Statistica e Laboratorio

1. Introduzione alla statistica e all'analisi dei dati

Paolo Vidoni

Dipartimento di Scienze Economiche e Statistiche
Università di Udine
via Tomadini 30/a - Udine
paolo.vidoni@uniud.it

https://elearning.uniud.it/

Sommario

Informazioni generali

- **Orario delle lezioni:** lunedì ore 8.30-10.30 (aula C3), martedì ore 11.30-13.30 (aula C3), giovedì ore 13.30-15.30 (Lab Inf A028-A029).
- **Lezioni ed esercitazioni:** si svolgeranno prevalentemente nelle giornate di lunedì e martedì; le esercitazioni sono collocate all'interno delle ore dedicate alle lezioni teoriche.
- **Laboratorio:** le lezioni di laboratorio informatico con R, curate dalla dott.ssa Valentina Mameli, si svolgeranno prevalentemente nella giornata di giovedì.
- **Ricevimento studenti:** mercoledì ore 09.00-11.00 presso il Dipartimento di Scienze Economiche e Statistiche, via Tomadini 30/a o su Teams. È possibile fissare ulteriori appuntamenti contattando il docente.
- **Materiale didattico:** disponibile su `https://elearning.uniud.it/`.
- **Prerequisiti:** non ci sono prerequisiti formali, ma è richiesta la conoscenza dei contenuti degli insegnamenti di Analisi matematica e di Fondamenti di scienza dei dati e laboratorio.

Programma del corso

- Introduzione alla statistica e alla analisi dei dati (Lezione 1)
- Statistica descrittiva
  - Analisi univariate (Lezione 2)
  - Analisi multivariate (Lezione 3)
- Calcolo delle probabilità
  - Probabilità elementare (Lezione 4)
  - Variabili casuali (Lezione 5)
  - Modelli probabilistici (Lezione 6)
- Inferenza statistica
  - Statistiche campionarie (Lezione 7)
  - Stima puntuale e stima intervallare (Lezione 8)
  - Verifica delle ipotesi (Lezione 9)
  - Modello di regressione lineare (Lezione 10)

I metodi e i modelli verranno ripresi e applicati nella parte di laboratorio, utilizzando il software statistico R (`https://www.r-project.org/`).

Organizzazione del corso

- Le lezioni teoriche saranno supportate da *slides* e le esercitazioni saranno dedicate allo svolgimento di esercizi.
- Lo studio di R verrà sviluppato durante le ore di laboratorio.
- L'esame consiste in una *Prova scritta* che si svolge in due parti:
  - *Parte 1: teoria ed esercizi* (peso 75%) con quattro esercizi (5 punti ognuno), che possono anche richiedere qualche nozione su R, e due domande che riguardano la teoria (6 punti ognuna);
  - *Parte 2: laboratorio* (peso 25 %) con una analisi di dati da svolgere con R (che si svolgerà utilizzando i PC del laboratorio o il proprio laptop).
- La prova scritta risulta sufficiente se il voto di entrambe le parti risulta maggiore o uguale a 18.
- Gli studenti che sono risultati sufficienti alla prova scritta possono confermare il voto oppure chiedere di svolgere una *Prova orale facoltativa*.
- La *Prova scritta - parte 1: teoria ed esercizi* può venire sostituita da tre provette svolte durante il corso (tutte tre con voto maggiore o uguale a 18; è ammesso il recupero di una su tre).

Bibliografia e sussidi didattici

Alcuni testi di consultazione:

- Walpole, R.E., Myers, H.R., Myers, S.L. e Ye, K.E. (2020). *Probabilità e Statistica per Ingegneria e Scienze. Strumenti e Applicazioni in R*, 9a Ed., Pearson.
- Navidi, W. (2006). *Probabilità e Statistica per l'Ingegneria e le Scienze*, McGraw-Hill.
- Iacus, S.M., Masarotto, G. (2013). *Laboratorio di Statistica con R*, II ed., Mc Graw-Hill.
- Crivellari, F. (2006). *Analisi Statistica dei Dati con R*, Apogeo.
- *OpenIntro Statistics* (testi, laboratori, video, forum, ecc.) disponibile alla pagina web `https://www.openintro.org/`.

Dalla pagina web `https://elearning.uniud.it/` si possono scaricare:

- le *slides* riferite alle lezioni teoriche;
- gli esercizi riferiti ai vari argomenti trattati;
- dispense, script, dataset e materiale informativo su R;
- i temi d'esame passati;
- ulteriori materiali (tavole, formulario, ecc.)

Che cos'è la statistica?

**(tratto da una pubblicazione dell'*American Statistical Association*
www.amstat.org/careers/whatisstatistics.cfm)**

- "Statistics is the science of learning from data, and of measuring, controlling, and communicating uncertainty" (Davidian, M. and Louis, T. A., 10.1126/science.1218685).
- Lo studio di un fenomeno di interesse richiede spesso l'analisi di **informazioni espresse in forma quantitativa** (**i dati**).
- La statistica è una matematica applicata che, pur avendo come riferimento concreto i dati e il particolare fenomeno di interesse, interviene con principi e metodologie proprie.
- La statistica è di supporto a varie discipline, quali l'economia, la finanza, la sociologia, la medicina, l'ingegneria, la biologia, ecc.
- La statistica fornisce concetti e strumenti per evidenziare gli aspetti rilevanti racchiusi nei dati e per quantificare la forza delle conclusioni che si possono dedurre da tale analisi.

I passi principali di una analisi statistica

- Formulazione del problema: studiare il contesto, specificare gli obiettivi dell'analisi, tradurre il problema in termini matematico-statistici.
- Raccogliere e organizzare i dati: dati osservazionali o sperimentali, unità statistiche, dati mancanti, codifica e organizzazione dei dati.
- Analisi iniziale dei dati: sintesi numeriche e grafiche per iniziare a esplorare i dati.
- Analisi completa dei dati.
- Presentazione dei risultati.

"The formulation of a problem is often more essential than its solution which may be merely a matter of mathematical or experimental skill" (Albert Einstein)

Cosa fanno gli statistici?

- Il mondo sta diventando sempre più "quantitativo" e molte professioni si basano sulle evidenze che si possono trarre dai dati e sull'utilizzazione di opportuni metodi introdotti per tale finalità.
- Gli statistici definiscono e utilizzano metodi e modelli per raccogliere a analizzare dati, con l'obiettivo di ricavare informazioni utili sul fenomeno di interesse.
- Gli statistici sono spesso chiamati a collaborare con esperti di altre discipline per affrontare problemi sia di natura scientifica che pratica e devono necessariamente essere a conoscenza degli aspetti più rilevanti del particolare contesto di applicazione dei metodi.
- "The best thing about being a statistician is that you get to play in everyone else's backyard" (John Tukey, Bell Labs, Princeton Univ.).

Moneyball (2011): `https://www.youtube.com/watch?v=KWPhV6PUr9o`

Hans Rosling (TED2006): `https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen`

Il *dream job* del prossimo decennio?

Hal R. Varian, chief economist di Google, l'ha definito come "il lavoro più sexy del prossimo decennio":
`https://www.youtube.com/watch?v=pi472Mi3VLw`

Le 7 professioni ICT più richieste, Il Sole 24 Ore (23 marzo 2017)

CARRIERE DEL FUTURO

Dal data scientist al cloud architect, i 7 lavori che serviranno di più nell'Europa digitale

C'è chi le chiama ancora "carriere del futuro". Ma sono lavori che si stanno già facendo largo sul mercato internazionale e italiano, con un ritmo di crescita che supera quello dei settori tradizionali. Si sta parlando dei digital jobs, le professioni dell'Ict che nascono e si evolvono di pari passo con la cosiddetta rivoluzione digitale. Secondo dati della Commissione europea, il fabbisogno di risorse potrebbe oscillare tra i 500mila e le 700mila posizioni entro il 2020. Quali sono i ruoli e le competenze più ambite? Dai data scientist, gli "scienziati" dei Big Data, agli sviluppatori e agli specialisti di Cloud. Ecco le sette professioni che potrebbero crescere di più nei prossimi quattro anni.

–di **Alberto Magnani** | 23 marzo 2017

1/7 Le 7 professioni Ict più richieste / Data scientist

Fino a qualche anno fa in pochi avrebbero saputo descrivere il lavoro dei data scientist: "gli scienziati dei dati" che raccolgono e trasformano in informazioni utili i dati del Web. Oggi sono una risorsa sempre più ambita dalle imprese, chiamate ad analizzare i flussi della Rete e ricavare dati preziosi per potenziare il business di una società o migliorare l'efficienza dei servizi. In Italia la figura è ancora poco diffusa, se è vero che – dati del Politecnico di Milano – solo il 30% delle imprese ha assunto un data scientist all'interno del proprio organico. All'estero, in compenso, il fenomeno è già esploso: un report della società di consulenza Deloitte pronostica una "carenza" di un milione di analisti di dati a livello globale entro il 2018. L'equivalente di due volte i professionisti Ict che saranno richiesti dall'intera Europa entro il 2020.

Il **data scientist**, con competenze trasversali in statistica, matematica e informatica, è tra le figure ICT più ricercate del mondo del lavoro, con una domanda che supera di gran lunga la disponibilità di candidati.

Sommario

Business analytics

Business analytics

From Wikipedia, the free encyclopedia

Not to be confused with Business analysis.

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(October 2010) (Learn how and when to remove this template message)*

Business analytics (**BA**) refers to the skills, technologies, practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning.[1] Business analytics focuses on developing new insights and understanding of business performance based on data and statistical methods. In contrast, business intelligence traditionally focuses on using a consistent set of metrics to both measure past performance and guide business planning, which is also based on data and statistical methods.[*citation needed*]

Business analytics makes extensive use of statistical analysis, including explanatory and predictive modeling,[2] and fact-based management to drive decision making. It is therefore closely related to management science. Analytics may be used as input for human decisions or may drive fully automated decisions. Business intelligence is querying, reporting, online analytical processing (OLAP), and "alerts."

In other words, querying, reporting, OLAP, and alert tools can answer questions such as what happened, how many, how often, where the problem is, and what actions are needed. Business analytics can answer questions like why is this happening, what if these trends continue, what will happen next (predict), and what is the best outcome that can happen (optimize).[3]

Contents [hide]

Examples of application [edit]

Banks, such as Capital One, use data analysis (or *analytics*, as it is also called in the business setting), to differentiate among customers based on credit risk, usage and other characteristics and then to match customer characteristics with appropriate product offerings. Harrah's, the gaming firm, uses analytics in its customer loyalty programs. E & J Gallo Winery quantitatively analyses and predicts the appeal of its wines. Between 2002 and 2005, Deere & Company saved more than $1 billion by employing a new analytical tool to better optimize inventory.[3] A telecoms company that pursues efficient call center usage over customer service may save money as well.

Types of analytics [edit]

- Decision Analytics: supports human decisions with visual analytics that the user models to reflect reasoning.[4]
- Descriptive Analytics: gains insight from historical data with reporting, scorecards, clustering etc.
- Predictive Analytics: employs predictive modelling using statistical and machine learning techniques
- Prescriptive Analytics: recommends decisions using optimization, simulation, etc.

Basic domains within analytics [edit]

- Behavioral analytics
- Cohort Analysis
- Collections analytics
- Contextual data modeling - supports the human reasoning that occurs after viewing "executive dashboards" or any other visual analytics
- Cyber analytics
- Enterprise Optimization
- Financial services analytics
- Fraud analytics
- Health care analytics
- Marketing analytics
- Pricing analytics
- Retail sales analytics
- Risk & Credit analytics
- Supply Chain analytics
- Talent analytics
- Telecommunications
- Transportation analytics
- Customer Journey Analytics
- Market Basket Analysis

Business e (big) data

February 11, 2012

The Age of Big Data

By STEVE LOHR

GOOD with numbers? Fascinated by data? The sound you hear is opportunity knocking.

A report last year by the McKinsey Global Institute, the research arm of the consulting firm, projected that the United States needs 140,000 to 190,000 more workers with "deep analytical" expertise and 1.5 million more data-literate managers, whether retrained or hired.

Welcome to the Age of Big Data. The new megarich of Silicon Valley, first at Google and now Facebook, are masters at harnessing the data of the Web — online searches, posts and messages — with Internet advertising. At the World Economic Forum last month in Davos, Switzerland, Big Data was a marquee topic. A report by the forum, "Big Data, Big Impact," declared data a new class of economic asset, like currency or gold.

Non solo business: social data analytics

Social data analysis

From Wikipedia, the free encyclopedia

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(April 2011) (Learn how and when to remove this template message)*

Social data analysis is the data-driven analysis of how people interact in social contexts, often with data obtained from social networking services. The goal may be to simply understand human behavior or even to propagate a story of interest to the target audience. Techniques may involve understanding how data flows within a network, identifying influential nodes (people, entities etc.), or discovering trending topics.

Social data analysis usually comprises two key steps: 1) gathering data generated from social networking sites (or through social applications), and 2) analysis of that data, in many cases requiring real-time (or near real-time) data analysis, measurements which understand and appropriately weigh factors such as influence, reach, and relevancy, an understanding of the context of the data being analyzed, and the inclusion of time horizon considerations. In short, social data analytics involves the analysis of social media in order to understand and surface insights which is embedded within the data.[1]

Social data analysis can provide a new slant on business intelligence where social exploration of data can lead to important insights that the user of analytics did not envisage/explore. The term was introduced by Martin Wattenberg in 2005[2] and recently also addressed as big social data analysis in relation to big data computing.

Systems are available to assist users in analyzing social data. They allow users to store data sets and create corresponding visual representations. The discussion mechanisms often use frameworks such as a blogs and wikis to drive this social exploration/Collaborative intelligence.

Contents [hide]

Obtaining social data [edit]

Social networking services are increasingly popular with the development of Web 2.0. Many of these services provide APIs that allow easy access to their data by responding to user queries with the requested data in the form of XML or JSON formatted strings. In order to protect privacy of their users, services such as Facebook require that the person requesting data has the necessary data access permissions. Services may also charge users for access to their data. Sources of social data include Twitter, Facebook, news websites, Wikipedia and We Feel Fine.

Some APIs only allow access to data in small quantities, hence indexing the data in bulk can become a challenge. Six_Apart was the first social media company to provide a (free) firehose of content for all the posts in their network (provided over XMPP). Twitter later came along and provided a firehose as did companies like Spinn3r, Datasift, and GNIP.

Methods of analysis [edit]

In most cases, we want to find out the relationships between social data and another event or we want to get interesting results from social data analyses to predict some events. There are some outstanding articles in this field, including *Twitter Mood Predicts The Stock Market*,[3] *Predicting The Present With Google Trends*[4] etc. In order to accomplish these goals, we need the appropriate methods to do the analyses. Usually, we use statistic methods, methods of machine learning or methods of data mining to do the analyses.

Universities all over the world are opening graduate program in Social Data Analysis.

Key concepts [edit]

When talking about social data analytics, there are a number of factors it's important to keep in mind (which we noted earlier):[1]

- **Sophisticated Data Analysis**: what distinguishes social data analytics from sentiment analysis is the depth of the analysis. Social data analysis takes into consideration a number of factors (context, content, sentiment) to provide additional insight.
- **Time consideration**: windows of opportunity are significantly limited in the field of social networking. What's relevant one day (or even one hour) may not be the next. Being able to quickly execute and analyze the data is an imperative.
- **Influence Analysis**: understanding the potential impact of specific individuals can be key in understanding how messages might be resonating. It's not just about quantity, it's also very much about quality.
- **Network Analysis**: social data is also interesting in that it migrates, grows (or dies) based on how the data is propagated throughout the network. It's how viral activity starts—and spreads.

See also [edit]

- Data Analysis
- Big Data
- Business intelligence
- Collaborative intelligence
- Social analytics
- IBM jStart
- Social data revolution
- Economic and Social Data Service

Sommario

Software statistici

- Le analisi statistiche richiedono strumenti software opportuni, ovvero i così detti **software statistici**.
- I fogli di calcolo (Microsoft Excel e altri) sono strumenti utili, ma solo per analisi semplici. I software statistici sono strumenti molto più potenti e flessibili.
- Esiste una varietà notevole di software statistici commerciali: IBM SPSS, STATA, SAS, ecc.
- L'ambiente *open-source* R (`http://www.r-project.org`) è per molti aspetti il migliore strumento ad oggi disponibile.

Il software statistico R

[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Mailing Lists
Bug Tracking
Development Site
Conferences
Search

R Foundation

Foundation
Board
Members
Donors
Donate

Documentation

Manuals
FAQs
The R Journal
Books
Certification
Other

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- **R version 3.2.5 (Very, Very Secure Dishes)** has been released on 2016-04-14. This is a rebadging of the quick-fix release 3.2.4-revised.
- Beta test period for version 3.3.0 has been extended to accommodate new Windows toolchain for CRAN. Final release rescheduled for Tuesday 2016-05-03.
- **Notice XQuartz users (Mac OS X)** A security issue has been detected with the Sparkle update mechanism used by XQuartz. Avoid updating over insecure channels.
- **R version 3.2.4 (Very Secure Dishes)** has been released on Thursday 2016-03-10.
- **R version 3.3.0 (Supposedly Educational) prerelease versions** will appear starting Monday 2016-03-14. Final release is scheduled for Thursday 2016-04-14.
- The **R Logo** is available for download in high-resolution PNG or SVG formats.
- **useR! 2016**, will take place at Stanford University, CA, USA, June 27 - June 30, 2016.
- **The R Journal Volume 7/2** is available.
- **R version 3.2.3 (Wooden Christmas-Tree)** has been released on 2015-12-10.
- **R version 3.1.3 (Smooth Sidewalk)** has been released on 2015-03-09.

R (programming language)

From Wikipedia, the free encyclopedia

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(January 2016)*

R is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing.[3] The R language is widely used among statisticians and data miners for developing statistical software[4] and data analysis.[5] Polls, surveys of data miners, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years.[6]

R is an implementation of the S programming language combined with lexical scoping semantics inspired by Scheme.[7] S was created by John Chambers while at Bell Labs. There are some important differences, but much of the code written for S runs unaltered.[8]

R was created by Ross Ihaka and Robert Gentleman[9] at the University of Auckland, New Zealand, and is currently developed by the *R Development Core Team*, of which Chambers is a member. R is named partly after the first names of the first two R authors and partly as a play on the name of S.[10]

R is a GNU project.[11] The source code for the R software environment is written primarily in C, Fortran, and R.[12] R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems. While R has a command line interface, there are several graphical front-ends available.[13]

Contents [hide]

R

Paradigm	multi-paradigm: array, object-oriented, imperative, functional, procedural, reflective
Designed by	Ross Ihaka and Robert Gentleman
Developer	R Core Team[1]
First appeared	1993; 23 years ago[2]
Stable release	3.2.5 / April 14, 2016; 1 day ago
Typing discipline	Dynamic
License	GNU General Public License
Website	r-project.org
Influenced by	
	S, Scheme, XLispStat
Influenced	
	Julia

R Programming at Wikibooks

- Le applicazioni e gli esercizi presentati durante il corso verranno svolti anche utilizzando il software R.
- R è un software libero per le analisi statistiche con notevoli potenzialità dal punto di vista computazionale e grafico.
- Il progetto è iniziato a metà degli anni '90 per opera di due ricercatori che lavoravano in Nuova Zelanda: **R**obert Gentleman and **R**oss Ihaka.
- R è un sistema open source basato sul linguaggio di programmazione S. È un delle implementazioni esistenti di S; l'altra è il software commerciale S-PLUS.
- R studio (`https://www.rstudio.com/`) è un sistema di sviluppo integrato per R, che comprende la console, un editor che permette l'esecuzione diretta del codice, finestre grafiche e per la gestione dei comandi, per il debugging e l'organizzazione dell'ambiente di lavoro.

Lo sviluppo di R

- R è sviluppato con continuità da un gruppo di ricercatori che forma l'*R Core Group*, ma si avvale del contributo di molti volontari.

  Il numero di coloro che utilizzano quotidianamente R è dell'ordine di centinaia di migliaia, almeno; molti più del numero di coloro che utilizzano i software commerciali come SAS, STATA or IBM SPSS.
- Sono stati sviluppati moltissimi pacchetti aggiuntivi (*R packages*).
- Questa immensa ricchezza rende R il più importante e utilizzato software statistico, sia in ambito accademico che imprenditoriale.

  Negli Stati Uniti, R viene usato abitualmente da compagnie come *Google*, *Pfizer* e *Bank of America*.
- Le ragioni per cui viene utilizzato R in questo corso sono: versatilità, interattività, popolarità, libertà.

Le caratteristiche fondamentali di R

- Ambiente sviluppato per l'analisi di dati e la modellazione statistica.
- Notevoli potenzialità grafiche.
- È basato su un linguaggio di programmazione orientato agli oggetti, facilmente sviluppabile dagli utilizzatori.
- È libero e open source: gli utilizzatori possono accedere al codice e modificarlo liberamente.
- È multi piattaforma e quindi si può utilizzare su diversi sistemi operativi, con anche alcune opzioni disponibili per tablet e smartphone.
- Si può utilizzare per analizzare dataset molto grandi, utilizzando semplici interfacce con i principali sistemi per la gestione di database.
- Può interfacciarsi facilmente con altri linguaggi di programmazione come C, C++, Fortran e Java.

Documentazione per R

- Oltre alla pagina web principale, ci sono alcuni archivi come ad esempio il CRAN (Comprehensive R Archive Network) da cui si può scaricare il software (`https://cran.r-project.org/`), mailing list, forum, blog, pagine GitHub, ecc.
- Molta documentazione è disponibile sia sul CRAN che nel web.
- Si possono citare molti riferimenti bibliografici, tra i quali:

  Iacus, S. M. e Masarotto, G. (2007). *Laboratorio di Statistica con R*, 2a Ed., McGraw-Hill.

  Ieva, F., Masci, C. e Paganoni, A.M. (2016). *Laboratorio di Statistica con R*, 2a Ed., Pearson.

  Wickham, H. and Grolemund, G. (2017). *R for Data Science*, O'Reilly. (`https://r4ds.had.co.nz/`).

  Long, J.D, and Teetor, P. (2011). *R Cookbook*, 2nd Ed., O'Reilly. (`https://rc2e.com/`).

Sommario

I dati

I dati si ottengono sia tramite **osservazione** sia tramite **sperimentazione**.

I **dati osservazionali** esistono in natura e vengono rilevati direttamente per come si presentano. Sono spesso osservazioni di caratteristiche antropometriche, demografiche, socio-economiche, ma non solo. Un'importante classe di dati osservazionali è rappresentata dai risultati di censimenti o di sondaggi d'opinione.

Esempi di osservazione:

- rilevazione dell'età, statura, genere e gruppo sanguigno dei residenti nel Comune di Udine al 31 dicembre 2017;
- rilevazione dei dati relativi alle vendite di una certa azienda nel mese appena concluso;
- misurazione del livello di vari inquinanti nell'aria;
- monitoraggio delle visite ad un certo sito web in un certo periodo di tempo.

I **dati sperimentali** sono creati in circostanze controllate. L'esperimento può essere replicato un numero di volte arbitrario, mantenendo fede ad un determinato protocollo sperimentale.

Esempi di sperimentazione:

- pesatura di una modesta quantità di reagente con una bilancia di precisione;
- valutazione del grado di efficacia di un nuovo farmaco;
- analisi del grado di affidabilità di un componente elettronico;
- estrazione di un campione di individui da una popolazione nota.

Sia nei dati ottenuti tramite sperimentazione che tramite osservazione si rileva usualmente la presenza di una certa **variabilità**.

Unità statistiche e popolazione

I dati rappresentano l'informazione disponibile su certe caratteristiche di una **popolazione**, ovvero l'intera collezione di **unità statistiche** sulle quali si cerca l'informazione.

È necessario individuare in modo non equivoco la popolazione di interesse. Si possono considerare:

- **popolazioni reali**, che sono costituite da unità che hanno un'esistenza fisica simultanea al momento della rilevazione; sono popolazioni effettive e quindi *finite*. Possono essere esaminate in modo completo (**censimento**) o parziale (**campionamento**).
- **popolazioni virtuali**, che hanno un'esistenza concettuale e sono evocate dalla potenziale replicabilità a piacere della sperimentazione; sono (potenzialmente) *infinite* e quindi esaminabili solo in modo parziale (**campionamento**), considerando il numero finito di volte con cui la sperimentazione viene ripetuta.

Esempio. Un esempio di popolazione reale è l'insieme dei residenti nel Comune di Udine al 31 dicembre 2017. Un esempio di popolazione virtuale è l'insieme di tutte le possibili replicazioni (potenzialmente infinite) della pesata di una quantità di reagente con una bilancia di precisione. ◊

Censimento e campionamento

Censimento: si esaminano *tutte* le unità di una *popolazione reale*, con riferimento a determinate caratteristiche di interesse.

Anche per popolazioni reali i censimenti sono raramente effettuati. Molto più spesso si estrae un campione.

Campionamento: si esamina un *sottoinsieme* finito di unità statistiche, appartenenti ad una *popolazione reale o virtuale*, selezionate mediante l'esperimento di campionamento.

L'**esperimento di campionamento** è un particolare esperimento, assimilabile all'estrazione casuale di alcuni elementi da un'urna.

È un **esperimento casuale** (**aleatorio**), dal momento che risultano possibili una pluralità di esiti (campioni osservati) e prima di effettuare il campionamento non è possibile individuare con certezza quale potenziale campione verrà selezionato (**variabilità campionaria**).

Affinché il campione porti informazioni sull'intera popolazione, la sua *estrazione* deve essere *casuale*.

- Il campione va scelto in modo che rifletta le caratteristiche della popolazione.
- Esistono vari **piani di campionamento**, il più semplice è il **campionamento casuale semplice**, assimilabile all'*estrazione casuale con reinserimento* di elementi da un'urna.

Per l'inerente replicabilità dell'estrazione del campione, i dati campionari vanno interpretati come *sperimentali* anche se sono di tipo osservazionale.

Il **calcolo delle probabilità** fornisce gli strumenti matematici per lo studio di esperimenti casuali, e in particolare degli esperimenti di campionamento.

Le popolazioni reali possono essere studiate per via campionaria o censuaria, mentre per le popolazioni virtuali la strategia campionaria è la sola possibile.

Anche quando si conduce un'indagine di tipo campionario l'obiettivo non muta: si desidera acquisire informazione sull'intera popolazione (reale o virtuale), con riferimento a particolari caratteristiche di interesse.

Statistica: una definizione più precisa

I metodi statistici si possono dividere in due grandi classi.

Statistica descrittiva: metodi per la descrizione, la presentazione e la sintesi dei dati disponibili, al fine di individuarne la struttura essenziale.

Le finalità sono principalmente di tipo descrittivo, poiché si sintetizzano le informazioni disponibili, che riguardano la totalità della popolazione.

Anche quando i dati disponibili rappresentano un campione estratto da una popolazione, nella statistica descrittiva non se ne tiene conto.

Statistica inferenziale: metodi per ricavare dai dati campionari informazioni sulla popolazione di riferimento e per quantificare la fiducia da accordare a tali informazioni.

Le informazioni estraibili da un campione possono essere riferite alla popolazione con, inevitabilmente, un certo grado di **incertezza**

A tal fine, si utilizzano il linguaggio e i metodi del **calcolo delle probabilità**, ovvero quel ramo della matematica che permette di trattare l'incertezza.

Le tre discipline, calcolo delle probabilità, statistica descrittiva e statistica inferenziale, hanno strette relazioni reciproche.

Esempio: una gara di mountain bike

Si considerano i tempi, in minuti, riportati da 24 ciclisti che hanno partecipato ad una gara di *mountain bike*.

No.	Tempo	No.	Tempo	No.	Tempo	No.	Tempo
1	13.20	7	20.45	13	27.49	19	28.10
2	13.25	8	20.47	14	27.50	20	29.09
3	14.01	9	20.50	15	27.54	21	31.34
4	14.02	10	21.06	16	27.55	22	32.28
5	16.58	11	21.12	17	27.58	23	32.48
6	17.00	12	25.38	18	28.04	24	35.44

La popolazione reale di riferimento, costituita dai 24 ciclisti che hanno concluso la gara, viene esaminata in modo completo (**censimento**).

La caratteristica di interesse è il tempo impiegato per concludere la prova. ◊

Esempio: prevenire la spina bifida

L'*acido folico* è una vitamina del gruppo B, abbondante nelle verdure.

La *spina bifida* è un difetto di saldatura della colonna vertebrale che si manifesta durante la crescita embrionale e comporta gravi conseguenze.

Per studiare l'efficacia preventiva di una dieta ricca di acido folico nei primi mesi di gravidanza, a 2000 donne sono stati somministrati 800 microgrammi al giorno di tale vitamina ed a altre 2000 donne un placebo.

Si sono ottenuti e seguenti risultati:

	Acido folico	Placebo
Malati	0	6
Sani	2000	1994

La popolazione di riferimento è costituita dall'insieme di tutte le donne in gravidanza, comparabili per abitudini e stili di vita.

La popolazione viene esaminata in modo parziale (**campionamento**). Il campione è costituito dalle 4000 donne che sono entrate nella sperimentazione.

Si è interessati a capire se la somministrazione preventiva dell'acido folico in gravidanza risulta efficace per diminuire il numero di insorgenze della patologia in questione. ◊

Esempio: misurazioni con errore

Si effettuano 20 misurazioni, ripetute nelle medesime condizioni sperimentali, di un determinato oggetto con uno strumento affetto da errore non sistematico

9.85	10.02	9.91	10.08	9.61	9.94	9.96	10.06	10.09	9.45
9.89	10.13	9.87	9.85	10.14	10.07	9.60	10.15	9.84	9.97

La popolazione virtuale di riferimento è costituita dall'insieme di tutte le infinite, potenziali replicazioni della misurazione dell'oggetto in esame.

La popolazione viene esaminata in modo parziale (**campionamento**). Il campione è costituito dalle 20 misurazioni che sono state effettuate.

Si è interessati a determinare la vera dimensione dell'oggetto. ◊

Le principali indagini statistiche italiane

L'ISTAT (`https://www.istat.it/`) svolge circa 200 indagini ogni anno su tematiche di carattere socio-demografico, economico e ambientale.

Le indagini censuarie svolte periodicamente dall'ISTAT sono:

- il censimento generale della popolazione e delle abitazioni;
- il censimento generale dell'industria, del commercio, dei servizi e dell'artigianato;
- il censimento generale dell'agricoltura.

Tra le indagini campionarie a carattere periodico si ricordano:

- l'indagine trimestrale sulle forze di lavoro;
- l'indagine sui consumi delle famiglie;
- l'indagine multiscopo.