# Introduction to Statistics

8-

Thu. 9:00~11:45

Rstudio.

# Data Science

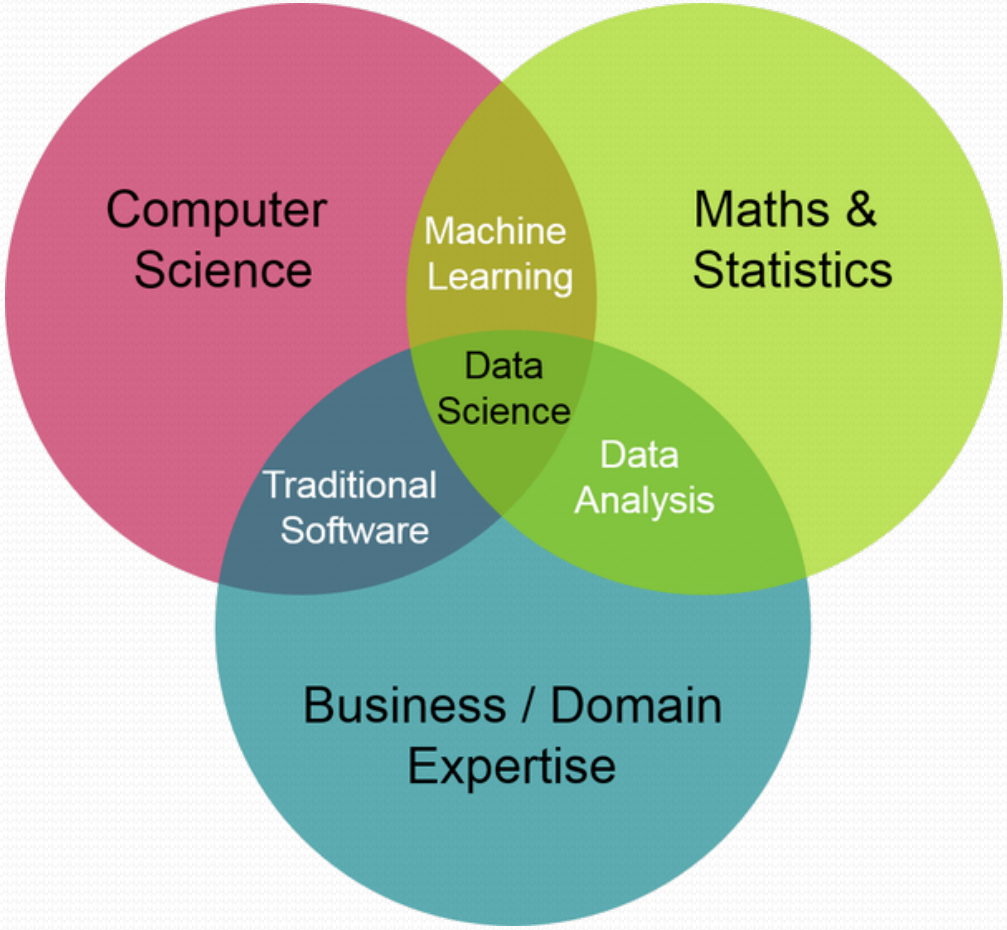
- Virginia Rometty, former chairman, president and chief executive officer of IBM said the following at Northwestern University's 157th commencement ceremony in 2015: "*What steam was to the 18th century, electricity to the 19th and hydrocarbons to the 20th, data will be to the 21st century. That's why I call data a new natural resource.*"

- Yuval Harari, author of Sapiens and Homodeus, mentioned "Dataism"

# Data Science

- With the rapid advances in computing and storage capacity, we now live in the “Big Data” era, where data are collected in real time and analyzed in science, business, industry, and government.
- Data Science has emerged as a prominent field of study because it can provide valuable insights using Big Data for making informed decisions in business, human health, security etc.
- There is considerable value in training students in the practice of extracting information from data.

# Data Science



# Data Science

- Data Science encompasses a wide range of concepts, methodologies and algorithms involved in collecting, managing, visualizing, analyzing, and transforming *data* into information, knowledge-creation and decision-making. + communications.
- The word *Data* refers to data collection, storage and management and retrieval, while the word *Science* deals with modeling, analysis, inference, interpretation, and decision-making.
- Voluminous data is being regularly collected and analyzed in science, business, industry, as well as by government and society at large.

# What is Statistics?

- Principle and methodology for designing the process of data collection, summarizing and interpreting the data, and draw conclusions or generalities

- Science of understanding data and of making decisions in the face of variability and uncertainty

# Origin of Statistics

- Status: Latin for ‘State’
- 統計
- First data record: Kushim Tablet (Uruk period, 3400-3000 BC)
- Clay tablet: used to record transactions of barley

Source: Wikipedia

# Statistics Procedures

- Research questions
- Design and data collection
  - Sample survey or experiment?
  - How to choose people (subjects) for the study, and how many?
  - How to conduct an experiment?
- Description - Graphical and numerical methods for summarizing the data.
- Estimation & Inference - Methods for estimating the quantities of interest and making predictions about a population (total set of subjects of interest), based on a sample (subset of the sample on which study collects data).
- Draw conclusions and make decisions

# Examples of Research Qs:

- How can we study whether a new therapy is better than a standard therapy for treating depression?
- How (if at all) is happiness associated with income, job satisfaction, family situation, social life, religious beliefs, political ideology?
- Can we predict college GPA using IQ, average time studying per week, high school GPA, SAT scores, number of hours spent on social activities, …?

# More Examples:

- Which parts of brain are associated with an eye movement task?
- Can we predict NFL draft rankings using college performance?
- Is there a difference in voting patterns across gender, age, region, etc?
- Can we predict disease or no disease using genes?

# Two Important Points

- **Uncertainty:** "Statistics is never having to say you are certain …"
- We say that statistics can be used to predict, but it is very important to know that we cannot predict with certainty. Statistical conclusions involve uncertainty; it may even be incorrect at times.
- Survey example: https://projects.fivethirtyeight.com/biden-approval-rating/
- Conclusion: "likely"
- Data ≠ information

  Data = signal (information) + noise

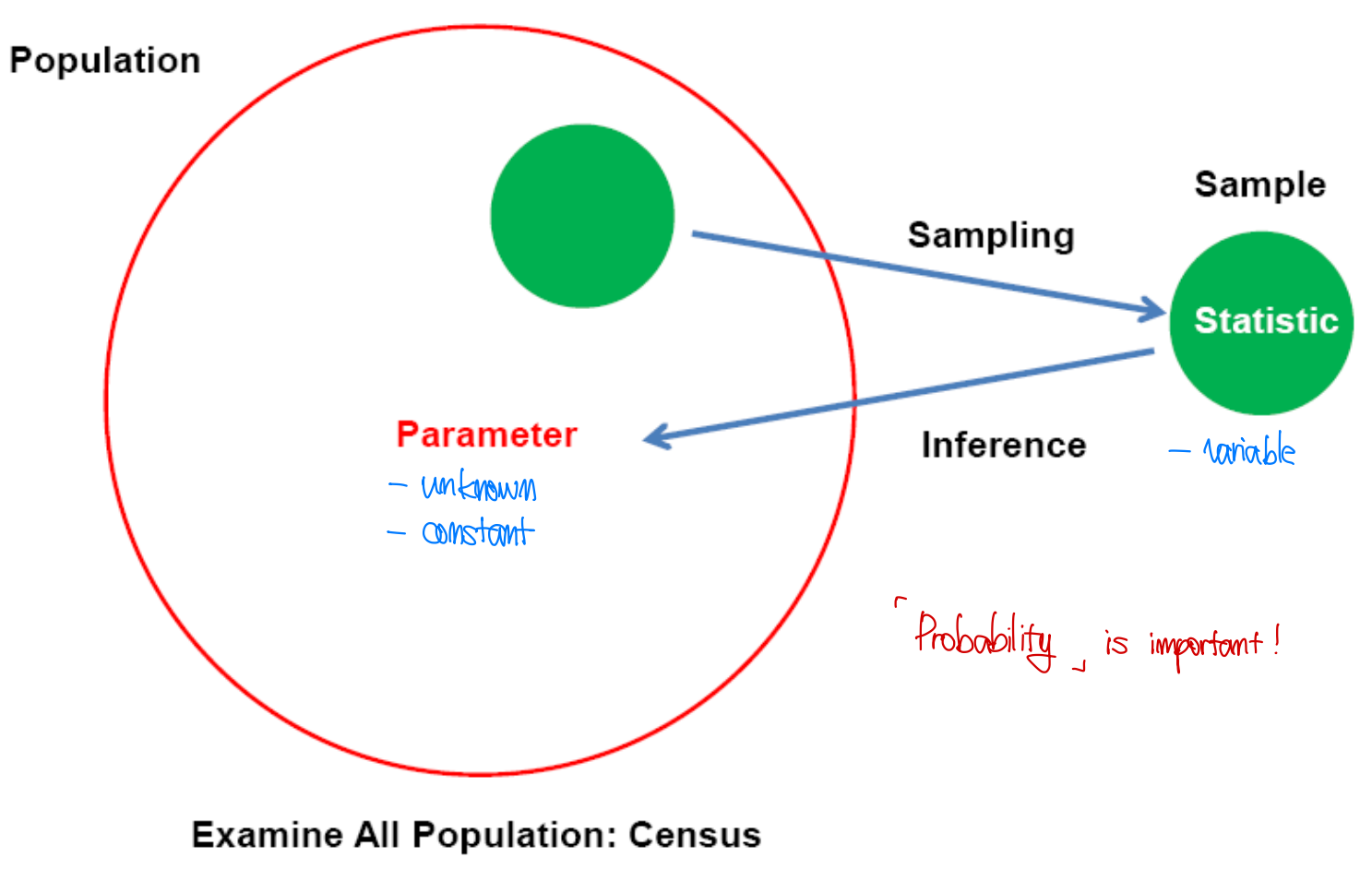
# Two Important Points

- Difference between Mathematics and Statistics
- Mathematics is *deductive* in nature:
- $x<2$, $y<3$: premise then $x+y<5$: conclusion
- From the fact $x<2$, $y<3$ we arrive at the conclusion $x+y<5$ with certainty. There is no element of uncertainty involved in this process. This conclusion always holds. Statistics is *inductive* in nature. Conclusions are derived based on data.

# Population vs Sample, Parameter vs Statistic

- **Population**: Complete collection of elements (scores, Yes/No, etc.) to be studied in the investigation.
- **Sample**: A subset of the population.
- **Parameter**: A numerical measure depending on *all* the units of a population describing some characteristic of the population.
  - Population mean (or median or some other measure)
  - Population proportion (or percentage)
- **Statistic**: A numerical measurement depending only on the *sampled* units describing some characteristic of the sample to shed light on the population analog, namely, the parameter.

# Population vs Sample, Parameter vs Statistic



A/D of all adults " n adults

# Population vs Sample, Parameter vs Statistic

Proportion of all adults " n adults

- Example 1:

A Realmeter poll, taken in July, 2023, reported the results from a sample of n=2,507 adults nationwide.
Interviewers asked the following question:

"*Do you approve or disapprove of the way our president is handling his job?*"

Statistics

Of the 2,507 adults in the sample, 38.1% responded by stating they approve of the President's handling of his job.

# Population vs Sample, Parameter vs Statistic

- Example 2:

A building contractor has a chance to buy an old lot of 5000 used bricks at an auction. She feels that in order for the lot of bricks to be worth purchasing for her current project, the percentage of all cracked bricks should be small. However, she does not have enough time to inspect all 5000 bricks. The person in charge of selling the bricks conjectures that only about 2% of all bricks are cracked. She agrees to buy the lot but only if the percentage of all cracked bricks is not more than 2%. She plans to take a sample of 100 bricks and determine the proportion of these bricks that are cracked. As a result of the sample, 3% of the bricks are cracked.

# Collecting Data

- **Sample survey**: Sample people from a population and interview them.
- A good sample will reproduce the characteristics of interest in the population as closely as possible. Some samples can be biased, they favor one side or they do not represent the entire population. For example:
  - Voluntary response sample: consists of people who choose themselves by responding to a general appeal (e.g., online polls, mail-in surveys, etc.). The problem with such samples is that people with strong opinions are usually the only ones which respond. This gives biased results!
  - Convenient sample: chooses individuals that are easiest to contact. The researcher makes no attempt, or only a limited attempt, to ensure that this sample is an accurate representation of some larger group or population.

# Collecting Data

- To avoid this, statisticians use **Randomization**. Random means not predictable or no discernable pattern, so we must use a randomization scheme rather than our (biased) judgement to get our sample. So how can we use random samples to help us predict/determine what is going on in the population? This is where probability comes in.

# Sampling Methods: SRS

- To reduce bias, one can use simple random sampling (SRS). Simple random sampling produces estimates which are unbiased. That is, the values of the statistic neither overestimate nor underestimate the value of the population parameter.
- A SRS is a sampling design where each sample of size n has an equal chance of being selected.
- With an SRS, because each sample of size n has the same chance of being selected, we cannot obtain information for separate "groups" of individuals (e.g., individuals of different gender, race, income class, religion, etc.).

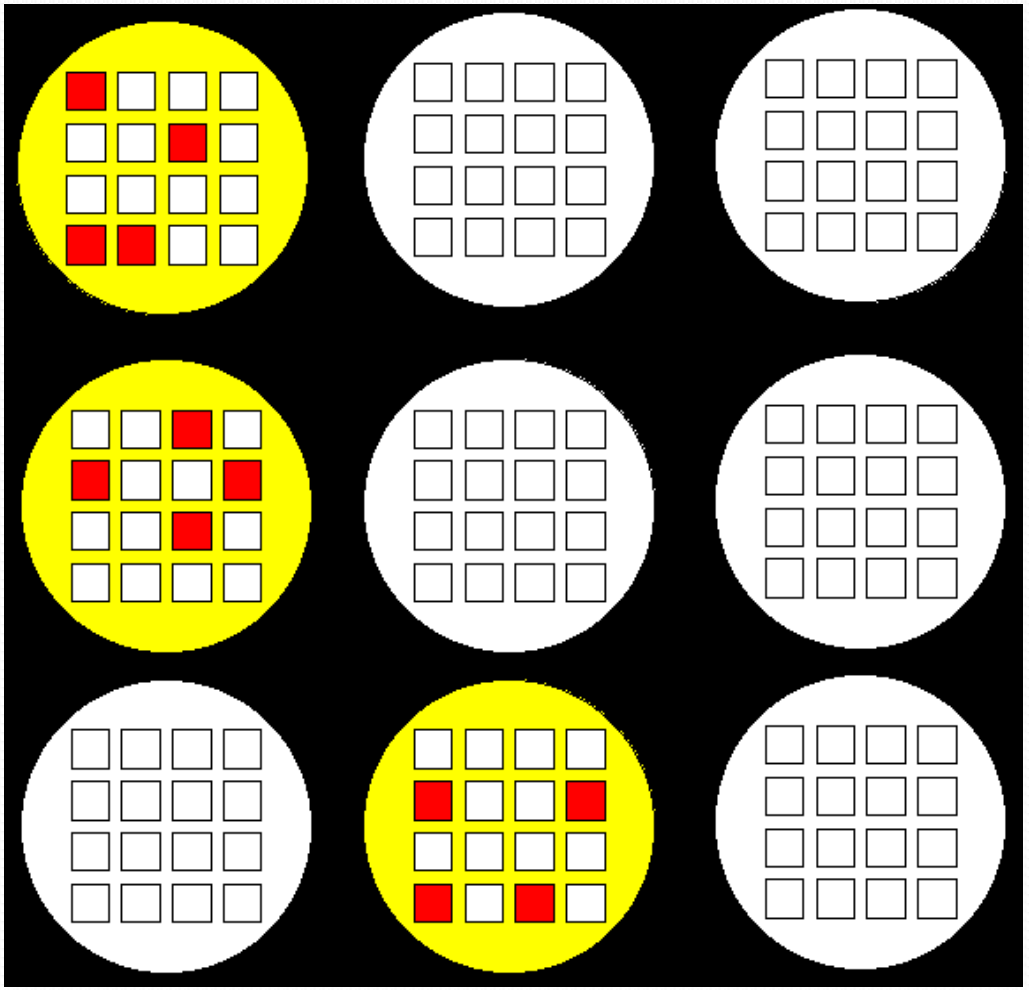
# Stratified Random Sampling

- First, divide the sampling frame into distinct groups or strata.
- Take simple random samples (SRS) within each stratum and combine these to make up the complete sample.
- Example: Suppose a public opinion poll designed to estimate the proportion of voters who favor spending more tax revenue on an improved ambulance service is to be conducted in a certain county. The county contains two cities and a rural area. The population of interest for the poll is all adults of voting age who reside in the county. A stratified random sample of adults residing in the county can be obtained by selecting an SRS of adults from each city and another SRS of adults from the rural area. That is, the two cities and the rural area represent three strata from which we obtain simple random samples.

# Cluster Sampling

Two stage of SRS.

- The population units are aggregated into larger sampling units called *Clusters*. A SRS of clusters is selected. Then from each selected cluster, we select a random sample of all or part of the subunits in each selected cluster.



# Experimental Designs vs. Observational Studies

- Observational Study: This involves collecting and analyzing data WITHOUT RANDOMLY assigning treatments to experimental units. Surveys are a common example of this. Can't measure causality.
  Ex. 담배와 암의 관계

- Designed Experiment: A treatment is RANDOMLY imposed on individual subjects in order to observe whether the treatment causes a change in the response.
  Ex. 백신 실험

# Experimental Designs vs. Observational Studies

- Errors that arise in designed experiments and observational studies

- Random sampling errors: these occur naturally in a sampling process
  - we can't avoid this type of error
  - these errors are well-understood when good sampling techniques are used
  - it is possible to determine to what degree these errors affect our outcome
  - these have the potential to be bigger in smaller samples than in larger ones
- Nonsampling errors: other than the sampling process
  - these are more problematic than sampling errors and should be of concern
  - it is virtually impossible to determine to what degree these errors affect our outcome
  - these can be minimized by using appropriate survey methodologies (using a random sample, e.g.) or appropriate experimental designs

# Experimental Designs vs. Observational Studies

- Examples of nonsampling errors:
- measurement error: Inaccuracies in measuring the true variable values due to, for example, the fallibility of the measurement instrument, data entry errors, or respondent errors.
- lurking variables: These are characteristics which affect the variable of interest but are not included in the study. Both observational studies and designed experiments are affected by these; however, designed experiments make it possible to control for lurking variables.
- nonresponse bias: An individual selected for a survey study does not respond. Certain groups may be less likely (or more likely) to respond, and the results could be biased.
- voluntary response bias: Respondents choose to take part in the survey study. People with strong opinions are more likely to reply, and the results could be biased.

# Example: 1936 Literary Digest Poll

- The magazine *Literary Digest* had been predicting U.S. presidential elections with great success. For this particular election, the magazine sent out questionnaires to 10 million people taken from lists of car owners, magazine subscribers, telephone directories, and registered voters. About 2.3 million responses were received. The result was, Roosevelt: 43% and Landon: 57%.However, When Election Day arrived, voters re-elected Roosevelt with an overwhelming 61 percent of the vote. What went wrong?

# Lurking Variables

- A teacher at an elementary school (grades 1 through 5) measures the heights of children on the playground and then makes a scatter plot of the children's heights and reading test scores. She finds that taller students tend to have higher reading scores.
- Would the association be positive or negative?
- What might be a lurking variable that affects both height and reading scores?

Age, Grade

# Variables

- **Variable**: a characteristic that can vary in value among subjects in a sample or a population.

- We need to know types of variables or measures, in order to evaluate the appropriateness of the statistical techniques used, and consequently whether the conclusions derived from them are valid. In other words, you can't tell whether the results in a particular medical research study are credible unless you know what types of variables or measures have been used in obtaining the data.

# Qualitative Variables

- **Qualitative (Categorical) variables**: involve assigning non-numerical items into groups or categories. The qualitative characteristic or classification group of an item is an attribute. Some examples of qualitative variable are:
  - The pizza was delivered on time

    Categorical Variable: Delivery Result

    Attribute: On Time, Not On Time
  - The survey responses include disagree, neutral, or agree.

    Categorical Variable: Survey Response

    Attribute: Disagree, Neutral, Agree
  - This car comes in black, white, red, blue, or yellow.

    Categorical Variable: Color

    Attribute: Black, White, Red, Blue, or Yellow.

* It orders variables

# Scales of Measurements

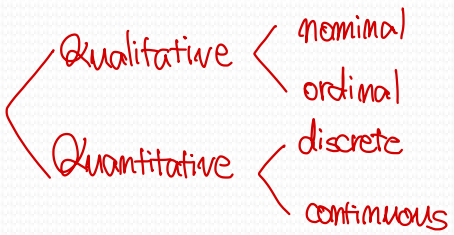
- Nominal Scale: What does the word “nominal" comes from? It has to do with naming. So, nominal comes from name and that is all you can do with variables measured on nominal scales (nominal variables). The important thing is there is no measure of distance between the values. You're either married or not married. The answer is determined, yes or no. So, there is no question of how far apart in a quantitative sense those categories are.
- Some other examples are sex (male, female), race (Black, Hispanic, Oriental, White, other), political party (democrat, republican, other), blood type (A, B, AB, O), and pregnancy status (pregnant, not pregnant).

# Scales of Measurements

- Ordinal Scale: Ordinal implies order (ranking). So the things being measured are in some order. For example, you don't rank male and female as higher and lower. But you do rank stages of cancer, for example, as higher and lower. You can rank pains as higher or lower.
- Ordinal scales both name and order. Some other examples of ordinal scales are rankings (e.g., soccer top 6 teams, pop music top 10 songs), order of finish in a race (first, second, third, etc.), and cancer stage (stage I, stage II, stage III).

# Quantitative Variables

- **Quantitative variables**: result from measurement or numerical estimation. These measurements yield discrete or continuous variables.
- **Discrete variables** vary only by whole numbers such as the number of students in a class (variable: class size).
- **Continuous variables** vary to any degree, limited only by the precision of the measurement system. Some examples include the width of a desk, the time to complete a task, or the height of students (variables: length, time, and height).



# Experimental Designs vs. Observational Studies

- Sometimes, a researcher is interested in finding association between multiple variables, e.g. the effect of GRE scores on graduate school admission
- **Response variable:** the focus of a question in a study or experiment.
- **Explanatory variable** (predictors): one that explains changes in that variable. It can be anything that might affect the response variable.

# Why Study Statistics?

- One answer: You need it to understand research findings based on data in psychology, medicine, business, …
- “The best thing about being a statistician is you get to play in everyone's back yard." - John Tukey, famous statistician.
- Another answer: In a competitive job market, understanding how to deal with quantitative information provides an important advantage. (“The sexy job of the next 10 years will be statistician" Hal Varian, chief economist at Google)
- Broader answer: In your everyday life, it will help you make sense of what to heed and what to ignore in statistical information provided in news reports, medical research, surveys, political campaigns, advertisements, …

# Why Study Statistics?

- Thomas J. Sargent, winner of the 2011 Nobel Prize in Economics, said
- “As scientists, scientists in many fields need to speak with statistical data… Artificial intelligence is actually statistics, but with a very gorgeous phrase, in fact, is statistics. Many of the formulas are very old, but we say that all artificial intelligence uses statistics to solve problems.”

# Why Study Statistics?

- Sci-fi writer Ted Chiang said:
- “There was an exchange on Twitter a while back where someone said, ‘What is artificial intelligence?’ And someone else said, ‘A poor choice of words in 1954,’” he says. “And, you know, they’re right. I think that if we had chosen a different phrase for it, back in the ’50s, we might have avoided a lot of the confusion that we’re having now.” So if he had to invent a term, what would it be? His answer is instant: applied statistics. “It’s genuinely amazing that . . . these sorts of things can be extracted from a statistical analysis of a large body of text,” he says. But, in his view, that doesn’t make the tools intelligent. Applied statistics is a far more precise descriptor, “but no one wants to use that term, because it’s not as sexy”.