ØAMET2200
Business Decision Making Using Data
Lecture 1

Instructor: Fenella Carpena

August 23, 2019

Agenda for Today

- Syllabus and Course Information
- Review Chapters 1-12 of Textbook (Stine and Foster)
  - Probability
  - Random Variables
  - The Normal Distribution
- Practice Problems
- Introduction to Stata

Course Logistics

- Fridays, 8:30AM-12:10PM, Stensberggata 26-28 Room X158
- Each lecture divided into three 1-hour blocks

  Block 1 08:30-09:30, then 20-minute break

  Block 2 09:50-10:50, then 20-minute break

  Block 3 11:10-12:10, end of lecture

- Office Hours: To be determined.
- Email Policy: For any questions, **post a discussion** on Canvas
  - I reserve the right to ignore questions emailed directly to me and not posted on Canvas
  - Exceptions: sensitive information, administrative questions that concern only yourself. Email `fenella.carpena@oslomet.no` with `[ØAMET2200]` in the subject line.

Course Materials

- Textbook: Stine and Fosters *Statistics for Business: Decision Making and Analysis*, Pearson New International Edition
  - We will follow the textbook closely
  - The course is highly cumulative: **if you fall behind, it will be difficult to catch up**
- Software: programming using Stata
- Problem Sets: There will be 4 problem sets (*arbeidskrav*)
  - Submit individually or in groups of up to 4 students
  - Each graded 0 or 1 points; Approval = 1 point
  - Need all problem sets approved before you can take final exam
- Final Exam: Can be answered **only in English**

Course Schedule

Lecture	Date	Topics	Reading
1	August 23	Introduction and Review	Review Chapters 1-12
2	August 30	Samples and Surveys Sampling Variation and Quality	Chapter 13 Chapter 14.1-14.2
3	September 6	Confidence Intervals Statistical Tests	Chapter 15 Chapter 16
4	September 13	Comparison	Chapter 17
5	September 20	Linear Patterns Curved Patterns	Chapter 19 Chapter 20
6	October 4	Simple Regression Model Regression Diagnostics	Chapter 21 Chapter 22
7	October 11	Multiple Regression	Chapter 23
8	October 18	Building Regression Models	Chapter 24
9	October 25	Categorical Explanatory Variables	Chapter 25
10	November 1	Time Series	Chapter 27
11	November 8	Regressions with Big Data Catch Up and Review	Chapter 29

The course plan is subject to change throughout the semester.
There is no lecture on September 27.

What is this course about?

Getting data analysis right

- Businesses have access to **massive** amounts of data
- Many behaviors used to be **too personal** or **unobserved**
- Now, collected on a regular and continuous basis through everyday **technology**
  - Online browsing activity
  - Cellphone location
  - Usage of apps
  - Wearable devices (e.g., Apple Watch, Fitbit)
  - Smart home devices (e.g., Google Home, Amazon Echo)
- Lots of **opportunities** for businesses using these data

Examples abound of how valuable data can be

THE WALL STREET JOURNAL.
Europe Edition ▾ | March 4, 2019 | Print Edition | Video

TECHNOLOGY

On Orbitz, Mac Users Steered to Pricier Hotels

Orbitz has found that Apple users spend as much as 30% more a night on hotels, so the online travel site is starting to show them different, and sometimes costlier, options than Windows visitors see. Dana Mattioli has details on The News Hub. Photo: Bloomberg.

By Dana Mattioli

Updated Aug. 23, 2012 6:07 p.m. ET

Orbitz Worldwide Inc. has found that people who use Apple Inc.'s Mac computers spend as much as 30% more a night on hotels, so the online travel agency is starting to show them different, and sometimes costlier, travel options than Windows visitors see.

The Orbitz effort, which is in its early stages, demonstrates how tracking people's online activities can use even seemingly innocuous information—in this case, the fact that customers are visiting Orbitz.com from a Mac—to start predicting their tastes and

But it's not just about having data

March 28, 2014 11:38 am

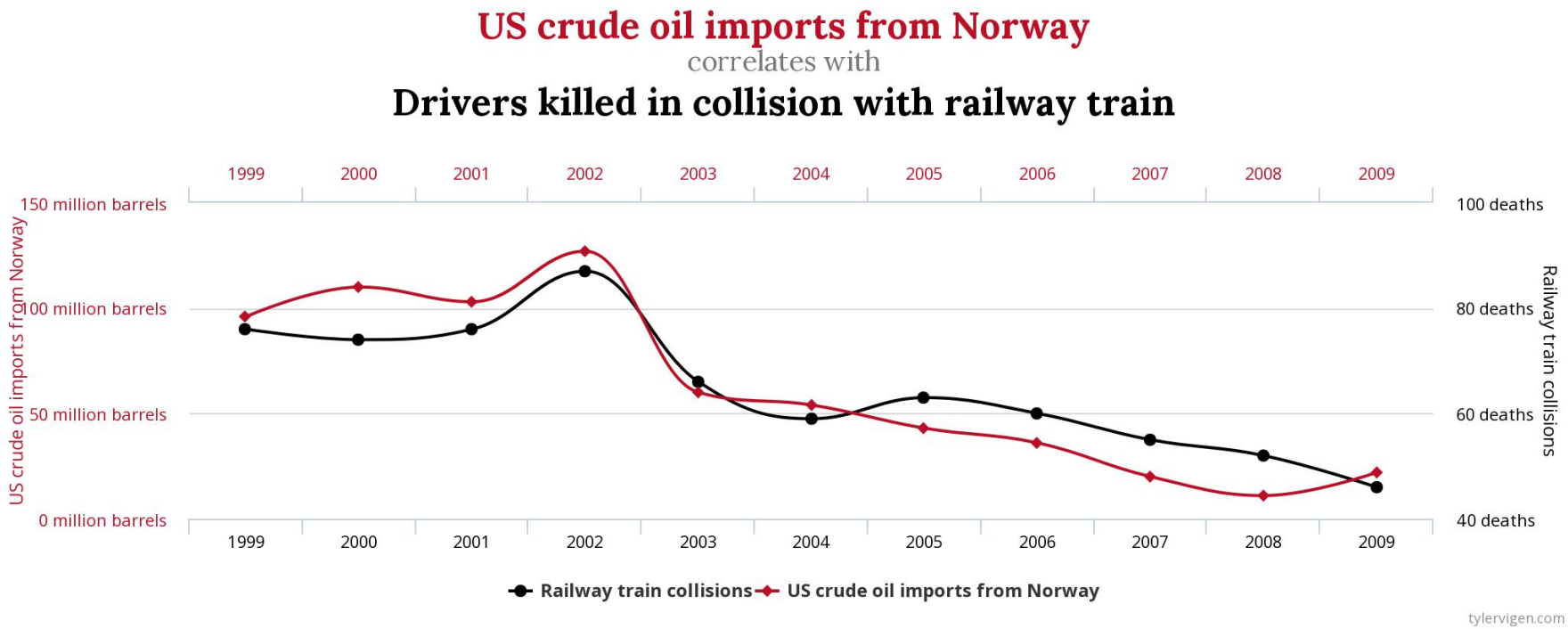
Big data: are we making a big mistake?

By Tim Harford

Big data is a vague term for a massive phenomenon that has rapidly become an obsession with entrepreneurs, scientists, governments and the media

"There are a lot of small data problems that occur in big data," says Spiegelhalter. "They don't disappear because you've got lots of the stuff. They get worse."

Possibilities for untold numbers of spurious correlations



Course Objectives

- Analyzing data requires careful knowledge and constant vigilance.
- It's knowing how to **produce** it, how to **use** it, and how to **interpret** it.
- Even more important: **critical thinking** when making business decisions under limited information.
- Overall goal: prepare you for you professional careers by making you better producers and consumers of data.

Course Overview

Part 1 Data You Want vs. Data You Have (Lectures 2-3)

Part 2 Making Decisions Based on Inference (Lectures 3-4)

Part 3 Experiments and Causality (Lectures 5-6)

Part 4 Building Regression Models (Lectures 7-11)

Part 1: Data you want vs. Data you have

- Lectures 2-3
- Business decision-makers never have all the data they want
  - about their customers
  - about their employees' productivity
  - about their competitors
- If you can only collect **a sample of data**, what can you say about what is generally going on?
- Topics covered:
  - Samples and Surveys (Chapter 13)
  - Sampling Variation and Quality (Chapter 14.1-14.2)
  - Confidence Intervals (Chapter 15)

Part 1: Data you want vs. Data you have

Questions of Interest

- Suppose you want to **market a new app**.
  - How do you learn if (potential) customers will like it?
  - How do you learn how much (potential) customers are willing to pay?
- You need to sample from a population
  - How you **select your sample** matters
  - Who selects into your sample matters
- What can you say about potential **customers you will never observe**?

Part 2: Making Decisions Based on Inference

- Lectures 3-4
- Given **limited data**, business decision-makers need to choose how to act on it.
  - You have some benchmark or two groups you want to compare.
  - How do you know any difference isn't due to noise?
  - What is the likelihood of making a mistake in your decision?
- We need **statistical tests** that provide us with some clarity.
- Topics covered:
  - Statistical Tests (Chapter 16)
  - Comparison (Chapter 17)

Part 2: Making Decisions Based on Inference

Questions of Interest

- General Motors observes some faulty ignitions switches.
  - Should it issue a recall?
- Samsung pilots a new production process in 10% of its manufacturing plants.
  - Should the company adopt this process more broadly?
- Philips Healthcare develops a new medical diagnostic device.
  - Does it predict disease better than existing devices?

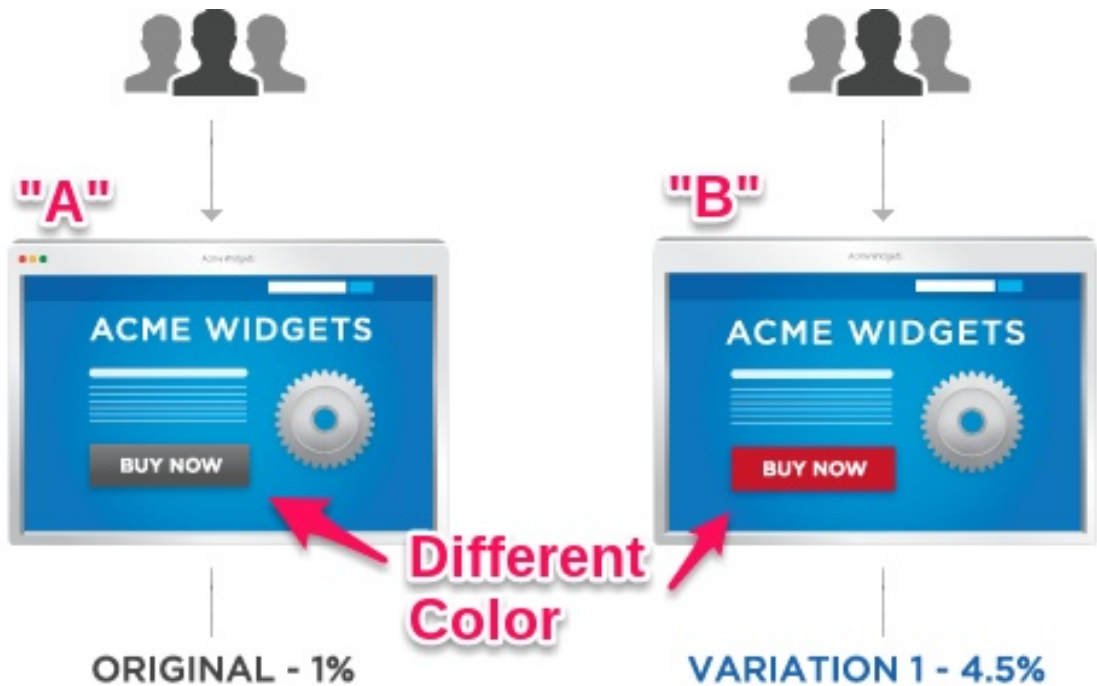
Part 3: Experiments and Causality

- Lectures 5-6
- In many comparisons you make, you want to attribute differences to some **causal factor**.
- **Experiments** are a uniquely powerful and transparent method of testing causal effects.
- Topics covered:
  - Experiment Design
  - Hawthorne Effects

Part 3: Experiments and Causality

Questions of Interest

- A/B Testing: You want to **redesign** a website to increase sales.
- Does a red Buy Now button make customers more likely to buy a product than a gray Buy Now button?



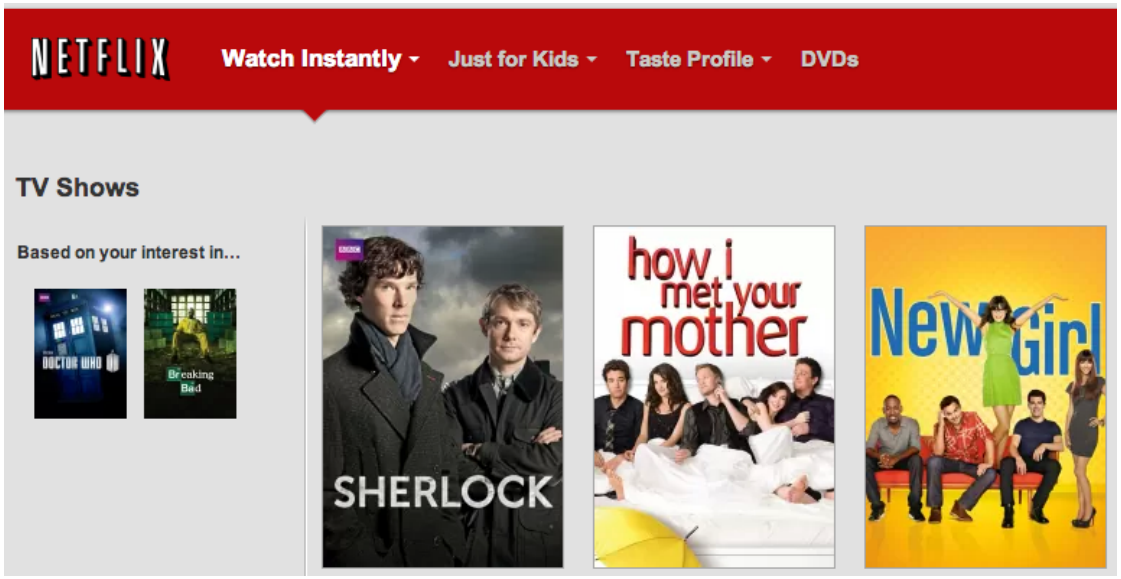
Part 4: Building Regression Models

- Lectures 7-11
- You want to build a model to **forecast** outcomes using data that you have.
- What is the **best model** you can build?
- Topics covered:
  - Linear and Curved Patterns (Chapter 19-20)
  - Simple Regression Model (Chapter 21-22)
  - Multiple Regression (Chapter 23-25)
  - Time Series (Chapter 27)
  - Regressions with Big Data (Chapter 29)

Part 4: Building Regression Models

Questions of Interest

- You want to recommend a new product to customers.



- You want to estimate the effect of competitors on your sales.
- You want to predict the performance of a stock.

Questions?

Random Variables

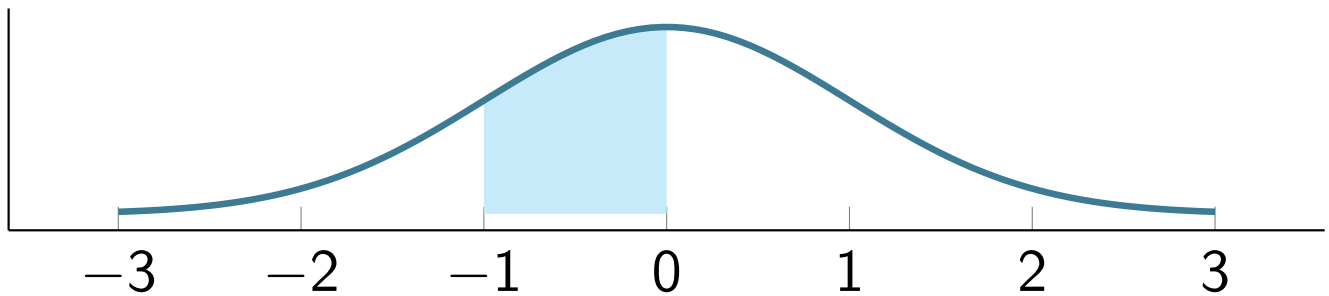
- Random variables represent the outcome of a random process
  - Can be either **discrete** or **continuous**
  - We denote a r.v. using capital letters X, Y, Z, etc.
  - Lower-case letters e.g., x, indicate a generic possible value of the random variable
- Examples of random variables
  - The # obtained from rolling a fair die (discrete)
  - Total # of "Heads" when a coin is tossed two times (discrete)
  - Consumption of natural gas (continuous)
  - Weight of a bag of Sørland potato chips (continuous)
- Can you think of other examples of random variables? Is the example you provided a discrete or continuous?

Probability Distribution

- Each r.v. is associated with a function that maps every potential outcome to a probability
  - Discrete: **probability mass function** (PMF)
  - Continuous: **probability density function** (PDF)
- Example of a PMF: Let X be the number obtained from rolling a fair die. Then the PMF of X is:

x	1	2	3	4	5	6
$P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6

- Example of a PDF: Let Z be a standard normal r.v. The area under the PDF is $P(-1 \leq Z \leq 0) = 0.3413$.



Joint Distributions

- The **joint probability distribution** of two discrete random variables X and Y gives the probability for simultaneous outcomes, $P(X = x, Y = y)$.
- Example (Worksheet 1, Exercise 4): Customers at MAX Burgers buy both burgers and drinks. The following is the joint distribution of the number of burgers (X) and drinks (Y) bought by customers.

		X	
		1 burger	2 burgers
Y	1 drink	0.40	0.20
	2 drinks	0.10	0.25
	3 drinks	0.00	0.05

- What does $P(X = 1, Y = 2)$ mean in words? What is its value?

Independence

- Two discrete r.v.'s X and Y are **independent** if and only if $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$ for all x and y.
- X doesn't give us information about Y, and vice versa.
- The opposite of independent is **dependent**.
- Return to the MAX Burgers example. Are X and Y independent?

		X	
		1 burger	2 burgers
Y	1 drink	0.40	0.20
	2 drinks	0.10	0.25
	3 drinks	0.00	0.05

Features of Random Variables: Expectation, Variance, SD

- **Expected value**: weighted average of all possible values.
  - Notation: $E(X)$, μ, or μ_X
  - Discrete case: $E(X) = \sum_{i=1}^{N} x_i \cdot P(X = x_i)$, where N is the total number of possible values of X

- **Variance**: a measure of how much the distribution is tightly centered around its mean.
  - Notation: $var(X)$, σ^2, or σ_X^2
  - $var(X) = E[(X - \mu)^2] = E(X^2) - \mu^2$

- **Standard deviation**: the positive square root of the variance.
  - Notation: $sd(X)$, σ, or σ_X
  - $sd(X) = \sqrt{var(X)}$

Worksheet 1, Exercise 4: MAX Burgers

		X	
		1 burger	2 burgers
Y	1 drink	0.40	0.20
	2 drinks	0.10	0.25
	3 drinks	0.00	0.05

(c) What is $E(X)$? $Var(X)$? $SD(X)$?

Worksheet 1, Exercise 4: MAX Burgers

		X	
		1 burger	2 burgers
Y	1 drink	0.40	0.20
	2 drinks	0.10	0.25
	3 drinks	0.00	0.05

(d) What is $E(Y)$? $Var(Y)$? $SD(Y)$?

Features of Random Variables: Covariance and Correlation

- **Covariance**: measures the amount of linear dependence between two r.v.'s
  - Notation: $cov(X, Y)$ or σ_{XY}
  - $cov(X, Y) = E(X - \mu_X)(Y - \mu_Y) = E(XY) - \mu_X \mu_Y$

- **Correlation**: a scale-free measure of linear dependence between two r.v.'s
  - Notation: $corr(X, Y)$, ρ, or ρ_{XY}
  - $corr(X, Y) = \frac{cov(X,Y)}{sd(X)sd(Y)}$

Worksheet 1, Exercise 4: MAX Burgers

		X	
		1 burger	2 burgers
	1 drink	0.40	0.20
Y	2 drinks	0.10	0.25
	3 drinks	0.00	0.05

(c) What is $cov(X, Y)$?

(d) What is $corr(X, Y)$?

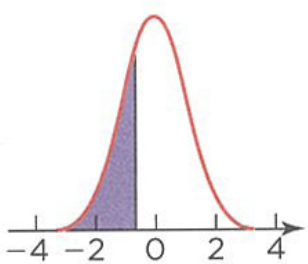
The Normal Distribution

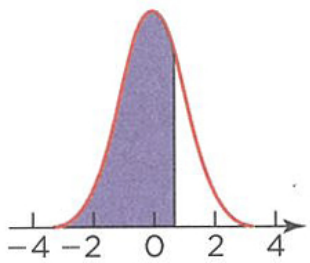
- Really important for this course!
- If X is a **normally distributed r.v.**, we write $X \sim \mathcal{N}(\mu, \sigma^2)$.
- Special case: **standard normal distribution**, $\mu = 0$, $\sigma^2 = 1$
- **Important property:** Any normal r.v. $X \sim \mathcal{N}(\mu, \sigma^2)$ can be transformed to a standard normal r.v. using the formula

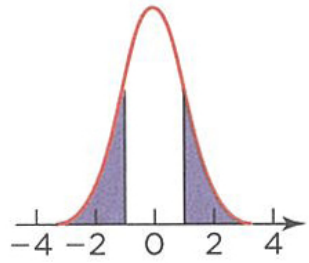
$$Z = \frac{X - \mu}{\sigma}$$

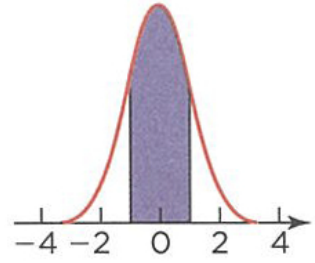
- Why do we care? This transformation allows us to find the probabilities of any normal r.v. using the standard normal table (also called Z Table).

How do we read the Z Table?









z	$P(Z \leq -z)$	$P(Z \leq z)$	$P(\|Z\| > z)$	$P(\|Z\| \leq z)$
0	0.50	0.50	1	0
0.0502	0.48	0.52	0.96	0.04
0.1004	0.46	0.54	0.92	0.08
0.1510	0.44	0.56	0.88	0.12
0.2019	0.42	0.58	0.84	0.16
0.2533	0.40	0.60	0.80	0.20

(a) $P(Z \leq -0.0502)$

(b) $P(Z \leq 0.2533)$

(c) $P(|Z| > 0.1510)$

(d) $P(|Z| \leq 0.2019)$

Worksheet 1, Exercise 9: Tire Manufacturer

A tire manufacturer warranties its tires to last at least 20,000 miles or "you get a new set of tires." In its experience, a set of these tires lasts on average 26,000 miles with SD 5,000 miles. Assume that wear is normally distributed. The manufacturer profits $200 on each set sold, and replacing a set costs the manufacturer $400.

(a) What is the probability that a set of tires wears out before 20,000 miles?

Worksheet 1, Exercise 9: Tire Manufacturer

A tire manufacturer warranties its tires to last at least 20,000 miles or "you get a new set of tires." In its experience, a set of these tires lasts on average 26,000 miles with SD 5,000 miles. Assume that wear is normally distributed. The manufacturer profits \$200 on each set sold, and replacing a set costs the manufacturer \$400.

(b) What is the probability that the manufacturer turns a profit on selling a set to one customer?