LECTURE 1: TYPES OF DATA

AP STATS

> There are two goals when presenting data: convey your story and establish credibility.
>
> Edward Tufte

Statistics is the mathematical science concerned with collection, analysis, interpretation, and presentation of data. We will need to familiarize ourselves with how data is presented and how to interpret data. When presenting data, you need to *be very specific about the context* in which you're getting your data from. The language of statistics is *probability*, which we will study later in greater detail. The language of probability, however is that of *set theory*.

1. Intro to Sets

In math, we view everything as `objects` and mathematics is about *how* these `objects` interact with each other.

A **set** is a collection of `objects`. A familiar example of sets are Number sets:

$\mathbb{N} = \{0, 1, 2, 3, \ldots\}$	**N**atural numbers (positive whole numbers and 0)
$\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, 4 \ldots\}$	**Z**ahlen = integers (non-decimals)
$\mathbb{Q} = \{\text{fractions}\} = \left\{\frac{p}{q} : p, q \in \mathbb{Z}, q \neq 0\right\}$	**Q**uotient (all fractions and all decimals that repeat)
$\mathbb{R} = (-\infty, \infty)$	**R**eal numbers (all numbers that are on the number line)

The notation $\in$ is "in". If A is a set, we say that $x \in A$ if x is an `object` that is in A but we use the notation $x \notin A$ if x cannot be found in A.

Example 1.

(a) $-3 \in \mathbb{Z}$ but $-3 \notin \mathbb{N}$.
(b) 10th grade $\in$ high school
(c) $\infty \notin \mathbb{R}$ (this says infinity is not in the Real Numbers set)

A *set* is a collection of objects. The objects in that collection are called *members* or *elements* of that set. If A is a set, we write $m \in A$ if m is an element of A and $m \notin A$ if m is not in A.

To describe all of the elements in a set, make sure to open and close with curly braces $\{set\ members\}$.

Example 2.

(a) $A = \{1, 2, 3\}$ is a set containing three elements, namely $1, 2$ and 3.

(b) Note that you should only list distinct elements in a set. For example

$$\{1, 2, 3, 2\} = \{1, 2, 3\}$$

since both sets are made up of three members 1, 2 and 3.

(c) Sets do not necessarily have to be numbers, e.g., $\left\{☕, 😄, 2\right\}$ is a set whose members are ☕, 😄, and the number 2.

(d) $B = \big\{1, 2, \{1, 2\}\big\}$ is the set consisting of precisely of numbers 1 and 2 and of the set $\{1, 2\}$. This set has exactly three distinct elements, it is not the same as $\{1, 2\}$, and it is not the same as $\big\{\{1, 2\}\big\}$.[1]

The Empty Set. The *empty set* is a unique set and is *the* set with no elements. We denote it by $\emptyset$ and we write $\emptyset = \{\}$. For any object x, we have $x \notin \emptyset$; e.g. $\emptyset \notin \emptyset$.

The only thing that matters to a set is its (distinct) members; the order in which the members are listed or if there are repetitions in the set do not matter.[2]

Since we ignore these differences, we have

$$\{a, c, b, a, b, d, a, b, b, d, c\} = \{a, b, c, d\} .$$

Sets defined by Proposition. Let A be a set and let $P(x)$ be some proposition (a statement that can be either true or false) whose truth value depends on elements $x \in A$. We use the following notation to define such sets:

$$\big\{x \in A : P(x)\big\} \quad \text{or} \quad \big\{x \in A \mid P(x)\big\} .$$

Both notations read "the set of elements x in A such that $P(x)$ is true". The : and the | in set notation reads "such that".

Remark. There is no such thing as a set that will contain *everything* (there is no such thing as an absolutely universal set).

If A and B are sets, the *intersection* is the set

$$A \cap B \stackrel{\text{def}}{=} \{x : x \in A \ \text{ and } x \in B\}$$

of elements that can be found in both A **and** B.

---

[1]View the set member $\{1, 2\}$ as its own object in the set B, i.e., we have $\{1, 2\} \in B$.

[2]We can define *ordered sets* and *multisets* to track these differences respectively, but we won't be concerned with these in this course.

The *union* of A and B is the set

$$A \cup B \stackrel{\text{def}}{=} \{x : x \in A \text{ or } x \in B\}$$

and it is the set of elements that is in A or is in B (or both![3]).

Finally, the *complement* of B in A is

$$A \setminus B \stackrel{\text{def}}{=} \{x : x \in A \text{ and } x \notin B\}$$

which is all of the members who are exclusively in A and not in B.

2. Constructing New Sets from Old

Let A and B be sets.

Recall that we can construct new sets from A and B, namely

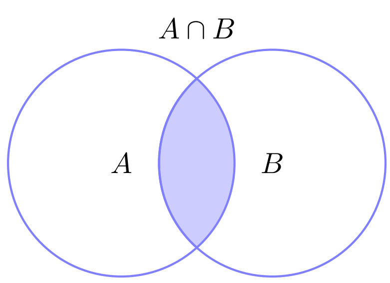
$A \cup B = \{x : x \in A \;\text{ or } x \in B\}$	"A **union** B"
$A \cap B = \{x : x \in A \;\text{ and } x \in B\}$	"A **intersect** B"
$A \setminus B = \{x \in A : x \notin B\}$	"A **complement** B"

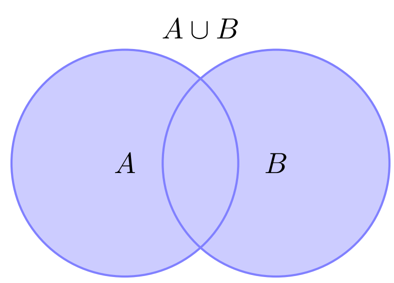
That is, the **union** is all objects that are contained in either set A or set B (or both!). Remember that the : and the | in sets are synonymous

The **intersection** is all objects that are in both A and B.

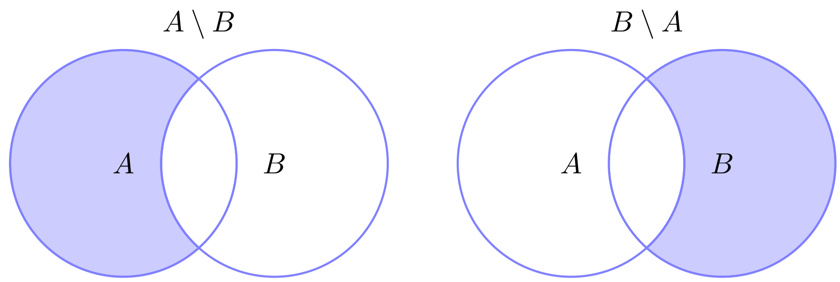
The **complement** is all objects that are in A but strictly not in B.

Pictorally, we have





[3]we use inclusive "or" in math (unless specified otherwise). That is, when your parents ask you "do you want ice cream or cake?", you saying "both!" is a valid response; this is the type of "or" we use in math. The other type of "or" is *exclusive*, as in "is it Monday or Tuesday?" You saying "both!" is not a valid response; we don't use this "or" in math- be careful with your language!



3. First Definitions

Definition. A ***random variable*** is a variable that is random.

More concretely, a random variable is a function $X : \Omega \to \mathbb{R}$ where Ω is the sample space.

This *very helpful* definition is the basis of probability as it is a function going from a *sample space* (the set of all possible events) to real number line.

Although we will give concrete definitions to the terms *sample space* and *random variable*, the ideas are pretty intuitive:

Example 1. If we flip 2 coins (labeling them Coin 1 and Coin 2), then the *sample space* Ω is going to be

$$\Omega = \{HH, HT, TH, TT\}$$

Let X be the number of heads that appears. This X is the random variable as its inputs are the possible events that can happen and its outputs are real numbers:

$$\begin{gathered}
\text{if } \omega = HH \implies X(\omega = HH) = 2 \\
\text{if } \omega = HT \implies X(\omega = HT) = 1 \\
\text{if } \omega = TH \implies X(\omega = TH) = 1 \\
\text{if } \omega = TT \implies X(\omega = TT) = 0
\end{gathered}$$

A few stats-definitions: the data we record are individual observations of a *variable*. That is,

- **Individuals** are the objects described by the data (these are the events themselves)
- **Variables** are *characteristics* of an individual that we record (a set of special events)

We will work with two variable types we'll work with:

(1) **Categorical Variables**: describes a particular quality or characteristic and we can create *categories*.
(2) **Quantitative (Numerical) Variables**: associates the variable to a number type.
    - discrete variables (integer values)
    - continuous variables (continuous range of values)

4. Categorical Data

If we are dividing up the data by using *names* or *labels* then you are using a **categorical variable**.

> A **categorical variable** is a variable that describes a quality or characteristic. You divide the data into smaller **categories**. The information collected is called **categorical data**.

Some examples of categorical variables:

- ◇ X = methods of getting to school
  - ∘ Categories could include: car, bike, schoolbus, walking, etc.
- ◇ X = color of eyes
  - ∘ Categories could include: red, brown, blue, green, etc.
- ◇ X = gender
  - ∘ Categories could include: male, female, nonbinary

When dealing with categorical data, you will be expected to:

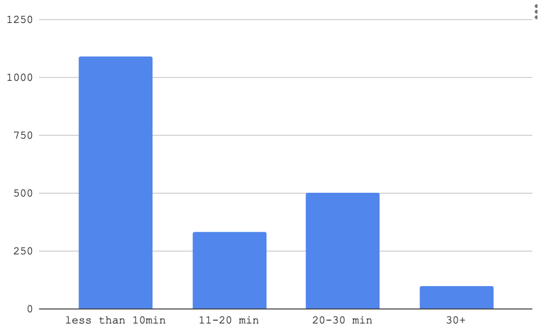
(1) create and interpret bar graph data
(2) provide *frequency* and *relative frequencies*

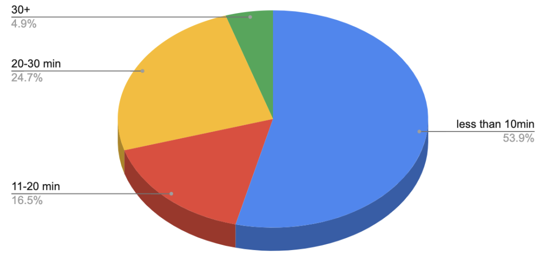
Since the data is split into smaller categories, the most efficient ways of communicating categorical data visually has been through the use

- ◇ Pie charts
- ◇ Bar graphs
- ◇ Segmented Bar graphs

Example 1. In a survey of 2023 families, 1090 of the families drive less than 10 minutes to drop their kids off at school, 333 families drive between 11-20 minutes, 500 families drive between 21-30 minutes, and 100 families drive 30+ minutes.

- Since we are assigning numerical labels to organize the data (as opposed to numerical values to *measure* the data), this is categorical data.
- It's easy to communicate the survey results using only a bar graph or pie chart.





We can use a **frequency table** to organize the table:

driving time	no. of families	relative frequency	percentage
≤ 10 min	1090	$1090/2023$	53.88%
$11-20$ min	333	$333/2023$	16.46%
$21-30$ min	500	$500/2023$	24.72%
$30+$ min	100	$100/2023$	4.94%

5. Quantitative Data

If we are assigning a numerical value to each individual object in the dataset, then you are using a **quantitative variable** (or numerical variable).

> A **quantitative (numerical) variable** is a variable that has a numerical value. The information collected is called **numerical data**.

Some examples of quantitative variables:

- ◇ X = number of people in a family
  - ○ Numerical values could be $1, 2, 3, \ldots$ (discrete variable)
- ◇ X = score out of 100 `True` or `False` on a test
  - ○ Numerical values could be any integer between $0-100$ (discrete variable)
- ◇ X = weight of newborns
  - ○ lots of babies so we typically use a range of values such as 0.5kg to 0.8kg (continuous variable)

As noted before, we break up quantitative variables into two more categories: (1) discrete and (2) continuous. We must always keep in mind what type of numerical variable we are working with since we might get different answers for the same question if we change the variable type!

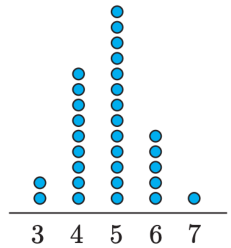
For discrete quantitative data we organize the data by using

- Σ Frequency tables
- Σ Dot plots
- Σ Stemplots

- **frequency table**

Number	*Tally*	*Freq.*
3	||	2
4	𝍸 ||||	9
5	𝍸 𝍸 |||	13
6	𝍸	5
7	|	1

- **dot plot**



- **stemplot**

Example:

Stem	*Leaf*
0	9
1	7 1
2	8 3 6 7 6 4
3	9 3 5 5 6 8 2 1
4	7 9 3 4 2
5	1

When there's a lot of data involved, you'll probably want to use a calculator or computer. However, we should also know how to do the basics without CAS.

Example 1. The data set below is the test scores (out of 100) for a Stats test for 50 students:

56	29	78	67	68	69	80	89	92	71
58	66	56	88	81	70	73	63	74	38
67	64	62	55	56	75	90	92	47	44
59	64	89	62	51	87	89	76	59	88
72	80	95	68	80	64	53	43	61	39

To organize this data, let's use a Stem and leaf plot:

$Stem$	$Leaf$
2	9
3	$8, 9$
4	$3, 4, 7$
5	$1, 5, 6, 6, 6, 8, 9, 9$
6	$1, 2, 2, 3, 4, 4, 4, 6, 7, 7, 8, 8, 9$
7	$0, 1, 2, 3, 4, 5, 6, 8$
8	$0, 0, 0, 1, 7, 8, 8, 9, 9$
9	$0, 2, 2, 5,$

From this, we get a decent picture of what the distribution of the numerical data looks like.

For continuous quantitative data, we organize the data using

$\int$ histograms
$\int$ cumulative frequency graphs

These require more in-depth discussions so we'll discuss them again.

Question 1. Consider the following variables that can be used to study the different *populations*. For each variable, classify the variable as *categorical*, or *discrete quantitative*, or *continuous quantitative*.

(a) Countries:
   (1) population size
   (2) time zone
   (3) average rainfall
   (4) life expectancy
   (5) mean income
   (6) literacy rate
   (7) capital city
   (8) largest river
(b) Peoples:
   (1) age
   (2) height
   (3) gender
   (4) ethnicity
   (5) income
   (6) literacy
   (7) marital status
   (8) high school GPA