

Managerial Statistics

Lecture

Basic data concepts

Instructor Maxim Massenkoff

Today

Definitions

Data

Data types

Samples and populations

Another example

Activity

Definitions

We start with some key definitions:

- ▶ **Data:** Facts and figures collected, analyzed, and summarized for presentation and interpretation
- ▶ **Statistics:** The analysis and interpretation of data
- ▶ **Population:** the group under study (e.g. all people in the United States; businesses that make shoes)
- ▶ **Sample:** a portion of the population used for analysis
- ▶ **A statistic:** a property of a sample

Example

Political polling

- ▶ **Population:** Voters. The group we fundamentally want to know about
- ▶ **Sample:** the people we surveyed
- ▶ **A statistic:** the proportion who will vote for Biden

What is data?

- ▶ Here's an example of data (or, a **dataset**)

<u>Name</u>	<u>Age</u>	<u>Earnings</u>	<u>Full-time</u>
Oakley	22	\$20,000	no
Lennon	50	\$40,000	yes
Royal	31	\$50,000	yes
Skyler	37	\$15,000	no

- ▶ This is a **sample** of workers from ABC Corp, which has more than four employees
- ▶ **Statistics** help us summarize and interpret data
- ▶ An example of a **statistic**: half of the people here are full-time.

What is data?

Key definitions

<u>Name</u>	<u>Age</u>	<u>Earnings</u>	<u>Full-time</u>
Oakley	22	\$20,000	no
Lennon	50	\$40,000	yes
Royal	31	\$50,000	yes
Skyler	37	\$15,000	no

- ▶ **Observation**: a single measurement (a row)
- ▶ **Unit** (also known as an **Element**): the entities that the data refer to (Oakley, Lennon)
- ▶ **Variable**: a characteristic of interest (Name, Age)
- ▶ **n**: total observations (in this case, 4)

What is data?

<u>Name</u>	<u>Age</u>	<u>Earnings</u>	<u>Full-time</u>
Oakley	22	\$20,000	no
Lennon	50	\$40,000	yes
Royal	31	\$50,000	yes
Skyler	37	\$15,000	no

- ▶ Whenever you start working with a new dataset, **know what a row represents**
- ▶ In this case, 1 row = 1 person



U.S. Department
of Transportation
**National Highway
Traffic Safety
Administration**



DOT HS 812 827

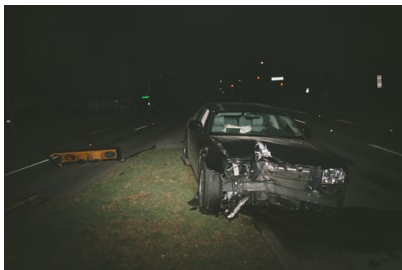
September 2019

Fatality Analysis Reporting System (FARS) Analytical User's Manual, 1975-2018

Fatality Analysis Reporting System

The US records data on every auto fatality

- ▶ What do we expect each observation to be?
 - ▶ An accident
- ▶ How else might the data be structured?
 - ▶ Vehicle, people
- ▶ What are some useful variables to have?
 - ▶ Angle of accident, Age of driver, Speed, Model of car



Fatality Analysis Reporting System

What can we learn?



White

Definitions

Example: The Pfizer vaccine trial

- ▶ What was the **data**?: information on whether people had the vaccine or a placebo, whether they had COVID, and whether it was a severe case
- ▶ **Population**: Human adults
- ▶ **Sample**: the people who participated in the study
- ▶ Key **statistic**: Comparison of the health outcomes in the two groups, namely efficacy of 95%

Definitions

Example: The Pfizer vaccine trial

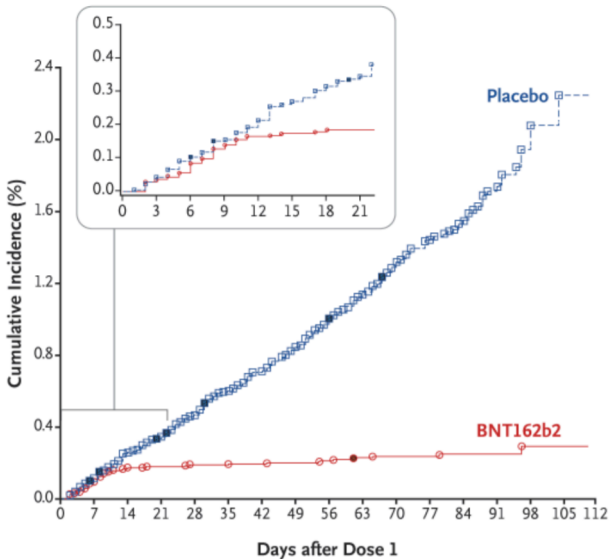
Data in Pfizer trial looked very roughly like this

Patient ID	Vaccine	COVID test	Test Date
1	Yes	Negative	3/15/2020
1	Yes	Negative	3/28/2020
2	No	Positive	3/17/2020
3	Yes	Negative	3/21/2020
⋮	⋮	⋮	⋮

Fundamentally not too fancy

The Pfizer trial

Main results



Data types

Let's think more about the car accident data to learn about different data types

- ▶ **Quantitative**: Numeric data. Examples: Age of driver, speed, hour of crash.
- ▶ **Categorical** (also known as **qualitative**): Data indicating groups. Sedan, limo, pickup.
 - ▶ Special kind: **Indicator (dummy)**: only takes on two values, 0 or 1. E.g. Alcohol-involved
- ▶ **Ordinal**: weather conditions (1=least favorable, 7=most favorable)

Data sources

- ▶ **Observational data:** data collected (e.g., through a survey) without altering any variables
 - ▶ Example: the US Census
 - ▶ Unemployment rate
 - ▶ Inflation
- ▶ **Experimental data:** data collected while *manipulating* certain variables
 - ▶ A drug trial generates experimental data (there is experimental control over medical treatment, the variable of interest)
 - ▶ Amazon running “A/B” tests of different webpage designs

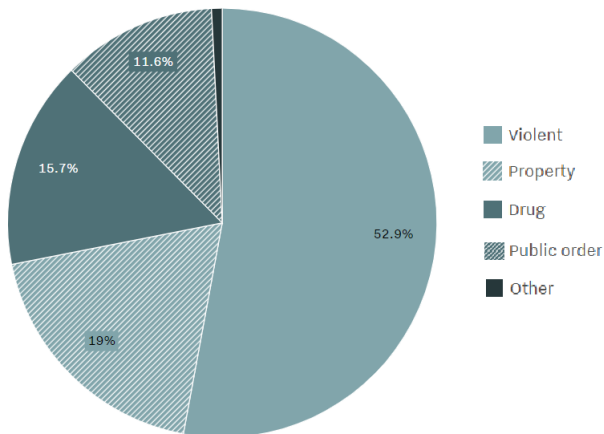
Cross sectional vs. Time series data

- ▶ **Cross-sectional data:** A *snapshot*. One observation per unit. Cannot use it to study a given unit changing over time.
 - ▶ Heights of every new recruit at enlistment
 - ▶ Annual tax revenue for each of the 50 states in 2020
- ▶ **Time series or panel data:** Variables on the same units at multiple points in time
 - ▶ Each college student's GPA freshman-senior year (4 observations per unit)
 - ▶ Annual tax revenue for each of the 50 states in the years 2000-2020

Examples: Cross sectional data

Why people are in state prison

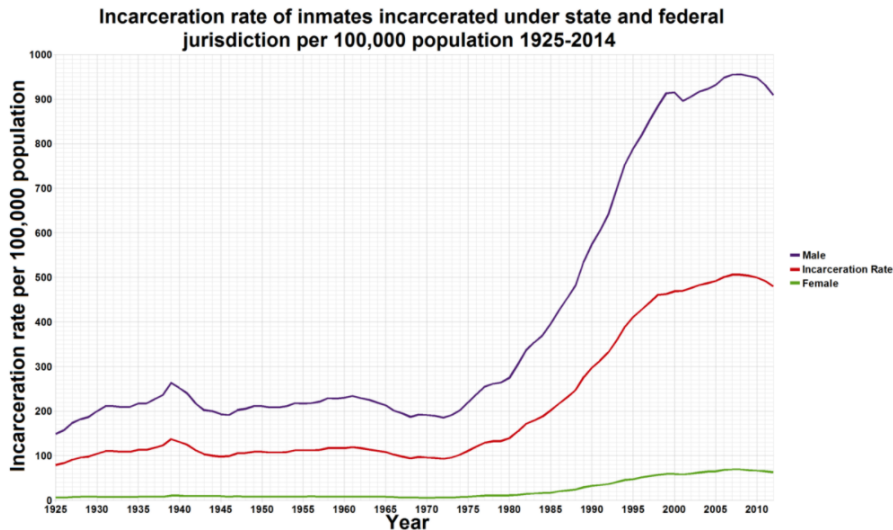
Percent of state prisoners in 2014, based on most serious offense



This is a **categorical** variable

source

Examples: Time series



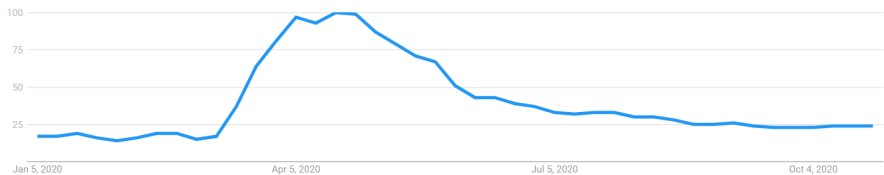
This is a **quantitative** variable

source

Examples: Time series

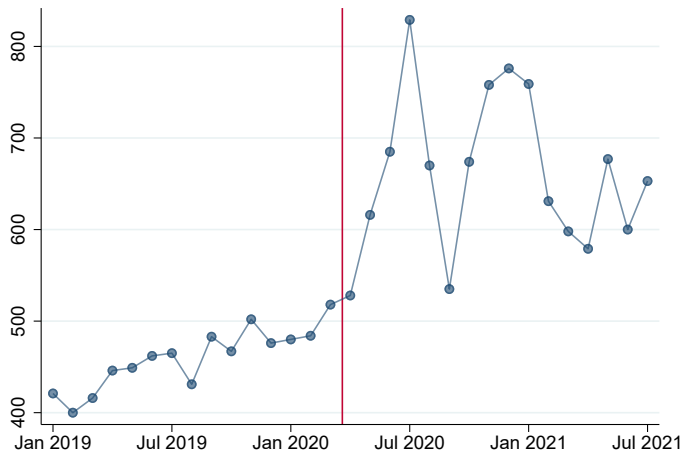
Google searches for “sourdough” in 2020

Interest over time ?



Examples: Time series

Monthly vehicle thefts in San Francisco



Samples and populations

- ▶ A key distinction in statistics is whether we're working with a **sample** or the full **population**
- ▶ The **population** is the entire group of interest.
- ▶ Example: What's the US unemployment rate? The population of interest is every American in the labor force
- ▶ If Bureau of Labor Statistics calls 50,000 people to calculate the unemployment rate, their estimate is based on a **sample**

Sampling

How should we choose our sample?

- ▶ In a simple random sample, all units in the population have an equal probability of being selected
- ▶ Samples can be drawn **with or without replacement**
- ▶ Without replacement: someone can only be drawn once

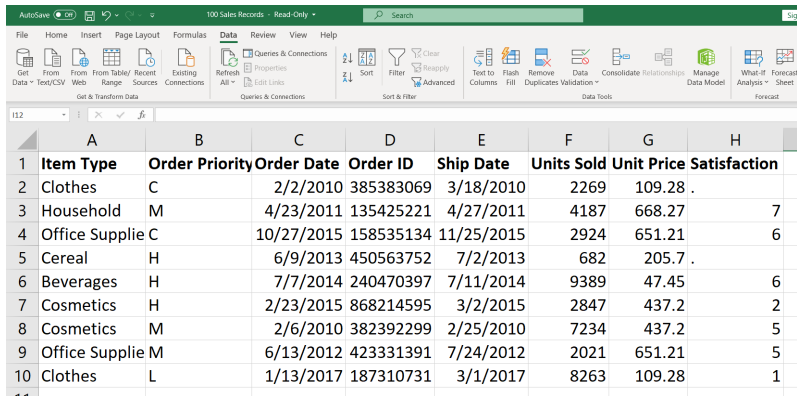
Sampling

Seatbelt survey

- ▶ It's hard to survey US drivers, but we know that about 50% of American drivers killed in car accidents are not wearing seatbelts
- ▶ We conclude that about 50% of American drivers wear seatbelts
- ▶ Any issues with this conclusion?

Let's work through one more example

Excel screenshot



Excel screenshot showing a data table with 10 rows and 8 columns. The ribbon is set to 'Data' with 'Sort & Filter' selected. The table contains sales data with columns: Item Type, Order Priority, Order Date, Order ID, Ship Date, Units Sold, Unit Price, and Satisfaction.

	A	B	C	D	E	F	G	H
1	Item Type	Order Priority	Order Date	Order ID	Ship Date	Units Sold	Unit Price	Satisfaction
2	Clothes	C	2/2/2010	385383069	3/18/2010	2269	109.28	
3	Household	M	4/23/2011	135425221	4/27/2011	4187	668.27	7
4	Office Supplie	C	10/27/2015	158535134	11/25/2015	2924	651.21	6
5	Cereal	H	6/9/2013	450563752	7/2/2013	682	205.7	
6	Beverages	H	7/7/2014	240470397	7/11/2014	9389	47.45	6
7	Cosmetics	H	2/23/2015	868214595	3/2/2015	2847	437.2	2
8	Cosmetics	M	2/6/2010	382392299	2/25/2010	7234	437.2	5
9	Office Supplie	M	6/13/2012	423331391	7/24/2012	2021	651.21	5
10	Clothes	L	1/13/2017	187310731	3/1/2017	8263	109.28	1

► *Always know what each row represents*

► In this case a sales order

Let's work through one more example

Excel screenshot

The screenshot shows an Excel spreadsheet with a table of 10 sales records. The columns are labeled A through H. Below the table, three blue arrows point from specific columns to data type labels: 'Categorical' points to column A (Item Type), 'Quantitative' points to column E (Ship Date), and 'Ordinal' points to column H (Satisfaction).

	A	B	C	D	E	F	G	H
1	Item Type	Order Priority	Order Date	Order ID	Ship Date	Units Sold	Unit Price	Satisfaction
2	Clothes	C	2/2/2010	385383069	3/18/2010	2269	109.28	
3	Household	M	4/23/2011	135425221	4/27/2011	4187	668.27	7
4	Office Supplie	C	10/27/2015	158535134	11/25/2015	2924	651.21	6
5	Cereal	H	6/9/2013	450563752	7/2/2013	682	205.7	
6	Beverages	H	7/7/2014	240470397	7/11/2014	9389	47.45	6
7	Cosmetics	H	2/23/2015	868214595	3/2/2015	2847	437.2	2
8	Cosmetics	M	2/6/2010	382392299	2/25/2010	7234	437.2	5
9	Office Supplie	M	6/13/2012	423331391	7/24/2012	2021	651.21	5
10	Clothes	L	1/13/2017	187310731	3/1/2017	8263	109.28	1

Categorical

Quantitative

Ordinal

► Each row is a sales order

Activity (if time)

- ▶ Next, a group activity
- ▶ Let's look through some news articles and figure out how our core concepts relate

News examples for group activity

Look through some news articles on research. Report the:

- ▶ Unit of observation: imagining the data analyzed, what would each row represent?
- ▶ Sample: which units are analyzed, and what is the implicit population?
- ▶ Key variable: one of the variables central to the analysis? Is it qualitative or quantitative discrete/continuous?
- ▶ Was the data observational or experimental?
- ▶ Was the data cross-sectional or time series?
- ▶ Any criticism of the study

News examples for group activity

Example: my MN3040 survey

- ▶ Unit of observation: **student**
- ▶ Sample: **people who responded to the survey**, population: all MN3040 students
- ▶ Key variable: **cups of caffeine, quantitative discrete**
- ▶ Was the data observational or experimental? **Observational**
- ▶ Was the data cross-sectional or time series? **Cross-sectional**
- ▶ Any criticism of the study: **responses?**

News examples for group activity

Look through some news articles on research. Report the:

1. Unit of observation: imagining the data analyzed, what would each row represent?
2. Sample: which units are analyzed?
3. Key variable: one of the columns central to the analysis?
Qualitative or quantitative discrete/continuous?
4. Was the data observational or experimental?
5. Was the data cross-sectional or time series?
6. Any criticism of the study

News examples for group activity

Look through some news articles on research. Report the:

1. Unit of observation: imagining the data analyzed, what would each row represent?
2. Sample: which units are analyzed?
3. Key variable: one of the columns central to the analysis?
Qualitative or quantitative discrete/continuous?
4. Was the data observational or experimental?
5. Was the data cross-sectional or time series?
6. Any criticism of the study

In 5-10 minutes I'll have you come back and explain the study and these different features to the class.

Article 1 (link)



PUBLIC RELEASE: 22-APR-2009

New study shows chewing gum can lead to better academic performance in teenagers

Higher math scores seen in classroom setting

EDELMAN PUBLIC RELATIONS

WHAT: New research from Baylor College of Medicine indicates a positive effect of chewing gum on academic performance in teenagers.¹ The study examined whether chewing Wrigley sugar-free gum can lead to better academic performance in a "real life" classroom setting. Major findings include:

- The researchers found that students who chewed gum showed an **increase in standardized math test scores and their final grades were better compared** to those who didn't chew gum.
 - Students who chewed gum had a significantly greater increase in their standardized math test scores after 14 weeks of chewing gum in math class and while doing homework compared to those who did not chew gum. Chewing gum was associated with a three percent increase in standardized math test scores, a small but statistically significant change.
 - Students who chewed gum had final grades that were significantly better than those who didn't chew gum.

Today's competitive testing environment has parents and students looking for approaches to improve academic performance, particularly as standardized test scores have become a mandatory requirement for assessing academic achievement. Together, these findings can be meaningful when related to small steps that can lead to better academic performance.

Why the #\$\$%! Do We Swear? For Pain Relief

Dropping the F-bomb or other expletives may not only be an expression of agony, but also a means to alleviate it

By Frederik Joelving on July 12, 2009

Bad language could be good for you, a new study shows. For the first time, psychologists have found that swearing may serve an important function in relieving pain.

The study, published today in the journal *NeuroReport*, measured how long college students could keep their hands immersed in cold water. During the chilly exercise, they could repeat an expletive of their choice or chant a neutral word. When swearing, the 67 student volunteers reported less pain and on average endured about 40 seconds longer.

READ THIS NEXT

MIND

Profanity Bleeps Physical Pain

July 13, 2009

NEWS

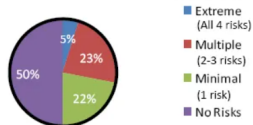
Study: Less risky behavior among students

Students in Hamilton County tend to have less risky behavior than they did 13 years ago, according to the Health Department.

Tuesday, October 4th 2011, 1:55 PM EDT

Updated: Tuesday, October 4th 2011, 1:58 PM EDT

Number of Current Risk Behaviors



Article 4 (link)

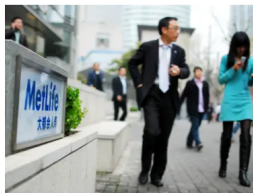
[HOME](#) > [MARKETS CONTRIBUTORS](#)

US Companies Are Getting Squeezed Hard In China



Adam Jourdan, Reuters Feb 28, 2013, 4:19 AM

SHANGHAI ([Reuters](#)) - Profitability and sales are harder to come by in China as U.S. firms face increasing competition from domestic and foreign players, said a U.S. business group survey on Thursday.



Yepoka Yeebo / Business Insider

An annual survey by the American Chamber of Commerce in Shanghai showed a majority of firms believed that competition had intensified, while the number who said they were profitable in 2012 dropped to 73 percent from 78 percent in 2011.

Article 5 ([link](#))

Beetroot Juice Boosts Stamina, New Study Shows

Date: August 7, 2009

Source: University of Exeter

Summary: Drinking beetroot juice boosts your stamina and could help you exercise for up to 16 percent longer. A new study shows for the first time how the nitrate contained in beetroot juice leads to a reduction in oxygen uptake, making exercise less tiring. The study reveals that drinking beetroot juice reduces oxygen uptake to an extent that cannot be achieved by any other known means, including training.

Share:

RELATED TOPICS

Health & Medicine

- > [Fitness](#)
- > [Pharmaceuticals](#)
- > [Pharmacology](#)
- > [Staying Healthy](#)

Plants & Animals

- > [Agriculture and Food](#)
- > [Biology](#)
- > [Beer and Wine](#)
- > [Misc](#)

FULL STORY

Drinking beetroot juice boosts your stamina and could help you exercise for up to 16% longer. A University of Exeter led-study shows for the first time how the nitrate contained in beetroot juice leads to a reduction in oxygen uptake, making exercise less tiring.

The study reveals that drinking beetroot juice reduces oxygen uptake to an extent that cannot be achieved by any other known means, including training.

The research team believes that the findings could be of great interest to endurance athletes. They could also be relevant to elderly people or those with cardiovascular respiratory or

Thanks