

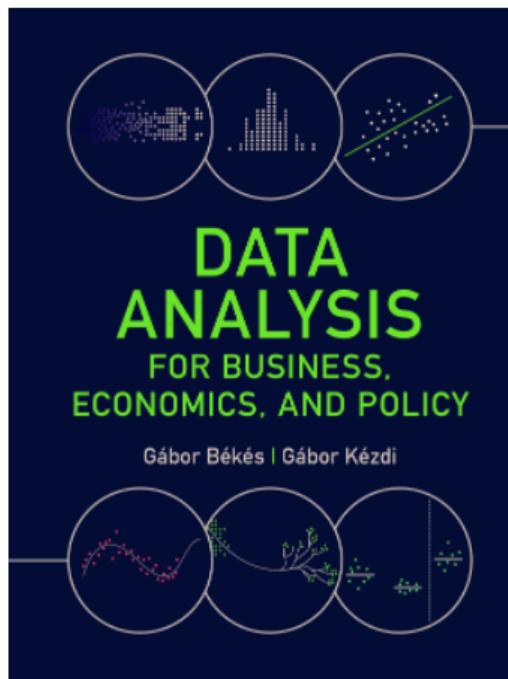
# 01 Origins of Data

Gábor Békés

Data Analysis 1 – MS Business Analytics: Exploration

2023

# Slideshow for the Békés-Kézdi Data Analysis textbook



- ▶ Cambridge University Press, 2021
- ▶ [gabors-data-analysis.com](http://gabors-data-analysis.com)
  - ▶ Download all data and code:  
[gabors-data-analysis.com/data-and-code/](http://gabors-data-analysis.com/data-and-code/)
- ▶ This slideshow is for Chapter 01

## Motivation

- ▶ Suppose, you want to understand the extent and patterns of differences in online and offline prices. A super project, the Billion Prices Project at MIT did a variety of data collection approaches such as crowd-sourcing platforms, mobile phone apps and web scraping methods.
  - ▶ Interested in understanding more about management practices? The World Management Survey is a major effort by academics to survey practices around the world - asking the same questions in many countries the same way.

# What is data

- ▶ Data is most straightforward to analyze if it forms a single data table.
- ▶ Format: Data table (matrix)
- ▶ A data table consists of *observations* and *variables*.
  - ▶ Observations are also known as cases, or rows
  - ▶ Variables are sometimes called features or covariates.
- ▶ In a data table the rows are the observations, columns are variables.
- ▶ Storage: comma separated values .csv (.txt) is simplest. Delimited can be anything: comma(,), semicolon (;) or other (|)
- ▶ A dataset is a collection of data tables, typically related / used in a project
  - ▶ 10 data tables, same topic for 10 different years

# Basics: data structure and quality

## Data structures

- ▶ Cross-sectional (xsec) data have information on many units observed at the same time.
- ▶ Time series (tseries) data have information on a single unit observed many times.
- ▶ Multi-dimensional (panel) data have multiple dimensions.
  - ▶ Many cross-sectional units observed many times
  - ▶ Units observed in different space

## Data structures

A bit more on multi-dimensional - panel (xt) data

- ▶ A common type of panel data has many units, each observed multiple times. Such data is sometimes called *longitudinal data*, or cross-section-time-series data, sometimes abbreviated as *xt data*.
- ▶ Example: countries observed repeatedly for several years
- ▶ In xt data tables observations are identified by two ID variables: one for the cross-sectional units, one for time.
- ▶ xt data is *balanced* if all cross-sectional units are observed at the very same time periods. It is called unbalanced if some cross-sectional units are observed more times than others.

## Finding a good deal among hotels: data collection

- ▶ Welcome to Vienna, Austria
- ▶ hotels dataset
- ▶ Collected from a price comparison website + anonymized.
- ▶ Vienna, 2017 November weekday,  $N = 428$
- ▶ For each hotel the data includes information on the location of the hotel, the price on the night in focus in EUR, average customer rating, stars of the hotel, and distance to the city center.



Image: [en.wikipedia.org/wiki/File:Montage\\_of\\_Vienna.jpg](https://en.wikipedia.org/wiki/File:Montage_of_Vienna.jpg)

## Data structures

Table: List of observations

hotel_id	accom_type	country	city	city_actual	dist	stars	rating	price
21894	Apartment	Austria	Vienna	Vienna	2.7	4	4.4	81
21897	Hotel	Austria	Vienna	Vienna	1.7	4	3.9	81
21901	Hotel	Austria	Vienna	Vienna	1.4	4	3.7	85
21902	Hotel	Austria	Vienna	Vienna	1.7	3	4	83
21903	Hotel	Austria	Vienna	Vienna	1.2	4	3.9	82

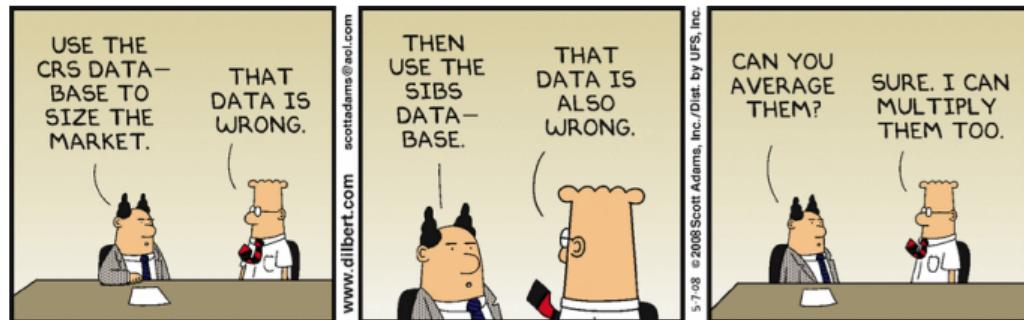
Source: hotels dataset. Vienna, for a 2017 November weekday

List of five observations with key variable values:

- ▶ ‘accom\_type’ is the type of accommodation.
- ▶ ‘city’ is the city based on the search, city\_actual is the municipality.

# Data quality is key

- ▶ Data quality is key
- ▶ Garbage-in-garbage-out:  
If our data is useless to answer our question the results of our analysis are bound to be useless...
- ▶ ... no matter how fancy method we apply to it.



## Data quality and your question

Data quality is generally a subjective notion!

- ▶ First you have to specify what is your (research) question!
- ▶ What do you want to explore or understand?
- ▶ If you have a clear answer, then you can decide on your data quality!

However, there are some objective measures to decide if you have your question!

# Data quality

1. Content - what is the substance a variable captures.
  - ▶ Just because a variable is called something it doesn't necessarily measure that (e.g., "product quality", "socio-economic status").
2. Validity - how close the actual content of the variable to the intended content.
3. Reliability. If we were to measure the same variable multiple times for the same observation it should give the same result.
4. Comparability of measurement - how similarly the same variable is measured across different observations.
5. Coverage -what proportion of the observations in focus are in the data.
  - ▶ Complete coverage (rare).
  - ▶ Incomplete coverage (almost always).
6. Unbiased selection - if coverage incomplete the observations that are included in the data should be similar to all observations that were intended to be covered.

## SIDENOTE

- ▶ This is not the type of class where you will have to memorize a list
- ▶ But you should be able to evaluate the quality of the data you work with
- ▶ Always know your data.
  - ▶ Data quality is key (remember: garbage in, garbage out).
  - ▶ Data quality is determined by how the data was collected.

## Data analysts should know their data

- . Data analysts should know their data
  - ▶ How data was born
  - ▶ All details of measurement that may be relevant for their analysis

To this end, consider having

- ▶ README.txt that describes where dataset comes from
- ▶ VARIABLES.xls that provides basic information on your variables

# Data collection

# Data collection

- ▶ Automated data collection
- ▶ Survey
- ▶ Administrative / Census
- ▶ Big Data

## Collecting data from existing sources

- ▶ Data, or information that can be turned into data, is collected by someone else
- ▶ For purposes different from the purpose of our analysis
- ▶ Data quality consequences
  - ▶ May not contain variables that we need
  - ▶ Validity of main variables may be high or low
  - ▶ Potential selection bias if incomplete coverage of observations
- ▶ Frequent advantages
  - ▶ Inexpensive
  - ▶ Often many observations
  - ▶ Can have complete coverage

# Data collection: Digital

## Automated data collection

- ▶ Application Programming Interface, or API – directly load data into a statistical software.
  - ▶ API is a software intermediary, or an interface,
  - ▶ It allows programs, or scripts, to talk to each other.
- ▶ API widely used in many context.
  - ▶ Macro data: FRED - St Louis Fed at [research.stlouisfed.org/docs/api/fred/](https://research.stlouisfed.org/docs/api/fred/), also World Bank, etc.
  - ▶ Micro data such as weather at: [openweathermap.org/api](https://openweathermap.org/api)
- ▶ Data collection limited to dataset.
- ▶ Typically additional info available.

## Data collection: Digital

- ▶ Collecting data from online platform
- ▶ html code includes data, can be found, analyzed and collected
  - ▶ online services
  - ▶ code in R (rvest) / Python (beautiful soup), Selenium (many languages)
- ▶ Need extensive cleaning
- ▶ Once a procedure is ready (code, script), can be repeated
- ▶ Data collection limited to what is on a site

## Data collection: Administrative

- ▶ Business transactions
- ▶ Government records, taxes, social security
- ▶ Often: census - records on the population
- ▶ Many advantages
  - ▶ Often great coverage, few missing values, high quality content
  - ▶ Many well defined and documented variables
- ▶ Some disadvantages
  - ▶ Variables defined for business/government purposes. May not fit in analysis plans
  - ▶ Often not detailed/specific enough
  - ▶ Biggest problem is very limited access

## Finding a good deal among hotels: data collection

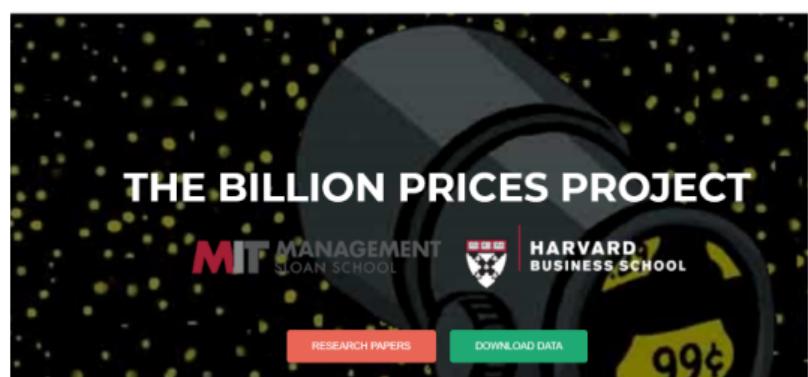
- ▶ The dataset on hotels in Vienna was collected from a price comparison website, by web scraping.
- ▶ On a specific date
- ▶ The purpose of the website is not facilitating data analysis...
- ▶ No other potential source
- ▶ Good quality, but noise, needed work to make it ready for analysis.
- ▶ Coverage is good but not full. Hotels advertising on these websites are not a random sub-sample. Which are the hotels that are left out?

## Comparing online and offline prices: data collection

- ▶ The Billion Prices Project - academic initiative - product prices collected
- ▶ This course: Cavallo (2017, AER)
- ▶ 56 large multi-channel retailers in 10 countries.
- ▶ price levels identical about 72 percent of the time.
- ▶ Price changes are not synchronized but have similar frequencies and average sizes.

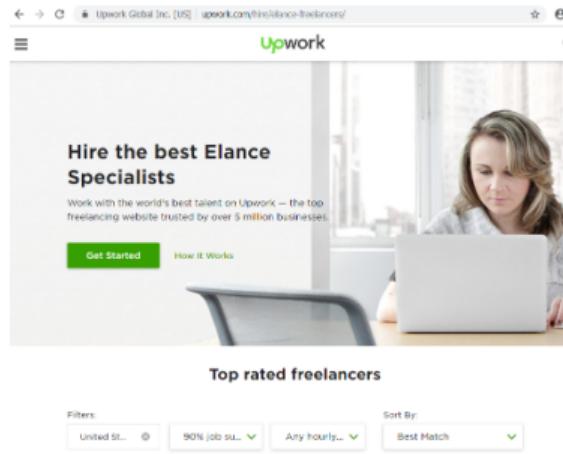
The Billion Prices Project

Home Our Public Data Our Research



## Comparing online and offline prices: data collection

- ▶ BPP is about measuring prices for the same products sold through different channels
- ▶ Mixed methods
- ▶ Offline data collectors by Mechanical Turk / Upwork
- ▶ Online prices were scraped
- ▶ Project managers focusing on collecting info on exactly the same products on approximately during the same time



## Data quality - billion prices project data

1. Content - what product, what price
2. Validity - intention is price of target product available at store.  
**What could go wrong?**

## Data quality - billion prices project data

1. Content - what product, what price
2. Validity - intention is price of target product available at store.  
*What could go wrong?*
3. Reliability. Timing is very difficult especially if price change frequently
4. Comparability in measurement- are products *equally* well identified? Laptop vs cheese
5. Coverage. Not universal. Project plan choice.
6. Unbiased selection. Time consuming planning. If electronic goods, need a typical set of TVs, phones etc.

# Survey and sampling

## Data collection: Survey

- ▶ Surveys collect data by asking people (*respondents*) and recording their answers.
- ▶ Answers to a *questionnaire* are short and easily transformed into variables.
- ▶ Major advantage: you can ask exactly what you want to know
- ▶ There are two major kinds of surveys: self-administered surveys and interviews.
- ▶ Web, telephone, in person, mix - computer aided interview.
- ▶ Choice of data collection approach matters a great deal.
- ▶ Self-administered survey
  - ▶ cheap and efficient, can use visual aids.
  - ▶ What could go wrong?

# Sampling

- ▶ Sometimes we can collect data on all observations we want
- ▶ Those all observations are called the population
  - ▶ All employees in an organization
  - ▶ All countries on Earth
- ▶ More often we don't because it's impractical or prohibitively expensive
- ▶ Sampling is when we purposefully collect data on a subset of the population
  - ▶ A sample is a subset of the population
  - ▶ Sampling is the process that selects that subset

## Representative sample

- ▶ A sample is good if it represents the population
- ▶ A sample is representative of a population if
  - ▶ all important variables have very similar distributions in the sample and the population
  - ▶ all patterns in the sample are very similar to the patterns in the population
- ▶ Examples
  - ▶ The age distribution of a sample of employees is the same as the age distribution of all employees
  - ▶ The average online - offline price difference is the same in the sample of stores in the sample as in all stores with both online and offline sales

# How can we tell if a sample is representative

- ▶ Never for sure
  - ▶ We know the distributions and patterns in the sample but not in the population
  - ▶ The very reason to have a sample is because we can't collect the same data on all observations in the population
- ▶ Benchmarking
  - ▶ We may know a few distributions or patterns in the population
  - ▶ Those should be similar in the sample
  - ▶ Example: proportion female employees in the sample and among all employees
- ▶ Knowing the process of sampling
  - ▶ Random sampling is known to lead to representative samples with high likelihood

## Sampling: Random sampling

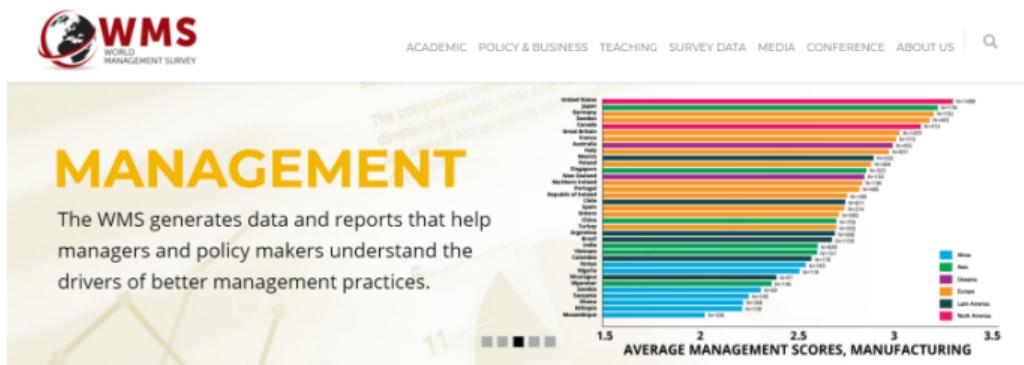
- ▶ *Random sampling* is a selection rule that is independent of any important variable
- ▶ Random sampling is the process that most likely leads to representative samples.
- ▶ Any other methods may lead to biased selection.
- ▶ Important is the independence of the rule from anything important for the analysis
- ▶ Examples
  - ▶ Good: people with odd-numbered birth dates (a 50% sample)
  - ▶ Good: the first half of a list of firms that were sorted by a random number generated by the computer
  - ▶ Bad: the first half of a list of people by alphabetical order
  - ▶ Bad: firms that were established in the most recent years.

## Random sampling is best

- ▶ Provided sample is large enough.
- ▶ In small samples (dozens) anything is possible
- ▶ It's the sample size that matters not how large a fraction it is of the total population size.
- ▶ Sample of a few thousand observations may equally well represent populations of fifty thousand, ten million, or three hundred million.
- ▶ The required sample size depends on details of what you want to measure
- ▶ MORE on this is DA4 (Winter)

## Management quality and firm size: data collection

- ▶ What causes superior performance of some countries? What causes superior performance of some firms in some countries?
- ▶ Many potential arguments: Institutions that lead to competitive markets. Education that helps research - yields new patents
- ▶ [www.worldmanagementsurvey.org](http://www.worldmanagementsurvey.org) - Survey on firms and management.



## Management quality and firm size: data collection

- ▶ Ask 10K+ manufacturing firms (also public sector)
- ▶ Developing management questions
  - ▶ Scorecard for 18 monitoring, targets and incentives practices
  - ▶ Approx 45 minute phone interview of manufacturing plant managers
- ▶ Obtaining unbiased comparable responses ("Double-blind")
  - ▶ Interviewers do not know the company's performance
  - ▶ Managers are not informed (in advance) they are scored
  - ▶ Run from London, with same training and country rotation
- ▶ Getting firms to participate in the interview
  - ▶ Introduced as "Lean-manufacturing" interview, no financials
  - ▶ Run by 100+ MBAs (credible with business experience)

## Management quality and firm size: data collection

Example question: "how is performance tracked?"

- ▶ (1): Measures tracked do not indicate directly if overall business objectives are being met. Certain processes are not tracked at all.
- ▶ (3): Most key performance indicators are tracked formally. Tracking is overseen by senior management.
- ▶ (5): Performance is continuously tracked and communicated, both formally and informally, to all staff using a range of visual management tools.

## Management quality and firm size: data collection

- ▶ Survey quality assessment
- ▶ Content of each score - based on information gathered in a standardized way translated to scores by the interviewers using standardized rules.
- ▶ Validity, reliability and comparability - **How to think about assessment?**
- ▶ What would be an alternative? Pros and Cons?

# What is different with Big Data?

- ▶ Big Data refers to: (i) massive (very large) datasets that are (ii) often automatically and continuously collected and stored, and (iii) may be of complex nature.
- (i) Very large. Billions of observations. (Bigger than what fits into your computer.)
  - ▶ Warning: just because sample is large, it is not necessarily representative!!!!
- (ii) Automatic collection. Not for your analytic purpose - unlike a survey. Data collected by apps, sensors.
- (iii) Complex - text (video, music/noise), network, multidimensional, maps

# What is different and what is he same with Big Data?

Some of these are kind of cryptic for now; we will clarify them in subsequent chapters

- ▶ Different
  - ▶ A particular source of uncertainty of the results of an analysis is greatly reduced
  - ▶ Rare or more nuanced patterns can be uncovered
  - ▶ Practical challenges
    - ▶ Some challenges may be solved by working with a random subsample
- ▶ Same
  - ▶ Need to represent entire population if incomplete coverage
    - ▶ Example: Big Data with 75% coverage with a selection bias leads to biased results
    - ▶ Non-big data from same population with 1% random sample leads to good results

# Sample selection bias

- ▶ The sample you collect is different to the population
- ▶ This difference is crucial in the story
- ▶ Example: Predicting presidential election
  - ▶ 1936: Literary Digest. FD Roosevelt vs Landon. 10m people asked. 2m replied. Biggest poll ever. Landon was predicted win 57%
  - ▶ 1948 Chicago Tribune. Dewey predicted beat Truman. Used phone registry.
  - ▶ What could have gone wrong?

## Legal and ethical aspects

- ▶ Data collection - ethical and legal constraints
- ▶ Especially with sensitive information
- ▶ GDPR

Always communicate with the source owner(s) and or with legal professional if you are planning to use seemingly sensitive data!

## Data collection: hard, time-consuming, costly.

- ▶ Collecting data is a tedious task, and costly as well.
- ▶ Usually it is much harder than expected, with many on-the-field problems.
- ▶ Worth getting some experience!

# AI and data collection, wrangling

- ▶ Data collection and management often behind walls
- ▶ AI can help write code to web-scrape etc
- ▶ AI is great to give a first impression of your dataset, incl. quality, data structure
- ▶ AI is helpful to discuss sampling ideas
- ▶ AI needs context to do good, and will not have proper domain knowledge
- ▶ AI needs supervision

# Main takeaway

- ▶ Know your data
  - ▶ How it was born,
  - ▶ What its main advantages are;
  - ▶ What its main disadvantages are.
- ▶ Data quality determines the results of your analysis
  - ▶ Data quality is determined by how the data was born.
- ▶ Data is stored in data tables
  - ▶ Rows are observations
  - ▶ Columns are variables
- ▶ Data may come from
  - ▶ Existing sources (admin, transactions, web scraping)
  - ▶ Collected purposefully for the analysis (surveys)

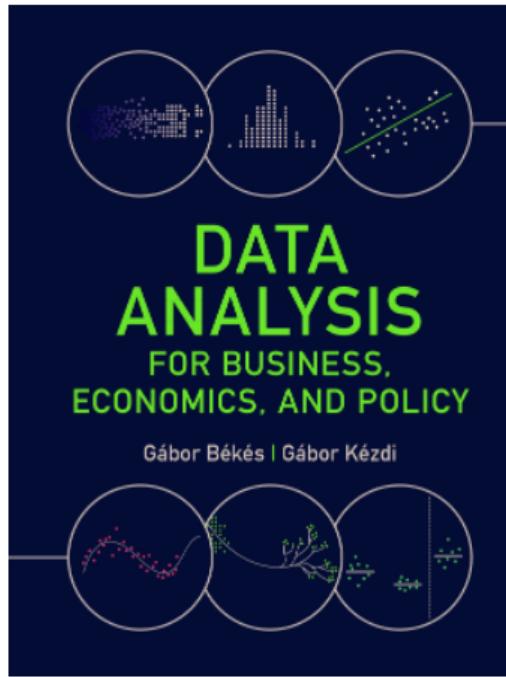
## 02 Preparing data for analysis

Gábor Békés

Data Analysis 1: Exploration

2023

# Slideshow for the Békés-Kézdi Data Analysis textbook



- ▶ Cambridge University Press, 2021
- ▶ [gabors-data-analysis.com](http://gabors-data-analysis.com)
  - ▶ Download all data and code:  
[gabors-data-analysis.com/data-and-code/](http://gabors-data-analysis.com/data-and-code/)
- ▶ This slideshow is for Chapter 02

## Motivation

- ▶ Does immunization of infants against measles save lives in poor countries? Use data on immunization rates in various countries in various years from the World Bank. How should you store, organize and use the data to have all relevant information in an accessible format that lends itself to meaningful analysis?
- ▶ You want to know, who has been the best manager in the top English football league. Have downloaded data on football games and on managers. To answer your question you need to combine this data. How should you do that? And are there issues with the data that you need to address?

## Variable types: Qualitative vs quantitative

- ▶ Data can be born (collected, generated) in different form, and our variables may capture the quality or the quantity of a phenomenon.
- ▶ Quantitative variables are born as numbers. Typically take many values.
  - ▶ also called numeric variables
  - ▶ special case is time (date)
- ▶ Qualitative variables, also called categorical variables, take on a few values, with each value having a specific interpretation (belonging a category).
  - ▶ Another name used is categorical or factor variable.
  - ▶ binary variable (YES/NO) is special case.

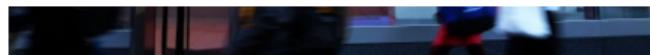
## Variable types - binary

- ▶ A special case is a binary variable, which can take on two values
  - ▶ ...yes/no answer to whether the observation belongs to some group. Best to represent these as 0 or 1 variables: 0 for no, 1 for yes.
  - ▶ Sometimes binary variables with 0-1 values are called dummy variables or an indicator
  - ▶ Flag - binary showing existence of some issue (such as missing value for another variable, presence in another dataset)

## Variable types - formal definition

1. Nominal qualitative variables take on values that *cannot* be unambiguously ordered. **Color, brands**
2. Ordinal, or ordered variables take on values that are *unambiguously ordered*. All quantitative variables can be ordered; some qualitative variables can be ordered, too. **Grades**
3. "Interval" variables are ordered variables, with a *difference between values that can be compared*. **Degree Celsius. Price in dollar.**
4. "Ratio" (or "scale") variables are interval variables with the additional property: their ratios mean the same regardless of the magnitudes. This additional property also implies a meaningful zero in the scale. **Distance in miles. Price in dollar.**

# Storing variables: Example the Washington Post (2016)



(Jewel Samad/AFP/Getty Images)

By Christopher Ingraham  
August 26, 2016

A surprisingly high number of scientific papers in the field of genetics contain errors introduced by Microsoft Excel, according to an analysis recently published in the journal *Genome Biology*.

A team of Australian researchers analyzed nearly 3,600 genetics papers published in a number of leading scientific journals — like *Nature*, *Science* and *PLoS One*. As is common practice in the field, these papers all came with supplementary files containing lists of genes used in the research.

The Australian researchers found that roughly 1 in 5 of these papers included errors in their gene lists that were due to Excel automatically converting gene names to things like calendar dates or random numbers.

*[This new model for training scientists could create a conflict of interest]*

You see, genes are often referred to in scientific literature by symbols — essentially shortened versions of full gene names. The gene "Septin 2" is typically shortened as SEPT2. "Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Protein Ligase" gets merrily shortened to MARCH1.

What you type	What you see	How Excel stores it
MARCH1	1-MAR	42430
SEPT2	2-SEP	42615

<https://www.washingtonpost.com/news/wonk/wp/2016/08/26/an-alarming-number-of-scientific-pa>

# Data wrangling (data munging)

Data wrangling is the process of transforming raw data to a set of data tables that can be used for a variety of downstream purposes such as analytics.

## [1] Understanding and storing

- ▶ start from raw data
- ▶ understand the structure and content
- ▶ create tidy data tables
- ▶ understand links between tables

## [2] Data cleaning

- ▶ understand features, variable types
- ▶ filter duplicates
- ▶ look for and manage missing observations
- ▶ understand limitations

# The tidy data approach

A useful concept of organizing and cleaning data is called the *tidy data* approach:

1. Each observation forms a row.
2. Each variable forms a column.
3. Each type of observational unit forms a table.
4. Each observation has a unique identifier (ID)

Advantages:

- ▶ standard data tables that turn out to be easy to work with.
- ▶ finding errors and issues with data are usually easier with tidy data tables
- ▶ transparent, which helps other users to understand
- ▶ easy to extend. New observations added as new rows; new variables as new columns.

## Simple tidy data table

Table: A simple tidy table

	Variables/columns		
	hotel_id	price	distance
Observations/rows	21897	81	1.7
	21901	85	1.4
	21902	83	1.7

Source: [hotels-vienna data](#). Vienna, 2017 November weekend.

## Tidy data table of multi-dimensional data

- ▶ The *tidy approach* - store xt data in data tables with each row referring to one cross-sectional unit observed in one time period.
- ▶ One row is one observation '*it*'.
- ▶ This is sometimes called the *long format* for xt data.
- ▶ The next row then may be the same cross-sectional unit observed in the next time period.
- ▶ Important and difficult task for analysts is to figure out the structure of multi-dimensional data and create tidy data tables.
  
- ▶ Also used: *wide format* - one row would refer to one cross-sectional unit, and different time periods are represented in different columns. Good for presenting and some analysis. Not to keep data.

## Displaying immunization rates across countries

- ▶ xt panel of countries with yearly observations,
- ▶ Downloaded from the World Development Indicators data website maintained by the World Bank.
- ▶ Illustrate the data structure focusing on the two ID variables (country and year) and two other variables, immunization rate and GDP per capita.

## Displaying immunization rates across countries – WIDE

Country	imm2015	imm2016	imm2017	gdppc2015	gdppc2016	gdppc2017
India	87	88	88	5743	6145	6516
Pakistan	75	75	76	4459	4608	4771

Wide format of country-year panel data, each row is one country, different years are different variables. imm: rate of immunization against measles among 12–13-month-old infants. gdppc: GDP per capita, PPP, constant 2011 USD. Source: world-bank-vaccination data

## Displaying immunization rates across countries – LONG

Country	Year	imm	gdppc
India	2015	87	5743
India	2016	88	6145
India	2017	88	6516
Pakistan	2015	75	4459
Pakistan	2016	75	4608
Pakistan	2017	76	4771

Note: Tidy (long) format of country-year panel data, each row is one country in one year. imm: rate of immunization against measles among 12–13-month-old infants. gdppc: GDP per capital, PPP, constant 2011 USD. Source: world-bank-vaccination data.

## Relational database

- ▶ The relational database is a concept of organizing information.
- ▶ It is a data structure that allows you map a concept set of information into a set of tables
- ▶ Each table is made up of rows and columns
- ▶ Each row is a record (observation) identified with a unique identifier *ID* (also called *key*).
- ▶ Rows (observations) in a table can be linked to rows in other tables with a column for the unique ID of the linked row (*foreign ID*)
  
- ▶ Define these tables, understand structure
- ▶ Merge tables when needed

## Identifying successful football managers

- ▶ Who have been the best football managers in England?
- ▶ We combine data from two sources for this analysis, one on teams and games, and one on managers.
- ▶ Data covers 11 seasons of English Premier League (EPL) games – 2008/2009 to 2018/2019
- ▶ The data comes from the website [www.football-data.co.uk](http://www.football-data.co.uk).
- ▶ Each observation is a single game. Key variables are
  - ▶ the date of the game
  - ▶ name of the home team, the name of the away team,
  - ▶ goals scored by the home team, goals scored by the away team

## Identifying successful football managers

Table: Games data

Date	HomeTeam	AwayTeam	Home goals	Away goals
2018-08-19	Brighton	Man United	3	2
2018-08-19	Burnley	Watford	1	3
2018-08-19	Man City	Huddersfield	6	1
2018-08-20	Crystal Palace	Liverpool	0	2
2018-08-25	Arsenal	West Ham	3	1
2018-08-25	Bournemouth	Everton	2	2
2018-08-25	Huddersfield	Cardiff	0	0

Source: football data.

## Identifying successful football managers

- ▶ Is this a tidy data table?

## Identifying successful football managers

- ▶ Is this a tidy data table?
- ▶ It is.
- ▶ Each observation is a game, and each game is a separate row in the data table.
- ▶ Three ID variables identify each observation: date, home team, away team. The other variables describe the result of the game.
- ▶ From the two scores we know who won, by what margin, how many goals they scored, and how many goals they conceded.

## Identifying successful football managers

- ▶ Could we have an alternative tidy table?

## Identifying successful football managers

- ▶ Could we have an alternative tidy table?
- ▶ There is an alternative way to structure the same data table, which will serve our analysis better
- ▶ In this data table, each row is a game played by a team.
- ▶ It includes variables from the perspective of that team: when played, who the opponent was, and what the score was.

# Identifying successful football managers

Table: Games data - long table version

Date	Team	Opponent team	Goals	Opponent goals	Home/away	Points
2018-08-19	Brighton	Man United	3	2	home	3
2018-08-19	Burnley	Watford	1	3	home	0
2018-08-19	Man City	Huddersfield	6	1	home	3
2018-08-19	Man United	Brighton	2	3	away	0
2018-08-19	Watford	Burnley	3	1	away	3
2018-08-19	Huddersfield	Man City	1	6	away	0

## Identifying successful football managers

- ▶ Also a tidy data table, albeit a different one.
- ▶ It has twice as many rows as the original data table: Each game appears twice in this data table, once for each of the playing team's perspectives.
- ▶ New variable to denote whether the team at that game was the home team or the away team.
- ▶ Now we have two ID variables, one denoting the team, and one denoting the date of the game. The identity of the opponent team is a qualitative variable.
  
- ▶ Tidy data has some key features. But a given multi-dimensional data may be stored as tidy in multiple ways.

## Identifying successful football managers

- ▶ Our second data table is on managers.
- ▶ One row is one manager-team relationship.
- ▶ Each manager may feature more than once in this data table if they worked for multiple teams.
- ▶ For each observation, we have the name of the manager, their nationality, the name of the team (club), the start time of the manager's work at the team, and the end time.

## Identifying successful football managers

Table: Managers data

Name	Nat.	Club	From	Until
Arsene Wenger	France	Arsenal	1 Oct 1996	13 May 2018
Unai Emery	Spain	Arsenal	23 May 2018	Present*
Ron Atkinson	England	Aston Villa	7 June 1991	10 Nov 1994
Brian Little	England	Aston Villa	25 Nov 1995	24 Feb 1998
John Gregory	England	Aston Villa	25 Feb 1998	24 Jan 2002
Dean Smith	England	Aston Villa	10 Oct 2018	Present*
Alan Pardew	England	Crystal Palace	2 Jan 2015	22 Dec 2016
Alan Pardew	England	Newcastle	9 Dec 2010	2 Jan 2015

Source: football data. Present = 01 July 2019

## Identifying successful football managers

- ▶ This is a relational dataset.
- ▶ One data table with team-game observations, and one data table with manager-team observations.
- ▶ To work with the data, we need to create a workfile, which is a single data table that is at the team-game level with the additional variable of who the manager was at the time of that game.
- ▶ but before we do, need to merge them...

## Relational data and linking data tables across observations

- ▶ Organize and store data in tidy data tables with appropriate ID variables,
- ▶ how to combine such table into a workfile to run our analysis?
- ▶ The process of pulling different variables from different data tables for well-identified entities to create a new data table is called linking, joining, merging, or matching.

## Relational data and link data tables across observations

Matching (joining) depends on data structure

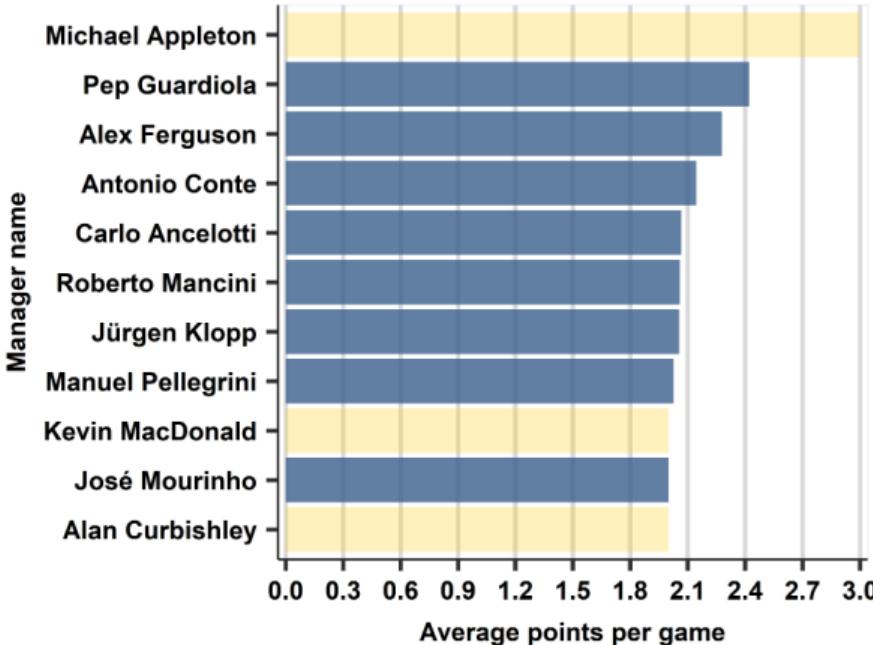
- ▶ one-to-one (1:1) matching: merging tables with the same type of observations.
  - ▶ Football teams and stadium now.
- ▶ many-to-one (m:1) or one-to-many (1:m) matching when in one of the data tables, a value may be matched to more than one values in the other table.
  - ▶ Football teams and their players now - many players in a team.
- ▶ many-to-many (m:m) matching when values in both tables could be matched to many others.
  - ▶ Football teams and their manager ever - some managers worked for multiple teams.

## Identifying successful football managers

- ▶ Started with a relational dataset.
- ▶ Merged two data tables and created a work file
- ▶ With the workfile at hand, we can describe it.
- ▶ The workfile has 8360 team-game observations: in each of the 11 seasons, 20 teams playing 38 games (19 opponent teams twice;  $11 \times 20 \times 38 = 8360$ ).
- ▶ There are 137 managers in the data.

## Identifying successful football managers

- ▶ Remember: data is 11 seasons, EPL.
- ▶ spells at teams: if a manager worked for two teams, we consider it two cases.
- ▶ Success: average points per game
- ▶ Above 2.0



## Complex data - tidy data- summary

- ▶ Creating a tidy data - generating a set of data tables that are easy to understand, combine and extend in the future.
- ▶ If relational data, IDs are essential
- ▶ Often raw data will not come in a tidy format, and you will need to work understanding the structure, relationships and find the individual ingredients.
- ▶ For analysis work, need to combine tidy data tables



# AI and data structure

Generative AI (LLM) is good and helpful

- ▶ understands your data structure
- ▶ helps you combine datasets (suggests code / does it)
- ▶ helps even summarize the data

# AI and data structure

Generative AI (LLM) is good and helpful

- ▶ understands your data structure
- ▶ helps you combine datasets (suggests code / does it)
- ▶ helps even summarize the data

Pay attention

- ▶ Check and debug a lot
- ▶ It does not know what you want. May need keep control.
- ▶ Best is to be very specific and focus on help with coding

# Data wrangling: cleaning

- ▶ Entity resolution:
  - ▶ Dealing with duplicates
  - ▶ ambiguous identification
  - ▶ non-entity rows
- ▶ Missing values

## Wrangling: Filter out Duplicates

- ▶ *duplicates*: some observations appearing more than once in the data.
- ▶ Duplicates may be the result of human error (when data is entered by hand), or the features of data source (e.g., data scraped from classified ads with some items posted more than once).
- ▶ Often, easy process. Just check and get rid of repeated observations
- ▶ Sometimes, same observation is featured number of times.
  - ▶ Need to investigate. Makes sense / an error?
  - ▶ Example: Daily stock quotes, some stock features twice, with different price.
  - ▶ Decision- what to keep. Sometimes no clear-cut way, but usually no big deal.

## Entity identification and resolution

- ▶ More generally, you would need to have unique IDs
- ▶ It could be that two observations belong to two entities although ID is the same.
  - ▶ example: John Smith – there may be many
  - ▶ need to figure out, maybe assign unique IDs in raw data
- ▶ It could be that two observations have different ID but belong to same entity
  - ▶ need to figure out and have a single ID
- ▶ Unique IDs crucial. Numerical IDs are better

## Entity resolution example

Team ID	Unified name	Original name
19	Man City	Manchester City
19	Man City	Man City
19	Man City	Man. City
19	Man City	Manchester City F.C.
20	Man United	Manchester United
20	Man United	Manchester United F.C.
20	Man United	Manchester United Football Club
20	Man United	Man United

Source: various sources

## Getting rid of non-entity observations

- ▶ Rows that do not belong to an entity we want in the data table.
- ▶ Find them and drop them
- ▶ Such as: a summary row in a table that adds up, or averages, variables across all, or some, entities.
- ▶ Case study: a data table downloaded from the World Bank on countries often includes observations on larger regions, such as Sub-Saharan Africa

## Missing values

- ▶ A frequent and important issue with variables is *missing values*.
- ▶ Missing values mean that the value of a variable is not available for some, but not all, observations.
  
- ▶ Scope: How much missing?
- ▶ Reason: Why missing?

## Missing values: scope and reason

### Key issues

1. Look at content of data - related to data quality (esp. coverage)
2. Missing values need to be identified.
  - ▶ Easy: "NA" (for "not available"), a dot ".", an empty space "".
  - ▶ Hard - binary 0 for no, 1 for yes, 9 for missing
  - ▶ Hard - percent 0-100, 9999 for missing
  - ▶ Hard numeric, range is 1-100000, 9999999999 for missing
3. Missing value is encoded. But when aggregate must pay attention.
4. Missing values should be counted. Missing values mean fewer observations with valid information. May actually have a lot fewer observations to work with than the size of the original dataset.
5. The third issue is potential selection bias. Is data missing at random?

## Missing values - Understanding the selection process

- ▶ Random: When missing data really means no information, it may be the result of errors in the data collection process. Rare.
- ▶ In some other cases, missing just means "zero" or "no". In these instances, we should simply recode (replace) the missing values as "zero" or as "no".
- ▶ Often, values are missing systematically. Some survey respondents may not know the answer to a question or refuse to answer it, and such respondents are likely to be different from those who provide valid answers.

## Missing values: what can we do?

Two basic options:

1. Restrict the analysis to observations with non-missing values for all variables used in the analysis.
2. *Imputation* - Fill in some value for the missing values, such as the mean or median.

## Missing values: Some practical advice

- ▶ Focus on more fully filled variables. Often, simpler.
- ▶ Sometimes, informative if missing - create a new variable (called flag) to capture missing value and use this variable instead of the original.
  - ▶ For example, the original variable is a text of the Twitter handler.
  - ▶ Here, the binary variable that is 1 if the person has an account and 0 otherwise could be more interesting.
- ▶ Automatic missing variable filling packages. Advanced only.
- ▶ Always be conservative, impute if absolutely necessary!

## Missing values: Some practical advice

- ▶ For binary variables: zero if yes/no.
- ▶ For qualitative nominal variables, you may add missing as a new value: white, blue red and missing.
  - ▶ What if binary qualitative?
- ▶ For ordinal variables, you may add missing as new value or recode missing to a neutral variable: high, average, low, with missing recoded as average.
- ▶ For quantitative variables - you may recode with mean or median
- ▶ if impute, create a flag and use it analysis
  
- ▶ Always be conservative, impute if absolutely necessary!

# Practical data management

- ▶ Structure of files
- ▶ Naming files

CloudPleasers by Forrest Brazeal



"It has come to our attention that some of you are live-tweeting this event with #NameCon15, two digit year, when it should always be #NameCon2015, four digit year..."

## Naming files

- ▶ Use file names that simple, machine readable. No spaces, punctuation(.,;), accented characters or capitalized letters
- ▶ Human readable. Deliberate use of “\_” (part of info) and “-” (help read)
- ▶ Computer ordering friendly. Use numbers early on. Left pad (01 not 1)
- ▶ dates used in ISO: YYYY-MM-DD.
- ▶ Good examples
  - ▶ "bekes-kezdi\_textbook-presentation\_ceu2018.pdf"
  - ▶ "ch02\_organizing-data\_world-bank-download\_2017-06-01.tex"
- ▶ Never use:
  - ▶ thesis.pdf, mytext.doc
  - ▶ calculations1112018.xls, Gábor's-01.nov.19\_work.pdf.

## Structure of files

It is good practice to structure the data files at three levels.

- ▶ Raw data files
- ▶ Clean and tidy data files
- ▶ Workfile(s) for analysis

You may also organize your folders - maybe even before we start the analysis

- ▶ Wrangling (code)
- ▶ Analysis (code)
- ▶ Output (graphs, tables)
- ▶ Report (pdf, markdown)

## Data wrangling: common steps

1. Write a code - it can be repeated and improved later
2. Understand the structure of the dataset, create data tables, recognize links. Draw a schema.
3. Start by looking into the data table(s) to spot issues
4. Store data in tidy data tables. Make sure one row in the data is one observation and manage duplicates
5. Get each variable in an appropriate format
6. Have a description of variables
7. Make sure values are in meaningful ranges; correct non-admissible values or set them as missing
8. Identify missing values and store them in an appropriate format. Make edits if needed.
9. Document every step of data cleaning

# AI and data wrangling

Generative AI (LLM) is good and helpful

- ▶ understands your variables
- ▶ finds potential problems

<https://chat.openai.com/share/c1f54966-d1ce-4412-acd6-d2f16f9abb53>

# AI and data wrangling

Generative AI (LLM) is good and helpful

- ▶ understands your variables
- ▶ finds potential problems

<https://chat.openai.com/share/c1f54966-d1ce-4412-acd6-d2f16f9abb53>

Pay attention

- ▶ You shall know more and may want different steps
- ▶ It does not know what you want. May need keep control.
- ▶ Eventually copy code – keep reproducible

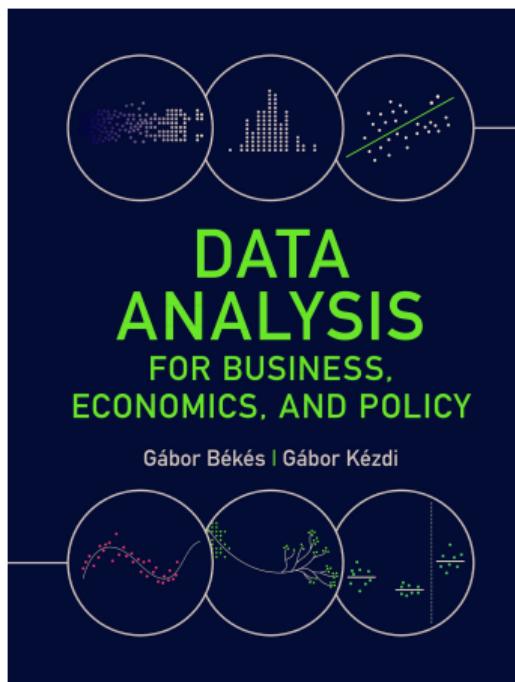
# 03 Exploratory data analysis

Gábor Békés

Data Analysis 1: Exploration

2023

# Slideshow for the Békés-Kézdi Data Analysis textbook



- ▶ Cambridge University Press, 2021
- ▶ [gabors-data-analysis.com](http://gabors-data-analysis.com)
  - ▶ Download all data and code:  
[gabors-data-analysis.com/data-and-code/](http://gabors-data-analysis.com/data-and-code/)
- ▶ This slideshow is for Chapter 03

## Motivation

Understand the market conditions for hotels in Vienna, using prices.

- ▶ How should you start the analysis itself?
  - ▶ How to describe the data and present the key features?
  - ▶ How to explore the data and check whether it is clean enough for (further) analysis?

# Exploratory data analysis (EDA) - describing variables

5 reason to do EDA!

1. To check data cleaning (part of iterative process)
2. To guide subsequent analysis (for further analysis)
3. To give context of the results of subsequent analysis (for interpretation)
4. To ask additional questions (for specifying the (research) question)
5. Offer simple, but possibly important answers to questions.

# Key tasks: describe variables

Look at key variables

- ▶ what values they can take and
- ▶ how often they take each of those values.
- ▶ are there extreme values

Describe what you see

- ▶ Descriptive statistics - key features summarized
- 
- ▶ to understand variables you work with
  - ▶ to make comparisons

# Variable description, histograms

## Frequency of values

- ▶ The *frequency* or more precisely, *absolute frequency* or *count*, of a value of a variable is simply the number of observations with that particular value.
- ▶ The *relative frequency* is the frequency expressed in relative, or percentage, terms: the *proportion* of observations with that particular value among all observations.
- ▶ Practical note: With missing values – proportion can be relative to all observations OR only observations with non-missing values (usual choice).

# Probabilities and frequencies

- ▶ *Probability* is general a concept that is related to relative frequency.
- ▶ Probability is a measure of the likelihood of an *event*.
- ▶ An event is something that may or may not happen.
- ▶ Probabilities are always between zero and one.
- ▶ Probability as a generalization of relative frequencies in datasets.
- ▶ Probabilities are more general than relative frequencies as they can describe events without datasets.

# The distribution and the histogram

A key part of EDA is to look at (empirical) distribution of most important variables.

- ▶ All variables have a *distribution*.
- ▶ The distribution of a variable tells the frequency of each value of the variable in the data.
- ▶ May be expressed in terms of absolute frequencies (number of observations) or relative frequencies (percent of observations).
- ▶ The distribution of a variable completely describes the variable as it occurs in the data.
- ▶ independent from values the other variables may show.

# Histograms

Histogram reveals important properties of a distribution.

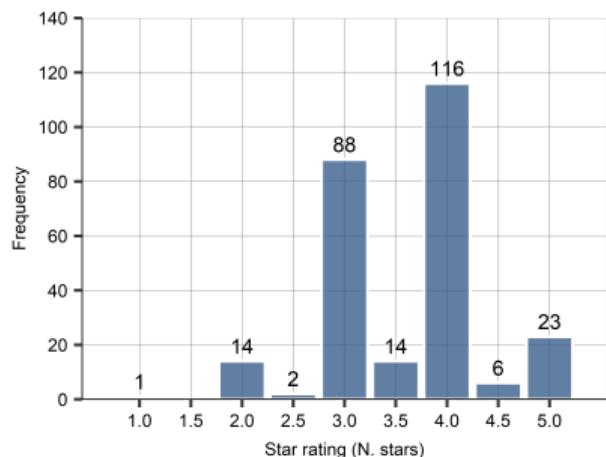
- ▶ Number and location of *modes*: these are the peaks in the distribution that stand out from their immediate neighborhood.
- ▶ Approximate regions for *center* and *tails*
- ▶ *Symmetric* or not - asymmetric distributions have a long left tail or a long right tail
- ▶ *Extreme values*: values that are very different from the rest. Extreme values are at the far end of the tails of histograms.

## Extreme values

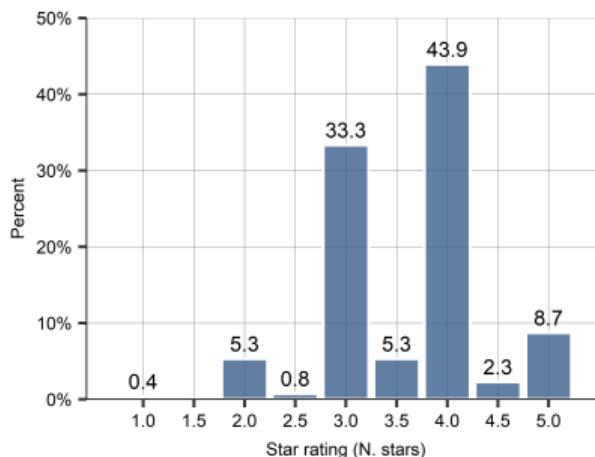
- ▶ Some variables have extreme values: substantially larger or smaller values for one or a handful of observations than the values for the rest of the observations.
  - ▶ Need conscious decision.
    - ▶ Is this an error? (drop or replace)
    - ▶ Is this not an error but not part of what we want to talk about? (drop)
    - ▶ Is this an integral feature of the data? (keep)

# Hotel price histograms

(a) Absolute frequency (count)



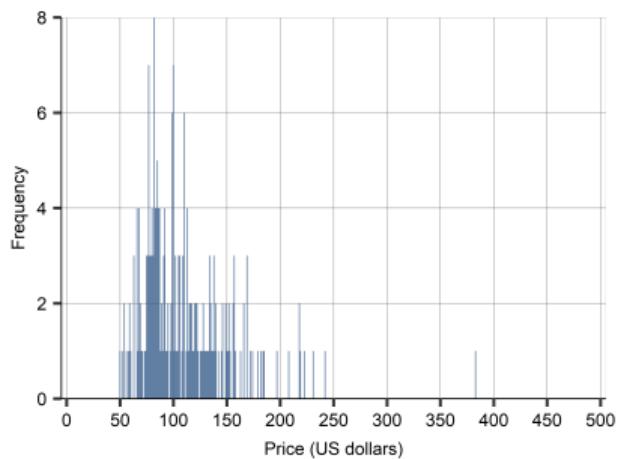
(b) Relative frequency (percent)



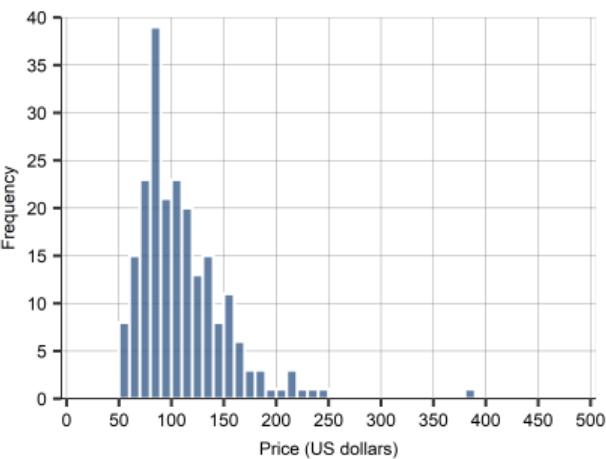
Source: [hotels-vienna dataset](#). Vienna, Hotels only, for a 2017 November weekday

## Hotel price histograms

(a) Histogram: individual values



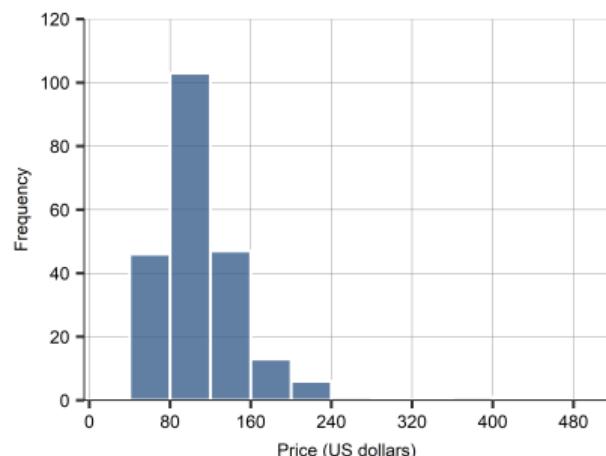
(b) Histogram: 20\$ bins



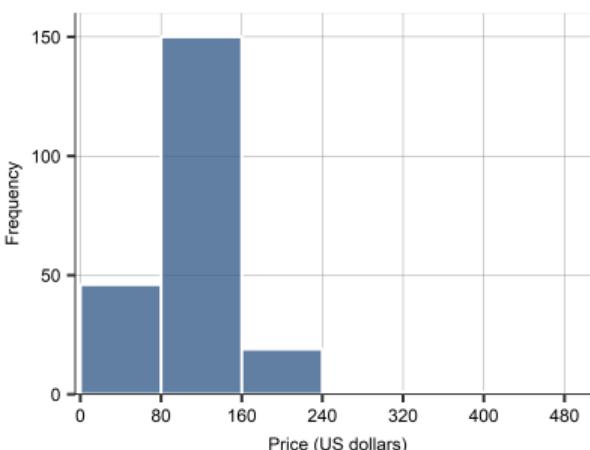
Note: Panel (a) just shows individual values - help see where most values are. Panel (b) is a histogram with 20\$ bins - more useful to capture frequencies. Source: hotels-vienna dataset. Vienna, 3-4 stars hotels only, for a 2017 November weekday

# Hotel price histograms

(a) Histogram: 40\$ bins



(b) Histogram: 80\$ bins

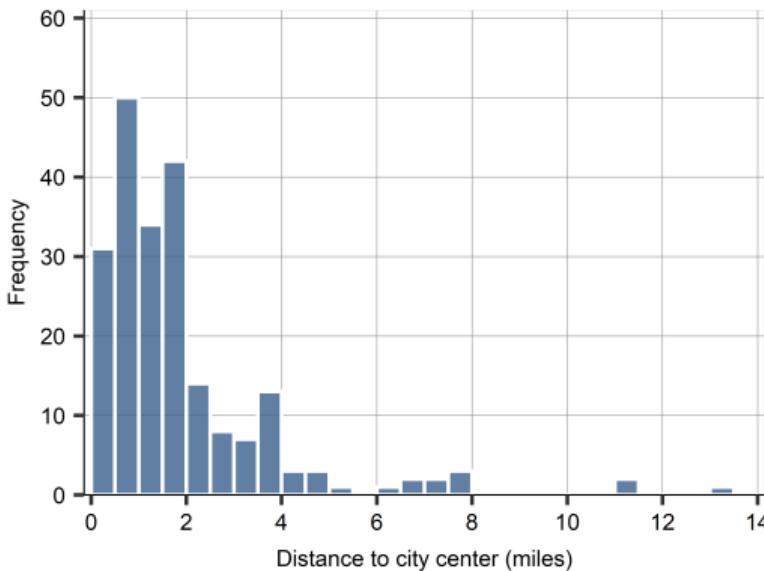


Note: *Bin size matters. Wider bins suggest a more gradual decline in frequency.*

## Hotel density plot

- ▶ Vienna all hotels, 3-4 stars
- ▶ Use absolute frequency (count)
- ▶ For this histogram we use 0.5-mile-wide bins. This way we can see the extreme values in more detail
- ▶ Dropped very far - likely not Vienna

Figure: Histogram of distance to the city center.

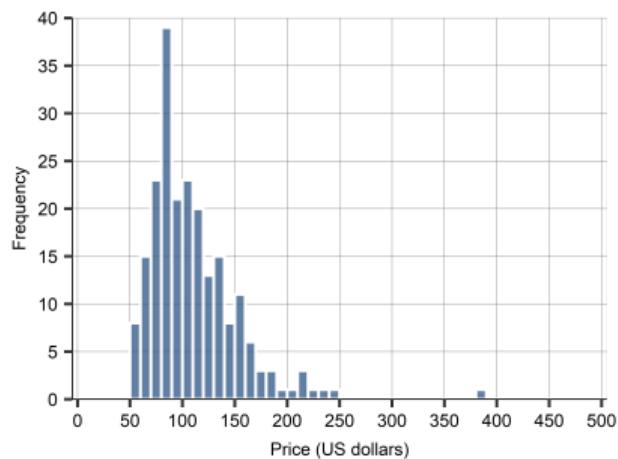


## Hotel prices

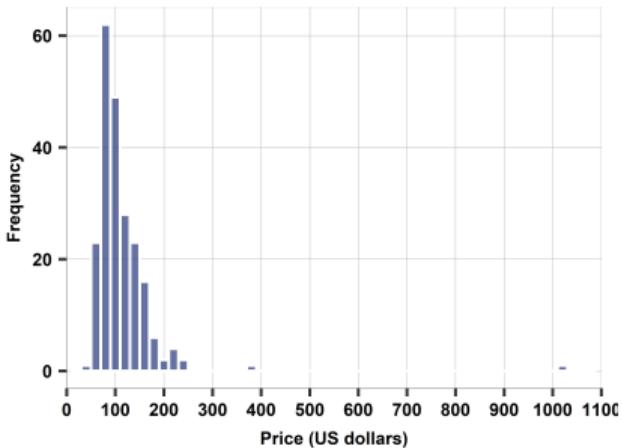
- ▶ Vienna all hotels, 3-4 stars
- ▶ Use absolute frequency (count)
- ▶ We go back to prices
- ▶ How to decide what to include? -> check observation!

## Hotel price histograms

(a) Histogram: 20\$ bins as seen



(b) Histogram: including extreme value above 1000\$



Source: [hotels-vienna dataset](#). Vienna, 3-4 stars hotels only, for a 2017 November weekday

## EDA and cleaning - Vienna hotels

1. Start with full data  $N=428$
2. Tabulate key qualitative variables
3. accommodation type - could be apartment, etc. Focus on hotels.  $N=264$
4. stars - focus on 3, 3.5 4 stars, as lower bit not well covered, luxury could vary a lot.  $N=218$
5. Look at quantitative variables, focus on extreme values.
6. Start with price.  $p=1012$  likely error drop. keep others  $N= 217$
7. Distance: some hotels are far away. define cutoff. drop beyond 8km  $N=214$
8. check why hotels could be far away. Find variable city\_actual. Tabulate. Realise few hotels are not in Vienna. Drop them.  $N=207$
9. So, the final cut: Hotels, 3 to 4 stars, below 1000 euros, less than 8km from center, in Vienna actual  $N=207$ .

# Summary statistics

## Summary statistics

- ▶ For any given variable, a *statistic* is a meaningful number that we can compute from a dataset.
- ▶ Basic *summary statistics* describe the most important features of distributions of variables.
- ▶ Many of you know this. I briefly cover it

## Summary statistics: Sample mean

The most used statistic is the *mean*:

$$\bar{x} = \frac{\sum x_i}{n} \quad (1)$$

where  $x_i$  is the value of variable  $x$  for observation  $i$  in the dataset that has  $n$  observations in total. Two key features

$$\overline{x + a} = \bar{x} + a \quad (2)$$

$$\overline{x \cdot b} = \bar{x} \cdot b \quad (3)$$

# The Expected value

- ▶ The expected value is the value that one can expect for a randomly chosen observation
- ▶ The notation for the expected value is  $E[x]$ .
- ▶ For a quantitative variable, the expected value is the mean
- ▶ For a qualitative variable, it can only be determined if transformed to a number
  - ▶ Male/Female binary variable. Expected value could be probability / relative frequency of females.
  - ▶ Quality of hotel: 1 to 5 stars, mean can be calculated, but its meaning is less straightforward.
    - ▶ What is the assumption for getting the mean as number?

## Summary statistics: The median and other quantiles

- ▶ *quantiles*: a quantile is the value that divides the observations in the dataset to two parts in specific proportions.
- ▶ The *median* is the middle value of the distribution - half the observations have lower value and the other half have higher value.
- ▶ *Percentiles* divide the data into two parts along a certain percentage.
  - ▶ The first percentile is the value below which one percent of the observations are and 99 percent above.
- ▶ *Quartiles* divide the data into two parts along fourths.
  - ▶ 1st quartile has one quarter of the observations below and three quarters above; it is the  $25^{th}$  percentile.
  - ▶ 2nd quartile has two quarters of the observations below and two quarters above; this is the median, and also the  $50^{th}$  percentile.

## Summary statistics: The mode

- ▶ The *mode* is the value with the highest frequency in the data.
- ▶ Some distributions are unimodal, others have multiple modes.
- ▶ Multiple modes are apart from each other, each standing out in its "neighborhood", but they may have different frequencies.

## Summary statistics: central tendency

- ▶ The mean, median and mode are different statistics for the *central value* of the distribution
- ▶ Central tendency.
  - ▶ The mode is the most frequent value
  - ▶ The median is the middle value
  - ▶ The mean is the value that one can expect for a randomly chosen observation.

## Summary statistics: spread of distributions

- ▶ *spread of distributions* is also often used in analysis.
- ▶ Statistics that measure the spread of distributions are the range, inter-quantile ranges, the standard deviation and the variance.
- ▶ The *range* is the difference between the highest value (the maximum) and the lowest value (the minimum) of a variable.
- ▶ The *inter-quantile ranges* is the difference between two quantiles- the third quartile (the 75<sup>th</sup> percentile) and the first quartile (the 25<sup>th</sup> percentile).
- ▶ The 90- 10 percentile range gives the difference between the 90<sup>th</sup> percentile and the 10<sup>th</sup> percentile.

## Summary statistics: standard deviation

- ▶ The most widely used measure of spread is the *standard deviation*. Its square is the *variance*.
- ▶ Variance is the average squared difference of each observed value from the mean.

$$Var[x] = \frac{\sum(x_i - \bar{x})^2}{n} \quad (4)$$

$$Std[x] = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}} \quad (5)$$

## Summary statistics: standard deviation

- ▶ The variance is a less intuitive measure. At the same time, the variance is easier to work with, because it is a mean value itself.
- ▶ The standard deviation (SD) captures the typical difference between a randomly chosen observation and the mean.
  - ▶ Not exactly the average but similar
  - ▶ Same unit of measure (ie dollars)
- ▶ Two distributions with same mean. If SD is higher - more dispersed the data
- ▶ In Finance, SD and variance are measures of price volatility of an asset.
  - ▶ High volatility: price jumps up and down.

$$Std[x] = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}} \quad (6)$$

## Using the standard deviation to define standardized values

The standard deviation is often used to re-calculate differences between values in order to express those in terms of typical distance.

$$x_{\text{standardized}} = \frac{(x - \bar{x})}{\text{Std}[x]} \quad (7)$$

- ▶ *standardized value of a variable* shows the difference from the mean in units of standard deviation.
- ▶ For example: a standardized value of one shows a value is one standard deviation larger than the mean; a standardized value of negative one shows a value is one standard deviation smaller than the mean

## Summary statistics: skewness

- ▶ A distribution is *skewed* if it isn't symmetric.
- ▶ It may be skewed in two ways, having a *long left tail* or having a *long right tail*.
- ▶ Example: hotel price distributions having a long right tail - such as in hotel price distribution.
- ▶ Skewness and the prevalence of extreme values are related. With distributions with long tails, values far away from all other values are more likely.
- ▶ When extreme values are important for the analysis, skewness of distributions is important, too.

## Summary statistics: skewness measure

Simplest measure is *mean–median measure of skewness*.

$$\text{Skewness} = \frac{(\bar{x} - \text{median}(x))}{\text{Std}[x]} \quad (8)$$

- ▶ When the distribution is symmetric its mean and median are the same.
- ▶ When it is skewed with a long right tail the mean is larger than the median: the few very large values in the right tail tilt the mean further to the right.
- ▶ When a distribution is skewed with a long left tail the mean is smaller than the median
- ▶ To make this measure comparable across various distributions use a standardized measure
- ▶ If multiplied by 3, and then it's called *Pearson's second measure of skewness*.

## Visualizing summary statistics

- ▶ Measures of central value: Mean (average), median, other quantiles (percentiles), mode.
- ▶ Measures of spread: Range, inter-quantile range, variance, standard deviation.
- ▶ Measure of skewness: The mean–median difference.
  
- ▶ The box plot is a visual representation of many quantiles and extreme values.
- ▶ The violin plot mixes elements of a box plot and a density plot.

# Visualizing summary statistics

Figure: Boxplot

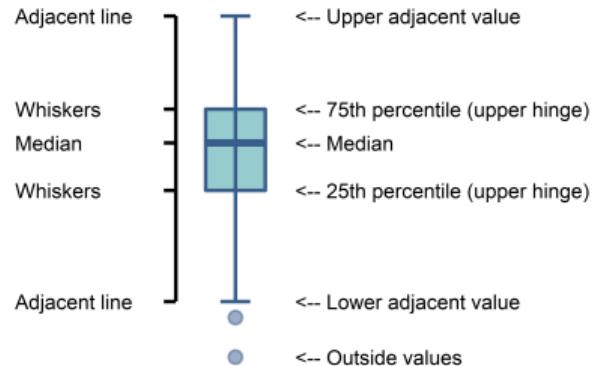
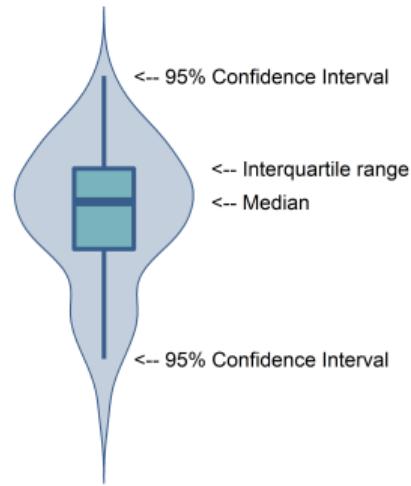


Figure: Violinplot



## Density plots

- ▶ *Density plots* - also called *kernel density estimates*
- ▶ alternative to histograms - instead of bars density plots show continuous curves.
- ▶ Instead of bars, density plots show continuous curves. We may think of them as curves that wrap around the corresponding histograms.
- ▶ density plots complementing histograms - some believe density plots allow for easier comparison of distributions across groups in the data.

## Vienna vs London

- ▶ Compare two cities, how hotel markets vary
- ▶ Vienna, London
- ▶ 3-4 star hotels, only "Hotels" (no apartments), below 1000 dollars.
- ▶ Focus on actual city=Vienna and actual city=London (exclude nearby related villages).
- ▶ Use hotels-europe dataset.
- ▶ N=207 for Vienna, N=435 for London
- ▶ Graphical vs comparison table

# London vs Vienna

Figure: Vienna Austria

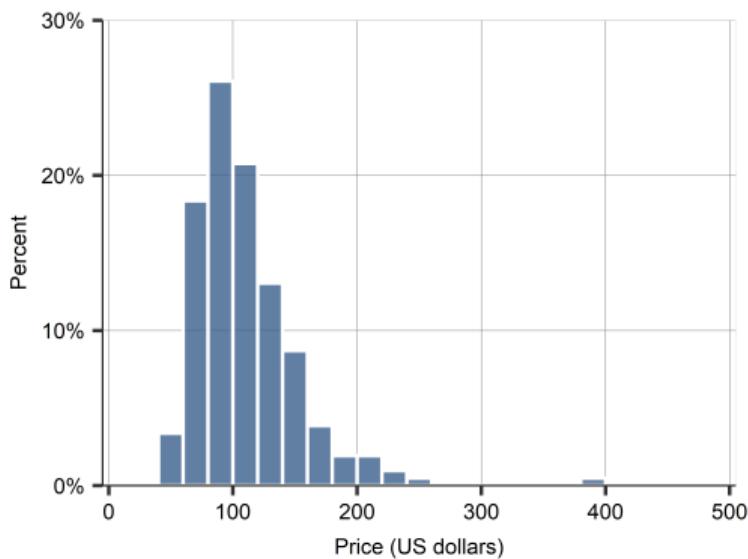
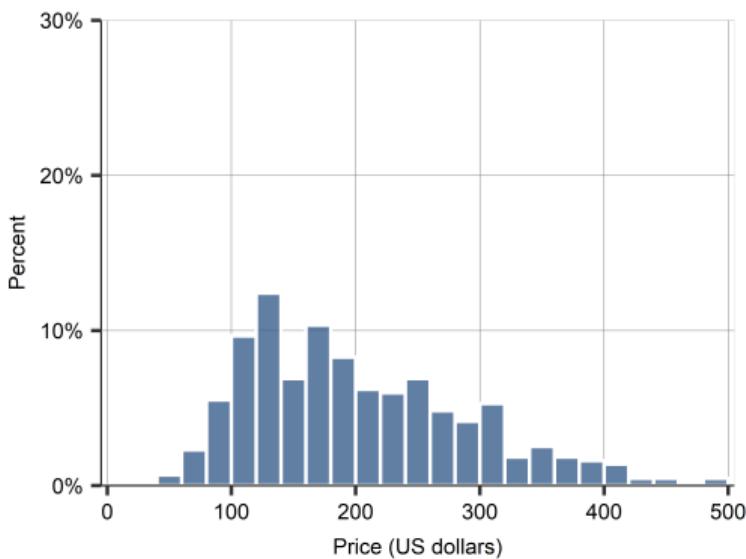
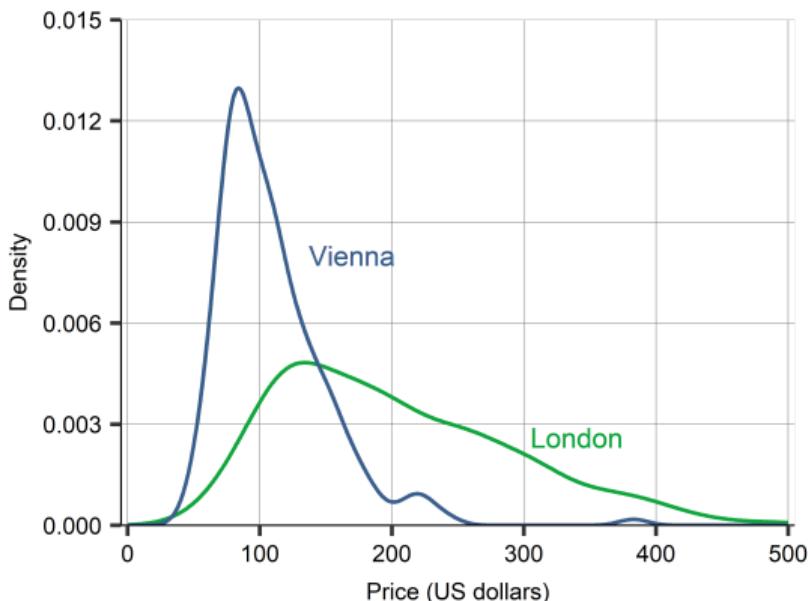


Figure: London, UK



## The density plot

- ▶ Density plot
- ▶ Less reliable than histogram
- ▶ But key points good be read off
- ▶ Easy when comparison



## Case study hotels: descriptive statistics

Table: Descriptive statistics for hotel prices in two cities.

City	N	Mean	Median	Min	Max	Std	Skew
London	435	202.36	186	49	491	88.13	0.186
Vienna	207	109.98	100	50	383	42.22	0.236

Source: hotels-europe dataset. Vienna and London, weekday, November 2017

## Vienna vs London

- ▶ Compare two cities, how hotel markets vary
- ▶ Graphical vs comparison table - **Advantage / disadvantage?**
- ▶ Both help define key messages: (1) describe and (2) explain/make sense.
  - ▶ Hotel prices in London tend to be substantially higher on average.
  - ▶ London prices are also more spread, with a minimum close to the Vienna minimum, but many hotels above 200 dollars
  - ▶ These together imply that there are many hotels in London with a price comparable to hotel prices in Vienna, but there are also many hotels with substantially higher prices

# Distributions

# Theoretical distributions

Theoretical distributions are distributions of variables with idealized properties.

- ▶ Show frequencies for theoretical distributions and not for empirical ones.
- ▶ The likelihood of each value in a more abstract setting - hypothetical "dataset" or "population," or the abstract space of the possible realizations of events.
- ▶ Theoretical distributions are fully captured by few *parameters*: these are statistics determine the whole distributions

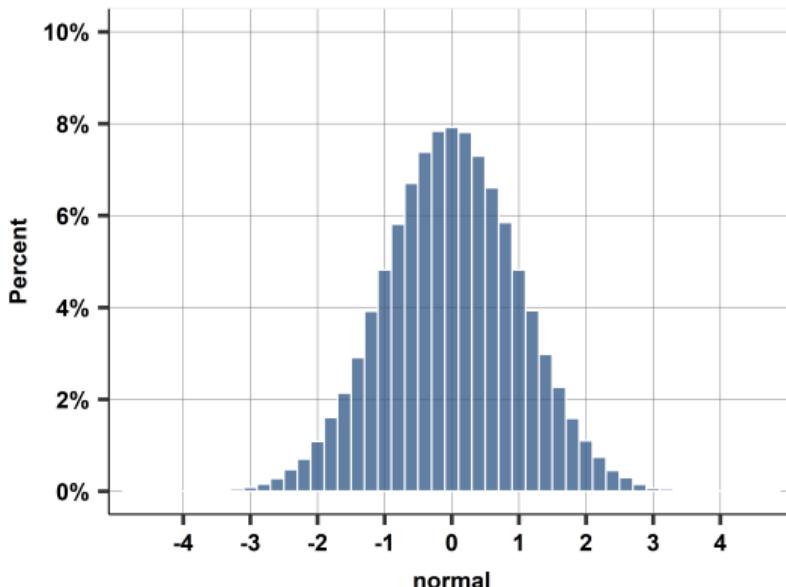
# Theoretical distributions

Theoretical distributions can be helpful

- ▶ Have well-known properties!
- ▶ If variable in our data well approximated by a theoretical distribution → attribute properties to the variable
- ▶ Real life, many variables surprisingly close to theoretical distributions.
- ▶ Will be useful when generalizing from data - **Class 05**

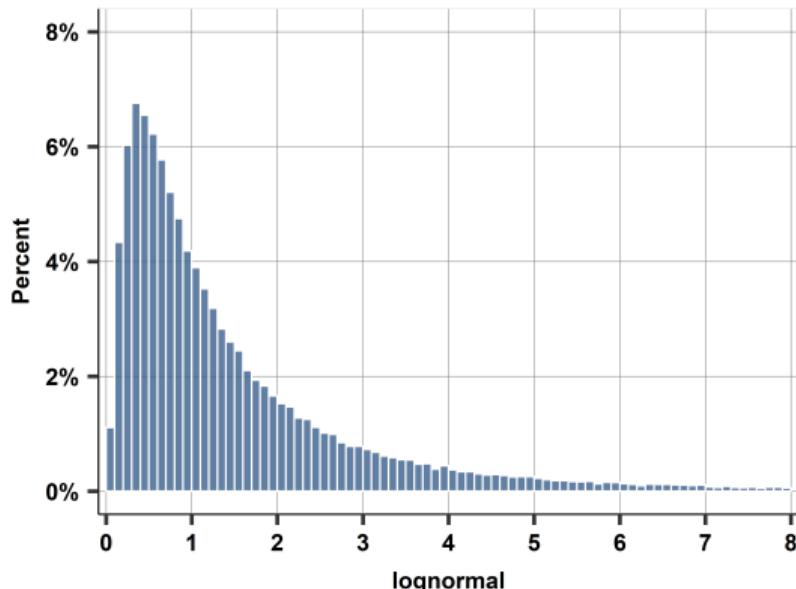
# The Normal distribution

- ▶ Histogram is bell-shaped
- ▶ Outcome (event), can take any value
- ▶ Distribution is captured by two parameters
  - ▶  $\mu$  is the mean
  - ▶  $\sigma$  the standard deviation
- ▶ Symmetric = median, mean (and mode) are the same.
- ▶ Example: height of people, IQs, ect.



# The log-normal distribution

- ▶ Asymmetrically distributed with long right tails.
- ▶ start from a normally distributed RV ( $x$ ), transform it:  $(e^x)$  and the resulting variable is distributed log-normal.
- ▶ Always non-negative
- ▶ Example distributions of income, or firm size.



## A few more points on the Normal and log-normal

- ▶ Many many variables in real life are close to normal
- ▶ Especially when based on elementary things which are added up
- ▶ Not good approximation when
  - ▶ some reasons for non-symmetry
  - ▶ extreme values are important
- ▶ Variables are well approximated by the log-normal if they are the result of many things *multiplied* (the natural log of them is thus a sum).

# Income and log-income

Figure: income

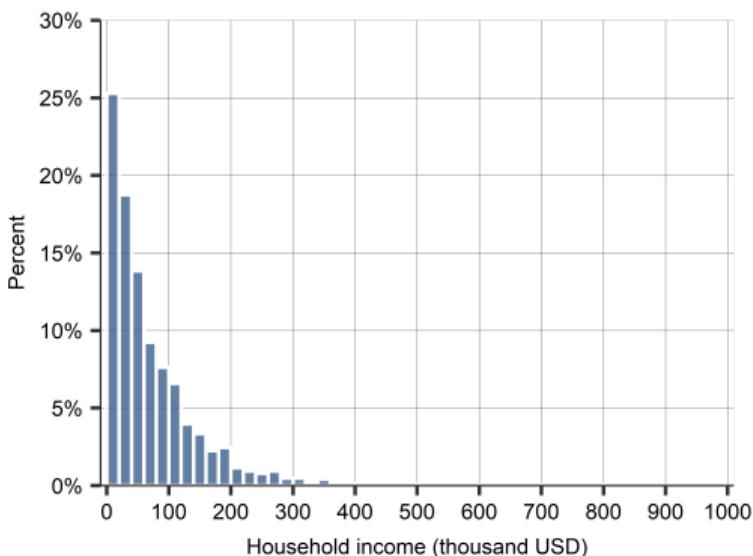
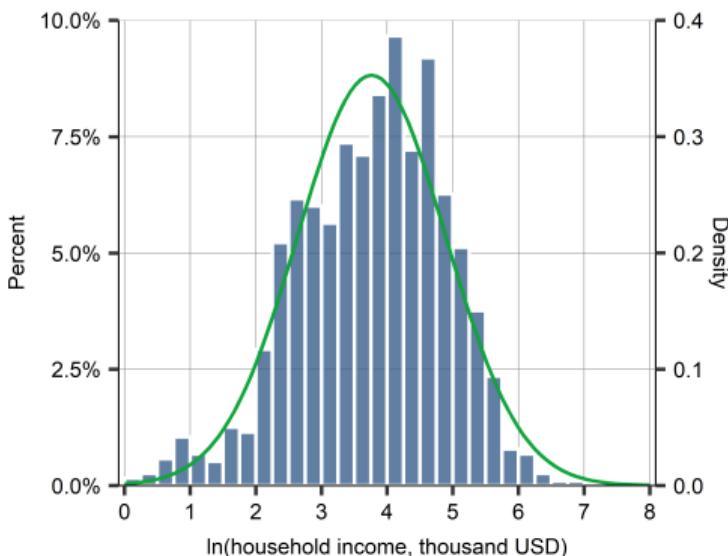
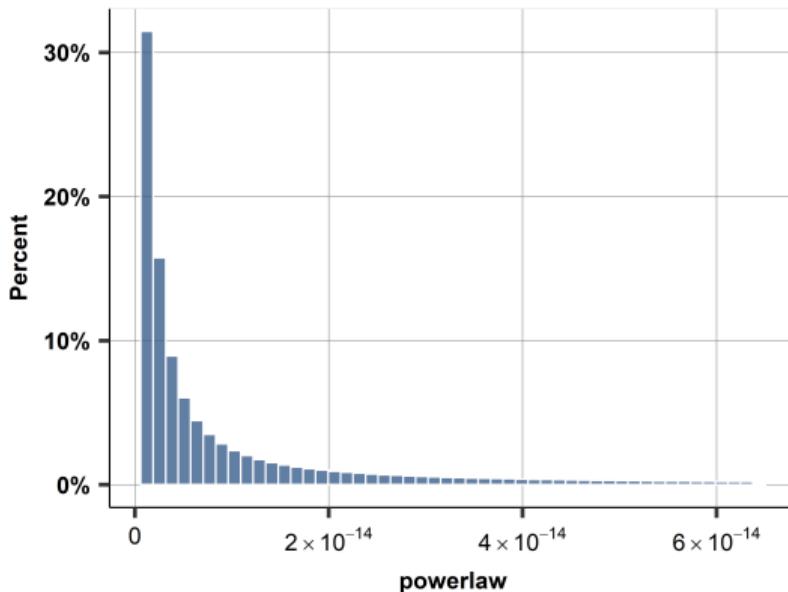


Figure: log income



# The power law distribution

- ▶ Also called as Pareto distribution
- ▶ Very large extreme values - well approximated
- ▶ Relative frequency of close-by values are the same along large and small values
- ▶ Real world: many examples, but often not the whole distribution
- ▶ Example: frequency of words, city population, wealth



# Data vizualization

## Data visualization: Steps

- ▶ We shall make conscious decisions and not let default settings guide us.
- ▶ Usage - what you want to show and to whom – deciding on purpose, focus, and audience
- ▶ Pick a geometric object – decide how information is conveyed: we need to choose a geometric object to visualize the information we want to show.
- ▶ Encode information – choose details of the object (color, height)
- ▶ Settle on scaffolding: – supporting features of the graph such as axes, labels, and titles.
- ▶

This is a very brief overview, more in Chapter 03

## Data visualization: usage

- ▶ What is the purpose, what message you want to convey and to whom?
- ▶ As a general principle, one graph should convey one message.
- ▶ Be explicit about the purpose of the graph and the target audience: general audience vs specialist
- ▶ For a specialist audience, more complicated graphs are okay.

## Data visualization: geoms and encoding

- ▶ Geometric object: Pick an object suitable for the information to be conveyed. [A line showing value over time.](#)
- ▶ May be one or more geoms. [Dots for years and a trend line](#)
- ▶ Encoding: Pick one encoding only [Position of the line](#) Don't apply different colors or shades

## Data visualization: process

- ▶ Decide on geometric objects, and build graphs from bottom up. Advanced, dataviz experts
- ▶ Decide on a type of graph (such as bar chart), and define its elements (geoms). Top down. Most social science / business.
- ▶ Graph type – Can pick a standard object to convey information [Histogram: bars to show frequency](#)

# Data visualization: scaffolding

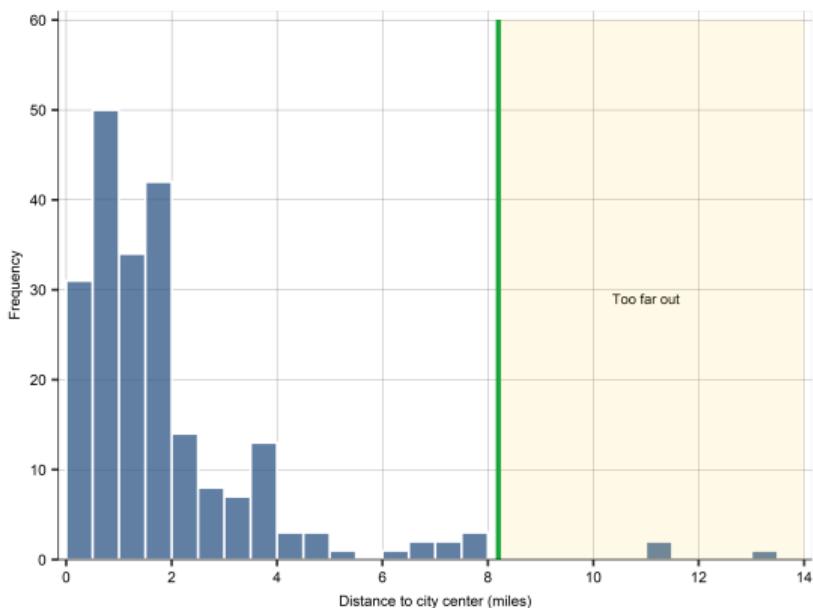
- ▶ How to present elements that support understanding.
- ▶ Make sure, a graph has
  - ▶ Title
  - ▶ Axis title and labels
  - ▶ Legend
- ▶ Content as well as format, such as font type and size.

## Data visualization: annotations

- ▶ if there is something else we want to emphasize.
- ▶ additional information can help put the graph into context or emphasize some part of it
- ▶ Colored area, circled observations, arrow+text, etc

## Data visualization: example

- ▶ Usage: to show distribution for general audience
- ▶ Encoding is bars (histogram), bin size set at 20
- ▶ Axes labelled with title + grid
- ▶ annotation: far away hotels



## Summary steps of EDA

1. First focus on the most important variables. Go back to look at others if subsequent analysis suggests to.
2. For qualitative variables, list relative frequencies.
3. For quantitative variables, look at histograms. May decide for transformation, find extreme values, learn about key aspects of data.
4. Check for extreme values. Decide what to do with them.
5. Look at summary statistics. It may prompt actions, such as focusing on some part of the dataset.
6. Do further exploration if necessary (time series data, comparisons across groups of observations, etc.)

## Content for exam

- ▶ In class, we'll not cover the whole chapter.
- ▶ 3.C1, 3.8, 3.C2 Please read at home. I'll assume you have.
- ▶ We cover 3.9, but make sure you also go through it carefully. I'll assume you have.
- ▶ 3.U1. Please read it. Not part of exam, but very good to know. Especially if more metrics.

Here is the chatgpt link <https://chat.openai.com/c/62f67c7d-62e0-463a-9576-d307fb0>

# 04 Comparison and correlation

Gábor Békés

Data Analysis 1: Exploration

2023

y and x  
ooooooo

A1  
ooooooo

A2  
ooo

Quantitative  
o

A3  
oooooooooooo

Stat. dependence  
oooooooooooo

A4  
oo

The score  
ooooooo

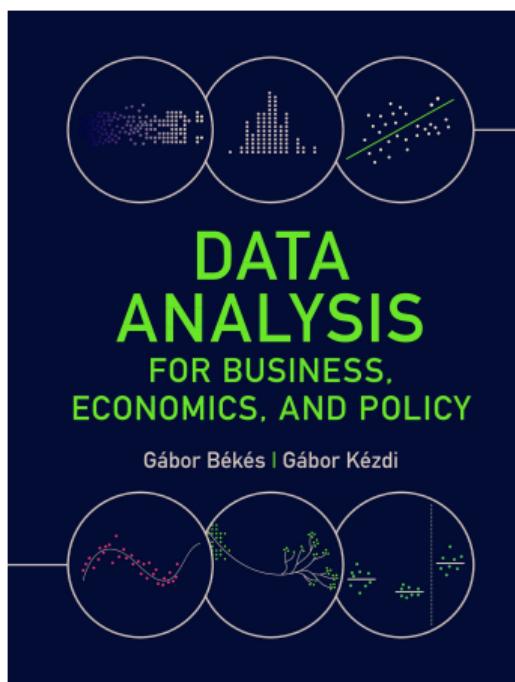
A5  
o

Variation in x  
ooooooo

Sum  
o

Logs  
ooooooo

# Slideshow for the Békés-Kézdi Data Analysis textbook



- ▶ Cambridge University Press, 2021
- ▶ [gabors-data-analysis.com](http://gabors-data-analysis.com)
  - ▶ Download all data and code:  
[gabors-data-analysis.com/data-and-code/](http://gabors-data-analysis.com/data-and-code/)
- ▶ This slideshow is for Chapter 04

y and x oooooooo	A1 ooooooo	A2 ooo	Quantitative o	A3 oooooooooooo	Stat. dependence oooooooooooo	A4 oo	The score oooooo	A5 o	Variation in x oooooo	Sum o	Logs ooooooo
---------------------	---------------	-----------	-------------------	--------------------	----------------------------------	----------	---------------------	---------	--------------------------	----------	-----------------

## Motivation

*Are larger companies better managed? Answering this question may help in benchmarking management practices in a specific company, assessing the value of a company, or estimating the potential benefits of a merger between two companies.*

*To answer this question you downloaded data from the World Management Survey. How should you use the data to measure firm size and the quality of management? How should you assess whether larger companies are better managed?*

y and x	A1 ●oooooo	A2 ooooo	Quantitative o	A3 oooooooooooo	Stat. dependence oooooooooooo	A4 oo	The score oooooo	A5 o	Variation in x oooooo	Sum o	Logs ooooooo
---------	---------------	-------------	-------------------	--------------------	----------------------------------	----------	---------------------	---------	--------------------------	----------	-----------------

## The y and the x

- ▶ Much of data analysis is built on comparing values of a  $y$  variable by values of an  $x$  variable, or more  $x$  variables.
- ▶ Uncover the patterns of association: whether and how observations with particular values of one variable ( $x$ ) tend have particular values of the other variable ( $y$ ).
- ▶ The role of  $y$  is different from the role of  $x$ .
  - ▶ it's the values of  $y$  we are interested in
  - ▶ compare observations that are different in their  $x$  values.
- ▶ It is our decision to pick  $y$

y and x	A1 ○●○○○○	A2 ○○○	Quantitative ○	A3 ○○○○○○○○○○○○	Stat. dependence ○○○○○○○○○○○○	A4 ○○	The score ○○○○○	A5 ○	Variation in x ○○○○○	Sum ○	Logs ○○○○○○
---------	--------------	-----------	-------------------	--------------------	----------------------------------	----------	--------------------	---------	-------------------------	----------	----------------

## The y and the x

- ▶ This asymmetry comes from the goal of our analysis.
- ▶ Goal 1: predicting the value of a  $y$  variable with the help of other variables - many  $x$  variables, such as  $x_1, x_2, \dots$
- ▶ The prediction itself takes place when we know the values of those other variables but not the  $y$  variable.
- ▶ Goal 2: learn about the effect of a causal variable  $x$  on an outcome variable  $y$ .
- ▶ What the value of  $y$  would be if we could change  $x$

y and x	A1 oo•ooo	A2 ooo	Quantitative o	A3 oooooooooooo	Stat. dependence oooooooooooo	A4 oo	The score oooooo	A5 o	Variation in x oooooo	Sum o	Logs ooooooo
---------	--------------	-----------	-------------------	--------------------	----------------------------------	----------	---------------------	---------	--------------------------	----------	-----------------

# Conditioning, conditional distributions

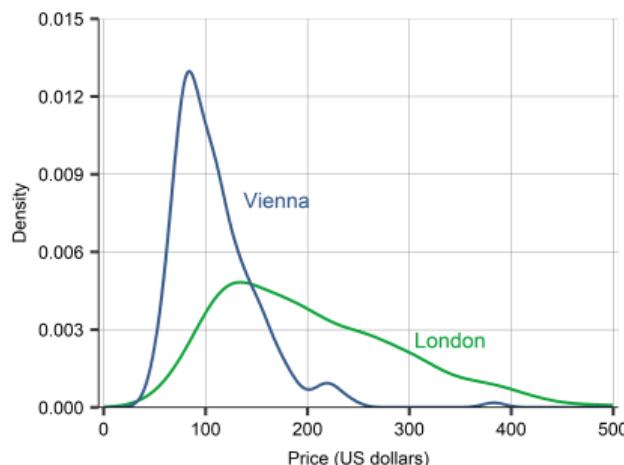
y and x	A1 ○○○●○○	A2 ○○○	Quantitative ○	A3 ○○○○○○○○○○○○	Stat. dependence ○○○○○○○○○○○○	A4 ○○	The score ○○○○○	A5 ○	Variation in x ○○○○○	Sum ○	Logs ○○○○○
---------	--------------	-----------	-------------------	--------------------	----------------------------------	----------	--------------------	---------	-------------------------	----------	---------------

## Comparison and conditioning

- ▶ Comparison → conditioning
- ▶ We compare  $y$ , by values of  $x$  → we condition  $y$  on  $x$ .
  - ▶  $x$  (by the values of which we make comparisons) → conditioning variable.
  - ▶  $y$  → outcome variable.
- ▶ Compare salaries of workers ( $y$ ) with low and high level of education ( $x$ ) →
  - ▶ salary is the outcome
  - ▶ education is the conditioning variable.

## Comparisons and conditional distributions

- ▶ The conditional distribution of a variable is the distribution of the outcome variable given the conditioning variable.
  - ▶ Straightforward concept if the conditioning variable is qualitative (simple if binary)
  - ▶ Comparing distributions
  - ▶ Conditional distribution of prices ( $y$ , conditional on  $x$  - the city=Vienna and city=London)



## Conditional statistic

- ▶ Conditional mean = mean of a variable for each value of the conditioning variable.
- ▶ The conditional expectation of variable  $y$  for different values of variable  $x$  is

$$E[y|x]$$

- ▶ This is a function: for a value of  $x$ , the conditional expectation gives number that is the expected value (mean, average) of variable  $y$  for observations that have that  $x$  value
- ▶ It gives different values based on the conditioning variable  $x$

## Case Study - Management quality and firm size

- ▶ Management quality and firm size: describing patterns of association
- ▶ Whether, and to what extent, larger firms are better managed.
- ▶ Answering this question can help understand why some firms are better managed than others.
  - ▶ Size itself may be a cause ....
  - ▶ Whether firm size is an important determinant of better management can inform policy questions such ...
- ▶ Data from the World Management Survey to investigate our question.

y and x	A1	A2	Quantitative	A3	Stat. dependence	A4	The score	A5	Variation in x	Sum	Logs
oooooooo	oooooo	ooo	o	oooooooooooo	oooooooooooo	oo	oooooo	o	oooooo	o	oooooo

## Case Study - Management quality and firm size

- ▶ Interviews by CEO/senior managers, based on that a score is given
- ▶ Management quality is measured as management score.
- ▶ Each score is an assessment by the survey interviewers of management practices in a particular domain
  - ▶ tracking and reviewing performance or
  - ▶ time horizon and breadth of targets, etc
- ▶ Measured on a scale of 1 (worst practice) to 5 (best practice).
- ▶ Normalized - standardized score

y and x	A1	A2	Quantitative	A3	Stat. dependence	A4	The score	A5	Variation in x	Sum	Logs
oooooooo	oo•ooo	ooo	o	oooooooooooo	oooooooooooo	oo	ooooooo	o	oooooo	o	ooooooo

## Case Study - Management quality and firm size

- ▶ Take 18 individual measures and average
- ▶ Our measure of the quality of management is the simple average of these 18 scores = "the" management score.
- ▶ By construction, the range of the management score is between 1 and 5 because.

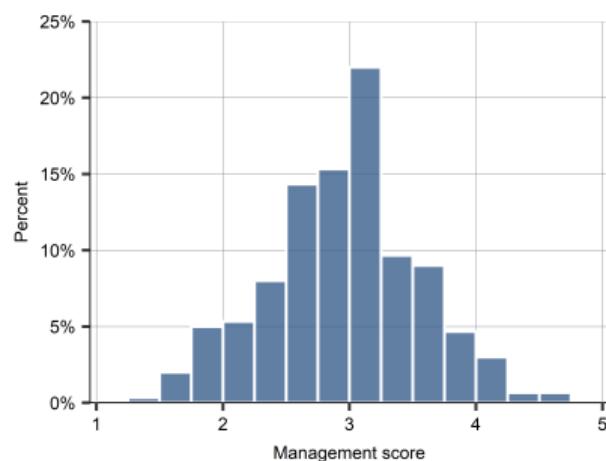
## Case Study - Management quality and firm size

- ▶ Data from the World Management Survey to investigate our question.
- ▶ Survey introduced in class 01
- ▶ In this case study we analyze a cross-section of Mexican firms from the 2013 wave of the survey.
- ▶ Only firms with 100 – 5000 employees, N=300
- ▶ The  $y$  = measure of the quality of management. The  $x$  = measure of firm size.
- ▶ Firm size = number of employees

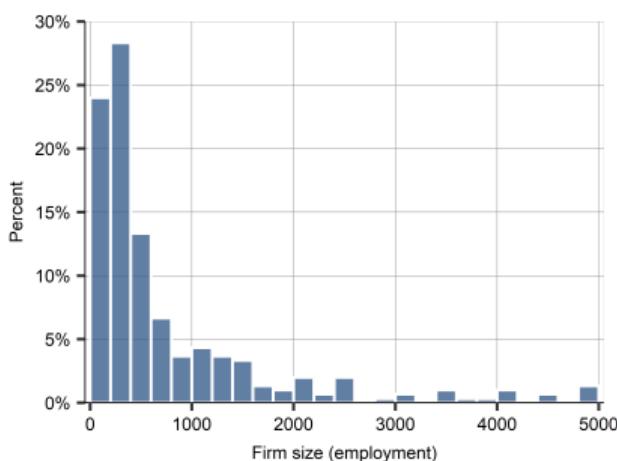
y and x  
ooooooA1  
oooo●●A2  
oooQuantitative  
oA3  
ooooooooooooStat. dependence  
ooooooooooooA4  
ooThe score  
oooooooA5  
oVariation in x  
ooooooSum  
oLogs  
ooooooo

## Case Study - Management quality and firm size

(a) Management score



(b) Firm size (number of employees)



Note: Source: Management quality is an average score of 18 variables. Firm size is number of employees. wms-management-survey data. Mexican sample,  $n=300$ .

y and x  
ooooooA1  
oooooo●A2  
oooQuantitative  
oA3  
ooooooooooooStat. dependence  
ooooooooooooA4  
ooThe score  
ooooooA5  
oVariation in x  
ooooooSum  
oLogs  
oooooo

## Case Study - Management quality and firm size

- ▶ Management score: The mean is 2.9, the median is 3, and the standard deviation is 0.7.
- ▶ Firm size: The range of employment is 100 to 5000. The mean is 760 and the median is 350, skewness with a long right tail. Some large firms, but not extreme, kept as is.

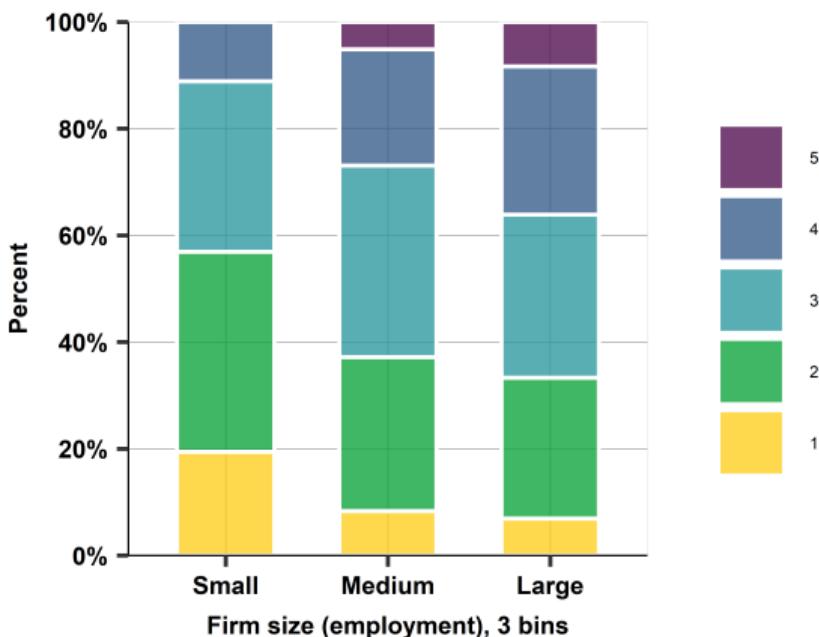
## Case Study - Management quality and firm size

- ▶ Conditional probabilities
- ▶ Three bins of firm size. By number of employees: small (100–199, N=72), medium (200–999, N=156), large (1000, N=72)
- ▶ Take a single measure: Lean management score, with values 1,2,3,4,5.
- ▶ Thus, for each score variable we have 15 conditional probabilities: the probability of each of the 5 values of  $y$  by each of the three values of  $x$  – e.g.,  
 $P(y = 1|x = \text{small})$ .

## Case Study - Management quality and firm size

- ▶ Lean management score 1–5
- ▶ Firm size: small, medium, large
- ▶ Conditional probability:
  - ▶ share of score=1 conditional on being a small firm is about 20%.
  - ▶ share of score=5 conditional on being a large firm is about 10%.
- ▶ Shows a pattern of association

Note: Source: *Management quality is an average score of 18 variables. Firm size is number of employees. wms-management-survey data. Mexican sample, n=300.*



y and x	A1 oooooo	A2 oo●○	Quantitative	A3 oooooooooooo	Stat. dependence oooooooooooo	A4 oo	The score oooooo	A5 o	Variation in x oooooo	Sum o	Logs oooooo
---------	--------------	------------	--------------	--------------------	----------------------------------	----------	---------------------	---------	--------------------------	----------	----------------

## Case Study - Management quality and firm size

- ▶ Conditional probabilities - can calculate average
- ▶ Three bins of employment: small (100–199, N=72), medium (200–999, N=156), large (1000, N=72)
- ▶ Mean management score is 2.68 for small firms, 2.94 for medium sized ones, and it is 3.18 for large.
- ▶ First simple evidence: larger firms have better management.

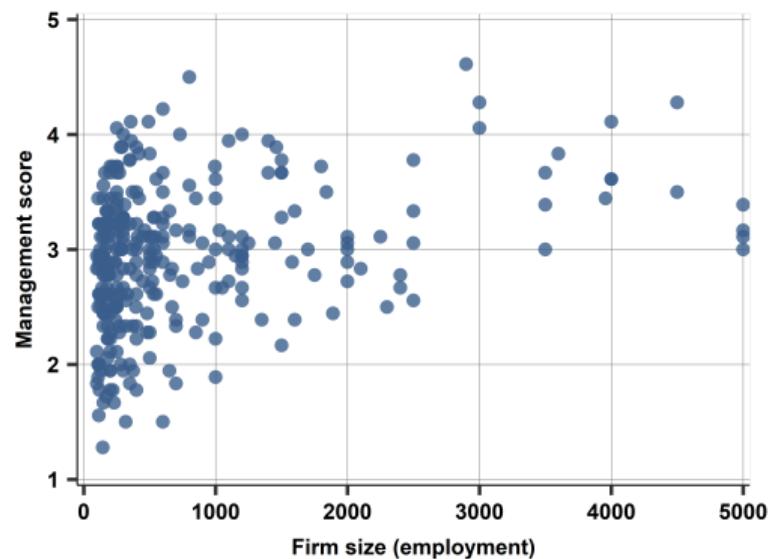
## Conditional and joint distributions of two quantitative variables

- ▶ Two variables, many values
- ▶ The joint distribution of two variables shows the probabilities (frequencies) of each value combination of the two variables.
- ▶ A scatterplot is a two-dimensional graph with the values of each of the two variables measured on its two axes, and dots entered for each observation in the dataset with the combination of the values of the two variables.
- ▶ Works when dataset relatively small.
- ▶ For larger samples, we can bin values, and use "bin scatter"
- ▶ Bin scatter shows conditional means for bins we created

## Case Study - Management quality and firm size

- ▶ Conditional mean and joint distribution
- ▶ How our management quality variable ( $y$ : the management score) is related to our firm size variable ( $x$ : employment)
- ▶ Scatterplot
- ▶ Bin-scatter

## Case Study - Management quality and firm size



- ▶ Scatterplot
- ▶ Both x and y axis qualitative
- ▶ Each dot is an observation
- ▶ Full information on distributions

Note: Source: Management quality is an average score of 18 variables. Firm size is number of employees. wms-management-survey data. Mexican sample, n=300.

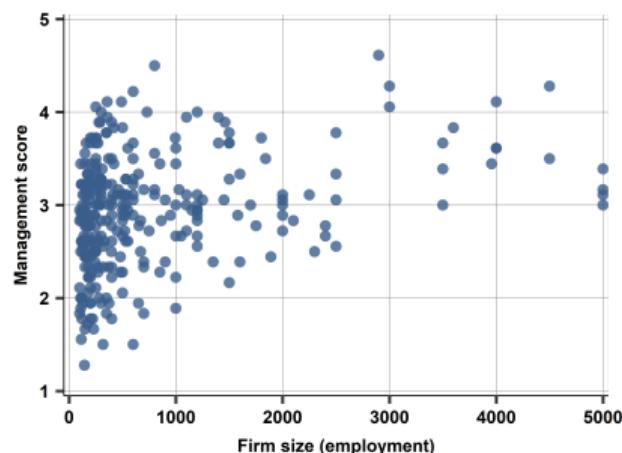
y and x	A1 oooooo	A2 ooo	Quantitative o	A3 oo•oooooooo	Stat. dependence oooooooooo	A4 oo	The score oooooo	A5 o	Variation in x oooooo	Sum o	Logs oooooo
---------	--------------	-----------	-------------------	-------------------	--------------------------------	----------	---------------------	---------	--------------------------	----------	----------------

## Case Study - Management quality and firm size

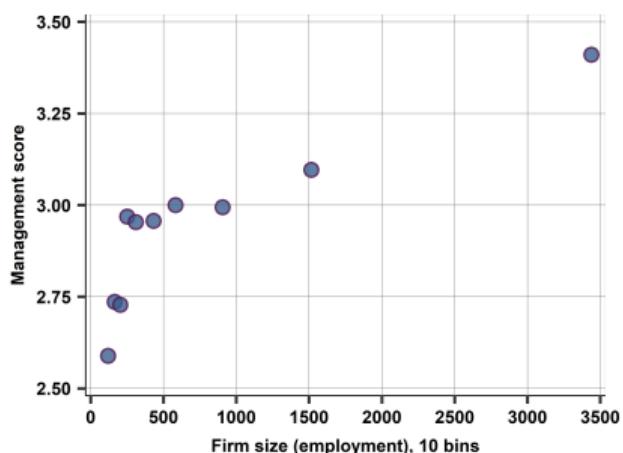
- ▶ Bin-scatter mean  $y$  conditional on three  $x$  bins and ten  $x$  bins.
- ▶ Bin-scatter: cut the distribution into 10 parts, with equal number of firms
- ▶ Show average management score as a point corresponding to the midpoint in the employment bin (e.g., 120 for the 100–120 bin).
- ▶ Dots NOT equally spread out - more frequent where more observations!

## Case Study - Management quality and firm size

(a) Scatterplot



(b) 10 Bin-scatter



Note: Source: Management quality is an average score of 18 variables. Firm size is number of employees. wms-management-survey data. Mexican sample, n=300.

y and x	A1 oooooo	A2 ooo	Quantitative o	A3 oooo●oooo	Stat. dependence oooooooo	A4 oo	The score oooooo	A5 o	Variation in x oooooo	Sum o	Logs oooooo
---------	--------------	-----------	-------------------	-----------------	------------------------------	----------	---------------------	---------	--------------------------	----------	----------------

## Case Study - Management quality and firm size

- ▶ Some positive association is shown, but not easy to read
- ▶ Bin-scatter - positive overall, but most for small vs medium.
- ▶ Difference in mean management quality tends to be smaller when comparing bins of larger size, suggesting a positive but nonlinear, concave pattern of association
  - ▶ (a positive concave function increases at a decreasing rate)

y and x	A1 ooooooo	A2 ooo	Quantitative o	A3 oooooo•oooo	Stat. dependence oooooooooo	A4 oo	The score oooooo	A5 o	Variation in x oooooo	Sum o	Logs ooooooo
---------	---------------	-----------	-------------------	-------------------	--------------------------------	----------	---------------------	---------	--------------------------	----------	-----------------

## Boxplot and violinplot

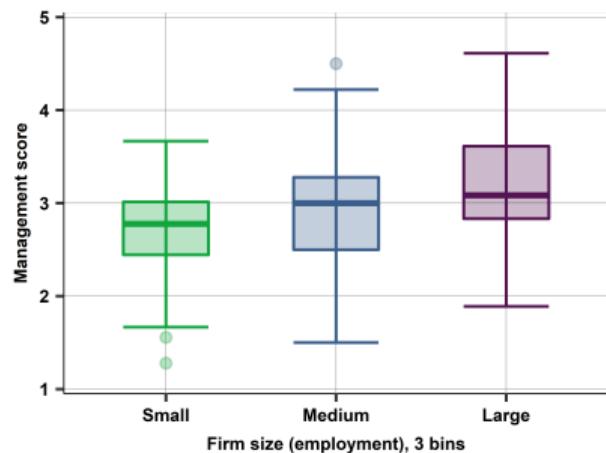
Boxplot: shows means and some features of the distribution.

- ▶ Median value of the variable, placed within a box.
- ▶ The upper side of the box is the third quartile (the 75<sup>th</sup> percentile) and the lower side is the first quartile (the 25<sup>th</sup> percentile).
- ▶ additional info: 1.5 times the inter-quartile range added to the third quartile and subtracted from the first quartile.
- ▶ may show individual values outside

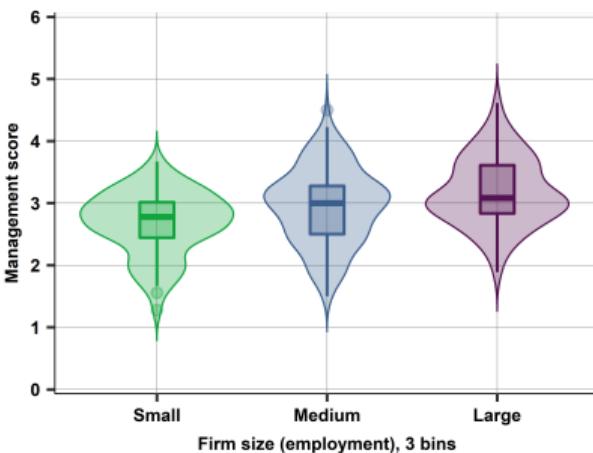
Violin plot - similar but shows density plot (kernel density as seen before)

## Case Study - Management quality and firm size

(a) Box plot



(b) Violin plot



Note: *Employee retention rates: The probability of staying with the firm, in the two experimental groups. Mean denoted in boxplot. Source: wms-management dataset*

y and x	A1 ooooooo	A2 ooo	Quantitative o	A3 oooooooo●oo	Stat. dependence oooooooooo	A4 oo	The score oooooo	A5 o	Variation in x oooooo	Sum o	Logs ooooooo
---------	---------------	-----------	-------------------	-------------------	--------------------------------	----------	---------------------	---------	--------------------------	----------	-----------------

## Case Study - Management quality and firm size

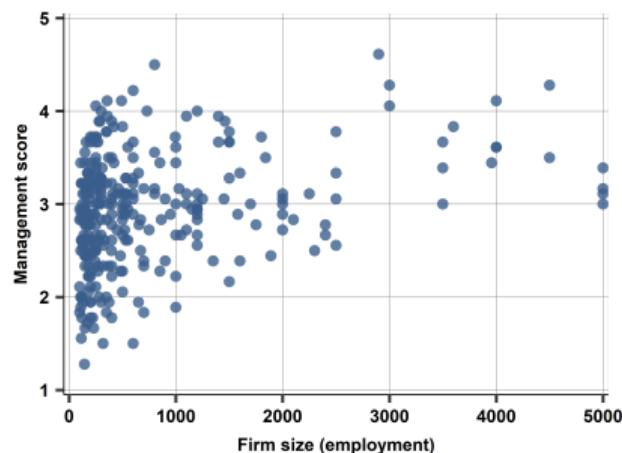
- ▶ Can look at differences along distributions to see role of potential skewness, extreme values
- ▶ Boxplot and violin plot by three bins for the latter
- ▶ show conditional distribution (conditional on  $x$  being in the actual bin)
- ▶ Both the box plots and the violin plots reveal that the median management score is higher in larger firms, reflecting the same positive association as the bin scatters and the scatterplot.
- ▶ That positive pattern is true when we compare almost any statistic of the management score: median, upper and lower quartiles, minimum and maximum.
- ▶ These figures also show that the spread of management score is somewhat smaller in smaller firms, but that difference in spread appears small.

## Case Study - Management quality and firm size - detour

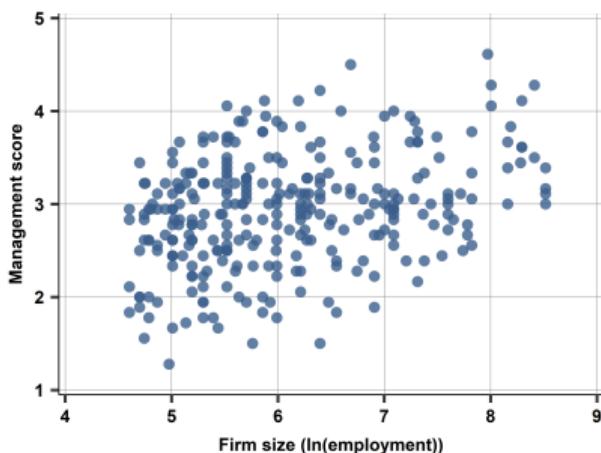
- ▶ To make the patterns more visible we apply a transformation
- ▶ The idea here is that the distribution of employment is very skewed, closer to lognormal than normal, so we take the natural logarithm of employment.
- ▶ Stretching the employment differences between firms at lower levels of employment and compressing those differences at higher levels.
- ▶ This scatterplot leads to a more spread out picture, reflecting the more symmetric distribution of the x variable.
- ▶ Here the positive association between mean management score and (log) employment is more visible.

## Case Study - Management quality and firm size

(a) Scatterplot



(b) Scatterplot with Log units



Note: Source: *Management quality is an average score of 18 variables. Firm size is number of employees. wms-management-survey data. Mexican sample, n=300.*

# Statistical dependence, correlation

y and x	A1 oooooo	A2 ooo	Quantitative o	A3 oooooooooooo	Stat. dependence o●oooooooooo	A4 oo	The score oooooo	A5 o	Variation in x oooooo	Sum o	Logs oooooo
---------	--------------	-----------	-------------------	--------------------	----------------------------------	----------	---------------------	---------	--------------------------	----------	----------------

## Dependence and independence

- ▶ Dependence of two variables -  $y$  and  $x$  means that the conditional distributions of  $y$  - conditional on  $x$  - are not the same ( $x$  is the conditioning variable).
- ▶ Independence of  $y$  and  $x$  means the opposite: the distribution of  $y$  on  $x$  is the same, regardless of the value of  $x$ .
- ▶ Dependence of  $y$  and  $x$ , may take many forms.
  - ▶ When the value of  $x$  is different,  $y$  may be more or less spread out
  - ▶ when the value of  $x$  is different the mean of  $y$  is different.

y and x	A1	A2	Quantitative	A3	Stat. dependence	A4	The score	A5	Variation in x	Sum	Logs
oooooooo	ooooooo	ooo	o	oooooooooooo	oo●oooooooo	oo	oooooo	o	oooooo	o	ooooooo

## Mean dependence

- ▶ Mean-dependence: conditional expectation  $E[y|x]$  varies with the value of  $x$ .
- ▶ Mean-dependence is the extent to which conditional expectations (means) differ.
- ▶ Two variables are positively mean-dependent if the average of one variable tends to be larger when the value of the other variable is larger, too.
- ▶ Covariance and Correlation Coefficient are measures of mean dependence.

y and x	A1 ooooooo	A2 ooo	Quantitative o	A3 oooooooooooo	Stat. dependence ooo•oooooo	A4 oo	The score oooooo	A5 o	Variation in x oooooo	Sum o	Logs ooooooo
---------	---------------	-----------	-------------------	--------------------	--------------------------------	----------	---------------------	---------	--------------------------	----------	-----------------

## Covariance

The formula for the covariance between two variables  $x$  and  $y$  both observed in a dataset with  $n$  observations is:

$$\text{Cov}[x, y] = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n} \quad (1)$$

- ▶ for each observation  $i = 1 \dots n$
- ▶ The product within the sum in the numerator multiplies the deviation of  $x$  from its mean ( $x_i - \bar{x}$ ) with the deviation of  $y$  from its mean ( $y_i - \bar{y}$ )
- ▶ The entire formula is the average of these products across all observations.

## Covariance

$$Cov[x, y] = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n}$$

- ▶ If a positive deviation of  $x$  from its mean goes with a positive deviation of  $y$  from its mean the product is positive. Thus, the average of this product across all observations is positive.
- ▶ The more often a positive  $x_i - \bar{x}$  goes together with a positive  $y_i - \bar{y}$  the larger positive is the covariance.
- ▶ Or, the larger are the positive deviations that go together the larger the covariance.

y and x	A1 oooooo	A2 ooo	Quantitative o	A3 oooooooooooo	Stat. dependence oooooo•oooo	A4 oo	The score oooooo	A5 o	Variation in x oooooo	Sum o	Logs oooooo
---------	--------------	-----------	-------------------	--------------------	---------------------------------	----------	---------------------	---------	--------------------------	----------	----------------

## The correlation coefficient

$$\text{Corr}[x, y] = \frac{\text{Cov}[x, y]}{\text{Std}[x]\text{Std}[y]} \quad (2)$$

$$-1 \leq \text{Corr}[x, y] \leq 1 \quad (3)$$

- ▶ The correlation coefficient is the standardized version of the covariance.
- ▶ The covariance may be any positive or negative number, while the correlation coefficient is bound to be between negative one and positive one.

y and x	A1 oooooo	A2 ooo	Quantitative o	A3 oooooooooooo	Stat. dependence ooooooo●ooo	A4 oo	The score oooooo	A5 o	Variation in x oooooo	Sum o	Logs oooooo
---------	--------------	-----------	-------------------	--------------------	---------------------------------	----------	---------------------	---------	--------------------------	----------	----------------

## Dependence, mean-dependence, correlation

- ▶ If two variables are independent, they are also mean-independent and thus the conditional expectations are all the same:  $E[y|x] = E[y]$  of any value of  $x$ .
- ▶ Is this true the other way around?

y and x	A1 oooooo	A2 ooo	Quantitative o	A3 oooooooooooo	Stat. dependence oooooooo●○	A4 oo	The score oooooo	A5 o	Variation in x oooooo	Sum o	Logs oooooo
---------	--------------	-----------	-------------------	--------------------	--------------------------------	----------	---------------------	---------	--------------------------	----------	----------------

## Dependence, mean-dependence, correlation

- ▶ If two variables are independent, they are also mean-independent and thus the conditional expectations are all the same:  $E[y|x] = E[y]$  of any value of  $x$ .
- ▶ But the reverse is not true.
- ▶ Can have zero correlation but mean dependence (e.g., a symmetrical U-shaped conditional expectation has an average of zero),
- ▶ Can have zero correlation and zero mean dependence without complete independence (e.g., the spread of  $y$  may be different for different values of  $x$ ).

y and x	A1 ooooooo	A2 ooo	Quantitative o	A3 oooooooooooo	Stat. dependence oooooooo●	A4 oo	The score oooooo	A5 o	Variation in x oooooo	Sum o	Logs ooooooo
---------	---------------	-----------	-------------------	--------------------	-------------------------------	----------	---------------------	---------	--------------------------	----------	-----------------

## Dependence, mean-dependence, correlation

- ▶ Covariance or the correlation coefficient allow for all kinds of variables, including binary variables and ordered qualitative variables as well as quantitative variables.
- ▶ The covariance and the correlation coefficient will always be zero if the two variables are mean-independent, positive if positively mean-dependent, and negative if negatively mean-dependent.
- ▶ However, they are more appropriate measures for quantitative variables. That's because the differences  $y_i - \bar{y}$  and  $x_i - \bar{x}$  make more sense when  $y$  and  $x$  are quantitative variables.

## Case Study - Management quality and firm size

- ▶ The covariance between firm size and the management score is 177.
- ▶ The standard deviation of firm size is 977, the standard deviation of management score is 0.6.
- ▶ Positive mean-dependence: firm size tends to be higher at firms with better management.
- ▶ the correlation coefficient is  $0.30 (177 / (977 * 0.6))$ .
- ▶ This suggests a positive and moderately strong association.
- ▶ Management quality–firm size correlation varies considerably across industries?

## Case Study - Management quality and firm size

Table: Measures of management quality and their correlation with size by industry

Industry	Management-firm size correlation	Observations
Auto	0.50	26
Chemicals	0.05	69
Electronics	0.33	24
Food, drinks, tobacco	0.05	34
Materials, metals	0.32	50
Textile, apparel	0.29	43
Wood, furniture, paper	0.28	29
Other	0.44	25
All	0.30	300

Note: Employee retention rates: The probability of staying with the firm, in the two experimental groups. Source: working from home dataset.

## Measuring a latent concept with many observed variables

- ▶ Often a concept is hard, even impossible, to measure.
- ▶ Latent variables - while we can think of them as a variable there is no single observed variable to measure them.
- ▶ Quality of management at a firm - it is a concept that may be measured with a collection of variables, not a single one of them
- ▶ IQ - measured by a series of quiz-like questions.
- ▶ The problem here is how to combine multiple observed variables

y and x	A1 ooooooo	A2 ooo	Quantitative o	A3 oooooooooooo	Stat. dependence oooooooooooo	A4 oo	The score o●ooooo	A5 o	Variation in x oooooo	Sum o	Logs ooooooo
---------	---------------	-----------	-------------------	--------------------	----------------------------------	----------	----------------------	---------	--------------------------	----------	-----------------

## Condensing information

If a dataset has more than one variable aimed to measure the same latent variable how should we combine them? Alternatives:

- ▶ Use one observed variable only
- ▶ Take the average (or sum) of all observed variables
- ▶ Use principal component analysis (PCA) to combine all observed variables

y and x	A1 ooooooo	A2 ooo	Quantitative o	A3 oooooooooooo	Stat. dependence oooooooooooo	A4 oo	The score oo•ooo	A5 o	Variation in x oooooo	Sum o	Logs ooooooo
---------	---------------	-----------	-------------------	--------------------	----------------------------------	----------	---------------------	---------	--------------------------	----------	-----------------

## Condensing information 1: Using a single variable

- ▶ Using one measured variable and exclude the rest has the advantage of easy interpretation.
- ▶ It has the disadvantage of discarding potentially useful information contained in the other measured variables.
- ▶ Can be often a sensible start

y and x	A1 ooooooo	A2 ooo	Quantitative o	A3 oooooooooooo	Stat. dependence oooooooooooo	A4 oo	The score ooo•ooo	A5 o	Variation in x oooooo	Sum o	Logs ooooooo
---------	---------------	-----------	-------------------	--------------------	----------------------------------	----------	----------------------	---------	--------------------------	----------	-----------------

## Condensing information 2: Using a sum

- ▶ Taking the average of all measured variables makes use of all information.
- ▶ If all measured using the *same scale* this approach, simple and a natural interpretation
- ▶ When variables measured in different scales, simple average is difficult to interpret and meaningless

## Condensing information 2: Using a sum

- ▶ Taking the average of all measured variables makes use of all information.
- ▶ If all measured using the *same scale* this approach, simple and a natural interpretation
- ▶ When variables measured in different scales, simple average is difficult to interpret and meaningless
- ▶ Need bring it to common scale - standardization: subtracting the mean and dividing with the standard deviation (class 03)
- ▶ The result is a series of variables with zero mean and standard deviation of one
- ▶ This standardized measure is called a "z-score" or "score"

y and x	A1 ooooooo	A2 ooo	Quantitative o	A3 oooooooooooo	Stat. dependence oooooooooooo	A4 oo	The score oooo●o	A5 o	Variation in x ooooooo	Sum o	Logs ooooooo
---------	---------------	-----------	-------------------	--------------------	----------------------------------	----------	---------------------	---------	---------------------------	----------	-----------------

## Condensing information 3 using different weights

- ▶ Principal component analysis (PCA) is a method to give potentially different weights to the observable variables for creating a weighted average.
- ▶ The weights are constructed in such a way that observed variables that are better measures receive higher weights.
- ▶ PCA finds out which observed variable is a better measure by examining how they would be correlated with the weighted average.
- ▶ Bit of black box method

y and x	A1 ooooooo	A2 ooo	Quantitative o	A3 oooooooooooo	Stat. dependence oooooooooooo	A4 oo	The score ooooo●	A5 o	Variation in x oooooo	Sum o	Logs ooooooo
---------	---------------	-----------	-------------------	--------------------	----------------------------------	----------	---------------------	---------	--------------------------	----------	-----------------

## We suggest using the z-score

- ▶ Use z-score - simple average of multiple observed variables after making sure that they are measured on the same scale
- ▶ Simple, easy to understand
- ▶ Transparent
- ▶ Typically marginally different to PCA
- ▶ Pay attention
  - ▶ Look at correlation signs, you may check it first (PCA is better here)
  - ▶ Sensitive to extreme values

## Case Study - Management quality and firm size

- ▶ The latent concept here is the overall quality of management.
- ▶ The observable variables are the 18 "score" variables.
- ▶ Each of the 18 scores were measured on a scale of 1 (worst practice) to 5 (best practice).
- ▶ The overall management score variable was the simple average of these.
- ▶ Looking at correlation of each of the sixteen observed variables with the average score [0.45 to 0.73]. Close, not same.
- ▶ PCA could be slightly better. In practice average score and PCA very similar.

y and x	A1 oooooo	A2 ooo	Quantitative o	A3 oooooooooooo	Stat. dependence oooooooooooo	A4 oo	The score oooooo	A5 o	Variation in x ●oooooo	Sum o	Logs oooooo
---------	--------------	-----------	-------------------	--------------------	----------------------------------	----------	---------------------	---------	---------------------------	----------	----------------

## Comparison and variation in x

- ▶ Variation in the conditioning variable is necessary to make comparisons.
- ▶ If no variation in the conditioning variable
  - ▶ all observations have the same values
  - ▶ impossible to make comparisons
- ▶ Example: Uncover the effect of price changes on sales → need many observations with different price values.
- ▶ Generalization: The more variation is there in the conditioning variable the better are the chances for comparison.

y and x	A1 ooooooo	A2 ooo	Quantitative o	A3 oooooooooooo	Stat. dependence oooooooooooo	A4 oo	The score ooooooo	A5 o	Variation in x o•ooooo	Sum o	Logs ooooooo
---------	---------------	-----------	-------------------	--------------------	----------------------------------	----------	----------------------	---------	---------------------------	----------	-----------------

## Comparison and variation in x

- ▶ source of variation in the conditioning variable
- ▶ why values of the conditioning variable may differ across observations.
- ▶ Option 1: experimental data
- ▶ Option 2: observational data

y and x	A1 ooooooo	A2 ooo	Quantitative o	A3 oooooooooooo	Stat. dependence oooooooooooo	A4 oo	The score oooooo	A5 o	Variation in x oo•ooo	Sum o	Logs ooooooo
---------	---------------	-----------	-------------------	--------------------	----------------------------------	----------	---------------------	---------	--------------------------	----------	-----------------

## Comparison in Experimental data

- ▶ We have an intervention or treatment.
- ▶ Value of the conditioning variable differs across observations because the person running the experiment made them different. Also: treatment variable
- ▶ There is controlled variation - a rule deciding treatment
- ▶ Experiment - comparing one or more outcome variables across the various values of a treatment variable
- ▶ Example: drug trial
  - ▶ Medical experiment - some patients receive the drug while others receive a placebo
  - ▶ Outcome is recovery from the illness or not
  - ▶ Control (treatment) variable is gets the drug or not

y and x	A1 ooooooo	A2 ooo	Quantitative o	A3 oooooooooooo	Stat. dependence oooooooooo	A4 oo	The score oooooo	A5 o	Variation in x ooo•ooo	Sum o	Logs ooooooo
---------	---------------	-----------	-------------------	--------------------	--------------------------------	----------	---------------------	---------	---------------------------	----------	-----------------

## Comparison with observational data

- ▶ Most data used in business, economics and policy analysis are observational.
- ▶ In observational data, no variable is fully controlled.
- ▶ Typical variables in such data are the results of the decisions
- ▶ The source of variation in these variables may have multiple sources
- ▶ People's choices, decisions, interactions, expectations, etc.
- ▶ Compare the value of the outcome variable for different values of the conditioning variable.
- ▶ Much harder interpretation

# Source of variation important for causal analysis

## Experimental data

- ▶ Easy - if conditioning variable is experimentally controlled -
- ▶ Made sure that differences in the outcome variable are due to that variable only
- ▶ Example. Randomly give aspirin vs placebo.
- ▶ Any difference in stroke likelihood is due to treatment

## Observational data

- ▶ Hard - many other things may be different when the value of the conditioning variable differs
- ▶ Example: observe people aspirin taking routine
- ▶ Any difference in stroke likelihood could be for other reasons
- ▶ E.g. Aspirin takers may have chosen to take Aspirin because they experienced a stroke already

y and x	A1 ooooooo	A2 ooo	Quantitative o	A3 oooooooooooo	Stat. dependence oooooooooooo	A4 oo	The score oooooo	A5 o	Variation in x ooooo●	Sum o	Logs ooooooo
---------	---------------	-----------	-------------------	--------------------	----------------------------------	----------	---------------------	---------	--------------------------	----------	-----------------

## AI and patterns

- ▶ AI is great to give you a first review of patterns – similar to a few lines of code, or `panda profiler` in Python
- ▶ judgment of correlation (weak, strong) is often wrong.
- ▶ you need to know what pattern to pursue

*See separate pdf for example*

## Summary

- ▶ Be explicit about what  $y$  and  $x$  are in your data and how they are related to the question of your analysis.  $E[y|x]$  is mean  $y$  conditional on  $x$ .
- ▶ For qualitative variables, correlation can be shown by summarizing conditional probabilities (frequencies).
- ▶ For quantitative variables, scatterplots offer a visual insight to the pattern of the relationship.
- ▶ The correlation coefficient captures a simple measure of mean dependence.

## Functional form: ln transformation

- ▶ Comparsion: Frequent nonlinear patterns better approximated with  $y$  or  $x$  transformed by taking relative differences:
- ▶ In cross-sectional data usually there is no natural base for comparison.
- ▶ Taking the natural logarithm of a variable is often a good solution in such cases.
- ▶ When transformed by taking the natural logarithm, differences in variable values we *approximate relative differences*.
  - ▶ Log differences works because differences in natural logs approximate percentage differences!

## Logarithmic transformation - interpretation

- ▶  $\ln(x)$  = the natural logarithm of  $x$ 
  - ▶ Sometimes we just say  $\log x$  and mean  $\ln(x)$ . Could also mean log of base 10. Here we use  $\ln(x)$
- ▶  $x$  needs to be a positive number
  - ▶  $\ln(0)$  or  $\ln(\text{negative number})$  do not exist
- ▶ Log transformation allows for comparison in relative terms – percentages!

Claim:

$$\ln(x + \Delta x) - \ln(x) \approx \frac{\Delta x}{x}$$

- ▶ The difference between the natural log of two numbers is approximately the relative difference between the two for small differences.

## Logarithmic Functions of $y$ and/or $x$

- ▶  $\ln(x)$  = the natural logarithm of  $x$ 
  - ▶ Sometimes we just say  $\log x$  and mean  $\ln(x)$ . Could also mean log of base 10. Here we use  $\ln(x)$
- ▶  $x$  needs to be a positive number
  - ▶  $\ln(0)$  or  $\ln(\text{negative number})$  do not exist
- ▶ Log transformation allows for comparison in relative terms (percentage), because:

$$\ln(x+y) - \ln(x) = \ln\left(1 + \frac{\Delta x}{x}\right) \approx \frac{\Delta x}{x} \quad (4)$$

- ▶ Numerically:
  - ▶  $\ln(1.01) = 0.0099 \approx 0.01$
  - ▶  $\ln(1.1) = 0.095 \approx 0.1$

## Logarithmic Functions of $y$ and/or $x$

- ▶ Alternatively, from the relationship in calculus:

$$\frac{d\ln(x)}{dx} = \frac{1}{x}$$

- ▶ So that, for small  $\Delta x$ ,

$$\frac{\ln(x + \Delta x) - \ln(x)}{\Delta x} \approx \frac{1}{x}$$

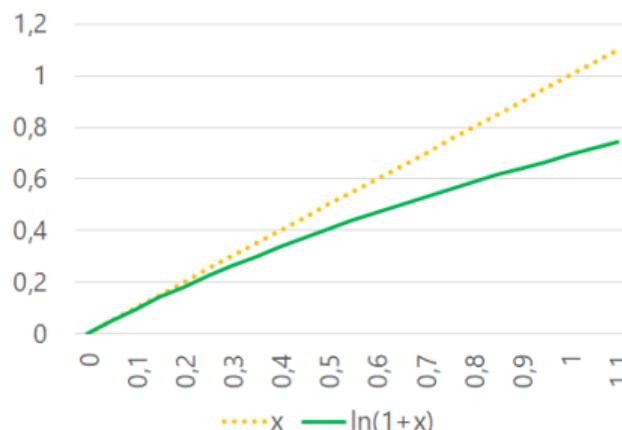
- ▶ And thus

$$\ln(x + \Delta x) - \ln(x) \approx \frac{\Delta x}{x}$$

- ▶ i.e., the difference between the natural log of two numbers is approximately the relative difference between the two
  - ▶ For small differences

# Log Approximation of Small Relative Diffs

- ▶ Log differences approximate small relative differences
- ▶ When  $x$  is small
  - ▶ 0.3 or smaller
  - ▶ the log approximation is close
- ▶ But for larger  $x$ , there is a difference,
  - ▶ And may have to calculate percentage change by hand



## Ln(x) vs percentage

- ▶ Log differences approximate relative differences (percent)
- ▶ log difference - we mean  $\ln(x)$ 
  - ▶ x needs to be a positive number
- ▶ Log differences approximate small relative differences - say below 0.3
  - ▶ A difference of 0.1 log units corresponds to a 10% difference
  - ▶ For larger positive differences, the log difference is smaller
    - ▶ A log difference of +1.0 corresponds to a +170% difference
  - ▶ For larger negative differences, the log difference is larger in absolute value
    - ▶ A log difference of -1.0 corresponds to a -63% difference

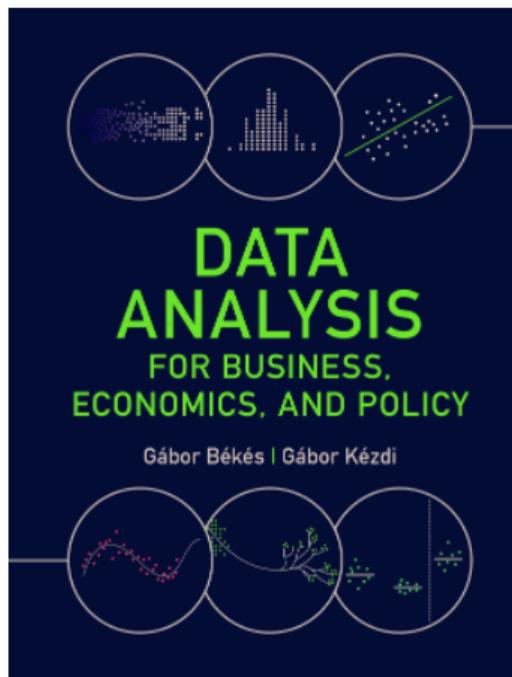
# 05 Generalizing from data

Gábor Békés

Data Analysis 1: Exploration

2023

# Slideshow for the Békés-Kézdi Data Analysis textbook



- ▶ Cambridge University Press, 2021
- ▶ [gabors-data-analysis.com](http://gabors-data-analysis.com)
  - ▶ Download all data and code: [gabors-data-analysis.com/data-and-code/](http://gabors-data-analysis.com/data-and-code/)
- ▶ This slideshow is for Chapter 05

## Generalization

- ▶ Sometimes we analyze a dataset with the goal of learning about patterns in that dataset alone.
- ▶ In such cases there is no need to generalize our findings to other datasets.
- ▶ Example: We search for a good deal among offers of hotels, all we care about are the observations in our dataset.
- ▶ Often we analyze a dataset in order to learn about patterns that may be true in other situations.
- ▶ We are interested in finding it the relationship between
  - ▶ Our dataset
  - ▶ The situation we care about

## Generalization

- ▶ Generalize the results from a single dataset to other situations.
- ▶ The act of generalization is called *inference*: we infer something from our data about a more general phenomenon because we want to use that knowledge in some other situation.
- ▶ Aspect 1: statistical inference
- ▶ Aspect 2: external validity

## Statistical inference

- ▶ Uses statistical methods to make inference.
- ▶ Well-developed and powerful toolbox that helps generalizing to situations similar to our data.
- ▶ Similar to ours = general pattern represented by our dataset.
- ▶ The general pattern is an abstract thing that may or may not exist.
- ▶ If we can assume that the general pattern exists, the tools of statistical inference can be very helpful.

## General patterns 1: Population and representative sample

- ▶ The cleanest example of representative data is a representative sample of a well-defined *population*.
- ▶ A sample is representative of a population if the distribution of all variables is very similar in the sample and the population.
- ▶ Random sampling is the best way to achieve a representative sample.

## General patterns 2: No population but general pattern

The concept of representation is less straightforward in other setups.

- ▶ Using data with observations from the past to uncover a pattern that may be true for the future.
- ▶ Generalizing patterns observed among some products to other, similar products.

There isn't necessarily a "population" from which a random sample was drawn on purpose. Instead, we should think of our data as one that represents a general pattern.

- ▶ There is a general pattern, each year is a random realization.
- ▶ There is a general pattern, each product is a random version, all represented by the same general pattern.

## External validity

- ▶ Assessing whether our data represents the same general pattern that would be relevant for the situation we truly care about.
- ▶ Externally valid case: the situation we care about and the data we have represent the same general pattern
- ▶ With external validity, our data can tell what to expect.
- ▶ No external validity: whatever we learn from our data, may turn out to be not relevant at all.

# The process of inference

The process of inference

1. Consider a statistic we may care about, such as the mean.
2. Compute its *estimated value* from a dataset
3. Infer the value in the population / in the general pattern, that our data represents.

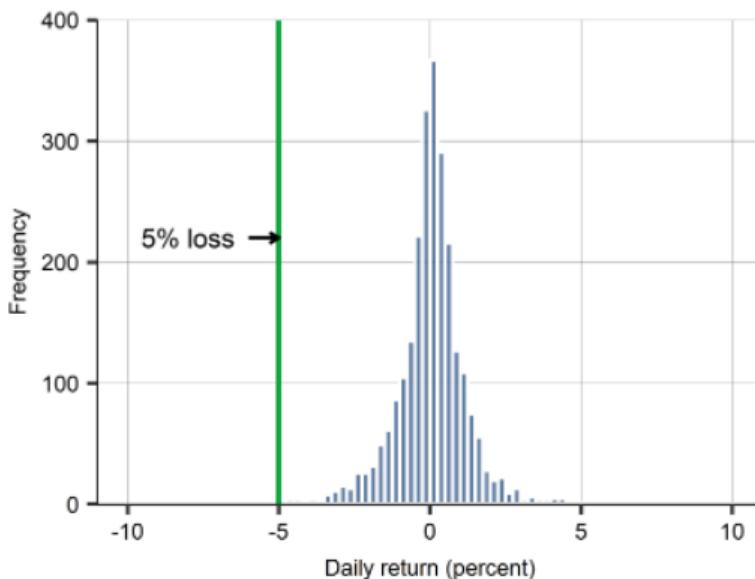
It is good practice to divide the inference problem into two.

1. Use statistical inference to learn about the population, or general pattern, that our data represents.
2. Assess external validity: define the population, or general pattern we are interested in and assess how it compares to the population, or general pattern, that our data represents.

## Stock market returns: Inference

- ▶ Task: Assess the likelihood of experiencing a loss of certain magnitude on an investment portfolio from one day to the next day
- ▶ Predict the frequency of a loss of certain magnitude for the coming calendar year
- ▶ The investment portfolio is the S&P 500, a US stock market index
- ▶ Data: day-to-day returns on the S&P 500, defined as percentage changes in the closing price of the index between two consecutive days
- ▶ 11 years: 25 August 2006 to 26 August 2016. It includes 2,519 days.

## Histogram of daily returns



Note: *S&P 500 market index. Day-to-day (gaps ignored) changes, in percentage. From August 25 2006 to August 26 2016.*

## Stock market returns: Inference

- ▶ To define "loss", we take a day-to-day loss exceeding 5 percent.
- ▶ "loss" is a binary variable, taking 1 when the day-to-day loss exceeds 5 percent and zero otherwise.
- ▶ The statistic in the data is the proportion of days with such losses.
- ▶ It is 0.5 percent in this dataset
  - ▶ the S&P500 portfolio lost more than 5 percent of its value on 0.5 percent of the days between August 25 2006 and August 26 2016.
- ▶ Inference problem: How can we generalize this finding? What can we infer from this 0.5 percent chance for the next calendar year?

## Repeated samples

- ▶ Repeated samples - the conceptual background to statistical inference
- ▶ Our data = one example of many datasets that could have been observed.
- ▶ Many datasets = samples drawn from the population (general pattern)
- ▶ Example 1: Simplest repeated samples
  - ▶ Data is a small set: 1, 2, 3, 4, 5
  - ▶ Pick all possible pairs as repeated samples
  - ▶ → exercise
- ▶ Example 2: Cars
  - ▶ Population is 20,000 yellow Toyota cars sold in Austria in 2019
  - ▶ Create a random sample of 1,000 cars drawn from the population
  - ▶ Repeat it many (say 10,000) times

## Repeated samples

- ▶ The goal of statistical inference is learning the value of a statistic in the population (or general pattern) represented by our data.
- ▶ The statistic has a distribution: its value may differ from sample to sample.
  - ▶ Simple case: mean of pairs of numbers vary across repeated samples
- ▶ The distribution of the statistic of interest is called its sampling distribution

## Repeated samples

- ▶ Standard deviation in this distribution: spread across repeated samples
- ▶ The standard error (SE) of the statistic = the standard deviation of the sampling distribution
- ▶ Any particular estimate is likely to be an erroneous estimate of the true value.
- ▶ The magnitude of that typical error is one SE.

## Repeated samples properties

The sampling distribution of a statistic (e.g. mean) is the distribution of this statistic across repeated samples.

The sampling distribution has three important properties

1. Unbiasedness: The average of the values in repeated samples is equal to its true value (=the value in the entire population / general pattern).
2. Asymptotic normality: The sampling distribution is approximately normal. With large sample size, it is very very close.
3. Root-n convergence: The standard error (the standard deviation of the sampling distribution) is smaller the larger the samples, with a proportionality factor of the square root of the sample size.

## Repeated samples

- ▶ Easier concept: When our data is sample from a well-defined population - many other samples could have turned out instead of what we have.
  - ▶ Example: Mexican firms - random sample - population of firms
- ▶ Harder concept: no clear definition of population. We think of a general pattern we care about.
  - ▶ The data of returns on an investment portfolio may be thought of as a particular realization of the history of returns that could have turned out differently.

## Case study as illustration

- ▶ Introduce the idea of repeated samples

## Stock market returns: A simulation

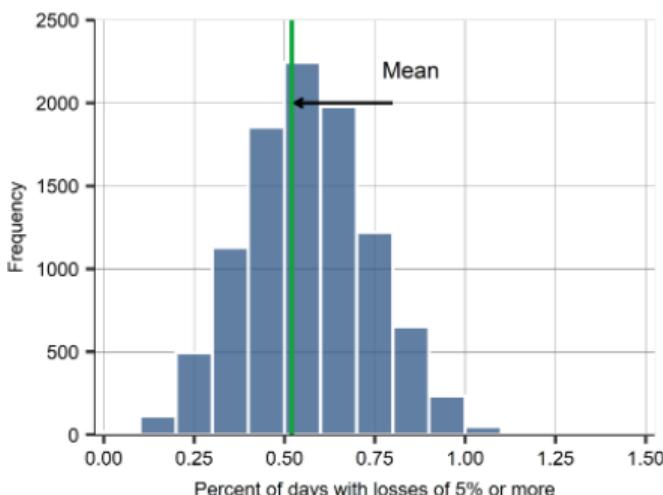
- ▶ We can not rerun history many many times...
- ▶ Simulation exercise - to better understand how repeated samples work
- ▶ Suppose the 11-year dataset is *the population* - the fraction of days with 5%+ losses is 0.5% in the entire 11 years' data. That's the true value.
- ▶ Assume we have only three years (900 days) of daily returns in our dataset.
- ▶ Task: estimate the true value of the fraction in the 11-year period from the data we have using a simulation exercise.
  1. many data table with three years' worth of observations may be created from the 11 years' worth of data,
  2. compute the fraction of days with 5%+ losses in data tables
  3. learn about the true value

## Stock market returns: A simulation

- ▶ Do simple random sampling: days are considered one after the other and are selected or not selected in an independent random fashion.
  - ▶ This sampling destroys the time series nature
  - ▶ This is OK because daily returns are (almost) independent across days in the original dataset
- ▶ We do this 10,000 times....

## Stock market returns: A simulation

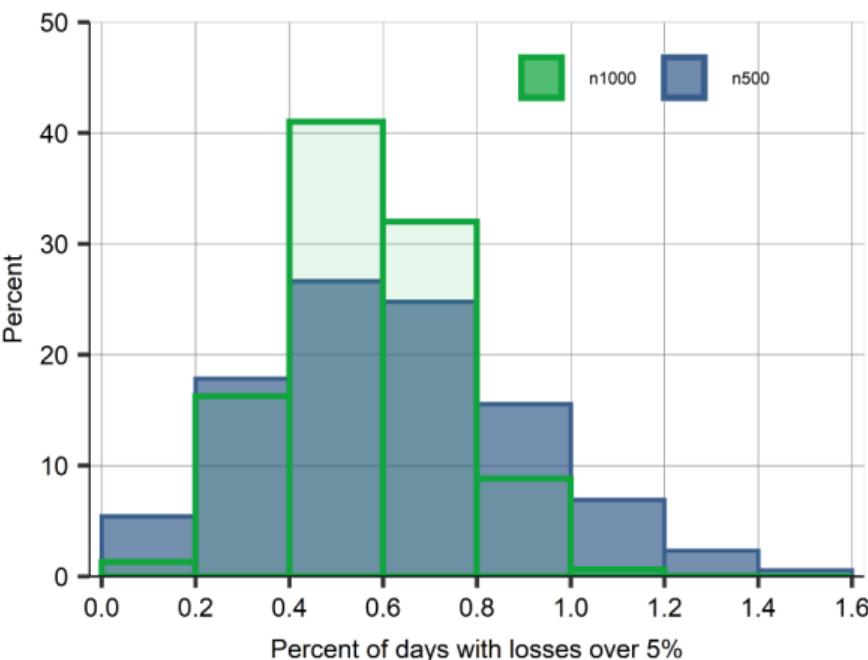
- ▶ percent of days with losses of 5% or more.
- ▶ histogram created from the 10,000 random samples, each w/ 900 obs, drawn from entire dataset
- ▶ distribution has some spread: smallest realization is 0.1 %, while the largest is smaller than 1.25 %



Histogram of the proportion of days with losses of 5 percent or more, across repeated samples of size n=900. 10,000 random samples. Source: sandp-stocks data. S&P 500 market index.

## Stock market returns: Sampling distributions

- ▶ Proportion of days with losses of 5 percent or more
- ▶ Repeated samples in two simulation exercises, with  $n=500$  and  $n=1,000$ . (10,000 random samples)
- ▶ Kernel density (goes to minus / can cut it at 0)
- ▶ Role of sample size: smaller sample: skewed; higher standard deviation



## The standard error and the confidence interval

- ▶ Confidence interval (CI) - measure of statistical inference.
  - ▶ Recall: Statistical inference - we analyze a dataset to infer the true value of a statistic: its value in the population, or general pattern, represented by our data.
- ▶ The CI defines a range where we can expect the true value in the population, or the general pattern.
- ▶ CI gives a range for the true value with a probability
- ▶ Probability tells how likely it is that the true value is in that range
- ▶ Probability - data analysts need to picks it, such as 95%

## The standard error and the confidence interval

- ▶ The “95 percent CI” gives the range of values where we think that true value falls with a 95 percent likelihood.
- ▶ Viewed from the perspective of a single sample, the chance (probability) that the truth is within the CI measured around the value estimated from that single sample is 95 percent.
- ▶ Also: we think that with 5 percent likelihood, the true value will fall outside the confidence interval.

## The standard error and the confidence interval

- ▶ Confidence interval - symmetric range around the estimated value of the statistic in our dataset.
  - ▶ Get estimated value.
  - ▶ Define probability
  - ▶ Calculate CI with the use of SE
- ▶ 95 percent CI is the  $\pm 1.96SE$  (but we use  $\pm 2SE$ ) interval around the estimate from the data.
  - ▶ 90% CI is the  $\pm 1.6SE$  interval, the 99 % CI is the  $\pm 2.6SE$

## Calculating the standard error

An important consequence of evidence from the repeated sample exercise:

- ▶ In reality, we don't get to observe the sampling distribution. Instead, we observe a single dataset
- ▶ That dataset is one of the many potential samples that could have been drawn from the population, or general pattern
- ▶ Good news: We can get a very good idea of how the sampling distribution would look like - good estimate of the standard error - even from a single sample.
- ▶ Getting SE – Option 1: Use a formula
- ▶ Getting SE – Option 2: Simulate by a new method, called 'bootstrapping'

## Calculating the standard error

Consider the statistic of the sample mean.

- ▶ Assume the values of  $x$  are independent across observations in the dataset.
- ▶  $\bar{x}$  is the estimate of the true mean value of  $x$  in the general pattern/population
- ▶ Sampling distribution is approximately normal, with the true value as its mean.

The standard error formula for the estimated  $\bar{x}$  is

$$SE(\bar{x}) = \frac{1}{\sqrt{n}} Std[x] \quad (1)$$

where  $Std[x]$  is the standard deviation of the variable  $x$  in the data and  $n$  is the number of observations in the data.

## The standard error formula

- ▶ The standard error is larger...
  - ▶ the larger the standard deviation of the variable.
  - ▶ the smaller the sample and
- ▶ For intuition, consider  $SE(\bar{x})$  vs  $Std[x]$ .
- ▶ Think back to the repeated samples simulation exercise:
  - ▶  $SE(\bar{x})$  = the standard error of  $\bar{x}$  is the standard deviation of the various  $\bar{x}$  estimates across repeated samples.
  - ▶ The larger the standard deviation of  $x$  itself, the more variation we can expect in  $\bar{x}$  across repeated samples.

## Stock market returns: The standard error formula

Let's consider our example of 11-years' of data on daily returns on the S&P 500 portfolio.

- ▶ The calculated statistics,  $P(\text{loss} > 5\%) = 0.5\%$
- ▶ The  $SE [P(\text{loss} > 5\%)]$  is calculated by,
  - ▶ The size of the sample is  $n = 2,519$  so that  $1/\sqrt{n} = 0.02$ .
  - ▶ The standard deviation of the fraction of  $SD [P(\text{loss} > 5\%)] = 0.07$ .
  - ▶ So the  $SE = 0.07 * 0.02 = 0.0014$  (0.14 percent).
- ▶ Can calculate the 95 percent CI:
  - ▶  $CI = [0.5 - 2 * SE, 0.5 + 2 * SE] = [0.22, 0.78]$
- ▶ This means that in the general pattern represented by the 11-year history of returns in our data, we can be 95 percent confident that daily losses of more than 5 percent occur with a 0.2 to 0.8 percent chance.

## Take a quick stop to summarize the idea of CI

- ▶ We are interested in generalizing from our data. Statistical inference.
- ▶ Consider a statistic such as the sample mean  $\bar{x}$
- ▶ Take a 95% confidence interval - where we can expect to see the true value
- ▶  $CI = \text{statistic} +/− 2SE$ .
- ▶ We have a formula for the SE calculated from our data only using the standard deviation and sample size.
- ▶ Using the CI, we can now do statistical inference, generalize for the population / general pattern we care about.

## External validity

- ▶ We discussed statistical inference: CI - uncertainty about the true value of the statistic in the population / general pattern that our data represents.
- ▶ What is the population, or general pattern, we care about?
- ▶ How close is our data to this?
- ▶ External validity is the concept that captures the similarity of our data to the population/general pattern we care about.
- ▶ High external validity: if our data is close to the population or the general pattern we care about.
- ▶ External validity is as important as statistical inference. However, it is not a statistical question.

## External validity

- ▶ The most important challenges to external validity may be collected in three groups:
- ▶ Time: we have data on the past, but we care about the future
- ▶ Space: our data is on one country, but interested how a pattern would hold elsewhere in the world
- ▶ Sub-groups: our data is on 25-30 year old people. Would a pattern hold on younger / older people?

## External validity

- ▶ Daily 5%+ loss probability - 95 percent CI [0.2, 0.8] in our sample. This captures uncertainty for samples like ours.
- ▶ If the future one year will be like the past 11 years in terms of the general pattern that determines returns on our investment portfolio.
- ▶ However, external validity may not be high - not sure what the future holds.
- ▶ Our data: 2006-2016 dataset includes the financial crisis and great recession of 2008-2009. It does not include the dotcom boom and bust of 2000-2001. We have no way to know which crisis is representative to future crises to come.
- ▶ Hence, the real CI is likely to be substantially wider.

## External validity: Example

- ▶ Manager and firm size evidence in Mexico
- ▶ How to think about external validity?
- ▶ What do you think AI would help
- ▶ Try at home:
  - ▶ Explain data, share a statistic, the CI, and ask about external validity.
  - ▶ More direct: think about a use case and ask for that special case
  - ▶

## External validity in Big Data

- ▶ Big data: very large  $N$
- ▶ Statistical inference not really important - CI becomes very narrow
- ▶ External validity remains as important
  
- ▶ 1.) Large sample DOES NOT mean representative sample
- ▶ 2.) Big data as result of actions - nature of things may change as people alter behavior, outside conditions change

# The bootstrap

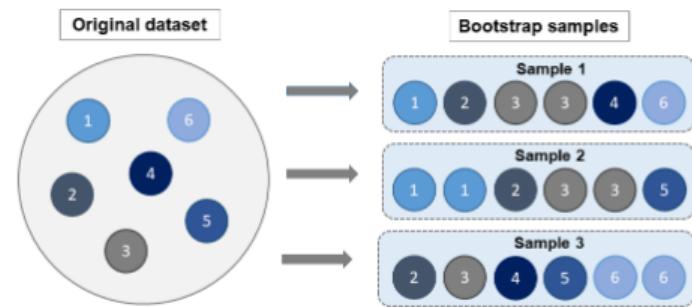
- ▶ Bootstrap is a method to create synthetic samples that are similar but different
- ▶ An method that is very useful in general.
- ▶ It is essential for many advanced statistics applications such as machine learning
- ▶ **More in Chapter 05**

## The bootstrap

- ▶ The bootstrap method takes the original dataset and draws many repeated samples of the size of that dataset.
- ▶ The trick is that the samples are drawn *with replacement*.
- ▶ The observations are drawn randomly one by one from the original dataset; once an observation is drawn it is “replaced” to the pool so that it can be drawn again, with the same probability as any other observation.
- ▶ The drawing stops when it reaches the size of the original dataset.
- ▶ The result is a sample of the same size as the original dataset, yielding a single *bootstrap sample*.

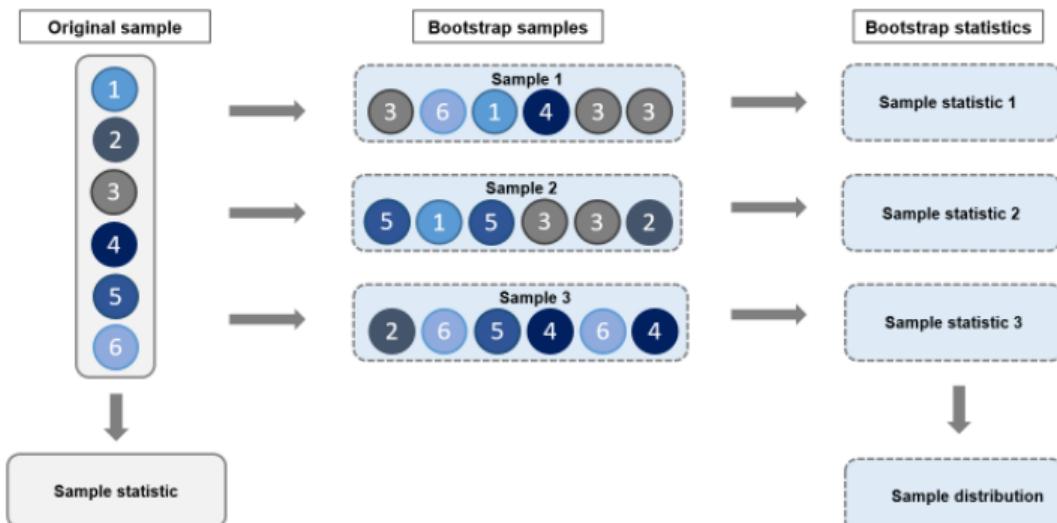
# The bootstrap

- ▶ A bootstrap sample is always the same size the original
- ▶ it includes some of the original observations multiple times,
- ▶ it does not include some of other original observations.
- ▶ We typically create 500 - 10,000 samples
- ▶ Computationally intensive but feasible, relatively fast.



# The bootstrap

- ▶ We have a dataset (the sample), can compute a statistic (e.g. mean)
- ▶ Create many bootstrap samples, and get a mean value for each sample
- ▶ Bootstrap estimate of  $SE = \text{standard deviation of statistic}$  based on bootstrap samples' estimates.



## The bootstrap method and bootstrap SE

- ▶ The bootstrap method creates many repeated samples that are different from each other, but each has the same size as the original dataset.
- ▶ The distribution of a statistic across these repeated bootstrap samples is a good approximation to the sampling distribution we are after
  - ▶ ... what the distribution would look like across datasets similar to the original dataset.
- ▶ Bootstrap gives a good approximation of the standard error, too.
- ▶ The bootstrap estimate (or the estimate from the bootstrap method) of the standard error is simply the standard deviation of the statistic across the bootstrap samples.

## Stock market returns: The Bootstrap standard error

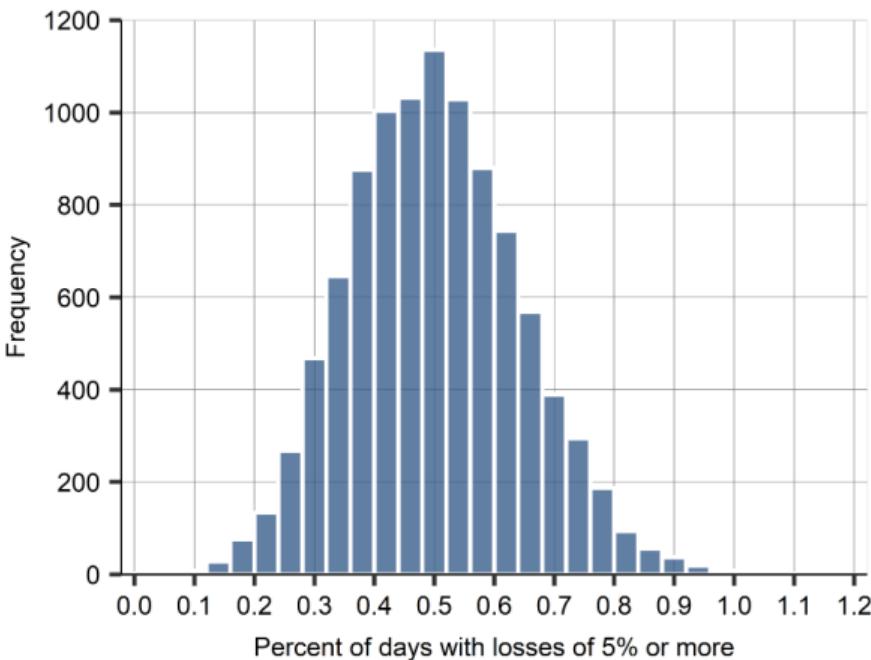
- ▶ We estimate the standard error by bootstrap.
- ▶ Let's consider our example of 11-years' of data on daily returns on the S&P 500 portfolio.
- ▶ Do the process —————>
- ▶ End up with a new a dataset: one observations / bootstrap sample.  
Only variable is the estimated proportion in a sample
- ▶ The SE is simply the standard deviation of those estimated values in this new dataset.

### The process

1. Take the original dataset and draw a bootstrap sample.
2. Calculate the proportions of days with 5%+ loss in that sample.
3. Save that value.
4. Then go back to the original dataset and take another bootstrap sample.
5. Calculate the proportion of days with 5%+ loss and save that value, too.
6. And so on, repeated many times.

## Stock market returns: The Bootstrap standard error

- ▶ 10,000 bootstrap samples with 2,519 observations
- ▶ The proportion of days with 5+ percent loss.
- ▶ Varied 0.1 percent to 1.2 percent. Mean=Median= 0.5
- ▶ Standard deviation across the bootstrap samples = 0.14
- ▶ CI: the 95 percent CI is [0.22, 0.78].



## Stock market returns: The Bootstrap standard error

- ▶ This means that in the general pattern represented by the 11-year history of returns in our data, we can be 95 percent confident that daily losses of more than 5 percent occur with a 0.22 to 0.78 percent chance.
- ▶ SE formula and bootstrap gave the same exact answer
- ▶ Under some conditions, this is what we expect
  - ▶ Large enough sample size
  - ▶ Observations independent
  - ▶ ... (other we overlook now)

## Generalization - Summary

- ▶ Generalization is a key task - finding beyond the actual dataset.
- ▶ This process is made up of discussing statistical inference and external validity.
- ▶ Statistical inference generalizes from our dataset to the population using a variety of statistical tools.
- ▶ External validity is the concept of discussing beyond the population for a general pattern we care about; an important but typically somewhat speculative process.

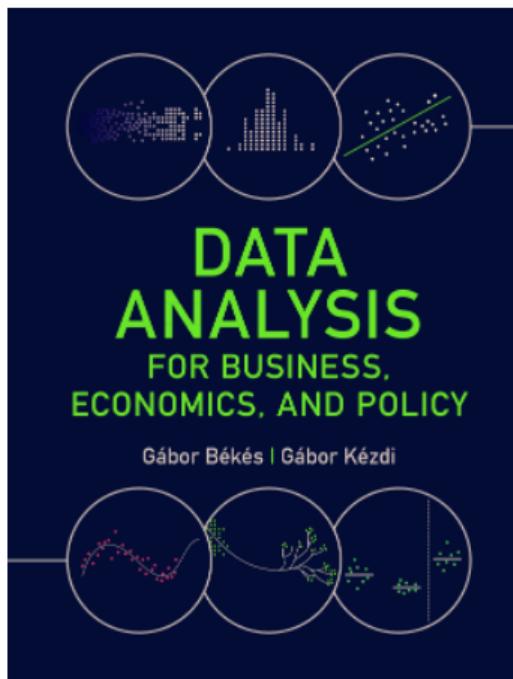
# 06 Testing hypotheses

Gábor Békés

Data Analysis 1: Exploration

2023

# Slideshow for the Békés-Kézdi Data Analysis textbook



- ▶ Cambridge University Press, 2021
- ▶ [gabors-data-analysis.com](http://gabors-data-analysis.com)
  - ▶ Download all data and code:  
[gabors-data-analysis.com/data-and-code/](http://gabors-data-analysis.com/data-and-code/)
- ▶ This slideshow is for Chapter 06

# Motivation

- ▶ The internet allowed the emergence of specialized online retailers while brick-and-mortar shops also sell goods on the main street. How to measure price inflation in the age of these options?
- ▶ To help answer this, we can collect and compare online and offline prices of the same products and test if they are the same.

Hypothesis  
o●oooooooooooo

A1  
ooooooo

The t-test  
oooooooooo

Making a decision  
oooooooooooooooo

p-value  
ooooooo

A2  
ooo

Multiple test  
oooooooooooo

Big Data  
o

Sum  
o

Extra  
oooo

# The logic of hypothesis testing

# The logic of hypothesis testing

- ▶ A hypothesis is a statement about a general pattern, of which we are not sure if true or not.
- ▶ Hypothesis testing = analyze our data to make a decision on the hypothesis
- ▶ Reject the hypothesis if there is enough evidence against it.
- ▶ Don't reject it if there isn't enough evidence against it.
- ▶ We may not have enough evidence against a hypothesis
  - ▶ if the hypothesis is true
  - ▶ or it is not true only the evidence is weak
- ▶ Important asymmetry here: rejecting a hypothesis is a more conclusive decision than not rejecting it.

Hypothesis  
oooooooooooo

A1  
oooooooo

The t-test  
oooooooooooo

Making a decision  
oooooooooooooooooooo

p-value  
ooooooo

A2  
ooo

Multiple test  
oooooooooooo

Big Data  
o

Sum  
o

Extra  
oooo

## The logic of hypothesis testing: inference

- ▶ Testing a hypothesis: making inference with a focus on a specific statement.
- ▶ Can answer questions about the population, or general pattern, represented by our data.
- ▶ It is an inference: have to assess external validity

# The logic of hypothesis testing: the setup

- ▶ Define the *the statistic we want to test,  $s$*  (e.g. mean).
- ▶ We are interested in the true value of  $s$ ,  $s_{true}$ .
- ▶ This is statistical inference, so the true value means the value in the population, or general pattern represented by our data.
- ▶ The value the statistic in our data is its estimated value, denoted by a hat on top  $\hat{s}$ .

# The logic of hypothesis testing: $H_0$ vs $H_A$

- ▶ Formally stating the question as two competing hypotheses of which only one can be true: a null hypothesis  $H_0$  and an alternative hypothesis  $H_A$ .
- ▶ Formulated in terms of the unknown true value of the statistic.
- ▶ The null specifies some value/ range; the alternative specifies other possible values.
- ▶ Together, the null and the alternative cover all the possibilities we are interested in
- ▶ One example is null:  $s$  is zero, alternative:  $s$  is not zero.

$$H_0 : s_{true} = 0$$

$$H_A : s_{true} \neq 0$$

# The logic of hypothesis testing: $H_0$ vs $H_A$

- ▶ Our case study research question: Do the online and offline prices of the same products differ or are they the same?
- ▶ We have the price difference as our statistic and  $H_0 : s_{true} = 0$
- ▶ Testing a hypothesis = see if there is enough evidence in our data to reject the null.

# The logic of hypothesis testing: Null protected

- ▶ Testing a hypothesis = see if there is enough evidence in our data to reject the null.
- ▶ The null is protected: it has to be hard to reject it otherwise the conclusions of hypothesis testing would not be strong.

# The logic of hypothesis testing: The criminal court example

- ▶ Logic of testing like a criminal court procedure.
  - ▶ Decide if the accused is guilty or innocent of a certain crime.
  - ▶ Assumption of innocence: accused judged guilty only if enough evidence against innocence
  - ▶ Even though the accused in court because of suspicion of guilt.
- ▶ To translate this procedure to the language of hypothesis testing,
  - ▶  $H_0$  is that the person is innocent
  - ▶  $H_A$  is that the person is guilty.

# The logic of hypothesis testing: $H_0$ vs $H_A$

- ▶ Two-sided alternative: The case when we test if  $H_A : s_{true} \neq 0$  - allows for  $s_{true}$  to be either greater than zero or less than zero. Not interested if the difference is positive or negative.

$$H_0 : s_{true} = 0$$

$$H_A : s_{true} \neq 0$$

- ▶ One-sided alternative: interested if a statistic is positive or not.

$$H_0 : s_{true} \leq 0$$

$$H_A : s_{true} > 0$$

# Summary of the logic of hypothesis testing

- ▶  $H_A$  is (often) what I wanna prove
- ▶  $H_0$  is what I wanna reject so that I can prove  $H_A$
- ▶  $H_0$  is not rejected
  - ▶ not enough evidence or
  - ▶ true (ie  $H_A$  is false)
- ▶ I can never say  $H_0$  is true.

## Case Study - Comparing online and offline prices: Testing hypotheses

- ▶ Question: Do the online and offline prices of the same products differ?
- ▶ this data includes 10 to 50 products in each retail store included in the survey (the largest retailers in the U.S. that sell their products both online and offline).
- ▶ The products were selected by the data collectors in offline stores, and they were matched to the same products the same stores sold online.
- ▶ Let define our statistic as the difference in average prices.

## Case Study - Comparing online and offline prices: Testing hypotheses

- ▶ Descriptive statistics of the difference
- ▶ Each product  $i$  has both an online and an offline price in the data,  $p_{i,online}$  and  $p_{i,offline}$ ,  $pdiff$  is their difference:

$$pdiff_i = p_{i,online} - p_{i,offline} \quad (1)$$

The statistic with  $n$  observations (products) in the data, is:

$$s = \overline{pdiff} = \frac{1}{n} \sum_{i=1}^n (p_{i,online} - p_{i,offline}) \quad (2)$$

## Case Study - Comparing online and offline prices: Testing hypotheses

- ▶ The average of the price differences is equal to the difference of the average prices
- ▶  $s$  statistic also measures the difference between the average of online prices and the average of offline prices among products with both kinds of price

$$\frac{1}{n} \sum_{i=1}^n (p_{i,online} - p_{i,offline}) = \frac{1}{n} \sum_{i=1}^n p_{i,online} - \frac{1}{n} \sum_{i=1}^n p_{i,offline}$$

## Case Study - Comparing online and offline prices: Testing hypotheses

### Descriptive statistics of the difference

- ▶ The mean difference is USD -0.05: online prices are, on average, 5 cents lower in this dataset.
- ▶ Spread around this average: Std: USD 10
- ▶ Extreme values matter: Range: -380 — USD +415.
- ▶ Of the 6439 products, 64% have the same online and offline price, for 87%, the difference within  $\pm 1$  dollars.

## Case Study - Comparing online and offline prices: the setup

### External validity

- ▶ The products in the data may not represent all products sold at these stores.
  - ▶ Could be a bias. **Example?**
- ▶ Strictly: The general pattern of the statistic represented by this dataset is average online-offline price differences in large retail store chains for the kind of products that data collectors would select with a high likelihood.
- ▶ More broadly: price differences among *all* products in the U.S. sold both online and offline by the same retailers.
  - ▶ Need an assumption. **What would it be?**

## Case Study - Comparing online and offline prices: the setup

Do average prices differ in the general pattern represented by the data?

$$H_0 : s_{true} = \bar{p}_{online\ true} - \bar{p}_{offline\ true} = 0 \quad (3)$$

$$H_A : s_{true} = \bar{p}_{online\ true} - \bar{p}_{offline\ true} \neq 0 \quad (4)$$

Hypothesis  
oooooooooooo

A1  
ooooooo●

The t-test  
oooooooooo

Making a decision  
oooooooooooooooo

p-value  
ooooooo

A2  
ooo

Multiple test  
oooooooooooo

Big Data  
o

Sum  
o

Extra  
oooo

# Testing

# The logic of hypothesis testing

- ▶ The t-test is the testing procedure based on the t-statistic
- ▶ We compare the estimated value of the statistic  $\hat{s}$  (our best guess of  $s$ ) to zero.
- ▶ Evidence to reject the null = based on difference between  $\hat{s}$  and zero.
- ▶ Reject the null if difference large = it is unlikely to be zero.
- ▶ Not reject the null if the difference is small = not enough evidence against it.
- ▶ Need to define "large"/"small" (*next*)

# T-test

- ▶ The test statistic is a statistic that measures the distance of the estimated value from what the true value would be if  $H_0$  was true.
- ▶ Uses estimated value of  $s$  ( $\hat{s}$ ) and the standard error of estimate ( $SE(\hat{s})$ ).
  - ▶ SE is the scaling (normalization)
- ▶ Consider  $H_0 : s_{true} = 0, H_A : s_{true} \neq 0$ . The t-statistic for this hypotheses is:

$$t = \frac{\hat{s}}{SE(\hat{s})} \tag{5}$$

- ▶ The test statistic summarizes all the information needed to make the decision.
- ▶ When hypotheses are about value of one coefficient the test statistic = t-statistic

Hypothesis  
oooooooooooo

A1  
ooooooo

The t-test  
oo●ooooo

Making a decision  
oooooooooooooooo

p-value  
oooooo

A2  
ooo

Multiple test  
oooooooooooo

Big Data  
o

Sum  
o

Extra  
oooo

## T-test

When  $\hat{s}$  is the average of a variable  $x$ , the t-statistic is simply

$$t = \frac{\bar{x}}{SE(\bar{x})} \quad (6)$$

When  $\hat{s}$  is the average of a variable  $x$  minus a number, the t-statistic is

$$t = \frac{\bar{x} - \text{number}}{SE(\bar{x})} \quad (7)$$

When  $\hat{s}$  is the difference between two averages, say,  $\bar{x}_A$  and  $\bar{x}_B$ , the t-statistic is

$$t = \frac{\bar{x}_A - \bar{x}_B}{SE(\bar{x}_A - \bar{x}_B)} \quad (8)$$

# T-test

- ▶ If  $\hat{s} > 0$  = the t-statistic is positive; if  $\hat{s} < 0$  = the t-statistic is negative.
- ▶ With a two-sided alternative ( $H_A : s_{true} \neq 0$ ) it is the magnitude not the sign of the t-statistic that matters.
- ▶ If  $\hat{s} = 0$  then  $t = 0$ .
  - ▶ In reality it's never *exactly* zero.
  - ▶ But expect  $\hat{s}$  estimate to be *close* to zero.
- ▶ If the null is incorrect and thus  $s_{true}$  is *not* zero -> we expect the  $\hat{s}$  estimate to be far from zero.

# T-test

- ▶ We standardize distance with  $SE(\bar{x})$
- ▶ May use  $SE(\bar{x}) = \sqrt{\frac{1}{n}} Std[x]$ .
- ▶ SE formula may be more complicated
- ▶ Sometimes no appropriate SE formula for a statistic interested in -> Need bootstrap estimation.

# Ask ChatGPT

*Can you show me the formula for a t-test for a difference in the means of a variable x in two samples. Also show in latex.*

- ▶ What you need to know is that there should be one, different from what you have seen.
- ▶ Mostly correct, but be able to check...

# T-test for the difference in two sample means [extra]

Let us consider two independent samples,  $x_1$  and  $x_2$ :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{Std_{x1}^2}{n_1} + \frac{Std_{x2}^2}{n_2}}}$$

Where:

$t$  is the t-statistic.

$\bar{x}_1$  and  $\bar{x}_2$  are the sample means.

$Std_{x1}$  and  $Std_{x2}$  are the standard deviations in  $x_1$  and  $x_2$ .

$n_1$  and  $n_2$  are the sample sizes of  $x_1$ ,  $x_2$ .

Hypothesis  
oooooooooooo

A1  
ooooooo

The t-test  
oooooooo●

Making a decision  
oooooooooooooooo

p-value  
oooooo

A2  
ooo

Multiple test  
oooooooooooo

Big Data  
o

Sum  
o

Extra  
oooo

# Generalization

# Making a decision

- ▶ In hypothesis testing the decision is based on a clear rule specified in advance.
- ▶ A decision rule makes the decision straightforward + transparent
- ▶ Helps avoid personal bias: put more weight on the evidence that supports our prejudices.
- ▶ Clear decision rules are designed to minimize the room for such temptations.

## Making a decision: decision rule

- ▶ The decision rule = comparing the test statistic to a pre-defined critical value.
- ▶ Is test statistic is large enough to reject the null.
- ▶ Null rejected if the test statistic is larger than the critical value
- ▶ Critical value - between being too strict or too lenient.
- ▶ When we make the decision, we may be right or wrong, don't know: need to think

# Making a decision

- ▶ We can be right in our decision in two ways:
  - ▶ we reject the null when it is not true,
  - ▶ or we do not reject the null when it is true.
- ▶ We can be wrong in our decision in two ways, too:
  - ▶ we reject the null even though it is true,
  - ▶ or we do not reject the null even though it is not true.

	$H_0$ is true	$H_0$ is false
Don't reject the null	True negative	False negative - Type II error
Reject the null	False positive - Type I error	True positive

# Making a decision

- ▶ We say that our decision is a *false positive* if we reject the null when it is true.
  - ▶ “positive” because we take the active decision to reject the protected null.
  - ▶ medical: person has the condition that they were tested against
  - ▶ False positive = type-I error;
- ▶ Our decision is a *false negative* if we do not reject the null even though we should.
  - ▶ “negative” because we do not take the active decision
  - ▶ medical: result is “negative” = not have the condition
  - ▶ False negative =type-II error.

# Making a decision

- ▶ False positives and false negatives: both wrong, but not equally.
- ▶ Testing procedure protects the null: reject it only if evidence is strong
- ▶ The background assumption - wrongly rejecting the null (a false positive) is a bigger mistake than wrongly accepting it (a false negative).
- ▶ Decision rule (critical value) is chosen in a way that makes false positives rare.

# Making a decision

- ▶ A commonly applied critical value for a t-statistic is  $\pm 2$  (or 1.96):
  - ▶ reject the null if the t-statistic is smaller than  $-2$  or larger than  $+2$ ;
  - ▶ don't reject the null if the t-statistic is between  $-2$  and  $+2$ .
- ▶ With  $\pm 2$  critical value - 5% is the probability of false positives - we have 5% as the probability that we would reject the null if it was true (False positive).
  - ▶  $\text{Prob}(\text{t-statistic} < -2)$  or  $\text{Prob}(\text{t-statistic} > 2)$  are both appr 2.5%
  - ▶ If the null is true: Probability t-statistic is below  $-2$  or above  $+2$  is 5%
- ▶ If we make the critical values  $-2.6$  and  $+2.6$  the chance of the false positive is 1%.

# Critical values and generalization

- ▶ Can set other critical values that correspond to different probabilities of a false positive.
- ▶ That choice of 5% means that we tolerate a 5% chance for being wrong when rejecting the null
- ▶ Data analysts avoid biases when testing hypotheses: use the same critical value regardless of the data and hypothesis they are testing.

# Critical values and generalization

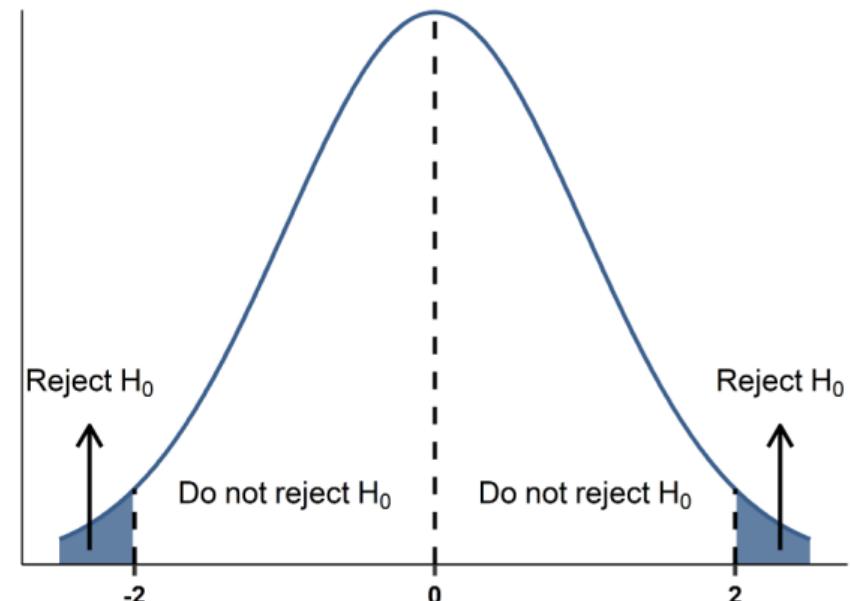
- ▶ Where does this  $2SD - 5\%$  come from?
- ▶ We can calculate the likelihood of a false positive because we know what the sampling distribution of the test statistic would be if the null were true.
- ▶ The sampling distribution of a statistic is its distribution across repeated samples
  - ▶ of the same size from the same population.
- ▶ The sampling distribution of an average is approximately normal, its mean is equal to the true mean, and its standard deviation is called the standard error.

## Critical values and generalization

- ▶ How would the sampling distribution look if the null hypothesis were true:
- ▶ Distribution of the t-statistic would be standard normal  $N(0, 1)$
  
- ▶ The t-statistic has the average in its numerator, so that its distribution is also approximately normal,
- ▶ The t-statistic SD=1 because because the t-statistic is standardized – it has the SE of  $\hat{s}$  in the denominator
  - ▶ Note: Small sample ( $<30$ ), the normal approximation to the distribution of the t-statistic is not very good. Instead, the distribution is closer to "t-distribution" )

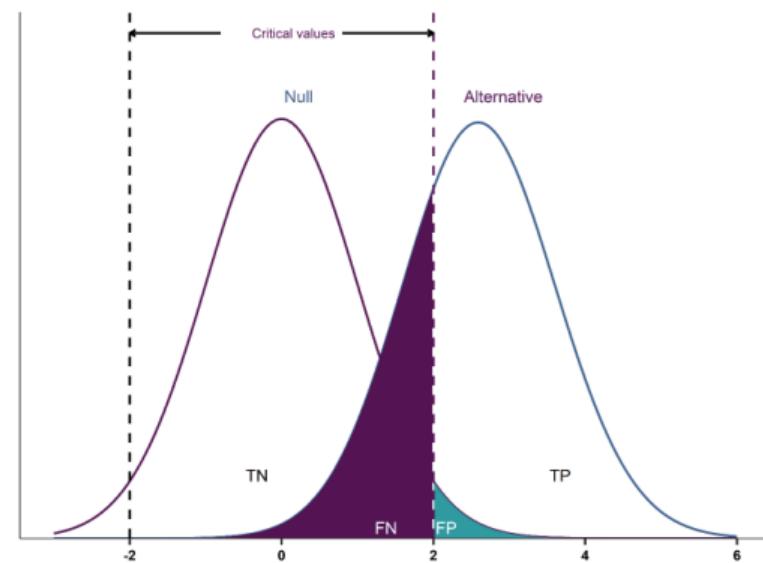
# Sampling distribution of the test statistic when the null is true

- ▶ Distribution of the t-statistic close to  $N(0, 1)$
- ▶ Prob t-statistic  $< -2$  or  $> 2$  is approximately 2.5%. Prob t-statistic is  $< -2$  or  $> +2$  is 5% if the null is true. (Two-sided alternative)
- ▶ 5% = probability of false positives if we apply the critical values of  $\pm 2$



# False negative (FN)

- ▶ Fixing the chance of FP affects the chance of FN at the same time.
- ▶ A FN arises when the t-statistic is within the critical values and we don't reject the null even though the null is not true.
- ▶ Making a FN call more likely when harder to make a decision
  - ▶ Sample is small
  - ▶ The difference between true value and null is small



# Size and power of the test

## Under the null:

- ▶ *Size of the test:* the probability of committing a false positive.
- ▶ *Level of significance:* The maximum probability of false positives we tolerate.

When we fix the level of significance at 5% and end up rejecting the null, we say that the statistic we tested is significant at 5%

## Under the alternative:

- ▶ *Power of the test:* the probability of avoiding a false negative
- ▶ Being different from the null can be in many ways...
- ▶ High power is more likely when
  - ▶ The sample is large and the dispersion is small.
  - ▶ The further away the true value is from what's in a null.

We usually fix the level of significance at 5% and hope for a high power of the test.

# Making a decision

- ▶ We know the sampling distribution of the test statistic if the null is true—> can calculate the likelihood of a false positive
- ▶ Recall: sampling distribution of an average value is approximately normal,
  - ▶ mean = being equal to the true mean value,
  - ▶ the standard deviation being equal to its standard error.
- ▶ The distribution of the t-statistic is standard normal distribution  $N(0,1)$ 
  - ▶ It has mean zero because  $s_{true} = 0$  if the null is true.
  - ▶ It has standard deviation one because the standard deviation of the sampling distribution of  $\hat{s}$  is  $SE(\hat{s})$ , and the t-statistic is  $\hat{s}/SE(\hat{s})$ .

## Recap

- ▶ In hypothesis testing we make decisions by a rule
  - ▶ A false positive is a decision to reject the null hypothesis when it is in fact true.
  - ▶ A false negative is a decision not to reject the null hypothesis when it is in fact not true.
- ▶ The level of significance is the maximum probability of a false positive that we tolerate.
- ▶ The power of the test is the probability of avoiding a false negative.
- ▶ In statistical testing we fix the level of significance of the test to be small (5%, 1%) and hope for high power.
- ▶ Tests with more observations have more power in general.

# The p-value

- ▶ The p-value makes testing easier - captures info for reject/accept calls.
  - ▶ Instead of calculating test statistics and specify critical values, we can make an informed decision based on the p-value only.
- ▶ p-value is the smallest significance level at which we can reject  $H_0$  given the value of the test statistic in the sample.
  - ▶ *the p-value is the probability that the test statistic will be as large as, or larger than, what we calculate from the data, if the null hypothesis is true.*
- ▶ The p-value tells us the largest probability of a false positive.
- ▶ The p-value depends on
  1. the test statistic,
  2. the critical value
  3. the sampling distribution of the test statistic

Hypothesis  
oooooooooooo

A1  
ooooooo

The t-test  
oooooooooo

Making a decision  
oooooooooooooooooooo

p-value  
o●oooo

A2  
ooo

Multiple test  
oooooooooooo

Big Data  
o

Sum  
o

Extra  
oooo

## Recap: p vs power

- ▶ p-value = probability rejecting the null while it is true (probability of avoiding FP).
- ▶ Power = probability rejecting the null while it is false (probability of avoiding FN)

# The p-value

- ▶ If the p-value is 0.05 the maximum probability that we make a false positive decision is 5%.
  - ▶ If we are willing to take that chance, we should reject the null; if we are not, we should not.
  - ▶ If the p-value is, say, 0.001 there is at most a 0.1% chance of being wrong if we were to reject the null.
- ▶ We can never be certain!  $p$  is never zero.
- ▶ For a reject/accept decision, one should pick a level of significance before the test
- ▶ What we can accept depends on the setting: what is the cost of a false positive.

Hypothesis  
oooooooooooo

A1  
ooooooo

The t-test  
oooooooooo

Making a decision  
oooooooooooooooooooo

p-value  
oooo●ooo

A2  
ooo

Multiple test  
oooooooooooo

Big Data  
o

Sum  
o

Extra  
oooo

## What p-value to pick?

- ▶ p-value is about a trade-off. Large (10-15%) or small (1%) depends on scenarios
- ▶ Guilty beyond reasonable doubt?
- ▶ Proof of concept?

# What p-value to pick?

- ▶ p-value is about a trade-off. Large (10-15%) or small (1%) depends on scenarios
- ▶ Guilty beyond reasonable doubt?
- ▶ Pick a conservative value, like 1% or lower
- ▶ Proof of concept?
- ▶ It's great if it works at 5%, but even 10-15% means it's much more likely to be true
  - ▶ May lead to doing more experimentation, increase sample size

# One-sided t-test, calculating p-value

- ▶ One sided test: having an inequality in  $H_A$
- ▶  $H_0 : s_{true} \geq 0$  against  $H_A : s_{true} < 0$
- ▶ Equality always part of the null
  
- ▶ In order to reject  $H_0$ , we need to reject each and every value in favor of  $s < 0$
- ▶ Hardest value to reject against is  $s = 0$  against  $s < 0$ 
  - ▶ this is why equality is part of the null
- ▶ Difference to two sided: we only care about being wrong on one side,
  - ▶ the probability of FP is smaller (=half)
  - ▶ t-test of two-sided hypotheses — the p-value as the sum of two probabilities – we only have half the probability of error
- ▶ Practically: run a two-sided test, calculate p-value and take its half.

## Case Study - Comparing online and offline prices: Testing hypotheses

- ▶ Let's fix the level of significance at 5%.
  - ▶ Doing so we tolerate a 5% chance for a false positive.
  - ▶ Allow a 5% chance to be wrong if we reject the null hypothesis of zero average price difference.
- ▶ A 5% level of significance translates to  $\pm 2$  bound for the t-statistic.
- ▶ The value of the statistic in the dataset is -0.054. Its standard error is 0.124.
- ▶ The CI is  $-0.054 \pm 2 * 0.124 = [-0.30, +0.19]$ 
  - ▶ Thus the t-statistic is 0.44. This is well within  $\pm 2$ .
  - ▶ Don't reject the null hypothesis of zero difference.
- ▶ We do not say we proved it's zero. We showed we cannot tell it apart from zero.

## Case Study - Comparing online and offline prices: Testing hypotheses

- ▶ Conclude that the average price difference is not different from zero in the general pattern represented by the data.
- ▶ Large dataset, good power. What we see in t-statistic is not because of very small sample size
- ▶ It is still possible that prices are indeed different, just the difference is very small. A few cent difference would not matter economically ...

## Case Study - Comparing online and offline prices: Testing hypotheses

- ▶ The p-value of the test is 0.66.
- ▶ That means that the smallest level of significance at which we can reject the null is 66%.
- ▶ The chance that we would make a mistake if we rejected the null is at most 66%.
- ▶ So we don't reject the null

Hypothesis  
oooooooooooo

A1  
ooooooo

The t-test  
oooooooooo

Making a decision  
oooooooooooooooo

p-value  
oooooo

A2  
ooo

Multiple test  
●oooooooooooo

Big Data  
o

Sum  
o

Extra  
oooo

# Multiple test

## Multiple testing: motivation

- ▶ Medical dataset: data on 400 patients
- ▶ A particular heart disease binary variable and 100 feature of life style (sport, eating, health background, socio-economic factors)
- ▶ Look for a pattern – is the heart disease equally likely for poor vs rich, take vitamins vs not, etc.
- ▶ You test one-by-one
- ▶ You find that for half a dozen factors, there is a difference
- ▶ Any special issue?

# Multiple testing

- ▶ The pre-set level of significance / p-value are defined for a single test
- ▶ In many cases, you will consider doing many many tests.
  - ▶ Different measures (mean, median, range, etc)
  - ▶ Different products, retailers, countries
  - ▶ Different measures of management quality
- ▶ For multiple tests, you cannot use the same approach as for a single one.

Hypothesis  
oooooooooooo

A1  
ooooooo

The t-test  
oooooooo

Making a decision  
oooooooooooooooo

p-value  
oooooo

A2  
ooo

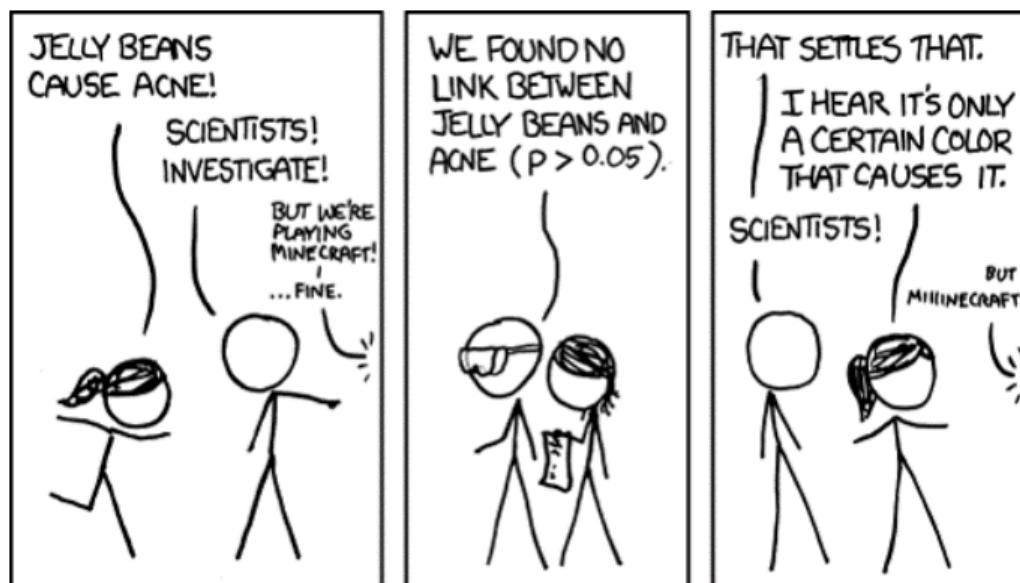
Multiple test  
ooo•oooo

Big Data  
o

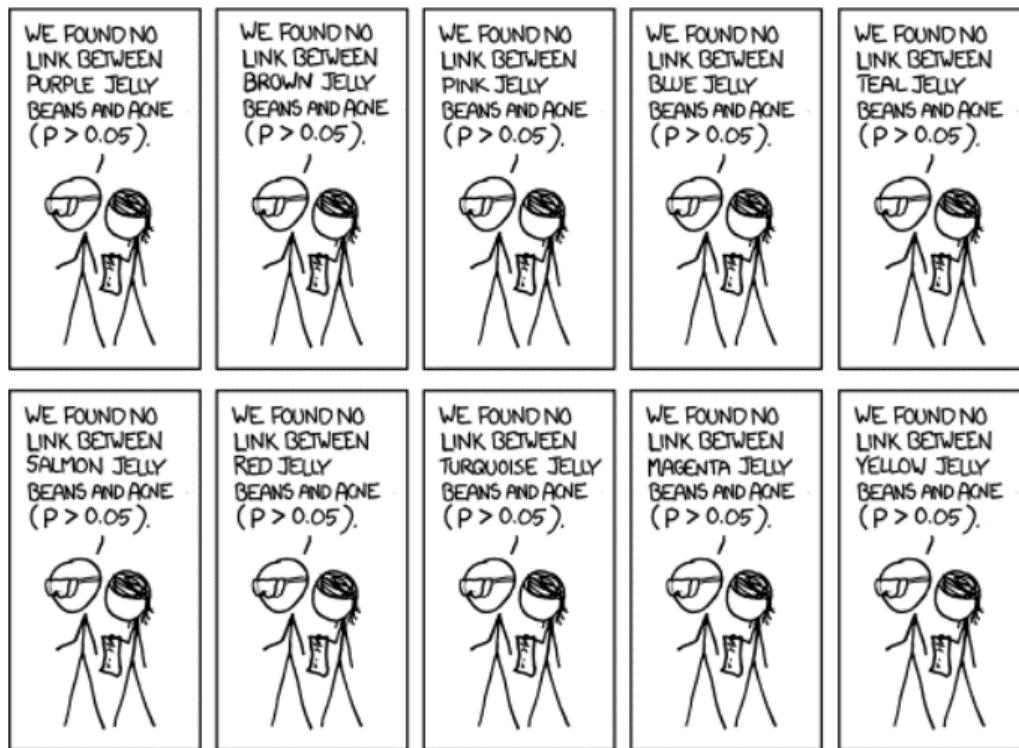
Sum  
o

Extra  
oooo

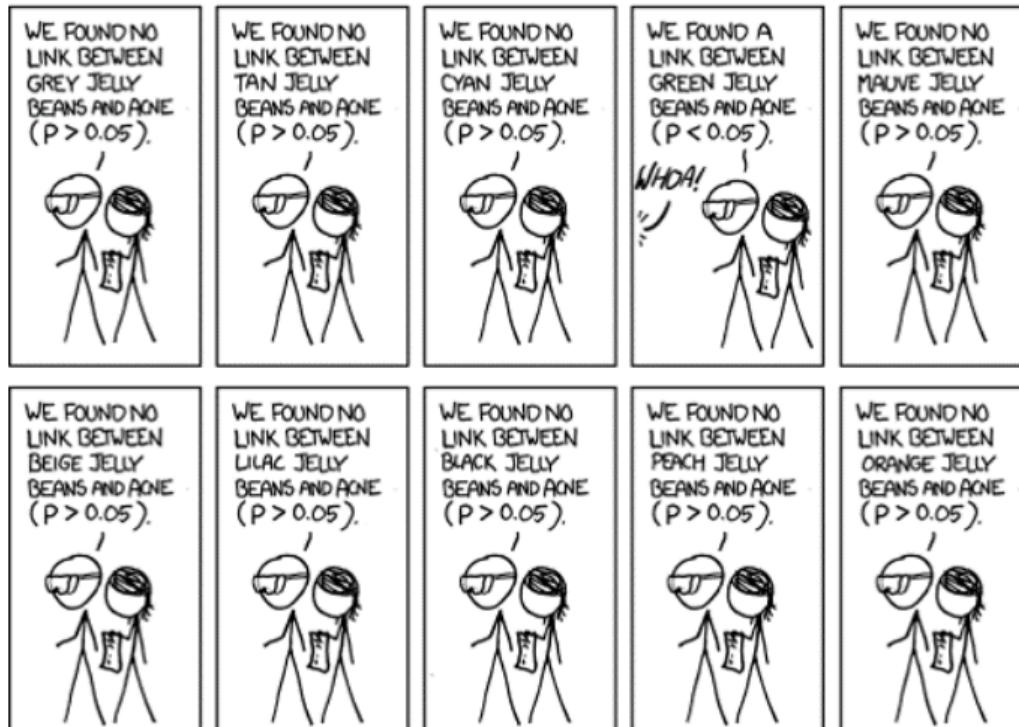
## Multiple testing - a serious example



# Multiple testing - a serious example



# Multiple testing - a serious example



Hypothesis  
oooooooooooo

A1  
ooooooo

The t-test  
oooooooo

Making a decision  
oooooooooooooooo

p-value  
oooooo

A2  
ooo

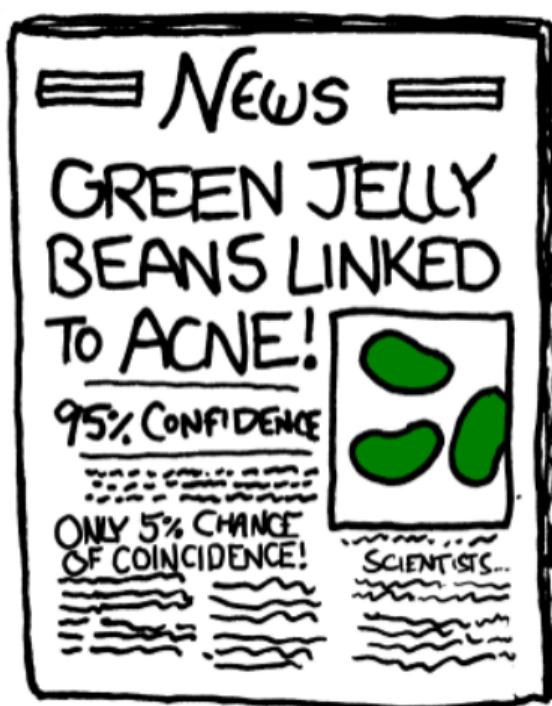
Multiple test  
oooooo•oooo

Big Data  
o

Sum  
o

Extra  
oooo

## Multiple testing - a serious example



## Multiple testing

- ▶ Consider a situation in which we test 100 hypotheses.
- ▶ Assume that all of those 100 null hypotheses are true.
  - ▶ Set significance - we accept 5% chance to be wrong when rejecting the null. That means that we tolerate if we are wrong 5 out of 100 times.
  - ▶ We can expect the null to be rejected 5 times when we test our 100 null hypotheses, all of which are true.
  - ▶ In practice that would appear in 5 out of the 100 tests
  - ▶ We could pick those five null hypotheses and say there is enough evidence to reject.
  - ▶ But that is wrong: we started out assuming that all 100 nulls are true.
- ▶ Simply by chance, we will see cases when we would reject the null, but we should not

Hypothesis  
oooooooooooo

A1  
ooooooo

The t-test  
oooooooooo

Making a decision  
oooooooooooooooooooo

p-value  
ooooooo

A2  
ooo

Multiple test  
oooooooooooo●○

Big Data  
o

Sum  
o

Extra  
oooo

# p-hacking

- ▶ Practice of doing many tests, and picking what works...

## Multiple testing

- ▶ There are various ways to deal with probabilities of false positives when testing multiple hypotheses.
- ▶ Often complicated.
- ▶ Solution 1: If you have a few dozens of cases, just use a strict criteria (such as 0.1-0.5% instead than 1-5%) for rejecting null hypotheses.
- ▶ A very strict such adjustment is the Bonferroni correction that suggests dividing the single hypothesis value by the number of hypotheses.
  - ▶ For example, if you have 20 hypotheses and aim for a  $p=.05$
  - ▶ reject the null only if you get a  $p=0.05/20=0.0025$
  - ▶ It is typically two strict

# Testing when data is very big

- ▶ Very large datasets – statistical inference lose relevance.
- ▶ Millions of observations generalizing to the general pattern does not add much.
- ▶ That is true for testing hypotheses, too.
- ▶ So: if you have millions of observations, just look at meaningful difference - do not worry about hypotheses testing (unless you care about very very small differences)

## Summary

Testing in statistics means making a decision about the value of a statistic in the general pattern represented by the data.

- ▶ Hypothesis starts with explicitly stating  $H_0$  and  $H_A$ .
- ▶ A statistical test rejects  $H_0$  if there is enough evidence against it; otherwise it does not reject it.
- ▶ Testing multiple hypotheses at the same time is a tricky business; it pays to be very conservative with rejecting the null.

Hypothesis  
oooooooooooo

A1  
ooooooo

The t-test  
oooooooo

Making a decision  
oooooooooooooooo

p-value  
oooooo

A2  
ooo

Multiple test  
oooooooooooo

Big Data  
o

Sum  
o

Extra  
●oooo

# Extra

## A special case in testing: the one sided-alternative

- ▶ Have only one of the inequalities in the alternative
- ▶ This leads to focusing on one side of the test statistic only
- ▶ Two most frequent examples are
  - ▶  $H_0 : s_{true} \leq 0$  against  $H_A : s_{true} > 0$
  - ▶  $H_0 : s_{true} \geq 0$  against  $H_A : s_{true} < 0$ .
- ▶ Having zero is key. If we can reject zero, we can reject anything below (above)
  - ▶ Test  $H_0 : s_{true} \leq 0$  vs  $H_A : s_{true} > 0 \rightarrow H_0 : s_{true} = 0$  vs  $H_A : s_{true} > 0$ .
  - ▶ Test  $H_0 : s_{true} \geq 0$  vs  $H_A : s_{true} < 0 \rightarrow H_0 : s_{true} = 0$  vs  $H_A : s_{true} < 0$ .

## One sided-alternative

- ▶ Focusing on deviations in one direction means that we care about one half of the sampling distribution of the test statistic.
- ▶ With  $H_0 : s_{true} \leq 0$  against  $H_A : s_{true} > 0$ , we care about whether  $\hat{s}$  is large positive enough in order to reject the null; if it is negative we don't reject it.
- ▶ The probability of a false positive is smaller in this case. We don't reject the null if the test statistic falls in the region that is specified in the null hypothesis.
- ▶ Thus, we make a false positive decision only half of the times.
- ▶ t-test of two-sided hypotheses — the p-value can be thought of as the sum of two probabilities
- ▶ So we only have half the probability of error

## One sided-alternative

Therefore, the practical way to testing one-sided hypotheses is a two-step procedure.

1. If the test statistic is in the region of the null don't reject the null.

This happens if  $\hat{s}$  is in the region of the null (e.g.,  $\hat{s} < 0$  for  $H_0 : s_{true} \leq 0$  against  $H_A : s_{true} > 0$ );

2. If the test statistic is in the region of the alternative proceed with testing the usual way with some modification.

Ask the software to calculate the p-value of the null hypothesis of the equality (for example,  $H_0 : s_{true} = 0$  if the true null is  $H_0 : s_{true} \leq 0$ ) and divide the p-value by two.