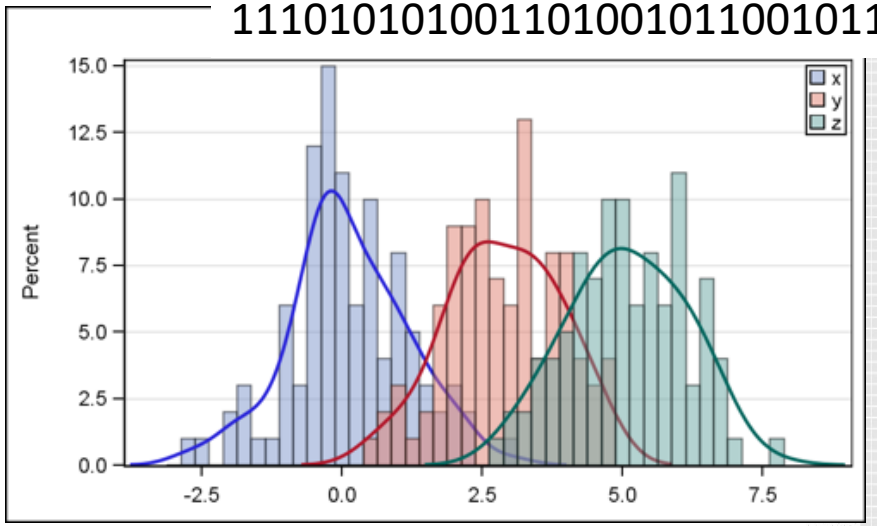
Statistics and data analysis

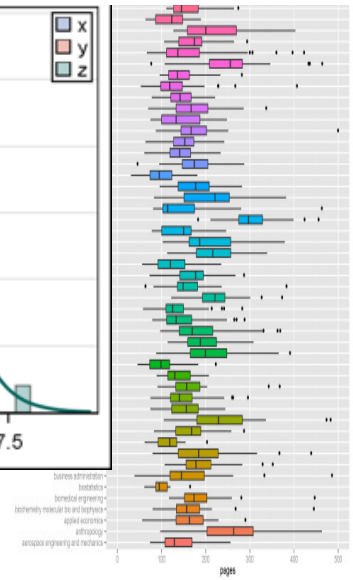
Zohar Yakhini

IDC, Herzeliya

Intro Class

00100111010101001010101001001000101010001010111110101101001100100111101010100110100101100101101100101





Yakh

Data analysis

Analysis of data is the process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making and inference of principles.

Analysis of data is a process of applying mathematical tools to extract useful and faithful information from observed data

Statistics is a scientific domain that investigates and characterizes tools that can be used in the process of analyzing data and in communicating and interpreting analysis results

IDC HERZLIYA

Yakhini TASHPA

Data analysis:
In analyzing observed data from the month of September we found that Covid19 positive testing rate in Haifa was 5.3% and in Jerusalem was 7.1%.

Statistics:
Do these results represent a significant difference?
Could they represent some random effect?

Further data analysis:
Are the differences related to any demographic parameters?
Get data about such potential parameters from more locations and compute correlations.

Statistics:
Are the observed correlations significant?
Can they be the result of some random effect?
Are there confounding factors involved?

IDC HERZLIYA

Yakhini TASHPA

Statistics and data analysis in the age of computers

The story of statistics is changing since:

- More efficient algorithms and computers make deeper and more elaborate calculations possible and practical
- The scope of data is changing in many respects, most notably volume

IDC HERZLIYA

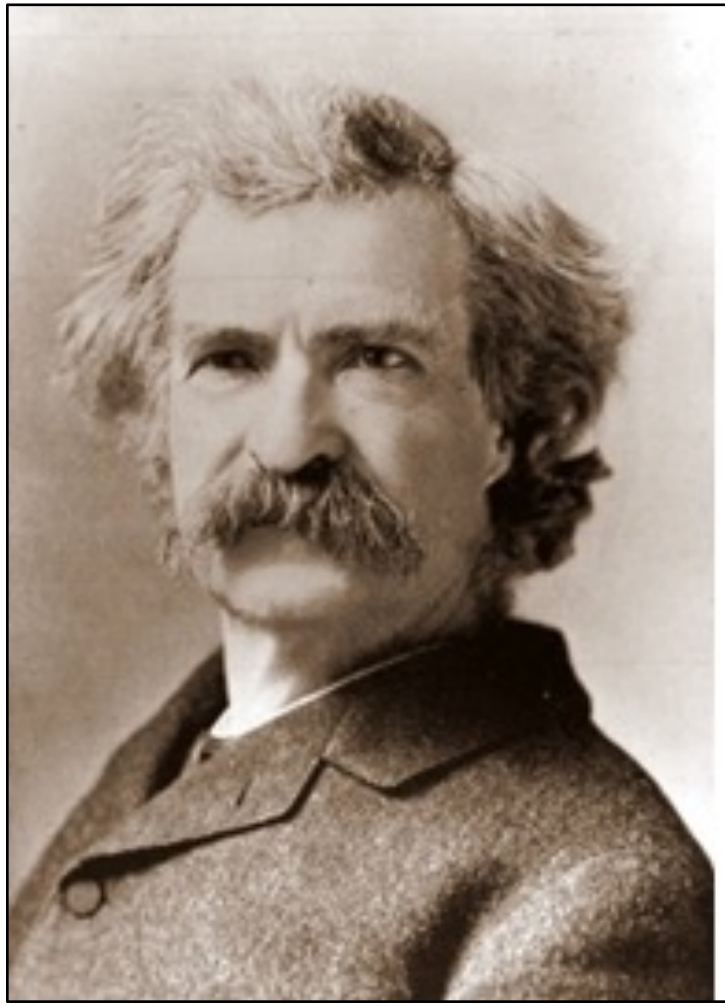
We will see many examples, including intro examples in this class …

Yakhini TASHPA

Statistics - defined

Yakhini TASHPA

Statistics



There are three kinds of lies: lies, damned lies, and statistics.

–Mark Twain

Chapters from My Autobiography

The science of effectively drawing conclusions from data

OR

The science of effectively and convincingly lying

IDC HERZLIYA

Yakhini TASHPA

Statistics is a common bond supporting all other sciences as well as all quantitative social and business investigations.
It provides standards of empirical proof and a language for communicating results in these domains.

The process of statistical investigation includes:

- Designing experiments to maximize information
- Using models to describe observations and assess their significance
- Efficiently and effectively answering questions of interest
- Verifying the validity of the process
- Snooping around for more to be learned …

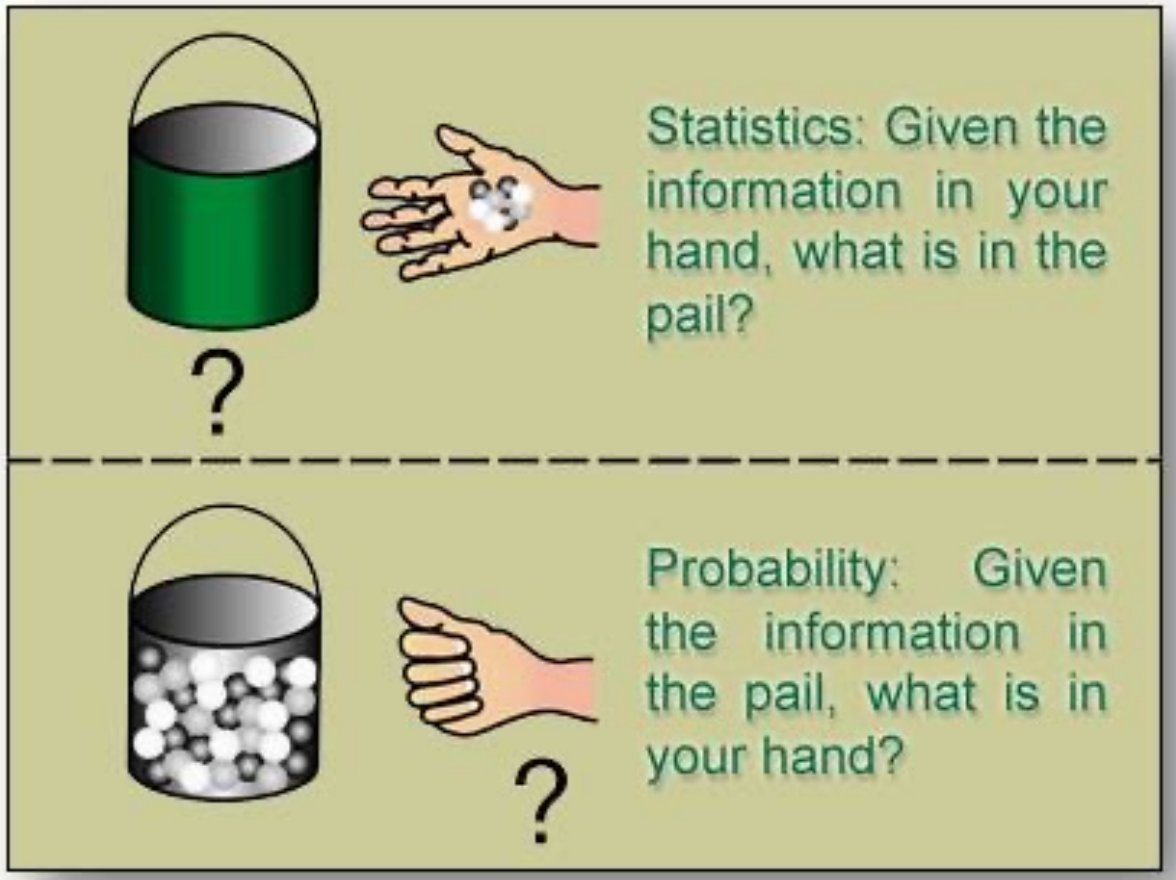
IDC HERZLIYA

Adapted from F Ramsey and D Schafer, Oregon State Univ

Probability theory and statistics

Statistics – given observations, what can we say about the underlying mechanism/system that gave rise to these observations?

Probability – assuming a model - what is the expected behavior of observations from the model?



In the age of computers, there are two separate aspects of the statistical enterprise:

1. Algorithmic developments aimed to draw conclusions from data.
   For example:
   Efficiently computing correlations for millions of quantitative records associated with users of a system or with a population of members of a healthcare service provider
   OR
   Using decision trees or random forests to make predictions about quantities of interest

2. Inference and assessment methodology.
   Aimed to test the validity of the results of the algorithm and to provide arguments to support the conclusions.

The science of effectively drawing conclusions from data
OR
The science of effectively and convincingly lying

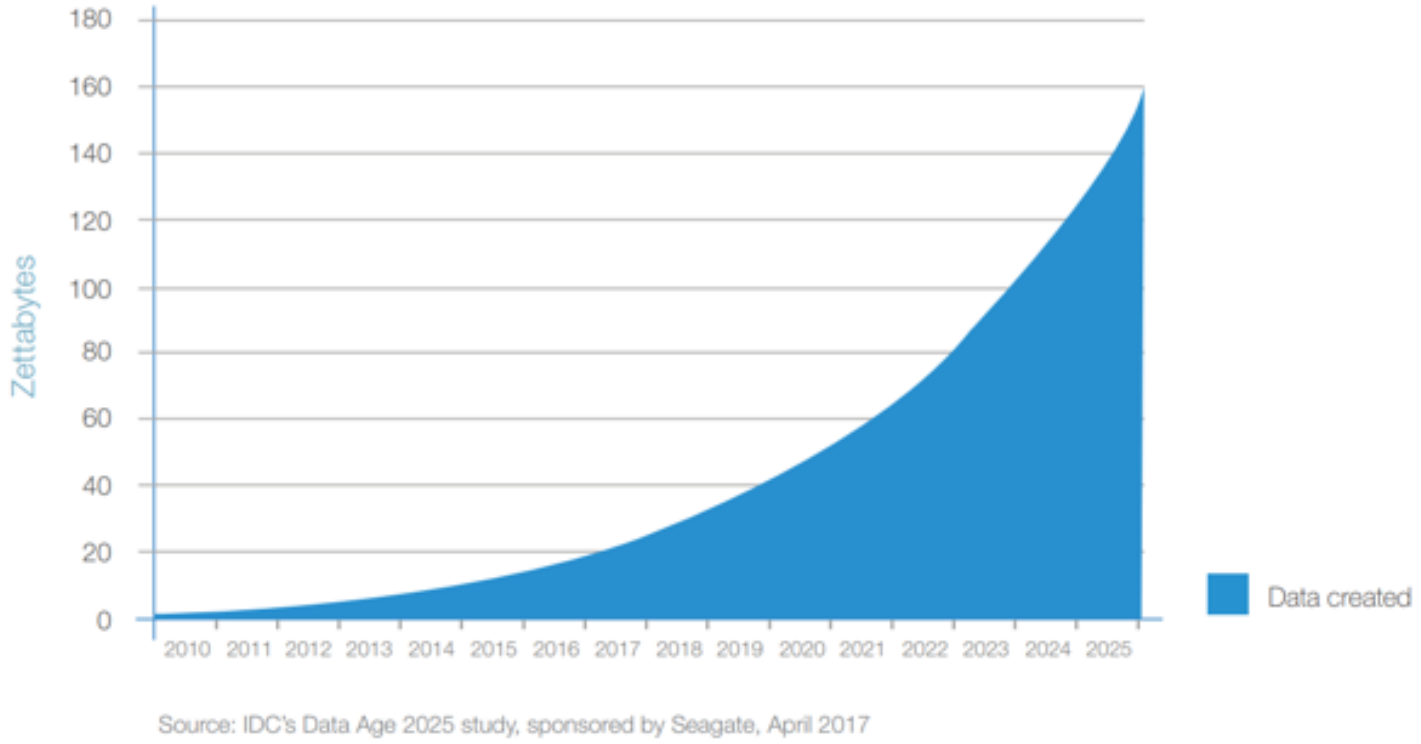
IDC HERZLIYA

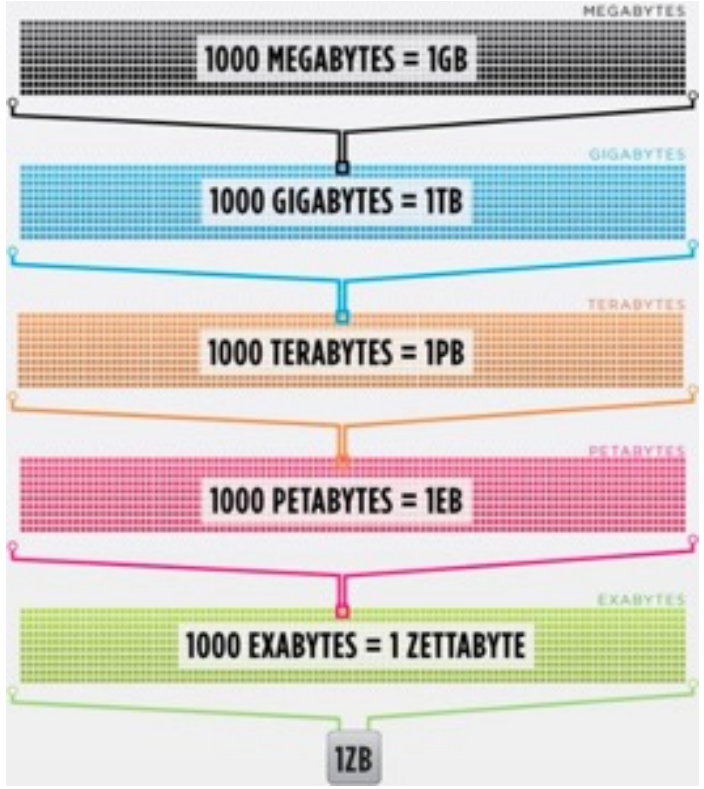
Adapted from Efron and Hastie, Stanford Univ

Yakhini TASHPA

DATA

- Data is eating the world: 163 Zettabytes will be created in 2025





BIG DATA

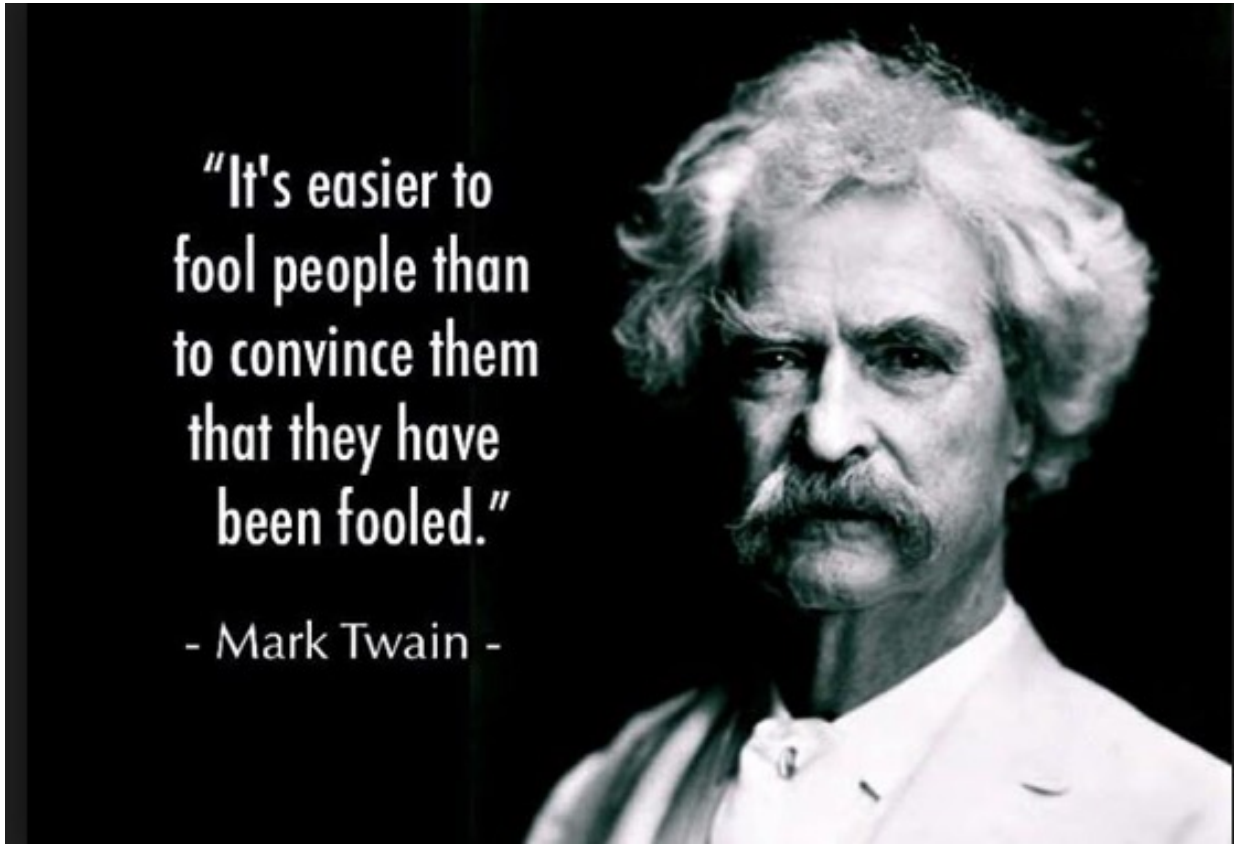
Yes - its a buzz word …
But still – the utilization of data collected in many disciplines, from physics and molecular biology to environmental and social science, as well as by business oriented organizations, involves many steps and each one of them needs to be done using **efficient** and **effective** methods to get to desirable results:

- Data collection and acquisition; Experiment design
  - + Garbage in garbage out
  - + Confounding factors, Representation
- Data storage, processing
  - + Accessibility and interaction
  - + Data cleaning
- **Statistical data analysis and efficient accurate inference**
- Conclusions
  - + **Visualization**
  - + **Reporting**
- Feedback
  - + Into data collection and other steps above

IDC HERZLIYA

Yakhini TASHPA

- Can we make statistical arguments simple and convincing?
- Visualization and presentation
- Clear and simple statements
- Rigorous and accurate methodologies
- The data scientist should know, to the greatest possible extent, what is standing behind any stated conclusion.

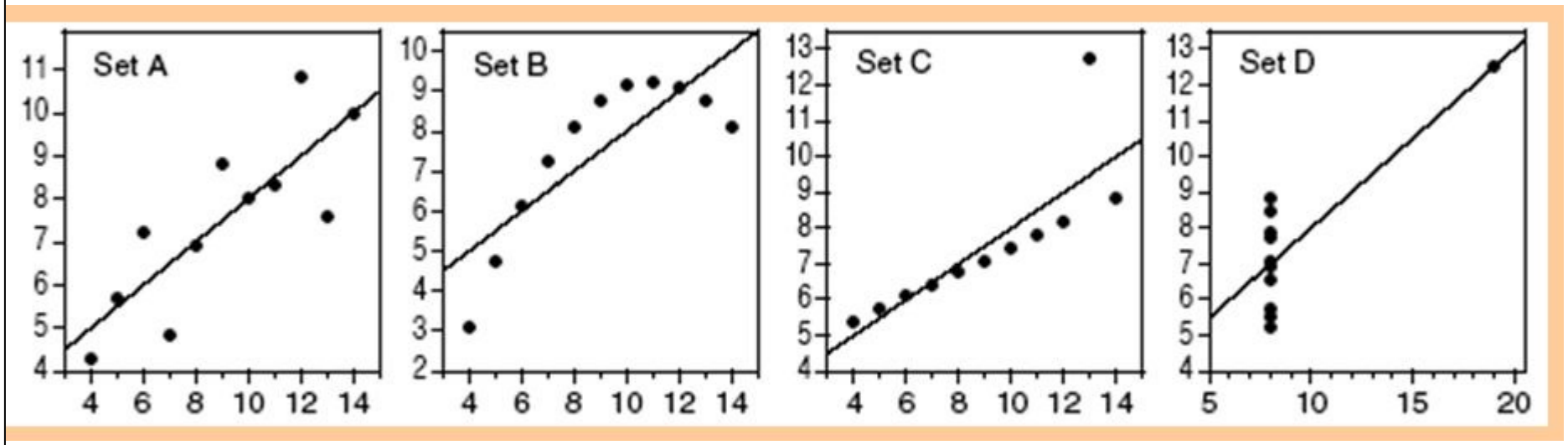


IDC HERZLIYA

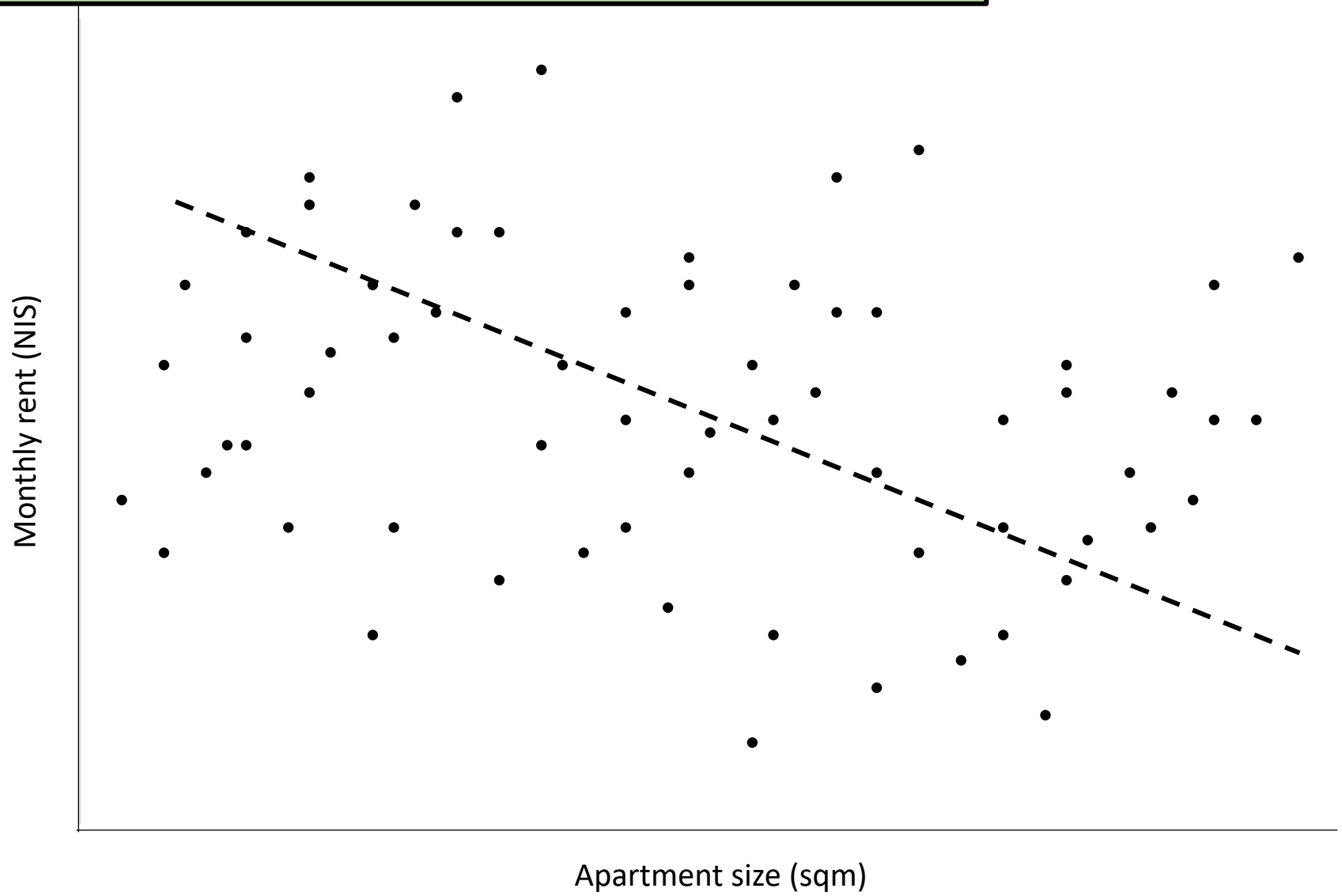
Yakhini TASHPA

Effective inference – Example 1

Applying the same formal statistical tool on all of these yields the same result – high correlation between the variables.
But clearly – the actual conclusion should be different for each one of them



Example 2 –correlations??



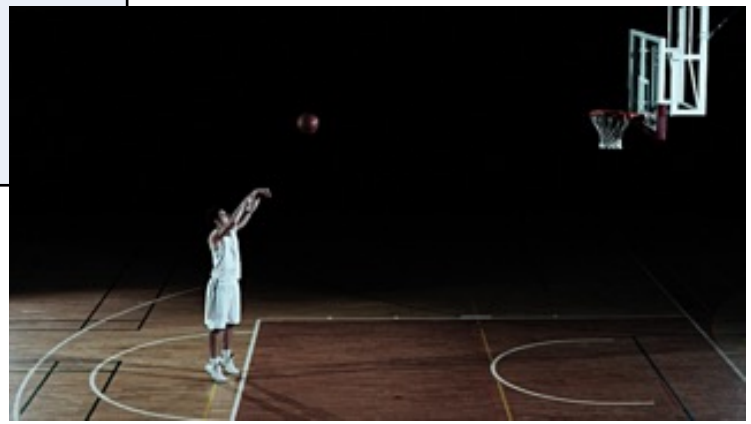
IDC HERZLIYA

Yakhini TASHPA

Basketball

The Randomistan basketball association conducted a test in its top basketball leagues.
They recorded the number of basketball players who managed to score 5/5 free shots from the line.
The data is segmented by height and by gender.

	Less than 1.70m	1.70-1.90	Taller than 1.90
Women	4/6	4/6	8/9
Men	1/2	1/2	23/27

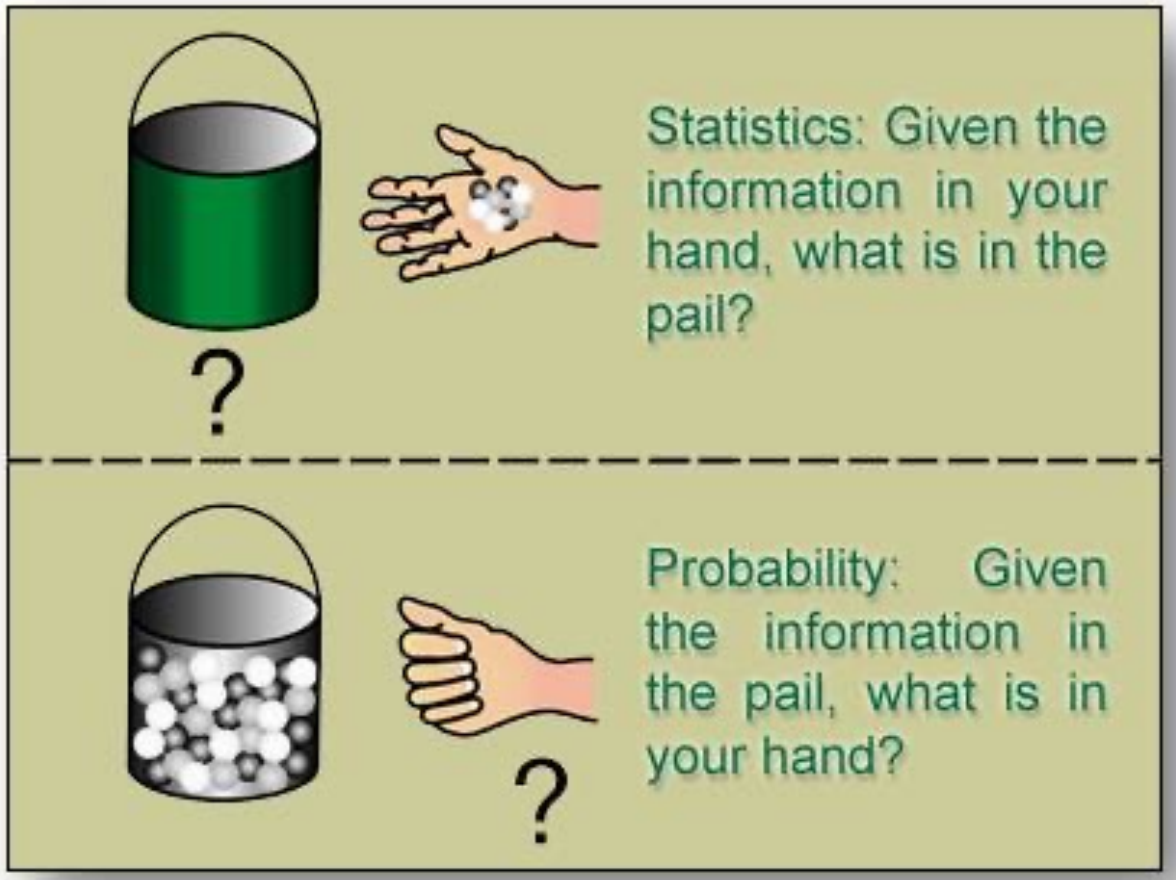


Who scores better from the line in Randomistan – men or women?

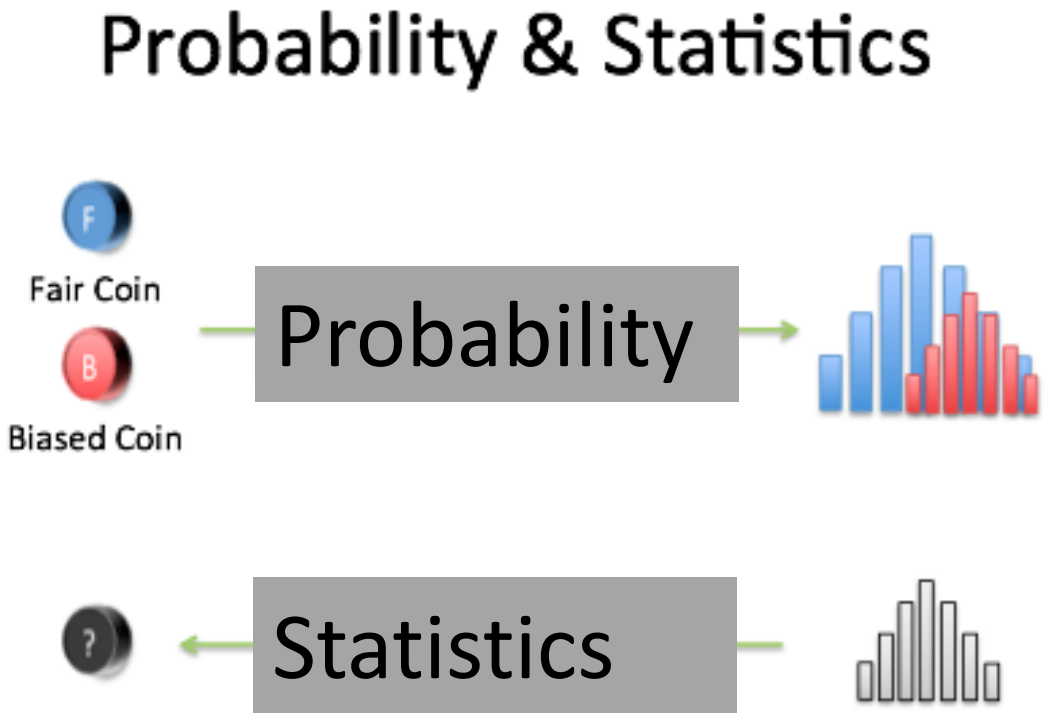
Probability theory and statistics

Statistics – given observations, what can we say about the underlying mechanism/system that gave rise to these observations?

Probability – assuming a model - what is the expected behavior of observations from the model?



Inference



Probability
Given a model – determine the probability of occurrence of various data related events (including functions)

Statistics
Given observed data – Infer a model mechanism that could generate it; Quantify the inference

IDC HERZLIYA

Course structure and formalities

- Time – Thu 1830-21hrs (H) OR Fri 845-1115 (E), with a 15 mins break
- 13 weeks
- 6 recitations: recitation will be once every two weeks (on average) – exact dates and times in Moodle (**first** recitation on Fri 11/11 and Wed 16/11)
- Prof Zohar Yakhini, office hours: C123 CS Bdg, Please email to coordinate.
  Ben Galili, office hours: C314 CS Bdg, Please email to coordinate.
- zohar.yakhini@gmail.com , ben.galili@post.runi.ac.il
- Shuli Finely, Saar Buchnik.
  Course Assistants will publish their office hours
- There will be 4 HW assignments that include theoretical aspects as well as practical data analysis mini-projects
- HW schedule will evolve as we go.
- HWAs will be due 2-3 weeks after assigned; Must be in pairs.
- Each HWA will be 18pts from the 100 grade points.
- The exam will be 28pts (MUST pass to pass class)

Topics to be covered

1. Introduction and review of probability theory
2. Important distributions
3. Statistical independence and what it means; Marginals and copulas
4. Data presentation and visualization
5. The binomial distribution and CLT
6. Foundations of statistical inference: confidence intervals, p-values, hypothesis testing
7. Parameter estimation and the EM algorithm
8. Correlation measures and how to use and misuse them
9. The hypergeometric distribution, inc examples and approximation schemes
10. Ranked lists, Wilcoxon rank sum, mHG
11. Multiple testing, Bonferroni and FDR corrections

12. Survival analysis, KM curves and the Mantel-Haenzel test
13. Entropy and information, KL and distances between distributions