Lecture Notes EMSE 4765:
DATA ANALYSIS - Probability Review

Chapter 1: Why Probability and Statistics?

Version: 01/08/2021

Text Book: A Modern Introduction to Probability and Statistics, Understanding Why and How

By: F.M. Dekking. C. Kraaikamp, H.P.Lopuhaä and L.E. Meester

1 Why Probability and Statistics?

1.1 Example — Biometry: Iris Recognition…

- **Coin -Toss Experiment 1: Suppose we take a coin, toss it and record the outcome.** A person will be identified with the outcome of this one coin toss.

- **Question:** Let's pick two people. What is the probability they are the "same"?

- **Coin -Toss Experiment 2: Suppose we take a coin, toss it twice and record the outcomes.** A person will be identified with the outcomes of these two coin tosses.

- **Question:** Let's pick two people. What is the probability they are the "same"?

- **Coin -Toss Experiment 3: Suppose we take a coin, toss it thrice and record the outcomes.** A person will be identified with the outcomes of these three coin tosses.

- **Question:** Let's pick two people. What is the probability they are the "same"?

1 Why Probability and Statistics?

1.1 Example — Biometry: Iris Recognition...

Hmm, why not use a large sequence of coin tosses as identification for individuals?

- Suppose we do that. **How can we measure how different two people are?**
- ***n*-distance: The fraction or relative frequency of coin tosses that are different** in a series of n coin tosses between two individuals.

Coin Toss	John	Mary	Different?
1	0	0	0
2	0	0	0
3	1	1	0
4	1	1	0
5	0	1	1
6	1	0	1
7	0	1	1
8	1	1	0
9	1	1	0
10	0	1	1
		Sum	4

$$10\text{-}Distance(John, Mary) = \frac{4}{10},$$
$$Notation: H_{10}(J, M) = \frac{4}{10}$$

Hamming distance between two strings of equal length is **the number of positions** at which the corresponding symbols are different.

1 Why Probability and Statistics?

1.1 Example — Biometry: Iris Recognition...

$$\begin{cases} H_{10}(J,M) = 0 & \Rightarrow \text{John and Mary are the same person} \\ H_{10}(J,M) > 0 & \Rightarrow \text{John and Mary are not the same person} \end{cases}$$

- Denote $J_i \in \{0,1\}$: Outcome of John's i-th coin toss experiment.
- Denote $M_i \in \{0,1\}$: Outcome of Marie's i-th coin toss experiment.

We have: $\boldsymbol{Pr(J_i = M_i) = \frac{1}{2}, i = 1, \ldots, 10}$

- **Question:** "What is the probability that John and Mary are the same person?" and "What is the probability that John and Mary are not the same person?"
- **Said differently:** What is $Pr\{H_{10}(J,M) = 0\}$ and $Pr\{H_{10}(J,M) > 0\}$?

$$\boldsymbol{Pr\{H_{10}(J,M) = 0\} = (\tfrac{1}{2})^{10}, Pr\{H_{10}(J,M) > 0\} = 1 - (\tfrac{1}{2})^{10}}$$

1 Why Probability and Statistics?

1.1 Example — Biometry: Iris Recognition...

- **Suppose we do not need that level of accuracy and we define:**

$$\begin{cases} H_{10}(J,M) \leq \tau & \Rightarrow \text{John and Mary are the same person} \\ H_{10}(J,M) > \tau & \Rightarrow \text{John and Mary are not the same person} \end{cases}$$

- **Question:** What are the different possible values for $H_{10}(J,M)$?

$$0, 0.1, 0.2, \ldots, 0.7, 0.8, 0.9, 1.$$

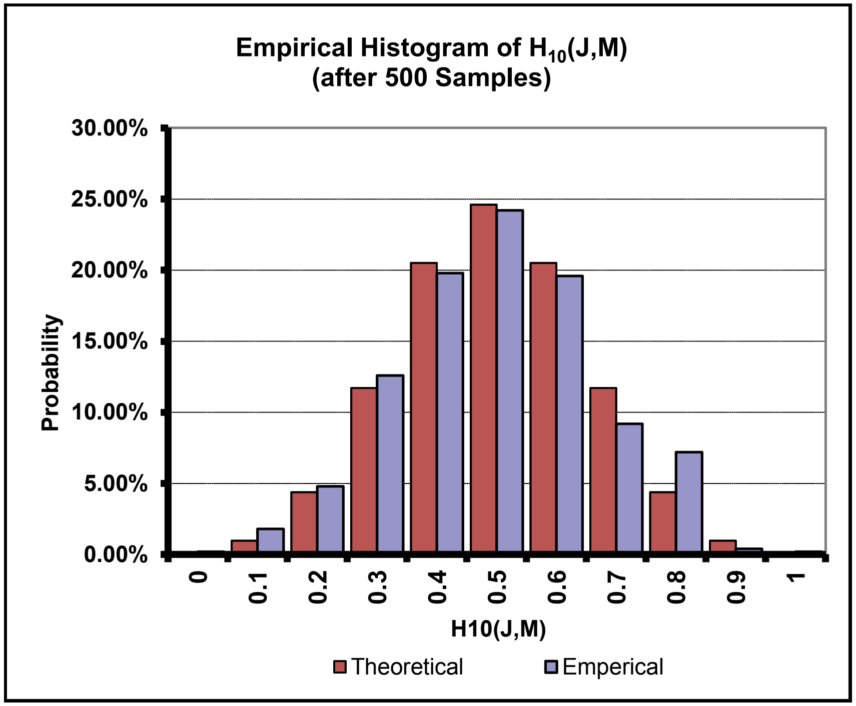
- **Suppose we set $\tau = 0.3$:** "What is the probability that John and Mary are the same person?" **Said differently:** What is $Pr\{H_{10}(J,M) \leq 0.3\}$?

- **Introducing $X = 10 \times H_{10}(J,M) \Rightarrow X \sim Bin(10, \frac{1}{2})$**

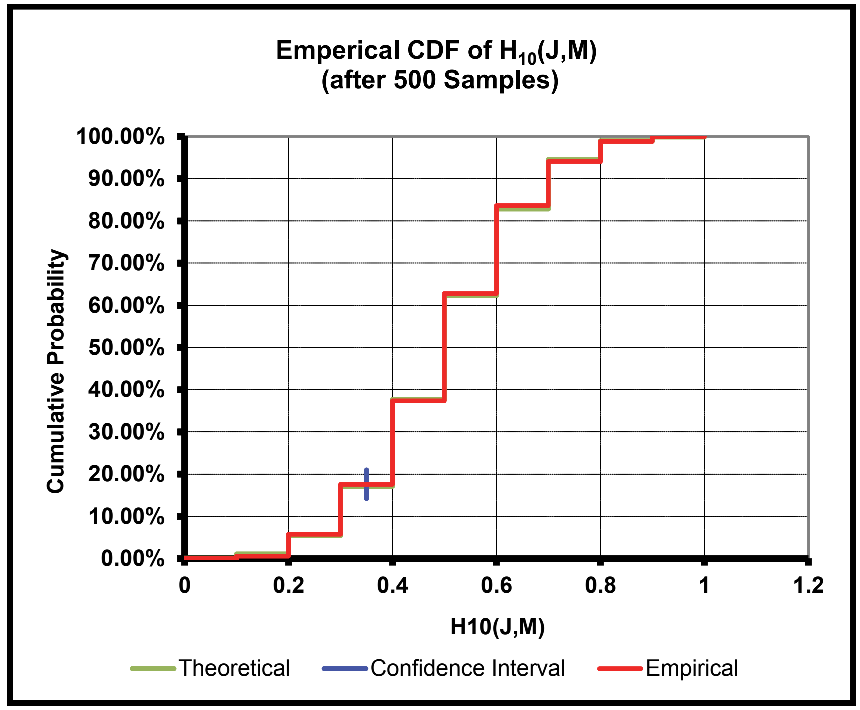
$$\sum_{k=0}^{3} Pr\{H_{10}(J,M) = \frac{k}{10}\} = Pr(X \leq 3) = \sum_{k=0}^{3} \binom{10}{k} \left(\frac{1}{2}\right)^{k} \left(\frac{1}{2}\right)^{10-k}$$

1 Why Probability and Statistics?

1.1 Example — Biometry: Iris Recognition. . .



Probability Mass Function



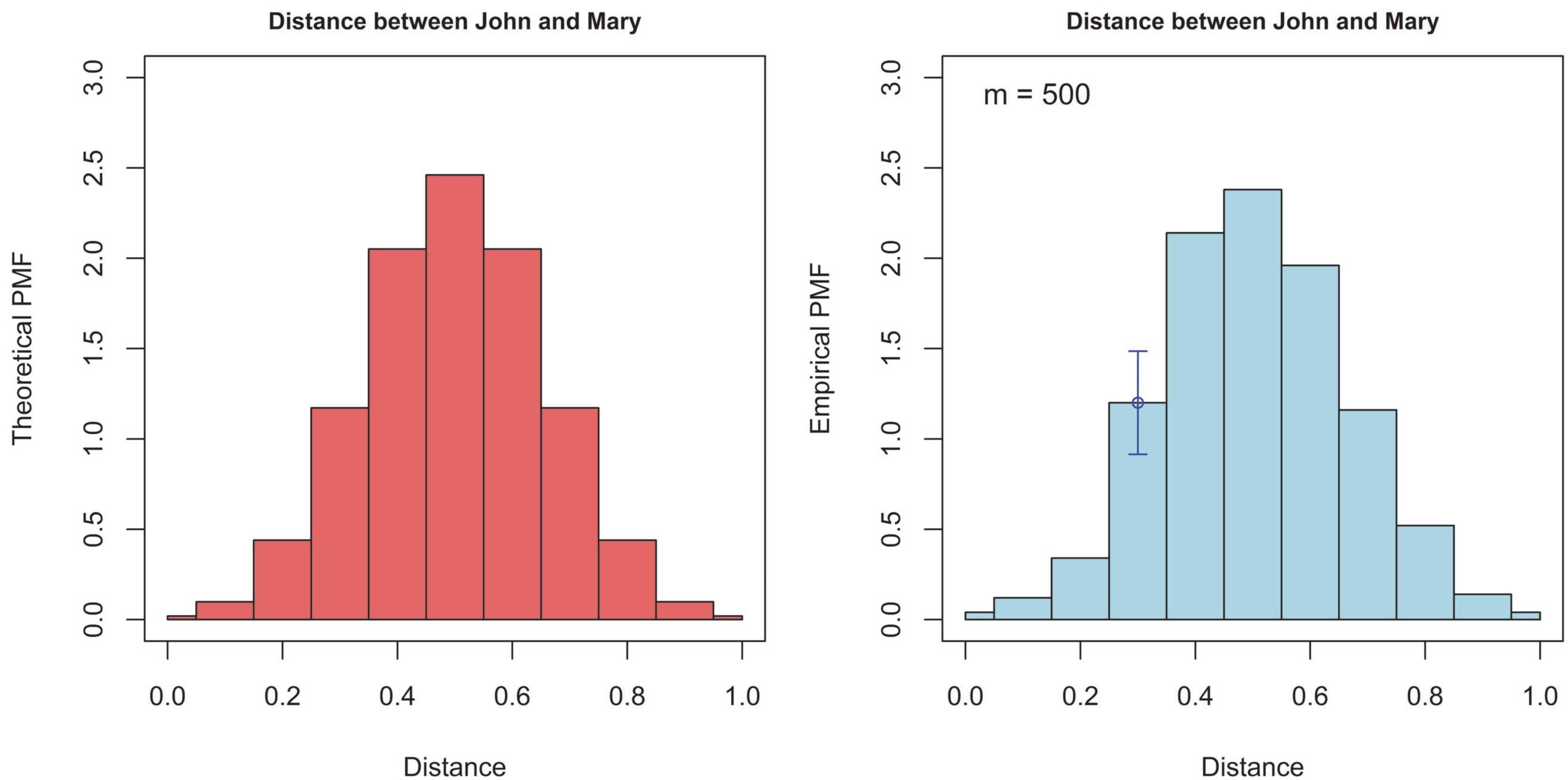
Cumulative Distribution Function

Statistics is about estimating distributions from data

1 Why Probability and Statistics?

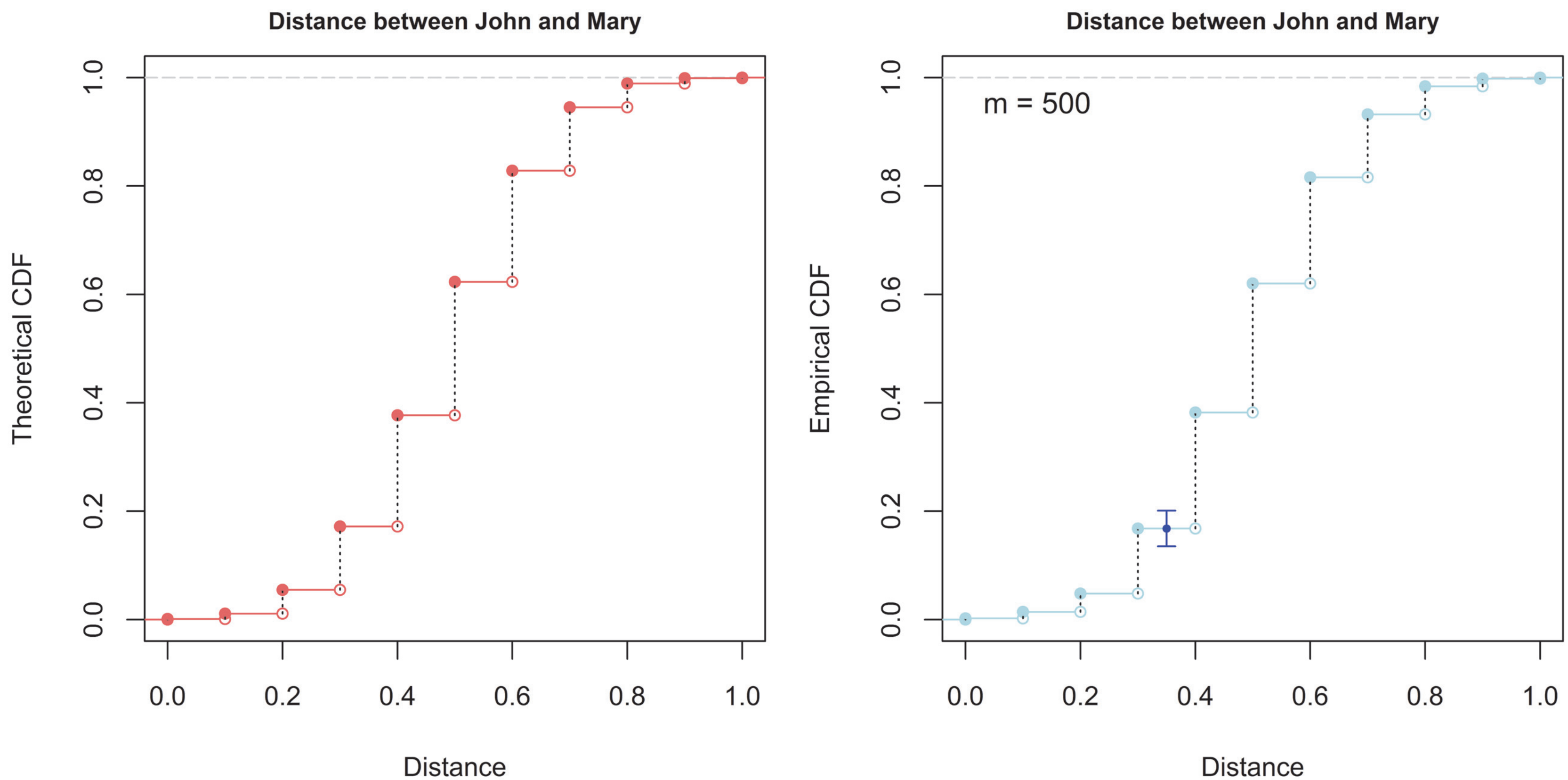
1.1 Example — Biometry: Iris Recognition. . .

Same Analysis in R in file "John_and_Mary.R"



Statistics is about estimating distributions from data

Same Analysis in R in file "John_and_Mary.R"

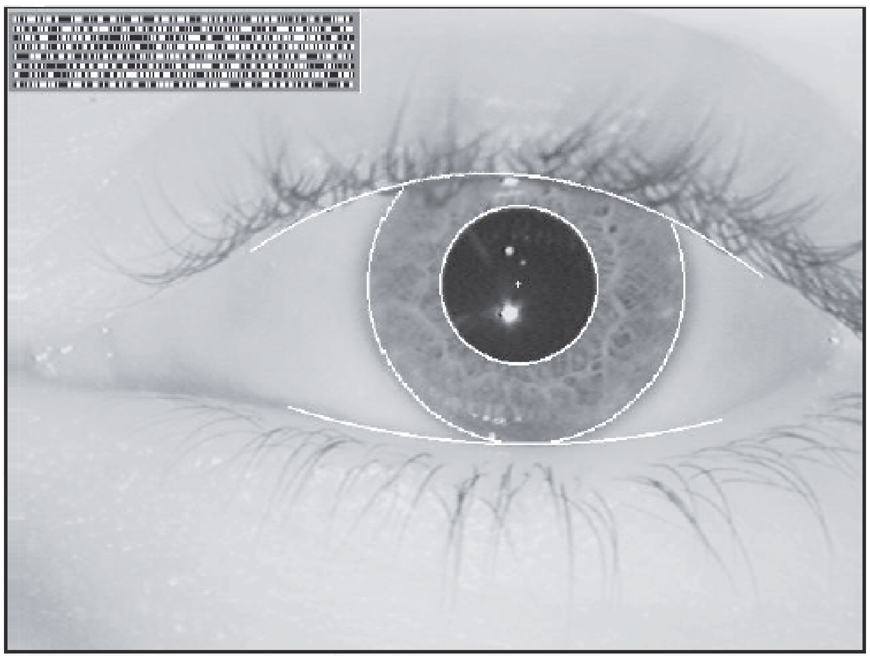


Statistics is about estimating distributions from data

1 Why Probability and Statistics?

1.1 Example — Biometry: Iris Recognition. . .

- **Iris recognition technology:** Based on **the visible qualities of the iris**. Converting these—via a video camera—into an **"iris code"**results into just 2048 bits. That is, a sequence of "2048 zeros and ones" or "a sequence of 2048 coin tosses" defined by your eye



Thus every individual is born with 2048 outcomes of coin tosses that are unique to him/her!

Source: *How Iris Recognition Works* by John Daugman, PhD, OBE University of Cambridge, The Computer Laboratory, Cambridge CB2 3QG, U.K. www.CL.cam.ac.uk/users/jgd1000/

1 Why Probability and Statistics?

1.1 Example — Biometry: Iris Recognition...

- **J. Daugman[1] concluded that of the 2048 "coin toss outcomes", 266** may **be considered as uncorrelated** observing "heads" **about 50% of the time.** Thus, **the iris code** may be seen as the outcome of **266 coin tosses with a "fair" coin.**

Hence, at birth a sequence of 266 coin tosses is encoded in our irises and this iris code can be used for identification purposes.

- **Question:** How many "different" persons are possible using this iris code?

Answer: $2^{266} \approx 1.18 \times 10^{80}$

Current World Population

7,593,430,009

http://www.worldometers.info/world-population/

[1] J. Daugman. Wavelet demodulation codes, statistical independence, and pattern recognition. In *Institute of Mathematics and its Applications, Proc. 2nd IMA-IP: Mathematical Methods, Algorithms, and Applications (Blackledge and Turner, Eds)*, pages 244–260. Horwood, London, 2000.

1 Why Probability and Statistics?

1.1 Example — Biometry: Iris Recognition...

- **The Definition "Two iris codes X and Y are identical"** $=$ "$H_{266}(X,Y) = 0$" is **not practical** since the measurement is not without error. **The Definition "Two iris codes X and Y are identical"** $=$ "Distance $< \tau$" is **practical**.

- **To reduce error**, Daugman takes 7 Iris pictures and evaluates min distance for different eyes, and max distance for same eyes.

- Next, Daugman uses an **empirical analysis** to set **$\tau = 0.342$ yielding a probabily of false miss identification:**

$$Pr(\text{"}J \neq M\text{"} | J = M) \approx 1e-6$$

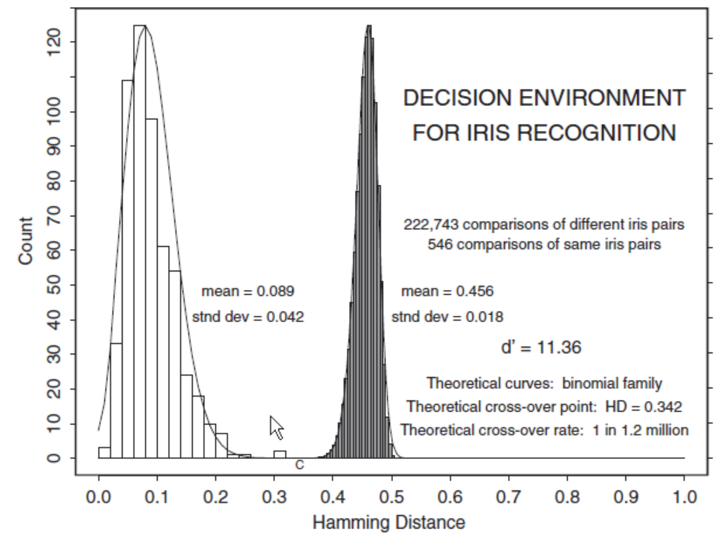


Fig. 1.1. Comparison of same and different iris pairs.

Source: J.Daugman. *Second IMA Conference on Image Processing: Mathematical Methods, Algorithms and Applications*, 2000. © Ellis Horwood Publishing Limited.

Analysis reconstruction in R in file "Iris_Analysis.R"