Statistica Aziendale
prof.ssa M. Grazia Pittau
Tipologie di dati economici

Dipartimento di Scienze Statistiche - Sapienza Università di Roma

a.a. 2024-25

Dipartimento
di Scienze Statistiche
Sapienza
Università di Roma

Perchè studiare Statistica? I

1. "I can live with doubt and uncertainty... I have approximate answers and possible beliefs and different degrees of certainty about different things, and I'm not absolutely sure of anything..." (Richard Feynman)
2. https://thisisstatistics.org/hindsight-is-2023-for-former-statistics-and-data-science-students/
3. https://ischoolonline.berkeley.edu/data-science/study-applied-statistics/
4. In un mercato del lavoro sempre più competitivo, riuscire a capire come **gestire l'informazione disponibile** in termini quantitativi costituisce un enorme vantaggio;

Perchè studiare Statistica? II

5. Tra i migliori lavori elencati nel 2024:
   U.S. News & World Report Announces the 2024 Best Jobs:
   Of the six STEM positions noted in the top 10, nurse practitioner took the No. 1 spot, statistician landed at No. 2
   statistician (Uses statistical methods to collects and analyzes data and to help solve real-world problems in business, engineering, healthcare, or other fields):
   data scientist: (Combines information technology, statistical analysis and other disciplines to interpret trends from data)
6. si veda anche il sito:
   `http://blog.revolutionanalytics.com/2017/05/best-job-i2017-statistican.html`
   Al primo posto troviamo la professione di Statistico (e solo al quinto quello di Data Scientist).
   Il **miglior** lavoro nel 2017? **Statistician Tops List of Best Jobs in 2017**!

Perchè studiare Statistica? III

https://www.amstat.org/news-listing/2021/10/08/statistician-tops-list-of-best-jobs-in-2017.
"The high demand for data scientists and statisticians comes from a growing emphasis on collecting and evaluating massive quantities of data. The opportunities for professionals trained in these fields are tremendous, as the IT sector, healthcare, business–and any sector that collects consumer information can put these numbers to use.
Data science is a relatively new field, which promises to revolutionize industries from business to government, health care to academia. A growing number of universities offer data science degree programs."

7 "Data Scientist: The Sexiest Job of the 21st Century": https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century Harvard Business Review, October 2012.

Perchè studiare Statistica? IV

8. Recentemente: `https://thisisstatistics.org/statistician-2019-top-job-america/` "USA TODAY (April 2019) named "statistician" as the #5 top job in America for its "very good work environment, very low stress at work, and very good projected employment growth–making it one of very few careers to receive the highest marks available for all three categories."
9. For Today's Graduate, Just One Word: **Statistics** (New York Times, August 5, 2009): `https://cacm.acm.org/news/36754-for-todays-graduate-just-one-word-statistics/fulltext`


HBR.ORG
Harvard Business Review
OCTOBER 2012
46 The Big Idea
The True Measures Of Success
Michael J. Mauboussin
84 International Business
10 Rules for Managing Global Innovation
Keeley Wilson and Yves L. Doz
93 Leadership
What Ever Happened To Accountability?
Thomas E. Ricks
GETTING CONTROL OF
BIG DATA
How vast new streams of information are changing the art of management
PAGE 59

Perchè R! I

1 La richiesta di lavori con conoscenza di R e Sas

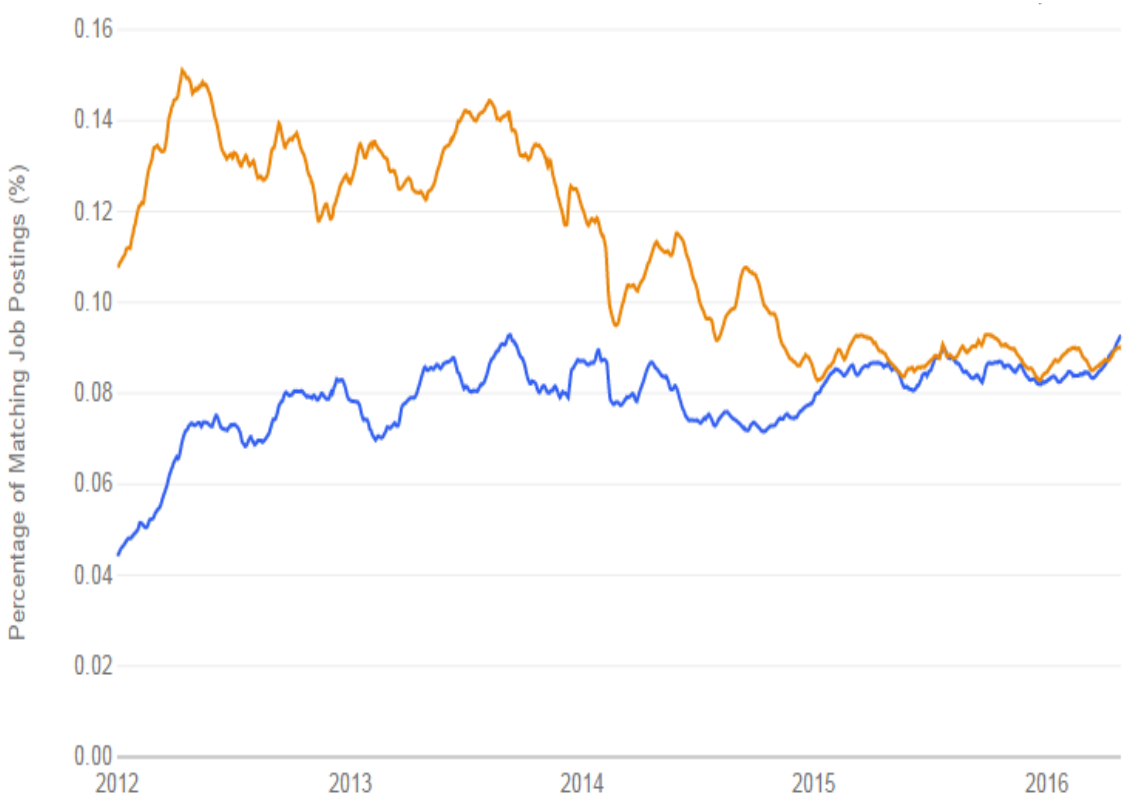


Figura: Trends in jobs requiring R modeling and SAS modeling: R in blue, SAS in yellow. (Statistic.com Jobs, 9/2/2016)

Perchè R! II

2. https://intellipaat.com/blog/sas-versus-r/
3. https://blog.revolutionanalytics.com/2019/09/devops-and-r.html
4. Data Analysts Captivated by R's Power (New York Times, Jan. 6, 2009): https://www.nytimes.com/2009/01/07/technology/business-computing/07program.html

Ci si può fidare della Statistica? Quando possiamo parlare di casua-effetto? I

1. Quando i dati mentono (Focus, Maggio 2016). Siamo in un'era che annega nei numeri. Che a volte si prestano a interpretazioni surreali (pag. 108):
   - Lo sapevate che la favola dei bambini portati dalle cicogne nacque in seguito a uno studio statistico? A fine Ottocento, nei Paesi Bassi, si scoprì che se in città arrivavano più cicogne nascevano anche più bambini, e fu così che si diffuse la ben nota leggenda... In realtà, le case dove c'era un nuovo nato venivano riscaldate di più e il calore attirava le cicogne. Il rapporto - in questo caso specifico - c' era, ma era indiretto (pag. 108).

Ci si può fidare della Statistica? Quando possiamo parlare di casua-effetto? II

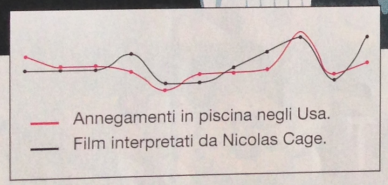
- Nel 1958, William Phillips, docente di economia a Londra, pubblicò un articolo su inflazione e disoccupazione; i dati parlavano da soli: ad alti livelli di inflazione corrispondevano bassi livelli di disoccupazione, e viceversa. Questo collegamento diventò famoso come "curva di Phillips" e orientò le politiche economiche occidentali almeno fino agli anni '70... quando esplose la stagflazione (cioè elevata inflazione con forte disoccupazione). Insomma, la relazione "scoperta" da Phillips non esisteva. I due fenomeni si influenzano a vicenda, questo è vero, ma di certo non basta che uno aumenti per far calare l'altro. In altre parole, era stata individuata quella che gli specialisti chiamano una "correlazione spuria" (pp 108, 109).
- Uno studente americano della Harvard Law School, di nome Tyler Vigen, ha dimostrato nel suo blog quanto sia facile, partendo da masse di dati abbastanza grandi, trovare correlazioni spurie di ogni tipo.
  `https://tylervigen.com/spurious-correlations`

Ci si può fidare della Statistica? Quando possiamo parlare di casua-effetto? III

- Secondo Andrew Gelman, uno statistico della Columbia University di New York, al pubblico piacciono i risultati corredati da rigorosi dati statistici, com'è accaduto per la "scoperta" che nel periodo dell'ovulazione le donne preferiscono vestirsi di rosso. Peccato che anche questa sia una correlazione spuria. "E come se la gente prendesse i dati statistici come una scusa per spegnere il cervello", ha detto Gelman al settimanale britannico New Scientist. `http://www.slate.com/articles/health_and_science/science/2013/07/statistics_and_psychology_multiple_comparisons_give_spurious_results.html`

2. Cosa vuole dire? E' fondamentale riconoscere casi correlazione spuria...

Esempi di correlazione spuria: Focus, Maggio 2016



TERRORIZZATI DAL CINEMA? Il numero di persone affogate in piscina negli Usa, fantasiosamente correlato al numero di film di Nicolas Cage, dal 1999 al 2009.

tutti rigorosamente corretti e controllabili, anche se è ovvio che, per esempio, le importazioni di greggio dalla Norvegia negli Stati Uniti non hanno rapporti con le collisioni fra treni e automobili (*v. disegno nella pagina precedente*).
Secondo Andrew Gelman, uno statistico della Columbia University di New York, al pubblico piacciono i risultati corredati da rigorosi dati statistici, com'è accaduto per la "scoperta" che nel periodo dell'ovulazione le donne preferiscono vestirsi di rosso. Peccato che anche questa sia una correlazione spuria. «È come se la gente prendesse i dati statistici come una scusa per spegnere il cervello», ha detto

esponenziale dei ricercatori, tutti che sgomitano per vedere pubblicate le proprie nuovissime ricerche. «Questo significa che nessuno controlla mai i risultati delle ricerche altrui, nemmeno quando ci va di mezzo la salute», pensò qualche anno fa John Ioannidis, oggi docente a Stanford e ormai considerato un esperto mondiale sulla credibilità, o meno, delle ricerche in campo medico.

PANIERE SBILANCIATO. Nel 2005, Ioannidis scelse le 49 scoperte mediche più importanti dei 13 anni precedenti e ne replicò 34. Risultato? Ben 14 degli articoli esaminati (il 41%) erano arrivati a

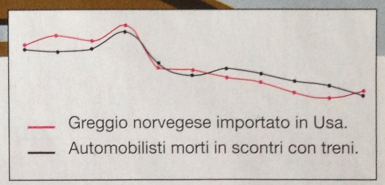
più una polemica t
sulla situazione e
racconta Gian Ma
dell'Eurispes. «L'Is
ca ufficiale, dava u
zione, mentre la no
pratica andava in u
Non erano le voci
sbagliate, ma il lo
le spese per la cas
9% pur comprend
tuo), le spese cond
ristrutturazioni...
flazione sembrava
aveva iniziato a g
fine riconobbe l'er
ma di pesatura. «

Esempi di correlazione spuria: Focus, Maggio 2016

serio.
te che
cico-
stati-
Bassi,
più ci-
ni, e fu
enda...
o nato
ore at-
questo
etto.
nte di
artico-

parlavano da soli: ad alti livelli di inflazione corrispondevano bassi livelli di disoccupazione, e viceversa. Questo collegamento diventò famoso come "curva di Phillips" e orientò le politiche economiche occidentali almeno fino agli Anni '70... quando esplose la stagflazione (cioè elevata inflazione con forte disoccupazione). Insomma, la relazione "scoperta" da Phillips non esisteva. I due fenomeni si influenzano a vicenda, questo è vero, ma di certo non basta che uno aumenti per far calare l'altro. In altre parole, era

DURE A MORIRE. La lezione però non è bastata, e molti studiosi continuano a fare scoperte che a un secondo esame si rivelano coincidenze dello stesso tipo. Di recente, per esempio, potreste aver letto che i genitori di bell'aspetto hanno una probabilità più alta di avere figlie femmine. Non credeteci: è una correlazione spuria. Uno studente americano della Harvard Law School, di nome Tyler Vigen, ha dimostrato nel suo blog quanto sia facile, partendo da masse di dati abbastanza grandi, trovare correlazioni spurie di ogni tipo. I suoi grafici (alcuni dei



COINCIDENZE "PERICOLOSE". Alcuni fenomeni statistici a prima vista sembrano collegati. Qui, le vittime di scontri con treni calano nel tempo come l'importazione negli Usa di greggio dalla Norvegia.

DIPARTIMENTO
DI SCIENZE STATISTICHE
SAPIENZA
UNIVERSITÀ DI ROMA

Cosa si intende per Statistica I

- **Statistica: L'arte e la scienza di imparare dai dati**;
- Insieme di metodologie finalizzare alla raccolta e all'analisi dei dati;
- In altri termini, la statistica è l'insieme dei metodi per:
  1. **progettare**: pianificare come devono essere raccolti i dati necessari per le ricerche;
  2. **descrivere**: sintetizzare i dati in formati semplici (attraverso metodi di sintesi e **grafici**) e facilmente leggibili limitando al massimo la loro distorsione e la perdita di informazioni;
  3. **inferire**: formulare previsioni basate sui dati raccolti. I dati sono dati di **tipo campionario**.
- Esempio: Le performance degli studenti dipendono dalla quantità di denaro speso per studente, dalla dimensione delle classi oppure dal salario degli insegnanti?

Tipologie di dati economici I

- Dati **dati macro-economici** (comportamenti collettivi o *aggregati*): si riferiscono ad unità a cui non si riconoscono capacità decisionali (tipicamente riguardano un sistema economico nel suo complesso: paesi, regioni ecc.).
- **micro-economici** (comportamenti individuali): si riferiscono ad un campione (o eventualmente ad un universo) di "individui" o unità assimilabili ad individui in quanto a capacità di prendere decisioni (es. famiglie, imprese). Solo in casi particolari abbiamo a disposizione l'universo in dati micro (registri di popolazione, di imprese, di studenti...).
- I dati macroeconomici si basano spesso su un processo di aggregazione. Questo processo di aggregazione dei dati (non del tutto assente nei microdati–si pensi ai dati sul consumo o sul reddito familiare), crea un sorta di *livellamento nei dati*. A seguito di questo effetto le variabili a livello aggregato presentano molto spesso un andamento più "liscio".

Tipologie di dati economici II

- Un classico esempio di dati macro-economici riguarda il PIL pro capite misurato per i paesi dell'UE in un determinato anno.

L'Italia nel contesto Europeo

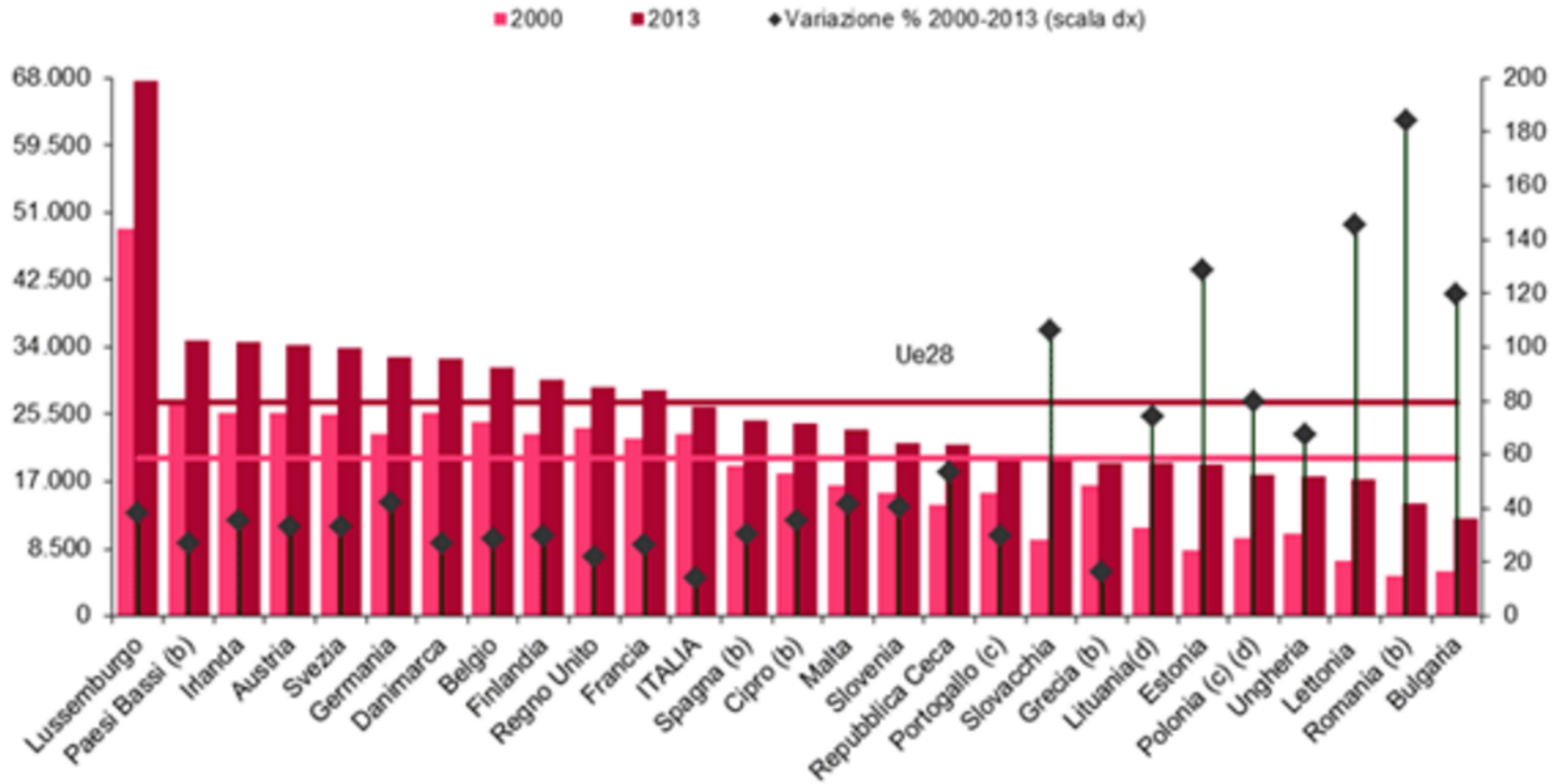


Figura: Pil pro capite nei paesi UE Anni 2000 e 2013 (in parità di potere d'acquisto e variazioni percentuali), Istat, 2014.

Dati micro-economici I

- Noi ci concentreremo sui dati a livello micro: i cosidetti **microdati**.
- I microdati (sia che siano dati relativi a singoli individui che dati a livello di singola azienda) provengono o da indagini campionarie o sono dati censuari. Questi dati sono spesso chiamati **dati osservazionali**, per distinguerli dai **dati sperimentali**.

Ecological Fallacy I

- **ecological fallacy**: distorsione (bias) che si può verificare quando le conclusioni circa un'eventuale associazione a livello individuale vengono dedotte da un'associazione che esiste e viene misurata a livello aggregato (in genere dai gruppi a cui gli individui appartengono).
- Gelman et al. hanno dimostrato che la correlazione tra reddito e propensione al voto per il *Grand Old Party*, negativa a livello aggregato, risulta positiva a livello individuale (*Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do*, 2010).
- Questa discordanza (visibile solamente se vengono analizzati anche dati micro) è possibile in quanto nulla garantisce che dai dati aggregati si possano fare conclusioni relative ai singoli individui o a sotto-insiemi di quegli stessi aggregati (per esempio alle contee). L'assunzione che si possano fare tali conclusioni è detta in statistica **fallacia ecologica**.

Ecological Fallacy II

- Presenza di **Contextual effects**. Effetti che esistono se la relazione tra una variabile risposta e un predittore è funzione del livello medio di una variabile indipendente nel gruppo. In altri termini, se c'è un'interazione micro-macro tra le X (con i coefficienti di regressione dei singoli gruppi che sono associati ai valori macro di una qualche variabile indipendente).

Dati osservazionali I

- Le **fonti principali** da cui derivano i dati osservazionali sono le indagini sulle famiglie, sulle imprese e le fonti amministrative:
  Istat: rilevazione del sistema dei conti delle imprese
  `http://indata.istat.it/sci/`;
  Banca Italia: Indagine sulle imprese industriali e dei servizi e Indagine sui bilanci delle famiglie italiane
  `https://www.bancaditalia.it/pubblicazioni/indagine-famiglie/index.html`;
  Istat: Archivio Asia
  `http://www.istat.it/it/archivio/archivio+asia`)
- I dati di Marketing possono possono essere raccolti nei punti vendita di supermercati e/o centri commerciali, tra i nuovi e i vecchi clienti.
- Tra le nuove fonti di dati ricordiamo Internet e le "interviste" on line.
- Il termine **dati osservazionali** si riferiscono in genere a dati provenienti da una **procedura di campionamento** da una popolazione finita.

Dati osservazionali II

- Nello schema campionario più semplice, il **campionamento casuale semplice**, la probabilità di estrarre l'unità i da una popolazione di ampiezza N risulta pari a $1/N$ per tutte le unità i.
- Esistono tuttavia schemi campionari più complessi (campionamento stratificato a più stadi), a cui fanno riferimento tutte le principali indagini campionarie in cui le **unità possono avere una diversa probabilità di essere estratte**.
- La maggiorparte delle indagini importanti fornisco i cosiddetti **pesi campionari**, che sono pari all'inverso della probabilità di inclusione delle unità nel campione. Questi pesi devono essere usati per ottenere una stima non distorta dei parametri della popolazione.
- Esistono problemi di *sample selection*, di *mancate risposte*, di *errori di misura*, di *attrition*.
- Esistono **diverse tipologie** di dati osservazionali:

Dati osservazionali III

1. **dati cross-section** o **sezionali**: sono ottenuti osservando uno stesso insieme di variabili $\mathbf{w}$ di caratteristiche su un campione S_t in diversi tempi o intervalli temporali. Sebbene sia impossibile intervistare un insieme di famiglie tutte allo stesso istante di tempo, i dati cross-section rappresentano una immagine "istantanea" delle caratteristiche di ciascun elemento di un sottoinsieme della popolazione, in base alle quali si fa inferenza sull'intera popolazione.
   - Se la **popolazione è stazionaria** (ovvero tutti parametri della popolazione sono costanti nel tempo $\theta_t = \theta$), allora l'inferenza sul parametro θ_t usando un campione S_t risulta valida anche quando $t' \neq t$.
   - Se per esempio le decisioni passate possono influenzare un certo comportamento di consumo, avremmo una struttura di dipendenza, che non può essere modellata se non si hanno dati di consumo nel tempo. Questo costituisce il limite maggiore dei dati sezionali.

Dati osservazionali IV

2. **dati cross-section ripetuti**: sono ottenuti da una **serie di campioni indipendenti** S_t, con $t = 1, \cdots, T$. (Esempio GSS, ESS).

   - Dal momento che il disegno campionario non prevede la ripetizione delle unità nel campione, non è presente alcuna informazione circa la dinamica di dipendenza.
   - Se la popolazione è considerata stazionaria, i dati sezionali ripetuti sono ottenuti attraverso un processo di campionamento che in qualche modo può essere paragonato ad una campionamento con ripetizione da una popolazione costante.
   - Se invece si pensa che la popolazione di riferimento sia non stazionaria, le cross-section ripetute sono in qualche modo legate, in modo da riflettere i cambiamenti della popolazione nel tempo.

Dati osservazionali V

3. **Dati panel o longitudinali**: sono ottenuti selezionando inizialmente un campione S e su questo stesso campione vengono misurate le stesse caratteristiche per una sequenza di periodi pari $t = 1, \cdots, T$. (es.: PSID, EU-SILC, Banca d'Italia).
   - I dati vengono raccolti intervistando gli individui e collezionando sia dati attuali che dati riferiti al passato delle medesime unità statistiche, oppure seguendo gli individui nel tempo una volta che sono stati inseriti nell'indagine.
   - In questo caso abbiamo una sequenza di vettori di dati $\{\mathbf{w}_1, \cdots, \mathbf{w}_t\}$ che sono usati per fare inferenza sia sul comportamento della popolazione o anche su un particolare campione di individui.
4. Numerosi sono i **limiti che hanno questi dati**. I dati cross-section non forniscono dati tali da poter modellare eventuali interdipendenze temporali.
5. Al contrario, i dati longitudinali, soprattutto se fanno riferimento ad un periodo di tempo sufficientemente lungo, permettono di modellare relazioni sia di tipo statico che dinamico.

Dati osservazionali VI

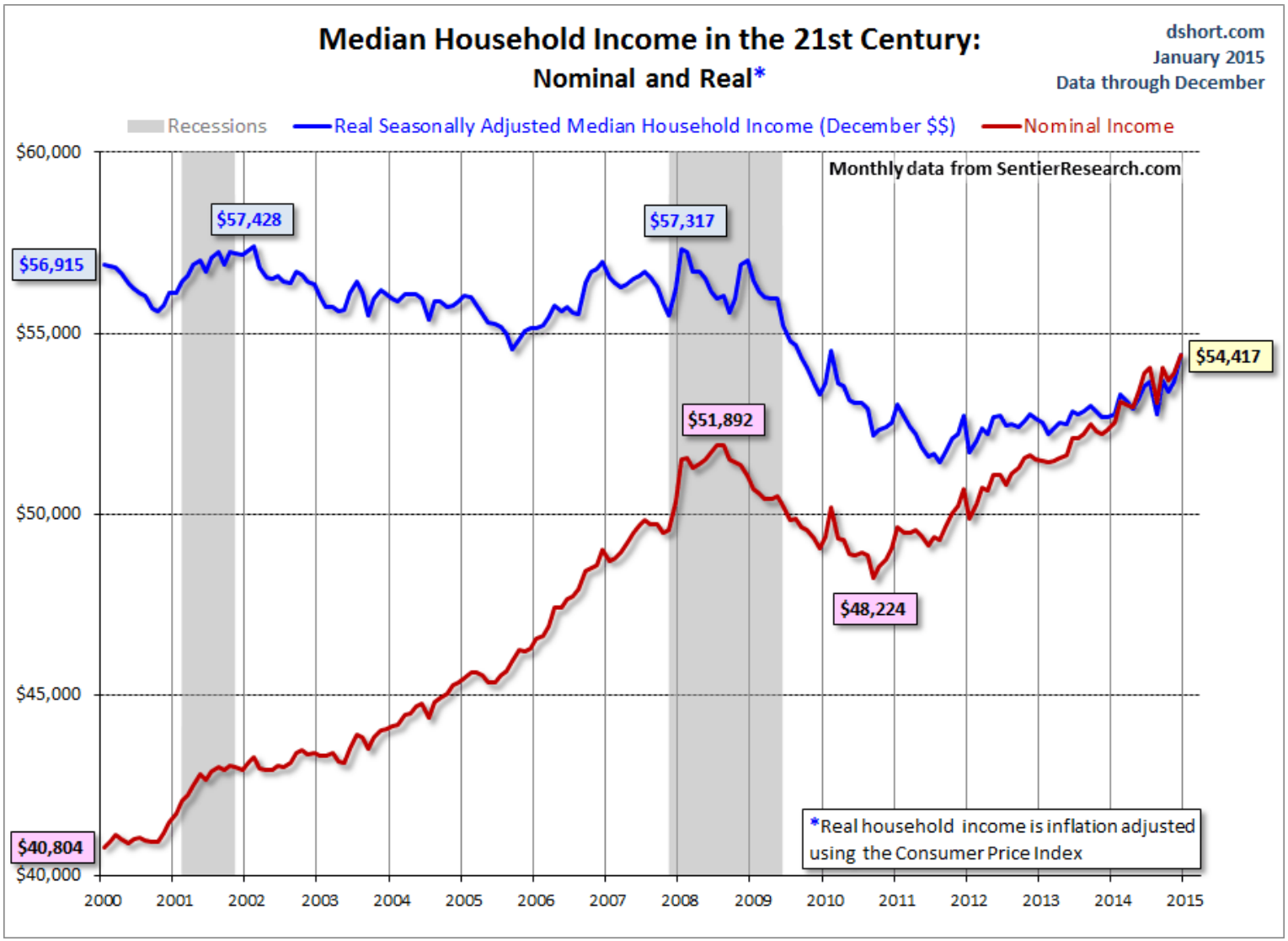
6. I dati longitudinali tuttavia non sono esenti da problemi. Il primo problema riguarda la rappresentatività del panel. Per analizzare comportamenti dinamici è particolarmente importante tenere gli individui (famiglie o imprese) nel panel il più a lungo possibile.
7. In realtà i dati panel soffrono del cosiddetto problema di *sample attriction*, semplicemente dovuto alla stanchezza di individui, famiglie e imprese di stare nel panel per periodi lunghi.
8. Questo problema comporta due tipi di complicazioni:
    1. Il panel diventa non bilanciato, ovvero si "perdono" unità statistiche che escono fuori dal panel;
    2. potrebbe essere che le unità statistiche rimaste nel panel siano "atipiche" e di conseguenza il campione non è più rappresentativo della popolazione.
9. **Serie storiche o temporali**: rilevazione del fenomeno sottostante che stiamo misurando (**variabile**) in specifici momenti nel tempo (ad esempio annualmente).

- Solitamente dati macro-economici.

Dati osservazionali VII

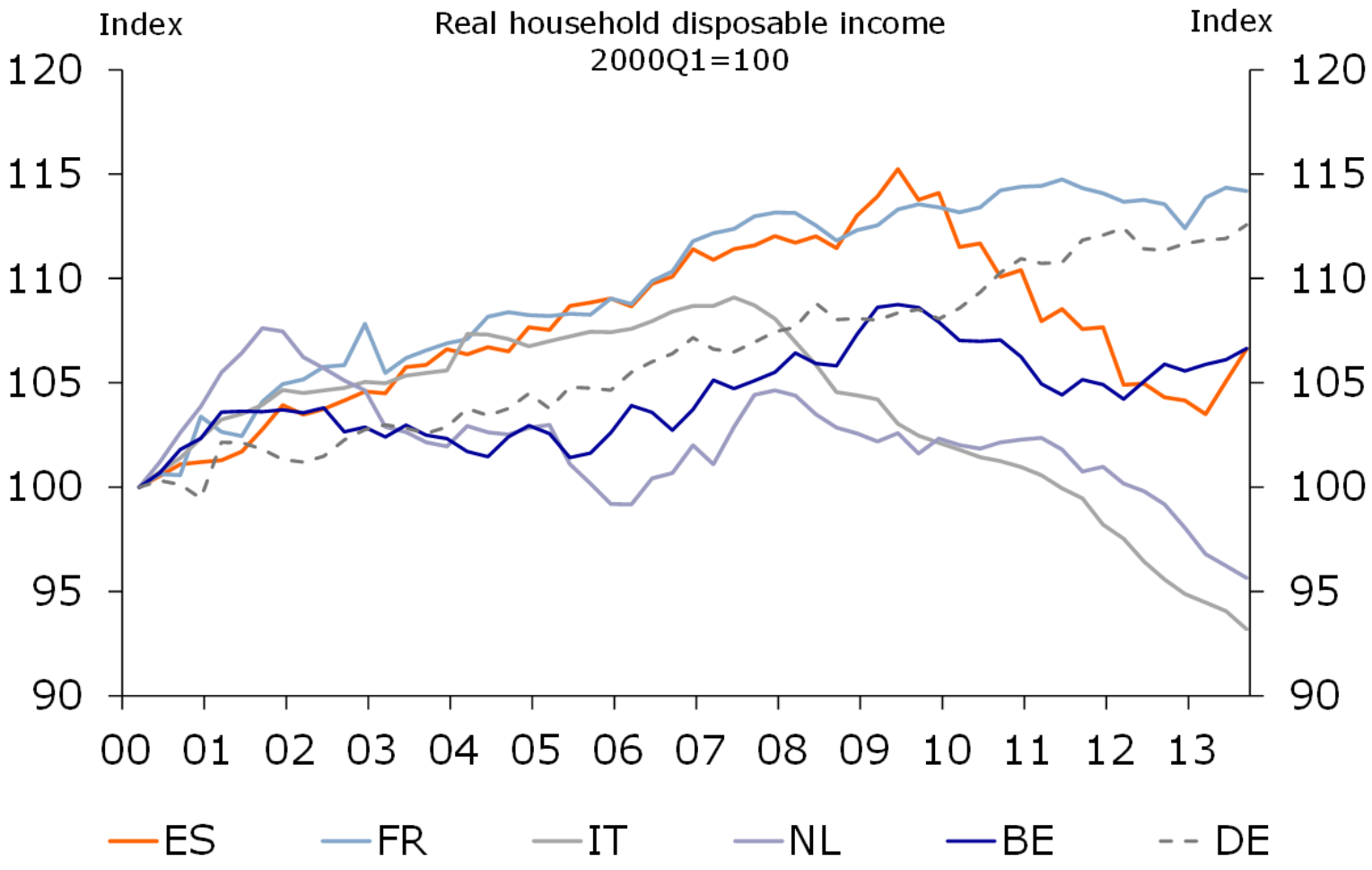
- Le serie storiche possono essere osservate a diverse **frequenze**. Le frequenze generalmente utilizzate sono: **annuale** (cioè la variabile viene osservata ogni anno), **trimestrale**, **mensile**, **settimanale**, **giornaliera**.
- Notazione: $Y_t, t = 1, 2, \ldots T$ (tempo criterio ordinatore dei dati).
- nello studio di una serie storica ha un ruolo fondamentale l'ordinamento temporale, nel senso che i T valori osservati sono ordinati rispetto al tempo t e perciò non sono scambiabili;
- lo scambio delle osservazioni distrugge le informazioni sulla evoluzione del fenomeno nel tempo;
- osservazioni dipendenti: i dati presentano "regolarità" o persistenze legate alla posizione dell'osservazione nella sequenza;
- per analisi di serie storiche utilizzo di metodi statistici diversi da quelli utilizzati con dati di tipo cross-section.

Reddito mediano delle famiglie negli Stati Uniti–andamento reale e nominale



Dipartimento di Scienze Statistiche
Sapienza Università di Roma

Reddito disponibile in alcuni paesi europei–Numeri indice 2000Q1=100



Dati sperimentali I

- I **dati sperimentali** si differenziano principalmente dai **dati osservazionali** in quanto i dati sperimentali fanno riferimento ad un esperimento che, in linea generale, può essere monitorato e tenuto rigorosamente sotto controllo.
- Negli esperimenti risulta quindi possibile far variare una variabile casuale di interesse tenendo **sotto controllo** le altre variabili indipendenti (predittori o covariate) e valutare l'effetto di queste variazioni su un variabile dipendente o *outcome*.
- Nelle ricerche sperimentali il principale obiettivo è quello di **manipolare** una o più variabile indipendenti ed esaminare gli effetti di questi cambiamenti sulla variabile indipendente. Dal momento che risulta possibile manipolare la(le) variabile(i) indipendente, la ricerca sperimentale ha il grande vantaggio di poter identificare una **relazione di causa-effetto** tra variabili.

Dati sperimentali II

- Esempio su studi ramdomizzati: consideriamo 100 studenti che hanno fatto l'esame scritto di Matematica 2. La variabile dipendente è il voto finale, mentre le variabili dipendenti sono l'eventuale possibilità di una revisione (misurata in ore) e il voto preso a Matematica 1. Potremmo considerare un disegno sperimentale manipolando il tempo di revisione. Lo sperimentatore divide **casualmente** i 100 studenti in due gruppi uguali da 50. Al primo gruppo si richiede la consegna del compito senza revisione, agli studenti del secondo gruppo invece si da la possibilità di una revisione di 50 minuti. A questo punto si confrontano i voti dei due gruppi.

Dati sperimentali III

- Al contrario nelle **ricerche non sperimentali il ricercatore non può manipolare le variabili indipendenti**. Non solo talvolta non è proprio possibile ma sebbene possibile potrebbe anche non essere eticamente corretto. Per esempio un ricercatore potrebbe essere interessato a valutare gli effetti dell'uso di droghe (variabili indipendenti) su certi tipi di comportamento (variabile dipendente). Ma per quanto possibile non è legale chiedere ad un individuo di assumere droghe per valutare gli effetti che queste hanno su alcuni comportamenti. Non è quindi possibile valutare propriamente la relazione causa-effetto ma è comunque possibile valutare e stimare l'associazione o la relazione che esiste tra queste variabili.

Dati sperimentali IV

- Quindi i dati osservazionali non provengono da "esperimenti", non possono essere assolutamente né controllati, né manipolati, e si lascia aperta la possibilità che esistano dei cosidetti *confounding factors*, ovvero **variabili che sono correlate sia con la variabile indipendente che dipendente**. Questi fattori rendono più complicato identificare e studiare la relazione a cui siamo interessati, in quanto potrebbero implicare la presenza di una eventuale relazione spuria.
- Per esempio se vogliamo valutare l'effetto di una data medicina x sul sonno, risulta indispensabile che lo sperimentatore effettui l'esperimento in modo tale che i pazienti dormano solo per effetto della medicina x e non per mezzo di altri fattori. Tutti gli altri fattori sono considerate variabili che "confondono" l'esperimento.

Dati sperimentali V

- Quando si vuole studiare l'effetto della relazione istruzione–salari usando dati osservazionali, si deve accettare l'idea che la variabile anni di istruzione di ciascun individuo è essa stessa una variabile risposta ottenuta a seguito di decisione individuale e di conseguenza non si possono considerare gli anni di istruzione come dei valori assegnati casualmente agli individui dallo sperimentatore.