Introduction to Statistics

STAT 330 - Iowa State University

Outline

In this lecture students will be introduced to the field of Statistics. Some terminology will be presented as well as different subtopics of Statistics.

Terminology

Population and Sample

Population: All individuals (or items) of interest.

- Typically, we want to learn *something* about the population
- Usually impossible to get information from entire population

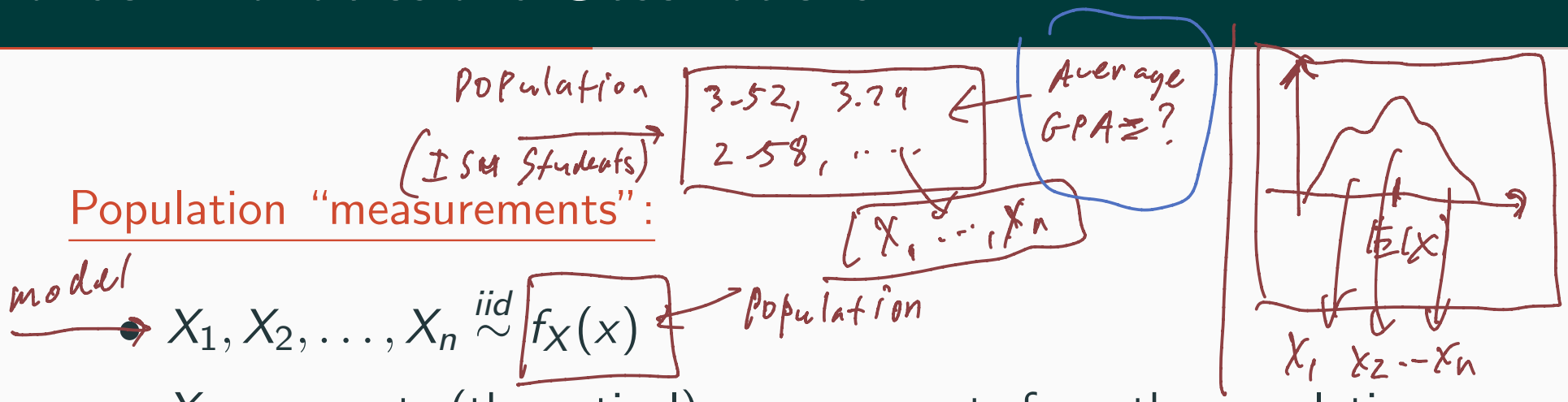
Sample: A subset of the population

- Since samples are much smaller than population, it possible to actually get information about the sample
- This information is called *"data"* or *"sample data"*.

Statistics: (as a field)

- Use probability to learn about the real world (population) from data (sample).
- Assuming random mechanism generated the data allows us to use probability.

Random Variables and Observations



Population "measurements":

- $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_X(x)$
- X_i represents (theoretical) measurements from the population.
- $f_x(x)$ represents the population distribution.

Observations:

- $x_1, x_2, \ldots, x_n$ are the observed data (or realization of the random variables).
- x_i are actual measurements from the sample.

Use the observed values $(x_1, \ldots, x_n)$ to learn about the population.

Example

Example 1: A machine fills bottles of water. We are interested in the amount of water filled in the bottles.

- X_i = amount of water filled in bottle i for $i = 1, \ldots, n$
- X_i follows *some* distribution (with *some* parameters).
- But, its impossible to measure *every* bottle that the machine fills

So take a *sample* of n bottles from the machine, and measure the amount of water in them.

- Gives observed values $x_1, \ldots, x_n = (500.01, \ldots, 499.80,)$
- Use this information to understand how much water the machine fills in general (population)

Drawing Samples

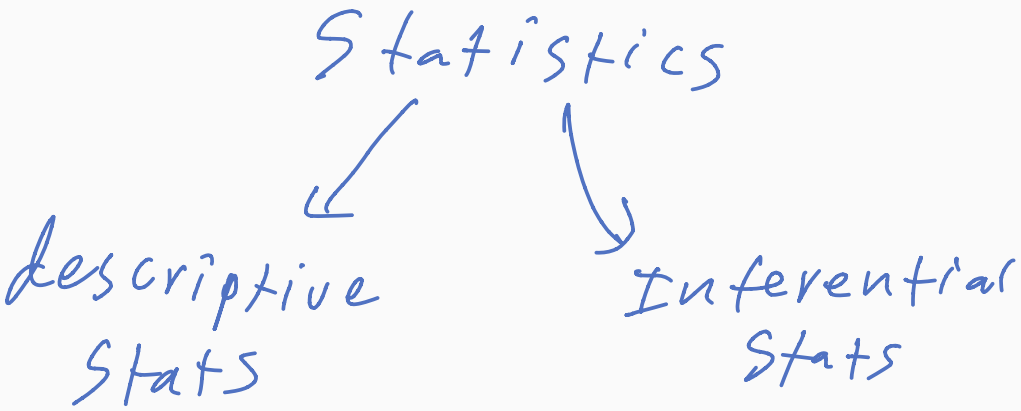
Typically, we assume a *simple random sample (SRS)* is drawn from the population to create our sample $X_1, \ldots X_n \overset{iid}{\sim} f_X(x)$

- All subsets of same size are equally likely to be chosen
- Guarantees the sample is representative of population

  $\rightarrow$ leads to good inferences
- If not, we will introduce *bias* in our sample

  $\rightarrow$ inferences can be way off from the truth

  $\rightarrow$ leads to untrustworthy results



Descriptive Statistics

Descriptive Statistics

Once we have obtained data from the sample, what comes next?

Descriptive Statistics: Describe/summarize key features of the data

- Graphics → visualize the data, describe shape, etc

- Numbers → numerical summaries of quantities of interest

Averages, proportions, - - -

No conclusions are made yet. We just want an idea of what the data looks like.

Inference (learning)

Estimation

Inferential Statistics: Draw conclusions about the population/distribution that generate the data.

1. Estimation: Estimate the parameters of the probability distribution that generated the data

   - In probability portion of the course, we assumed we knew the parameters of the distribution to answer questions

     - Ex: Get average of 5 hits per hour to a website. $X = \#$ of hits in next hour. $X \sim Pois(5)$. What is $P(X < 3)$?

   - In statistics, parameters are unknown and need to be estimated by the data

     - Confidence intervals
     - Hypothesis testing

Prediction

2. Prediction: Estimate parameters of a data model, then use model to predict values for new observations

Example 2: X = ACT score; Y = Freshman GPA

We model relationship as between X and Y as:

$$Y = f(X) + \epsilon$$

Use data to learn about the form of $f(X)$, "fit" a model, and then we can predict the GPA of a new student based on their ACT score.

$$\hat{Y}_{new} = \hat{f}(x)$$

Recap

Students should now be familiar will some terminology that we will encounter in the field of Statistics as well as topics for remaining lectures.