Topic 1: Introduction to Statistics & Describing Data with Graphs/Numerical Measures

Optional Reading: Chapter 1-3

Xiner Zhou

Department of Statistics

University of California, Davis

- What is Statistics?
- How to describe univariate categorical data?
- How to describe univariate numerical data?

What is Statistics? Why we need it?

Birth:1662

John Grant (founder of demography) and William Petty (Economist), developed early statistical methods to analyz demographic data, esp. 1st life table giving probability of survival to each age

18th century

the evolution of statistics was directly due to the need of collect information about European states, particularly demographics such as population, for the purpose such as governance and taxation

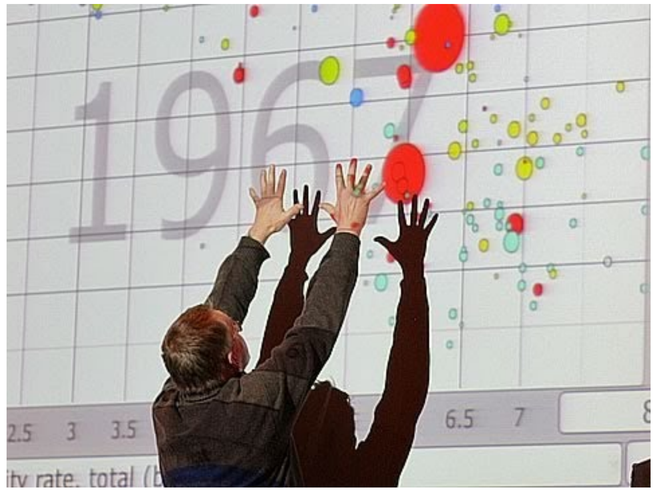
from German statistisch (adjective), Statistik (noun).

Modern world

Statistics has applications in almost every facet of our daily life, and almost all fields of a scientific or commercial nature

Examples of faculty and Data Scientists working on:

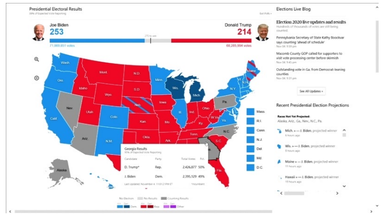
- 200 years of history through health and wealth
- Using statistics to treat chronic illness
- Building a better NBA team through analytics
- Bringing life to global health statistics



Overview of statistical analysis workflow:

Specify the question to be answered

- An analysis can only be as good as the question asked



Pollster: who will get the more votes on the election day?

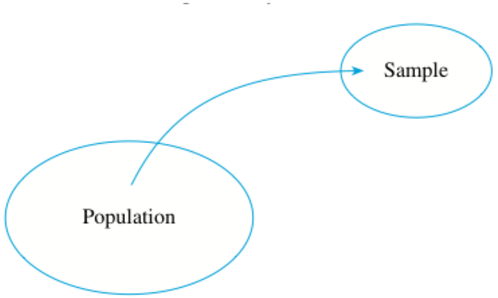
Identify the population of interest

- Population is the whole body of measurements of interest
- The question can often be summarized as parameters of the interest

Every eligible voters in US is the population
Proportion of voter who support Democrat is the parameter of interest
But can you possibly interview every voter?

Sample: is a subset of measurement or data drawn from the population

- We try to describe or predict the behavior of the population on the basis of information obtained from a representative sample from the population.



Use sample information to infer parameters corresponding to the question

- Descriptive statistics (Topic 1):
    - To summarize and describe the important characteristics via graphs or numerical quantities

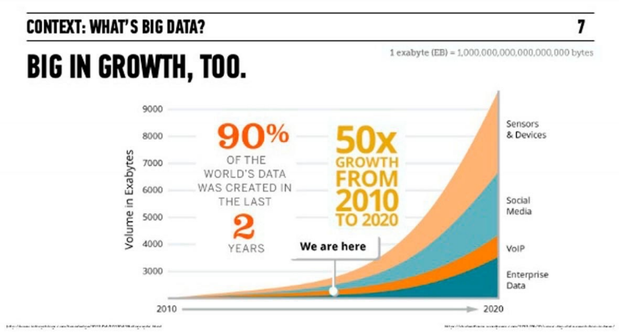
- Inferential Statistics (Topic 4-6):
    - To make inference (draw conclusion, make prediction, make decision about a hypothesis……) about the parameter, from the information contained in the sample we have

Data

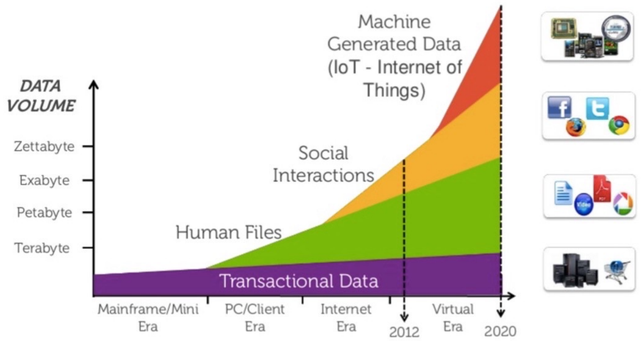
Statistics is an old subject, Data Science is a new emerging star.

How to analyze data?

Driven by our ability to collect much more data
But data is data



The Explosion of Data



A typical data set:

	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	7500	7.34	36	A	MD	70000	rent
2	25000	9.43	60	B	OH	254000	mortgage
3	14500	6.08	36	A	MO	80000	mortgage
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	3000	7.96	36	A	CA	34000	rent

Variable:

- is a characteristics that change or varies for different individual or objects
- By columns

Observation:

- By rows

How many variables we have:

- When a single variable is measured — univariate data
- When two variables are measured — bivariate data
- When more than 2 variables are measured — multivariate data

Types of Variables

Qualitative or Categorical:

Measure a quality or characteristic that can be categorized

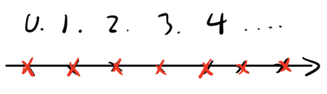
- Political affiliation: Republican, Democrat, Independent
- Taste ranking: excellent, good, fair, poor
- Color of an MM'S candy: brown, yellow, red, orange, green, blue

Quantitative or Numerical:

Measure a numerical quantity

- Prime interest rate
- Number of passengers on a flight from Los Angeles to New York City Weight of a package ready to be shipped
- Volume of orange juice in a glass

Discrete:
can only take finite or countable number of values



Continuous:
can take infinitely many values on a line interval

1. The most frequent use of your microwave oven (reheating, defrosting, warming, other)
2. The number of consumers who refuse to answer a telephone survey
3. The door chosen by a mouse in a maze experiment (A, B, or C)
4. The winning time for a horse running in the Kentucky Derby
5. The number of children in a fifth-grade class who are reading at or above grade level

Why different types of variable?
Different types of data require different methods of description and analysis

How to describe univariate categorical data?

In a survey about public education, 400 school administrators were asked to rate the quality
A B A A D C B A .....

Good visual description:

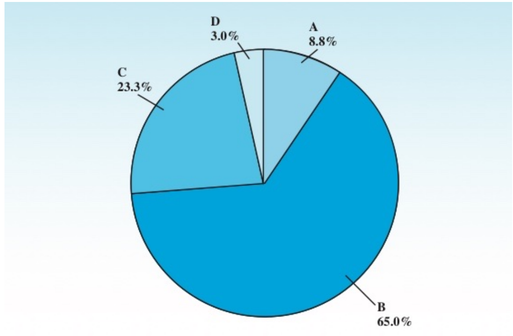
- "How often" each value occurred?
    - The frequency, or number of measurements in each category
    - The relative frequency, or proportion of measurements in each category
    - The percentage of measurements in each category

$$\text{Relative frequency} = \frac{\text{Frequency}}{n}$$

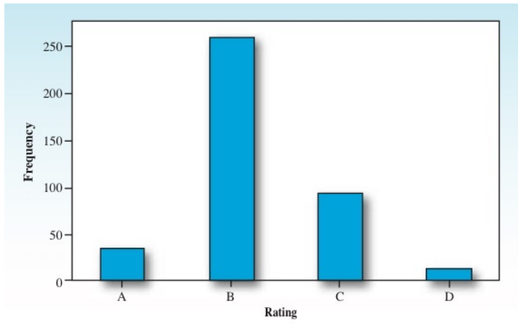
$$\text{Percent} = 100 \times \text{Relative frequency}$$

Rating	Frequency	Relative Frequency	Percent
A	35	35/400 = .09	9%
B	260	260/400 = .65	65%
C	93	93/400 = .23	23%
D	12	12/400 = .03	3%
Total	400	1.00	100%

Pie chart



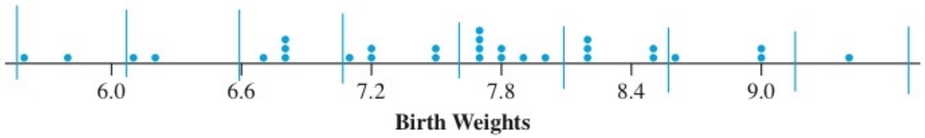
Bar chart



How to describe univariate numerical data?

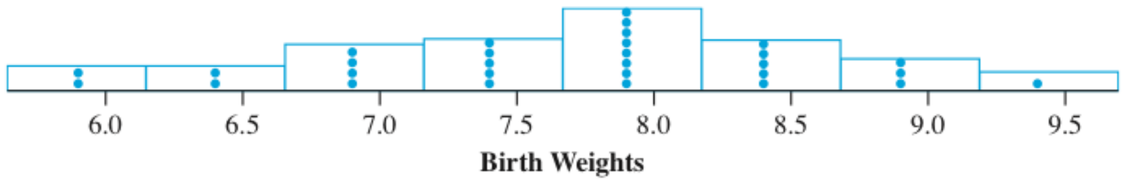
Birth Weights of 30 Full-Term Newborn Babies

7.2	7.8	6.8	6.2	8.2
8.0	8.2	5.6	8.6	7.1
8.2	7.7	7.5	7.2	7.7
5.8	6.8	6.8	8.5	7.5
6.1	7.9	9.4	9.0	7.8
8.5	9.0	7.7	6.7	7.7



Histogram:

- Divide the range of observations into subintervals of equal length
- Use frequency/ relative frequency/percentage to represent “how often” observations fall into each subintervals
- Proceed as bar chart
- It’s essentially a bar graph for continuous variable, by discretizing



Histogram

Graphs are great: A picture worth a thousand words
But they are not precise enough to be useful alone

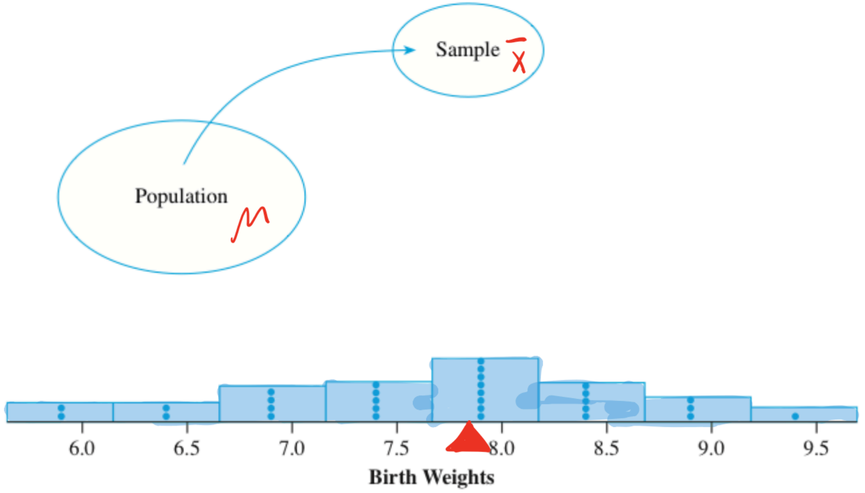
Numerical measures:
To describe and summarize the distribution of data

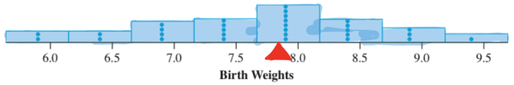
1. Measure of center —- Mean

Sample $x_1, x_2, \ldots, x_n$:

sample mean is $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

Population: population mean μ



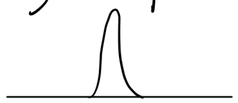


how far away the data
tends to spread around the center

Variability or dispersion

- design a test

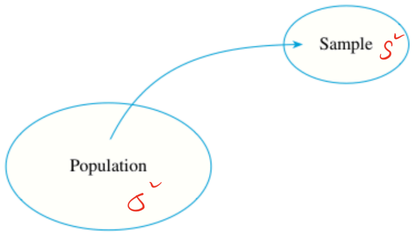
- quality control for manufacturing a product



2. Measure of Variability: Variance

Sample $x_1, x_2, \ldots, x_n$:
sample variance is $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$

Population: population variance σ^2



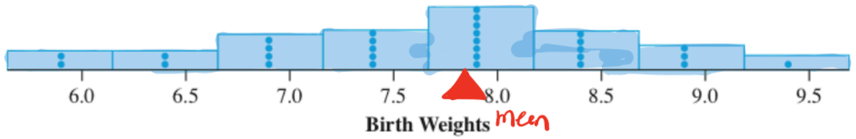
Variance will be large for high variable data, small for less variable data

But, variance is measured in terms of square of original units of measurements -> take square root

3. Standard deviation

Sample $x_1, x_2, \ldots, x_n$:
sample standard deviation is $S^2 = \sqrt{S^2}$

Population: population standard deviation σ

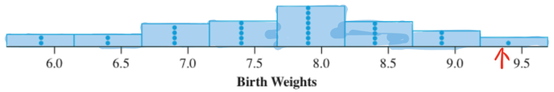


standard deviation : typical deviation from the mean

4. Measure of Relative Standing

Birth Weights of 30 Full-Term Newborn Babies

7.2	7.8	6.8	6.2	8.2
8.0	8.2	5.6	8.6	7.1
8.2	7.7	7.5	7.2	7.7
5.8	6.8	6.8	8.5	7.5
6.1	7.9	9.4	9.0	7.8
8.5	9.0	7.7	6.7	7.7



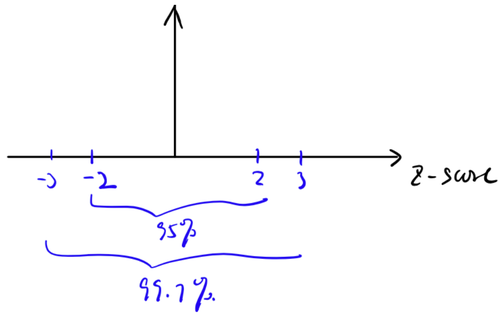
Birth Weights

the position of one observation relative to others

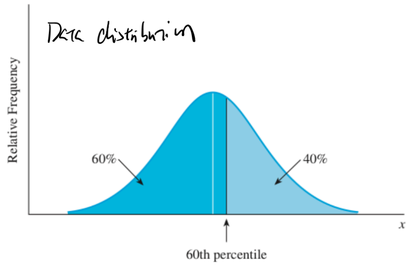
$$\text{Z-score} = \frac{x - \bar{x}}{S}$$

Measures the distance between an observation x and the sample mean, in units of standard deviation

We will see later, z-score is a tool for determining how likely an observation to occur under normal circumstances.



The pth percentile is the value that is greater than p% of measurements, and is less than the remaining (100-p)% of measurements.

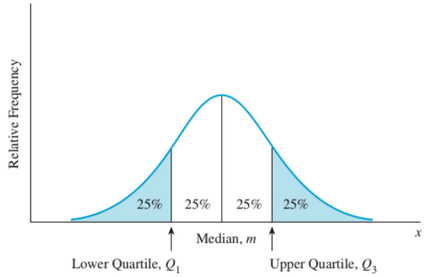


the value that is > 60% of the measurements
< 40% of the measurements

Important percentiles:

- 25th percentile = lower percentile =first quartile =Q1
- 50th percentile = median =second quartile
- 75th percentile = upper percentile =third quartile =Q3

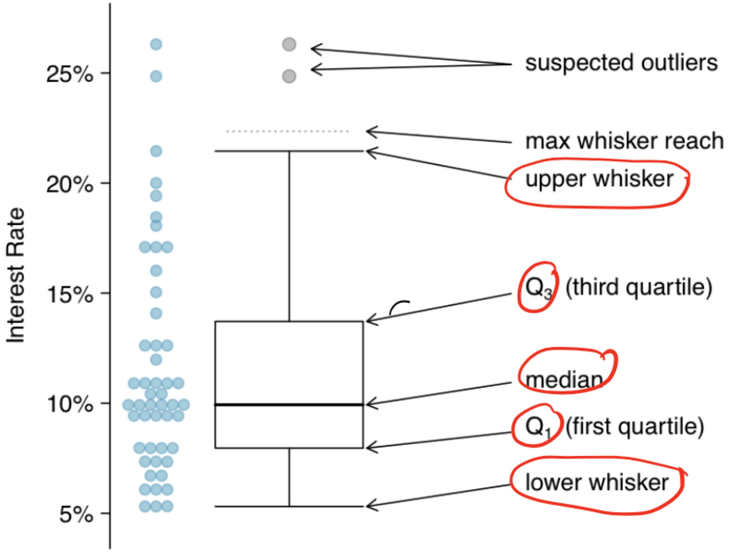
=> divide data into 4 sets, each containing equal proportion of measurements



range of the "middle 50%" of distribution

Interquartile range (IQR)
IQR=Q3-Q1

A graph based on just the notion of percentiles is the box plot:
Summarize a set of measurements using 5 numbers



- The box captures "middle 50% of the data
- Whiskers attempt to capture the range that can be considered not too extreme
    - Upper whisker = Q3+1.5 IQR
    - Lower whisker = Q1-1.5 IQR
- Any observation beyond the whiskers is unusually distant from the rest of the data, called outliers
    - Outliers can due to: data collection error, data entry error, or interesting property