentono allo studente di comprendere se le competenze sviluppate fino a quel punto sono sufficienti rispetto alle richieste dell'esame.

La prima fase dello studio, che è sicuramente individuale, è quella in cui è necessario acquisire le conoscenze teoriche relative ai problemi che saranno presentati all'esame. La seconda fase di studio, che può essere facilitata da scambi con altri e da incontri di gruppo, porta ad acquisire la capacità di applicare le conoscenze: è necessario capire come usare un software (R) per applicare i concetti statistici alla specifica situazione del problema che si vuole risolvere. Le due fasi non sono però separate: il saper fare molto spesso ci aiuta a capire meglio.

Sviluppare un metodo di studio efficace

> Memorization is not learning. — Richard Phillips Feynman

Avendo insegnato molte volte in passato un corso introduttivo di analisi dei dati ho notato nel corso degli anni che gli studenti con l'atteggiamento mentale che descriverò qui sotto generalmente ottengono ottimi risultati. Alcuni studenti sviluppano naturalmente questo approccio allo studio, ma altri hanno bisogno di fare uno sforzo per maturarlo. Fornisco qui sotto una breve descrizione del "metodo di studio' " che, nella mia esperienza, è il più efficace per affrontare le richieste di questo insegnamento (Burger & Starbird, 2012).

- Dedicate un tempo sufficiente al materiale di base, apparentemente facile; assicuratevi di averlo capito bene. Cercate le lacune nella vostra comprensione. Leggere presentazioni diverse dello stesso materiale (in libri o articoli diversi) può fornire nuove intuizioni.

- Gli errori che facciamo sono i nostri migliori maestri. Istintivamente cerchiamo di dimenticare subito i nostri errori. Ma il miglior modo di imparare è apprendere dagli errori che commettiamo. In questo senso, una soluzione corretta è meno utile di una soluzione sbagliata. Quando commettiamo un errore questo ci fornisce un'informazione importante: ci fa capire qual è il materiale di studio sul quale dobbiamo ritornare e che dobbiamo capire meglio.

- C'è ovviamente un aspetto "psicologico" nello studio. Quando un esercizio o problema ci sembra incomprensibile, la cosa migliore da fare è dire: "mi arrendo", "non ho idea di cosa fare!". Questo ci rilassa: ci siamo già arresi, quindi non abbiamo niente da perdere, non dobbiamo più preoccuparci. Ma non dobbiamo fermarci qui. Le cose "migliori" che faccio (se ci sono) le faccio quando non ho voglia di lavorare. Alle volte, quando c'è qualcosa che non so fare e non ho idea di come affontare, mi dico: "oggi non ho proprio voglia di fare fatica", non ho voglia di mettermi nello stato mentale per cui "in 10 minuti devo risolvere il problema perché dopo devo fare altre cose". Però ho voglia di *divertirmi* con quel problema e allora mi dedico a qualche aspetto "marginale" del problema, che so come affrontare, oppure considero l'aspetto più difficile del problema, quello che non so come risolvere, ma invece di cercare di risolverlo, guardo come altre persone hanno affrontato problemi simili, opppure lo stesso problema in un altro contesto. Non mi pongo l'obiettivo "risolvi il problema in 10 minuti", ma invece quello di farmi un'idea "generale" del problema, o quello di capire un caso più specifico e più semplice del problema. Senza nessuna pressione. Infatti, in quel momento ho deciso di non lavorare (ovvero, di non fare fatica). Va benissimo se "parto per la tangente", ovvero se mi metto a leggere del materiale che sembra avere poco a che fare con il problema centrale (le nostre intuizioni e la nostra curiosità solitamente ci indirizzano sulla strada giusta). Quando faccio così, molto spesso trovo la soluzione del problema che mi ero posto e, paradossalmente, la trovo in un tempo minore di quello che, in precedenza, avevo dedicato a "lavorare" al problema. Allora perché non faccio sempre così? C'è ovviamente l'aspetto dei "10 minuti" che non è sempre facile da dimenticare. Sotto pressione, possiamo solo agire in maniera automatica, ovvero possiamo solo applicare qualcosa che già sappiamo fare. Ma se dobbiamo imparare qualcosa di nuovo, la pressione è un impedimento.

- È utile farsi da soli delle domande sugli argomenti trattati, senza limitarsi a cercare di risolvere gli esercizi che vengono assegnati. Quando studio qualcosa mi viene in mente: "se questo è vero, allora deve succedere quest'altra cosa". Allora verifico se questo è vero, di solito con una simulazione. Se i risultati della simulazione sono quelli che mi aspetto, allora vuol dire che ho capito. Se i risultati sono diversi da quelli che mi aspettavo, allora mi rendo conto di non avere capito e ritorno indietro a studiare con più attenzione la teoria che pensavo di avere capito – e ovviamente mi rendo conto che c'era un aspetto che avevo frainteso. Questo tipo di verifica è qualcosa che dobbiamo fare da soli, in prima persona: nessun altro può fare questo al posto nostro.

- Non aspettatevi di capire tutto la prima volta che incontrate un argomento nuovo.[1] È utile farsi una nota mentalmente delle lacune nella vostra comprensione e tornare su di esse in seguito per carcare di colmarle. L'atteggiamento naturale, quando non capiamo i dettagli di qualcosa, è quello di pensare: "non importa, ho capito

[1] Ricordatevi inoltre che gli individui tendono a sottostimare la propria capacità di apprendere (Horn & Loewenstein, 2021).

in maniera approssimativa questo punto, non devo preoccuparmi del resto". Ma in realtà non è vero: se la nostra comprensione è superficiale, quando il problema verrà presentato in una nuova forma, non riusciremo a risolverlo. Per cui i dubbi che ci vengono quando studiamo qualcosa sono il nostro alleato più prezioso: ci dicono esattamente quali sono gli aspetti che dobbiamo approfondire per potere migliorare la nostra preparazione.

- È utile sviluppare una visione d'insieme degli argomenti trattati, capire l'obiettivo generale che si vuole raggiungere e avere chiaro il contributo che i vari pezzi di informazione forniscono al raggiungimento di tale obiettivo. Questa organizzazione mentale del materiale di studio facilita la comprensione. È estremamente utile creare degli schemi di ciò che si sta studiando. Non aspettate che sia io a fornirvi un riepilogo di ciò che dovete imparare: sviluppate da soli tali schemi e tali riassunti.
- Tutti noi dobbiamo imparare l'arte di trovare le informazioni, non solo nel caso di questo insegnamento. Quando vi trovate di fronte a qualcosa che non capite, o ottenete un oscuro messaggio di errore da un software, ricordatevi: "Google is your friend".

Corrado Caudek

Febbraio 2022

CAPITOLO 1

Distribuzioni di v.c. continue

Dopo avere introdotto con una simulazione il concetto di funzione di densità nel Capitolo **??**, prendiamo ora in esame alcune delle densità di probabilità più note. La più importante di esse è sicuramente la distribuzione Normale.

1.1 Distribuzione Normale

Non c'è un'unica distribuzione Normale, ma ce ne sono molte. Tali distribuzioni sono anche dette "gaussiane" in onore di Carl Friedrich Gauss (uno dei più grandi matematici della storia il quale, tra le altre cose, scoprì l'utilità di tale funzione di densità per descrivere gli errori di misurazione). Adolphe Quetelet, il padre delle scienze sociali quantitative, fu il primo ad applicare tale densità alle misurazioni dell'uomo. Karl Pearson usò per primo il termine "distribuzione Normale" anche se ammise che questa espressione "ha lo svantaggio di indurre le persone a credere che le altre distribuzioni, in un senso o nell'altro, non siano normali."

Limite delle distribuzioni binomiali

Iniziamo con un un breve excursus storico. Nel 1733, Abraham de Moivre notò che, aumentando il numero di prove in una distribuzione binomiale, la distribuzione risultante diventava quasi simmetrica e a forma campanulare. Con 10 prove e una probabilità di successo di 0.9 in ciascuna prova, la distribuzione è chiaramente asimmetrica.

```
N <- 10
x <- 0:10
y <- dbinom(x, N, 0.9)
binomial_limit_plot <-
  tibble(x = x, y = y) %>%
  ggplot(aes(x = x, y = y)) +
  geom_bar(
    stat = "identity", color = "black", size = 0.2
  ) +
  xlab("y") +
  scale_x_continuous(breaks = c(0, 2, 4, 6, 8, 10)) +
  ylab("Binomial(y | 10, 0.9)")
binomial_limit_plot
```

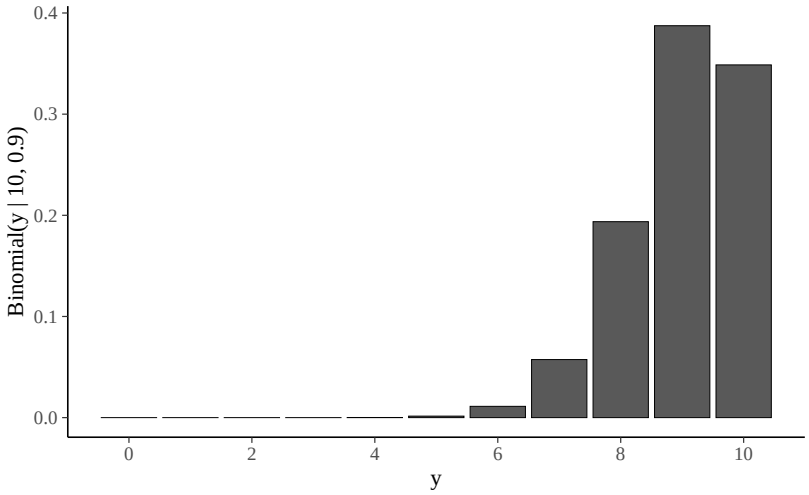


Figura 1.1: *Probabilità del numero di successi in* $N = 10$ *prove bernoulliane indipendenti, ciascuna con una probabilità di successo di 0.90. Il risultato è una distribuzione* $\text{Bin}(y \mid 10, 0.9)$. *Con solo dieci prove, la distribuzione è fortemente asimmetrica negativa.*

Ma se aumentiamo il numero di prove di un fattore di 100 a $N = 1000$, senza modificare la probabilità di successo di 0.9, la distribuzione assume una forma campanulare quasi simmetrica. Dunque, de Moivre scoprì che, quando N è grande, la funzione Normale (che introdurremo qui sotto), nonostante sia la densità di v.a. continue, fornisce una buona approssimazione alla funzione di massa di probabilità binomiale.

```
N <- 1000
x <- 0:1000
y <- dbinom(x, N, 0.9)
binomial_limit_plot <-
  tibble(x = x, y = y) %>%
  ggplot(aes(x = x, y = y)) +
  geom_bar(stat = "identity", color = "black", size = 0.2) +
  xlab("y") +
  # scale_x_continuous(breaks = c(0, 2, 4, 6, 8, 1000)) +
  ylab("Binomial(y | 1000, 0.9)") +
  xlim(850, 950)
binomial_limit_plot
```

La distribuzione Normale fu scoperta da Gauss nel 1809 e, storicamente, è intimamente legata al metodo dei minimi quadrati – si veda l'Appendice **??**. Il Paragrafo successivo illustra come si possa giungere alla Normale mediante una simulazione.

1.2 La Normale prodotta con una simulazione

McElreath (2020) presenta un esempio che illustra come sia possibile giungere alla distribuzione Normale mediante una simulazione. Supponiamo che vi siano mille persone tutte allineate su una linea di partenza. Quando viene dato un segnale, ciascuna persona lancia una moneta e fa un passo in avanti oppure all'indietro a seconda che sia uscita testa o croce. Supponiamo che la lunghezza di ciascun passo vari da 0 a 1 metro. Ciascuna persona lancia una moneta 16 volte e dunque compie 16 passi.

Alla conclusione di queste passeggiate casuali (*random walk*) non possiamo sapere con esattezza dove si troverà ciascuna persona, ma possiamo conoscere con certezza le caratteristiche della distribuzione delle mille distanze dall'origine. Per esempio, possiamo predire in maniera accurata la proporzione di persone che si sono spostate in avanti

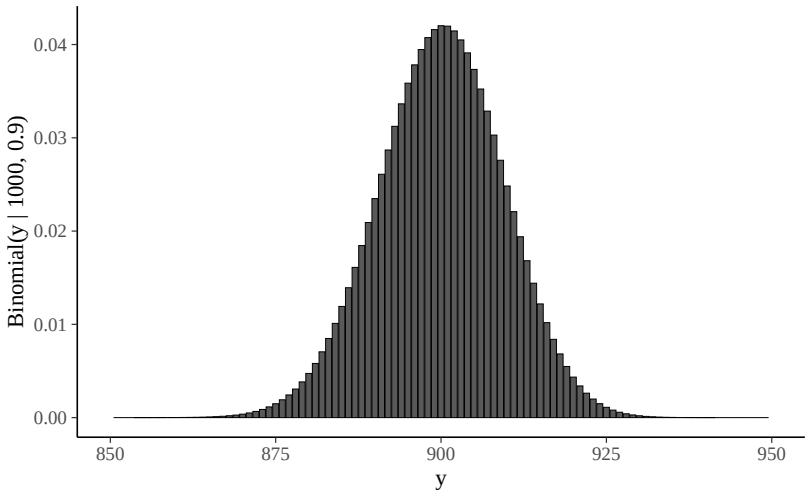


Figura 1.2: *Probabilità del numero di successi in* $N = 1000$ *prove bernoulliane indipendenti, ciascuna con una probabilità di successo di 0.90. Il risultato è una distribuzione* $\text{Bin}(y \mid 1000, 0.9)$. *Con mille prove, la distribuzione è quasi simmetrica a forma campanulare.*

oppure all'indietro. Oppure, possiamo predire accuratamente la proporzione di persone che si troveranno ad una certa distanza dalla linea di partenza (es., a 1.5 m dall'origine).

Queste predizioni sono possibili perché tali distanze si distribuiscono secondo la legge Normale. È facile simulare questo processo usando R. I risultati della simulazione sono riportati nella figura 1.3.

```
set.seed(4)
pos <-
  replicate(100, runif(16, -1, 1)) %>%
  as_tibble() %>%
  rbind(0, .) %>%
  mutate(step = 0:16) %>%
  gather(key, value, -step) %>%
  mutate(person = rep(1:100, each = 17)) %>%
  group_by(person) %>%
  mutate(position = cumsum(value)) %>%
  ungroup()

ggplot(
  data = pos,
  aes(x = step, y = position, group = person)
) +
  geom_vline(xintercept = c(4, 8, 16), linetype = 2) +
  geom_line(aes(color = person < 2, alpha = person < 2)) +
  scale_color_manual(values = c("gray", "black")) +
  scale_alpha_manual(values = c(1 / 5, 1)) +
  scale_x_continuous(
    "Numero di passi",
    breaks = c(0, 4, 8, 12, 16)
  ) +
  labs(y = "Posizione") +
  theme(legend.position = "none")
```

Un kernel density plot delle distanze ottenute dopo 4, 8 e 16 passi è riportato nella figura 1.4. Nel pannello di destra, al kernel density plot è stata sovrapposta una densità

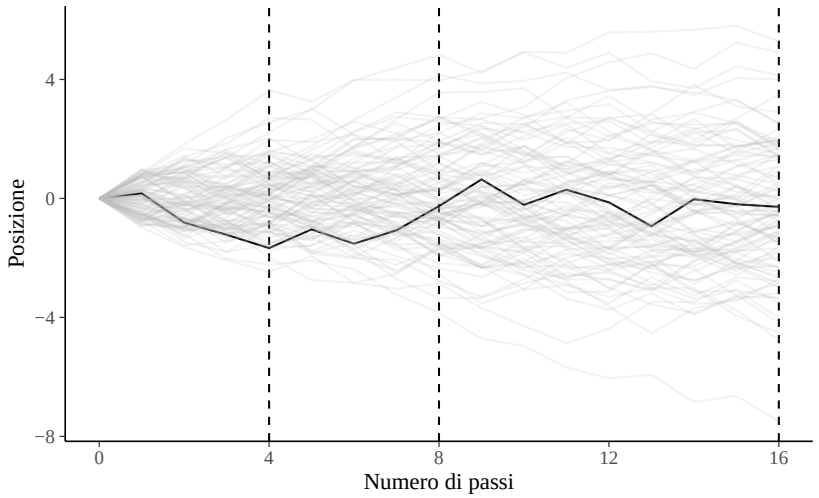


Figura 1.3: *Passeggiata casuale di 4, 8 e 16 passi. La spezzata nera indica la media delle distanze dall'origine come funzione del numero di passi.*

Normale di opportuni parametri (linea tratteggiata).

```
p1 <-
  pos %>%
  filter(step == 4) %>%
  ggplot(aes(x = position)) +
  geom_line(stat = "density", color = "black") +
  labs(title = "4 passi")

p2 <-
  pos %>%
  filter(step == 8) %>%
  ggplot(aes(x = position)) +
  geom_density(color = "black", outline.type = "full") +
  labs(title = "8 passi")

sd <-
  pos %>%
  filter(step == 16) %>%
  summarise(sd = sd(position)) %>%
  pull(sd)

p3 <-
  pos %>%
  filter(step == 16) %>%
  ggplot(aes(x = position)) +
  stat_function(
    fun = dnorm,
    args = list(mean = 0, sd = sd),
    linetype = 2
  ) +
  geom_density(color = "black", alpha = 1 / 2) +
  labs(
    title = "16 passi",
    y = "Densità"
  )
```

```
(p1 | p2 | p3) & coord_cartesian(xlim = c(-6, 6))
```

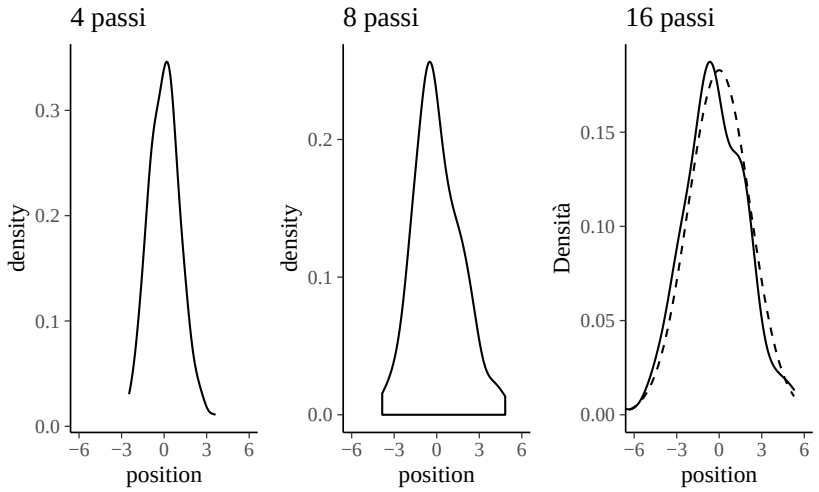


Figura 1.4: *Kernel density plot dei risultati della passeggiata casuale riportata nella figura precente, dopo 4, 8 e 16 passi. Nel pannello di destra, una densità Normale di opportuni parametri è sovrapposta all'istogramma lisciato.*

Questa simulazione mostra che qualunque processo nel quale viene sommato un certo numero di valori casuali, tutti provenienti dalla medesima distribuzione, converge ad una distribuzione Normale. Non importa quale sia la forma della distribuzione di partenza: essa può essere uniforme, come nell'esempio presente, o di qualunque altro tipo. La forma della distribuzione da cui viene realizzato il campionamento determina la velocità della convergenza alla Normale. In alcuni casi la convergenza è lenta; in altri casi la convergenza è molto rapida (come nell'esempio presente).

Da un punto di vista formale, diciamo che una variabile casuale continua Y ha una distribuzione Normale se la sua densità è

$$f(y; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}, \tag{1.1}$$

dove $\mu \in \mathbb{R}$ e $\sigma > 0$ sono i parametri della distribuzione.

La densità normale è unimodale e simmetrica con una caratteristica forma a campana e con il punto di massima densità in corrispondenza di μ.

Il significato dei parametri μ e σ che appaiono nella (1.1) viene chiarito dalla dimostrazione che

$$\mathbb{E}(X) = \mu, \qquad \text{Var}(X) = \sigma^2. \tag{1.2}$$

Esempio 1.1. La rappresentazione grafica di quattro densità Normali tutte con media 0 e con deviazioni standard 0.25, 0.5, 1 e 2 è fornita nella figura 1.5.

Concentrazione

È istruttivo osservare il grado di concentrazione della distribuzione Normale attorno alla media:

$$\begin{aligned} P(\mu - \sigma < X < \mu + \sigma) &= P(-1 < Z < 1) \simeq 0.683, \\ P(\mu - 2\sigma < X < \mu + 2\sigma) &= P(-2 < Z < 2) \simeq 0.956, \\ P(\mu - 3\sigma < X < \mu + 3\sigma) &= P(-3 < Z < 3) \simeq 0.997. \end{aligned}$$

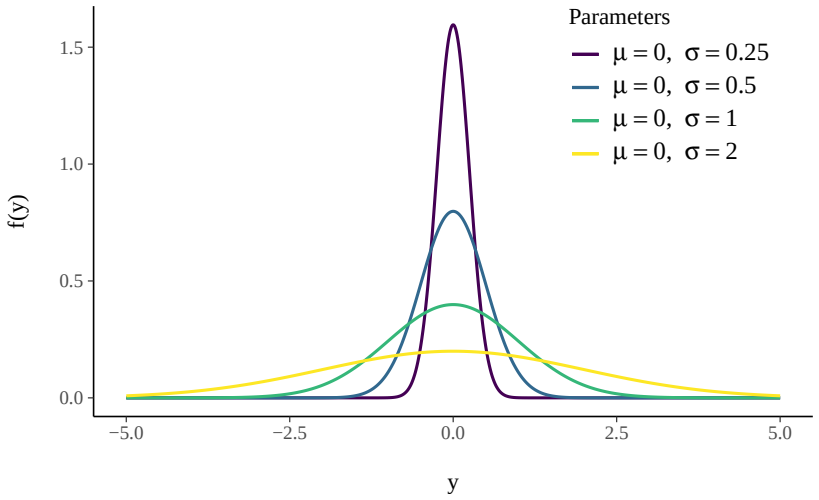


Figura 1.5: *Alcune distribuzioni Normali.*

Si noti come un dato la cui distanza dalla media è superiore a 3 volte la deviazione standard presenti un carattere di eccezionalità perché meno del 0.3% dei dati della distribuzione Normale presentano questa caratteristica.

Per indicare la distribuzione Normale si usa la notazione $\mathcal{N}(\mu, \sigma)$.

Funzione di ripartizione

Il valore della funzione di ripartizione di Y nel punto y è l'area sottesa alla curva di densità $f(y)$ nella semiretta $(-\infty, y]$. Non esiste alcuna funzione elementare per la funzione di ripartizione

$$F(y) = \int_{-\infty}^{y} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} dy, \tag{1.3}$$

pertanto le probabilità $P(Y < y)$ vengono calcolate mediante integrazione numerica approssimata. I valori della funzione di ripartizione di una variabile casuale Normale sono dunque forniti da un software.

Esempio 1.2. Usiamo Rper calcolare la funzione di ripartizione della Normale. La funzione `pnorm(q, mean, sd)` restituisce la funzione di ripartizione della Normale con media `mean` e deviazione standard `sd`, ovvero l'area sottesa alla funzione di densità di una Normale con media `mean` e deviazione standard `sd` nell'intervallo $[-\infty, q]$.

Per esempio, in precedenza abbiamo detto che il 68% circa dell'area sottesa ad una Normale è compresa nell'intervallo $\mu \pm \sigma$. Verifichiamo per la distribuzione del QI $\sim \mathcal{N}(\mu = 100, \sigma = 15)$:

```
pnorm(100 + 15, 100, 15) - pnorm(100 - 15, 100, 15)
#> [1] 0.683
```

Il 95% dell'area è compresa nell'intervallo $\mu \pm 1.96 \cdot \sigma$:

```
pnorm(100 + 1.96 * 15, 100, 15) - pnorm(100 - 1.96 * 15, 100, 15)
#> [1] 0.95
```

Quasi tutta la distribuzione è compresa nell'intervallo $\mu \pm 3 \cdot \sigma$:

```
pnorm(100 + 3 * 15, 100, 15) - pnorm(100 - 3 * 15, 100, 15)
#> [1] 0.997
```

Distribuzione Normale standard

La distribuzione Normale di parametri $\mu = 0$ e $\sigma = 1$ viene detta *distribuzione Normale standard*. La famiglia Normale è l'insieme avente come elementi tutte le distribuzioni Normali con parametri μ e σ diversi. Tutte le distribuzioni Normali si ottengono dalla Normale standard mediante una trasformazione lineare: se $Y \sim \mathcal{N}(\mu_Y, \sigma_Y)$ allora

$$X = a + bY \sim \mathcal{N}(\mu_X = a + b\mu_Y, \sigma_X = |b|\,\sigma_Y). \tag{1.4}$$

L'area sottesa alla curva di densità di $\mathcal{N}(\mu, \sigma)$ nella semiretta $(-\infty, y]$ è uguale all'area sottesa alla densità Normale standard nella semiretta $(-\infty, z]$, in cui $z = (y - \mu_Y)/\sigma_Y$ è il punteggio standard di Y. Per la simmetria della distribuzione, l'area sottesa nella semiretta $[1, \infty)$ è uguale all'area sottesa nella semiretta $(-\infty, 1]$ e quest'ultima coincide con $F(-1)$. Analogamente, l'area sottesa nell'intervallo $[y_a, y_b]$, con $y_a < y_b$, è pari a $F(z_b) - F(z_a)$, dove z_a e z_b sono i punteggi standard di y_a e y_b.

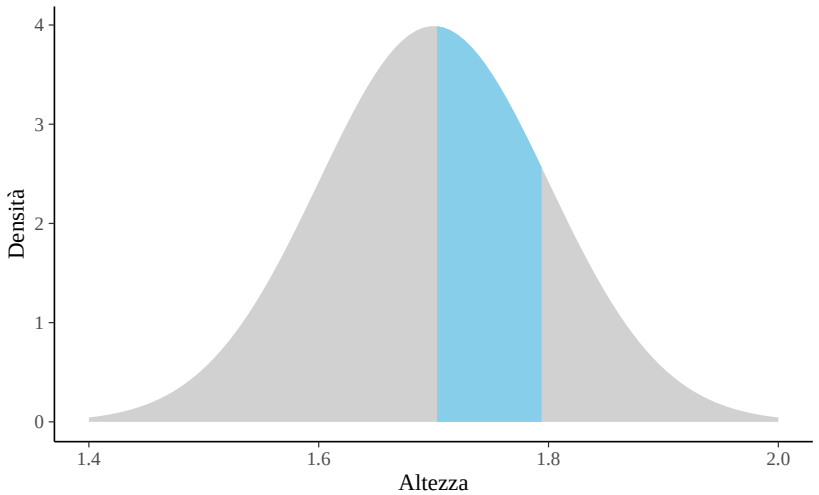
Si ha anche il problema inverso rispetto a quello del calcolo delle aree: dato un numero $0 \leq p \leq 1$, il problema è quello di determinare un numero $z \in \mathbb{R}$ tale che $P(Z < z) = p$. Il valore z cercato è detto *quantile* di ordine p della Normale standard e può essere trovato mediante un software.

Esempio 1.3. Supponiamo che l'altezza degli individui adulti segua la distribuzione Normale di media $\mu = 1.7$ m e deviazione standard $\sigma = 0.1$ m. Vogliamo sapere la proporzione di individui adulti con un'altezza compresa tra 1.7 e 1.8 m.

Il problema ci chiede di trovare l'area sottesa alla distribuzione $\mathcal{N}(\mu = 1.7, \sigma = 0.1)$ nell'intervallo $[1.7, 1.8]$:

```
df <- tibble(x = seq(1.4, 2.0, length.out = 100)) %>%
  mutate(y = dnorm(x, mean = 1.7, sd = 0.1))

ggplot(df, aes(x, y)) +
  geom_area(fill = "sky blue") +
  gghighlight(x < 1.8 & x > 1.7) +
  labs(
    x = "Altezza",
    y = "Densità"
  )
```



La risposta si trova utilizzando la funzione di ripartizione $F(X)$ della legge $\mathcal{N}(1.7, 0.1)$ in corrispondenza dei due valori forniti dal problema: $F(X = 1.8) - F(X = 1.7)$. Utilizzando la seguente istruzione

```
pnorm(1.8, 1.7, 0.1) - pnorm(1.7, 1.7, 0.1)
#> [1] 0.341
```

otteniamo il 31.43%.

In maniera equivalente, possiamo standardizzare i valori che delimitano l'intervallo considerato e utilizzare la funzione di ripartizione della normale standardizzata. I limiti inferiore e superiore dell'intervallo sono

$$z_{\text{inf}} = \frac{1.7 - 1.7}{0.1} = 0, \quad z_{\text{sup}} = \frac{1.8 - 1.7}{0.1} = 1.0,$$

quindi otteniamo

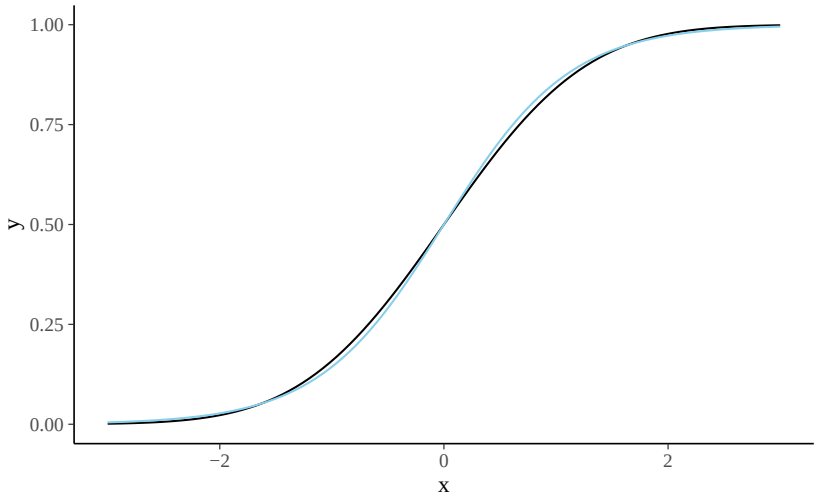
```
pnorm(1.0, 0, 1) - pnorm(0, 0, 1)
#> [1] 0.341
```

Il modo più semplice per risolvere questo problema resta comunque quello di rendersi conto che la probabilità richiesta non è altro che la metà dell'area sottesa dalle distribuzioni Normali nell'intervallo $[\mu - \sigma, \mu + \sigma]$, ovvero 0.683/2.

Funzione di ripartizione della variabile casuale normale standard e funzione logistica

Si noti che la funzione logistica (in blu), pur essendo del tutto diversa dalla Normale dal punto di vista formale, assomiglia molto alla Normale standard quando le due cdf hanno la stessa varianza.

```
tibble(x = c(-3, 3)) %>%
  ggplot(aes(x = x)) +
  stat_function(fun = pnorm) +
  stat_function(
    fun = plogis,
    args = list(scale = 0.56),
    col = "sky blue"
  )
```



1.3 Teorema del limite centrale

Laplace dimostrò il teorema del limite centrale (TLC) nel 1812. Il TLC ci dice che se prendiamo una sequenza di variabili casuali indipendenti e le sommiamo, tale somma tende a distribuirisi come una Normale. Il TLC specifica inoltre, sulla base dei valori attesi e delle varianze delle v.c. che vengono sommate, quali saranno i parametri della distribuzione Normale così ottenuta.

Teorema 1.1. *Si supponga che* $Y = Y_1, Y_2, \dots, Y_N$ *sia una sequenza di v.a. i.i.d. con* $\mathbb{E}(Y_n) = \mu$ *e* $\text{SD}(Y_n) = \sigma$*. Si definisca una nuova v.c. come la media di* Y*:*

$$Z = \frac{1}{N} \sum_{n=1}^{N} Y_n.$$

Con $N \to \infty$*,* Z *tenderà ad una Normale con lo stesso valore atteso di* Y_n *e una deviazione standard che sarà più piccola della deviazione standard originaria di un fattore pari a* $\sqrt{\frac{1}{\sqrt{N}}}$*:*

$$p_Z(z) \to \mathcal{N}\left(z \,\middle|\, \mu,\, \frac{1}{\sqrt{N}} \cdot \sigma\right). \tag{1.5}$$

Il TLC può essere generalizzato a variabili che non hanno la stessa distribuzione purché siano indipendenti e abbiano aspettative e varianze finite.

Molti fenomeni naturali, come l'altezza dell'uomo adulto di entrambi i sessi, sono il risultato di una serie di effetti additivi relativamente piccoli, la cui combinazione porta alla normalità, indipendentemente da come gli effetti additivi sono distribuiti. In pratica, questo è il motivo per cui la distribuzione normale ha senso come rappresentazione di molti fenomeni naturali.

1.4 Distribuzione Chi-quadrato

Dalla Normale deriva la distribuzione χ^2. La distribuzione ${\chi^2}_k$ con k gradi di libertà descrive la variabile casuale

$$Z_1^2 + Z_2^2 + \cdots + Z_k^2,$$

dove $Z_1, Z_2, \dots, Z_k$ sono variabili casuali i.i.d. con distribuzione Normale standard $\mathcal{N}(0, 1)$. La variabile casuale chi-quadrato dipende dal parametro intero positivo $\nu = k$ che ne identifica il numero di gradi di libertà. La densità di probabilità di ${\chi^2}_\nu$ è

$$f(x) = C_\nu x^{\nu/2-1} \exp(-x/2), \qquad \text{se } x > 0,$$

dove C_ν è una costante positiva.

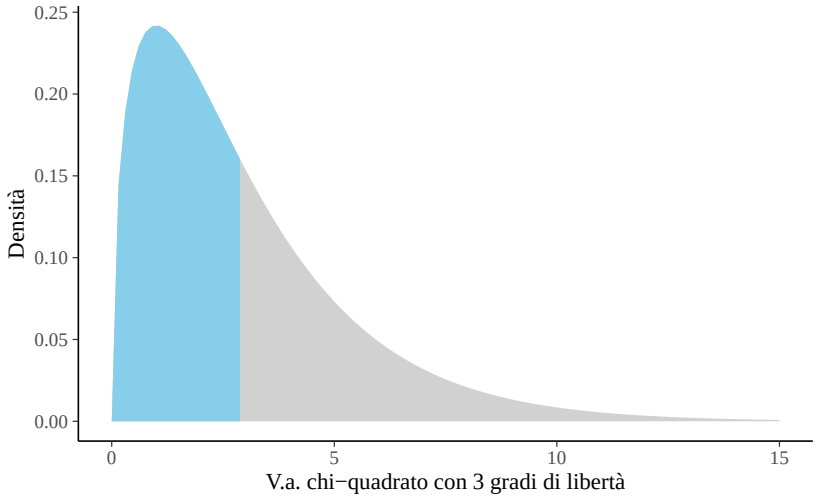
Proprietà

- La distribuzione di densità ${\chi^2}_\nu$ è asimmetrica.
- Il valore atteso di una variabile ${\chi^2}_\nu$ è uguale a ν.
- La varianza di una variabile ${\chi^2}_\nu$ è uguale a 2ν.
- Per $k \to \infty$, la ${\chi^2}_\nu \to \mathcal{N}$.
- Se X e Y sono due variabili casuali chi-quadrato indipendenti con ν_1 e ν_2 gradi di libertà, ne segue che $X + Y \sim \chi^2_m$, con $m = \nu_1 + \nu_2$. Tale principio si estende a qualunque numero finito di variabili casuali chi-quadrato indipendenti.

Esempio 1.4. Usiamo R per disegnare la densità chi-quadrato con 3 gradi di libertà dividendo l'area sottesa alla curva di densità in due parti uguali.

```r
df <- tibble(x = seq(0, 15.0, length.out = 100)) %>%
  mutate(y = dchisq(x, 3))

ggplot(df, aes(x, y)) +
  geom_area(fill = "sky blue") +
  gghighlight(x < 3) +
  labs(
    x = "V.a. chi-quadrato con 3 gradi di libertà",
    y = "Densità"
  )
```



Esempio 1.5. La figura 1.6 mostra alcune distribuzioni Chi-quadrato variando il parametro ν.

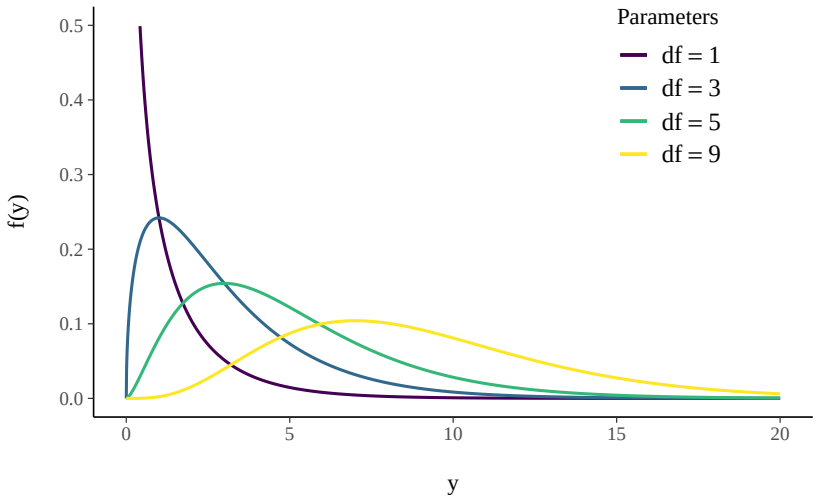


Figura 1.6: *Alcune distribuzioni Chi-quadrato.*

1.5 Distribuzione t di Student

Dalle distribuzioni Normale e Chi quadrato deriva un'altra distribuzione molto nota, la t di Student. Se $Z \sim \mathcal{N}$ e $W \sim \chi^2_\nu$ sono due variabili casuali indipendenti, allora il rapporto

$$T = \frac{Z}{\left(\frac{W}{\nu}\right)^{\frac{1}{2}}} \tag{1.6}$$

definisce la distribuzione t di Student con ν gradi di libertà. Si usa scrivere $T \sim t_\nu$. L'andamento della distribuzione t di Student è simile a quello della distribuzione Normale, ma ha una maggiore dispersione (ha le code più pesanti di una Normale, ovvero ha una varianza maggiore di 1).

Proprietà

La variabile casuale t di Student soddisfa le seguenti proprietà:

1. Per $\nu \to \infty$, t_ν tende alla normale standard $\mathcal{N}(0, 1)$.
2. La densità della t_ν è una funzione simmetrica con valore atteso nullo.
3. Per $\nu > 2$, la varianza della t_ν vale $\nu/(\nu - 2)$; pertanto è sempre maggiore di 1 e tende a 1 per $\nu \to \infty$.

Esempio 1.6. La figura 1.7 mostra alcune distribuzioni t di Student variando il parametro ν.

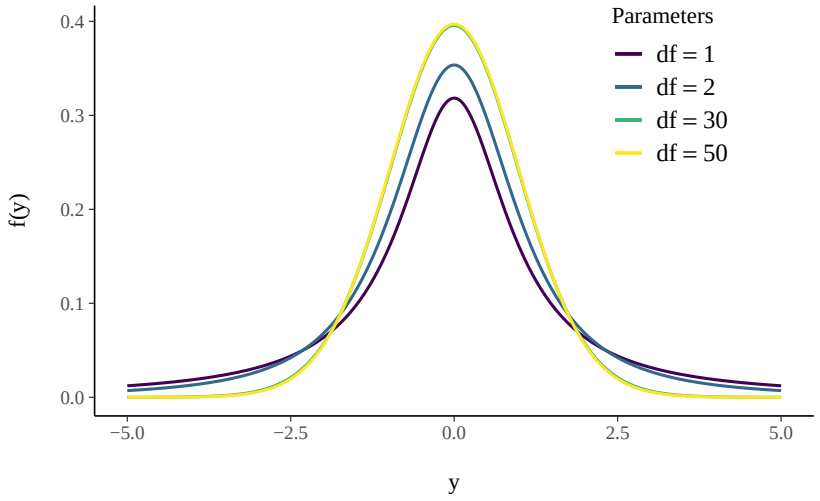


Figura 1.7: *Alcune distribuzioni t di Student.*

1.6 Funzione beta di Eulero

La funzione beta di Eulero è una funzione matematica, *non* una densità di probabilità. La menzioniamo qui perché viene utilizzata nella distribuzione Beta. La funzione beta si può scrivere in molti modi; per i nostri scopi, può essere scritta nel modo seguente:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}, \tag{1.7}$$

dove $\Gamma(x)$ è la funzione Gamma, ovvero il fattoriale discendente, cioè

$$x(x-1)(x-2)\dots(x-n+1).$$

1.7 Distribuzione Beta

Una distribuzione che viene usata per modellare percentuali e proporzioni è la distribuzione Beta in quanto è definita sull'intervallo $(0; 1)$ – ma non include i valori 0 o 1. La distribuzione Beta è una distribuzione estremamente flessibile e può assumere molti tipi di forme diverse (un'illustrazione è fornita dalla seguente GIF animata). Una definizione formale è la seguente.

Definizione 1.1. Sia π una variabile casuale che può assumere qualsiasi valore compreso tra 0 e 1, cioè $\pi \in [0, 1]$. Diremo che π segue la distribuzione Beta di parametri α e β, $\pi \sim \text{Beta}(\alpha, \beta)$, se la sua densità è

$$\begin{aligned} \text{Beta}(\pi \mid \alpha, \beta) &= \frac{1}{B(\alpha, \beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1} \quad \text{per } \pi \in [0, 1]\,, \end{aligned} \tag{1.8}$$

laddove $B(\alpha, \beta)$ è la funzione beta.

I termini α e β sono i parametri della distribuzione Beta e devono essere entrambi positivi. Tali parametri possono essere interpretati come l'espressione delle nostre credenze a priori relativamente alla probabilità di successo. Il parametro α rappresenta il numero di "successi" e il parametro β il numero di "insuccessi":

$$\frac{\text{Numero di successi}}{\text{Numero di successi} + \text{Numero di insuccessi}} = \frac{\alpha}{\alpha+\beta}\,.$$

Il rapporto $\frac{1}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+b)}{\Gamma(\alpha)\Gamma(\beta)}$ è una costante di normalizzazione:

$$\int_0^1 \pi^{\alpha-1}(1-\pi)^{\beta-1} = \frac{\Gamma(\alpha+b)}{\Gamma(\alpha)\Gamma(\beta)}\,. \tag{1.9}$$

Il valore atteso, la moda e la varianza di una distribuzione Beta sono dati dalle seguenti equazioni:

$$\mathbb{E}(\pi) = \frac{\alpha}{\alpha+\beta}\,, \tag{1.10}$$

$$\text{Mo}(\pi) = \frac{\alpha-1}{\alpha+\beta-2}\,, \tag{1.11}$$

$$\text{Var}(\pi) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}\,. \tag{1.12}$$

Attenzione alle parole. In questo contesto, il termine "beta" viene utilizzato con tre significati diversi:

- la distribuzione di densità Beta,
- la funzione matematica beta,
- il parametro β.

Esempio 1.7. Al variare di α e β si ottengono molte distribuzioni di forma diversa; per $\alpha = \beta = 1$ si ha la densità uniforme. Vari esempi di distribuzioni Beta sono mostrati nella figura 1.8.

Esempio 1.8. Nel disturbo depressivo la recidiva è definita come la comparsa di un nuovo episodio depressivo che si manifesta dopo un prolungato periodo di recupero (6-12 mesi) con stato di eutimia (umore relativamente normale). Supponiamo che una serie di studi mostri una comparsa di recidiva in una proporzione che va dal 20% al 60% dei casi,

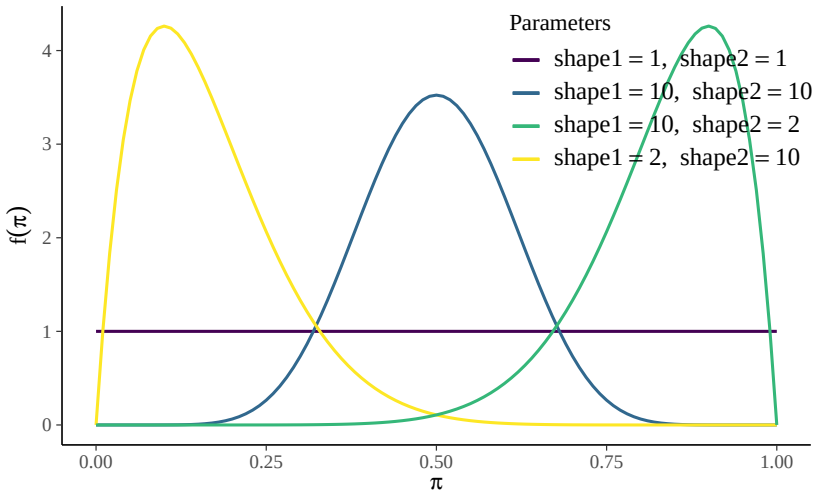
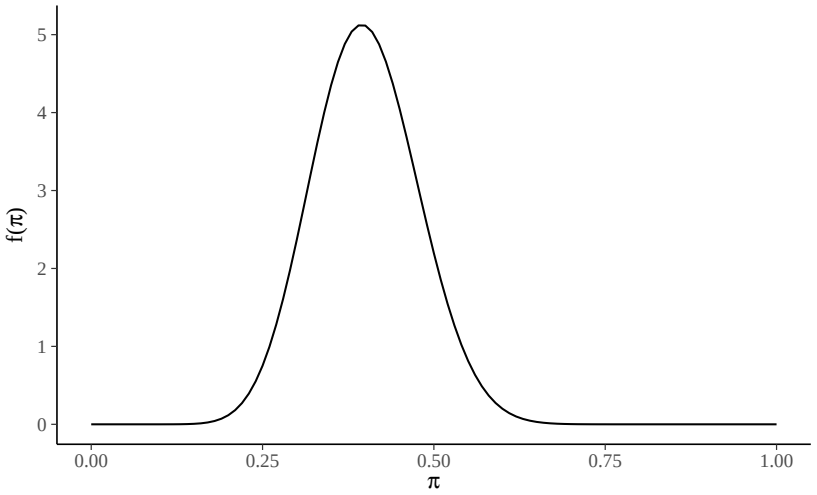


Figura 1.8: *Alcune distribuzioni Beta.*

con una media del 40% (per una recente discussione, si veda Nuggerud-Galeas et al., 2020). Sulla base di queste ipotetiche informazioni, è possibile usare la distribuzione Beta per rappresentare le nostre credenze a priori relativamente alla probabilità di recidiva. Per fare questo dobbiamo trovare i parametri della distribuzione Beta tali per cui la massa della densità sia compresa tra 0.2 e 0.6, con la media in corrispondenza di 0.4. Procedendo per tentativi ed errori, ed usando la funzione `bayesrules::plot_beta()`, un risultato possibile è $\mathrm{B}(16, 24)$.

```
find_pars <- function(ev, n) {
  a <- ev * n
  b <- n - a
  return(c(round(a), round(b)))
}

pars <- find_pars(.4, 40)
pars
#> [1] 16 24
bayesrules::plot_beta(pars[1], pars[2])
```



La media della distribuzione a priori diventa:

```
16 / (16 + 24)
#> [1] 0.4
```

e la moda è

```
(16 - 1) / (16 + 24 - 2)
#> [1] 0.395
```

Inoltre, la deviazione standard della distribuzione a priori diventa

```
sqrt((16 * 24) / ((16 + 24)^2 * (16 + 24 + 1)))
#> [1] 0.0765
```

uguale a circa 8 punti percentuali. Questo significa che le nostre credenze a priori rispetto la possibilità di recidiva tendono a deviare di circa 8 punti percentuali rispetto alla media della distribuzione a priori del 40%.

1.8 Distribuzione di Cauchy

La distribuzione di Cauchy è un caso speciale della distribuzione di t di Student con 1 grado di libertà. È definita da una densità di probabilità che corrisponde alla funzione, dipendente da due parametri θ e d (con la condizione $d > 0$),

$$f(x; \theta, d) = \frac{1}{\pi d} \frac{1}{1 + \left(\frac{x-\theta}{d}\right)^2}, \tag{1.13}$$

dove θ è la mediana della distribuzione e d ne misura la larghezza a metà altezza.

1.9 Distribuzione log-normale

Sia y una variabile casuale avente distribuzione normale con media μ e varianza σ^2. Definiamo poi una nuova variabile casuale x attraverso la relazione

$$x = e^y \quad \iff \quad y = \log x.$$

Il dominio di definizione della x è il semiasse $x > 0$ e la densità di probabilità $f(x)$ è data da

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{x} \exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\}. \tag{1.14}$$

Questa funzione di densità si chiama log-normale.

Il valore atteso e la varianza di una distribuzione log-normale sono dati dalle seguenti equazioni:

$$\mathbb{E}(x) = \exp\left\{\mu + \frac{\sigma^2}{2}\right\}. \tag{1.15}$$

$$\text{Var}(x) = \exp\left\{2\mu + \sigma^2\right\}\left(\exp\left\{\sigma^2\right\} - 1\right). \tag{1.16}$$

Si può dimostrare che il prodotto di variabili casuali log-normali ed indipendenti segue una distribuzione log-normale.

1.10 Distribuzione di Pareto

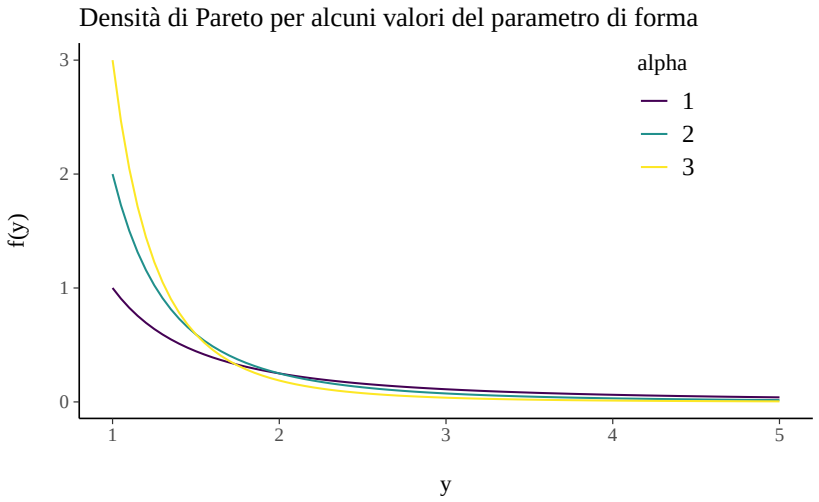
La distribuzione paretiana (o distribuzione di Pareto) è una distribuzione di probabilità continua e così chiamata in onore di Vilfredo Pareto. La distribuzione di Pareto è una distribuzione di probabilità con legge di potenza utilizzata nella descrizione di fenomeni sociali e molti altri tipi di fenomeni osservabili. Originariamente applicata per descrivere la distribuzione del reddito in una società, adattandosi alla tendenza che una grande porzione di ricchezza è detenuta da una piccola frazione della popolazione, la distribuzione di Pareto è diventata colloquialmente nota e indicata come il principio di Pareto, o "regola 80-20". Questa regola afferma che, ad esempio, l'80% della ricchezza di una società è detenuto dal 20% della sua popolazione. Viene spesso applicata nello studio della distribuzione del reddito, della dimensione dell'impresa, della dimensione di una popolazione e nelle fluttuazioni del prezzo delle azioni.

La densità di una distribuzione di Pareto é

$$f(x) = (x_m/x)^{\alpha},$$

dove x_m (parametro di scala) è il minimo (necessariamente positivo) valore possibile di X e α è un parametro di forma.

```
x <- seq(1, 5, 0.05)
a3 <- 3 / x^(4)
a2 <- 2 / x^(3)
a1 <- 1 / x^(2)
df <- bind_rows(
  tibble(x = x, p = a3, alpha = "3"),
  tibble(x = x, p = a2, alpha = "2"),
  tibble(x = x, p = a1, alpha = "1")
)
gg <- df %>%
  ggplot(aes(x = x, y = p, group = alpha)) +
  geom_line(aes(color = alpha)) +
  xlim(1, 5) +
  ggtitle("Densità di Pareto per alcuni valori del parametro di forma") +
  ylab("P(X = x)") +
  scale_colour_viridis(
    discrete = TRUE, labels = parse_format()
  ) +
  labs(
    x = "\ny",
    y = "f(y)\n"
  )
print(gg)
```



La distribuzione di Pareto ha una asimmetria positiva. Il supporto della distribuzione di Pareto è la retta reale positiva. Tutti i valori devono essere maggiori del parametro di scala x_m, che è in realtà un parametro di soglia.

Bibliografia

Burger, E. B. & Starbird, M. (2012). *The 5 elements of effective thinking*. Princeton University Press. (Cit. a p. viii).

Horn, S. & Loewenstein, G. (2021). Underestimating Learning by Doing. *Available at SSRN 3941441* (cit. a p. ix).

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd Edition). Boca Raton, Florida, CRC Press. (Cit. a p. 2).

Nuggerud-Galeas, S., Sáez-Benito Suescun, L., Berenguer Torrijo, N., Sáez-Benito Suescun, A., Aguilar-Latorre, A., Magallón Botaya, R. & Oliván Blázquez, B. (2020). Analysis of depressive episodes, their recurrence and pharmacologic treatment in primary care patients: A retrospective descriptive study. *Plos one*, *15*(5), e0233454 (cit. a p. 13).

Elenco delle figure

Abstract This document contains the material of the lessons of Psicometria B000286 (2021/2022) aimed at students of the first year of the Degree Course in Psychological Sciences and Techniques of the University of Florence, Italy.

Keywords Data science, Bayesian statistics.