Analisi dei dati con R

Test per la verifica di ipotesi e di significatività pura

Marco Alfo'

SAPIENZA
UNIVERSITÀ DI ROMA

Modello statistico

Un modello statistico rappresenta, in modo **semplice**, un fenomeno potenzialmente **complesso**.

L'idea di base è che l'osservazione empirica è prodotta da un *processo di generazione del dato* che può essere adeguatamente approssimato da un modello matematico **semplice**

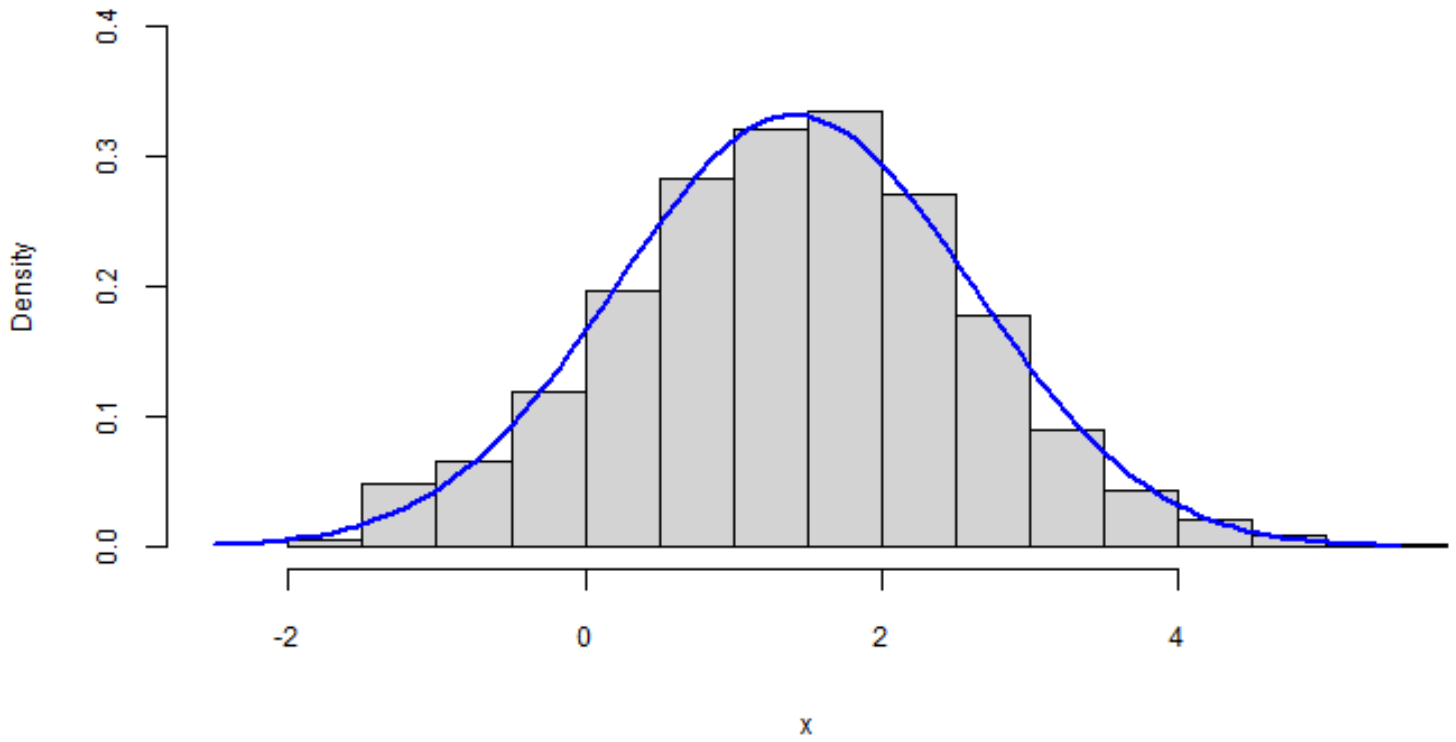
Tale approssimazione è basata sulla **conoscenza** del fenomeno e sulla necessità di **rappresentazione parsimoniosa** del dato osservato

La scelta di un modello permette di **ridurre la complessità** del compito, fornendo una struttura di riferimento ed un problema meglio **definito**

Tale ipotesi fornisce una **rappresentazione** della realtà e, quindi, i risultati ottenuti sono da considerarsi **condizionati a questa**

Il campione di dati osservati è sempre di dimensione **finita** ed i valori della/e variabile/i osservati sono sempre **discreti**

Modello statistico (es. grafico)



Modello statistico

L'istogramma precedente rappresenta una **popolazione reale**

La curva continua il **modello matematico** di approssimazione

Il modello serve a trarre, a partire dal campione osservato, informazioni sulla popolazione da cui questo è stato estratto (**istogramma**)

L'inferenza statistica parametrica si occupa, invece, di apprendere, dal campione a disposizione, le informazioni relative ai parametri della popolazione teorica (**curva continua**)

L'approssimazione facilita il compito, fornendo una struttura **continua**, **semplice** da trattare, nota, a meno di un numero finito di costanti che la caratterizzano (**parametri**)

$$X \sim f(\cdot | \boldsymbol{\theta})$$

Inferenza statistica parametrica

Utilizzando questo approccio, l'interesse si sposta dalla **conoscenza della popolazione**

all'**apprendimento** di quanta più **informazione** possibile sui parametri in base ai dati osservati

$$X \sim f(\cdot \mid \boldsymbol{\theta})$$

In generale, gli obiettivi possono essere

- Stima puntuale (**valore plausibile** sulla base del campione)
- Stima intervallare (**regione di valori plausibili** sulla base del campione, con una certa **confidenza**)
- Verifica di **ipotesi** (sul vettore dei parametri)

Verifica di ipotesi statistiche

Un'**ipotesi statistica** è una congettura/affermazione riguardante un parametro della popolazione di interesse

- **Ipotesi nulla** H_0: ipotesi da sottoporre a verifica, ritenuta vera fino a «prova contraria» e rifiutata solo se i dati osservati mostrano uno scostamento che può essere considerato *significativo*.

Un esempio di ipotesi nulla (Borra, Di Ciaccio, 2008) è l'affermazione che l'altezza media degli italiani nati nel 1980 sia pari a 175 cm,

$$H_0: \mu_X = 175$$

- **Ipotesi alternativa** H_1: affermazione che contraddice l'ipotesi nulla in una direzione specifica e che deve avere senso dal punto di vista empirico.

Un esempio di ipotesi alternativa è che l'altezza media degli italiani nati nel 1980 sia diversa da 175 cm:

$$H_1: \mu \neq 175$$

Sistema di ipotesi

- L'insieme dei possibili valori che il parametro di interesse θ può assumere costituisce il cosiddetto spazio parametrico Ω
- Quando si effettua un test di ipotesi, lo spazio parametrico viene suddiviso in due regioni (mutuamente esclusive) corrispondenti all'ipotesi **nulla** Ω_0 e all'ipotesi **alternativa** Ω_1
- La bipartizione dello spazio parametrico definisce il **sistema di ipotesi**

$$\begin{cases} \mathrm{H}_0 \colon \theta \in \Omega_0 \\ \mathrm{H}_1 \colon \theta \in \Omega_1 \end{cases}$$

La verifica di un'ipotesi statistica consiste nello stabilire se il campione osservato contiene ''**sufficiente**'' evidenza per rifiutare l'ipotesi nulla

Sistema di ipotesi

- Si parla di ipotesi *semplici o puntuali* se $\Omega_0 = \theta_0$ e/o $\Omega_1 = \theta_1$, ossia se le ipotesi definiscono in maniera puntuale (univoca) il valore del parametro di interesse (vincolo di uguaglianza)
- In questo caso, si ottiene il sistema di ipotesi

$$\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta = \theta_1 \end{cases}$$

- Si parla di *ipotesi compos(i)te* quando Ω_0 e/o Ω_1 inlcudono un insieme di possibili valori per il parametro di interesse (ad es. vincoli di disuguaglianza, non superiorità, non inferiorità)

$$\begin{cases} H_0: \theta \leq 0 \\ H_1: \theta > 0 \end{cases} \quad \text{oppure} \quad \begin{cases} H_0: \theta = 0 \\ H_1: \theta \neq 0 \end{cases}$$

Test di ipotesi

Come stabilire quando un campione contiene ''sufficiente'' evidenza per rifiutare l'ipotesi nulla?

L'inferenza statistica utilizza (almeno) due approcci al problema:

- regioni di **rifiuto** e **non rifiuto** → test per la **verifica di ipotesi** (parametriche)
- **p-value** → test di **significatività** (pura)

I due approcci provengono da tradizioni e sensibilità diverse, ma vengono, spesso, confusi

La confusione “operativa” non permette, però, di concentrare l’attenzione su ciò che ciascuno degli approcci **può** (oppure **non può**) fare

Test per la verifica di ipotesi parametriche

La distribuzione parametrica

- E’ bene considerare che l’ipotesi nulla specifica **completamente** il processo di generazione del campione osservato
  - quindi, mentre in generale

$$X \sim f\,(x \mid \theta\,)$$

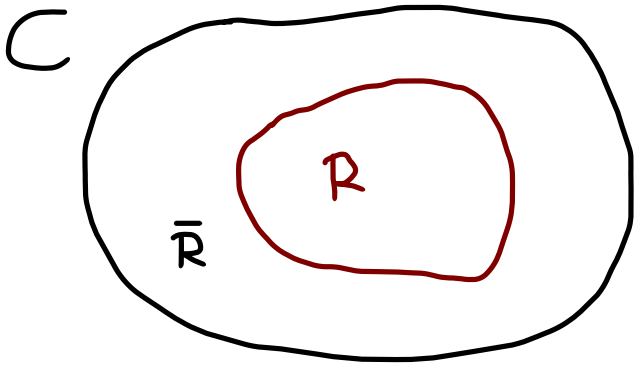
  - se si ipotizza che l’ipotesi nulla sia vera, ossia sub H_0

$$X \sim f\,(x \mid \theta_0\,)$$

- L’idea è quella di **misurare**, in modo opportuno, quanto il campione sia distante dal processo di generazione del dato postulato sotto H_0,
- nella ***direzione*** definita dall’ipotesi alternativa, H_1 che, quindi, non riveste lo stesso ruolo dell’ipotesi nulla

Lo spazio campionario

- Un test di ipotesi è una regola che permette di discriminare tra campioni che conducono, risp. non conducono, al rifiuto di H_0



- In figura, **C** rappresenta lo **spazio campionario**, cioè l'insieme di tutti i possibili campioni (di dimensione *n*), $\boldsymbol{x}_n = (x_1, \ldots, x_n)$
- Ogni campione che è possibile osservare è un elemento di C, $\boldsymbol{x}_n \in C$

La regione di rifiuto

- Un test di ipotesi è una regola che permette di discriminare tra campioni che conducono, risp. non conducono, al rifiuto di H_0
- Un test di ipotesi è una regola di decisione che suddivide **C** in due sottoinsiemi
  - **R,** detta anche **regione di rifiuto**, ossia l'insieme di tutti i campioni $\boldsymbol{x}_n = (x_1, \dots, x_n)$ che contengono "sufficiente" evidenza contraria all'ipotesi «nulla» H_0
  - $\overline{\boldsymbol{R}} = C \setminus \boldsymbol{R}$ detta anche **regione di non rifiuto**, ossia l'insieme di tutti i campioni $\boldsymbol{x}_n = (x_1, \dots, x_n)$ che **NON** contengono "sufficiente" evidenza contraria all'ipotesi «nulla» H_0
- Ovviamente, ogni decisione è soggetta, in condizioni di incertezza, ad una qualche forma di ***errore***

La statistica test

- La partizione dello spazio campionario viene determinata considerando una statistica test, ossia una funzione dei **SOLI** dati campionari

$$T(\boldsymbol{X}_n) = T(X_1, \dots, X_n)$$

- Il valore osservato della statistica test è indicato con

$$t_{obs} = T(\boldsymbol{x}_n)$$

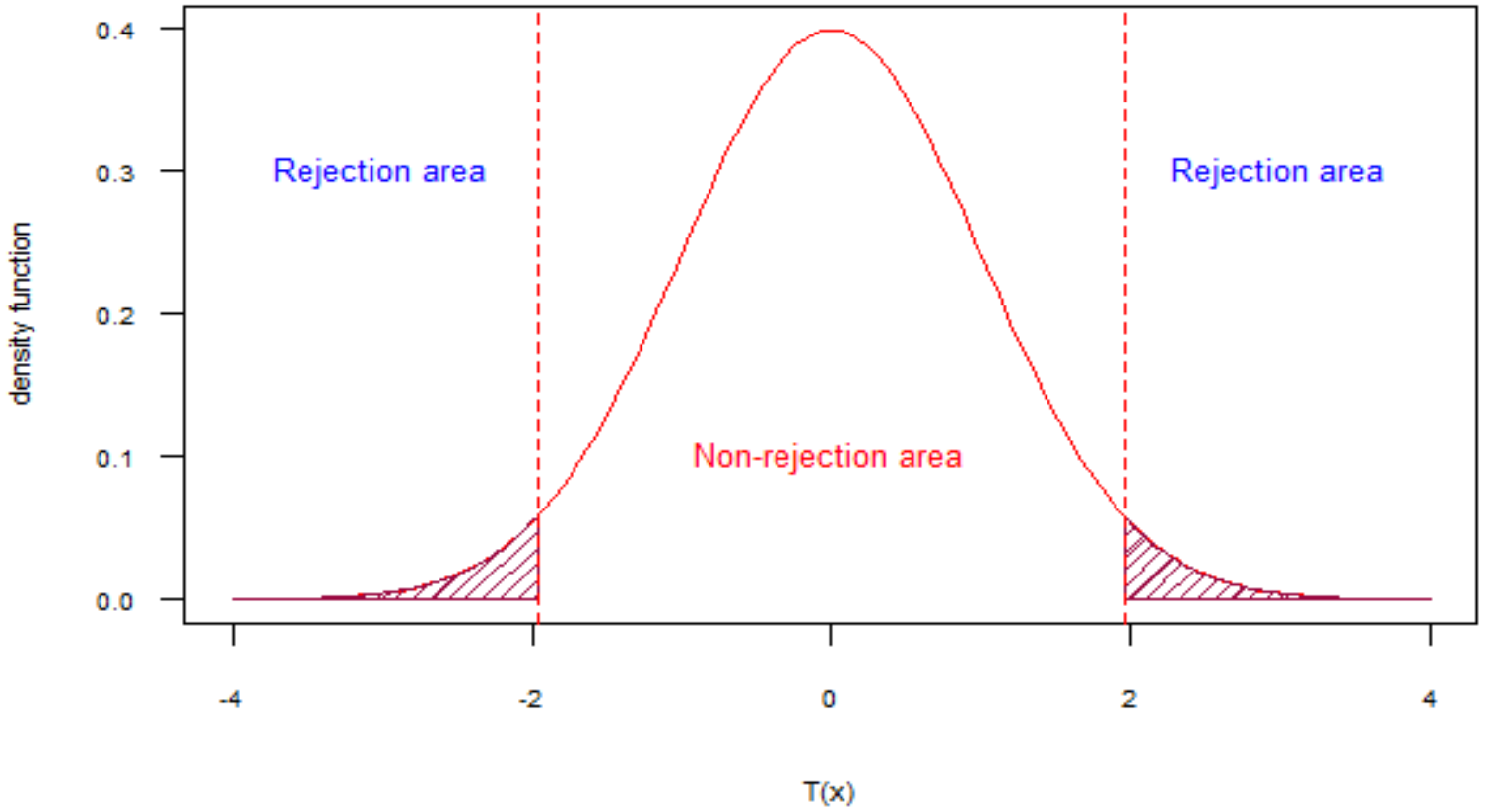
- Alla regione di rifiuto dello spazio campionario corrisponde la regione di rifiuto $\boldsymbol{R}_T$ per la statistica test, ossia

$$x_n \in \boldsymbol{R} \leftrightarrow t_{obs} \in \boldsymbol{R}_T$$

- Allo stesso modo, alla regione di non rifiuto dello spazio campionario corrisponde la regione di non rifiuto $\overline{\boldsymbol{R}}_T$ per la statistica test, ossia

$$\boldsymbol{x}_n \in \overline{\boldsymbol{R}} \leftrightarrow t_{obs} \in \overline{\boldsymbol{R}}_T$$

La statistica test (es. grafico)



Errori, definizione

- In una procedura di test, si possono commettere due tipi di **errore**

  - **I tipo/specie**: si rifiuta H_0 quando è vera ossia
  - **II tipo/specie**: non si rifiuta H_0 quando è falsa

- Il risultato del test può essere sintetizzato dalla tabella seguente

	$\overline{\boldsymbol{R}}_T$	$\boldsymbol{R}_T$
H_0 vera	✓	I specie
H_0 falsa	II specie	✓

- E' importante notare che gli errori di I e II tipo hanno un ruolo differente e, quindi, sono considerati in modo differente

Errori, test diagnostico

- L'esempio di un test diagnostico può essere utile per capire il ruolo dei due tipi di errore

  - **I tipo/specie** un soggetto sano è positivo al test (falso positivo, FP)
  - **II tipo/specie** un soggetto malato è negativo al test (falso negativo)

- Il risultato del test può essere sintetizzato dalla tabella seguente

	$negativo$	$positivo$
$sano$	TN	FP
$malato$	FN	TP

- E' chiaro, quindi, che l'errore di II tipo è, in qualche modo, **più grave** di quello di I tipo

La probabilità di errore

Come nel caso di ogni decisione **non banale**, gli errori di I e II tipo non possono essere eliminati

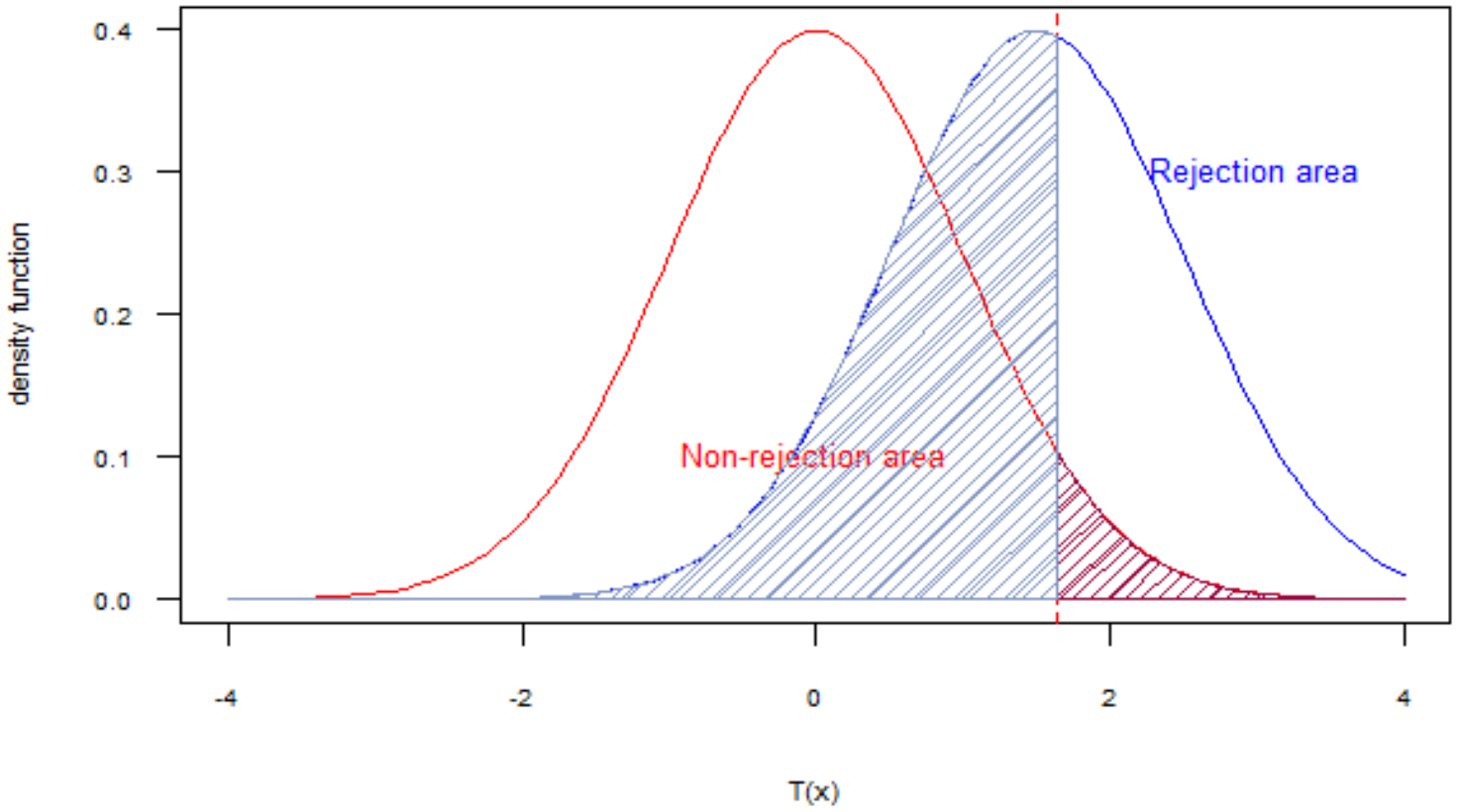
L'approccio «ideale» è una procedura capace di:

- minimizzare le rispettive probabilità:
  - $\alpha = \Pr(\boldsymbol{x}_n \in R \mid H_0) = \Pr(\mathrm{T}(\boldsymbol{x}_n) \in R_T \mid H_0)$
  - $\beta = \Pr(x_n \in \bar{R} \mid \bar{H}_0) = \Pr(\mathrm{T}(\boldsymbol{x}_n) \in \bar{R}_T \mid \bar{H}_0)$
- oppure massimizzare, fissato α, il complemento ad 1 di β

$$\gamma(\theta) = 1 - \beta(\theta) = 1 - \Pr(x_n \in \bar{R} \mid \theta),\ \theta \in \Theta$$

detta funzione di **potenza del test**

La probabilità di errore (es. grafico)



La scelta del test

- La soglia α definisce il **livello di significatività**,
- maggiore è il valore di α, più ampia è la regione di rifiuto dell'ipotesi nulla H_0

Il test **ottimo** dovrebbe avere valori bassi per α e β,

- purtroppo, α e β sono *implicitamente* legati in modo inverso (misurano l'area della regione di rifiuto e di non rifiuto, rispettivamente),
- e non possono essere **congiuntamente** minimizzati, per *n* fissato;
- si definisce, quindi, una procedura *ottimale* che, fissato un livello massimo per la probabilità di errore di primo tipo α, ricerca il test cui è associata la potenza massima.

Nella pratica…

Come si procede, in pratica?

A seconda degli obiettivi dell'analisi e del tipo di dati, si sceglie un modello statistico di riferimento, con vettore di parametri θ incognito;

- si definiscono le ipotesi da sottoporre a verifica, entrambe devono essere **plausibili**
- si seleziona la statistica test $T(\boldsymbol{X}_n)$,
- si fissa il livello di significatività α,
- si identificano le corrispondenti regioni di rifiuto/non rifiuto $\boldsymbol{R_T}$ e $\overline{\boldsymbol{R}}_{\boldsymbol{T}}$.

A questo punto, si passa al calcolo ed alla decisione basata sull'evidenza empirica…

Nella pratica…

Si calcola il valore della statistica test per il campione osservato

$$t_{obs} = T(\boldsymbol{x}_n)$$

e la decisione si basa sullo schema seguente:

- se il valore $t_{obs} \in \boldsymbol{R_T}$, si rifiuta H_0 e si definisce il risultato **statisticamente significativo**
- se $t_{obs} \in \overline{\boldsymbol{R}}_T$, si conclude che non c'è **sufficiente evidenza empirica** a sfavore di H_0, ed il risultato è dichiarato **statisticamente non significativo**.

NB

- In **nessuno** dei due casi il test si esprime sull'ipotesi nulla, ossia il test non aiuta a definirla **vera** oppure **falsa**
- In **ogni** caso, deve essere presente un'ipotesi alternativa **plausibile** per la quale si calcola il valore della funzione potenza

Test di significatività

I test di significatività

Un modo alternativo di condurre un test si basa sul concetto di ***p*-value**.

Il valore-p rappresenta la probabilità che la statistica test $T(\boldsymbol{X}_n)$ assuma valori **più estremi** di quello osservato nel campione analizzato, t_{obs}, se l'ipotesi nulla H_0 è vera.

Nel caso di statistiche test a **valori positivi**, si può scrivere

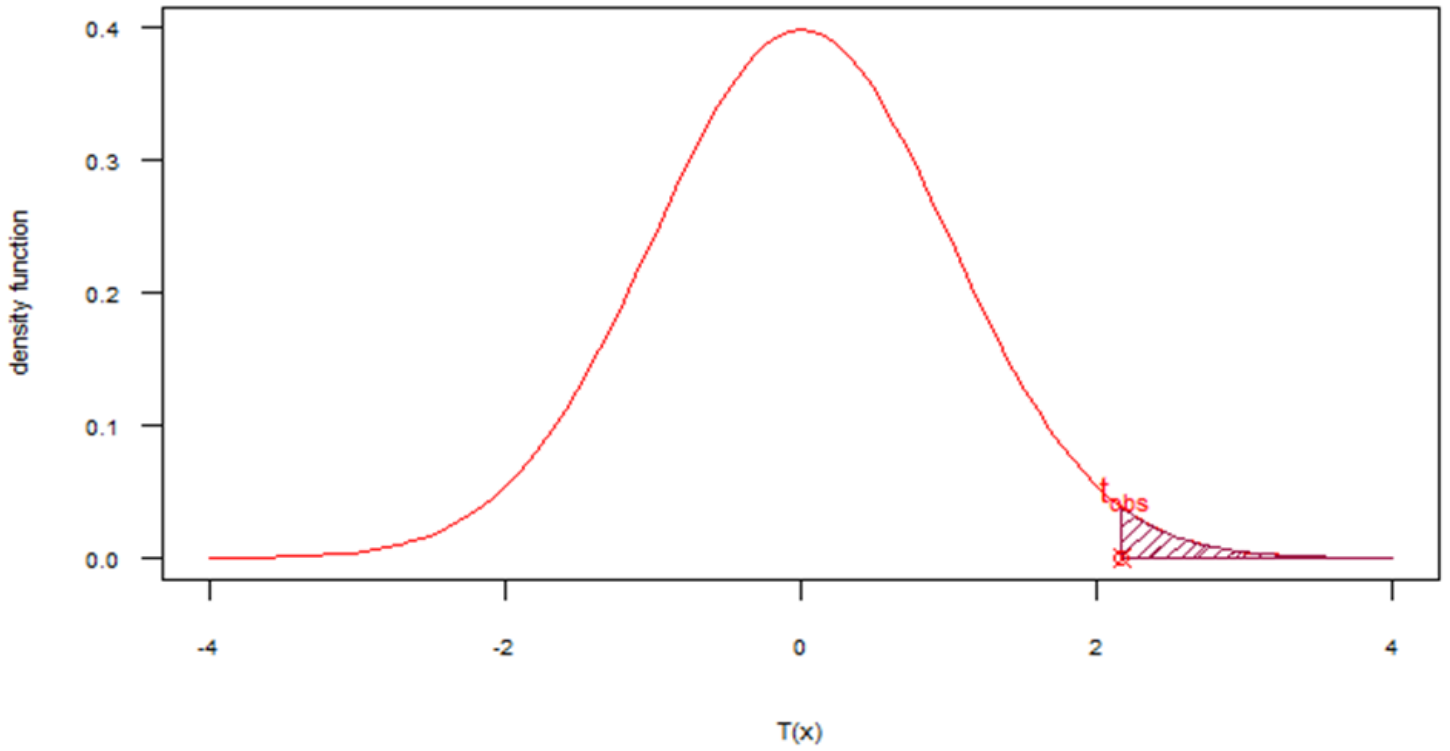
$$p = \Pr(T(\boldsymbol{X_n}) > t_{obs} \mid H_0)$$

Se si considera la statistica test come misura della **divergenza** del campione osservato dall'ipotesi nulla, minore è il valore-p, maggiore la **divergenza** del campione

Non può **MAI** essere interpretato come la probabilità che l'ipotesi nulla H_0 sia **vera**

poiché non riguarda l'ipotesi nulla, ma piuttosto il campione osservato, di cui esprime la **coerenza** all'ipotesi enunciata

I test di significatività (es. grafico)



I test di significatività in pratica…

Il valore-p viene confrontato con il **livello predeterminato** di significatività α per decidere se l'ipotesi nulla può essere rifiutata

Il processo di decisione può così essere schematizzato

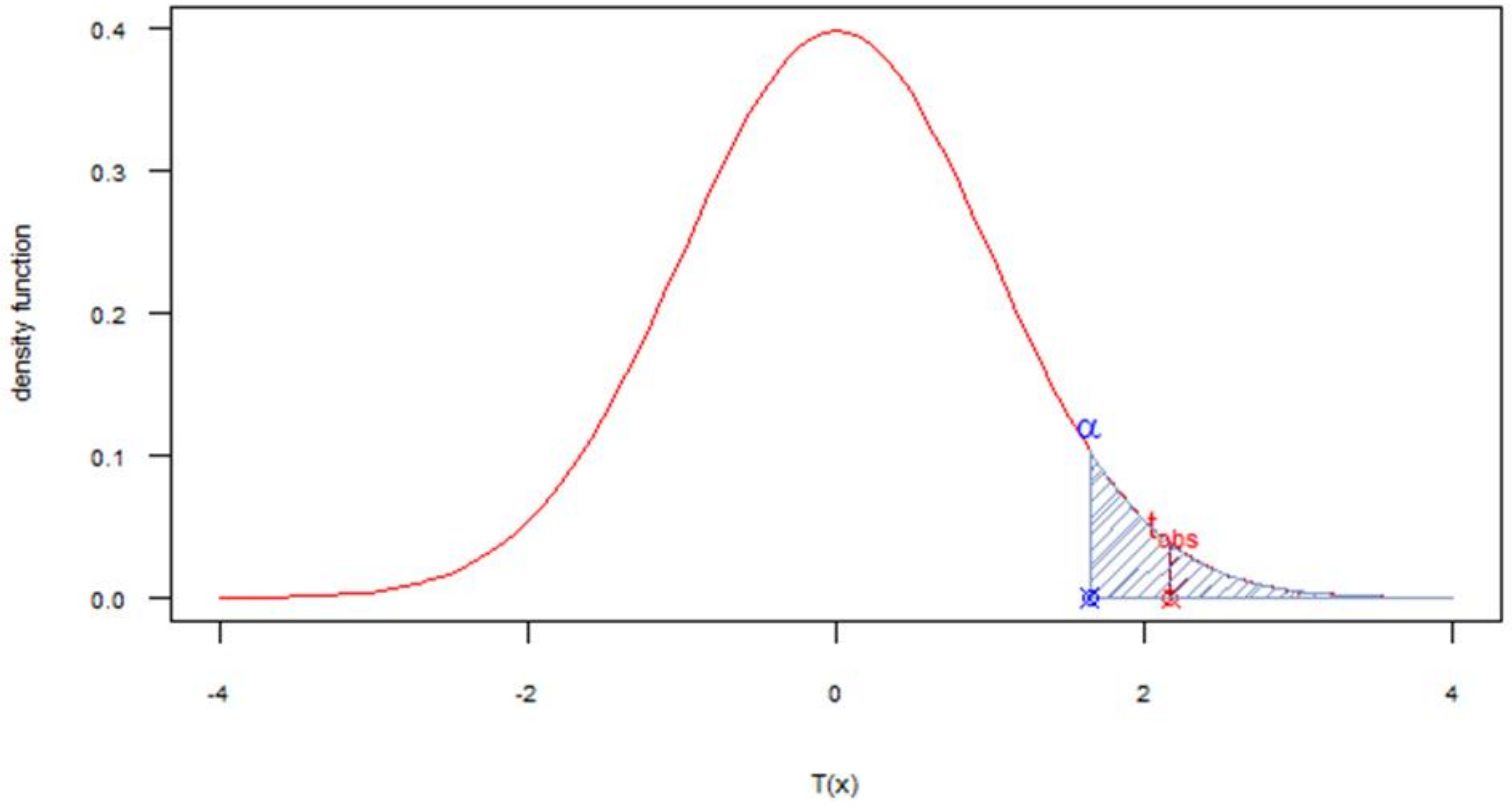
- se $p < \alpha$, il valore osservato della statistica test si trova nella zona di rifiuto $\boldsymbol{R_T}$ e **si può rifiutare** H_0 (al livello α)
- se $p > \alpha$, **non c'è sufficiente evidenza empirica** per rifiutare H_0

Il valore-p viene spesso riportato in letteratura insieme alle conclusioni tratte dal test d'ipotesi effettuato

Questo utilizzo non tiene conto dell'**errore di II specie** e della **potenza** del test

Spesso non c'è alcuna ipotesi alternativa **plausibile**

I test di significatività in pratica (es. grafico)



I test per la verifica di ipotesi e di significatività

- L'approccio basato sulle regioni di rifiuto e di non rifiuto può essere utilizzato in problemi a **bassa dimensione** ma risulta **difficilmente generalizzabile** a problemi più complessi

- In questi ultimi casi, **l'approccio grafico** non può essere di aiuto e questo riduce di molto l'applicabilità pratica dei test per la verifica di ipotesi

- Il valore-p, semplice da calcolare anche in problemi complessi, è utilizzato in modo **strumentale**, come ricordato, in un approccio *decisionale* in cui

- si evita di specificare **un'ipotesi alternativa credibile**

- e non si calcola il **valore della funzione potenza** che è un pilastro fondamentale di quella teoria

La significatività statistica…

L'approccio basato sul valore-p può essere utilizzato per sottolineare l'eventuale **incoerenza** tra campione ed ipotesi nulla

La significatività **statistica** non sempre implica significatività **sostanziale**

- la **differenza osservata** potrebbe non rappresentare un segnale sufficiente per il ricercatore
- Il valore-p potrebbe mostrare **incoerenza** tra campione ed ipotesi nulla, ma non tra campione ed ipotesi non molto distanti da questa
- al crescere della numerosità campionaria *n*, il valore-p tende a decrescere, poiché è calcolato **condizionatamente** ad un dato modello di rappresentazione
- in questo caso, dimensione crescente del campione significa volume di informazione a disposizione crescente…

La significatività statistica…

- Quindi, piuttosto che basare le nostre decisioni su un solo campione, sarebbe bene
- **diffondere** dati, metodi e strumenti così che
  - i risultati siano riproducibili (stessi dati e metodi di analisi)
  - i risultati siano replicabili (stessi metodi di analisi, diversi dati)
- cercando di arrivare ad una **decisione condivisa**, basata su più prove empiriche sotto le stesse condizioni
- che si dimostri **robusta** al passare del tempo e
- possa fornire una buona base per **accrescere** la nostra conoscenza
- Ioannidis (2005) *Why Most Published Research Findings Are False*

Un esempio, gli studi appaiati

- Si vuole sottoporre a **valutazione** un corso di formazione, impartito ad un collettivo selezionato
- Gli appartenenti al collettivo vengono valutati **prima** (A) e **dopo** (B) la loro partecipazione al corso di formazione
- Tra le due occasioni **(A vs B)** si suppone che l'**unico** cambiamento intervenuto sia la partecipazione al corso di formazione
  - a parte questo, i soggetti rimangono **invariati nelle loro caratteristiche**,
  - a ciascun soggetto $i = 1, \dots, n$ è associata la coppia (x_{iA}, x_{iB})
  - Il risultato del corso, per l'individuo è dato dalla differenza tra i valori di questa coppia, che può dipendere, solamente, dal corso

Un esempio, gli studi appaiati

- In questo caso, l'ipotesi potrebbe essere

$$\begin{cases} \mathrm{H}_0 : \mu_A = \mu_B \\ \mathrm{H}_1 : \mu_A \neq \mu_B \end{cases}$$

- dove i termini μ_A e μ_B rappresentano le medie del collettivo prima (A) e dopo (B) il corso di formazione
- Ovviamente, il sistema può essere tradotto in

$$\begin{cases} \mathrm{H}_0 : \mu_A - \mu_B = 0 \\ \mathrm{H}_1 : \mu_A - \mu_B \neq 0 \end{cases}$$

- dove

$$\mu_A - \mu_B = \mu(X_A - X_B)$$

- e la **statistica test** $T(\boldsymbol{X}_n) = T(\overline{\boldsymbol{X}_A - \boldsymbol{X}_B})$

Un esempio, gli RCT

- Si vuole sottoporre a **valutazione** un farmaco, impartito ad un collettivo selezionato
- Gli appartenenti al collettivo vengono associati, in modo casuale, (randomizzati) **al farmaco** (A) o **al farmaco** (B),
- I due farmaci non vengono somministrati allo stesso individuo, perché potrebbe esserci una qualche interazione tra i due
- In questo caso,
  - a ciascun soggetto $i = 1, \dots, n$ è associata un solo valorex_{iA} oppure x_{iB}
  - il risultato non può essere valutato per il singolo individuo
- allocazione casuale: tra i due gruppi **(A vs B)** si suppone **non ci sono differenze sistematiche** rispetto a variabili osservate o non osservate

Un esempio, gli RCT

- Anche in questo caso, l'ipotesi potrebbe essere

$$\begin{cases} \mathrm{H}_0 : \mu_A = \mu_B \\ \mathrm{H}_1 : \mu_A \neq \mu_B \end{cases}$$

- dove i termini μ_A e μ_B rappresentano le medie del collettivo sottoposto al farmaco A ed al farmaco B, rispettivamente
- Ovviamente, anche in questo caso il sistema può essere tradotto in

$$\begin{cases} \mathrm{H}_0 : \mu_A - \mu_B = 0 \\ \mathrm{H}_1 : \mu_A - \mu_B \neq 0 \end{cases}$$

- dove

$$\mu_A - \mu_B = \mu(X_A - X_B)$$

- ma la **statistica test** $T(\boldsymbol{X}_n) = T(\overline{\boldsymbol{X}}_A - \overline{\boldsymbol{X}}_B)$

Un esempio, gli studi osservazionali

- Si vuole sottoporre a **valutazione** una procedura chirurgica
- Si hanno a disposizione i risultati di un campione sottoposto **alla procedura** A o **alla procedura** B,
- Le due procedure non possono essere utilizzate sullo stesso individuo, e
- Il collettivo degli individui sottoposti ad **A** potrebbe differire dal collettivo di quelli sottoposti a **B**, per variabili osservate e/o non osservate
- In questo caso,
  - a ciascun soggetto $i = 1, \dots, n$ è associata un solo valore x_{iA} o x_{iB}
  - il risultato non può essere valutato per il singolo individuo
- allocazione non casuale: tra i due gruppi **(A vs B) possono esserci differenze sistematiche** rispetto a variabili osservate/non osservate

Un esempio, gli studi osservazionali

- Anche in questo caso, l'ipotesi potrebbe essere

$$\begin{cases} \mathrm{H}_0 : \mu_A = \mu_B \\ \mathrm{H}_1 : \mu_A \neq \mu_B \end{cases}$$

- dove i termini μ_A e μ_B rappresentano le medie del collettivo sottoposto alla procedura A ed alla procedura farmaco B, rispettivamente
- Ovviamente, anche in questo caso il sistema può essere tradotto in

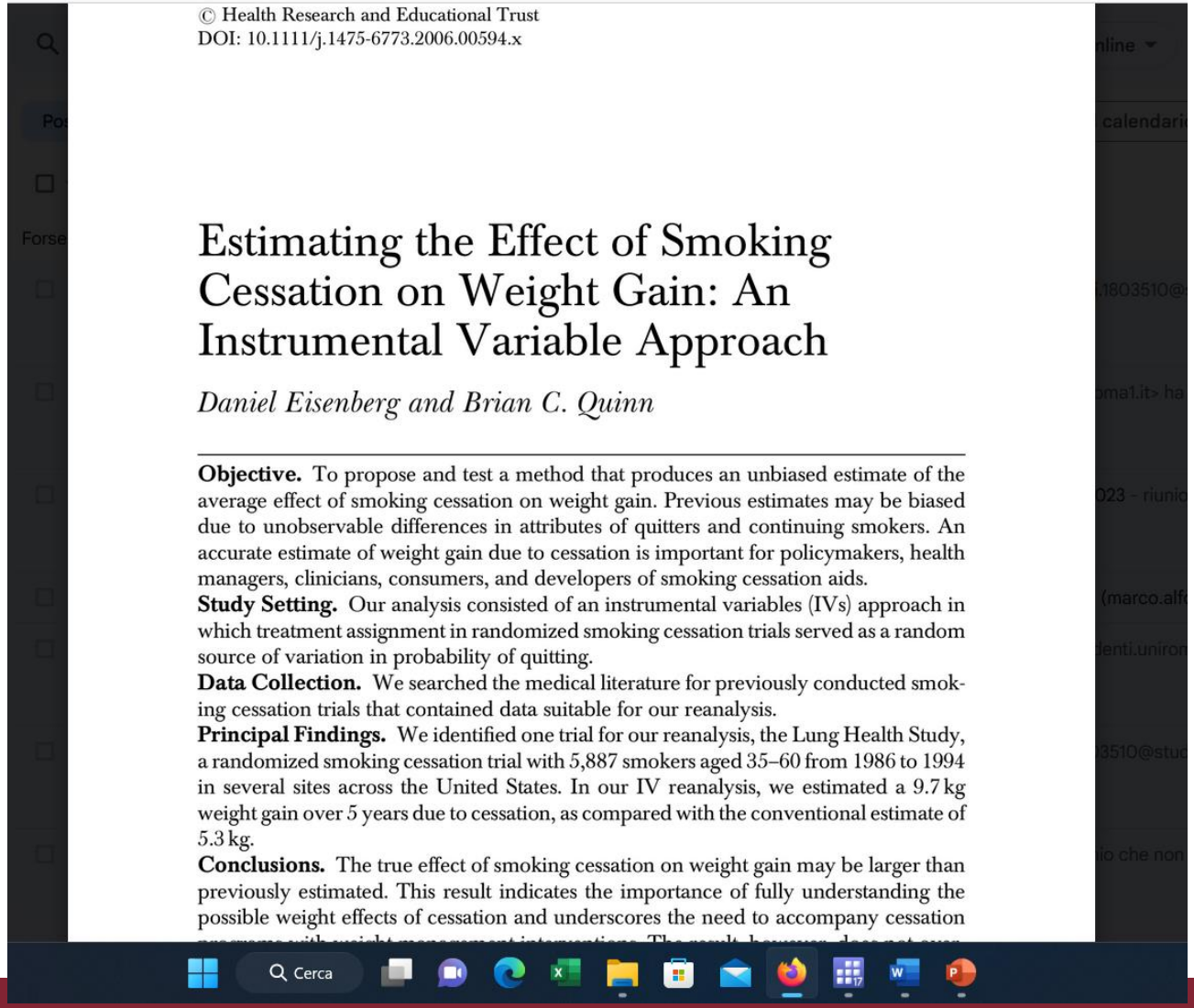
$$\begin{cases} \mathrm{H}_0 : \mu_A - \mu_B = 0 \\ \mathrm{H}_1 : \mu_A - \mu_B \neq 0 \end{cases}$$

- ma la **statistica test** $T(\boldsymbol{X}_n) = T(\overline{\boldsymbol{X}}_A - \overline{\boldsymbol{X}}_B)$ potrebbe non fornire l'informazione richiesta
- la differenza osservata potrebbe essere dovuta alla **differenze strutturali** tra collettivo A e collettivo B

Sui confronti…

- Il tema dei confronti è stato ampiamente dibattuto ed ha dato vita ad una branca specifica, quella di ***inferenza causale***
- Il punto centrale è che i confronti possono essere proposti sse **ceteris paribus**, altrimenti il risultato non può essere associato all'intervento (corso di formazione, farmaco, procedura chirurgica, etc.)
- Questo tema è molto attuale non solo in **biostatistica**, ma anche in **economia** (ad es. valutazione delle politiche)
- Gli studi **osservazionali**, così come i dati «**real-world**» pongono problemi spesso complessi, ma affascinanti
- Esempio: effetto dell'abitudine al fumo sul peso (i fumatori potrebbero essere **strutturalmente diversi** dai non fumatori)

Sui confronti...



DOI: 10.1111/j.1475-6773.2006.00594.x

Estimating the Effect of Smoking Cessation on Weight Gain: An Instrumental Variable Approach

Daniel Eisenberg and Brian C. Quinn

Objective. To propose and test a method that produces an unbiased estimate of the average effect of smoking cessation on weight gain. Previous estimates may be biased due to unobservable differences in attributes of quitters and continuing smokers. An accurate estimate of weight gain due to cessation is important for policymakers, health managers, clinicians, consumers, and developers of smoking cessation aids.

Study Setting. Our analysis consisted of an instrumental variables (IVs) approach in which treatment assignment in randomized smoking cessation trials served as a random source of variation in probability of quitting.

Data Collection. We searched the medical literature for previously conducted smoking cessation trials that contained data suitable for our reanalysis.

Principal Findings. We identified one trial for our reanalysis, the Lung Health Study, a randomized smoking cessation trial with 5,887 smokers aged 35–60 from 1986 to 1994 in several sites across the United States. In our IV reanalysis, we estimated a 9.7 kg weight gain over 5 years due to cessation, as compared with the conventional estimate of 5.3 kg.

Conclusions. The true effect of smoking cessation on weight gain may be larger than previously estimated. This result indicates the importance of fully understanding the possible weight effects of cessation and underscores the need to accompany cessation

Un esempio di studio appaiato

- Si vuole sottoporre a **valutazione** un corso di formazione, impartito ad un collettivo selezionato di n=30 studenti
- Gli studenti vengono valutati **prima** (A) e **dopo** (B) la loro partecipazione al corso di formazione

Id studente	Xa	Xb	(Xb-Xa)
1	22	25	3
2	27	29	2
3	27	28	1
4	22	25	3
5	26	28	2
6	20	28	8
7	23	25	2
8	27	29	2
9	19	27	8
10	21	24	3
11	20	26	6
12	23	25	2
13	23	24	1
14	…	…	…

Un esempio di studio appaiato

- Le singole righe riportano il risultato misurato, in trentesimi, sullo stesso individuo, prima (**A**) e dopo (**B**) il corso di formazione
- La differenza ha senso perché l'individuo rimane **costante** tra le due occasioni
- Risultato del test

Paired t-test

data: Xb and Xa

t = 7.8258, df = 29, p-value = 1.247e-08

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: 2.314458 3.952208

mean of the differences: 3.133333