MA/CS 109 Lecture 13
A “Toy” Example

Suppose you live in a town in New Hampshire with 5 voters. Suppose 3 of them are Republicans and 2 of them are Democrats.

A big city news paper reporter comes to town and wants to determine how many Republicans and how many Democrats are in the town, so they stand on a corner and ask 2 passing people who say they are voters what party they are in.
What information can they possibly get from this exercise?

Let’s make some simplifying assumptions—everyone in this town tells the truth and will answer any question asked.

Does the information the reporter receives from two people tell her anything for sure?

What are the possible situations?
Suppose the 5 people are A, B, C, D, E

Suppose A, B and C are Republicans and D, E are Democrats.

There are 5 ways to pick the first person and 4 ways to pick the second person, so 20 ways to pick 2 people, in order. Since we don’t care what order we choose the people, there are $20/2$ (since there are 2 ways to order 2 people) or 10 ways to pick a group of 2 people from 5 people.
## Possible outcomes

<table>
<thead>
<tr>
<th>People</th>
<th>Republicans</th>
<th>Democrats</th>
<th>Fraction Rep.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, B</td>
<td>2</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>A, C</td>
<td>2</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>A, D</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>A, E</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>B, C</td>
<td>2</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>B, D</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>B, E</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>C, D</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>C, E</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>D, E</td>
<td>0</td>
<td>2</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Of course, the big city news paper reporter will only see ONE of these pairs of people.

If she sees A,B or A,C or B,C then she only sees Republicans. She has no information that there are any Democrats in the town.

If she sees D,E then she sees only Democrats and has no information that there are any Republicans in the town.

If she sees any other pair of people, then she knows there are both Democrats and Republicans.
This isn’t much—however, she knows she is only asking 2 people in town, so she knows that whatever data she collects, there is still uncertainty. Can she quantify this uncertainty?

Let’s look at this problem “from the outside”—let’s look at all the ways the reporter’s survey of 2 people could come out.

This is a big a very clever change in point of view. Instead of thinking about which people she picks, think about which group of 2 people she picks from all the possible groups of 2.
## Possible outcomes

<table>
<thead>
<tr>
<th>People</th>
<th>Republicans</th>
<th>Democrats</th>
<th>Fraction Rep.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, B</td>
<td>2</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>A, C</td>
<td>2</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>A, D</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>A, E</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>B, C</td>
<td>2</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>B, D</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>B, E</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>C, D</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>C, E</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>D, E</td>
<td>0</td>
<td>2</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Each group of 2 people has a proportion of Republicans. One group has 0 Republicans, three groups have all Republicans and six groups are split evenly. So we observe the most common among the 2 person groups actually has the proportion of Republicans closest to the proportion of the whole town.
Of course, this is a not very useful to the reporter (yet) because she does not know this picture (if she did, she wouldn’t need to ask anybody in the first place)...but we have learned something.

Our first conclusion is:
If the reporter chooses a group of 2 people “at random”, then it is most likely she will choose a group that is a pretty good representation of the whole population.

But wait—what does “choose a group of 2 people at random” mean?? We need some vocabulary...
Vocabulary

To go on we need some more vocabulary to communicate efficiently.

The set of everyone of interest (in our case, voters in the New Hampshire town) is called the population.

A subset of that group (in our case, the people chosen by the reporter) is called a sample.
The number of individuals in a sample is called the **sample size**. (In our toy problem, the sample size is 2.)

**Notice the two uses of the word “sample”—this is not an accident.** The sample space in this toy problem is the set of samples of 2 people.

Sample space =\{AB, AC, AD, AE, BC, BD, BE, CD, CE, DE\}

When we say “choose a two person sample at random” we mean use the Equally likely probability model to describe the probability of a particular sample of 2 people being chosen.
More Vocabulary:

A number representing a quality of the entire population is called a parameter. (In our example, the fraction of the population that are Republicans is a parameter.)

A number representing a quality of a sample is called a statistic. (In our example, the fraction of the sample that are Republicans is a statistic.)
Back to our toy problem. We can now say that the probability that the reporter chooses a group of all Democrats is $1/10$, the probability that she chooses a group of all Republicans is $3/10$ and the probability of getting a group with proportion of Republicans 0.5 is $6/10 = 3/5$. It is by far most likely that she will choose a sample in this third group with statistic close to the population parameter.
This is reassuring, but still doesn’t help the reporter since she only knows information from the sample of two people that she picks.

Let’s look at a slightly bigger problem...
A Bigger Toy

Suppose the reporter goes to a bigger town—with 10 voters. Suppose this town has 6 Republicans and 4 Democrats.

The reporter repeats the experiment, but this time chooses a sample of 3 individuals.
Again, let’s step back and take a look at the big picture. We think of all the three person samples the reporter could pick, each possible sample is an outcome in our sample space of 3 person samples.

Again we look to see what the possible proportion of Republicans are in all the possible samples and which proportions are the most likely if the reporter chooses a three person sample at random (i.e., using the Equally likely outcomes model on the sample space).
How many three person samples are there? (How many outcomes in our sample space?)

There are $10 \times 9 \times 8$ ways to pick three people from 10 “in order”. There are $3 \times 2 \times 1 = 6$ ways to rearrange 3 people, so there are

$$\frac{(10 \times 9 \times 8)}{6} = 120$$

outcomes in our sample space (120 ways to pick and unordered group of 3 people from 10 people).

Suppose the people A, B, ... J, with A,B,C,D,E,F Republicans, G,H,I,J Democrats.
Not very useful in this form!...But as a histogram, we get very interesting information.

Number of 3 person samples

We see some familiar things...

First, most common statistic among the samples has fraction of Republicans closest to the actual parameter value for the population.
So choosing a three person sample at random (Equally likely outcomes) the reporter’s most likely choice is a sample with proportion of Republicans of 0.66.

There are still quite a few three person samples that give misleading information about the proportion of Republicans.
Second, the “shape” of the histogram is the same, sort of as as the previous example. High in the middle (near the actual proportion of Republicans) and low at the ends (far from the actual proportion of Republicans).
Before “going big”, looking at a huge town or the whole state, we need some more vocabulary—probably familiar—from descriptive statistics.

We all remember the three m’s...

Mean: The mean of a list of numbers $x_1,x_2,...,x_n$ is the average, $(x_1+x_2+...+x_n)/n$
Median: The median of a list of numbers $x_1, x_2, \ldots, x_n$ is the middle number. That is, if the numbers are in increasing order, so each $x_i$ is less than or equal to $x_{i+1}$, then the median is the middle number if $n$ is odd or the average of the two middle numbers if $n$ is even.

For example, if our list is 4, 2, 5, 5, 9, 3
Then we first put the numbers in order
2, 3, 4, 5, 5, 9, then, the median is $(4+5)/2=4.5.$
The mode is the most common number in a list. (we won’t use the mode…)

Mean and median are important because they measure the “middle” of your list (of your data). Differences between the mean and the median are particularly interesting because if the difference is large it indicates an asymmetry in the data.

The lists 0,0,0,0,0,0,0,0,0,100 has the same mean as 10,10,10,10,10,10,10,10,10,10,10,10, but very different median (if this were distribution of wealth, it would be a big deal…)
The **variance** and **standard deviation** are measures of the dispersion of a list of numbers —how “spread out” this numbers in the list are.

Given a list $x_1, x_2, \ldots, x_n$ with mean $M$, the variance is

$$\frac{(x_1-M)^2 + (x_2-M)^2 + \ldots + (x_n-M)^2)}{n}$$

(or the average of the square of the difference from the mean.)

Standard deviation = $\sqrt{\text{Variance}}$
The variance is a little mysterious... if we want to measure how far \( x_i \) is from the mean \( M \), then it makes sense to compute \( (x_i - M) \), but why square this?

Two reasons: One is that squaring makes this number positive. The three numbers 0,5,10 are more spread out than 4,5,6, but both lists have mean 5. If we just added \( (x_i-M) \) then the numbers above the mean would cancel the numbers below the mean...
Why square instead of using absolute value or 4\textsuperscript{th} power?...squaring is a lot easier to compute—so technical reasons involved in later computations.

For us, the mean measures the middle of our data and the variance and standard deviation measure the spread of our data around the mean.