

# Introduction to Statistics

## Math 130: Introductory Statistics

### 1

#### 1.1 Goals

##### Goals of this class meeting

- Define statistics.
- Understand different methods of describing data.
- Understand different measures of center.

#### 1.2 Definitions

##### Some necessary definitions

- Statistics: the study of how to use data to answer interesting questions.
- Population: the complete collection of all elements to be studied.
  - Example: Suppose someone asks, “How do people with cancer respond to a certain type of drug”.  
Population = all people in the world with cancer.
- Census: collection of data that includes *every* member of the population.
- Sample: a collection of data from a subset of members from a population.

#### 1.3 Fish

##### How many fish in a lake?

- How do you answer this question?
- Is it feasible to take a census? Collect every fish in the lake?
- Is it possible to answer this question with a sample?

### How many fish in a lake?

- Catch a bunch of fish (say 50), tag them, and toss them back.
- Come back later, catch a bunch of fish (say 100).
- Suppose of the 100 new fish, 10 of them have tags on them.
- One would conjecture then, that 10% of the total number of fish are tagged.

$$0.1 \times (\text{total number of fish}) = 50$$

$$\text{total number of fish} = \frac{50}{0.1} = 500$$

## 2 Introduction

### 2.1 Types of data

#### Parameter and statistic

- Can one conclude there are *exactly* 500 fish?
- A **parameter** is a (probably unknown) characteristic of a population.
- A **statistic** is an estimate of a parameter obtained from a sample.

#### Other types of data

- Quantitative data: numerical data representing counts or measurements.
- Qualitative data: characteristics that can be put into categories.
- Discrete data: number of possible outcomes is finite or “countable”.
- Continuous data: infinitely many possible values that correspond to some scale that contains no gaps, interruptions, or jumps.

#### And more types of data...

- Nominal data: consists of categories that cannot be ordered in a meaningful way.
- Ordinal data: order is meaningful, but not the distances between data values.
  - Excellent.
  - Very good.
  - Good.
  - Poor.

- Very poor.
- Interval data: order is meaningful, *and* distances are meaningful. However, there is *no natural zero*.
  - Examples: temperature, time.
- Ratio data: order, differences, and zero are all meaningful.
  - Examples: weight, prices, speed.

## 2.2 Thinking critically

### Thinking critically

- Loaded questions.
- Order of questions.
- Self selection bias.
- Correlation and causality.
- Confounding.
- Always think critically about whether the statistical analysis actually answers the question posed.

## 2.3 Experimental design

### Types of studies

- Observational study: data is observed and generated outside a lab.
  - Examples: economic and financial data.
- Experimental study: elements in the sample are controlled, researcher applies a treatment then proceeds to study the effects.
- Observational or experimental studies?
  - According to a survey, 65% of smokers who used a nicotine patch quit for at least 6 months.
  - Half of a group of people with Asthma were given a new medication, the other half were given a placebo.

### **Natural experiments**

- When data is observed, but a controlled experiment luckily happened “naturally”.
- Does small class size lead to better performance.
- Self selection problem.
- Maimonides’ rule: In Israeli public schools, classes are broken into two when enrollment exceeds 40.

### **Maimonides**



- Moses Maimonides (1135 - 1204) Jewish Rabbi, Physician, and Philosopher.
- His life’s work a big impact on Jewish thought and study.
- Believed small class sizes were important for effective education.

### **Types of studies: time**

- Cross-sectional studies: data are collected from one point in time.
- Time series studies: data is collected from one individual from many points in time.
- Panel study: data is collected from many individuals over many points in time.
- Retrospective study: data is collected from the past by examining records, interviews, etc.
- Prospective study: data begins to be collected and continues into the future.

## 3 Describing a sample

### 3.1 Frequency distributions

#### Frequency distributions

- **Frequency distribution:** lists data values (or groups of intervals) and the number of times those values (or groups) appear in the sample.
- Components of a frequency distribution:
  - Lower class limits: smallest numbers that can belong to a particular class.
  - Upper class limits: largest numbers that can belong to a particular class.
  - Class boundaries: halfway point between the upper class limit of one class, and the lower class limit of another class.
  - Class midpoints: the midpoints of the classes = average of the lower and upper class limits.
  - Class width = difference between two consecutive class boundaries.

#### Relative frequency distribution

- **Relative frequency distribution:** lists data values (or groups of intervals) and the *percentage* of times those values (or groups) appear in the sample.

$$\text{relative frequency} = \frac{\text{class frequency}}{\text{sum of all frequencies}}$$

- **Cumulative frequency distribution:** lists classes of data and the number of times that class *or below* appears in the sample.
  - Just give a running sum of the frequencies.

### 3.2 Histograms

#### Histograms

- Histogram: a bar graph that conveys the same information as a frequency distribution.
  - Vertical axis: frequencies.
  - Horizontal axis: classes of data.
- Relative frequency histogram
- Frequency polygon: line connecting the top of the bars.
- **O-give:** dumbest word I've ever heard, I would have called it a cumulative frequency polygon.

### Example with MS Excel

- Let's examine Problem 2.26 on page 54.
- Life span of light bulbs. There are 40 light bulbs from Manufacturer A, and 40 light bulbs from Manufacturer B.
- Construct a frequency distribution for Manufacturer A with the following lower class limits: 650, 750, 850, 950, 1050, 1150.
- In Excel create a column of "bins". Excel considers bins the largest number allowed in the class.
- Go to Data menu → Data Analysis → Histogram.
- Histograms are not supposed to have gaps.
  - Right-click on one of the bars, select **Format Data Series**, and reduce gap width to zero.

## 4 Measures of center

### 4.1 Mean, median, and mode

#### Measures of center

- Measures of center: some measure of the middle of the dataset.
- Arithmetic mean: Add up all values of a dataset and divide by the total number of values.

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

- Median: arrange data in numerical order, and pick the middle number.
  - If there are an even number of observations, take the average of the middle two.
- Mode: value in the dataset that occurs most frequently.
  - Bimodal distributions: when two values occur with the same greatest frequency.
  - Multi-modal distributions: when more than two values occur with the same greatest frequency.
  - When no value is repeated: no mode.

## 4.2 Other measures of center

### Other measures of center

- Mid-range: average between the largest and smallest values.
- Weighted mean: when different values are given different degrees of importance.

$$\text{weighted mean: } \bar{x} = \sum_{i=1}^n w_i x_i$$

- where the all the weights add to 1:  $\sum_{i=1}^n w_i = 1$ .
- With an arithmetic mean all weights are the same,  $w_i = 1/n$ .

## 4.3 Outliers and skewness

### Which measure is best?

- Outliers: numbers that appear to be very different from the rest of the sample.
  - Which measures of center are most sensitive to outliers?
- Skewness: when the distribution of values are not symmetric.
  - Skewed to the left: when data has a longer left tail, mean and median are to the left of the mode.
  - Skewed to the right: when data has a longer right tail, mean and median are to the right of the mode.
- Ordinal data?

## 5

### Next time...

- Homework:
  - Histograms: Section 2-4: problems 11, 12, 13.
  - Means, medians: Section 3-2: problems 5 through 12.
- Learn about measures of variation.
- Use what we know about center and variation to analyze measures of relative standing.