INTRODUCTION

To Data Science

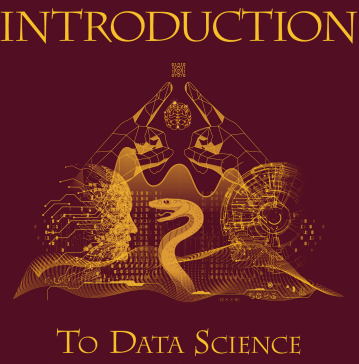
Introduction to Data Science

Probabilistic Thinking - Class 6

Giora Simchoni

gsimchoni@gmail.com and add #intro2ds in subject

Stat. and OR Department, TAU



Conditional Probability

INTRODUCTION

TO DATA SCIENCE

Reminder: Discrete Empirical Distributions

- Marginal distribution: $P(X = x_k), P(Y = y_l)$
- Joint distribution: $P(X = x_k, Y = y_l)$
- Conditional distribution: $P(Y = y_l|X = x_k), P(X = x_k|Y = y_l)$
- Most interesting: conditional distribution

Reminder: The Contingency Table

Y/X	x_1	x_2	x_3	**Total**
y_1	$P(x_1, y_1)$	$P(x_2, y_1)$	$P(x_3, y_1)$	$P(Y = y_1)$
y_2	$P(x_1, y_2)$	$P(x_2, y_2)$	$P(x_3, y_2)$	$P(Y = y_2)$
Total	$P(X = x_1)$	$P(X = x_2)$	$P(X = x_3)$	1

Reminder: Important Laws

Given two random variables X, Y, we have:

- Bayes' Law:

$$Pr(Y = y|X = x) = \frac{Pr(Y = y, X = x)}{Pr(X = x)} = \frac{Pr(X = x|Y = y)Pr(Y = y)}{Pr(X = x)}$$

- Law of total probability:

$$Pr(X = x) = \sum_y Pr(X = x, Y = y) = \sum_y Pr(X = x|Y = y)Pr(Y = y)$$

- The two are often combined to give:

$$Pr(Y = y_1|X = x) = \frac{Pr(X = x|Y = y_1)Pr(Y = y_1)}{\sum_y Pr(X = x|Y = y)Pr(Y = y)}$$

- In this way, knowing only about $Pr(X|Y)$ and $Pr(Y)$ teaches us about $Pr(Y|X)$

Example: Corona testing

Assume we have a Corona test described as **super accurate** meaning:

- 99% of people who are carriers test positive
- 99% of people who are healthy test negative

- Question: given I get randomly tested and get a positive test, what is the probability that I am actually a carrier?

- What if I add in this contingency table?

	predict healthy	**predict sick**	**Total**
really healthy	99000	1000	100000
really sick	1	99	100
Total	99001	1099	100100

- Answer: Assuming the disease is rare (say 1/1000 = 0.001), that probability is very small even for an *accurate* test!

Formulation in terms of conditional probabilities

- Define two Bernoulli variables:
  - $Y \in \{0, 1\}$ — carrier or not ; $X \in \{0, 1\}$ — positive test or not
- Given values:

$$Pr(X = 1|Y = 1) = Pr(X = 0|Y = 0) = 0.99 \ , \quad Pr(Y = 1) = 0.001$$

- We are interested in $Pr(Y = 1|X = 1)$, using our formulas above we get:

$$Pr(Y = 1|X = 1) = \frac{Pr(X = 1|Y = 1)Pr(Y = 1)}{Pr(X = 1)} = \frac{0.99 \cdot 0.001}{0.99 \cdot 0.001 + 0.01 \cdot 0.999}$$

- Conclusion: If you get a positive result in this accurate test, you still have $< 10\%$ chance of b actual carrier

Simpson's Paradox

INTRODUCTION

To Data Science

UC Berkeley Gender Bias Study

- A well known research studying relation between:
  - Gender
  - Admission to Berkeley
  - Department

When checking relation between Gender and Admission:

	Men	**Women**	**Total**
Not Admitted	150	220	370
Admitted	220	150	370
Total	370	370	540

Men applying are more likely to be admitted?

UC Berkeley Gender Bias Study

- When conditioning on Department, the conclusion reverses:

Department		Men	Women	Total
A	Not Admitted	50	200	250
	Admitted	20	100	120
B	Not Admitted	100	20	120
	Admitted	200	50	250
Total	Total	370	370	540

- Conclusion: Women applied more for department A, with lower admission rates!

- Simpson's Paradox: a trend appears marginally, and disappears or completely reverses when checking by group

Formulizing the UC Berkeley Example

- X – Gender (M/F)
- Y – Admission to Berkeley (yes/no)
- Z – Department (A/B)

- $Pr(Y = yes|X = M) > Pr(Y = yes|X = F)$, but:
- $Pr(Y = yes|X = M, Z = A) < Pr(Y = yes|X = F, Z = A)$
- $Pr(Y = yes|X = M, Z = B) < Pr(Y = yes|X = F, Z = B)$

How can that be?

The probabilities which are not conditioned on Department are still **weighted averages** (Bayes Law):

$$Pr(Y = yes|X = F) = Pr(Y = yes|X = F, Z = A) \cdot \mathbf{Pr(Z = A|X = F)} + Pr(Y = yes|X = F, Z = B) \cdot \mathbf{Pr(Z = B|X = F)}$$

The paradox occurs due to these weights! (women applying to more competitive departments)

Simpson's paradox: the effect of additional conditioning

- In general assume now we have three random variables X, Y, Z
- We can consider two conditional distributions: $Pr(Y|X)$ and $Pr(Y|X,Z)$
- The paradox, which is not really a paradox, says that we can reach "conflicting" conclusions, for example:
  - $Pr(Y=1|X=x_1) > Pr(Y=1|X=x_2)$
  - but: $Pr(Y=1|X=x_1, Z=z) < Pr(Y=1|X=x_2, Z=z), \quad \forall z$
- So in this situation, is x_1 or x_2 a better support for $Y=1$?
- Back to basic formulas, the key is the weighting:

$$Pr(Y=y|X=x) = \sum_z Pr(Y=y|X=x, Z=z)\mathbf{Pr(Z=z|X=x)}$$

Another example: Corona mortality

- We have two countries (call them *Italy* and *Germany*)
- In Italy the mortality rate among Corona patients is higher than in Germany
- But in Germany the mortality rate is higher than in Italy both among young patients and among old patients
- How can that be?

Another example: Corona mortality

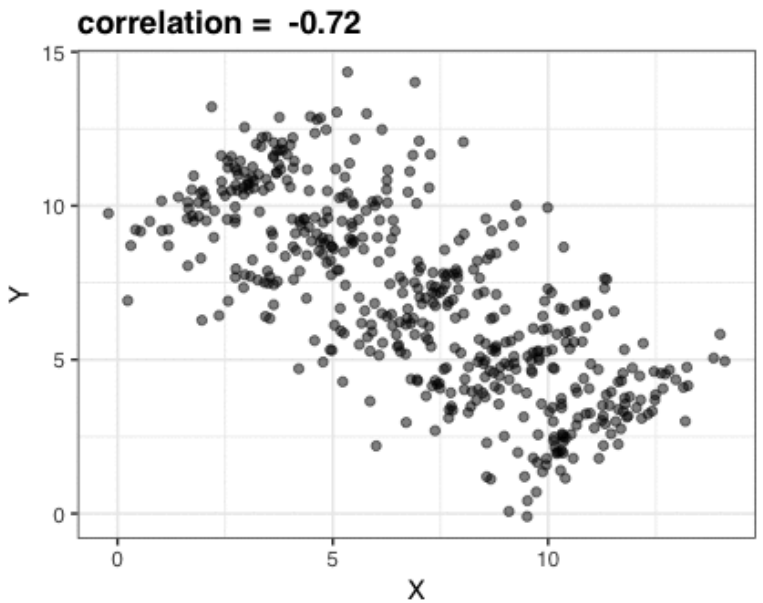
- Let Y = Dead/Alive, X = Country, Z = Young/Old and we are considering only Corona patients:
- $Pr(Y = D|X = I) > Pr(Y = D|X = G)$, but:
- $Pr(Y = D|X = I, Z = Yo) < Pr(Y = D|X = G, Z = Yo)$
- $Pr(Y = D|X = I, Z = O) < Pr(Y = D|X = G, Z = O)$

Solution: Germany has a lot of young patients, Italy a lot of old:

$$Pr(Z = Yo|X = G) \gg Pr(Z = Yo|X = I).$$

Simpson's paradox beyond binary variables

Continuous distributions with clusters (example in recitation)



source

When Y is numeric: we may be interested in conditional expectation, with "paradox":

- $\mathbb{E}(Y|X = x_1) > \mathbb{E}(Y|X = x_2)$, but:
- $\mathbb{E}(Y|X = x_1, Z = z) < \mathbb{E}(Y|X = x_2, Z = z), \quad \forall z$

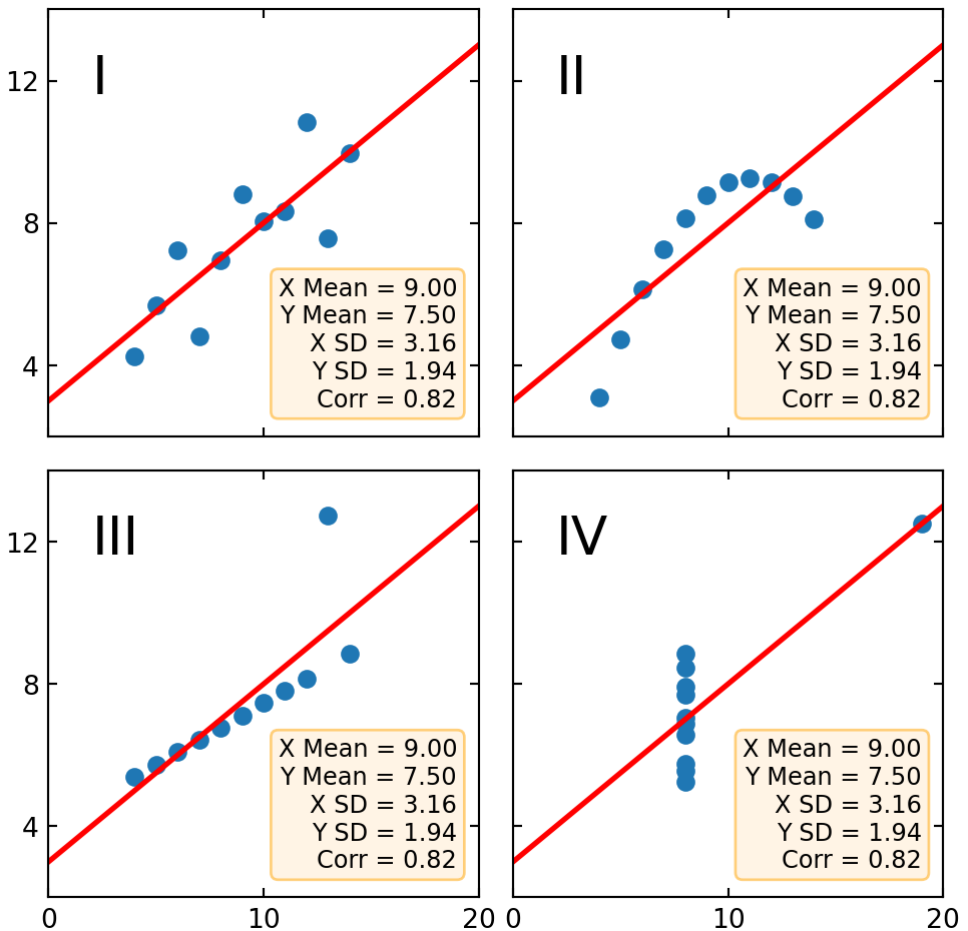
Anscombe's Quartet

INTRODUCTION

To Data Science

Anscombe's Quartet

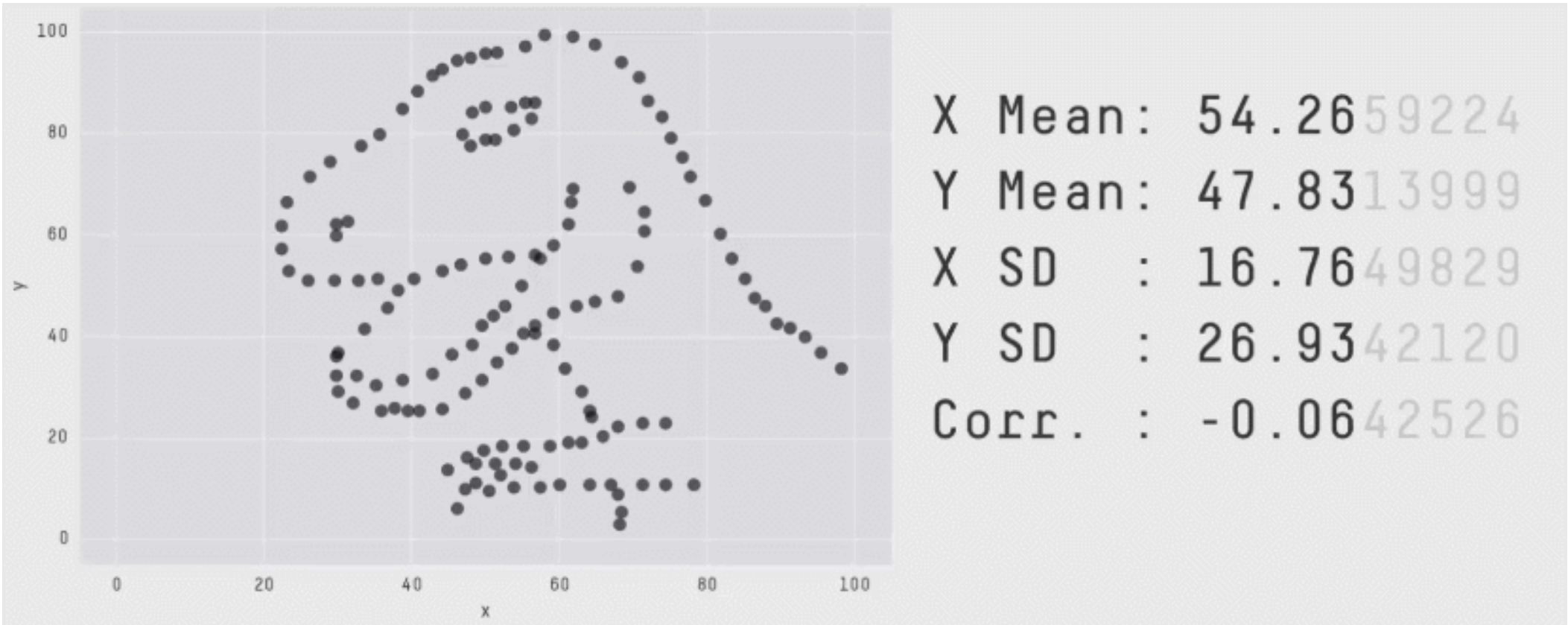
▶ Code



- Beware of looking at just summary statistics!
- Be careful of outliers!
- Plot first!

Datasaurus Dozen

A modern Anscombe's Quartet by Matejka and George (2017):



Pitfalls Summary

INTRODUCTION

TO DATA SCIENCE

Conclusions about probabilistic thinking

- Conditional distributions are very important for interpretation and not very intuitive sometimes
- It is critical to carefully consider which direction and level of conditioning is relevant to reasoning about data
- It is important to be able to write the information, questions and answers explicitly as statements about conditional probabilities or expectations, and use the laws of probability correctly

Dependence and causality

- A well known but often misunderstood fact is that correlation/dependence is not the same as causality
- Example: Assume we study X=smoking and Y=lung disease and find a strong correlation between them: people who smoke more have more lung disease
- Assume for now the connection is real and replicable in multiple studies
- Is it correct to conclude that smoking causes lung disease?

Causal and (example of) non-causal relationships

How and when can we infer causality?

- This is an area of active research, there are many theories and practical methods:
  - Clinical trials that guarantee found relations are causal
  - Instrumental variable methods
  - Causal inference methods by Judea Pearl and others
- We will not discuss these in more detail in this course

Intro to Data Science