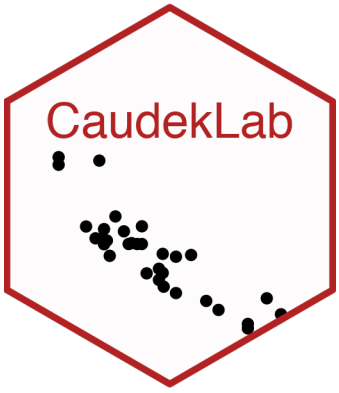
Psicometria

Corrado Caudek

2022-01-16

Questo documento è stato realizzato con:

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



Nel blog della mia pagina personale sono forniti alcuni approfondimenti degli argomenti qui trattati. https://ccaudek.github.io/caudeklab/

Indice

Data della versione presente: Gennaio 16, 2022.

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

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

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

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

Sono sicuro che, leggendo queste righe, a molti studenti sarà venuta in mente la seguente domanda: perché non chiediamo a qualche esperto di fare il "lavoro sporco" (ovvero le analisi statistiche) per noi, mentre noi (gli psicologi) ci occupiamo solo di ciò che ci interessa, ovvero dei problemi psicologici slegati dai dettagli "tecnici" della Data

science? La risposta a questa domanda è che non è possibile progettare uno studio 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 verrano trattati i temi base della Data science e verrà adottato un punto di vista bayesiano, che corrisponde all'approccio più recente e sempre più diffuso in psicologia.

Come studiare

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

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

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

Sviluppare un metodo di studio efficace

> *Memorization is not learning.*
>
> — Richard Phillips Feynman

Avendo insegnato molte volte in passato un corso introduttivo di analisi dei dati ho notato nel corso degli anni che gli studenti con l'atteggiamento mentale che descriverò qui sotto generalmente ottengono ottimi risultati. Alcuni studenti sviluppano naturalmente questo approccio allo studio, ma altri hanno bisogno di fare uno sforzo per maturarlo.

Fornisco qui sotto una breve descrizione del "metodo di studio" che, nella mia esperienza, è il più efficace per affrontare le richieste di questo insegnamento (Burger & Starbird, 2012).

- Dedicate un tempo sufficiente al materiale di base, apparentemente facile; assicuratevi di averlo capito bene. Cercate le lacune nella vostra comprensione. Leggere presentazioni diverse dello stesso materiale (in libri o articoli diversi) può fornire nuove intuizioni.
- Gli errori che facciamo sono i nostri migliori maestri. Istintivamente cerchiamo di dimenticare subito i nostri errori. Ma il miglior modo di imparare è apprendere dagli errori che commettiamo. In questo senso, una soluzione corretta è meno utile di una soluzione sbagliata. Quando commettiamo un errore questo ci fornisce un'informazione importante: ci fa capire qual è il materiale di studio sul quale dobbiamo ritornare e che dobbiamo capire meglio.
- C'è ovviamente un aspetto "psicologico" nello studio. Quando un esercizio o problema ci sembra incomprensibile, la cosa migliore da fare è dire: "mi arrendo", "non ho idea di cosa fare!". Questo ci rilassa: ci siamo già arresi, quindi non abbiamo niente da perdere, non dobbiamo più preoccuparci. Ma non dobbiamo fermarci qui. Le cose "migliori" che faccio (se ci sono) le faccio quando non ho voglia di lavorare. Alle volte, quando c'è qualcosa che non so fare e non ho idea di come affontare, mi dico: "oggi non ho proprio voglia di fare fatica", non ho voglia di mettermi nello stato mentale per cui "in 10 minuti devo risolvere il problema perché dopo devo fare altre cose". Però ho voglia di *divertirmi* con quel problema e allora mi dedico a qualche aspetto "marginale" del problema, che so come affrontare, oppure considero l'aspetto più difficile del problema, quello che non so come risolvere, ma invece di cercare di risolverlo, guardo come altre persone hanno affrontato problemi simili, opppure lo stesso problema in un altro contesto. Non mi pongo l'obiettivo "risolvi il problema in 10 minuti", ma invece quello di farmi un'idea "generale" del problema, o quello di capire un caso più specifico e più semplice del problema. Senza nessuna pressione. Infatti, in quel momento ho deciso di non lavorare (ovvero, di non fare fatica). Va benissimo se "parto per la tangente", ovvero se mi metto a leggere del materiale che sembra avere poco a che fare con il problema centrale (le nostre intuizioni e la nostra curiosità solitamente ci indirizzano sulla strada giusta). Quando faccio così, molto spesso trovo la soluzione del problema che mi ero posto e, paradossalmente, la trovo in un tempo minore di quello che, in precedenza, avevo dedicato a "lavorare" al problema. Allora perché non faccio sempre così? C'è ovviamente l'aspetto dei "10 minuti" che non è sempre facile da dimenticare. Sotto pressione, possiamo solo agire in maniera automatica, ovvero possiamo solo applicare qualcosa che già sappiamo fare. Ma se dobbiamo imparare qualcosa di nuovo, la pressione è un impedimento.
- È utile farsi da soli delle domande sugli argomenti trattati, senza limitarsi a cercare di risolvere gli esercizi che vengono assegnati. Quando studio qualcosa mi viene in mente: "se questo è vero, allora deve succedere quest'altra cosa". Allora verifico se questo è vero, di solito con una simulazione. Se i risultati della simulazione sono quelli che mi aspetto, allora vuol dire che ho capito. Se i risultati sono diversi da quelli che mi aspettavo, allora mi rendo conto di non avere capito e ritorno indietro a studiare con più attenzione la teoria che pensavo di avere capito – e ovviamente mi rendo conto che c'era un aspetto che avevo frainteso. Questo tipo di verifica è qualcosa che dobbiamo fare da soli, in prima persona: nessun altro può fare questo al posto nostro.
- Non aspettatevi di capire tutto la prima volta che incontrate un argomento nuovo.[1] È utile farsi una nota mentalmente delle lacune nella vostra comprensione e tornare su di esse in seguito per carcare di colmarle. L'atteggiamento naturale, quando

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

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

Corrado Caudek

Capitolo 1

Confronto tra due gruppi indipendenti

Il problema del confronto tra due gruppi indipendenti può essere formulato nei termini di un modello lineare nel quale la variabile x è dicotomica, ovvero assume solo due valori.

1.1 Modello lineare con una variabile dicotomica

Se x è una variabile dicotomica con valori 0 e 1, allora per il modello lineare $\mu_i = \alpha + \beta x_i$ abbiamo quanto segue. Quando $x = 0$, il modello diventa

$$\mu_i = \alpha$$

mentre, quando $X = 1$, il modello diventa

$$\mu_i = \alpha + \beta.$$

Ciò significa che il parametro α è uguale al valore atteso del gruppo codificato con $X = 0$ e il parametro β è uguale alla differenza tra le medie dei due gruppi (essendo la media del secondo gruppo uguale a $\alpha + \beta$). Il parametro β, dunque, codifica l'effetto di una manipolazione sperimentale o di un trattamento, e l'inferenza su β corrisponde direttamente all'inferenza sull'efficacia di un trattamento o di un effetto sperimentale.[1] L'inferenza su β, dunque, viene utilizzata per capire quanto "credibile" può essere considerato l'effetto di un trattamento o di una manipolazione sperimentale.

Un esempio concreto

Esaminiamo nuovamente i dati `kid_score` discussi da Gelman et al. (2020). La domanda della ricerca è se il QI del figlio (misurato sulla scala PIAT) è associato al livello di istruzione della madre.

Codifichiamo il livello di istruzione della madre (x) con una *variabile indicatrice* (ovvero, una variabile che assume solo i valori 0 e 1) tale per cui:

- $x = 0$: la madre non ha completato la scuola secondaria di secondo grado (scuola media superiore);
- $x = 1$: la madre ha completato la scuola media superiore.

Supponiamo che i dati siano contenuti nel data.frame `df`.

[1] Per "effetto di un trattamento" si intende la differenza tra le medie di due gruppi (per esempio, il gruppo "sperimentale" e il gruppo "di controllo").

```
library("rio")
df <- rio::import(here("data", "kidiq.dta"))
```

Calcoliamo le statistiche descrittive per i due gruppi:

```
df %>%
  group_by(mom_hs) %>%
  summarise(
    mean_kid_score = mean(kid_score),
    std = sqrt(var(kid_score))
  )
#> # A tibble: 2 × 3
#>   mom_hs mean_kid_score   std
#>    <dbl>          <dbl> <dbl>
#> 1      0           77.5  22.6
#> 2      1           89.3  19.0
```

Il punteggio medio PIAT è pari a 77.5 per i bambini la cui madre non ha il diploma di scuola media superiore e pari a 89.3 per i bambini la cui madre ha completato la scuola media superiore. Questa differenza suggerisce un'associazione tra le variabili, ma tale differenza potrebbe essere soltanto la conseguenza della variabilità campionaria, senza riflettere una caratteristica generale della popolazione. Come possiamo usare il modello statistico lineare per fare inferenza sulla differenza osservata tra i due gruppi? Non dobbiamo fare nient'altro che usare il modello lineare che abbiamo definito in precedenza.

```
modelString = "
data {
  int<lower=0> N;
  vector[N] y;
  vector[N] x;
}
transformed data {
  vector[N] x_std;
  vector[N] y_std;
  x_std = (x - mean(x)) / sd(x);
  y_std = (y - mean(y)) / sd(y);
}
parameters {
  real alpha_std;
  real beta_std;
  real<lower=0> sigma_std;
}
model {
  alpha_std ~ normal(0, 2);
  beta_std ~ normal(0, 2);
  sigma_std ~ cauchy(0, 2);
  y_std ~ normal(alpha_std + beta_std * x_std, sigma_std);
}
generated quantities {
  real alpha;
  real beta;
  real<lower=0> sigma;
  real cohen_d;
  alpha = sd(y) * (alpha_std - beta_std * mean(x) / sd(x))
```

```
          + mean(y);
  beta = beta_std * sd(y) / sd(x);
  sigma = sd(y) * sigma_std;
  cohen_d = beta / sigma;
}
"
writeLines(modelString, con = "code/simpleregstd.stan")
```

Come in precedenza, salviamo i dati in un oggetto di classe `list`:

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

Compiliamo il modello:

```
file <- file.path("code", "simpleregstd.stan")
mod <- cmdstan_model(file)
```

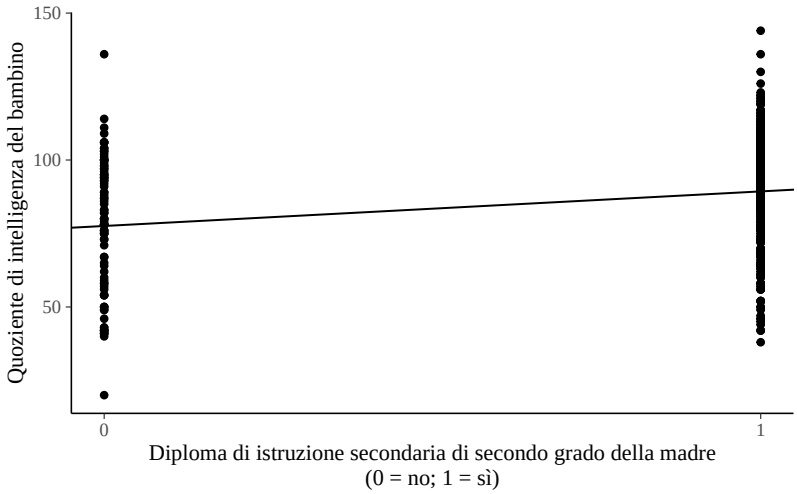
Adattiamo il modello ai dati:

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

Creiamo un grafico con i valori predetti dal modello lineare:

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

```
tibble(
  kid_score = df$kid_score,
  mom_hs = df$mom_hs
) %>%
  ggplot(aes(mom_hs, kid_score)) +
  geom_point() +
  geom_abline(intercept = mean(posterior$alpha), slope = mean(posterior$beta)) +
  labs(
    y = "Quoziente di intelligenza del bambino",
    x = "Diploma di istruzione secondaria di secondo grado della madre\n(0 = no; 1 = sì)"
  ) +
  scale_x_continuous(breaks=c(0, 1))
```



Le stime a posteriori dei parametri si ottengono con:

```
fit$summary(c("alpha", "beta", "sigma", "cohen_d"))
#> # A tibble: 4 × 10
#>   variable   mean median    sd   mad     q5    q95  rhat ess_bulk
#>   <chr>     <dbl>  <dbl> <dbl> <dbl>  <dbl>  <dbl> <dbl>    <dbl>
#> 1 alpha    77.6   77.5   2.08  2.06  74.1   81.0    1.00   16538.
#> 2 beta     11.8   11.7   2.35  2.34   7.88  15.6    1.00   16718.
#> 3 sigma    19.9   19.9   0.676 0.671 18.8   21.0    1.00   15949.
#> 4 cohen_d   0.592  0.591 0.120 0.119  0.393  0.788  1.00   16771.
#> # … with 1 more variable: ess_tail <dbl>
```

I risultati confermano ciò che ci aspettavamo:

- il coefficiente `alpha` = 77.56 corrisponde alla media del gruppo codificato con $x = 0$, ovvero la media dei punteggi PIAT per i bambini la cui madre non ha completato la scuola media superiore;
- il coefficiente `beta` = 11.76 corrisponde alla differenza tra le medie dei due gruppi, ovvero 89.32 - 77.55 = 11.77 (con piccoli errori di approssimazione).

La seguente chiamata ritorna l'intervallo di credibilità al 95% per tutti i parametri del modello:

```
rstantools::posterior_interval(as.matrix(stanfit), prob = 0.95)
#>                2.5%     97.5%
#> alpha_std   -0.094    0.0925
#> beta_std     0.144    0.3289
#> sigma_std    0.913    1.0437
#> alpha       73.432   81.6209
#> beta         7.135   16.3396
#> sigma       18.643   21.3029
#> cohen_d      0.357    0.8277
#> lp__      -208.906 -204.3240
```

Possiamo dunque concludere che i bambini la cui madre ha completato la scuola superiore ottengono in media circa 12 punti in più rispetto ai bambini la cui madre non ha completato la scuola superiore. L'intervallo di credibilità al 95% ci dice che possiamo essere sicuri al 95% che tale differenza sia di almeno 7 punti e possa arrivare fino a ben

16 punti. Per riassumere, possiamo concludere, con un grado di certezza soggettiva del 95%, che c'è un'associazione positiva tra il livello di scolarità della madre e l'intelligenza del bambino: le madri che hanno livello di istruzione più alto della media tendo ad avere bambini il cui QI è anch'esso più alto della media.

1.2 La dimensione dell'effetto

Nel caso di due gruppi indipendenti, la dimensione dell'effetto si può stimare con la statistica d di Cohen:

$$d = \frac{\bar{y}_1 - \bar{y}_2}{s}.$$

Nel caso presente, la differenza $\bar{y}_1 - \bar{y}_2$ corrisponde a al parametro β del modello lineare. Inoltre, una stima della deviazione starndard comune dei due gruppi è fornita dalla deviazione standard della regressione, ovvero dal parametro σ. Nel blocco `generated quantities` del modello Stan ho calcolato `cohen_d = beta / sigma`. Ciò significa che Stan calcolerà la distribuzione a posteriori del parametro `cohen_d`. Possiamo dunque riassumere la distribuzione a posteriori di `cohen_d` con un qualche indice di tendenza centrale (che sarà la nostra stima della dimensione dell'effetto) e calcolare l'intervallo di credibilità, per esempio al 95%. Questi risultati si ottengono con l'istruzione riportata di seguito:

```
posterior::summarise_draws(
  stanfit,
  ~quantile(.x, probs = c(0.025, 0.5, 0.975))
)
#> # A tibble: 8 × 4
#>   variable     `2.5%`       `50%`   `97.5%`
#>   <chr>         <dbl>       <dbl>     <dbl>
#> 1 alpha_std   -0.0940   -0.000366    0.0925
#> 2 beta_std     0.144     0.236       0.329
#> 3 sigma_std    0.913     0.974       1.04
#> 4 alpha       73.4      77.5        81.6
#> 5 beta         7.14     11.7        16.3
#> 6 sigma       18.6      19.9        21.3
#> 7 cohen_d      0.357     0.591       0.828
#> 8 lp__      -209.     -205.       -204.
```

I risultati dell'analisi bayesiana coincidono con quelli che si ottengono utilizzando la formula del d di Cohen con le medie dei due gruppi e una stima della varianza *pooled*. Il calcolo della statistica d di Cohen è fornita, ad esempio, dal pacchetto `effectsize`:

```
library("effectsize")
(d <- cohens_d(kid_score ~ mom_hs, data = df))
#> Cohen's d |         95% CI
#> -------------------------
#> -0.59     | [-0.83, -0.36]
#>
#> - Estimated using pooled SD.
```

Il fatto che l'output abbia un segno negativo dipende dal fatto che è stata sottratta la media maggiore dalla media minore (in altri termini, dobbiamo guardare il risultato in valore assoluto).

In conclusione, il valore d di Cohen di entità "media" [$d > 0.5$; Sawilowsky (2009)] può essere interpretato dicendo che la scolarità delle madri ha un'influenza non trascurabile sul QI dei bambini.

Bibliografia

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

Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press. (Cit. a p. 1).

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

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of modern applied statistical methods*, *8*(2), 26 (cit. a p. 5).

Elenco delle figure

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

Keywords Data science, Bayesian statistics.