STAT 201 Chapter 1

Introduction to Statistics

Excuses

- **I’m sleepy …**

  - Drink coffee / tea

- **I don’t like math.**

  - That’s fine. This is a statistics course. We focus more on why and how to use some methods to tell a beautiful story of data instead of giving you formula and numbers to plug into.

- **I’m shy.**

  - You’re going to have to talk to people in your entire life, use this as an opportunity to break out of your shell.

Before We Get to Statistics …

- You are all dumb.
- I am dumb.

- We are all going to school to learn and become less dumb.
- We should **NOT** be embarrassed to not understand something at first – it is a sign of intelligence and hard work to ask questions.

Have you ever learnt statistics?

- A. learnt, and still remember some
- B. learnt, but have forgotten everything
- C. Heard but never learnt
- D. What is statistics?

Asking Happiness

- There are around 100 million people in Japan and around 300 million people in US. A survey in Japan randomly chooses 1000 people to their level of happiness. Suppose we want to do the same survey in US, and we want our survey to be **as precise as** the one in Japan. How many people should we choose?
- A. 1000
- B. 2000
- C. 3000
- D. >3000

Drinking Soup

- There is a **small cup** of red soup
- There is a **big pot** of green soup

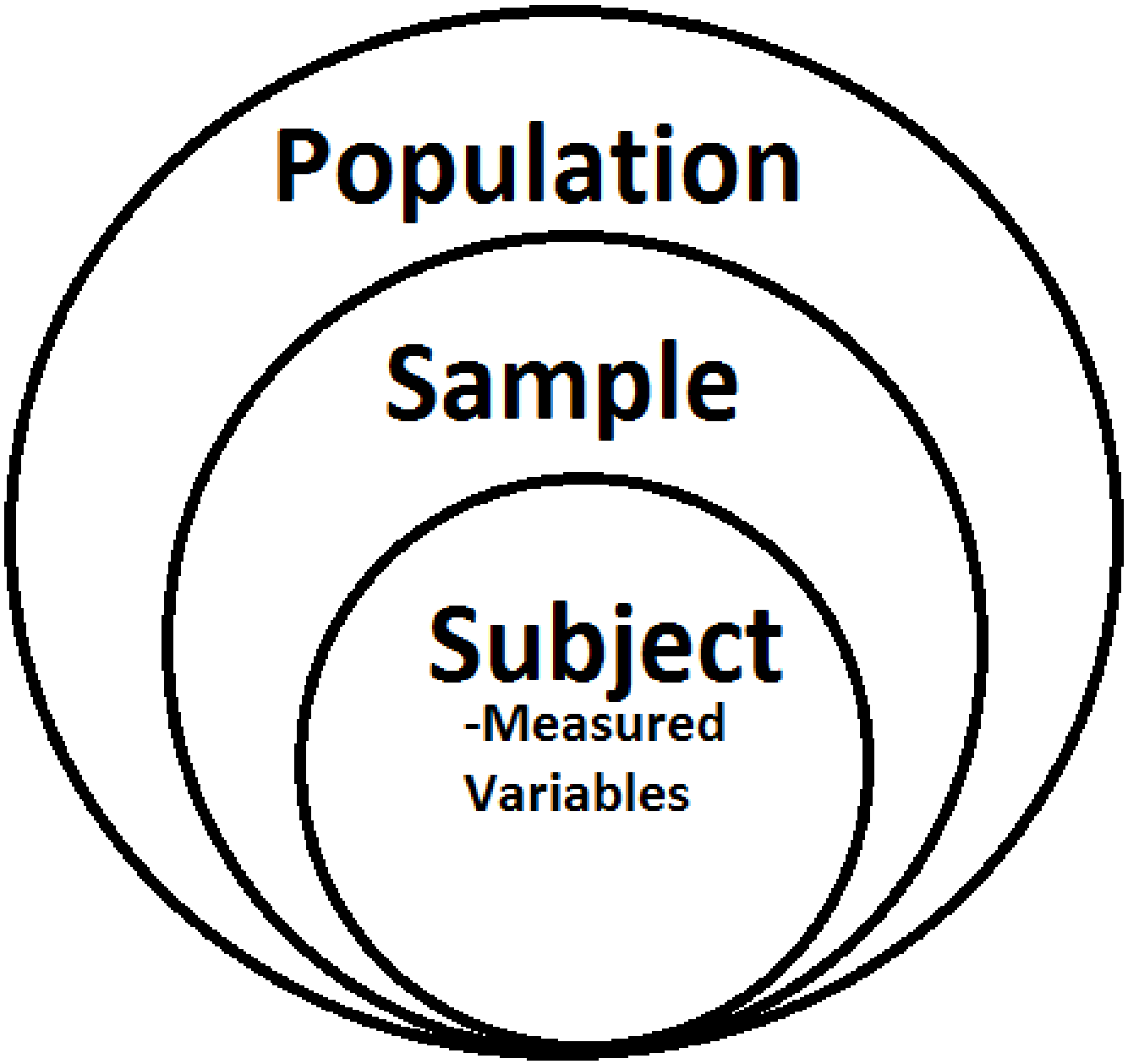
- You want to know which one tastes better. How much soup do you drink?

Statistics: dig out the truth from chaos

- Statistics measures uncertainty in real life.
- Statistics helps us make right decision!

Definition

- **Subject**: entities that we measure in a study
  - soup, people
- **Population**: the total set of subjects in which we are interested in
  - cups/pots of soup, US population
- **Sample**: the subset of the population for whom we have data, often randomly selected
  - a mouthful of soup, 1000 people in US
- **Variable**: any characteristic that is observed for the subject
  - delicious, happiness level, whatever we're measuring


Population
Sample
Subject
-Measured
Variables

Definitions

- **Statistic**: numerical summary of a sample (we know)

  - Mean(Average), median, etc.

- **Parameter**: numerical summary of a population (we don’t know)

  - Mean, median, variance, etc.

How to memorize?

- Statistic starts with an ‘s’, so it’s talking about the sample.
- Parameter starts with a ‘p’, so it’s talking about the population.

Example

- Old McDonald's farm has 5000 turkeys and we're interested in estimating the average weight of all the turkeys. Instead of weighing all 5000, we only weigh 100 randomly selected turkeys.

- Here a turkey is the subject, all the 5000 turkeys in old McDonald's farm make up the population, 100 selected turkeys make up the sample.
- What is the variable? Weight!

Example

- Last semester there were 514 STAT201 students. We wanted to approximate the average height of a STAT201 student.
- So we looked at 40 students and measured their height. It showed that the average height of the 40 students was 165 cm.
- After that, we found that the mandatory physicals record of all students, in which the average height of all 514 STAT201 students was 172 cm.
- What is Subject, Population, Sample, Variable, Statistic, Parameter?

Sample v.s. Population

- Subject: STAT201 student
- Population: all 514 STAT 201 students last semester
- Sample: the 40 students we selected and measured
- Variable: Height
- Statistic: sample mean = 165 cm
- Parameter: population mean = 172 cm

Major Components to Statistics

- **Design of Study**

  - What question are we answering?

- **Descriptive Statistics**

  - What summary can help us answer the question?

- **Inferential Statistics (or Statistical Inference)**

  - Can we predict or draw conclusions based on the data we have?

Design of the Study

- What is the research question?
- What is the population of interest?
- What is the variable of interest?
- How will the sample be selected? (Important)
- How will the data be collected?

- Essentially, what's the best way of going about using statistics to solve your problem?

Design of the Study: Sample Selection

- **Census**: collect data for every individual subject in the population

- The word “census” originated in ancient Rome from the Latin word *censere* (“to estimate”). The crucial role of census in the Roman Empire is to determine taxes. It was carried out every five years.
- Required by the US constitution, United States Census Bureau should hold census for every ten years.
- Problem: time, money, labor

Design of the Study: Sample Selection

- **Judgment**: collect a sample that an expert thinks is representative
  - Problem: there may be some **bias**
- **Convenience**: collect the sample that is easiest to access
  - Problem: bias, bias, lots of bias!

❑Note **Bias** is when the results of a sample are not representative of a population.

Design of the Study: Sample Selection

- **Volunteer**: Subjects choose to participate
  - Problem: sample will still be biased
  - Example: Medical experiments to test medication
- **Systematic**: Use a method to select
  - Problem: there may be a system we don't know
  - Example: check every fifth item produced (maybe every fifth item was made by the same machine)

Design of the Study: Sample Selection

- **Simple Random Sample (SRS)**: the sample is chosen in such a way that every subject is equally likely to be selected for the study
  - We prefer this method above all else
  - Problem: Sometimes this isn't feasible (e.g. mosquitoes)

Descriptive Statistics

- Later, we will explore the ways of describing the sample that has been collected through
  - Numerical summaries: Mean (average), median, mode, etc.
  - Graphical displays: Charts, graphs, etc.

Inferential Statistics (Statistical Inference)

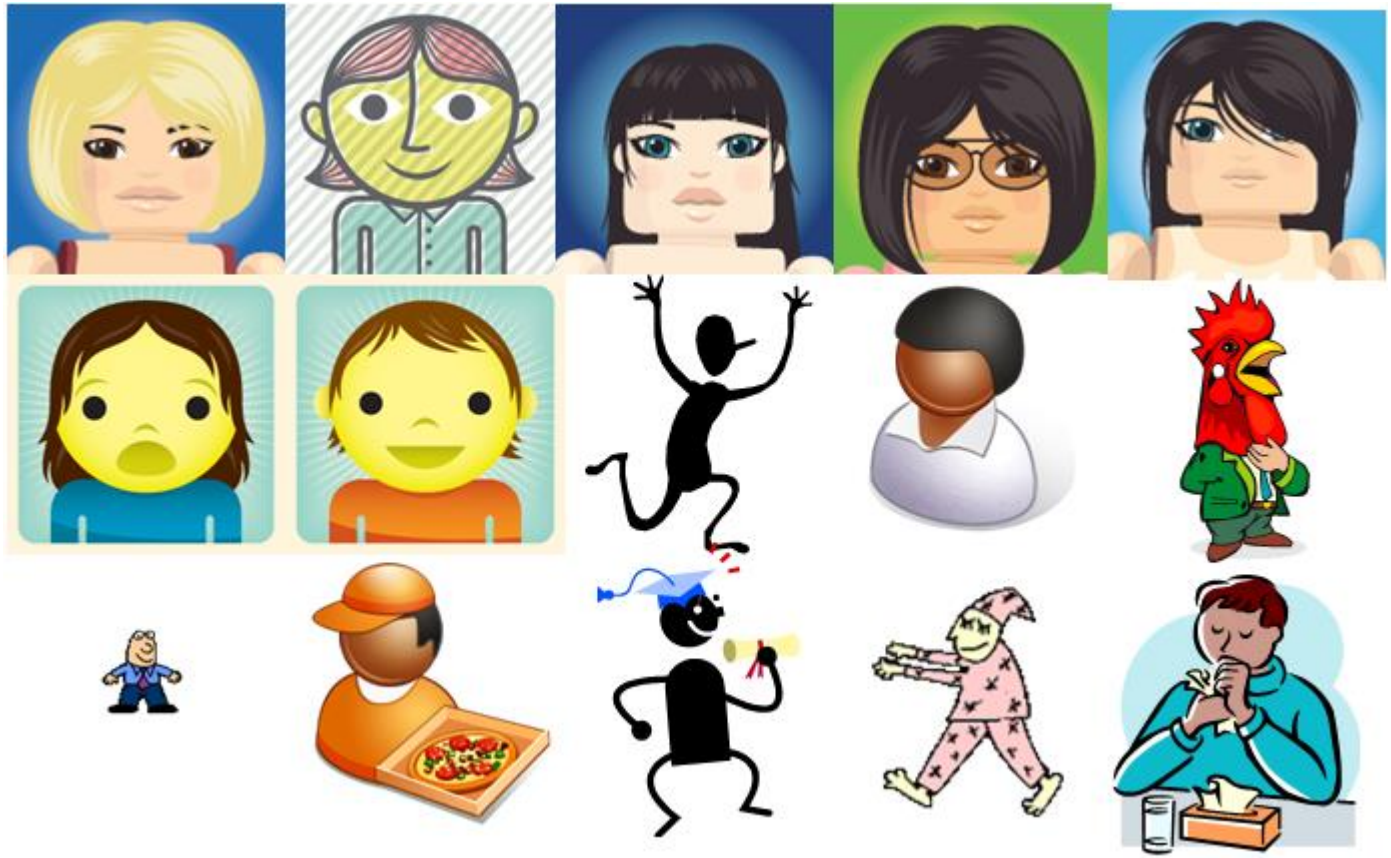
- Making predictions or drawing conclusions about a population of data from a sample

- This will be covered at the end of the semester, but involves everything we learn up to that point.

Example

- Let’s say there are fifteen clipart people in the world. We need to know, on average, how old a clipart person is.

Example: Population of 15 Clipart People



Example

- Interviewing a **population** of fifteen clipart people is too much work!
- Instead, we want to take a **sample** of the population and only interview them – that will be easier

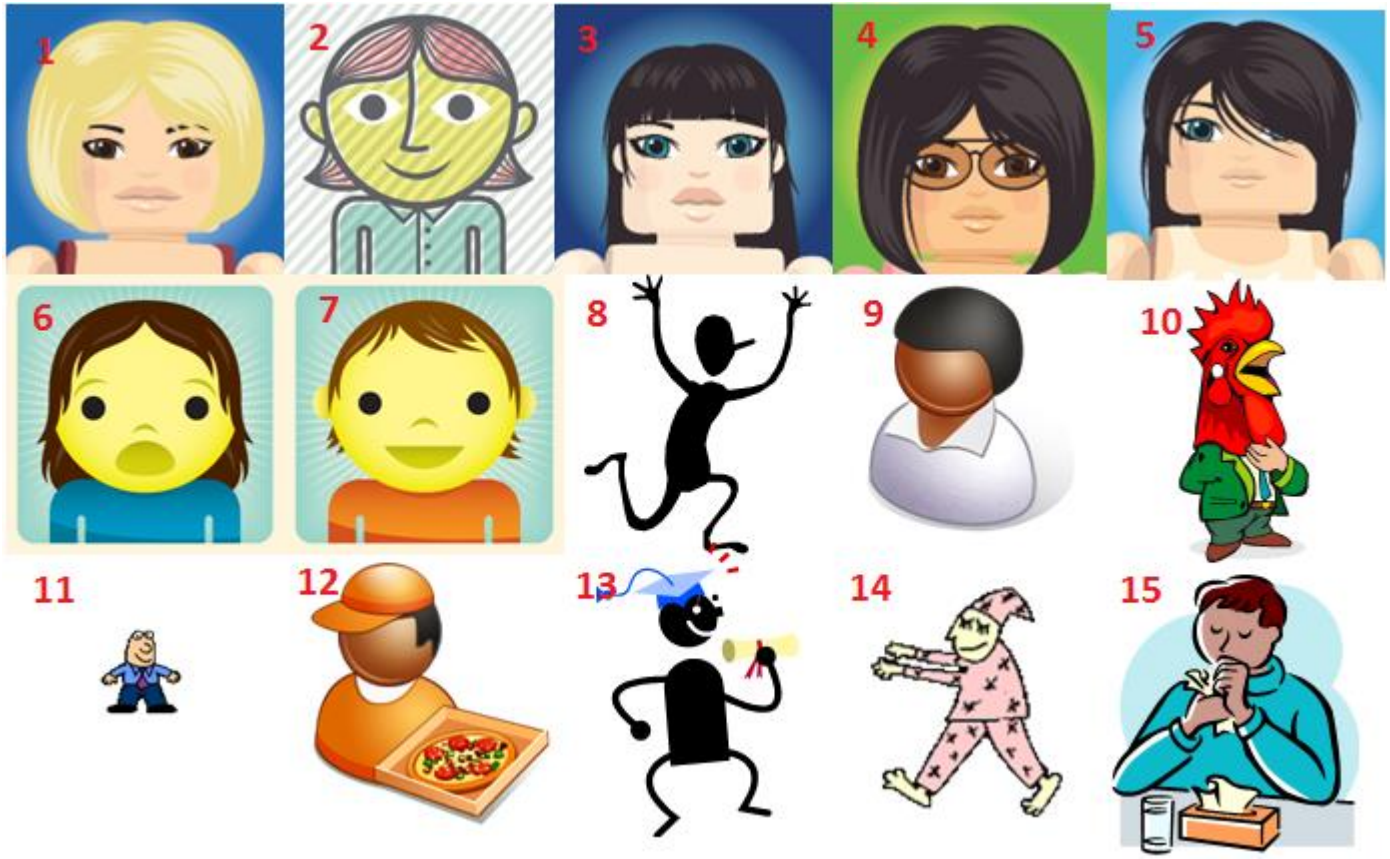
- **Mean Idea:** We can then look at the **statistics** of the **sample** and use that to make **inference** about the **population**.

Example: Sampling Method

- Let’s use a simple random sample (most favored)

  1) Number all the clipart people (population)

Example: Sampling Method



Example: Sampling Method

- Let's use a simple random sample (most favored)

  1) Number all the clipart people (population)

  2) Choose a sample size (let's make it 5)

  3) Get 5 random numbers from 1 – 15

  - write 1 – 15 on fifteen small pieces of papers

  - put them into a hat

  - randomly select 5

Example: Sampling Method

5

3

12

8

14

Example: Sample

Example: Descriptive Statistics

- Mean age of our sample:

  (25+42+17+27+62)/5=34.6 years old

Example: Inferential Statistics

- Can we say that the average age of all 15 clipart people is 34.6 years old?
- Can we predict the age of the next randomly chosen clipart person?

- We will answer these rousing questions when we get to Chapters 8 and 9

Example

- Subject: Clipart people
- Population: All fifteen clipart people
- Variable: age
- Type of Sampling: Simple Random Sample (SRS)
- Statistics: Mean of sample (34.6)
- Parameter: Mean of the population (unknown)