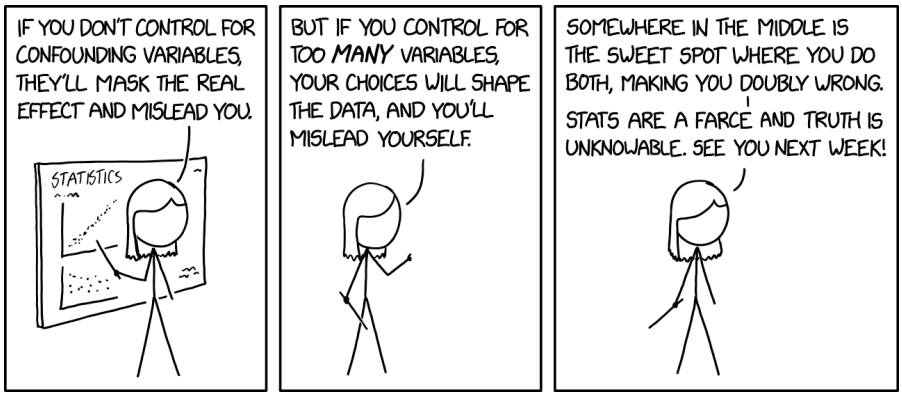
Corrado Caudek

Data Science per psicologi

Psicometria – AA 2021/2022



Indice

Elenco delle figure

Elenco delle tabelle

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. *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 un insieme di 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

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 pon-

te, 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 psicologico 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 verranno 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

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

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.

- 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 possia-

mo 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 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

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

un software, ricordatevi: “Google is your friend”!

Corrado Caudek
Marzo 2022

Parte I

Nozioni preliminari

1

Concetti chiave

La *data science* si pone all'intersezione tra statistica e informatica. La statistica è un insieme di metodi ugilizzati per estrarre informazioni dai dati; l'informatica implementa tali procedure in un software. In questo Capitolo vengono introdotti i concetti fondamentali.

1.1 Popolazioni e campioni

Popolazione. L'analisi dei dati inizia con l'individuazione delle unità portatrici di informazioni circa il fenomeno di interesse. Si dice popolazione (o universo) l'insieme Ω delle entità capaci di fornire informazioni sul fenomeno oggetto dell'indagine statistica. Possiamo scrivere $\Omega = \{\omega_i\}_{i=1,...,n} = \{\omega_1, \omega_2, ..., \omega_n\}$, oppure $\Omega = \{\omega_1, \omega_2, ...\}$ nel caso di popolazioni finite o infinite, rispettivamente.

L'obiettivo principale della ricerca psicologica è conoscere gli esiti psicologici e i loro fattori trainanti nella popolazione. Questo è l'obiettivo delle sperimentazioni psicologiche e della maggior parte degli studi osservazionali in psicologia. È quindi necessario essere molto chiari sulla popolazione a cui si applicano i risultati della ricerca. La popolazione può essere ben definita, ad esempio, tutte le persone che si trovavano nella città di Hiroshima al momento dei bombardamenti atomici e sono sopravvissute al primo anno, o può essere ipotetica, ad esempio, tutte le persone depresse che hanno subito o saranno sottoporsi ad un intervento di psicoterapia. Il ricercatore deve sempre essere in grado di determinare se un soggetto appartiene alla popolazione oggetto di interesse.

Una *sottopopolazione* è una popolazione in sé e per sé che soddisfa proprietà ben definite. Negli esempi precedenti, potremmo essere interessati alla sottopopolazione di uomini di età inferiore ai 20 anni o di pazienti depressi sottoposti ad uno specifico intervento psicologico. Molte questio-

ni scientifiche riguardano le differenze tra sottopopolazioni; ad esempio, confrontando i gruppi con o senza psicoterapia per determinare se il trattamento è vantaggioso. I modelli di regressione, introdotti nel Capitolo 6 riguardano le sottopopolazioni, in quanto stimano il risultato medio per diversi gruppi (sottopopolazioni) definiti dalle covariate.

Campione. Gli elementi ω_i dell'insieme Ω sono detti *unità statistiche*. Un sottoinsieme della popolazione, ovvero un insieme di elementi ω_i, viene chiamato *campione*. Ciascuna unità statistica ω_i (abbreviata con u.s.) è portatrice dell'informazione che verrà rilevata mediante un'operazione di misurazione.

Un campione è dunque un sottoinsieme della popolazione utilizzato per conoscere tale popolazione. A differenza di una sottopopolazione definita in base a chiari criteri, un campione viene generalmente selezionato tramite un procedura casuale. Il *campionamento casuale* consente allo scienziato di trarre conclusioni sulla popolazione e, soprattutto, di quantificare l'incertezza sui risultati. I campioni di un sondaggio sono esempi di campioni casuali, ma molti studi osservazionali non sono campionati casualmente. Possono essere *campioni di convenienza*, come coorti di studenti in un unico istituto, che consistono di tutti gli studenti sottoposti ad un certo intervento psicologico in quell'istituto. Indipendentemente da come vengono ottenuti i campioni, il loro uso al fine di conoscere una popolazione target significa che i problemi di rappresentatività sono inevitabili e devono essere affrontati.

1.2 Variabili e costanti

Definiamo *variabile statistica* la proprietà (o grandezza) che è oggetto di studio nell'analisi dei dati. Una variabile è una proprietà di un fenomeno che può essere espressa in più valori sia numerici sia categoriali. Il termine "variabile" si contrappone al termine "costante" che descrive una proprietà invariante di tutte le unità statistiche.

Si dice *modalità* ciascuna delle varianti con cui una variabile statistica può presentarsi. Definiamo *insieme delle modalità* di una variabile statistica l'insieme M di tutte le possibili espressioni con cui la variabile può manifestarsi. Le modalità osservate e facenti parte del campione si chiamano *dati* (si veda la Tabella 1.1).

Esempio 1.1. Supponiamo che il fenomeno studiato sia l'intelligenza. In uno studio, la popolazione potrebbe corrispondere all'insieme di tutti gli italiani adulti. La variabile considerata potrebbe essere il punteggio del test standardizzato WAIS-IV. Le modalità di tale variabile potrebbero essere $112, 92, 121, ...$ Tale variabile è di tipo quantitativo discreto.

Esempio 1.2. Supponiamo che il fenomeno studiato sia il compito Stroop. La popolazione potrebbe corrispondere all'insieme dei bambini dai 6 agli 8 anni. La variabile considerata potrebbe essere il reciproco dei tempi di reazione in secondi. Le modalità di tale variabile potrebbero essere $1/2.35, 1/1.49, 1/2.93, ...$ La variabile è di tipo quantitativo continuo.

Esempio 1.3. Supponiamo che il fenomeno studiato sia il disturbo di personalità. La popolazione potrebbe corrispondere all'insieme dei detenuti nelle carceri italiane. La variabile considerata potrebbe essere l'assessment del disturbo di personalità tramite interviste cliniche strutturate. Le modalità di tale variabile potrebbero essere i Cluster A, Cluster B, Cluster C descritti dal DSM-V. Tale variabile è di tipo qualitativo.

1.2.1 Variabili casuali

Il termine *variabile* usato nella statistica è equivalente al termine *variabile casuale* usato nella teoria delle probabilità. Lo studio dei risultati degli interventi psicologici è lo studio delle variabili casuali che misurano questi risultati. Una variabile casuale cattura una caratteristica specifica degli individui nella popolazione e i suoi valori variano tipicamente tra gli individui. Ogni variabile casuale può assumere in teoria una gamma di valori sebbene, in pratica, osserviamo un valore specifico per ogni individuo. Quando faremo riferiremo alle variabili casuali considerate in termini generali useremo lettere maiuscole come X e Y; quando faremo riferimento ai valori che una variabile casuale assume in determinate circostanze useremo lettere minuscole come x e y.

1.2.2 Variabili indipendenti e variabili dipendenti

Un primo compito fondamentale in qualsiasi analisi dei dati è l'identificazione delle variabili dipendenti (Y) e delle variabili indipendenti (X). Le variabili dipendenti sono anche chiamate variabili di esito o di risposta e le variabili indipendenti sono anche chiamate predittori o covariate. Ad esempio, nell'analisi di regressione, che esamineremo in seguito, la

domanda centrale è quella di capire come Y cambia al variare di X. Più precisamente, la domanda che viene posta è: se il valore della variabile indipendente X cambia, qual è la conseguenza per la variabile dipendente Y? In parole povere, le variabili indipendenti e dipendenti sono analoghe a "cause" ed "effetti", laddove le virgolette usate qui sottolineano che questa è solo un'analogia e che la determinazione delle cause può avvenire soltanto mediante l'utilizzo di un appropriato disegno sperimentale e di un'adeguata analisi statistica.

Se una variabile è una variabile indipendente o dipendente dipende dalla domanda di ricerca. A volte può essere difficile decidere quale variabile è dipendente e quale è indipendente, in particolare quando siamo specificamente interessati ai rapporti di causa/effetto. Ad esempio, supponiamo di indagare l'associazione tra esercizio fisico e insonnia. Vi sono evidenze che l'esercizio fisico (fatto al momento giusto della giornata) può ridurre l'insonnia. Ma l'insonnia può anche ridurre la capacità di una persona di fare esercizio fisico. In questo caso, dunque, non è facile capire quale sia la causa e quale l'effetto, quale sia la variabile dipendente e quale la variabile indipendente. La possibilità di identificare il ruolo delle variabili (dipendente/indipendente) dipende dalla nostra comprensione del fenomeno in esame.

Esempio 1.4. Uno psicologo convoca 120 studenti universitari per un test di memoria. Prima di iniziare l'esperimento, a metà dei soggetti viene detto che si tratta di un compito particolarmente difficile; agli altri soggetti non viene data alcuna indicazione. Lo psicologo misura il punteggio nella prova di memoria di ciascun soggetto.

In questo esperimento, la variabile indipendente è l'informazione sulla difficoltà della prova. La variabile indipendente viene manipolata dallo sperimentatore assegnando i soggetti (di solito in maniera causale) o alla condizione (modalità) "informazione assegnata" o "informazione non data". La variabile dipendente è ciò che viene misurato nell'esperimento, ovvero il punteggio nella prova di memoria di ciascun soggetto.

1.2.3 La matrice dei dati

Le realizzazioni delle variabili esaminate in una rilevazione statistica vengono organizzate in una *matrice dei dati*. Le colonne della matrice dei dati contengono gli insiemi dei dati individuali di ciascuna variabile statistica considerata. Ogni riga della matrice contiene tutte le informazioni

relative alla stessa unità statistica. Una generica matrice dei dati ha l'aspetto seguente:

$$D_{m,n} = \begin{pmatrix} \omega_1 & a_1 & b_1 & \cdots & x_1 & y_1 \\ \omega_2 & a_2 & b_2 & \cdots & x_2 & y_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \omega_n & a_n & b_n & \cdots & x_n & y_n \end{pmatrix}$$

dove, nel caso presente, la prima colonna contiene il nome delle unità statistiche, la seconda e la terza colonna si riferiscono a due mutabili statistiche (variabili categoriali; A e B) e ne presentano le modalità osservate nel campione mentre le ultime due colonne si riferiscono a due variabili statistiche (X e Y) e ne presentano le modalità osservate nel campione. Generalmente, tra le unità statistiche ω_i non esiste un ordine progressivo; l'indice attribuito alle unità statistiche nella matrice dei dati si riferisce semplicemente alla riga che esse occupano.

1.3 Parametri e modelli

Ogni variabile casuale ha una *distribuzione* che descrive la probabilità che la variabile assuma qualsiasi valore in un dato intervallo.[1] Senza ulteriori specificazioni, una distribuzione può fare riferimento a un'intera famiglia di distribuzioni. I parametri, tipicamente indicati con lettere greche come μ e α, ci permettono di specificare di quale membro della famiglia stiamo parlando. Quindi, si può parlare di una variabile casuale con una distribuzione Normale, ma se viene specificata la media $\mu = 100$ e la varianza $\sigma^2 = 15$, viene individuata una specifica distribuzione Normale – nell'esempio, la distribuzione del quoziente di intelligenza.

I metodi statistici parametrici specificano la famiglia delle distribuzioni e quindi utilizzano i dati per individuare, stimando i parametri, una specifica distribuzione all'interno della famiglia di distribuzioni ipotizzata. Se f è la PDF di una variabile casuale Y, l'interesse può concentrarsi

[1] In questo e nei successivi Paragrafi di questo Capitolo introduco gli obiettivi della *data science* utilizzando una serie di concetti che saranno chiariti solo in seguito. Questa breve panoramica risulterà dunque solo in parte comprensibile ad una prima lettura e serve solo per definire la *big picture* dei temi trattati in questo insegnamento. Il significato dei termini qui utilizzati sarà chiarito nei Capitoli successivi.

sulla sua media e varianza. Nell'analisi di regressione, ad esempio, cerchiamo di spiegare come i parametri di f dipendano dalle covariate X. Nella regressione lineare classica, assumiamo che Y abbia una distribuzione normale con media $\mu = \mathbb{E}(Y)$, e stimiamo come $\mathbb{E}(Y)$ dipenda da X. Poiché molti esiti psicologici non seguono una distribuzione normale, verranno introdotte distribuzioni più appropriate per questi risultati. I metodi non parametrici, invece, non specificano una famiglia di distribuzioni per f. In queste dispense faremo riferimento a metodi non parametrici quando discuteremo della statistica descrittiva.

Il termine *modello* è onnipresente in statistica e nella *data science*. Il modello statistico include le ipotesi e le specifiche matematiche relative alla distribuzione della variabile casuale di interesse. Il modello dipende dai dati e dalla domanda di ricerca, ma raramente è unico; nella maggior parte dei casi, esiste più di un modello che potrebbe ragionevolmente usato per affrontare la stessa domanda di ricerca e avendo a disposizione i dati osservati. Nella previsione delle aspettative future dei pazienti depressi che discuteremo in seguito (Zetsche et al., 2019), ad esempio, la specifica del modello include l'insieme delle covariate candidate, l'espressione matematica che collega i predittori con le aspattative future e qualsiasi ipotesi sulla distribuzione della variabile dipendente. La domanda di cosa costituisca un buon modello è una domanda su cui torneremo ripetutamente in questo insegnamento.

1.4 Effetto

L'*effetto* è una qualche misura dei dati. Dipende dal tipo di dati e dal tipo di test statistico che si vuole utilizzare. Ad esempio, se viene lanciata una moneta 100 volte e esce testa 66 volte, l'effetto sarà 66/100. Diventa poi possibile confrontare l'effetto ottenuto con l'effetto nullo che ci si aspetterebbe da una moneta bilanciata (50/100), o con qualsiasi altro effetto che può essere scelto. La *dimensione dell'effetto* si riferisce alla differenza tra l'effetto misurato nei dati e l'effetto nullo (di solito un valore che ci si aspetta di ottenere in base al caso soltanto).

1.5 Stima e inferenza

La stima è il processo mediante il quale il campione viene utilizzato per conoscere le proprietà di interesse della popolazione. La media campionaria è una stima naturale della media della popolazione e la mediana campionaria è una stima naturale della mediana della popolazione. Quando parliamo di stimare una proprietà della popolazione (a volte indicata come parametro della popolazione) o di stimare la distribuzione di una variabile casuale, stiamo parlando dell'utilizzo dei dati osservati per conoscere le proprietà di interesse della popolazione. L'inferenza statistica è il processo mediante il quale le stime campionarie vengono utilizzate per rispondere a domande di ricerca e per valutare specifiche ipotesi relative alla popolazione. Discuteremo le procedure bayesiane dell'inferenza nell'ultima parte di queste dispense.

1.6 Metodi e procedure della psicologia

Un modello psicologico di un qualche aspetto del comportamento umano o della mente ha le seguenti proprietà:

1. descrive le caratteristiche del comportamento in questione,
2. formula predizioni sulle caratteristiche future del comportamento,
3. è sostenuto da evidenze empiriche,
4. deve essere falsificabile (ovvero, in linea di principio, deve potere fare delle predizioni su aspetti del fenomeno considerato che non sono ancora noti e che, se venissero indagati, potrebbero portare a rigettare il modello, se si dimostrassero incompatibili con esso).

L'analisi dei dati valuta un modello psicologico utilizzando strumenti statistici.

Questa dispensa è strutturata in maniera tale da rispecchiare la suddivisione tra i temi della misurazione, dell'analisi descrittiva e dell'inferenza. Nel prossimo Capitolo sarà affrontato il tema della misurazione e, nell'ul-

tima parte della dispensa verrà discusso l'argomento più difficile, quello dell'inferenza. Prima di affrontare il secondo tema, l'analisi descrittiva dei dati, sarà necessario introdurre il linguaggio di programmazione statistica R (un'introduzione a R è fornita in Appendice). Inoltre, prima di potere discutere l'inferenza, dovranno essere introdotti i concetti di base della teoria delle probabilità, in quanto l'inferenza non è che l'applicazione della teoria delle probabilità all'analisi dei dati.

2

La misurazione in psicologia

Introduco il problema della misurazione in psicologia parlando dell'intelligenza. In quanto psicologi, siamo abituati a pensare alla misurazione dell'intelligenza, ma anche le persone che non sono psicologi sono ben familiari con la misurazione dell'intelligenza: tra le misurazioni delle caratteristiche psicologiche, infatti, la misurazione dell'intelligenza è forse la più conosciuta.

I test di intelligenza consistono in una serie di problemi di carattere verbale, numerico o simbolico. Come ci si può aspettare, alcune persone riescono a risolvere correttamente un numero maggiore di problemi di altre. Possiamo contare il numero di risposte corrette e osservare le differenze individuali nei punteggi calcolati. Scopriamo in questo modo che le differenze individuali nell'abilità di risolvere tali problemi risultano sorprendentemente stabili nell'età adulta. Inoltre, diversi test di intelligenza tendono ad essere correlati positivamente: le persone che risolvono un maggior numero di problemi verbali, in media, tenderanno anche a risolvere correttamente un numero più grande di numerici e simbolici. Esiste quindi una notevole coerenza delle differenze osservate tra le persone, sia nel tempo sia considerando diverse procedure di test e valutazione.

Avendo stabilito che ci sono differenze individuali tra le persone, è possibile esaminare le associazioni tra i punteggi dei test di intelligenza e altre variabili. Possiamo indagare se le persone con punteggi più alti nei test di intelligenza, rispetto a persone che ottengono punteggi più bassi, hanno più successo sul lavoro; se guadagnano di più; se votano in modo diverso; o se hanno un'aspettativa di vita più alta. Possiamo esaminare le differenze nei punteggi dei test di intelligenza in funzione di variabili come il genere, il gruppo etnico-razziale o lo stato socio-economico. Possiamo fare ricerche sull'associazione tra i punteggi dei test di intelligenza e l'efficienza dell'elaborazione neuronale, i tempi di reazione o la quantità di materia grigia all'interno della scatola cranica. Tutte queste

ricerche sono state condotte e gli psicologi hanno scoperto una vasta gamma di associazioni tra le misure dell'intelligenza e altre variabili. Alcune di queste associazioni sono grandi e stabili, altre sono piccole e difficili da replicare. In riferimento all'intelligenza, dunque, gli psicologi hanno condotto un enorme numero di ricerche ponendosi domande diverse. In quali condizioni si verificano determinati effetti? Quali variabili mediano o moderano le relazioni tra i punteggi dei test di intelligenza e altre variabili? Queste relazioni si mantengono stabili in diversi gruppi di persone? Le ricerche sull'intelligenza umana sono un campo in continuo sviluppo.

Tuttavia, tuttavia una domanda sorge spontanea: i test di intelligenza misurano davvero qualcosa e, in caso affermativo, che cos'è questo qualcosa? Infatti, dopo un secolo di teoria e ricerca sui punteggi dei test di intelligenza e, in generale, sui test psicologici, non sappiamo ancora con precisione cosa effettivamente questi test misurano. Queste considerazioni relative ai test di intelligenza ci conducono dunque alla domanda che ha motivato le precedenti considerazioni: cosa significa misurare un attributo psicologico? Questa è una domanda a cui è difficile rispondere, una domanda a cui è dedicata un'intera area di ricerca, quella della teoria della misurazione psicologica.

Non possiamo qui entrare nel merito delle complessità formali della teoria della misurazione psicologica – questo argomento verrà approfondito nei successivi insegnamenti sulla testistica psicologica. Ci limiteremo invece a presentare alcune nozioni di base su un tema centrale della teoria della misurazione psicologica: il tema delle scale delle misure psicologiche.

2.1 Le scale di misura

In generale possiamo dire che la teoria della misurazione si occupa dello studio delle relazioni esistenti tra due domini: il "mondo fisico" e il "mondo psicologico". Secondo la teoria della misurazione, la misurazione è un'attività rappresentativa, cioè è un processo di assegnazione di numeri in modo tale da preservare, all'interno del dominio numerico, le relazioni qualitative che sono state osservate nel mondo empirico. La teoria della misurazione ha lo scopo di specificare le condizioni necessarie per la costruzione di una rappresentazione adeguata delle relazioni empiriche all'interno di un sistema numerico. Da una prospettiva formale,

le operazioni descritte dalla teoria della misurazione possono essere concettualizzate in termini di mappatura tra le relazioni esistenti all'interno di due insiemi (quello empirico e quello numerico). Il risultato di questa attività è chiamato "scala di misurazione".

Una famosa teoria delle scale di misura è stata proposta da Stevens (1946). Stevens ci fa notare che, in linea di principio, le variabili psicologiche sono in grado di rappresentare (preservare) con diversi gradi di accuratezza le relazioni qualitative che sono state osservate nei fenomeni psicologici. Secondo la teoria di Stevens, possiamo distinguere tra quattro scale di misura: le scale nominali (*nominal scales*), ordinali (*ordinal scales*), a intervalli (*interval scales*), di rapporti (*ratio scales*). Tali scale di misura consentono operazioni aritmetiche diverse, come indicato nella tabella successiva, in quanto ciasuna di esse è in grado di "catturare" soltanto alcune delle proprietà dei fenomeni psicologici che intende misurare.

Scale di modalità	Operazioni aritmetiche
nominali	enumerare le classi di equivalenza e/o le frequenze per ciascuna classe di equivalenza
ordinali	enumerare le classi di equivalenza e/o le frequenze per ciascuna classe di equivalenza
intervallari	differenze (rapporti tra differenze)
di rapporti	rapporti diretti tra le misure

2.1.1 Scala nominale

Il livello di misurazione più semplice è quello della scala nominale. Questa scala di misurazione corrisponde ad una tassonomia. I simoboli o numeri che costituiscono questa scala non sono altro che i nomi delle categorie che utilizziamo per classificare i fenomeni psicologici. In base alle misure fornite da una scala nominale, l'unica cosa che siamo in grado di dire a proposito di una caratteristica psicologica è se essa è uguale o no ad un'altra caratteristica psicologica.

La scala nominale raggruppa dunque i dati in categorie qualitative *mutuamente esclusive* (cioè nessun dato si può collocare in più di una categoria). Esiste la sola relazione di equivalenza tra le misure delle u.s., cioè nella scala nominale gli elementi del campione appartenenti a classi

diverse sono differenti, mentre tutti quelli della stessa classe sono tra loro equivalenti: $x_i = x_j$ oppure $x_i \neq x_j$.

L'unica operazione algebrica che possiamo compiere sulle modalità della scala nominale è quella di contare le u.s. che appartengono ad ogni modalità e contare il numero delle modalità (classi di equivalenza). Dunque la descrizione dei dati avviene tramite le frequenze assolute e le frequenze relative.

A partire da una scala nominale è possibile costruire altre scale nominali che sono equivalenti alla prima trasformando i valori della scala di partenza in modo tale da cambiare i nomi delle modalità, ma lasciando però inalterata la suddivisione u.s. nelle medesime classi di equivalenza. Questo significa che prendendo una variabile misurata su scala nominale e cambiando i nomi delle sue categorie otteniamo una nuova variabile esattamente corrispondente alla prima.

2.1.2 Scala ordinale

La scala ordinale conserva la proprietà della scala nominale di classificare ciascuna u.s. all'interno di una e una sola categoria, ma alla relazione di equivalenza tra elementi di una stessa classe aggiunge la relazione di ordinamento tra le classi di equivalenza. Essendo basata su una relazione d'ordine, una scala ordinale descrive soltanto l'ordine di rango tra le modalità, ma non ci dà alcuna informazione su quanto una modalità sia più grande di un'altra. Non ci dice, per esempio, se la distanza tra le modalità a e b sia uguale, maggiore o minore della distanza tra le modalità b e c.

Esempio 2.1. Un esempio classico di scala ordinale è quello della scala Mohs per la determinazione della durezza dei minerali. Per stabilire la durezza dei minerali si usa il criterio empirico della scalfittura. Vengono stabiliti livelli di durezza crescente da 1 a 10 con riferimento a dieci minerali: talco, gesso, calcite, fluorite, apatite, ortoclasio, quarzo, topazio, corindone e diamante. Un minerale appartenente ad uno di questi livelli se scalfisce quello di livello inferiore ed è scalfito da quello di livello superiore.

2.1.3 Scala ad intervalli

La scala ad intervalli include le proprietà di quella nominale e di quella ordinale, e in più consente di misurare le distanze tra le coppie di u.s. nei termini di un intervallo costante, chiamato *unità di misura*, a cui viene attribuito il valore "1". La posizione dell'origine della scala, cioè il punto zero, è scelta arbitrariamente, nel senso che non indica l'assenza della quantità che si sta misurando. Avendo uno zero arbitrario, questa scala di misura consente valori negativi. Lo zero, infatti, *non* viene attribuito all'u.s. in cui la proprietà misurata risulta assente.

La scala a intervalli equivalenti ci consente di effettuare operazioni algebriche basate sulla differenza tra i numeri associati ai diversi punti della scala, operazioni algebriche non era possibile eseguire nel caso di misure a livello di scala ordinale o nominale. Il limite della scala ad intervalli è quello di non consentire il calcolo del rapporto tra coppie di misure. Possiamo dire, per esempio, che la distanza tra a e b è la metà della distanza tra c e d. Oppure che la distanza tra a e b è uguale alla distanza tra c e d. Non possiamo dire, però, che a possiede la proprietà misurata in quantità doppia rispetto b. Non possiamo cioè stabilire dei rapporti diretti tra le misure ottenute. Solo per le *differenze* tra le modalità sono dunque permesse tutte le operazioni aritmetiche: le differenze possono essere tra loro sommate, elevate a potenza oppure divise, determinando così le quantità che stanno alla base della statistica inferenziale.

Nelle scale ad intervalli equivalenti, l'unità di misura è arbitraria, ovvero può essere cambiata attraverso una dilatazione, operazione che consiste nel moltiplicare tutti i valori della scala per una costante positiva. Poiché l'aggiunta di una costante non altera le differenze tra i valori della scala, è anche ammessa la traslazione, operazione che consiste nel sommare una costante a tutti i valori della scala. Essendo la scala invariate rispetto alla traslazione e alla dilatazione, le trasformazioni ammissibili sono le *trasformazioni lineari*:

$$y' = a + by, \quad b > 0.$$

L'aspetto che rimane invariante a seguito di una trasformazione lineare è l'uguaglianza dei rapporti fra intervalli.

Esempio 2.2. Esempio di scala ad intervalli è la temperatura misurata in gradi Celsius o Fahrenheit, ma non Kelvin. Come per la scala nominale, è possibile stabilire se due modalità sono uguali o diverse: 30°C $\neq$

20°C. Come per la scala ordinale è possibile mettere due modalità in una relazione d'ordine: 30°C > 20°C. In aggiunta ai casi precedenti, però, è possibile definire una unità di misura per cui è possibile dire che tra 30°C e 20°C c'è una differenza di 30° - 20° = 10°C. I valori di temperatura, oltre a poter essere ordinati secondo l'intensità del fenomeno, godono della proprietà che le differenze tra loro sono direttamente confrontabili e quantificabili.

Il limite della scala ad intervalli è quello di non consentire il calcolo del rapporto tra coppie di misure. Ad esempio, una temperatura di 80°C non è il doppio di una di 40°C. Se infatti esprimiamo le stesse temperature nei termini della scala Fahrenheit, allora i due valori non saranno in rapporto di 1 a 2 tra loro. Infatti, 20°C = 68°F e 40°C = 104°F. Questo significa che la relazione "il doppio di" che avevamo individuato in precedenza si applicava ai numeri della scala centigrada, ma non alla proprietà misurata (cioè la temperatura). La decisione di che scala usare (Centigrada vs. Fahrenheit) è arbitraria. Ma questa arbitrarietà non deve influenzare le inferenze che traiamo dai dati. Queste inferenze, infatti, devono dirci qualcosa a proposito della realtà empirica e non possono in nessun modo essere condizionate dalle nostre scelte arbitrarie che ci portano a scegliere la scala Centigrada piuttosto che quella Fahrenheit.

Consideriamo ora l'aspetto invariante di una trasformazione lineare, ovvero l'uguaglianza dei rapporti fra intervalli. Prendiamo in esame, ad esempio, tre temperature: $20°C = 68°F$, $15°C = 59°F$, $10°C = 50°F$.

È facile rendersi conto del fatto che i rapporti fra intervalli restano costanti indipendentemente dall'unità di misura che è stata scelta:

$$\frac{20°C - 10°C}{20°C - 15°C} = \frac{68°F - 50°F}{68°F - 59°F} = 2.$$

2.1.4 Scala di rapporti

Nella scala a rapporti equivalenti la posizione dello zero non è arbitraria, ma corrisponde all'elemento dotato di intensità nulla rispetto alla proprietà misurata. Una scala a rapporti equivalenti si costruisce associando il numero 0 all'elemento con intensità nulla; viene poi scelta un'unità di misura u e, ad ogni elemento, si assegna un numero a definito come:

$$a = \frac{d}{u}$$

dove d rappresenta la distanza dall'origine. Alle u.s. vengono dunque assegnati dei numeri tali per cui le differenze e i rapporti tra i numeri riflettono le differenze e i rapporti tra le intensità della proprietà misurata.

Operazioni aritmetiche sono possibili non solo sulle differenze tra i valori della scala (come per la scala a intervalli equivalenti), ma anche sui valori stessi della scala. L'unica arbitrarietà riguarda l'unità di misura che si utilizza. L'unità di misura può cambiare, ma qualsiasi unità di misura si scelga, lo zero deve sempre indicare l'intensità nulla della proprietà considerata.

Le trasformazioni ammissibili a questo livello di scala sono dette trasformazioni di similarità:

$$y' = by, \quad b > 0.$$

A questo livello di scala, a seguito delle trasformazioni ammissibili, rimangono invariati anche i rapporti:

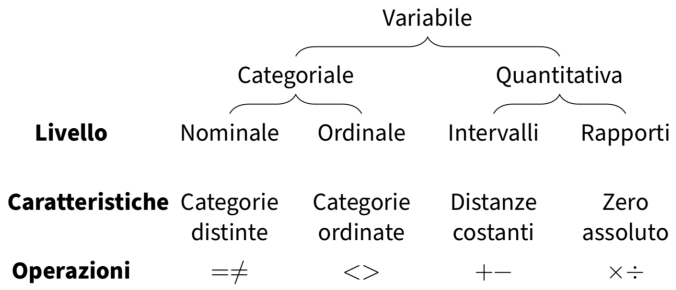
$$\frac{y_i}{y_j} = \frac{y'_i}{y'_j}.$$

2.2 Gerarchia dei livelli di scala di misura

Stevens (1946) parla di *livelli di scala* poiché i quattro tipi di scala di misura stanno in una precisa gerarchia: la scala nominale rappresenta il livello più basso della misurazione, la scala a rapporti equivalenti è invece il livello più alto.

Scale di modalità	Operazioni aritmetiche
nominali	enumerare le classi di equivalenza e/o le frequenze per ciascuna classe di equivalenza
ordinali	enumerare le classi di equivalenza e/o le frequenze per ciascuna classe di equivalenza
intervallari	differenze (rapporti tra differenze)
di rapporti	rapporti diretti tra le misure

Passando da un livello di misurazione ad uno più alto aumenta il numero di operazioni aritmetiche che possono essere compiute sui valori della scala, come indicato nella figura seguente.

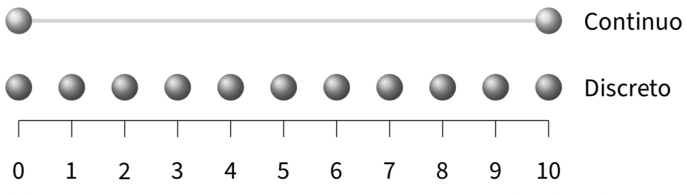


Per ciò che riguarda le trasformazioni ammissibili, più il livello di scala è basso, più le funzioni sono generali (sono minori cioè i vincoli per passare da una rappresentazione numerica ad un'altra equivalente). Salendo la gerarchia, la natura delle funzioni di trasformazione si fa più restrittiva.

2.3 Variabili discrete o continue

Le variabili a livello di intervalli e di rapporti possono essere discrete o continue. Le variabili discrete possono assumere alcuni valori ma non altri. Una volta che l'elenco di valori accettabili è stato specificato, non ci sono casi che cadono tra questi valori. Le variabili discrete di solito assumono valori interi.

Quando una variabile può assumere qualsiasi valore entro un intervallo specificato, allora si dice che la variabile è continua. In teoria, ciò significa che frazioni e decimali possono essere utilizzati per raggiungere un livello di precisione qualsiasi. In pratica, a un certo punto dobbiamo arrotondare i numeri, rendendo tecnicamente la variabile discreta. In variabili veramente discrete, tuttavia, non è possibile aumentare a piacimento il livello di precisione della misurazione.



Esempio 2.3. Il numero di biciclette possedute da una persona è una

variabile discreta poiché tale variabile può assumere come modalità solo i numeri interi non negativi. Frazioni di bicicletta non hanno senso.

2.4 Alcune misure sono migliori di altre

In psicologia, ciò che vogliamo misurare non è una caratteristica fisica, ma invece è un concetto teorico inosservabile, ovvero un costrutto.

> Un costrutto rappresenta il risultato di una fondata riflessione scientifica, non è per definizione accessibile all'osservazione diretta, ma viene inferito dall'osservazione di opportuni indicatori (Sartori, 2005).

Ad esempio, supponiamo che un docente voglia valutare quanto bene uno studente comprenda la distinzione tra le quattro diverse scale di misura che sono state descritte sopra. Il docente potrebbe predisporre un test costituito da un insieme di domande e potrebbe contare a quante domande lo studente risponde correttamente. Questo test, però, può o può non essere una buona misura del costrutto relativo alla conoscenza effettiva delle quattro scale di misura. Per esempio, se il docente scrive le domande del test in modo ambiguo o se usa una linguaggio troppo tecnico che lo studente non conosce, allora i risultati del test potrebbero suggerire che lo studente non conosce la materia in questione anche se in realtà questo non è vero. D'altra parte, se il docente prepara un test a scelta multipla con risposte errate molto ovvie, allora lo studente può ottenere dei buoni risultati al test anche senza essere in grado di comprendere adeguatamente le proprietà delle quattro scale di misura.

In generale non è possibile misurare un costrutto senza una certa quantità di errore. Poniamoci dunque il problema di determinare in che modo una misurazione possa dirsi adeguata.

2.4.1 Tipologie di errori

L'errore è, per definizione, la differenza tra il valore vero e il valore misurato della grandezza in esame. Gli errori sono classificati come sistematici (o determinati) e casuali (o indeterminati). Gli errori casuali sono fluttuazioni, in eccesso o in difetto rispetto al valore reale, delle singole determinazioni e sono dovuti alle molte variabili incontrollabili che influenzano ogni misura psicologica. Gli errori sistematici, invece, influiscono sulla misurazione sempre nello stesso senso e, solitamente, per una stessa quantità (possono essere additivi o proporzionali).

Le differenze tra le due tipologie di errori, sistematici e casuali, introducono i concetti di accuratezza e di precisione della misura. Una misura viene definita:

- *accurata*, quando vi è un accordo tra la misura effettuata ed il valore reale;
- *precisa* quando, ripetendo più volte la misura, i risultati ottenuti sono concordanti, cioè differiscono in maniera irrilevante tra loro.

La metafora del tiro a bersaglio illustra la relazione tra precisione e accuratezza.

Per tenere sotto controllo l'incidenza degli errori, sono stati introdotti in psicologia i concetti di attendibilità e validità.

Uno strumento si dice *attendibile* quando valuta in modo coerente e stabile la stessa variabile: i risultati ottenuti si mantengono costanti dopo ripetute somministrazione ed in assenza di variazioni psicologiche e fisiche dei soggetti sottoposti al test o cambiamenti dell'ambiente in cui ha luogo la somministrazione.

L'attendibilità di uno strumento, però, non è sufficiente: in primo luogo uno strumento di misura deve essere *valido*, laddove la validità rappresenta il grado in cui uno strumento misura effettivamente ciò che dovrebbe misurare. In genere, si fa riferimento ad almeno quattro tipi di validità.

- La *validità di costrutto* riguarda il grado in cui un test misura ciò per cui è stato costruito. Essa si suddivide in: validità convergente e validità divergente. La validità convergente fa riferimento alla concordanza tra uno strumento e un altro che misura lo stesso costrutto. La validità divergente, al contrario, valuta il grado di discriminazione tra strumenti che misurano costrutti differenti. Senza validità di costrutto le altre forme di validità non hanno senso.

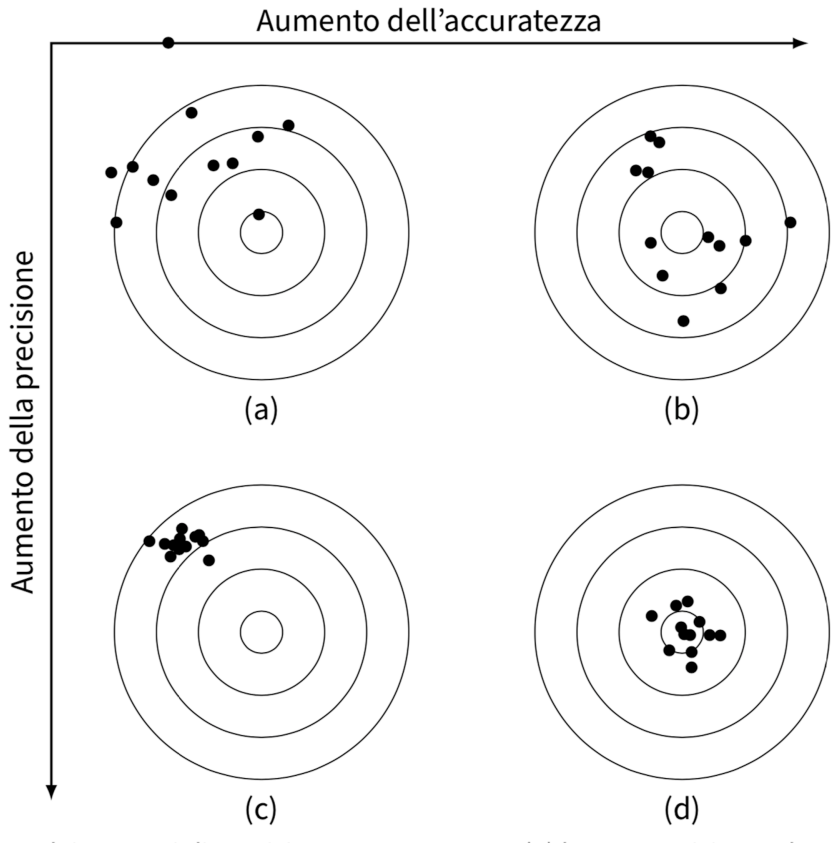


Figura 2.1: *Metafora del tiro al bersaglio.*

- In base alla *validità di contenuto*, un test fornisce una misura valida di un attributo psicologico se il dominio dell'attributo è rappresentato in maniera adeguata dagli item del test. Un requisito di base della validità di contenuto è la rilevanza e la rappresentatività del contenuto degli item in riferimento all'attributo che il test intende misurare.
- La *validità di criterio* valuta il grado di concordanza tra i risultati dello strumento considerato e i risultati ottenuti da altri strumenti che misurano lo stesso costrutto, o tra i risultati dello strumento considerato e un criterio esterno. Nella validità concorrente, costrutto e criterio vengono misurati contestualmente, consentendo un confronto immediato. Nella validità predittiva, il costrutto viene misurato prima e il criterio in un momento successivo, consentendo la valutazione della capacità dello strumento di predire un evento futuro.

- Infine, la *validità di facciata* fa riferimento al grado in cui il test appare valido ai soggetti a cui esso è diretto. La validità di facciata è importante in ambiti particolari, quali ad esempio la selezione del personale per una determinata occupazione. In questo caso è ovviamente importante che chi si sottopone al test ritenga che il test vada a misurare quegli aspetti che sono importanti per le mansioni lavorative che dovranno essere svolte, piuttosto che altre cose. In generale, la validità di facciata non è utile, tranne in casi particolari.

Conclusioni

Una domanda che uno psicologo spesso si pone è: "sulla base delle evidenze osservate, possiamo concludere dicendo che l'intervento psicologico è efficace nel trattamento e nella cura del disturbo?" Le considerazioni svolte in questo capitolo dovrebbero farci capire che, prima di cercare di rispondere a questa domanda con l'analisi statistica dei dati, devono essere affrontati i problemi della validità e dell'attendibilità delle misure (oltre a stabilire l'appropriato livello di scala di misura delle osservazioni). L'attendibilità è un prerequisito della validità. Se gli errori di misurazione sono troppo grandi, i dati sono inutili. Inoltre, uno strumento di misurazione può essere preciso ma non valido. La validità e l'attendibilità delle misurazioni sono dunque entrambe necessarie.

In generale, l'attendibilità e la validità delle misure devono essere valutate per capire se i dati raccolti da un ricercatore siano adeguati (1) per fornire una risposta alla domanda della ricerca, e (2) per giungere alla conclusione proposta dal ricercatore alla luce dei risultati dell'analisi statistica che è stata eseguita. È chiaro che le informazioni fornite in questo capitolo si limitano a scalfire la superficie di questi problemi. I concetti qui introdotti, però, devono sempre essere tenuti a mente e costituiscono il fondamento di quanto verrà esposto nei capitoli successivi.

Nozioni di base

3

Il calcolo delle probabilità

Una possibile definizione di teoria delle probabilità è la seguente: la teoria delle probabilità ci fornisce gli strumenti per prendere decisioni razionali in condizioni di incertezza, ovvero per formulare le migliori congetture possibili.

3.1 La probabilità come la logica della scienza

La figura 3.1 fornisce una rappresentazione schematica del processo dell'indagine scientifica. Possiamo pensare al progresso scientifico come alla ripetizione di questo ciclo, laddove i fenomeni naturali (e, ovviamente psicologici) vengono esplorati e i ricercatori imparano sempre di più sul loro funzionamento. Le caselle della figura descrivono le varie fasi del processo di ingagine scientifica, mentre lungo le frecce sono riportati i compiti che conducono i ricercatori da una fase alla successiva.

Consideriamo i compiti e le fasi dell'indagine scientifica. Iniziamo in basso a sinistra.

- *Invenzione e perfezionamento delle ipotesi.* In questa fase del processo scientifico, i ricercatori pensano ai fenomeni naturali, a ciò che è presente nella letteratura scientifica, ai risultati dei loro esperimenti, e formulano ipotesi o teorie che possono essere valutare mediante esperimenti empirici. Questo passaggio richiede innovazione e creatività.
- L'*inferenza deduttiva* procede in maniera deterministica dai fatti alle conclusioni. Ad esempio, se dico che tutti gli uomini sono mortali e che Socrate è un uomo, allora posso concludere deduttivamente che Socrate è mortale. Quando i ricercatori progettano gli esperimenti in base alle teorie, usano la logica deduttiva per dire: "Se A è vero, allora

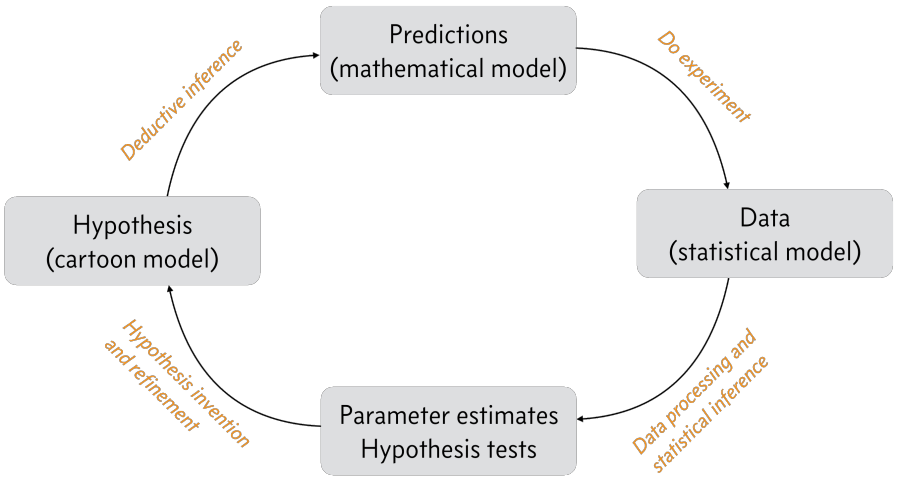


Figura 3.1: *Rappresentazione schematica del processo scientifico (figura adattata dalla Fig. 1.1 di P. Gregory, Bayesian Logical Data Analysis for the Physical Sciences, Cambridge, 2005).*

B deve essere vero", dove A è l'ipotesi teorica e B è l'osservazione sperimentale.

- *Esecuzione degli esperimenti.* Questa fase richiede molte risorse (tempo e denaro). Richiede anche innovazione e creatività. Nello specifico, i ricercatori devono pensare attentamente a come costruire l'esperimento necessario per verificare la teoria di interesse. Quale risultato dell'esperimento si ottengono i dati.
- L'*inferenza induttiva* procede dalle osservazioni ai fatti. Se pensiamo ai fatti come a ciò che governa o genera le osservazioni, allora l'induzione è una sorta di inferenza inversa. Supponiamo di avere osservato B. Questo rende A vero? Non necessariamente. Ma può rendere A più plausibile. Questo è un sillogismo debole. Ad esempio, si consideri la seguente coppia ipotesi/osservazioni.
  - A = L'iniezione di acque reflue dopo la fratturazione idraulica, nota come fracking, può portare a una maggiore frequenza di terremoti.
  - B = La frequenza dei terremoti in Oklahoma è aumentata di 100 volte dal 2010, quando il fracking è diventato una pratica comune.

  - Poiché B è stato osservato, A è più plausibile. A non è necessariamente vero, ma è più plausibile.

- L'*inferenza statistica* è un tipo di inferenza induttiva che è specificamente formulata come un problema inverso. L'inferenza statistica è quell'insieme di procedure che hanno lo scopo di quantificare quanto più plausibile sia A dopo aver osservato B. Per svolgere l'inferenza statistica è dunque necessario quantificare tale plausibilità. Lo strumento che ci consente di fare questo è la teoria delle probabilità.

L'inferenza statistica è l'aspetto del processo dell'indagine scientifica che è l'oggetto centrale di questo insegnamento. Il risultato dell'inferenza statistica è la conoscenza di quanto siano plausibili le ipotesi e le stime dei parametri sotto le ipotesi considerate. Ma l'inferenza statistica richiede una teoria delle probabilità, laddove la teoria delle probabilità può essere vista come una generalizzazione della logica. A causa di questa connessione con la logica e del suo ruolo cruciale nella scienza, E. T. Jaynes afferma infatti che la probabilità è la "logica della scienza". È dunque necessario esaminare preliminarmente alcune nozioni di base della teoria delle probabilità.

3.2 Che cos'è la probabilità?

La definizione della probabilità è un problema estremamente dibattuto ed aperto. Sono state fornite due possibili soluzioni al problema di definire il concetto di probabilità.

(a) La natura della probabilità è "ontologica" (ovvero, basata sulla metafisica): la probabilità è una proprietà della della realtà, del mondo, di come sono le cose, indipendentemente dalla nostra esperienza. È una visione che qualcuno chiama "oggettiva".

(b) La natura della probabilità è "epistemica" (ovvero, basata sulla conoscenza): la probabilità si riferisce alla conoscenza che abbiamo del mondo, non al mondo in sé. Di conseguenza è detta, in contrapposizione alla precedente definizione, "soggettiva".

In termini epistemici, la probabilità fornisce una misura della nostra incertezza sul verificarsi di un fenomeno, alla luce delle informazioni di-

sponibili. Potremmo dire che c'è una "scala" naturale che ha per estremi il vero (1: evento certo) da una parte ed il falso (0: evento impossibile) dall'altra. La probabilità è la quantificazione di questa scala: quantifica lo stato della nostra incertezza rispetto al contenuto di verità di una proposizione (ovvero, quantifica la plausibilità di una proposizione).

- Nell'interpretazione frequentista della probabilità, la probabilità $P(A)$ rappresenta la frequenza relativa a lungo termine nel caso di un grande numero di ripetizioni di un esperimento casuale sotto le medesime condizioni. L'evento A deve essere una proposizione relativa alle variabili casuali[1].
- Nell'interpretazione bayesiana della probabilità $P(A)$ rappresenta il grado di credenza, o plausibilità, a proposito di A, dove A può essere qualsiasi proposizione logica.

In questo insegnamento utilizzeremo l'interpretazione bayesiana della probabilità. Possiamo citare De Finetti, ad esempio, il quale ha formulato la seguente definizione "soggettiva" di probabilità la quale risulta applicabile anche ad esperimenti casuali i cui eventi elementari non siano ritenuti ugualmente possibili e che non siano necessariamente ripetibili più volte sotto le stesse condizioni:

Definizione 3.1. La probabilità di un evento E è la quota $p(E)$ che un individuo reputa di dover pagare ad un banco per ricevere "1" ovvero "0" verificandosi o non verificandosi E. Le valutazioni di probabilità degli eventi devono rispondere ai pricipi di equità e coerenza.

I principi di equità e coerenza sono definiti come segue.

Definizione 3.2. Una scommessa risponde ai pricipi di *equità* se il ruolo di banco e giocatore sono scambiabili in ogni momento del gioco e sempre alle stesse condizioni; *coerenza* se non vi sono combinazioni di scommesse che consentano (sia al banco che al giocatore) di realizzare perdite o vincite certe.

[1] Viene stressata qui l'idea che ciò di cui parliamo è qualcosa che emerge nel momento in cui è possibile ripetere l'esperimento casuale tante volte sotto le medesime condizioni. Le variabili casuali, infatti, forniscono una quantificazione dei risultati che si ottengono ripetendo tante volte l'esperimento casuale sotto le medesime condizioni.

Secondo de Finetti (1931), *nessuna scienza ci permetterà di dire: il tale fatto accadrà, andrà così e così, perché ciò è conseguenza di tale legge, e tale legge è una verità assoluta, ma tanto meno ci condurrà a concludere scetticamente: la verità assoluta non esiste, e quindi tale fatto può accadere e può non accadere, può andare così e può andare in tutt'altro modo, nulla io ne so. Quel che si potrà dire è questo: io prevedo che il tale fatto avverrà, e avverrà nel tal modo, perché l'esperienza del passato e l'elaborazione scientifica cui il pensiero dell'uomo l'ha sottoposta mi fanno sembrare ragionevole questa previsione.*

In altri termini, de Finetti ritiene che la probabilità debba essere concepita non come una proprietà "oggettiva" dei fenomeni ("la probabilità di un fenomeno ha un valore determinato che dobbiamo solo scoprire"), ma bensì come il "grado di fiducia – in inglese *degree of belief* – di un dato soggetto, in un dato istante e con un dato insieme d'informazioni, riguardo al verificarsi di un evento". Per denotare sia la probabilità (soggettiva) di un evento sia il concetto di *valore atteso* (che descriveremo in seguito), de Finetti (1970) utilizza il termine "previsione" (e lo stesso simbolo P): *la previsione [...] consiste nel considerare ponderatamente tutte le alternative possibili per ripartire fra di esse nel modo che parrà più appropriato le proprie aspettative, le proprie sensazioni di probabilità.*

3.3 Variabili casuali e probabilità di un evento

Esaminiamo qui di seguito alcuni concetti di base della teoria delle probabilità.

3.3.1 Variabili casuali

Sia Y il risultato del lancio di moneta equilibrata, non di un generico lancio di una moneta, ma un'istanza specifica del lancio di una specifica moneta in un dato momento. Definita in questo modo, Y è una *variabile casuale*, ovvero una variabile che assume valori diversi con probabilità diverse. Se la moneta è equilibrata, c'è una probabilità del 50% che il lancio della moneta dia come risultato "testa" e una probabilità del 50% che dia come risultato "croce".

Per facilitare la trattazione, le variabili casuali assumono solo valori numerici. Per lo specifico lancio della moneta in questione, diciamo, ad

esempio, che la variabile casuale Y assume il valore 1 se esce testa e il valore 0 se esce croce.

3.3.2 Eventi e probabilità

Nella teoria delle probabilità il risultato "testa" nel lancio di una moneta è chiamato *evento*.[2] Ad esempio, $Y = 1$ denota l'evento in cui il lancio di una moneta produce come risultato testa.

Il funzionale $Pr[\cdot]$ definisce la probabilità di un evento. Ad esempio, per il lancio di una moneta equilibrata, la probabilità dell'evento "il risultato del lancio della moneta è testa" è scritta come

$$Pr[Y = 1] = 0.5.$$

Se la moneta è equilibrata dobbiamo anche avere $Pr[Y = 0] = 0.5$. I due eventi $Y = 1$ e $Y = 0$ sono *mutuamente esclusivi* nel senso che non possono entrambi verificarsi contemporaneamente. Nella notazione probabilistica,

$$Pr[Y = 1 \; e \; Y = 0] = 0.$$

Gli eventi $Y = 1$ e $Y = 0$ di dicono *esaustivi*, nel senso che almeno uno di essi deve verificarsi e nessun altro tipo di evento è possibile. Nella notazione probabilistica,

$$Pr[Y = 1 \; o \; Y = 0] = 1.$$

Il connettivo logico "e" specifica eventi *congiunti*, ovvero eventi che possono verificarsi contemporaneamente (eventi *compatibili*) e per i quali, perciò, la probabilità della loro congiunzione è $Pr(A \; e \; B) > 0$. Il connettivo logico "o" specifica eventi *disgiunti*, ovvero eventi che non possono verificarsi contemporaneamente (eventi *incompatibili*) e per i quali, perciò, la probabilità della loro congiunzione è $P(A \; e \; B) = 0$.

[2] Per un ripasso delle nozioni di base della teoria degli insiemi, si veda l'Appendice **??**.

3.4 Spazio campionario e risultati possibili

Anche se il lancio di una moneta produce sempre uno specifico risultato nel mondo reale, noi possiamo anche immaginare i possibili risultati alternativi che si sarebbero potuti osservare. Quindi, anche se in uno specifico lancio la moneta dà testa ($Y = 1$), possiamo immaginare la possibilità che il lancio possa avere prodotto croce ($Y = 0$). Tale ragionamento controfattuale è la chiave per comprendere la teoria delle probabilità e l'inferenza statistica.

I risultati possibili che si possono osservare come conseguenza del lancio di una moneta determinano i valori possibili che la variabile casuale può assumere. L'insieme di tutti i risultati possibili è chiamato *spazio campionario*. Lo spazio campionario può essere concettualizzato come un'urna contenente una pallina per ogni possibile risultato del lancio della moneta. Su ogni pallina è scritto il valore della variabile casuale. Uno specifico lancio di una moneta – ovvero, l'osservazione di uno specifico valore di una variabile casuale – è chiamato *esperimento casuale*.

Il lancio di un dado ci fornisce l'esempio di un altro esperimento casuale. Supponiamo di essere interessati all'evento "il lancio del dado produce un numero dispari". Un *evento* seleziona un sottoinsieme dello spazio campionario: in questo caso, l'insieme dei risultati $\{1, 3, 5\}$. Se esce 3, per esempio, diciamo che si è verificato l'evento "dispari" (ma l'evento "dispari" si sarebbe anche verificato anche se fosse uscito 1 o 5).

3.5 Usare la simulazione per stimare le probabilità

I metodi basati sulla simulazione ci consentono di stimare le probabilità degli eventi in un modo diretto se siamo in grado di generare realizzazioni molteplici e casuali delle variabili casuali coinvolte nelle definizioni degli eventi. Per simulare il lancio di una moneta equilibrata in R iniziamo a definire un vettore che contiene i possibili risultati del lancio della moneta (ovvero i possibili valori della variabile casuale Y):

```
coin <- c(0, 1)
```

L'estrazione casuale di uno di questi due possibili valori (ovvero, la simulazione di uno specifico lancio di una moneta) si realizza con la funzione `sample()`:

```
sample(coin, size = 1)
#> [1] 0
```

In maniera equivalente, lo stesso risultato si ottiene mediante l'istruzione

```
rbinom(1, 1, 0.5)
#> [1] 1
```

Supponiamo di ripetere questo esperimento casuale 100 volte e di registrare i risultati così ottenuti. La stima della probabilità dell'evento $Pr[Y = 1]$ è data dalla frequenza relativa del numero di volte in cui abbiamo osservato l'evento di interesse $(Y = 1)$:

```
M <- 10
y <- rep(NA, M)
for (m in 1:M) {
  y[m] = rbinom(1, 1, 0.5)
}
estimate = sum(y) / M

cat("estimated Pr[Y = 1] =", estimate)
#> estimated Pr[Y = 1] = 0.5
```

Ripetiamo questa procedura 10 volte.

```
flip_coin <- function(M) {
  y <- rep(NA, M)
  for (m in 1:M) {
    y[m] = rbinom(1, 1, 0.5)
  }
  estimate <- sum(y) / M
```

```
  cat("estimated Pr[Y = 1] =", estimate, "\n")
}
```

```
for(i in 1:10) {
  flip_coin(10)
}
#> estimated Pr[Y = 1] = 0.5
#> estimated Pr[Y = 1] = 0.3
#> estimated Pr[Y = 1] = 0.7
#> estimated Pr[Y = 1] = 0.5
#> estimated Pr[Y = 1] = 0.5
#> estimated Pr[Y = 1] = 0.6
#> estimated Pr[Y = 1] = 0.5
#> estimated Pr[Y = 1] = 0.8
#> estimated Pr[Y = 1] = 0.4
#> estimated Pr[Y = 1] = 0.5
```

Dato che la moneta è equilibrata, la stima delle probabilità dell'evento $Pr[Y = 1]$ è simile a al valore che ci aspettiamo ($Pr[Y = 1] = 0.5$), ma il risultato ottenuto nelle varie simulazioni non è sempre esatto. Proviamo ad aumentare il numero di lanci in ciascuna simulazione:

```
for(i in 1:10) {
  flip_coin(100)
}
#> estimated Pr[Y = 1] = 0.44
#> estimated Pr[Y = 1] = 0.53
#> estimated Pr[Y = 1] = 0.43
#> estimated Pr[Y = 1] = 0.58
#> estimated Pr[Y = 1] = 0.5
#> estimated Pr[Y = 1] = 0.41
#> estimated Pr[Y = 1] = 0.51
#> estimated Pr[Y = 1] = 0.49
#> estimated Pr[Y = 1] = 0.5
#> estimated Pr[Y = 1] = 0.57
```

In questo secondo caso, gli errori tendono ad essere più piccoli della simulazione precedente. Cosa succede se in ciascuna simulazione esaminiamo

i risultati di 10,000 lanci della moneta?

```
for(i in 1:10) {
  flip_coin(1e4)
}
#> estimated Pr[Y = 1] = 0.5029
#> estimated Pr[Y = 1] = 0.4886
#> estimated Pr[Y = 1] = 0.4956
#> estimated Pr[Y = 1] = 0.49
#> estimated Pr[Y = 1] = 0.5032
#> estimated Pr[Y = 1] = 0.5051
#> estimated Pr[Y = 1] = 0.4928
#> estimated Pr[Y = 1] = 0.4968
#> estimated Pr[Y = 1] = 0.4991
#> estimated Pr[Y = 1] = 0.4976
```

Ora le stime ottenute sono molto vicine alla vera probabilità che vogliamo stimare (cioè 0.5, perché la moneta è equilibrata). I risultati delle simulazioni precedenti pongono dunque il problema di determinare quale sia il numero di lanci di cui abbiamo bisogno per assicurarci che le stime siano accurate (ovvero, vicine al valore corretto della probabilità)

3.6 La legge dei grandi numeri

La visualizzazione mediante grafici contribuisce alla comprensione dei concetti della statistica e della teoria delle probabilità. Un modo per descrivere ciò che accade all'aumentare del numero M di ripetizioni del lancio della moneta consiste nel registrare la stima della probabilità dell'evento $Pr[Y = 1]$ in funzione del numero di ripetizioni dell'esperimento casuale per ogni $m \in 1 : M$. Un grafico dell'andamento della stima di $Pr[Y = 1]$ in funzione di m si ottiene nel modo seguente.

```
nrep <- 1e4
estimate <- rep(NA, nrep)
flip_coin <- function(m) {
  y <- rbinom(m, 1, 0.5)
```

```
  phat <- sum(y) / m
  phat
}
for(i in 1:nrep) {
  estimate[i] <- flip_coin(i)
}
d <- data.frame(
  n = 1:nrep,
  estimate
)
d %>%
  ggplot(
    aes(x = n, y = estimate)
  ) +
  geom_line() +
  theme(legend.title = element_blank()) +
  labs(
    x = "Numero di lanci della moneta",
    y = "Stima Pr[Y = 1]"
)
```

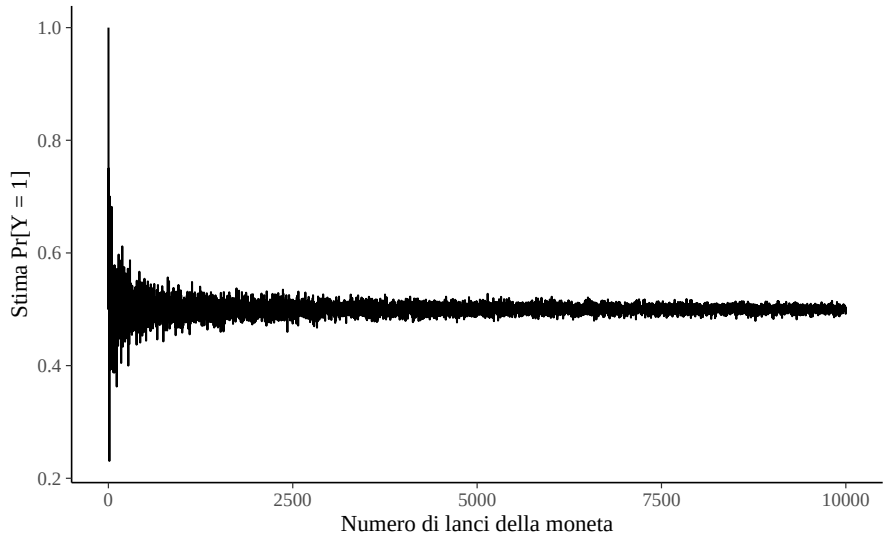


Figura 3.2: *Stima della probabilità di successo in funzione del numero di lanci di una moneta.*

Dato che il grafico 3.2 su una scala lineare non rivela chiaramente l'andamento della simulazione, utilizzeremo invece un grafico in cui sull'asse x è stata imposta una scala logaritmica. Con l'asse x su scala logaritmica, i valori tra 1 e 10 vengono tracciati all'incirca con la stessa ampiezza come nel caso dei valori tra 50 e 700, eccetera.

```
d %>%
  ggplot(
    aes(x = n, y = estimate)
  ) +
  geom_line() +
  scale_x_log10(
    breaks = c(1, 3, 10, 50, 200,
               700, 2500, 10000)
  ) +
  theme(legend.title = element_blank()) +
  labs(
    x = "Numero di lanci della moneta",
    y = "Stima Pr[Y = 1]"
)
```

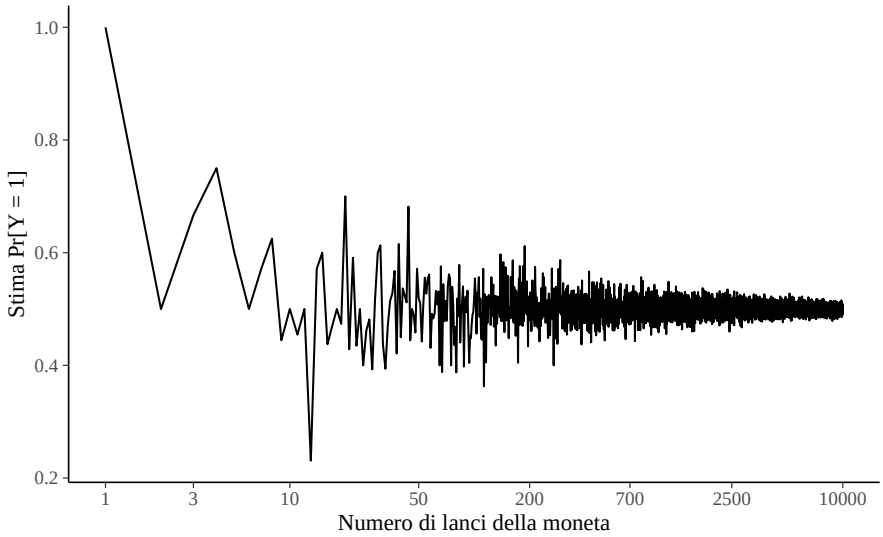


Figura 3.3: *Stima della probabilità di successo in funzione del numero di lanci di una moneta – scala logaritmica.*

La *legge dei grandi numeri* ci dice che all'aumentare del numero di ripetizioni dell'esperimento casuale la media dei risultati ottenuti tenderà ad avvicinarsi al valore atteso man mano che verranno eseguite più prove. Nel caso presente, la figura 3.3 mostra appunto che, all'aumentare del numero M di lanci della moneta, la stima di $Pr[Y = 1]$ tende a convergere al vero valore di 0.5.

3.7 Variabili casuali multiple

Le variabili casuali non esistono isolatamente. Abbiamo iniziato con una singola variabile casuale Y che rappresenta il risultato di un singolo, specifico lancio di una moneta equlibrata. Ma supponiamo ora di lanciare la moneta tre volte. Ciò suggerisce che possiamo avere le variabili casuali Y_1, Y_2, Y_3 che rappresentano i risultati di ciascuno dei lanci. Possiamo assumere che ogni lancio sia indipendente, ovvero che non dipenda dal risultato degli altri lanci. Ognuna di queste variabili Y_n per $n \in 1:3$ ha $Pr[Y_n = 1] = 0.5$ e $Pr[Y_n = 0] = 0.5$. Possiamo combinare più variabili casuali usando le operazioni aritmetiche. Se Y_1, Y_2, Y_3 sono variabili casuali che rappresentano tre lanci di una moneta equilibrata (o un lancio di tre monete equilibrate), possiamo definire la somma di tali variabili casuali come

$$Z = Y_1 + Y_2 + Y_3.$$

Possiamo simulare i valori assunti dalla variabile casuale Z simulando i valori di Y_1, Y_2, Y_3 per poi sommarli.

```
y1 <- rbinom(1, 1, 0.5)
y2 <- rbinom(1, 1, 0.5)
y3 <- rbinom(1, 1, 0.5)
c(y1, y2, y3)
#> [1] 0 0 1
z <- sum(c(y1, y2, y3))
cat("z =", z, "\n")
#> z = 1
```

ovvero,

```
y <- rep(NA, 3)
for (i in 1:3) {
  y[i] <- rbinom(1, 1, 0.5)
}
y
#> [1] 1 0 0
z <- sum(y)
cat("z =", z, "\n")
#> z = 1
```

oppure, ancora più semplicemente:

```
y <- rbinom(3, 1, 0.5)
y
#> [1] 0 1 1
z <- sum(y)
cat("z =", z, "\n")
#> z = 2
```

Possiamo ripetere questa simulazione $M = 1e5$ volte:

```
M <- 1e5
z <- rep(NA, M)
for(i in 1:M) {
  y <- rbinom(3, 1, 0.5)
  z[i] <- sum(y)
}
```

e calcolare una stima della probabilità che la variabile casuale Z assuma i valori 0, 1, 2, 3:

```
table(z) / M
#> z
#>      0      1      2      3
#> 0.1256 0.3750 0.3749 0.1245
```

Nel caso di 4 monete equilibrate, avremo:

```
M <- 1e5
z <- rep(NA, M)
for(i in 1:M) {
  y <- rbinom(4, 1, 0.5)
  z[i] <- sum(y)
}
table(z) / M
#> z
#>       0       1       2       3       4
#> 0.06213 0.25019 0.37400 0.25097 0.06271
```

Viene detta *variabile casuale discreta* una variabile casuale le cui modalità possono essere costituite solo da numeri interi:

$$\mathbb{Z} = \dots, -2, -1, 0, 1, 2, \dots$$

3.8 Funzione di massa di probabilità

È conveniente avere una funzione che associa ogni possibile valore di una variabile casuale alla sua probabilità. In generale, ciò è possibile se e solo se la variabile casuale è discreta, così com'è stata definita nel Paragrafo precedente.

Ad esempio, se consideriamo $Z = Y_1 + \cdots + Y_4$ come il numero di risultati "testa" in 4 lanci della moneta, allora possiamo definire la seguente funzione:

$$\begin{array}{lcll} p_Z(0) & = & 1/16 & \text{TTTT} \\ p_Z(1) & = & 4/16 & \text{HTTT, THTT, TTHT, TTTH} \\ p_Z(2) & = & 6/16 & \text{HHTT, HTHT, HTTH, THHT, THTH, TTTH} \\ p_Z(3) & = & 4/16 & \text{HHHT, HHTH, HTHH, THHH} \\ p_Z(4) & = & 1/16 & \text{HHHH} \end{array}$$

Il lancio di quattro monete può produrre sedici possibili risultati. Dato che i lanci sono indipendenti e le monete sono equilibrate, ogni possibile risultato è ugualmente probabile. Nella tabella in alto, le sequenze dei risultati possibili del lancio delle 4 monete sono riportate nella colonna

più a destra. Le probabilità si ottengono dividendo il numero di sequenze che producono lo stesso numero di eventi testa per il numero dei risultati possibili.

La funzione p_Z è stata costruita per mappare un valore u per Z alla probabilità dell'evento $Z = u$. Convenzionalmente, queste probabilità sono scritte come

$$p_Z(z) = \Pr[Z = z].$$

La parte a destra dell'uguale si può leggere come: "la probabilità che la variabile casuale Z assuma il valore z".

Una funzione definita come sopra è detta *funzione di massa di probabilità* della variabile casuale Z. Ad ogni variabile casuale discreta è associata un'unica funzione di massa di probabilità.

Una rappresentazione grafica della stima della funzione di massa di probabilità per l'esperimento casuale del lancio di quattro monete equilibrate è fornita nella figura 3.4.

```
set.seed(1234)
M <- 1e5
nflips <- 4
u <- rbinom(M, nflips, 0.5)
x <- 0:nflips
y <- rep(NA, nflips+1)
for (n in 0:nflips)
  y[n + 1] <- sum(u == n) / M
bar_plot <-
  data.frame(Z = x, count = y) %>%
  ggplot(
    aes(x = Z, y = count)
  ) +
  geom_bar(stat = "identity") +
  scale_x_continuous(
    breaks = 0:4,
    labels = c(0, 1, 2, 3, 4)
  ) +
  labs(
```

```
    y = "Probabilità stimata Pr[Z = z]"
)
bar_plot
```

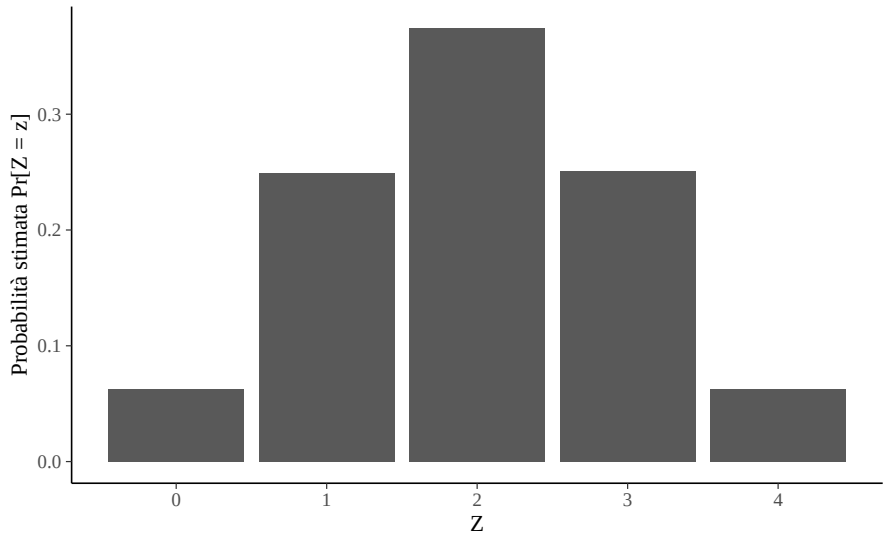


Figura 3.4: *Grafico di $M = 100\,000$ simulazioni della funzione di massa di probabilità di una variabile casuale definita come il numero di teste in quattro lanci di una moneta equilibrata.*

Se A è un sottoinsieme della variabile casuale Z, allora denotiamo con $P_z(A)$ la probabilità assegnata ad A dalla distribuzione P_z. Mediante una distribuzione di probabilità P_z è dunque possibile determinare la probabilità di ciascun sottoinsieme $A \subset Z$ come

$$P_z(A) = \sum_{z \in A} P_z(Z).$$

Esempio 3.1. Nel caso dell'esempio discusso nella Sezione 3.8, la probabilità che la variabile casuale Z sia un numero dispari è $Pr(\text{Z è un numero dispari}) = P_z(Z = 1) + P_z(Z = 3) = \frac{4}{16} + \frac{4}{16} = \frac{1}{2}$.

Considerazioni conclusive

In questo capitolo abbiamo visto come si costruisce lo spazio campionario di un esperimento casuale, quali sono le proprietà di base della probabilità e come si assegnano le probabilità agli eventi definiti sopra uno spazio campionario discreto. Abbiamo anche introdotto le nozioni di "variabile casuale", ovvero di una variabile che prende i suoi valori casualmente. E abbiamo descritto il modo di specificare la probabilità con cui sono presi i differenti valori, ovvero la funzione di distribuzione probabilistica $F(X) = Pr(X < x)$, e la funzione di massa di probabilità. Le procedure di analisi dei dati psicologici che discuteremo in seguito faranno un grande uso di questi concetti e della notazione qui introdotta.

4

Distribuzione predittiva a posteriori

Oltre ad una sintesi della distribuzione a posteriori attraverso il computo di indici caratteristici e alla verifica di ipotesi, un altro compito dell'analisi bayesiana è la predizione di nuovi dati futuri. Dopo aver osservato i dati di un campione e ottenuto le distribuzioni a posteriori dei parametri, è infatti possibile ottenere una qualche indicazione su come potrebbero essere i dati futuri. L'uso più immediato della stima della distribuzione dei possibili valori futuri della variabile di esito è la verifica del modello. Infatti, il modo più diretto per testare un modello è quello di utilizzare il modello per fare previsioni sui possibili dati futuri per poi confrontare i dati predetti con i dati effettivi. Questa pratica va sotto il nome di controllo predittivo a posteriori.

4.1 La distribuzione dei possibili valori futuri

La distribuzione dei possibili valori futuri della variabile di esito può essere predetta da un modello statistico sulla base della distribuzione a posteriori dei parametri, $p(\theta \mid y)$, avendo già osservato n manifestazioni del fenomeno y. Una tale distribuzione va sotto il nome di *distribuzione predittiva a posteriori* (*posterior predictive distribution*, PPD).

Quando vengono simulate le osservazioni della distribuzione predittiva a posteriori si usa la notazione y^{rep} (dove *rep* sta per *replica*) quando, nella simulazione, vengono utilizzate le stesse osservazioni di X che erano state usate per stimare i parametri del modello. Si usa invece la notazione $\tilde{y}$ per fare riferimento a possibili valori X che non sono contenuti nel campione osservato, ovvero, ad un campione di dati che potrebbe essere osservato in qualche futura occasione.

La distribuzione predittiva a posteriori viene usata per fare inferenze predittive. L'idea è che, se il modello ben si adatta bene ai dati del

campione allora, sulla base dei parametri stimati, dovremmo essere in grado di generare nuovi dati non osservati y^{rep} che risultano molto simili ai dati osservati y. I dati y^{rep} vengono concepiti come stime di $\tilde{y}$. La distribuzione predittiva a posteriorie è data da:

$$p(\tilde{y} \mid y) = \int_{\theta} p(\tilde{y}, \theta \mid y) d\theta = \int_{\theta} p(\tilde{y} \mid \theta, y) p(\theta \mid y) d\theta.$$

Supponendo che le osservazioni passate e future siano condizionalmente indipendenti dato θ, ovvero che $p(\tilde{y} \mid \theta, y) = p(\tilde{y} \mid \theta)$, possiamo scrivere

$$p(\tilde{y} \mid y) = \int_{\theta} p(\tilde{y} \mid \theta) p(\theta \mid y) d\theta. \tag{4.1}$$

La (4.1) descrive la nostra incertezza sulla distribuzione di future osservazioni di dati, data la distribuzione a posteriori di θ, ovvero tenendo conto della scelta del modello e della stima dei parametri mediante i dati osservati. Si noti che, nella (4.1), $\tilde{y}$ è condizionato da y ma non da ciò che è incognito, ovvero θ. La distribuzione predittiva a posteriori è invece ottenuta mediante marginalizzazione sopra le variabili da "scartare", ovvero sopra i parametri incogniti θ.

Un esempio formulato mediante il codice Stan può chiarire questo concetto. Consideriamo il codice relativo alla distribuzione predittiva a posteriori nel caso di un modello di regressione lineare classico con un solo predittore x. Il blocco *Model* sarà:

```
model {
 y ~ normal(x * beta + alpha, sigma);
}
```

Quello che è di interesse per la discussione presente è il blocco *Generated Quantities*. Tale blocco avrà questa forma:

```
generated quantities {
 real y_rep[N];
 for (n in 1:N) {
   y_rep[n] = normal_rng(x[n] * beta + alpha, sigma);
 }
}
```

La variabile `y_rep` è ciò a cui siamo interessati. Nel codice precedente, `x` è il vettore che contiene i valori della variabile indipendente nel campione di osservazioni esaminato. I parametri del modello di regressione sono `alpha` e `beta`; `sigma` è la stima dell'errore standard della regressione. Supponiamo che questi tre parametri siano degli scalari. Se lo fossero, per il valore `x` n-esimo, l'istruzione `normal_rng()` ritornerebbe un valore a caso dalla distribuzione normale con media $\alpha + \beta x_n$ e deviazione standard σ. Il ciclo `for()` ripete questa operazione N volte, ovvero tante volte quanti sono gli elementi del vettore `x` del campione. Quello che è stato detto sopra ci dà un'idea di quello che succederebbe se `alpha`, `beta` e `sigma` fossero degli scalari. Ma non lo sono. Per ciascuno dei tre paramatri abbiamo un numero molto alto di stime, ovvero l'approssimazione MCMC della distribuzione a posteriori. Poniamo che l'ampiezza campionaria N sia 30. Se `alpha`, `beta` e `sigma` fossero degli scalari, la distribuzione predittiva a posteriori sarebbe costituita da 30 valori y^{rep}, ovvero, non sarebbe nient'altro che $\hat{y} = \hat{\alpha} + \hat{\beta}x$. Ma `alpha`, `beta` e `sigma` non sono degli scalari: per ciascuno di questi parametri abbiamo un grande numero di stime, diciamo 2000. Dunque, quando `normal_rng()` estrae un valore a caso dalla distribuzione normale, i parametri della normale non sono fissi: per determinare μ prendiamo un valore a caso, chiamiamolo `beta'`, dalla distribuzione dei valori `beta` e un valore a caso, chiamiamolo `alpha'`, dalla distribuzione dei valori `alpha`. Avendo questi due valori, calcoliamo il valore $\mu'_n = \alpha' + \beta' x_n$. Lo stesso si può dire per σ'. A questo punto possiamo trovare il valore `y_n'` estraendo un valore a caso dalla distribuzione gaussiana di parametri μ' e σ'. Per l'n-esimo valore x possiamo ripetere questo processo tante volte. Se lo ripetiamo, ad esempio, 2,000 volte, per tutti e 30 i valori x del campione otterremo una matrice $30 \times 2,000$. In questo modo possiamo generare le previsioni del modello, ovvero y^{rep}, che includono due fonti di incertezza:

- la variabilità campionaria, ovvero il fatto che abbiamo osservato uno specifico insieme di valori (x, y); in un altro campione tali valori saranno diversi;
- la variabilità a posteriori della distribuzione dei parametri, ovvero il fatto che di ciascun parametro non conosciamo il "valore vero" ma solo una distribuzione (a posteriori) di valori.

Nel caso dell'esempio presente, l'integrale della (4.1) può essere interpretato dicendo che, nell'esempio della matrice di dimensioni $30 \times 2,000$, noi marginalizziamo rispetto alle colonne, ovvero, per ciascuna riga facciamo la media dei valori colonna. Otteniamo così un vettore di 30 osservazioni,

ovvero y^{rep}.

Quando, con metodi grafici, vengono esaminati i valori della distribuzione predittiva a posteriori, possiamo esaminare un numero arbitrario di previsioni. Per esempio, possiamo rappresentare graficamente 50 rette di regressione predette – o un qualsiasi altro numero. Questa rappresentazione grafica quantifica la nostra incertezza a posteriori relativamente (in questo esempio) all'orientamento della retta di regressione.

Esercizio 4.1. Illustreremo ora il problema di trovare la distribuzione $p(\tilde{y} \mid y)$ in un caso semplice, ovvero quello dello schema Beta-Binomiale. Nell'esempio, useremo un'altra volta i dati del campione di pazienti clinici depressi di Zetsche et al. (2019) – si veda l'Appendice **??**. Supponendo di volere esaminare in futuro altri 30 pazienti clinici, ci chiediamo: quanti di essi manifesteranno una depressione grave?

Se vogliamo fare predizioni su $\tilde{y}$ (il numero di "successi" previsti futuri) dobbiamo innanzitutto riconoscere che i possibili valori $\tilde{y} \in \{0, 1, \dots, 30\}$ non sono tutti egualmente plausibili. Sappiamo che $\tilde{y}$ è una v.c. binomiale con distribuzione

$$p(\tilde{y} \mid \theta) = \binom{30}{\tilde{y}} \theta^{\tilde{y}} (1-\theta)^{30-\tilde{y}} \, . \tag{4.2}$$

La v.c. $\tilde{y}$ dipende da θ, ma il parametro θ è esso stesso una variabile casuale. Avendo osservato $y = 23$ successi in $n = 30$ prove nel campione (laddove la presenza di una depressione grave è stata considerata un "successo"), e avendo assunto come distribuzione a priori per θ una Beta$(2, 10)$ (per continuare con l'esempio precedente), la distribuzione a posteriori di θ sarà una Beta$(25, 17)$:

```
bayesrules::summarize_beta_binomial(
  alpha = 2, beta = 10, y = 23, n = 30
)
#>       model alpha beta   mean mode      var      sd
#> 1     prior     2   10 0.1667  0.1 0.010684 0.10336
#> 2 posterior    25   17 0.5952  0.6 0.005603 0.07485
```

Per trovare la distribuzione sui possibili dati previsti futuri $\tilde{y}$ dobbiamo applicare la (4.1):

$$p(\tilde{y} \mid y = 23) = \int_0^1 p(\tilde{y} \mid \theta) p(\theta \mid y = 23) d\theta \, . \tag{4.3}$$

Per il modello Beta-Binomiale è possibile trovare una soluzione analitica alla (4.1).

Poniamo di avere osservato y successi in n prove e di utilizzare una distribuzione a priori Beta(a, b). Possiamo scrivere

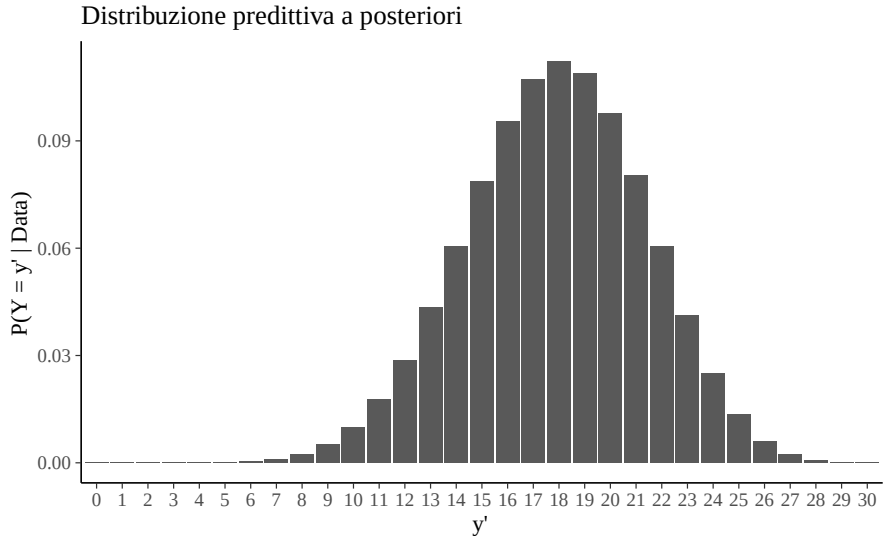
$$\begin{aligned}
p(\tilde{y} \mid y) &= \int_0^1 p(\tilde{y} \mid \theta) p(\theta \mid y) \, d\theta \\
&= \int_0^1 \binom{\tilde{n}}{\tilde{y}} \theta^{\tilde{y}} (1-\theta)^{\tilde{n}-\tilde{y}} \text{Beta}(a+y, b+n-y) \, d\theta \\
&= \binom{\tilde{n}}{\tilde{y}} \int_0^1 \theta^{\tilde{y}} (1-\theta)^{\tilde{n}-\tilde{y}} \frac{1}{B(a+y, b+n-y)} \theta^{a+y-1} (1-\theta)^{b+n-y-1} \\
&= \binom{\tilde{n}}{\tilde{y}} \frac{1}{B(a+y, b+n-y)} \int_0^1 \theta^{\tilde{y}+a+y-1} (1-\theta)^{\tilde{n}-\tilde{y}+b+n-y-1} \\
&= \binom{\tilde{n}}{\tilde{y}} \frac{B(\tilde{y}+a+y, b+n-y+\tilde{n}-\tilde{y})}{B(a+y, b+n-y)} \, .
\end{aligned} \tag{4.4}$$

Svolgendo i calcoli in R, per i dati dell'esempio otteniamo:

```
beta_binom <- function(rp) {
  val <- choose(np, rp) *
    beta(rp + a + y, b + n - y + np - rp) /
    beta(a + y, b + n - y)
  val
}

n <- 30
y <- 23
a <- 2
b <- 10
np <- 30
data.frame(
  heads = 0:np,
  pmf = beta_binom(0:np)
```

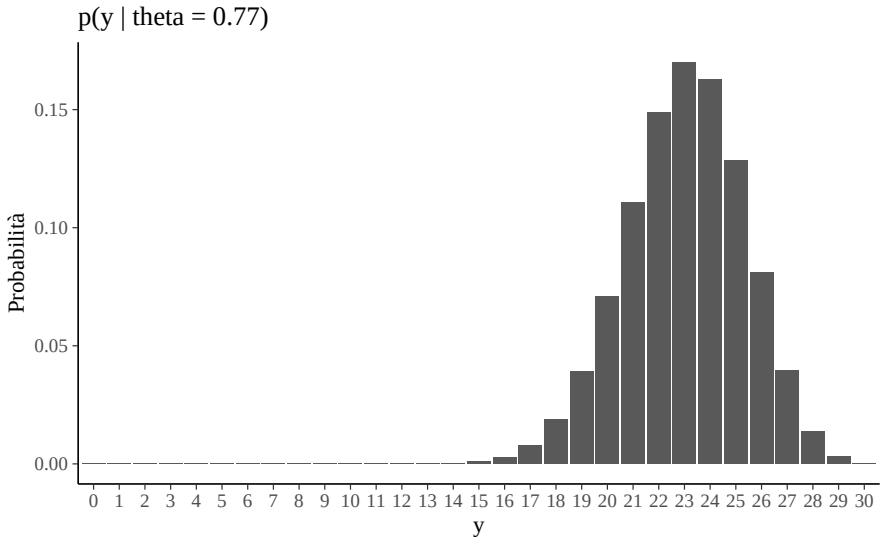
```
) %>%
  ggplot(aes(x = factor(heads), y = pmf)) +
  geom_col() +
  labs(
    title = "Distribuzione predittiva a posteriori",
    x = "y'",
    y = "P(Y = y' | Data)"
  )
```



È facile vedere come, in questo esempio, la distribuzione predittiva a posteriori $p(\tilde{y} \mid y)$ sia diversa dalla binomiale di parametro $\theta = 23/30$:

```
tibble(
  heads = 0:np,
  pmf = dbinom(x = 0:np, size = np, prob = 23 / 30)
) %>%
  ggplot(aes(x = factor(heads), y = pmf)) +
  geom_col() +
  labs(
    title = "p(y | theta = 0.77)",
    x = "y",
    y = "Probabilità"
```

```
)
```



In particolare, la $p(\tilde{y} \mid y)$ ha una varianza maggiore di $\text{Bin}(y \mid \theta = 0.77, n = 30)$. Questa maggiore varianza riflette le due fonti di incertezza che sono presenti nella (4.1): l'incertezza sul valore del parametro (descritta dalla distribuzione a posteriori) e l'incertezza dovuta alla variabilità campionaria (descritta dalla funzione di verosimiglianza). Possiamo concludere la discussione di questo esempio dicendo che, nel caso di 30 nuovi pazienti clinici, alla luce delle nostre credenze precedenti e dei dati osservati nel campione, ci aspettiamo di osservare 18 pazienti con una depressione severa, anche se è ragionevole aspettarci un numero compreso, diciamo, tra 10 e 25.

Una volta trovata la distribuzione predittiva a posteriori $p(\tilde{y} \mid y)$ diventa possibile rispondere a domande come: qual è la probabilità di depressione grave in almeno 10 dei 30 pazienti futuri? Rispondere a domande di questo tipo è possibile, ma richiede un po' di lavoro. Tuttavia, non è importante imparare scrivere il codice necessario a risolvere problemi di questo tipo perché, in generale, anche per problemi solo leggermente più complessi di quello discusso qui, non sono disponibili espressioni analitiche della distribuzione predittiva a posteriori. Invece, è possibile trovare una approssimazione numerica della $p(\tilde{y} \mid y)$ mediante simulazioni MCMC. Inoltre, se viene utilizzato un tale metodo, risulta facile

rispondere a domande simili a quella che abbiamo presentato sopra.

4.2 Metodi MCMC per la distribuzione predittiva a posteriori

Se svolgiamo l'analisi bayesiana con il metodo MCMC, le repliche $p(y^{rep} \mid y)$ (ovvero le stime delle possibili osservazioni future $p(\tilde{y} \mid y)$) possono essere ottenute nel modo seguente:

- campionare $\theta_i \sim p(\theta \mid y)$, ovvero campionare un valore del parametro dalla distribuzione a posteriori;
- campionare $y^{rep} \sim p(y^{rep} \mid \theta_i)$, ovvero campionare il valore di un'osservazione dalla funzione di verosimiglianza condizionata al valore del parametro definito nel passo precedente.

Se i due passaggi descritti sopra vengono ripetuti un numero sufficiente di volte, l'istogramma risultante approssimerà la distribuzione predittiva a posteriori che, in teoria (ma non in pratica) potrebbe essere ottenuta per via analitica (si veda il Paragrafo **??**).

Esercizio 4.2. Generiamo ora $p(y^{rep} \mid y)$ nel caso dell'inferenza su una proporzione.

Riportiamo qui sotto il codice Stan — si veda il Capitolo **??**.

```
modelString = "
data {
  int<lower=0> N;
  int<lower=0, upper=1> y[N];
}
parameters {
  real<lower=0, upper=1> theta;
}
model {
  theta ~ beta(2, 10);
  y ~ bernoulli(theta);
}
generated quantities {
```

```
  int y_rep[N];
  real log_lik[N];
  for (n in 1:N) {
    y_rep[n] = bernoulli_rng(theta);
    log_lik[n] = bernoulli_lpmf(y[n] | theta);
  }
}
"
writeLines(modelString, con = "code/betabin23-30-2-10.stan")
```

Si noti che nel nel blocco `generated quantities` sono state aggiunte le istruzioni necessarie per simulare y^{rep}, ovvero, `y_rep[n] = bernoulli_rng(theta)`. I dati dell'esempio sono:

```
data_list <- list(
  N = 30,
  y = c(rep(1, 23), rep(0, 7))
)
```

Compiliamo il codice Stan

```
file <- file.path("code", "betabin23-30-2-10.stan")
mod <- cmdstan_model(file)
```

ed eseguiamo il campionamento MCMC:

```
fit <- mod$sample(
  data = data_list,
  iter_sampling = 4000L,
  iter_warmup = 2000L,
  seed = SEED,
  cores = 4L,
  chains = 4L,
  parallel_chains = 4L,
  refresh = 0,
  thin = 1
)
```

Per comodità, trasformiamo l'oggetto `fit` in un oggetto di classe `stanfit`:

```
stanfit <- rstan::read_stan_csv(fit$output_files())
```

Il contenuto dell'oggetto `stanfit` può essere esaminato nel modo seguente:

```
list_of_draws <- extract(stanfit)
print(names(list_of_draws))
#> [1] "theta"   "y_rep"   "log_lik" "lp__"
```

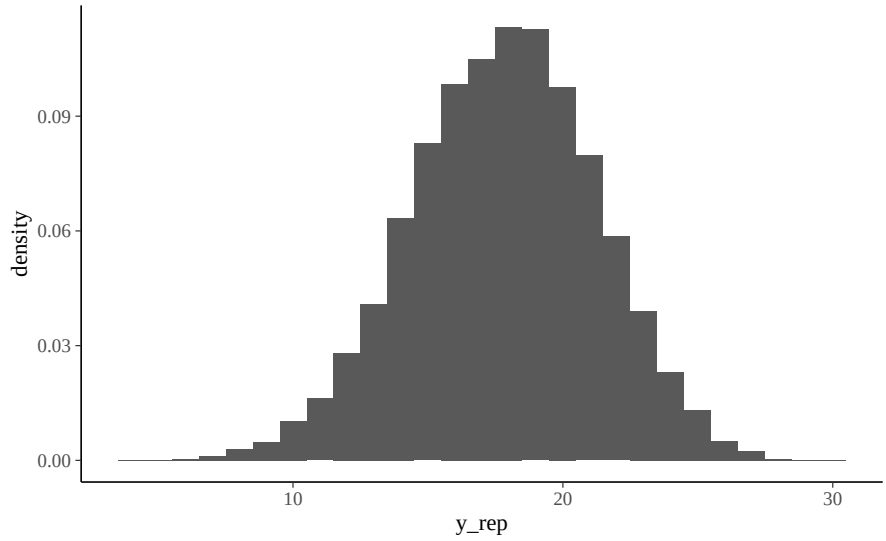
Dall'oggetto `list_of_draws` recuperiamo `y_rep`:

```
y_bern <- list_of_draws$y_rep
dim(y_bern)
#> [1] 16000    30
head(y_bern)
#>
#> iterations [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
#>       [1,]    1    1    0    0    1    0    1    0
#>       [2,]    0    1    0    1    0    1    1    0
#>       [3,]    1    1    0    0    1    1    1    1
#>       [4,]    1    0    1    1    0    1    0    1
#>       [5,]    1    1    0    1    1    1    0    1
#>       [6,]    1    0    1    1    1    0    0    1
#>
#> iterations [,9] [,10] [,11] [,12] [,13] [,14] [,15]
#>       [1,]    0     0     1     0     1     1     1
#>       [2,]    1     0     1     1     1     1     1
#>       [3,]    1     1     1     0     0     1     1
#>       [4,]    1     1     1     1     1     1     0
#>       [5,]    1     0     1     0     1     1     1
#>       [6,]    1     1     1     1     1     1     1
#>
#> iterations [,16] [,17] [,18] [,19] [,20] [,21] [,22]
#>       [1,]     1     1     0     0     0     0     1
#>       [2,]     1     1     0     0     0     1     1
#>       [3,]     0     0     1     1     1     1     1
```

```
#>       [4,]    1    0    1    0    0    0    0
#>       [5,]    0    1    1    1    1    0    1
#>       [6,]    1    0    1    1    0    0    0
#>
#> iterations [,23] [,24] [,25] [,26] [,27] [,28] [,29]
#>       [1,]    0    1    1    0    1    0    1
#>       [2,]    1    1    0    1    0    0    1
#>       [3,]    1    1    1    0    1    0    1
#>       [4,]    1    0    1    1    0    1    0
#>       [5,]    0    1    1    1    1    1    1
#>       [6,]    1    0    1    1    1    0    1
#>
#> iterations [,30]
#>       [1,]    1
#>       [2,]    1
#>       [3,]    1
#>       [4,]    1
#>       [5,]    0
#>       [6,]    1
```

Dato che il codice Stan definisce un modello per i dati grezzi (ovvero, per ciascuna singola prova Bernoulliana del campione), ogni riga di `y_bern` include 30 colonne, ciascuna delle quali corrisponde ad un campione (n = 16000 in questa simulazione) di possibili valori futuri $y_i \in \{0, 1\}$. Per ottenere una stima della distribuzione predittiva a posteriori `p(y_rep)`, ovvero, una stima della probabilità associata a ciascuno dei possibili numeri di "successi" in $N = 30$ nuove prove future, è sufficiente calcolare la proporzione di valori 1 in ciascuna riga:

```
tibble(y_rep = rowSums(y_bern)) %>%
  ggplot(aes(x = y_rep, after_stat(density))) +
  geom_histogram(binwidth = 1)
```



4.3 Posterior predictive checks

La distribuzione predittiva a posteriori viene utilizzata per eseguire i cosiddetti *controlli predittivi a posteriori* (*Posterior Predictive Checks*, PPC). Ricordiamo che la distribuzione predittiva a posteriori corrisponde alla simulazione di un campione di dati generati utilizzando le proprietà del modello adattato. Nei PPC si realizza un confronto grafico tra $p(y^{rep} \mid y)$ e i dati osservati y. Confrontando visivamente gli aspetti chiave dei dati previsti futuri y^{rep} e dei dati osservati y possiamo determinare se il modello è adeguato.

Oltre al confronto tra le distribuzioni $p(y)$ e $p(y^{rep})$ è anche possibile un confronto tra la distribuzione di varie statistiche descrittive, i cui valori sono calcolati su diversi campioni y^{rep}, e le corrispondenti statistiche descrittive calcolate sui dati osservati. Vengono solitamente considerate statistiche descrittive quali la media, la varianza, la deviazione standard, il minimo o il massimo. Ma confronti di questo tipo sono possibili per qualunque statistica descrittiva. Questi confronti sono chiamati PPC.

Esercizio 4.3. Esaminiamo ora un set di dati che non seguono la distribuzione normale (Gelman et al., 2020). I dati corrispondono ad una

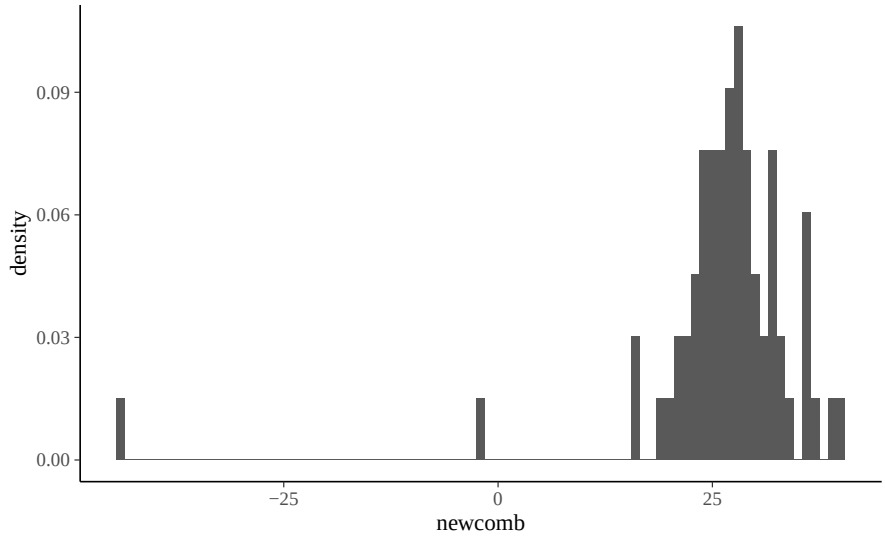
serie di misurazioni prese da Simon Newcomb nel 1882 come parte di un esperimento per stimare la velocità della luce. A questi dati verrà (inappropriatamente) adattata una distribuzione normale. L'obiettivo dell'esempio è quello di mostrare come i PPC possono rivelare la mancanza di adattamento di un modello ai dati.

I PPC mostrano che il modo più semplice per verificare l'adattamento del modello è quello di visualizzare y^{rep} insieme ai dati effettivi. Iniziamo a caricare i dati:

```
library("MASS")
data("newcomb")
```

Visualizziamo la distribuzione dei dati con un istogramma:

```
tibble(newcomb) %>%
  ggplot(aes(x = newcomb, after_stat(density))) +
  geom_histogram(binwidth = 1)
```



Creiamo un oggetto di tipo `list` dove inserire i dati:

```
data_list <- list(
  y = newcomb,
```

```
  N = length(newcomb)
)
```

Il codice Stan per il modello normale è il seguente:

```
modelString <- "
data {
  int<lower=0> N;
  vector[N] y;
}
parameters {
  real mu;
  real<lower=0> sigma;
}
model {
  mu ~ normal(25, 10);
  sigma ~ cauchy(0, 10);
  y ~ normal(mu, sigma);
}
generated quantities {
  vector[N] y_rep;
  for (n in 1:N) {
    y_rep[n] = normal_rng(mu, sigma);
  }
}
"
writeLines(modelString, con = "code/newcomb.stan")
```

Adattando il modello ai dati

```
file <- file.path("code", "newcomb.stan")
mod <- cmdstan_model(file)
fit <- mod$sample(
  data = data_list,
  iter_sampling = 4000L,
  iter_warmup = 2000L,
  seed = SEED,
  chains = 4L,
```

```
  cores = 4L,
  refresh = 0,
  thin = 1
)
```

otteniamo le seguenti stime dei parametri μ e σ:

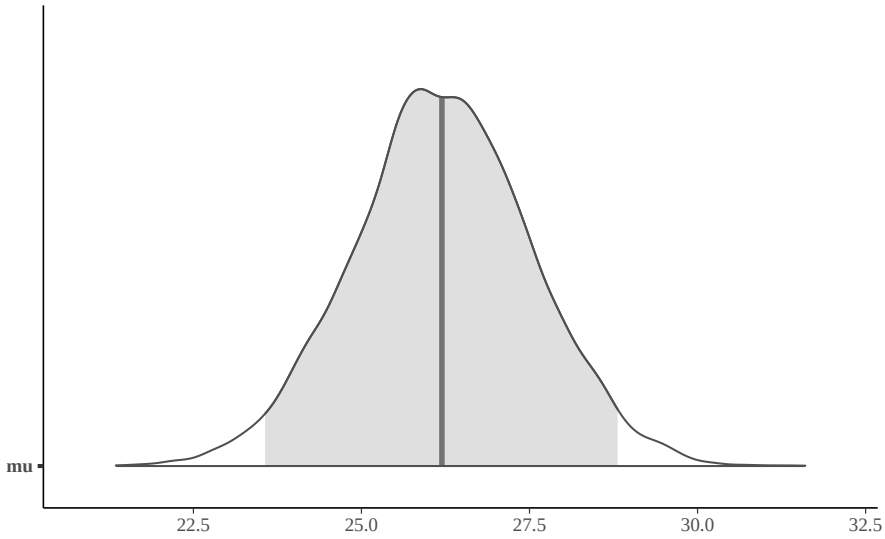
```
fit$summary(c("mu", "sigma"))
#> # A tibble: 2 x 10
#>   variable  mean median    sd   mad    q5   q95  rhat
#>   <chr>    <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
#> 1 mu        26.2   26.2 1.33  1.30  24.0   28.4  1.00
#> 2 sigma     10.9   10.8 0.958 0.943  9.40  12.5  1.00
#> # ... with 2 more variables: ess_bulk <dbl>,
#> #   ess_tail <dbl>
```

Trasformiamo `fit` in un oggetto `stanfit`:

```
stanfit <- rstan::read_stan_csv(fit$output_files())
```

La distribuzione a posteriori di μ è

```
mu_draws <- as.matrix(stanfit, pars = "mu")
mcmc_areas(mu_draws, prob = 0.95) # color 95% interval
```



Confrontiamo μ con la media di y:

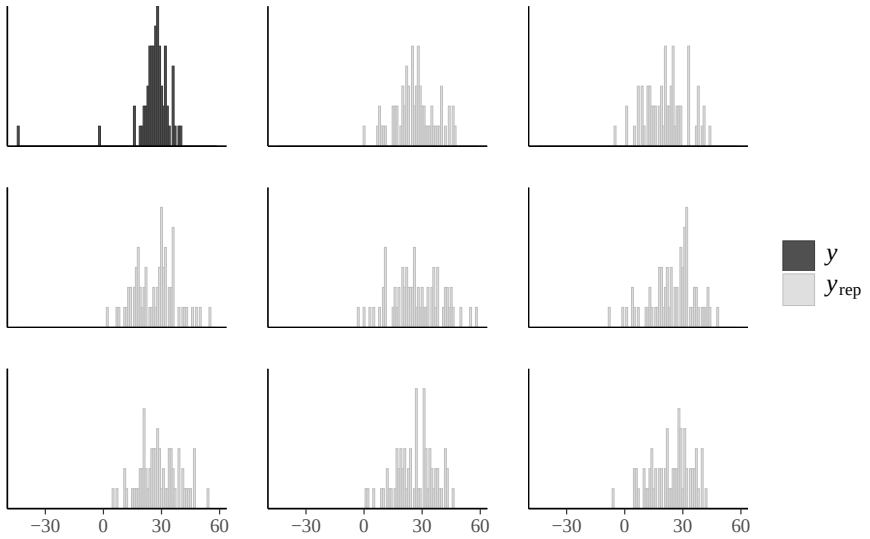
```
mean(newcomb)
#> [1] 26.21
```

Anche se trova la media giusta, il modello non è comunque adeguato a prevedere le altre proprietà della y. Estraiamo y^{rep} dall'oggetto `stanfit`:

```
y_rep <- as.matrix(stanfit, pars = "y_rep")
dim(y_rep)
#> [1] 16000    66
```

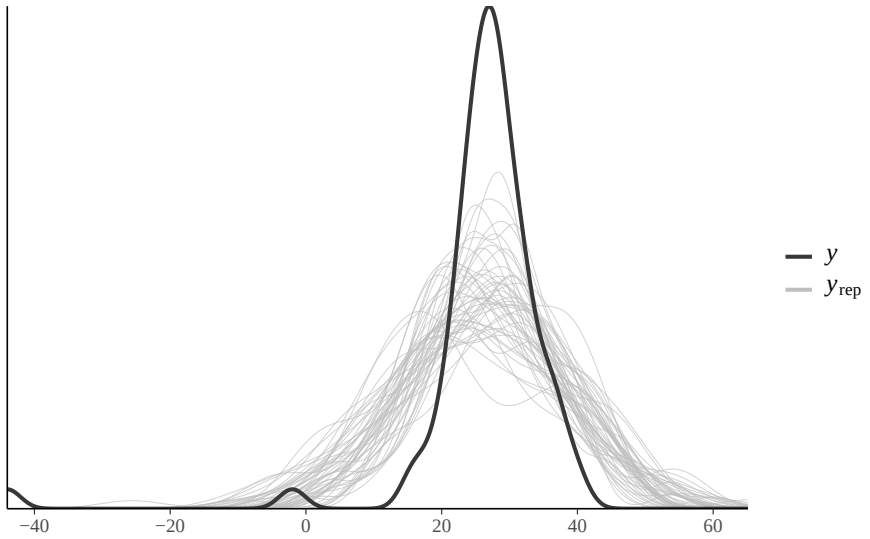
I valori `y_rep` sono i dati della distribuzione predittiva a posteriori che sono stati simulati usando gli stessi valori X dei predittori utilizzati per adattare il modello. Il confronto tra l'istogramma della y e gli istogrammi di diversi campioni y^{rep} mostra una scarsa corrispondenza tra i due:

```
ppc_hist(data_list$y, y_rep[1:8, ], binwidth = 1)
```



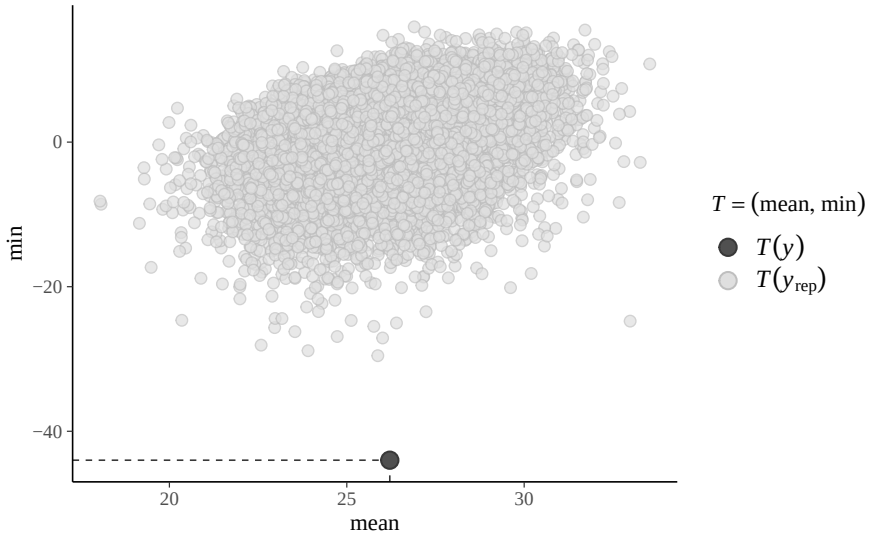
Alla stessa conclusione si giunge tramite un confronto tra la funzione di densità empirica della y e quella di diversi campioni y^{rep}:

```
ppc_dens_overlay(data_list$y, y_rep[1:50, ])
```



Generiamo ora i PPC per la media e il minimo della distribuzione:

```
ppc_stat_2d(data_list$y, y_rep, stat = c("mean", "min"))
```



Mentre la media viene riprodotta accuratamente dal modello (come abbiamo visto sopra), ciò non è vero per il minimo dela distribuzione. L'origine di questa mancanza di adattamento è il fatto che la distribuzione delle misurazioni della velocità della luce è asimmetrica negativa. Dato che ci sono poche osservazioni nella coda negativa della distribuzione, solo per fare un esempio, utilizzeremo ora un secondo modello che ipotizza una distribuzione t di Student:

```
modelString = "
data {
  int<lower=0> N;
  vector[N] y;
}
parameters {
  real mu;
  real<lower=0> sigma;
  real<lower=0> nu;
}
model {
```

```
  mu ~ normal(25, 10);
  sigma ~ cauchy(0, 10);
  nu ~ cauchy(0, 10);
  y ~ student_t(nu, mu, sigma);
}
generated quantities {
  vector[N] y_rep;
  for (n in 1:N) {
    y_rep[n] = student_t_rng(nu, mu, sigma);
  }
}
"
writeLines(modelString, con = "code/newcomb2.stan")
```

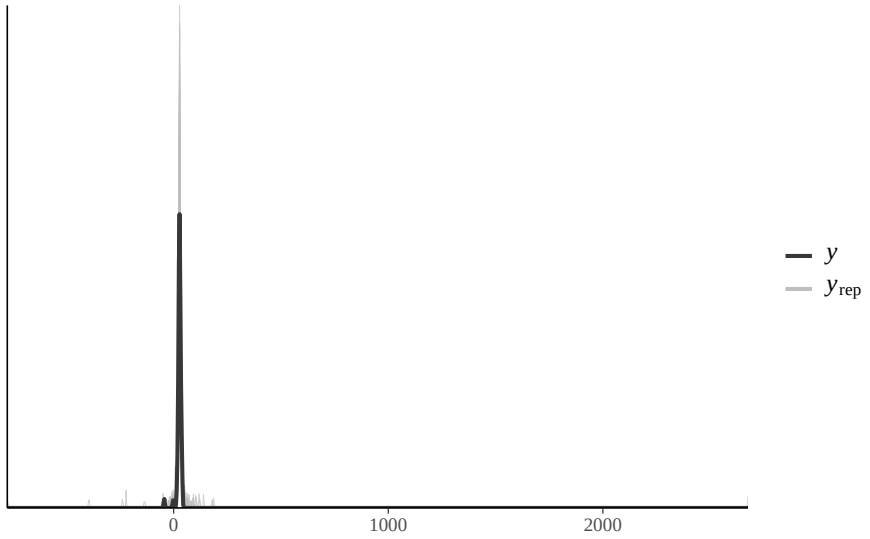
Adattiamo questo secondo modello ai dati.

```
file <- file.path("code", "newcomb2.stan")
mod <- cmdstan_model(file)
fit <- mod$sample(
  data = data_list,
  iter_sampling = 4000L,
  iter_warmup = 2000L,
  seed = SEED,
  chains = 4L,
  cores = 4L,
  parallel_chains = 2L,
  refresh = 0,
  thin = 1
)
#> Running MCMC with 4 parallel chains...
#>
#> Chain 1 finished in 0.3 seconds.
#> Chain 2 finished in 0.3 seconds.
#> Chain 3 finished in 0.3 seconds.
#> Chain 4 finished in 0.3 seconds.
#>
#> All 4 chains finished successfully.
#> Mean chain execution time: 0.3 seconds.
```

```
#> Total execution time: 0.5 seconds.
```

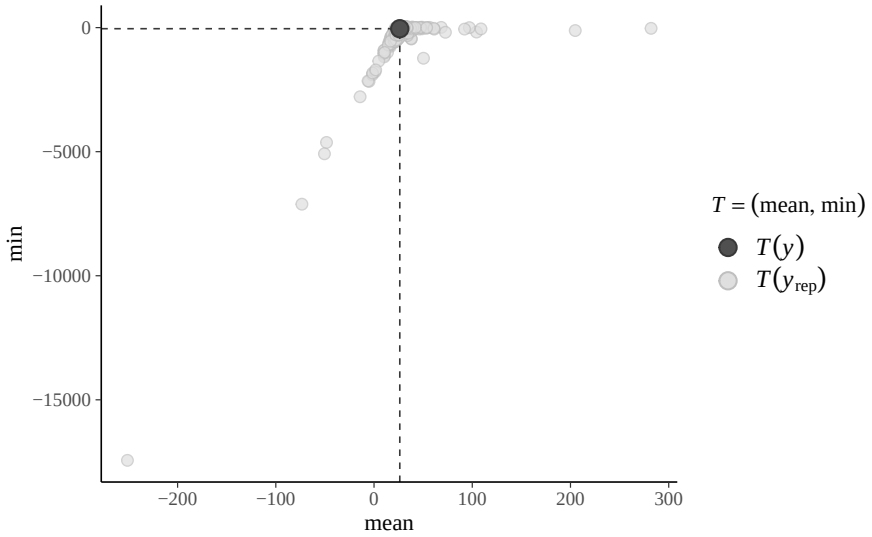
Per questo secondo modello il confronto tra la funzione di densità empirica della y e quella di diversi campioni y^{rep} risulta adeguato:

```
stanfit <- rstan::read_stan_csv(fit$output_files())
y_rep <- as.matrix(stanfit, pars = "y_rep")
ppc_dens_overlay(data_list$y, y_rep[1:50, ])
```



Inoltre, anche la statistica "minimo della distribuzione" viene ben predetta dal modello.

```
ppc_stat_2d(data_list$y, y_rep, stat = c("mean", "min"))
```



In conclusione, per le misurazioni della velocità della luce di Newcomb l'accuratezza predittiva del modello basato sulla distribuzione t di Student è chiaramente migliore di quella del modello normale.

Considerazioni conclusive

Questo capitolo presenta i controlli predittivi a posteriori. A questo proposito è necessario notare un punto importante: i controlli predittivi a posteriori, quando suggeriscono un buon adattamento del modello alle caratterische dei dati previsti futuri y^{rep}, non forniscono necessariamente una forte evidenza della capacità del modello di generalizzarsi a nuovi campioni di dati. Una tale evidenza sulla generalizzabilità del modello può solo essere fornita da studi di *holdout validation*, ovvero da studi nei quali viene utilizzato un *nuovo* campione di dati. Se i PPC mostrano un cattivo adattamento del modello ai dati previsti futuri, però, questo controllo fornisce una forte evidenza di una errata specificazione del modello.

5

Modello Normale-Normale

5.1 Distribuzione Normale-Normale con varianza nota

Per σ^2 nota, la v.c. gaussiana è distribuzione a priori coniugata della v.c. gaussiana. Siano $Y_1, \dots, Y_n$ n variabili casuali i.i.d. che seguono la distribuzione gaussiana:

$$Y_1, \dots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma).$$

Si vuole stimare μ sulla base di n osservazioni $y_1, \dots, y_n$. Considereremo qui solamente il caso in cui σ^2 sia supposta perfettamente nota.

Ricordiamo che la densità di una gaussiana è

$$p(y_i \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y_i - \mu)^2}{2\sigma^2}\right\}.$$

Essendo le variabili i.i.d., possiamo scrivere la densità congiunta come il prodotto delle singole densità e quindi si ottiene

$$p(y \mid \mu) = \prod_{i=1}^{n} p(y_i \mid \mu).$$

Una volta osservati i dati y, la verosimiglianza diventa

$$\begin{aligned}\mathcal{L}(\mu \mid y) = \prod_{i=1}^{n} p(y_i \mid \mu) = \\ &\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y_1-\mu)^2}{2\sigma^2}\right\} \times \\ &\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y_2-\mu)^2}{2\sigma^2}\right\} \times \\ &\vdots \\ &\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y_n-\mu)^2}{2\sigma^2}\right\}. \end{aligned} \tag{5.1}$$

Se viene scelta una densità a priori gaussiana, ciò fa sì che anche la densità a posteriori sia gaussiana. Supponiamo che

$$p(\mu) = \frac{1}{\tau_0\sqrt{2\pi}} \exp\left\{-\frac{(\mu-\mu_0)^2}{2\tau_0^2}\right\}, \tag{5.2}$$

ovvero che la distribuzione a priori di μ sia gaussiana con media μ_0 e varianza τ_0^2. Possiamo dire che μ_0 rappresenta il valore ritenuto più probabile per μ e τ_0^2 il grado di incertezza che abbiamo rispetto a tale valore.

Svolgendo una serie di passaggi algebrici, si arriva a

$$p(\mu \mid y) = \frac{1}{\tau_p\sqrt{2\pi}} \exp\left\{-\frac{(\mu-\mu_p)^2}{2\tau_p^2}\right\}, \tag{5.3}$$

dove

$$\mu_p = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \tag{5.4}$$

e

$$\tau_p^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}. \tag{5.5}$$

In altri termini, se la distribuzione a priori per μ è gaussiana, la distribuzione a posteriori è anch'essa gaussiana con valore atteso (a posteriori) μ_p e varianza (a posteriori) τ_p^2 date dalle espressioni precedenti.

In conclusione, il risultato trovato indica che:

- il valore atteso a posteriori è una media pesata fra il valore atteso a priori μ_0 e la media campionaria $\bar{y}$; il peso della media campionaria è tanto maggiore tanto più è grande n (il numero di osservazioni) e τ_0^2 (l'incertezza iniziale);
- l'incertezza (varianza) a posteriori τ_p^2 è sempre più piccola dell'incertezza a priori τ_0^2 e diminuisce al crescere di n.

5.2 Il modello Normale con Stan

Per esaminare un esempio pratico, consideriamo i 30 valori BDI-II dei soggetti clinici di Zetsche et al. (2019):

```
df <- data.frame(
  y = c(
    26.0, 35.0, 30, 25, 44, 30, 33, 43, 22, 43,
    24, 19, 39, 31, 25, 28, 35, 30, 26, 31, 41,
    36, 26, 35, 33, 28, 27, 34, 27, 22
  )
)
```

Calcoliamo le statistiche descrittive del campione di dati:

```
df %>%
  summarise(
    sample_mean = mean(y),
    sample_sd = sd(y)
  )
#>   sample_mean sample_sd
#> 1       30.93     6.607
```

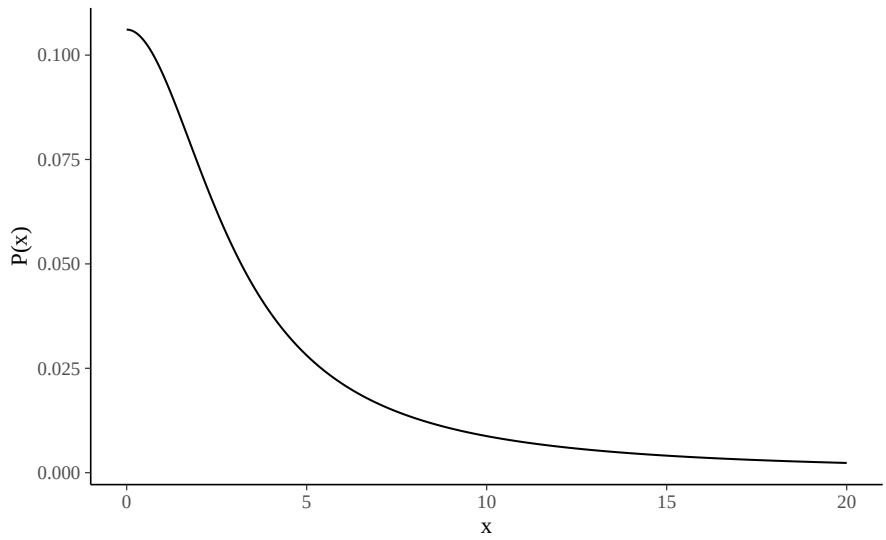
Nella discussione seguente assumeremo che μ e σ siano indipendenti. Assegneremo a μ una distribuzione a priori $\mathcal{N}(25, 2)$ e a σ una distribuzione a priori $Cauchy(0, 3)$.

Il modello statistico diventa:

$$
\begin{aligned}
Y_i &\sim \mathcal{N}(\mu, \sigma) \\
\mu &\sim \mathcal{N}(\mu_\mu = 25, \sigma_\mu = 2) \\
\sigma &\sim \text{Cauchy}(0, 3)
\end{aligned}
$$

In base al modello definito, la variabile casuale Y segue la distribuzione Normale di parametri μ e σ. Il parametro μ è sconosciuto e abbiamo deciso di descrivere la nostra incertezza relativa ad esso mediante una distribuzione a priori Normale con media uguale a 25 e deviazione standard pari a 2. L'incertezza relativa a σ è quantificata da una distribuzione a priori half-Cauchy(0, 5), come indicato nella figura seguente:

```
data.frame(x = c(0, 20)) %>%
  ggplot(aes(x)) +
  stat_function(
    fun = dcauchy,
    n = 1e3,
    args = list(location = 0, scale = 3)
  ) +
  ylab("P(x)") +
  theme(legend.position = "none")
```

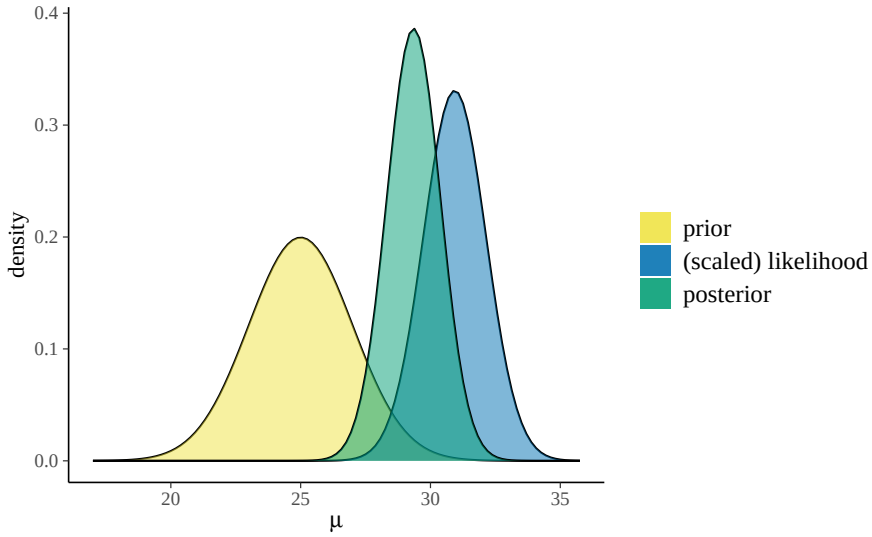


Dato che il modello è Normale-Normale, è possibile una soluzione analitica, come descritto in precedenza per il caso in cui σ è noto. In tali condizioni, la distribuzione a posteriori per μ può essere trovata con la funzione `bayesrules:::summarize_normal_normal()`:

```
bayesrules:::summarize_normal_normal(
  mean = 25, sd = 2, sigma = sd(df$y),
  y_bar = mean(df$y), n = 30
)
#>       model  mean  mode   var    sd
#> 1     prior 25.00 25.00 4.000 2.000
#> 2 posterior 29.35 29.35 1.067 1.033
```

La rappresentazione grafica della funzione a priori, della verosimiglianza e della distribuzione a posteriori per μ è fornita da:

```
bayesrules:::plot_normal_normal(
  mean = 25, sd = 2, sigma = sd(df$y),
  y_bar = mean(df$y), n = 30
)
```



La procedura MCMC utilizzata da Stan è basata su un campionamento Monte Carlo Hamiltoniano che non richiede l'uso di distribuzioni a priori

coniugate. Pertanto per i parametri è possibile scegliere una qualunque distribuzione a priori arbitraria.

Per continuare con l'esempio, poniamoci il problema di trovare le distribuzioni a posteriori dei parametri μ e σ usando le funzioni del pacchetto `cmdstanr`. Il modello statistico descritto sopra si può scrivere in Stan nel modo seguente:

```
modelString = "
data {
  int<lower=0> N;
  vector[N] y;
}
parameters {
  real mu;
  real<lower=0> sigma;
}
model {
  mu ~ normal(25, 2);
  sigma ~ cauchy(0, 3);
  y ~ normal(mu, sigma);
}
"
writeLines(modelString, con = "code/normalmodel.stan")
```

Si noti che, nel modello, il parametro σ è considerato incognito.

Sistemiamo i dati nel formato appropriato per potere essere letti da Stan:

```
data_list <- list(
  N = length(df$y),
  y = df$y
)
```

Leggiamo il file in cui abbiamo salvato il codice Stan

```
file <- file.path("code", "normalmodel.stan")
```

compiliamo il modello

```
mod <- cmdstan_model(file)
```

ed eseguiamo il campionamento MCMC:

```
fit <- mod$sample(
  data = data_list,
  iter_sampling = 4000L,
  iter_warmup = 2000L,
  seed = SEED,
  chains = 4L,
  parallel_chains = 2L,
  refresh = 0,
  thin = 1
)
```

Le stime a posteriori dei parametri si ottengono con:

```
fit$summary(c("mu", "sigma"))
#> # A tibble: 2 x 10
#>   variable  mean median    sd   mad    q5   q95  rhat
#>   <chr>    <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
#> 1 mu       29.3   29.3  1.11  1.10  27.4  31.0   1.00
#> 2 sigma     6.89   6.78 0.959 0.910  5.52  8.63  1.00
#> # ... with 2 more variables: ess_bulk <dbl>,
#> #   ess_tail <dbl>
```

oppure, dopo avere trasformato l'oggetto `fit` nel formato `stanfit`,

```
stanfit <- rstan::read_stan_csv(fit$output_files())
```

con

```
out <- rstantools::posterior_interval(as.matrix(stanfit), prob = 0.95)
out
#>          2.5%   97.5%
#> mu     26.990  31.349
```

```
#> sigma   5.321   9.079
#> lp__  -77.914 -74.276
```

Possiamo dunque concludere, con un grado di certezza soggettiva del 95%, che siamo sicuri che la media della popolazione da cui abbiamo tratto i dati è compresa nell'intervallo [26.99, 31.35].

5.3 Il modello normale con `quap()`

Ripetiamo l'analisi precedente usando le funzioni del pacchetto `rethinking` per trovare le distribuzioni a posteriori dei parametri μ e σ. Definiamo il modello statistico mediante la funzione `alist()`:

```
flist <- alist(
  y ~ dnorm(mu, sigma),
  mu ~ dnorm(25, 2),
  sigma ~ dcauchy(0, 3)
)
```

Le precedenti istruzioni `R` specificano una variabile casuale Y che si distribuisce come una Normale di parametri μ e σ; questa è la verosimiglianza. La distribuzione a priori del parametro μ è una Normale di media 25 e deviazione standard 2. La distribuzione a priori del parametro σ è una half-Cauchy di parametri `location` $= 0$ e `scale` $= 3$.

Usiamo la funzione `quap()` per ottenere l'approssimazione quadratica delle distribuzioni a posteriori di μ e σ:

```
set.seed(123)
m <- quap(
  flist,
  data = df
)
```

L'intervallo di credibilità al 95% è dato dalla funzione `precis()`:

```
out <- precis(m, prob = 0.95)
out
#>         mean     sd  2.5%  97.5%
#> mu    29.388 1.0633 27.30 31.472
#> sigma  6.501 0.8473  4.84  8.162
```

I risultati sono simili a quelli trovati in precedenza.

5.4 Il modello normale con `brms::brm()`

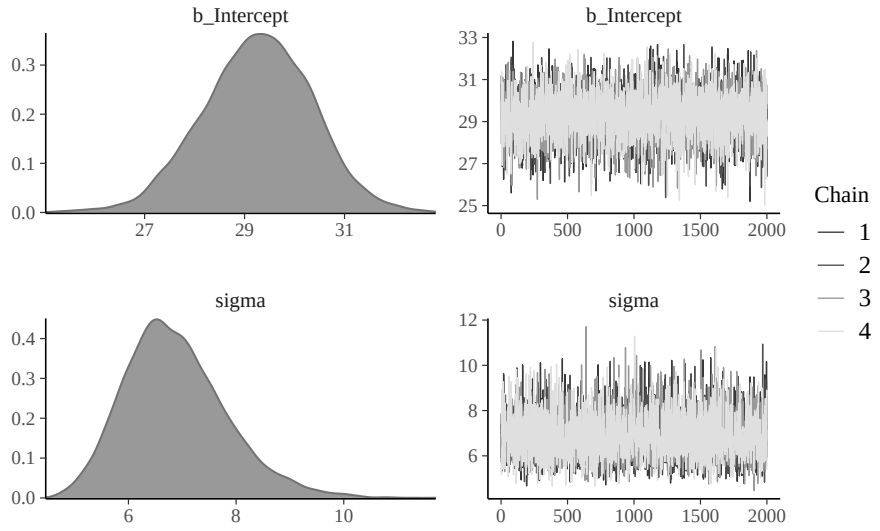
Stimiamo ora la distribuzione a posteriori di μ usando la funzione `brms::brm()`. In questo caso non è necessario scrivere il modello in forma esplicita, come abbiamo fatto usando linguaggio Stan. La sintassi specifiata di seguito viene trasformata in maniera automatica nel linguaggio Stan prima di adattare il modello ai dati:

```
fit3 <- brm(
  data = df,
  family = gaussian(),
  y ~ 1,
  prior = c(
    prior(normal(25, 2), class = Intercept),
    prior(cauchy(0, 3), class = sigma)
  ),
  iter = 4000,
  refresh = 0,
  chains = 4,
  backend = "cmdstanr"
)
#> Running MCMC with 4 chains, at most 8 in parallel...
#>
#> Chain 1 finished in 0.1 seconds.
#> Chain 2 finished in 0.1 seconds.
#> Chain 3 finished in 0.1 seconds.
#> Chain 4 finished in 0.1 seconds.
#>
```

```
#> All 4 chains finished successfully.
#> Mean chain execution time: 0.1 seconds.
#> Total execution time: 0.2 seconds.
```

I trace-plot si ottengono con l'istruzione seguente:

```
plot(fit3)
```



Le stime della distribuzione a posteriori si ottengono con la funzione `summary()`:

```
summary(fit3)
#>  Family: gaussian
#>   Links: mu = identity; sigma = identity
#> Formula: y ~ 1
#>    Data: df (Number of observations: 30)
#>   Draws: 4 chains, each with iter = 2000; warmup = 0; thin = 1;
#>          total post-warmup draws = 8000
#>
#> Population-Level Effects:
#>           Estimate Est.Error l-95% CI u-95% CI Rhat
#> Intercept    29.26      1.09    27.08    31.34 1.00
```

```
#>           Bulk_ESS Tail_ESS
#> Intercept     4559     4987
#>
#> Family Specific Parameters:
#>       Estimate Est.Error l-95% CI u-95% CI Rhat
#> sigma     6.88      0.94     5.27     9.01 1.00
#>       Bulk_ESS Tail_ESS
#> sigma     4366     4638
#>
#> Draws were sampled using sample(hmc). For each parameter, Bulk_ESS
#> and Tail_ESS are effective sample size measures, and Rhat is the potential
#> scale reduction factor on split chains (at convergence, Rhat = 1).
```

Nuovamente, i risultati sono molto simili a quelli ottenuti in precedenza.

Considerazioni conclusive

Questo esempio ci mostra come calcolare l'intervallo di credibilità per la media di una v.c. Normale. La domanda più ovvia di analisi dei dati, dopo avere visto come creare l'intervallo di credibilità per la media di un gruppo, riguarda il confronto tra le medie di due gruppi. Questo però è un caso speciale di una tecnica di analisi dei dati più generale, chiamate analisi di regressione lineare. Prima di discutere il problema del confronto tra le medie di due gruppi è dunque necessario esaminare il modello statistico di regressione lineare.

6

Introduzione al modello lineare

Lo scopo della ricerca è trovare le associazioni tra le variabili e fare confronti fra le condizioni sperimentali. Nel caso della psicologia, il ricercatore vuole scoprire le leggi generali che descrivono le relazioni tra i costrutti psicologici e le relazioni che intercorrono tra i fenomeni psicologici e quelli non psicologici (sociali, economici, storici, ...). Abbiamo già visto come la correlazione di Pearson sia uno strumento adatto a questo scopo. Infatti, essa ci informa sulla direzione e sull'intensità della relazione lineare tra due variabili. Tuttavia, la correlazione non è sufficiente, in quanto il ricercatore ha a disposizione solo i dati di un campione, mentre vorrebbe descrivere la relazione tra le variabili nella popolazione. A causa della variabilità campionaria, le proprietà dei campioni sono necessariamente diverse da quelle della popolazione: ciò che si può osservare nella popolazione potrebbe non emergere nel campione e, al contrario, il campione manifesta caratteristiche che non sono necessariamente presenti nella popolazione. È dunque necessario chiarire, dal punto di vista statistico, il legame che intercorre tra le proprietà del campione e le proprietà della popolazione da cui esso è stato estratto. Il modello lineare utilizza la funzione matematica più semplice per descrivere la relazione fra due variabili, ovvero la funzione lineare. In questo Capitolo vedremo come si possa fare inferenza sulla relazione tra due variabili mediante il modello lineare bayesiano. Inizieremo a descrivere le proprietà geometriche della funzione lineare per poi utilizzare questa semplice funzione per costruire un modello statistico secondo un approccio bayesiano.

6.1 La funzione lineare

Iniziamo con un ripasso sulla funzione di lineare. Si chiama *funzione lineare* una funzione del tipo

$$f(x) = a + bx, \tag{6.1}$$

dove a e b sono delle costanti. Il grafico di tale funzione è una retta di cui il parametro b è detto *coefficiente angolare* e il parametro a è detto *intercetta* con l'asse delle y [infatti, la retta interseca l'asse y nel punto $(0, a)$, se $b \neq 0$].

Per assegnare un'interpretazione geometrica alle costanti a e b si consideri la funzione

$$y = bx. \tag{6.2}$$

Tale funzione rappresenta un caso particolare, ovvero quello della *proporzionalità diretta* tra x e y. Il caso generale della linearità

$$y = a + bx \tag{6.3}$$

non fa altro che sommare una costante a a ciascuno dei valori $y = bx$. Nella funzione lineare $y = a + bx$, se b è positivo allora y aumenta al crescere di x; se b è negativo allora y diminuisce al crescere di x; se $b = 0$ la retta è orizzontale, ovvero y non muta al variare di x.

Consideriamo ora il coefficiente b. Si consideri un punto x_0 e un incremento arbitrario ε come indicato nella figura 6.1. Le differenze $\Delta x = (x_0 + \varepsilon) - x_0$ e $\Delta y = f(x_0 + \varepsilon) - f(x_0)$ sono detti *incrementi* di x e y. Il coefficiente angolare b è uguale al rapporto

$$b = \frac{\Delta y}{\Delta x} = \frac{f(x_0 + \varepsilon) - f(x_0)}{(x_0 + \varepsilon) - x_0}, \tag{6.4}$$

indipendentemente dalla grandezza degli incrementi Δx e Δy. Il modo più semplice per assegnare un'interpretazione geometrica al coefficiente angolare (o pendenza) della retta è dunque quello di porre $\Delta x = 1$. In tali circostanze infatti $b = \Delta y$.

6.2 L'errore di misurazione

Per descrivere l'associazione tra due variabili, tuttavia, la funzione lineare non è sufficiente. Nel mondo empirico, infatti, la relazione tra variabili

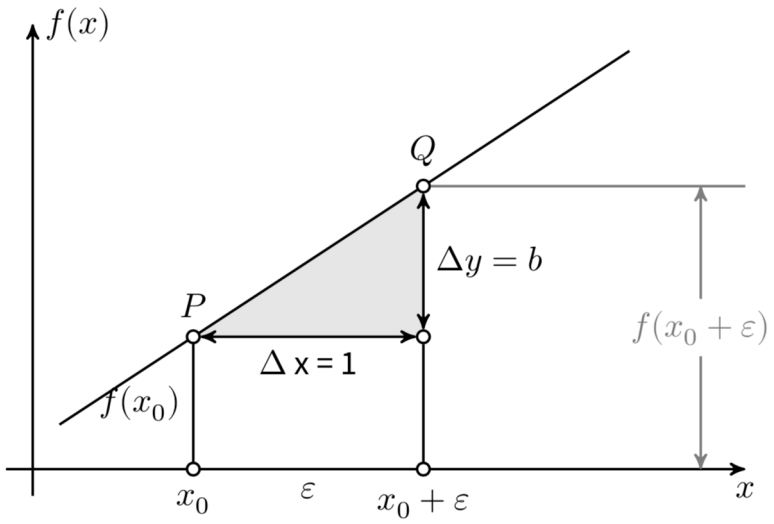


Figura 6.1: *La funzione lineare* $y = a + bx$.

non è mai perfettamente lineare. È dunque necessario includere nel modello lineare anche una componente d'errore, ovvero una componente della y che non può essere spiegata dal modello lineare. Nel caso di due sole variabili, questo ci conduce alla seguente formulazione del modello lineare:

$$y = \alpha + \beta x + \varepsilon, \tag{6.5}$$

laddove i parametri α e β descrivono l'associazione tra le variabili casuali y e x, e il termine d'errore ε specifica quant'è grande la porzione della variabile y che non può essere predetta nei termini di una relazione lineare con la x.

Si noti che la (6.5) consente di formulare una predizione, nei termini di un modello lineare, del valore atteso della y conoscendo x, ovvero

$$\hat{y} = \mathbb{E}(y \mid x) = \alpha + \beta x. \tag{6.6}$$

In altri termini, se i parametri del modello (α e β) sono noti, allora è possibile predire la y sulla base della nostra conoscenza della x. Per esempio, se conosciamo la relazione lineare tra quoziente di intelligenza

ed aspettativa di vita, allora possiamo prevedere quanto a lungo vivrà una persona sulla base del suo QI. Sì, c'è una relazione lineare tra intelligenza e aspettativa di vita (Hambrick, 2015)! Ma quando è accurata la previsione? Ciò dipende dal termine d'errore della (6.5). Il modello lineare fornisce un metodo per rispondere a domande di questo tipo[1].

6.3 Una media per ciascuna osservazione

In precedenza abbiamo visto come sia possibile stimare i parametri di un modello bayesiano Normale nel quale le osservazioni sono indipendenti e identicamente distribuite secondo una densità Normale,

$$Y_i \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma), \quad i = 1, \dots, n. \tag{6.7}$$

Il modello (6.7) assume che ogni Y_i sia una realizzazione della stessa $\mathcal{N}(\mu, \sigma^2)$. Da un punto di vista bayesiano[3], si assegnano distribuzioni a priori ai parametri μ e σ, si genera la verosimiglianza in base ai dati osservati e, con queste informazioni, si generano le distribuzione a posteriori dei parametri (Gelman et al., 2020):

$$\begin{aligned} Y_i \mid \mu, \sigma &\overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2) \\ \mu &\sim \mathcal{N}(\mu_0, \tau^2) \\ \sigma &\sim \text{Cauchy}(x_0, \gamma) \end{aligned}$$

È comune però che vengano però registrate altre variabili x_i che possono essere associate alla risposta di interesse y_i. La variabile x_i viene chiamata *predittore* (o variabile indipendente) in quanto il ricercatore è tipicamente interessato a predire il valore y_i a partire da x_i. Come si può estende il modello Normale della (6.7) per lo studio della possibile relazione tra y_i e x_i?

Il modello (6.7) assume una media μ comune per ciascuna osservazione Y_i. Dal momento che desideriamo introdurre una nuova variabile x_i che

[1] Per una discussione sugli aspetti di base del modello lineare, si veda il capitolo 7[2] di *Introduction to Modern Statistics*.

[3] Per un'introduzione alla trattazione frequentista del modello lineare, si veda l'Appendice **??**.

assume un valore specifico per ciascuna osservazione y_i, il modello (6.7) può essere modificato in modo che la media comune μ venga sostituita da una media μ_i specifica a ciascuna i-esima osservazione:

$$Y_i \mid \mu_i, \sigma \overset{ind}{\sim} \mathcal{N}(\mu_i, \sigma), \quad i = 1, \dots, n. \tag{6.8}$$

Si noti che le osservazioni $Y_1, \dots, Y_n$ non sono più identicamente distribuite poiché hanno medie diverse, ma sono ancora indipendenti come indicato dalla notazione `ind` posta sopra il simbolo $\sim$ nella (6.8)

6.3.1 Relazione lineare tra la media $y \mid x$ e il predittore

L'approccio che consente di mettere in relazione un predittore x_i con la risposta Y_i è quello di assumere che la media di ciascuna Y_i, ovvero μ_i, sia una funzione lineare del predittore x_i. Una tale relazione lineare è scritta come

$$\mu_i = \beta_0 + \beta_1 x_i, \quad i = 1, \dots, n. \tag{6.9}$$

Nella (6.9), ciascuna x_i è una costante nota (ecco perché viene usata una lettera minuscola per la x) e β_0 e β_1 sono parametri incogniti. Questi parametri che rappresentano l'intercetta e la pendenza della retta di regressione sono variabili casuali. Si assegna una distribuzione a priori a β_0 e a β_1 e si esegue l'inferenza riassumendo la distribuzione a posteriori di questi parametri.

In questo modello, la funzione lineare $\beta_0 + \beta_1 x_i$ è interpretata come il valore atteso della Y_i per ciascun valore x_i, mentre l'intercetta β_0 rappresenta il valore atteso della Y_i quando $x_i = 0$. Il parametro β_1 (pendenza) rappresenta invece l'aumento medio della Y_i quando x_i aumenta di un'unità. È importante notare che la relazione lineare (6.8) di parametri β_0 e β_1 descrive l'associazione tra *la media* μ_i e il predittore x_i. In altri termini, tale relazione lineare ci fornisce una predizione sul valore medio μ_i, non sul valore *effettivo* Y_i.

6.3.2 Il modello lineare

Sostituendo la (6.9) nella (6.8) otteniamo il modello lineare:

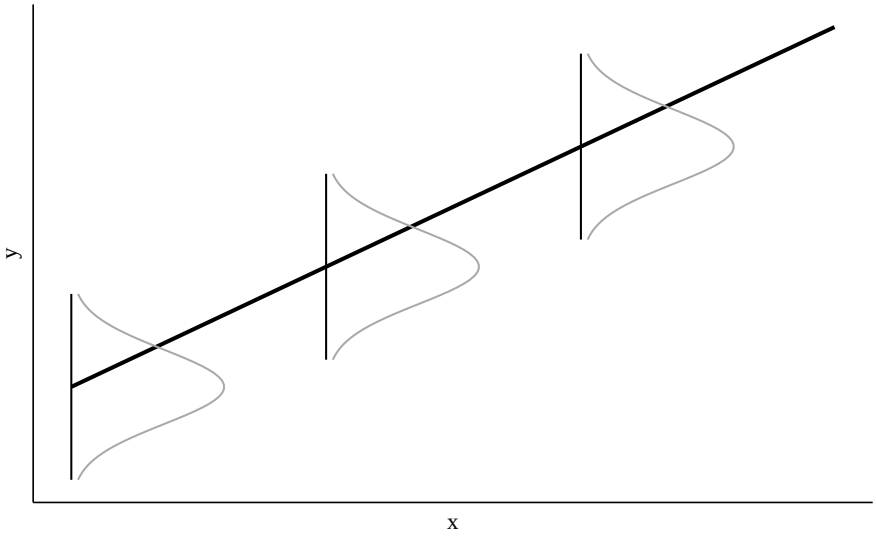
$$Y_i \mid \beta_0, \beta_1, \sigma \overset{ind}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma), \quad i = 1, \dots, n. \tag{6.10}$$

Questo è un caso speciale del modello di campionamento Normale, dove le Y_i seguono indipendentemente una densità Normale con una media $(\beta_0+\beta_1 x_i)$ specifica per ciascuna osservazione e con una deviazione standard (σ) comune a tutte le osservazioni. Poiché include un solo predittore (x), questo modello è comunemente chiamato *modello di regressione lineare semplice*.

In maniera equivalente, il modello (6.10) può essere formulato come

$$Y_i = \mu_i + \varepsilon_i, \quad i = 1, \dots, n, \tag{6.11}$$

dove la risposta media è $\mu_i = \beta_0 + \beta_1 x_i$ e i residui $\varepsilon_1, \dots, \varepsilon_n$ sono i.i.d. da una Normale con media 0 e deviazione standard σ.



Nel modello lineare, l'osservazione Y_i è una variabile casuale, il predittore x_i è una costante fissa, e β_0, β_1 e σ sono parametri incogniti. Utilizzando il paradigma bayesiano, viene assegnata una distribuzione a priori congiunta a $(\beta_0, \beta_1, \sigma)$. Dopo avere osservato le risposte $Y_i, i = 1, \dots, n$, l'inferenza procede stimando la distribuzione a posteriori dei parametri.

Osservazione. Nella costruzione di un modello di regressione bayesiano, è importante iniziare dalle basi e procedere un passo alla volta. Sia Y una variabile di risposta e sia x un predittore o un insieme di predittori. È possibile costruire un modello di regressione di Y su x applicando i seguenti principi generali:

- Stabilire se Y è discreto o continuo. Di conseguenza, identificare l'appropriata struttura dei dati (per esempio, Normale, di Poisson, o Binomiale).
- Esprimere la media di Y come funzione dei predittori x (per esempio, $\mu = \beta_0 + \beta_1 x$).
- Identificare tutti i parametri incogniti del modello (per esempio, μ, β_1, β_2).
- Valutare quali valori che ciascuno di questi parametri potrebbe assumere. Di conseguenza, identificare le distribuzioni a priori appropriate per questi parametri.

Nel caso di una variabile Y continua che segue la legge gaussiana e un solo predittore, ad esempio, il modello diventa:

$$\begin{aligned} Y_i \mid \beta_0, \beta_1, \sigma &\overset{ind}{\sim} \mathcal{N}(\mu_i, \sigma^2) \quad \text{con} \quad \mu_i = \beta_0 + \beta_1 x_i \\ \beta_0 &\sim \mathcal{N}(\mu_0, \sigma_0^2) \\ \beta_1 &\sim \mathcal{N}(\mu_1, \sigma_1^2) \\ \sigma &\sim \text{Cauchy}(x_0, \gamma) \ . \end{aligned}$$

Un algoritmo MCMC viene usato per simulare i campioni dalle distribuzioni a posteriori e, mediante tali campioni, si fanno inferenze sulla risposta attesa $\beta_0 + \beta_1 x$ per ciascuno specifico valore del predittore x. Inoltre, è possibile valutare le dimensioni degli errori di previsione mediante un indice sintetico della densità a posteriori della deviazione standard σ.

Considerazioni conclusive

Il modello lineare semplice viene usato per descrivere la relazione tra due variabili e per determinare il segno e l'intensità di tale relazione. Inoltre, il modello lineare ci consente di prevedere il valore della variabile dipendente in base ai valori della variabile indipendente. Il modello lineare semplice è in realtà molto limitato, in quanto descrive soltanto la relazione tra la variabile dipendente y e una sola variabile esplicativa x. Esso diventa molto più utile quando incorpora più variabili indipendenti. In questo secondo caso, però, i calcoli per la stima dei coefficienti del modello diventano più complicati. Abbiamo deciso di iniziare considerando il

modello lineare semplice perché, in questo caso, sia la logica dell'inferenza sia le procedure di calcolo sono facilmente maneggiabili. Nel caso più generale, quello del modello lineare multiplo (ovvero, con più di un predittore), la logica dell'inferenza rimane identica a quella discussa qui, ma le procedure di calcolo richiedono l'uso dell'algebra matriciale. Il modello lineare multiplo può includere sia regressori quantitativi, sia regressori qualitativi, utilizzando un opportuno schema di codifica. È interessante notare come un modello lineare multiplo che include una sola variabile esplicativa qualitativa corrisponde all'analisi della varianza ad una via; un modello lineare multiplo che include più di una variabile esplicativa qualitativa corrisponde all'analisi della varianza più vie. Possiamo qui concludere dicendo che il modello lineare, nelle sue varie forme e varianti, costituisce la tecnica di analisi dei dati maggiormente usata in psicologia.

A

Simbologia di base

Per una scrittura più sintetica possono essere utilizzati alcuni simboli matematici.

- $\log(x)$: il logaritmo naturale di x.
- L'operatore logico booleano $\wedge$ significa "e" (congiunzione forte) mentre il connettivo di disgiunzione $\vee$ significa "o" (oppure) (congiunzione debole).
- Il quantificatore esistenziale $\exists$ vuol dire "esiste almeno un" e indica l'esistenza di almeno una istanza del concetto/oggetto indicato. Il quantificatore esistenziale di unicità $\exists!$ ("esiste soltanto un") indica l'esistenza di esattamente una istanza del concetto/oggetto indicato. Il quantificatore esistenziale $\nexists$ nega l'esistenza del concetto/oggetto indicato.
- Il quantificatore universale $\forall$ vuol dire "per ogni."
- $\mathcal{A}, \mathcal{S}$: insiemi.
- $x \in A$: x è un elemento dell'insieme A.
- L'implicazione logica "$\Rightarrow$" significa "implica" (se …allora). $P \Rightarrow Q$ vuol dire che P è condizione sufficiente per la verità di Q e che Q è condizione necessaria per la verità di P.
- L'equivalenza matematica " $\Longleftrightarrow$ " significa "se e solo se" e indica una condizione necessaria e sufficiente, o corrispondenza biunivoca.
- Il simbolo | si legge "tale che."
- Il simbolo $\triangleq$ (o $:=$) si legge "uguale per definizione."
- Il simbolo Δ indica la differenza fra due valori della variabile scritta a destra del simbolo.
- Il simbolo $\propto$ si legge "proporzionale a."
- Il simbolo $\approx$ si legge "circa."
- Il simbolo $\in$ della teoria degli insiemi vuol dire "appartiene" e indica l'appartenenza di un elemento ad un insieme. Il simbolo $\notin$ vuol dire "non appartiene."
- Il simbolo $\subseteq$ si legge "è un sottoinsieme di" (può coincidere con l'insieme stesso). Il simbolo $\subset$ si legge "è un sottoinsieme proprio di."

- Il simbolo # indica la cardinalità di un insieme.
- Il simbolo $\cap$ indica l'intersezione di due insiemi. Il simbolo $\cup$ indica l'unione di due insiemi.
- Il simbolo $\emptyset$ indica l'insieme vuoto o evento impossibile.
- In matematica, argmax identifica l'insieme dei punti per i quali una data funzione raggiunge il suo massimo. In altre parole, $\text{argmax}_x f(x)$ è l'insieme dei valori di x per i quali $f(x)$ raggiunge il valore più alto.
- a, c, α, γ: scalari.
- x, y: vettori.
- X, Y: matrici.
- $X \sim p$: la variabile casuale X si distribuisce come p.
- $p(\cdot)$: distribuzione di massa o di densità di probabilità.
- $p(y \mid x)$: la probabilità o densità di y dato x, ovvero $p(y = Y \mid x = X)$.
- $f(x)$: una funzione arbitraria di x.
- $f(X; \theta, \gamma)$: f è una funzione di X con parametri θ, γ. Questa notazione indica che X sono i dati che vengono passati ad un modello di parametri θ, γ.
- $\mathcal{N}(\mu, \sigma^2)$: distribuzione gaussiana di media μ e varianza $sigma^2$.
- Beta(α, β): distribuzione Beta di parametri α e β.
- $\mathcal{U}(a, b)$: distribuzione uniforme con limite inferiore a e limite superiore b.
- Cauchy(α, β): distribuzione di Cauchy di parametri α (posizione: media) e β (scala: radice quadrata della varianza).
- $\mathcal{B}(p)$: distribuzione di Bernoulli di parametro p (probabilità di successo).
- Bin(n, p): distribuzione binomiale di parametri n (numero di prove) e p (probabilità di successo).
- $\mathbb{KL}(p \mid\mid q)$: la divergenza di Kullback-Leibler da p a q.

Bibliografia

de Finetti, B. (1931). Probabilismo. *Logos*, pages 163–219.

de Finetti, B. (1970). *Teoria delle probabilità: sintesi introduttiva con appendice critica*. Einaudi.

Gelman, A., Hill, J., and Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press.

Hambrick, D. (2015). Research confirms a link between intelligence and life expectancy. *Scientific American. Retrieved from http://www. scientificamerican. com/article/research-confirms-a-link-between-intelligence-and-life-expectancy*.

Horn, S. and Loewenstein, G. (2021). Underestimating learning by doing. *Available at SSRN 3941441*.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684):677–680.

Zetsche, U., Bürkner, P.-C., and Renneberg, B. (2019). Future expectations in clinical depression: Biased or realistic? *Journal of Abnormal Psychology*, 128(7):678–688.