STA 1013 : Statistics through Examples

Lecture 1: Overview, Terminology

Hwiyoung Lee

August 28, 2019

Department of Statistics, Florida State University

Overview

1. What is Statistics ?
2. Course Overview
3. Topic 1 : Terminology

What is Statistics ?

What is Statistics ?

Statistics :

- The practice or science of collecting and analyzing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.

- Collecting Data
- Presening Data
- Analyze Data
- Interpreting results and Drawing conclusions

Statistics helps researchers organize and interpret the data. And also Statistics is a useful decision making tool under uncertainty.

What we can do with the help of Statistics ?

- **Nuclear Science**

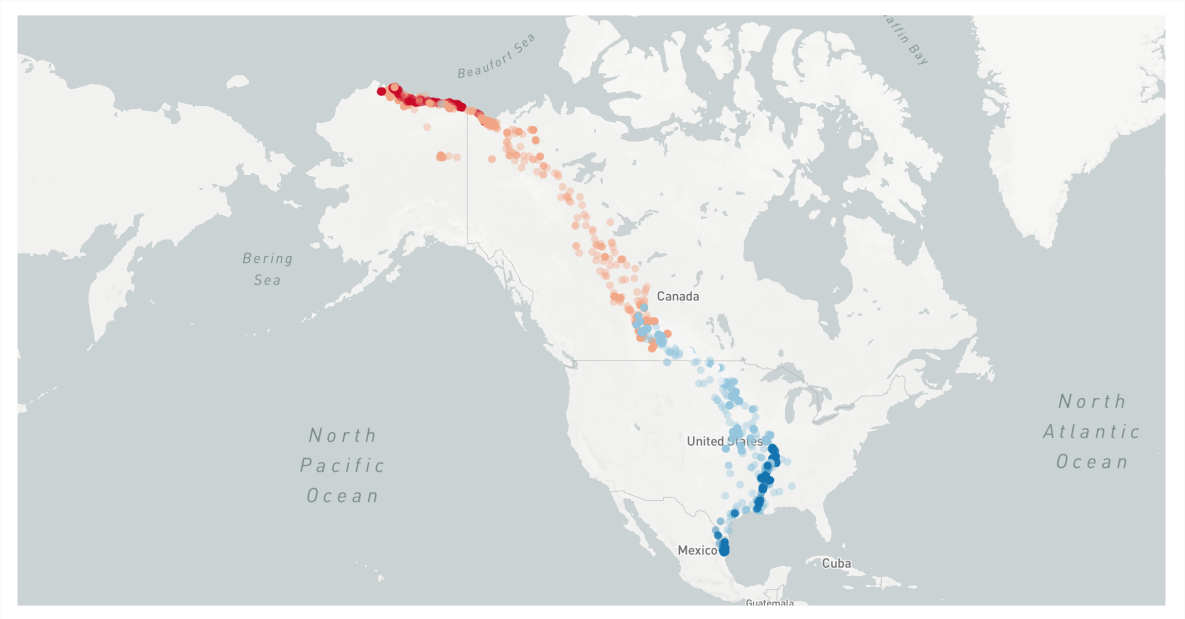
  Study the safety of nuclear power plants

- **Environmental Science**

  Evaulate the environmental impact of pollution

  Studies of plant and animal populations

  Estimate animal migration patterns

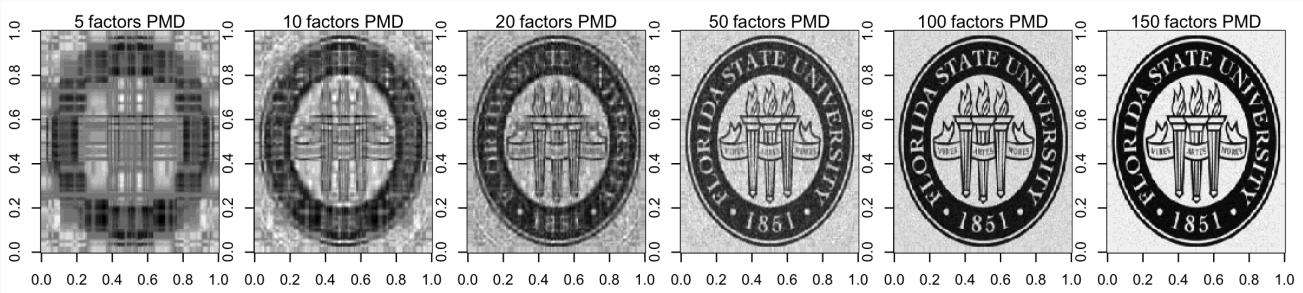


What we can do with the help of Statistics ?

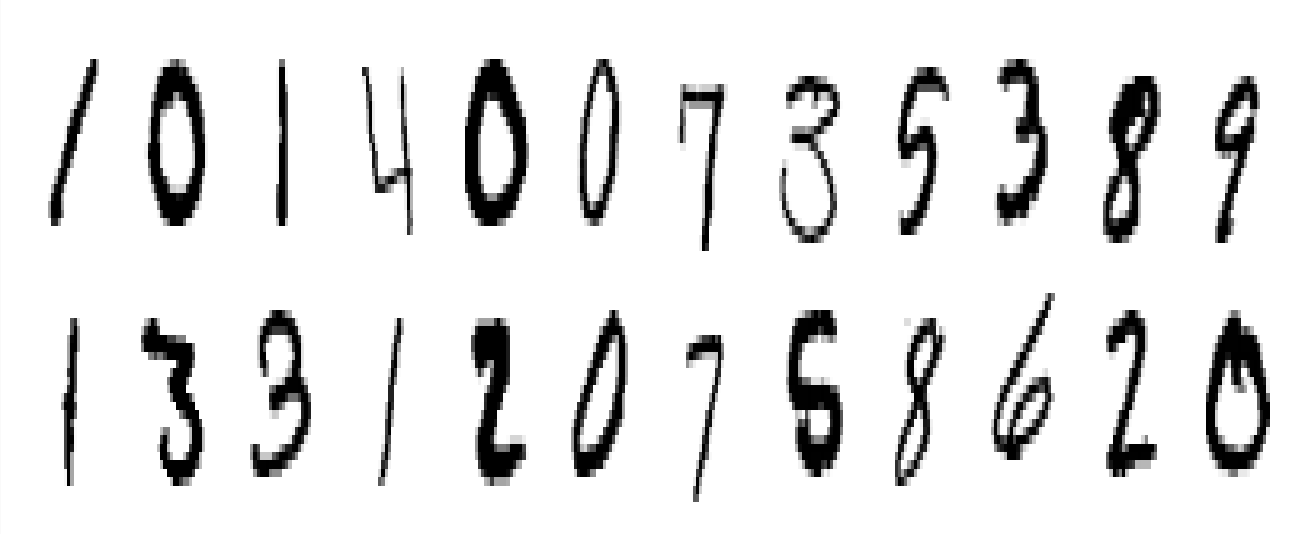
- **Machine Learning**

  Developing A.I system

  Image reconstruction



  Recognize handwritten digits



What we can do with the help of Statistics ?

- **Biostatistics, and Genetics**

  Determine the effectiveness of new drugs

  Determine the optimal amount of dosage of drug to treatment

  Analyzing Medical image data (Neuroimaging, MRI,...)

- **Economics**

  Estimate the next year unemployment rate

  Predict stock market trend, price

⋮

Nowadays Statistics is used in almost every field of science.

Course Overview

Course provides methods for

- **Design**

  Planning and carrying out research studies (Sampling methods)

- **Description**

  Visualizing and Summarizing data

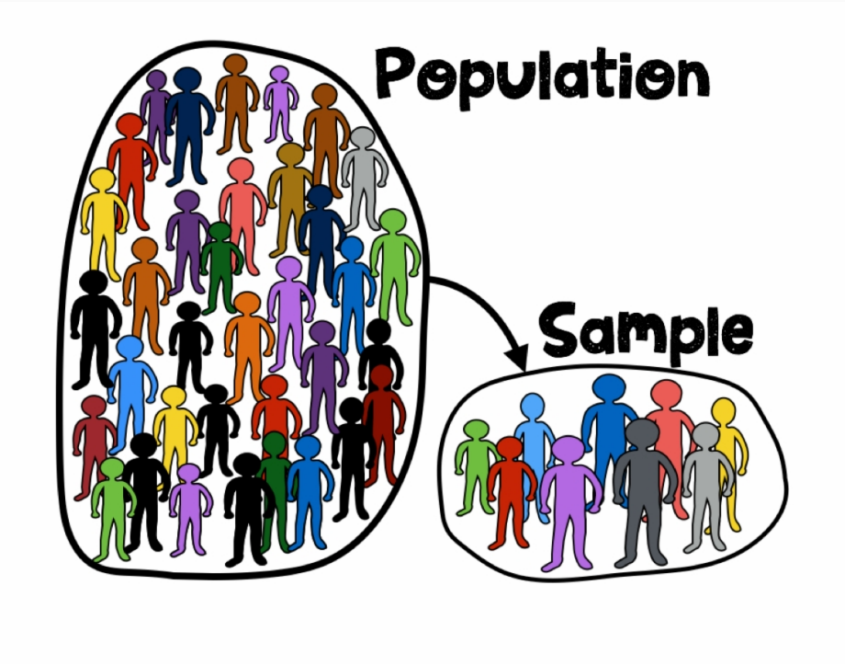
- **Inference**

  Making predictions : Estimate parameters, and response

  Making decision : Hypothesis testing

Research Design

Sampling → Topic 2



Descrption of Data

We might want to summarize data. This can be done

- **Visually** → Topic 3

  Histogram

  Boxplot

  Stem and leaf plot

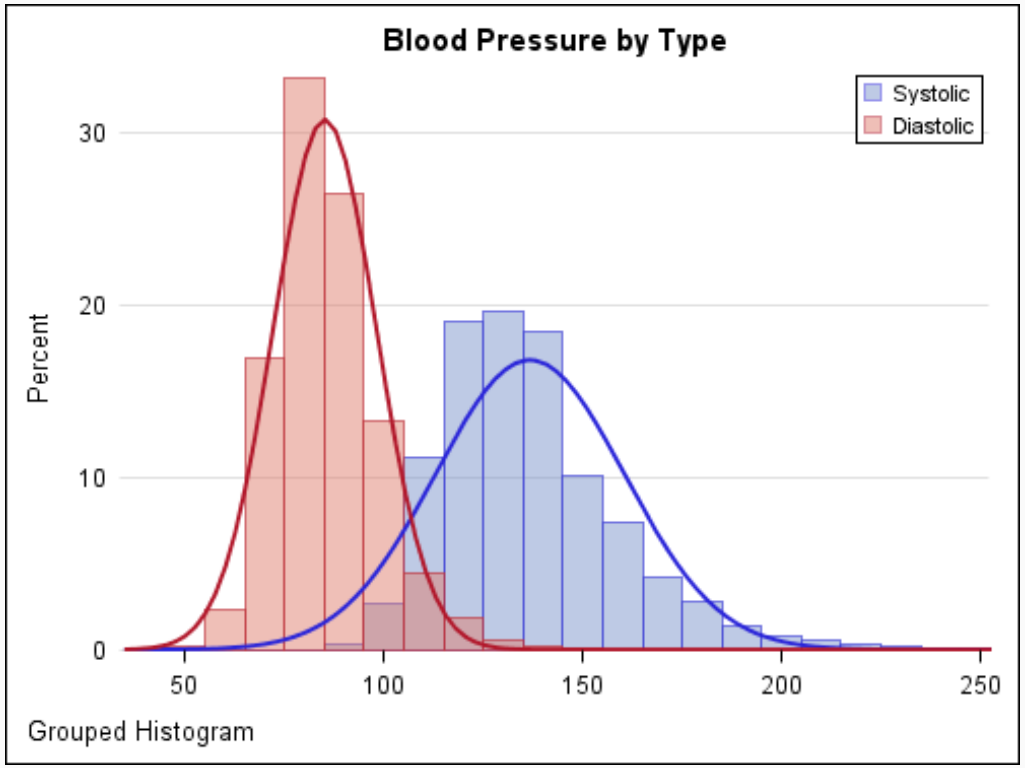
  ⋮

- **with Numbers** → Topic 4

  Center : Mean, Median, Mode

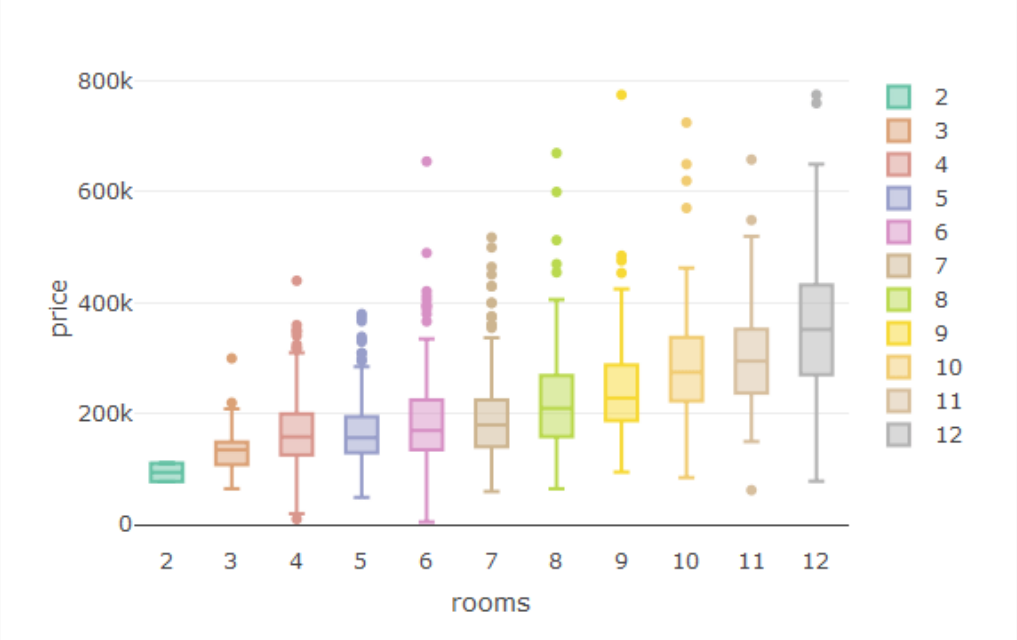
  Dispersion : Variance, Standard deviation, IQR

Histogram


Blood Pressure by Type
Systolic
Diastolic
Percent
30
20
10
0
50
100
150
200
250
Grouped Histogram

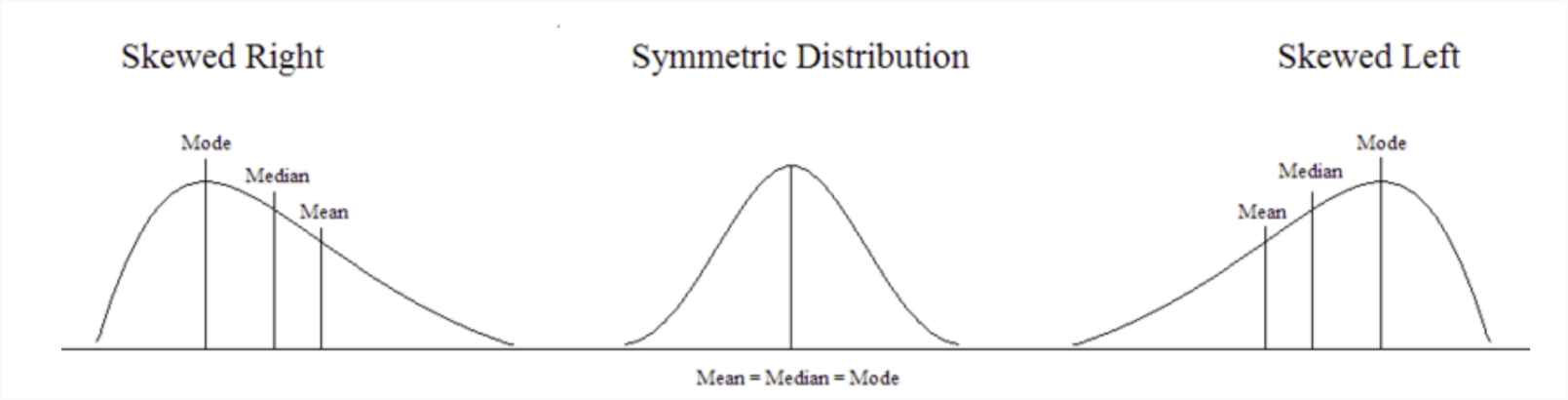
Visualization

Box plot

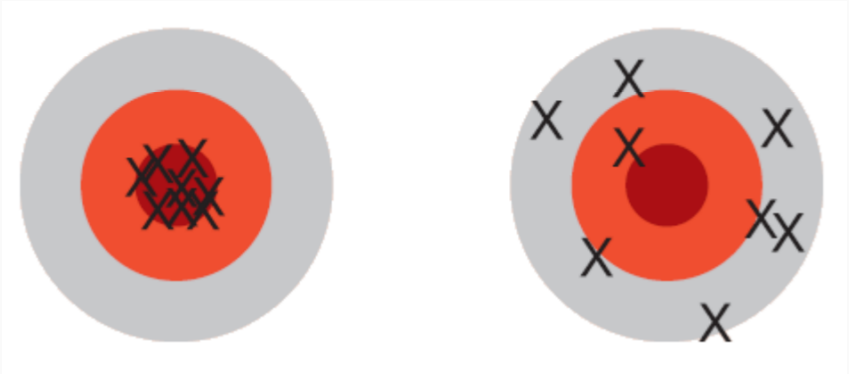


Summarizing data

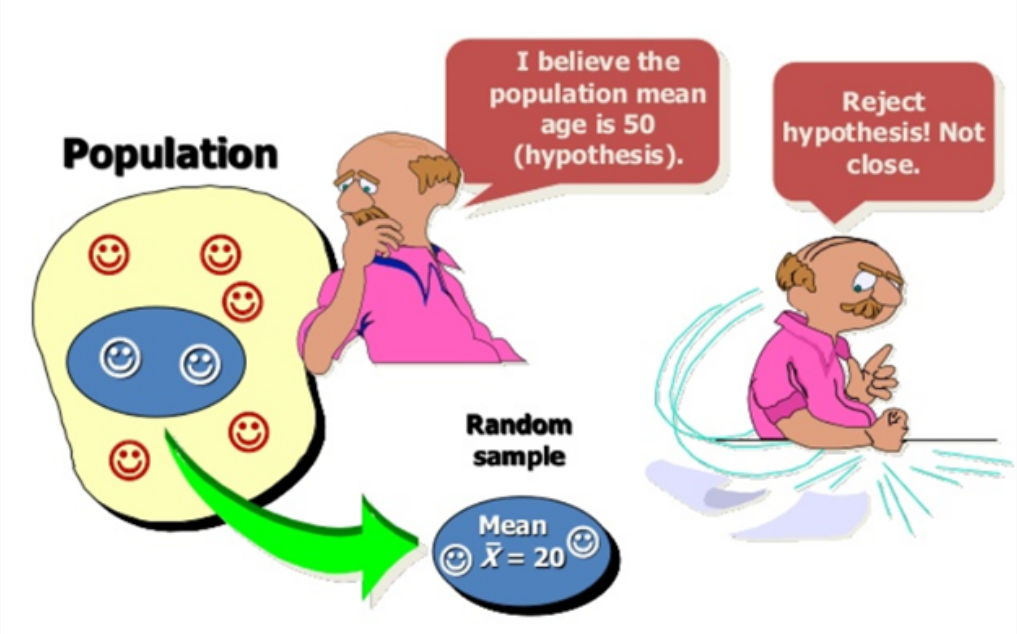
Measure of center



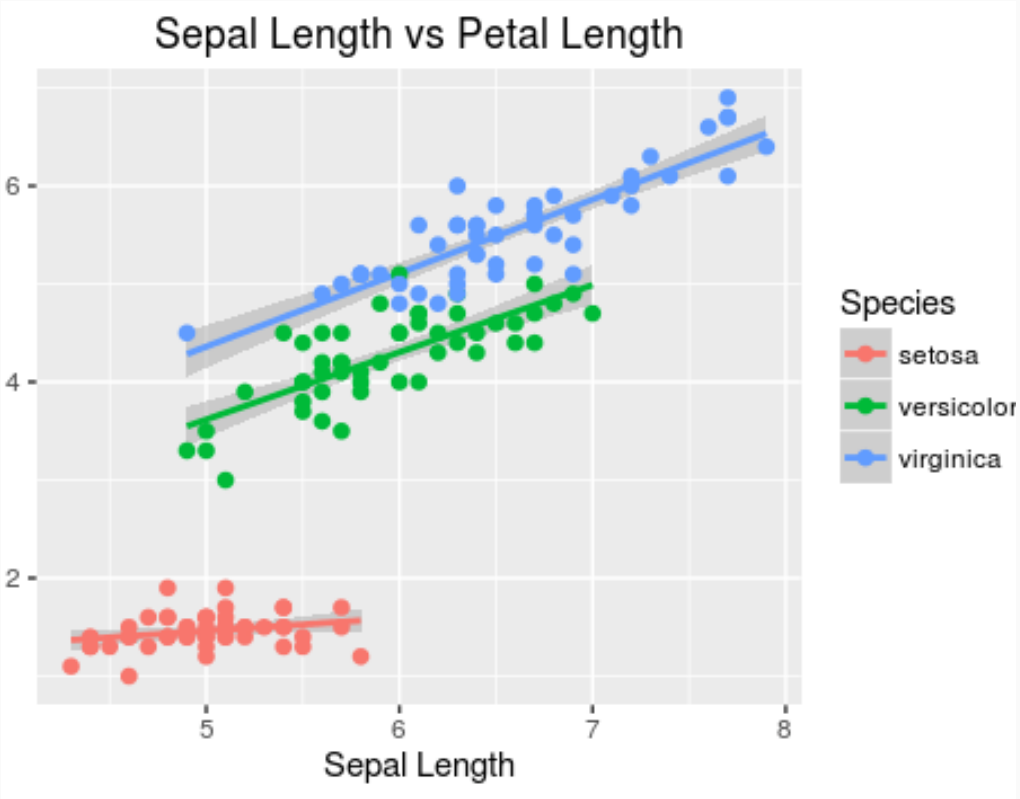
Measure of dispersion



Hypothesis Testing



Linear Regression



Topic 1 : Terminology

Data and Statistics

- **Statistics** is science of collecting, analyzing, and interpreting of **data**
- The goal of **Statistics** is to help researchers organize and interpret the **data**
- **Statistics** can turn raw **data** into useful knowledge
- **Statistical** graphs, and plots visualize, and summarize the **data**

What is Data ?

- Data is information
- Typically data takes the form of
  1. Observed measurements
     (e.g. height, temperature, GPA, annual income)
  2. Descriptions
     (e.g. marital status, gender, ethnicities)
- A lot of data is obtained from surveys and experiments

Examples of Data

- Height of freshmen in FSU
  : $5.6, 5.3, 5.8, 6.1, 5.4, 6.3, \cdots$
- Which pet do you have or would like to have ?
  : Dog, Cat, rabbit, Dog, Not answer, Ferret, $\cdots$
- Student eye colors in STAT1013
  : Brown, Brown, Brown, Blue, Brown, Hazel, $\cdots$
- GPA of students in Dept. Statistics.
  : $3.5, 3.6, 2.8, 3.8, 3.1, 2.6, \cdots$

Terminologies

Observation, Data Point : A single collected data value (a piece of information)

Variable : A variable, is a characteristic or property of an individual or unit whose value may change from one observation to another.

Population

A **population** is the entire overall group we are interested in.

Examples :

- Want to know average height of all U.S adult female
- Want to know average GPA of freshmen in FSU
- Want to know average lifetime of light bulbs produced by GE

Population

It is often **time-consuming** or **costly** to obtain data from the entire population.

- How many trees are there in Rocky Mountain National Park ?
- Want to know average size of fishes in Lake Ella ?
- Average blood sugar level of diabedic patients in U.S

No one knows!! But maybe we can gather some data and get an approximate idea.

Sample

We rarely have access to the entire population. Instead we rely on a subset of the population to use as a proxy for the population.

Sample is a subset of the entire population that we collect data on.

Sample size: The number of subjects involved in a study, i.e., the number of data points in the sample.

Examples of Sample

- Want to know average height of all **U.S adult female**

  → Measure height of 3,000 U.S adult female

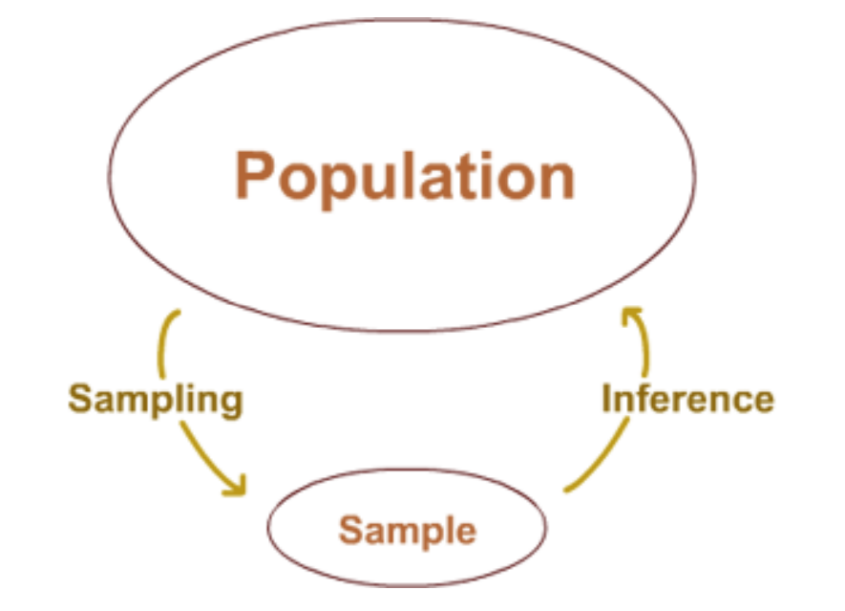
- Want to know average GPA of **university students in U.S**

  → Collect GPAs of 50 freshmen in Dirac library

- Average lifetime of **light bulbs produced by GE**

  → Examine the lifetimes of 1,000,000 GE light bulbs

Population vs Sample



In real life, we would obtain data from a sample instead of a population, and then use our knowledge of statistics to make **inferences** about the population from which the sample was drawn.

Parameter vs Statistic

Parameter : A numerical measurement that describes a characteristic of a population.
Most often parameters cannot be determined, because we do not have census data.

Statistic : A numerical measurement that describes a characteristic of a sample. Very often sample statistics are used as estimates for the corresponding population parameters.

Statistical Inference : The process of using sample statistics to draw conclusions about population parameters is known as statistical inference

Parameter vs Statistic

	Population Parameter	**Sample Statistic**
Size	N	n
Mean (average)	μ	$\bar{X}$
Variance	σ^2	s^2
Standard deviation	σ	s
Proportion / Percent	π	p
Correlation	ρ	r

Exercise

Identify Population parameter and Sample Statistic

1. We want to know what percent of this term's Statistics students are taking French. We randomly select 40 students in our Statistics class and ask if they are taking French, answer YES or NO. It turns out that 15 of them say YES. so we calculate that 15/40=37.5% of them are taking French. FSU administrators want a truer percent. so they poll all of this term's Statistics students and find that 29.8% of them are taking French.

Exercise

2. You want to know the mean age (average age) of all the persons living in your household. so you do the calculation and the mean turns out to be 36.4 years. Your teacher wants to estimate the mean age of all the persons living in your household so she asks you your and your father's age (20 and 50), so she says that her estimate is 35 years.

Exercise

3. You want to know the average lifespan of a cat. so you ask your veterinarian. She says based on the FDA research with 5,000 cats in Florida. It is 13 years.

Major concerns in statistical inference

We have to be careful when generalizing from a sample to a population. The descriptive characteristics of your sample may not accurately reflect the characteristics of the population you are studying.

Two important questions in statistical inference are:

- Is my sample big enough?
- Is my sample representative of the population of interest?

Small sample size

If our sample is small, we can't confidently draw inference to the population.

Example :

- Want to know average height for men in U.S
- We measure the height of 4 men in our class.

Can we get a good estimate for the ture average height of all male in U.S ? No

Smaller samples mean greater uncertainty when drawing inference to a population.

Sample is representative of the population of interest

Example :

- Want to know average height for men in U.S
- If our sample consists entirely of the FSU basketball team.

Can we get a good estimate for the ture average height of all male in U.S ? No

We will probably end up considerably overestimating the true average height of all US men