Stat 230 Introductory Statistics Course Introduction & Descriptive Statistics

Instructor: Dr. Yas Yamin

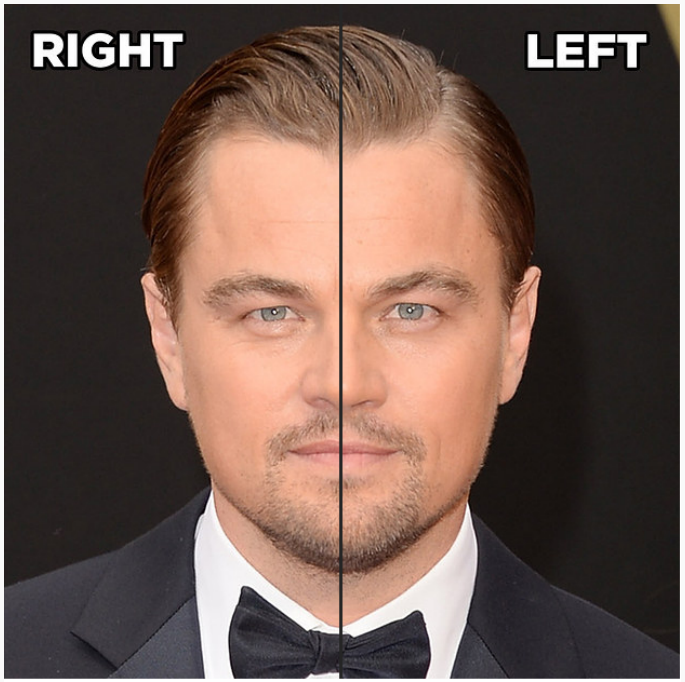
University of British Columbia Okanagan
yas.yamin@ubc.ca

What is Statistics? (and why you should care)

- The motivation behind statistics is that we are always confronted with real questions where we don't have all of the information, and we still need to answer those questions.
- Statistics provides the methods to study these kinds of problems and assists you in making decisions.
- It includes aspects of
  - the collection,
  - processing,
  - analysis,
    and
  - interpretation

  of data.
- Statistics aims to uncover the intrinsic characteristic or characteristics of a population through the use of samples, where the important information may be hard to see, because of the variablility in the sample measurements.

Example



- Is a face symmetric?
- Asymmetric, if you look closely at the detail of the right and left side.
- A question in digital surveillance might be: how do you match an observed face coming from a video with a database of facial measurements, where only one side of the face has been observed, and maybe only the other side of the face is recorded in the database?

`https://miro.medium.com/max/1252/1*_O2K9-zkalcXaCoKIVxhvA.png`

Data

- In order to find a matched face, we have to study the left and right sides of the faces from different people beforehand to determine the amount of variability in the measurements.
- The information on faces that is recorded for this kind of study is an example of data.
- There are many different types of data, corresponding to different types of information.
- Broadly, data can be categorical or numerical.
- In this course, we will most often be concerned with numerical measurements.

Examples of Statistical Questions

These might relate to

- watching the nightly news or reading a newspaper
- playing cards, board games, online games and gambling
- researching the risks of a medical drug
- observing signals coming from outside our galaxy
- understanding the risks of travel by air or car
- predicting when a region will have a major wildfire
- detecting climate change from atmospheric data

Types of Data —– Categorical

Binary Data can only take on two possible values

- Are you diabetic? yes or no.
- What is the outcome of flipping a coin? heads or tails.

Nominal Data can take on one of a limited (and usually fixed) number of possible values

- Eye colour: blue, green, brown, grey.
- Nationality: Irish, English, Scottish, French, Welsh, Canadian, etc.

Ordinal Data can be sorted by rank

- Questionaire response: Strongly disagree, Disagree, Neither agree nor disagree, Agree, Strongly agree
- Rating from 1 to 10
- Spiciness: mild, medium, hot.

Types of Data —- Numerical Data

Continuous data can take any real numerical value.

- volume of water inside a glass
- distance between cities
- temperature of a liquid at its boiling point

Discrete data typically take values in the set of counting numbers

- e.g. 0, 1, 2, 3,. . .
- number of children in a family
- number of bacteria on a piece of raw meat
- number of hurricanes that hit land each year

Some Vocabulary and Notation

- A population is the collection of objects or measurements that we wish to understand. Most of the time we are not able to study the whole population. Examples are
  - UBC Okanagan students
  - Temperatures at the Earth's surface
  - The life spans of bottle-nosed dolphins
- A sample consists of a number of observations.
  - The group of you sitting in this classroom is a (biased) sample of UBC students...and each of you individually are an observation
  - Temperature measurements recorded at a collection of weather stations around the world
  - The times from birth to death of a collection of bottle-nosed dolphins raised in captivity
- We often denote the observation measurements as $x_1, x_2, \ldots, x_n$. In this way, x_1 is the first observation, x_2 the second, and so on

Descriptive Statistics I

- Consider n observations: $x_1, x_2, \ldots, x_n$.
- We call this observed data $\{x_1, x_2, \ldots, x_n\}$, a *sample*
- The mean, or average, of the sample is denoted $\overline{x}$ and is given by

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

- The standard deviation of this sample s is given by

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$

Descriptive Statistics II

- The variance is square of the standard deviation s^2.
- The range of a sample is the difference between the minimum observation and the maximum observation

$$\text{range} = \text{maximum - minimum}$$

- When the observations are ordered from smallest to largest, the median is the 'middle' observation

Descriptive Statistics II

- The pth percentile is a value that has $100p\%$ of the ordered data falling below it, and $100(1-p)\%$ of the ordered data falling above it.
  - eg. the median is the 50th percentile

- We call the 25th percentile the first quartile (Q_1) and the 75th percentile the third quartile (Q_3)

- The interquartile range (IQR) is the size of the gap between the first and the third quartile. It is the 'distance' over which the 'middle half' of the data is spread.

$$\text{IQR} = \text{Q3 - Q1}$$

Example

Example 1

(meansd) Consider the following data: 6, 7, 8. What is the mean and standard deviation?

Example 2

missmean The mean of four numbers is 12. Three of these numbers are 2,11 and 19. What is the other number?

Exercises

1. Consider the data: 12, 9, 5, 8, 10, 1, 6, 11, 3, 7, 2, 4. What is the range, the median, Q1, Q3 and the IQR?

Exercise

Example 3

Consider the following data: 0.13, 0.25, 0.31, 0.44, 0.49, 0.51, 0.55, 0.59, 0.70, 0.81, 12.00. What is the IQR? Given that s=3.48, comment on the difference between s and the IQR.

Graphs

- Looking at data is a very useful first step.
- Often, a picture is the most useful way to summarize data.
- We will see various graphs as we go along, including histograms and scatterplots. These will be introduced as we need them.