

Statistica Aziendale

prof.ssa M. Grazia Pittau

Relazioni causali

Dipartimento di Scienze Statistiche - Sapienza Università di Roma

a.a. 2024-25

DIPARTIMENTO
DI SCIENZE STATISTICHE



SAPIENZA
UNIVERSITÀ DI ROMA

Differenza tra statistica ed econometria I

- *The Road Not Taken* (Frost, 1916)
Two roads diverged in a yellow wood,
And sorry I could not travel both
And be one traveler, long I stood
And looked down one as far as I could
To where it bent in the undergrowth;

.....

Two roads diverged in a wood, and I—
I took the one less traveled by,
And that has made all the difference.



Differenza tra statistica ed econometria II

■ **La strada non presa** (Frost, 1916)

Due strade divergevano in un bosco d'autunno e dispiaciuto di non poterle percorrere entrambe, essendo un solo viaggiatore, a lungo indugiai fissandone una, più lontano che potevo fin dove si perdeva tra i cespugli.

Poi presi l'altra, che era buona ugualmente e aveva forse l'aspetto migliore perché era erbosa e meno calpestata

sebbene il passaggio le avesse rese quasi uguali.

Ed entrambe quella mattina erano ricoperte di foglie che nessun passo aveva annerito

oh, mi riservai la prima per un altro giorno

anche se, sapendo che una strada conduce verso un'altra, dubitavo che sarei mai tornato indietro.

Differenza tra statistica ed econometria III

Lo racconterò con un sospiro
da qualche parte tra molti anni:
due strade divergevano in un bosco ed io -
io presi la meno battuta,
e questo ha fatto tutta la differenza.

- Usando la metafora dell'incrocio nella poesia di Robert Frost (1874–1963) possiamo iniziare a descrivere l'**effetto causato** da una scelta personale.
- La statistica si focalizza sulle procedure inferenziali: come conoscere le caratteristiche principali di una popolazione a partire da un campione casuale molto più piccolo rispetto all'intera popolazione.
- Gli statistici sono ovviamente consapevoli della casualità del campione e del fatto che campioni diversi possano dare risultati differenti.



Differenza tra statistica ed econometria IV



Differenza tra statistica ed econometria V

- L'econometria applicata si concentra essenzialmente sull'**inferenza causale**. Pur essendo consapevoli che esistono problemi importanti legati all'incertezza campionaria e all'inferenza statistica, lo scopo principale è differente.
- L'idea di base è quella di ragionare su due possibili *reliable statements* relativi a due possibili situazioni controfattuali. In altre parole l'obiettivo principale è quello di rispondere ad una domanda del tipo "Come cambierebbe la mia vita se io facessi una determinata scelta?"; "Come cambierebbe la società se attuassi questa politica?"; "Cosa sarebbe successo se avessi preso la strada più trafficata?"
Scelte che possono essere sia private che pubbliche.

Differenza tra statistica ed econometria VI

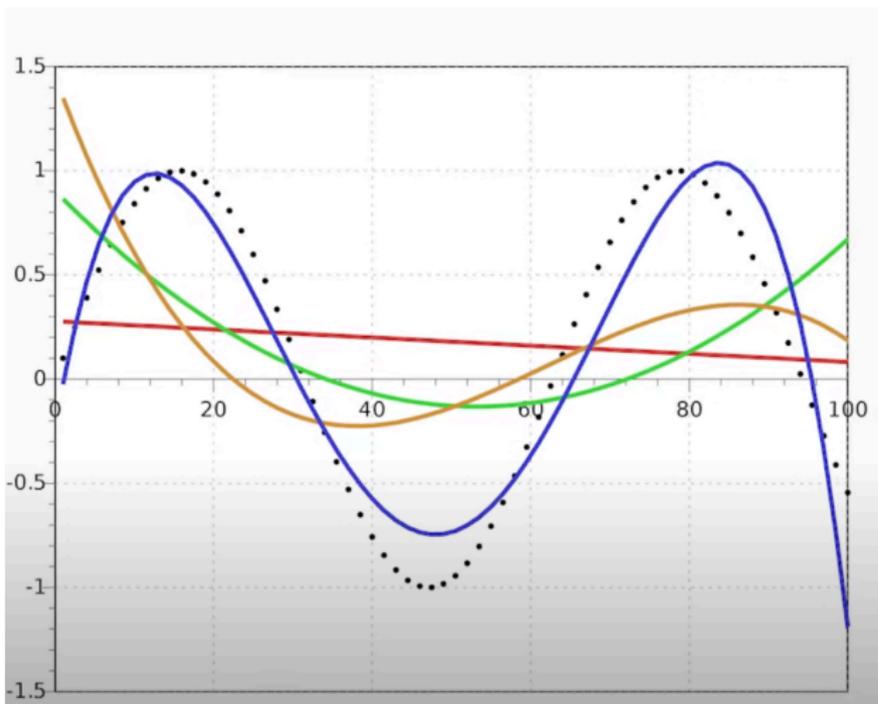
- Esiste un limite a quello che i dati statistici ci possono rivelare, pertanto è fondamentale affidarci anche a strumenti econometrici (i modelli appunto) che capire cosa succederebbe in un mondo controfattuale, una sorta di mondo parallelo.
- In questo corso ci occuperemo di scelte e conseguenze.



Differenza tra econometria e data science I

- La principale differenza è l'approccio al problema della previsione.
- I Data Scientists (DS) sono essenzialmente interessati a tutte quelle tecniche che assicurano il miglior fitting possibile.

Differenza tra econometria e data science II

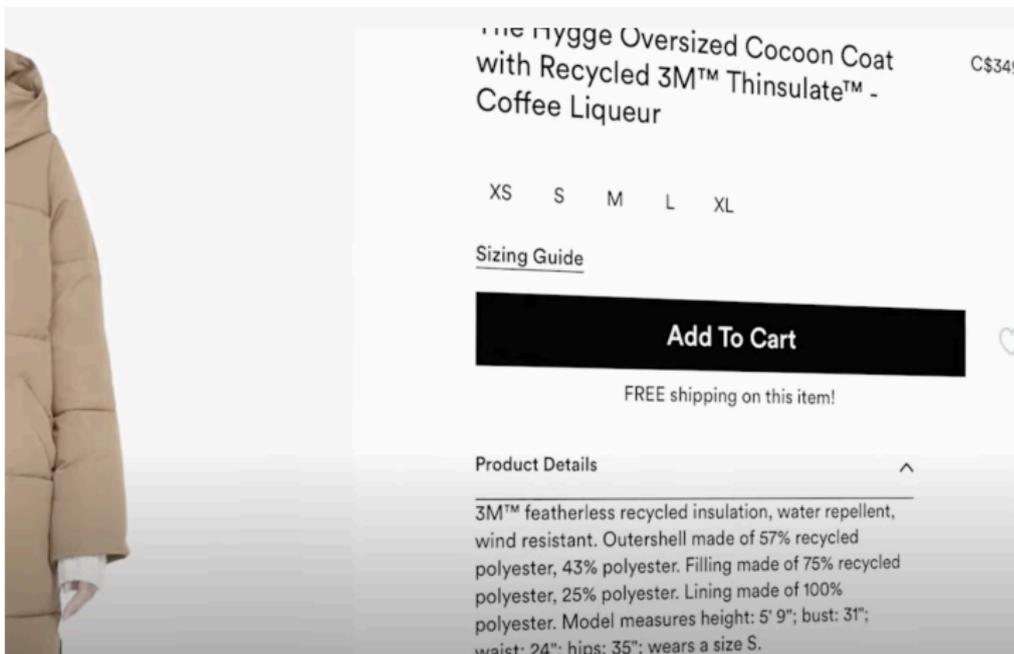


Differenza tra econometria e data science III

- Quindi ogni modello che si adatta bene ai dati è un “buon” modello. Se questo modello si basa sui dati del passato viene utilizzato dai DS per estrapolare comportamenti futuri.
- La maggiorparte delle analisi empiriche sono relative a problemi di Marketing e il modello che “fitta” meglio i dati del passato viene utilizzato per fare previsioni.
- Questa tipologia di modelli sono estremamente utili per il Marketing, in quanto ci permettono di capire chi comprerà cosa, chi farà delle scelte di acquisto ecc.



Differenza tra econometria e data science IV



The Hygge Oversized Cocoon Coat with Recycled 3M™ Thinsulate™ - Coffee Liqueur C\$349

XS S M L XL

[Sizing Guide](#)

Add To Cart

FREE shipping on this item!

Product Details ^

3M™ featherless recycled insulation, water repellent, wind resistant. Outershell made of 57% recycled polyester, 43% polyester. Filling made of 75% recycled polyester, 25% polyester. Lining made of 100% polyester. Model measures height: 5' 9"; bust: 31"; waist: 24"; hips: 35"; wears a size S.

Differenza tra econometria e data science V

- L'Econometria invece ha a che fare con problemi leggermente più complicati da un punto di vista concettuale.
- Gli econometrici sono particolarmente interessati a **relazioni causali**. Se manipolassimo qualcosa, sia essa l'assicurazione sanitaria, una politica monetaria, le tasse universitarie, come cambierebbe la società in risposta a questa manovra?
- Il passato non rappresenta più una guida così come lo è per i DS. Gli econometrici sono infatti consapevoli che le variazioni osservate sono influenzate non dalla storia passata ma da tantissime *confounding variables*, ovvero altre variabili/situazioni diverse da quelle che stiamo manipolando possono influenzare l'outcome.
- Pertanto osservare una relazione come emerge dai dati può essere fuorviante in quanto potrebbero esserci altri fattori non sempre facilmente controllabili.



Differenza tra econometria e data science VI

- Quindi noi lavoreremo su disegni di ricerca che vanno oltre il fitting di una curva. Non ci interessa il miglior fitting, piuttosto quali sono le conseguenze di una scelta effettuata. Per esempio quanto cambierà la nostra vita (in termini di successo, salario) in seguito alla scelta tra un'università molto prestigiosa e molto cara e un'università pubblica meno prestigiosa ma anche poco costosa?
- Un problema di questo tipo non è un problema di miglior fitting possibile, bensì un problema di relazione causale. Quali sono le conseguenze di una determinata scelta, sia essa pubblica o privata?



Econometria, big data e machine learning I

- I big data non influenzano molto l'econometria in genere e i modelli econometrici.
- La maggiorparte dei data-set che utilizzeremo per studiare le relazioni causali non sono particolarmente grandi.
- I grandi data set derivano essenzialmete da attività commerciali e quindi hanno un fortissimo impatto sulle ricerche di marketing.
- Per quanto molto popolari tra i giovani studiosi, importantissimi per alcune discipline, le relazioni causali non sono influenzate dai big data e dal machine learning.
- L'econometria esisteva prima dei big data, anche se molto teorica. Si pensi al teorema di Gauss Markov per esempio.
- In un approccio di analisi empirica come in questo corso, si parte da una domanda a cui si vuole dare una risposta.



Econometria, big data e machine learning II

- Molte compagnie importanti (google, Amazon, Trip Advisor...) sono interessati a capire cosa succederebbe se implementassero una data strategia. E' un problema di relazione causale. I DS sono bravissimi nel machine learning ma per rispondere ad una domanda di questo tipo abbiamo necessariamente bisogno di un modello econometrico.

Scelte e conseguenze I

- Consideriamo un problema che vi riguarda da vicino: le conseguenze sul vostro futuro in seguito alla scelta dell'università. Supponiamo che dobbiate decidere tra una prestigiosa e super costosa università privata e una buona ma non costosa università pubblica.
- Per poter decidere in modo consapevole dovremmo essere in grado di valutare l'*outcome* (salari, prestigio, successo) in entrambe le situazioni.



Scelte e conseguenze II



- Tuttavia nel momento in cui stiamo prendendo una decisione non possiamo vedere la fine della strada che non abbiamo percorso.

Scelte e conseguenze III

- La strada che non abbiamo percorso si chiama **controfattuale** e vorremmo provare a capire “cosa sarebbe successo se avessimo scelto diversamente e avessimo preso l'altra strada”.
- La strada controfattuale porta ad un *outcome* potenziale che non si realizza dal momento che abbiamo fatto una scelta diversa. Cosa sarebbe successo alla mia carriera se avessi fatto il PhD ad Harvard piuttosto che alla Sapienza?



Scelte e conseguenze IV



- Allo stesso tempo solo confrontando i due *outcome* possiamo scegliere in modo consapevole e trovare la strada migliore.

Scelte e conseguenze V

- Ma la fine della strada controfattuale è nascosta e non possiamo conoscere l'*outcome*. Come possiamo fare? Come posso immaginare il controfattuale?



Cloning Maschine I

- Immaginiamo un nostro **clone**, identico in tutto: stessi geni, stesso background, stessa identica esperienza fino al momento in cui le strade per l'università si dividono.
- Io scelgo una strada e il mio clone quella opposta.



Cloning Maschine II

- La presenza di un clone rende le due strade particolarmente “rivelatrici”, informative.
- L’idea di un **confronto bilanciato** è alla base dei modelli econometrici ed è stata battezzata **ceteris paribus, other things equal, parità di condizioni**.
- In presenza di un clone che sceglie la strada opposta possiamo parlare senza problemi di **causa-effetto**. Gli effetti sul nostro futuro causati dalla scelta dell’università.
- In questo mondo fantastico di cloni la sola differenza tra me e il clone è che il clone decide di prendere l’altra strada **all other things equal**.
- Le conseguenze (successo, salario) sono **causate** dalla scelta individuale di andare in una università privata o pubblica. Parliamo di **causa-effetto**, di **relazione causale**.



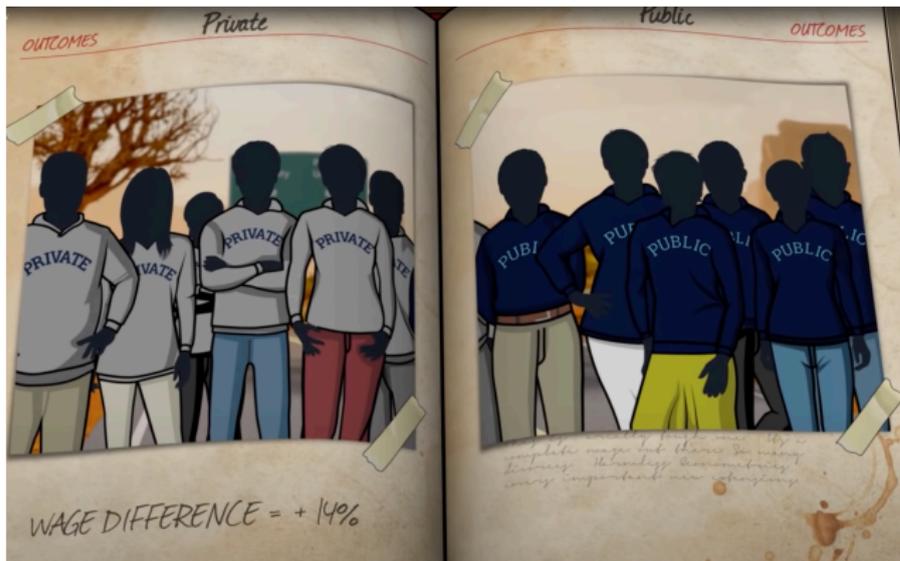
Cloning Maschine III

- Tuttavia non è ovviamente possibile (oltre a non essere etico e legale!) clonare tutte le persone. Quindi?
- In realtà non è necessario attivare una “cloning machine”, dal momento che i ricercatori hanno scoperto degli *strumenti* più intelligenti per scoprire cosa ci sarà alla fine del percorso.
- Gli strumenti degli econometrici devono in ogni caso tener conto delle parità di condizioni per poter essere utilizzato.

Ceteris Paribus I

- Consideriamo due gruppi di ex-studenti: un primo gruppo di studenti formato da coloro che scelgono l'università privata e un secondo gruppo formato da coloro che scelgono l'università pubblica.
- Confronto i loro salari e trovo che in media i salari degli alumni che hanno frequentato l'università privata sono più alti di circa il 14% dei salari degli alumni che hanno frequentato l'università pubblica.

Ceteris Paribus II



- Possiamo concludere che vale la pena spendere tanti soldi per l'università privata? L'università pubblica non è un buon investimento per il futuro?

Ceteris Paribus III

- Non possiamo parlare di relazione causale “Il 14% di differenza salariale è dovuta alla scelta universitaria”, a meno che non ci si trovi in una situazione di **ceteris paribus**.
- Quindi bisogna chiedersi: i due gruppi che stiamo confrontando sono davvero simili? I ragazzi hanno tutti uno stesso livello di conoscenza? Hanno tutti frequentato la stessa scuola? Hanno lo stesso background familiare? Le stesse ambizioni? Gli stessi skills?
- Sicuramente **no**.
- Quando nelle analisi empiriche ci si scorda di lavorare in condizioni di *ceteris paribus*, si incorre sicuramente in uno dei problemi più gravi che esistono: la distorsione da selezione, *selection bias*.

Selection bias I

- Per esempio chi proviene da una famiglia benestante o che ha studiato in una università importante come Harvard, Yale, MIT, verosimilmente sceglierà un'università privata ("Privilege bias", Forbes May 7, 2019).



Selection bias II

- Ogni qualvolta siamo interessati a dare una risposta a domande del tipo “La scelta dell’università ha implicazioni sul mio salario in futuro”? Oppure “Sottoscrivere un’assicurazione sanitaria implica un miglior stato di salute”? dobbiamo sempre pensare alla condizione di *ceteris paribus*. Altrimenti non ha nessun senso parlare di nesso causale.
- Ogni volta che la condizione di *ceteris paribus* non si verifica andiamo sempre incontro a situazioni di **selection bias**.
- Abbiamo detto che che gli alumni delle università private guadagnano circa il 14% in più rispetto a chi ha frequentato l’università privata. Possiamo quindi concludere che frequentare una università costosa e prestigiosa implica un salario più alto?
- Un’interpretazione di questo tipo è alquanto controversa. Perché?
- Riprendiamo i due gruppi di studenti che stanno scegliendo il loro futuro e *controlliamo* il loro voto di maturità (SAT) e il reddito disponibile delle famiglie di origine.

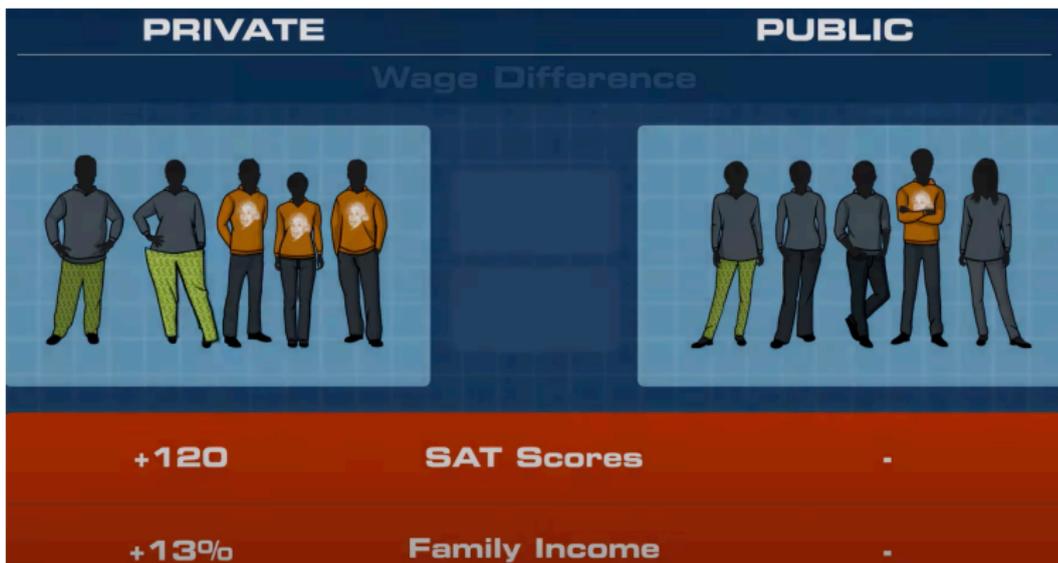


Selection bias III

- Hanno tutti lo stesso voto? Provengono tutti da famiglie con lo stesso reddito?
- Ci potremmo trovare in una situazione in cui gli studenti che scelgono la privata hanno mediamente voti più elevati, provengono da famiglie benestanti...



Selection bias IV



- Di conseguenza il vantaggio salariale del 14% potrebbe essere causato da **diseguaglianze** che esistevano **già prima** della scelta universitaria.

Selection bias V

- In assenza di parità di condizioni, e quindi in presenza di **selection bias**, rischiamo di interpretare risultati naïve come se fossero effetti causali.
 - I ragazzi che hanno punteggi mediamente più alti vengono accettati nelle università private con maggiore facilità.
 - I ragazzi che provengono da famiglie agiate scelgono le università private.
- Il confronto tra università privata e università pubblica in termini di causa-effetto ha senso solo quando i gruppi che stiamo confrontando sono identici (in media) in tutte le altre caratteristiche ad eccezione della scelta universitaria.
- Solo se tutte le altre caratteristiche sono uguali si può parlare di *ceteris paribus*.

Selection bias VI

- In questo esempio le università private sono più selettive e più costose rispetto a quelle pubbliche e quindi gli studenti che vanno alle università private non possono essere confrontati con coloro che frequentano università pubbliche.
- Questo è il modo in cui la selection bias ci **strega** facendoci credere quello che non è vero.
- Con **selection bias** si intendono tutte quelle situazioni in cui i confronti sono influenzati da differenze sistematiche relative a variabili diverse da quella di interesse (nel nostro esempio la scelta universitaria).
- La **selection bias** è un'insidia costante che si introduce con prepotenza in tutte le analisi empiriche *confondendo* i risultati e le conseguenti interpretazioni.
- Quali sono gli strumenti che esistono per combattere questa situazione senza necessariamente avere dei cloni virtuali?

Selection bias VII

- Come possiamo essere sicuri che i gruppi che stiamo confrontando siano comparabili e che le loro caratteristiche siano in media simili?

Esperimenti randomizzati I

- Primo strumento super potente che ha la statistica è dato dagli **esperimenti randomizzati** (Randomized Trials), ovvero dalla assegnazione del tutto casuale dei soggetti ai due gruppi.
- Come se l'ammissione ad Harvard o all'MIT avvenga in base al lancio di una moneta.
- L'assegnazione casuale (almeno con grandi numeri) ci assicura che le caratteristiche dei due gruppi siano in media simili (stessa % di donne, stessa % di gruppi etnici...).
- In una situazione di questo tipo non avremmo MAI problemi di selection bias.
- Gli esperimenti randomizzati sono tipici delle scienze sperimentali e molto utilizzati in medicina (esperimenti clinici randomizzati).

Esperimenti randomizzati II

- Esempio: il college americano militare West Point ha disegnato uno studio randomizzato per rispondere alla seguente domanda: “Computers e altri dispositivi elettronici dovrebbero essere ammessi in aula? Come la presenza di questi strumenti influenza il rendimento?”
- Il college West Point ha un insegnamento abbastanza tradizionale e non accetta la presenza di questi strumenti, contrariamente a molte altre università americane e non.
- Al fine dell'esperimento nel corso di Economia di base sono stati formati due gruppi di studenti: un gruppo che segue l'approccio tradizionale in cui è assolutamente vietato la presenza di strumenti elettronici e un altro gruppo in cui invece è ammessa la presenza di computer e altri strumenti.

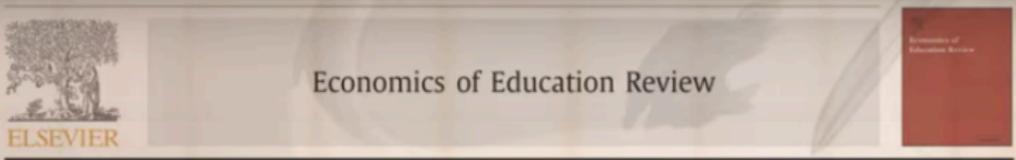
Esperimenti randomizzati III

- Il primo gruppo (insegnamento tradizionale) viene chiamato gruppo di **controllo**; il secondo gruppo (insegnamento con devices elettronici) gruppo **sperimentale**. Il trattamento *treatment effect* è in questo caso il permesso di utilizzare senza restrizioni di computer e strumenti elettronici in aula.

Esperimenti randomizzati IV



Esperimenti randomizzati V



Economics of Education Review

The impact of computer usage on academic performance: Evidence from a randomized trial at the United States Military Academy^{*}

Susan Payne Carter, Kyle Greenberg^{*}, Michael S. Walker

United States Military Academy, 607 Cullum Road, West Point, NY 10996, USA

<p>ARTICLE INFO</p> <hr/> <p><i>Article history:</i> Received 5 July 2016 Revised 8 December 2016 Accepted 9 December 2016 Available online 14 December 2016</p>	<p>ABSTRACT</p> <p>We present findings from a study that prohibited computer devices in randomly selected classrooms of an introductory economics course at the United States Military Academy. Average final exam scores among students assigned to classrooms that allowed computers were 0.18 standard deviations lower than exam scores of students in classrooms that prohibited computers. Through the use of two separate treatment arms, we uncover evidence that this negative effect occurs in classrooms where laptops and tablets are permitted without restriction and in classrooms where students are only permitted to use tablets that must remain flat on the desk.</p>
---	--

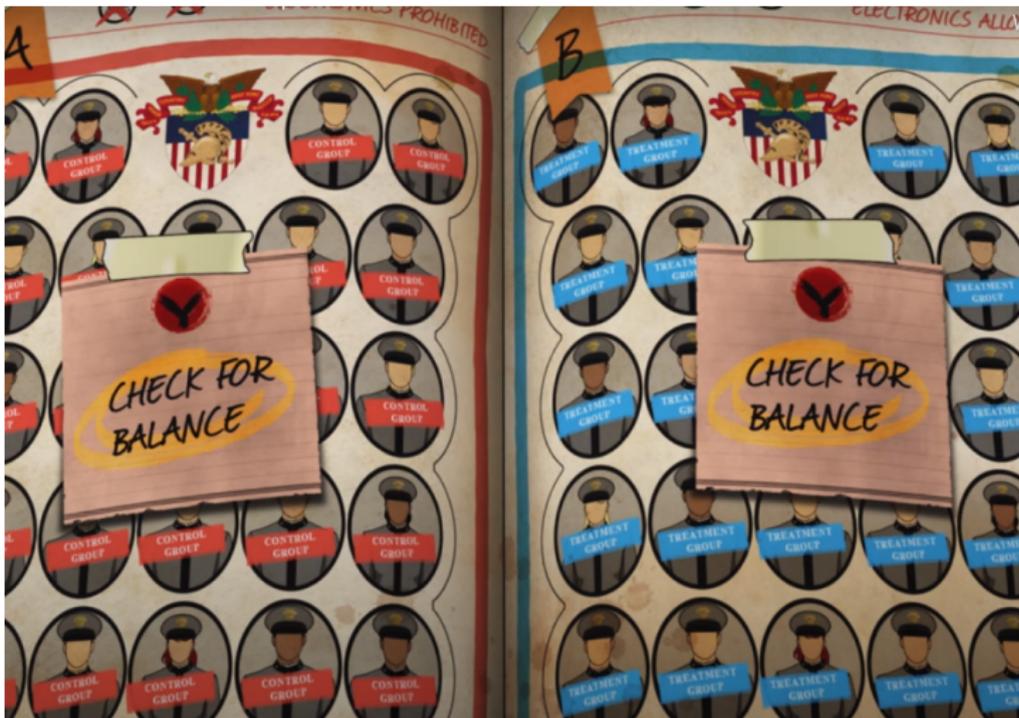
Published by Elsevier Ltd.

Esperimenti randomizzati VI

- Ogni domanda causale ha un chiaro *outcome*. Nel nostro esempio l'*outcome* finale è il punteggio all'esame di economia. Quindi ci si chiede: "Qual è l'effetto sul voto finale dell'esame di economia della presenza di strumenti elettronici" ?
- Gli studenti di economia di West Point sono stati assegnati al gruppo di controllo e al gruppo sperimentale in modo del tutto casuale. L'**assegnazione casuale** assicura le condizioni di *ceteris paribus* permettendo un ragionevole confronto.
- Grazie alle legge dei grandi numeri la randomizzazione assicura caratteristiche simili nei due gruppi.
- Quanto grandi devono essere i gruppi per essere sicuri delle parità di condizioni? Circa 250 unità possono essere sufficienti.
- Tuttavia non sempre i numeri sono così numerosi, quindi conviene sempre controllare se i due gruppi sono bilanciati, confrontando le caratteristiche degli studenti.



Esperimenti randomizzati VII



Esperimenti randomizzati VIII

	(1)	(2)
A. BASELINE CHARACTERISTICS		
Female	0.17	0.20
White	0.64	0.67
Black	0.11	0.10
Hispanic	0.13	0.13
Age	20.12 [1.06]	20.15 [1.00]
Prior military service	0.19	0.19
Division I athlete	0.29	0.40

Esperimenti randomizzati IX

- Il risultato di questo esperimento, assolutamente interpretabile in termini di casusa effetto ci dice che il punteggio medio dell'esame finale nel gruppo in cui gli strumenti elettronici erano permessi è risultato circa 0.28 unità di deviazione standard inferiore al punteggio complessivo avuto dal gruppo di controllo in cui gli strumenti elettronici non erano permessi in aula.
- Quindi la presenza in aula di strumenti elettronici implica distrazione e un punteggio mediamente più basso.
- Tuttavia l'assegnazione randomizzata tipica degli esperimenti non sempre è fattibile. Spesso è impraticabile, estremamente costosa e talvolta neanche etica (alcool, droga..).

Esperimenti randomizzati X



- Bibbia, libro di Daniele (Nabucodonosor, Re di Babilonia). “Test your servants for ten days. Give us nothing but vegetables to eat and water to drink. Then compare our appearance with that of the young men who eat the royal food, and treat your servants in accordance with what you see” (Daniel 1, 12–13).

Esperimenti randomizzati XI

- La ricerca sperimentale di Fisher (ML) e i suoi scritti hanno reso popolari gli esperimenti randomizzati (Ronald A. Fisher, *The Design of Experiments*, Oliver and Boyd, 1935).
- Quindi si pone la necessità di individuare nuove strategie, non estremamente costose che abbiano lo stesso potere, in termini di parità di condizioni, dell'assegnazione randomizzata.
- La possibile soluzione, nonché quella adottata dagli econometrici è **controllare per.....** Gli econometrici sono maniaci del controllo!

Controllare per... I

- L'obiettivo è quello di essere in grado di implementare delle strategie/metodologie statistiche che rendano due gruppi (che poi faranno scelte diverse per esempio sul piano universitario) il più simili tra loro in tutte le altre variabili ad eccezione della variabile *trattamento*.
- Quale possibile soluzione?
- Invece di confrontare i salari direttamente sui due gruppi di alumni: università privata *versus* università pubblica, andiamo ad individuare sottogruppi di alumni che hanno caratteristiche simili.
- Si individuano sottogruppi con medesime caratteristiche: stessa abilità, stesso background, stesso voto di laurea, famiglia con stesso reddito. E quindi il confronto viene fatto tra individui simili tra loro.
- Un grande passo avanti verso la parità di condizioni.
- Lo strumento più potente che abbiamo attualmente per lavorare in questo modo è la **regressione**.

Controllare per... II

- Il modello di regressione è lo strumento più adatto per confrontare i due gruppi di studenti **controllando per...**:
 - genere;
 - etnia;
 - background;
 - reddito familiare;
 - voto di laurea;
 - ...
- Ritornando al nostro esempio: una volta che abbiamo controllato per tutta una serie di caratteristiche, il gap salariale scende dal 14% al 9%.
- Ovvero solo il 9% delle differenze osservate è causato dalla scelta dell'università.
- Ma abbiamo davvero parità di condizioni? Abbiamo davvero controllato per tutto?

Controllare per... III

- Per esempio gli studenti che vanno ad Harvard, all' MIT sono in genere molto ambiziosi. Ma esiste un indice di ambizione, di "smartness"?
- Se non controlliamo per tutto, non abbiamo la condizione di *ceteris paribus*, non possiamo fare un confronto "mela contro mela".
- Ma non possiamo controllare per tutto....quindi?
- **Distorsione da variabili omesse!** E questo purtroppo succede sempre.
- Non è per nulla banale effettuare un confronto a parità di condizioni. La regressione dei nostri sogni ha molti più controlli di quelli che usualmente consideriamo (per esempio ambizione, motivazione, abilità...sono tutti fattori per cui non possiamo controllare).
- La regressione è uno strumento econometrico, non uno strumento magico. Tuttavia, nonostante dietro non ci sia la magia, i risultati sono spesso sorprendenti.



Controllare per... IV

- “Regression has the power to turn a clouded statistical night into a clear causal day” (Josh Angrist, MIT, Nobel Prize 2021).
- Prize motivation: “for their methodological contributions to the analysis of **causal relationships**”