Chapter 1. Introduction to Statistics and Data Analysis

Biostatistics for Engineers

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

Hansol Choi

August 24, 2023

Outline

Orientation

Course Information

- Instructor : Hansol Choi(apine@unist.ac.kr), 701-13, Bd 110.
- Textbook : Statistics for engineers and scientist, Walpole, R.E et al., **9TH ED. GLOBAL EDITION,** Pearson 2016,
- Office Hour: 13-16 Mon,Wed
- Lecture Hour: 10:30-11:45 Mon,Wed

Course Objectives

- To understand the fundamental concept of statistics.
- To learn statistical methods for Sciences and engineering applications.
- To gain insights on how to deal with multivariate data.
- To obtain basic concepts of statistical learning methods.

Grading

- Attendance 10%, Midterm 25%, Final 35%, Quizzes 20%, Assignments 10%.
- Attendance check is updated on the 'Smart Check attendance system' manually and regularly.
- If there is an error in the attendance check, the student is responsible for the proof; the evidence of the attendance must be submitted with request.
- The bonus points for any contribution to class will be added to quiz or assignments scores.

Assignment

- The assignments will be given for every subjectives.
- It will be posted on the course website.
- The due will be announced on the course website.
- Each assignment will be provided as as a Jupyter notebook format
- Python programming and pen-and-paper problems will be given.

Remarks

- Prerequisite: Basic Calculus (knowledge and technique about differentiation and integration of exponential and logarithm functions)
- Computer program: Python

Enjoy! ☺

Chaper 1. Introduction to Statistics and Data Analysis

Why do we need this all?

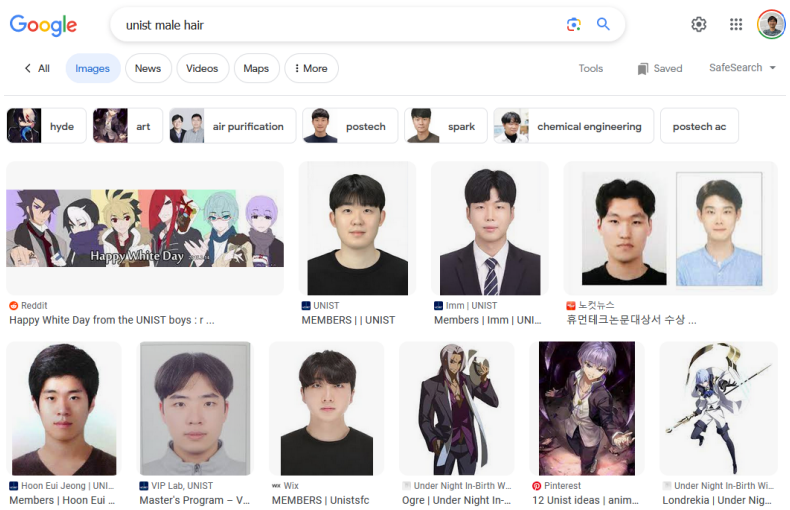
- Decision making in a scientific way.
  - Is a new drug A effective on cancer?
  - Is the new vaccine effective?
  - Is our new predictive model for weather forecast better than the existing one?
- We need to make a decision based on the data.
  - What is your question to decide answer?

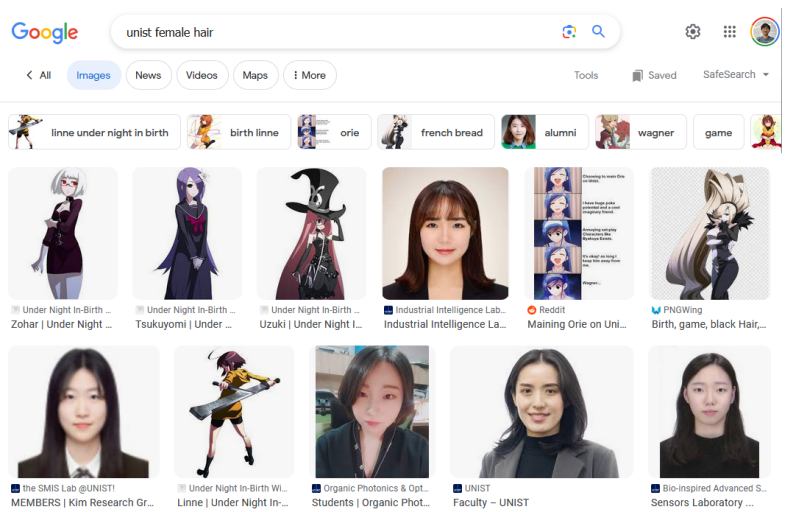
My hypothesis

- There are more male ghosts in Korea compared to male.
  - I really want to know.
  - Is it a proper scientific hypothesis?
  - Why not?
- You cannot gether data.
- You might be able to, but not in a scientific way.

My hypothesis 2

- Female human in UNIST have longer hairs on average, compared to male.
- How can I validate the hypothesis?

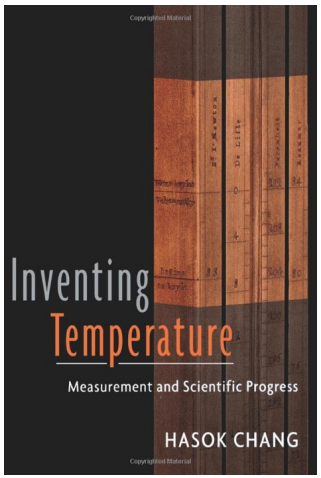




- Problem solved??
- Scientific way of data collection
- Scientific way of Measurements

Measurement

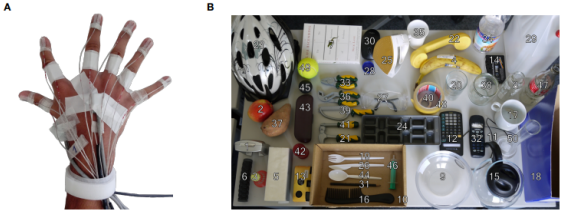
- Measurement is the quantification of attributes of an object or event, which can be used to compare with other objects or events. (Wikipedia)
- a process of determining how large or small a physical quantity is as compared to a basic reference quantity of the same kind(Wikipedia)



Scientific Measurement

- Describe how you did.
- Reproducibility of the measurement.
- Can be Experiment/Observation.

Object manipulation and common movement tasks. Experimental setup: The subjects were seated in front of a desk during the two tasks described below. An electromagnetic motion capture system (Polhemus Liberty 240/16-16) was used to record the hand and finger movements during the object manipulation and the common movement tasks (see Supplementary Fig. 5A). The hands were kept at 0.6 m distance from the main Polhemus system to maintain the recording noise below 0.005 mm. In total, 12 respectively 14 sensors were attached to the hand and fingers of five- or six-fingered subjects using medical tape. Every sensor measured three Cartesian coordinates for the position and three angles for the orientation relative to the main station. Each sensor was connected to the Polhemus system by plastic insulated aluminum wires. Two large sensors ($9 \times 11 \times 6\ mm^3$ at maximum positions, 9.1 g) were placed on the skin on top of the middle and thumb metacarpal bones. The others were small sensors (spherical, 17.3 mm length, 1.8 mm outer diameter, <1 g) which were placed at the distal and proximal phalanges of each finger. Measurements were recorded at 120 Hz.



Supplementary Figure 5: (A) Recording of finger movements with accurate electromagnetic motion capture system (Polhemus Liberty 240/16-16) during the object manipulation and the common movement tasks. (B) Objects used in the object manipulation task (in the order as they were given to the subjects in the experiment): 1 alarm clock, 2 apple, 3 badminton shuttle, 4 banana, 5 blackboard eraser, 6 bicycle handle, 7 book, 8 glass bottle 0.25l, 9 bowl, 10 cardboard box, 11 cable, 12 calculator, 13 disposable camera, 14 audio cassette, 15 coffee jar, 16 comb, 17 cup, 18 folder, 19 fork, 20 glass, 21 hammer, 22 handset, 23 helmet, 24 Ice cube mold, 25 iron, 26 knife, 27 safety glasses, 28 bottle cap, 29 milk bottle 5l, 30 mouse, 31 pencil, 32 phone, 33 pincers, 34 plastic bottle 1l, 35 plastic cup, 36 wrench, 37 sweet potato, 38 jam jar, 39 saw, 40 medical tape, 41 screw driver, 42 pencil sharpener, 43 spectacle case, 44 spoon, 45 squash ball, 46 syringe, 47 coca cola bottle, 48 double faced adhesive tape, 49 tennis ball, 50 vase.

Variability in Data

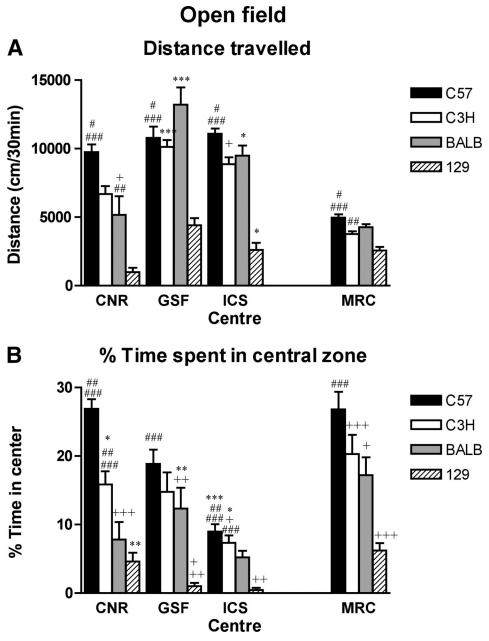


Figure: Mandillo et al. 2008 Physiological Genomics

- Diversity in the data.

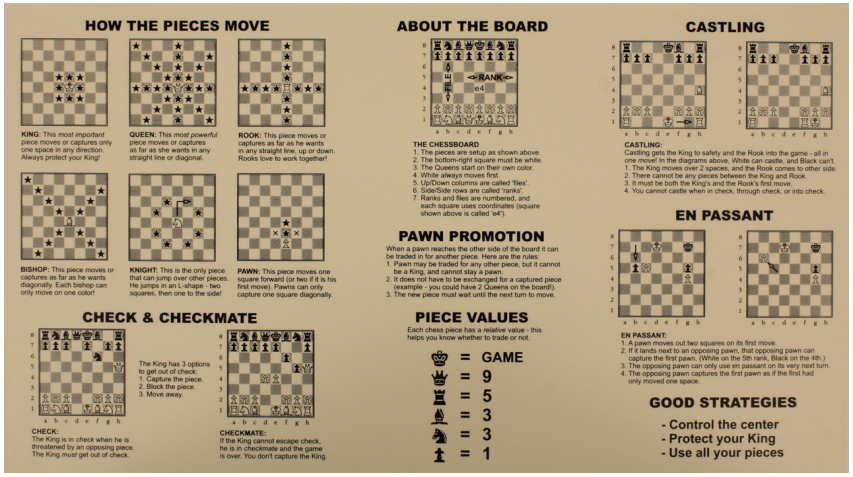
Descriptive statistics

- Simple description of the data and summary statistics.
  - eg) The size of T-shirt in this class room.
  - It will be useful to order the T-shirt
  - 33% of statistics class members are from Dept.Biology
  - It will be useful to decide the level of the class

Inferential statistics

- We need to decide something over population or factors based on the data.
  - We need to ”formalize” the decision process
  - Mathematical models.

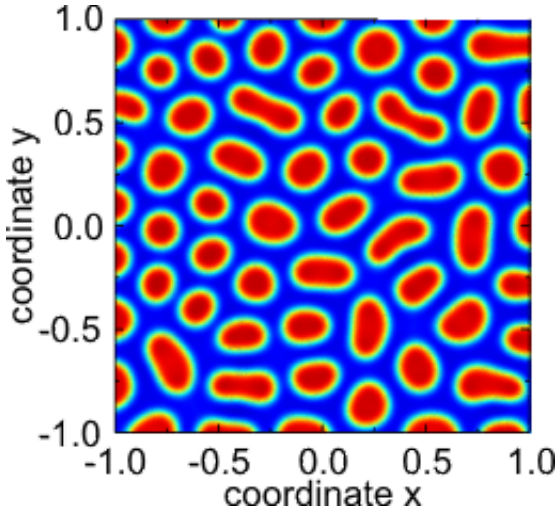
Model System: Formalization

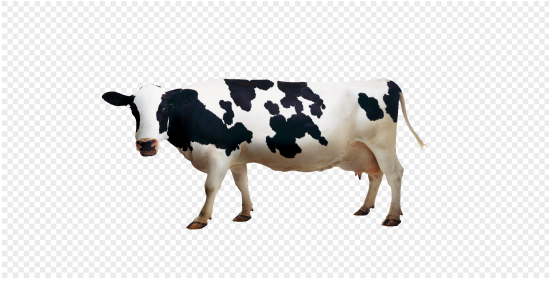


- A chess is composed of the following elements:
  - items: the board and pieces
  - a state: it describees the states of items
  - The Initial state
  - The rule: it decides what is possible and not.
- You can communicate the state of chess **without any ambiguity using the state.**
- **If the state of a chess is related with the state of a real observation, then we can use the chess as its model**

Model System: Turing's reaction diffusion model

$$\begin{pmatrix} \partial_t u \\ \partial_t v \end{pmatrix} = \begin{pmatrix} D_u & 0 \\ 0 & D_v \end{pmatrix} \begin{pmatrix} \partial_{xx} u \\ \partial_{xx} v \end{pmatrix} + \begin{pmatrix} F(u,v) \\ G(u,v) \end{pmatrix}$$

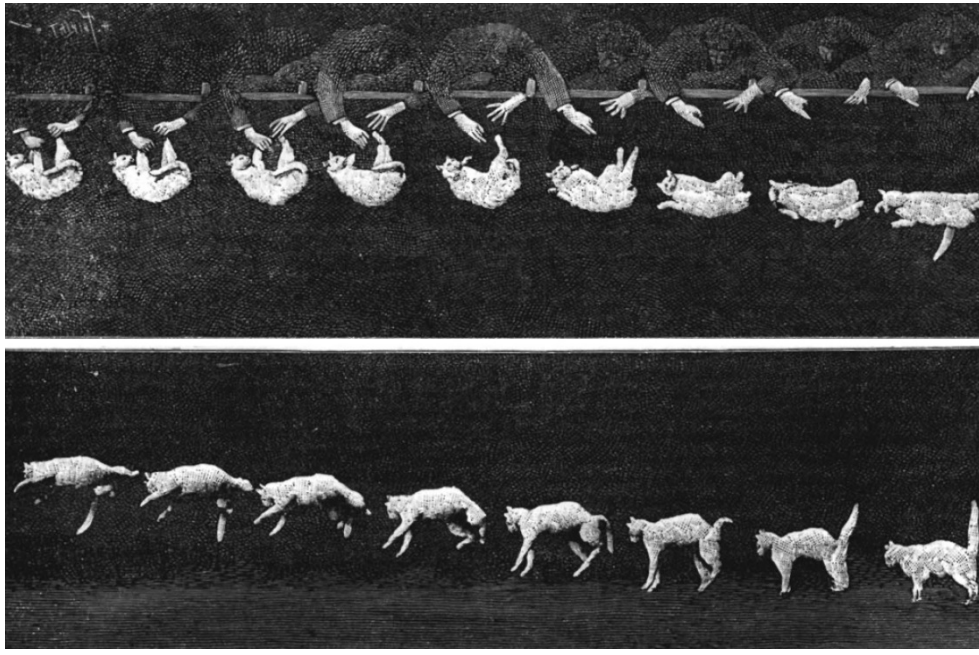




- Turing, 1952, The chemical basis of morphogenesis
- Highly Formalized Model for a Phenomenon in the World.

Mathematical model of the gravitation

- Mathematical model for the law of gravitation is $F = G\frac{m_1 m_2}{r^2}$
- The "model" is a kind of board game.
  - Items: G,m_1,m_2,r,F
  - The relationship : described by formula
- This model is found related with reality from the observation.



Statistical models

- Statistical Models: Formalization of Uncertainty
  - a mathematical model that embodies a set of statistical assumptions concerning the generation of sample data.
  - A statistical model represents, often in considerably idealized form, the data-generating process.
- A mathematical model is an abstract description of a concrete system using mathematical concepts and language.
- We know the exact properties of the mathematical models
- We use the statistical model to communicate about our data without ambiguity.

Probability: Chap02- Chap 06

- chap02 - chap 06.
- Our statistical models are based on the probability theory.
- How do we 'formalize' uncertainty in a mathematical way?
- How to represent them, how to manipulate them, and how to use them?
- It is a mathematical board game, which we can relate with the real world.

Assignments 1: Introduction

- Chap00.ipynb for prerequisites
- Chap01.ipynb for chapter 01. You will find the instruction in the file.
  - Read the book.
  - Pen and papers problems for chap 1
  - Python programming problems for chap 1
  - Python programming for reading data from file.