We need to use probability & statistics because of Variability & Uncertainty.

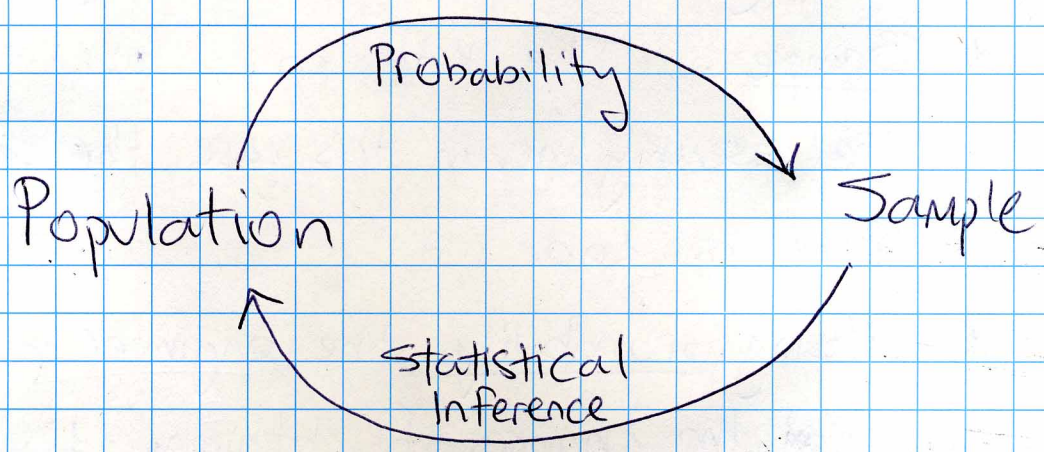
Ex1: A quality control engineer at an integrated circuit manufacturing plant takes a sample of 100 RAM chips from the assembly line and finds that 10 are defective. The company can tolerate 5% defective production in the long-run. The quality control engineer has to determine whether the long-run defective rate is within the tolerable range.

* Population: All possible RAM chips coming out of the manufacturing process.
* Sample: The RAM chips taken from the assembly line. In this case the sample size is 100.
* Using probability, the engineer computes that the chances of obtaining 10 out of 100 defective chips is 0.0167 if the long-term defect rate is 5%. Similary, he can compute that the chances of obtaining 10 or more defective chips is 0.0282.

* These small probabilities suggest that if the process indeed has a defective rate of 5% (or less) the particular sample collected by the engineer would rarely occur. <u>However it did occur!</u>

Therefore, the engineer determines that the process is very likely unacceptable. (or equivalently, the engineer says that he rejects the hypothesis that the process is acceptable at a certain confidence level)

This is called <u>statistical inference</u>.



Ex2: A neurologist wants to determine if a certain drug slows down the progress of Alzheimer's disease. It is known that Alzheimer's results in the abnormal enlargement of the ventricles (a compartment of the brain). The neurologist works together with an engineer who specializes in digital image processing to develop a computer program that automatically measures ventricular volume from Magnetic Resonance Images (MRI). The neurologist then recruits 20 Alzheimer's patients for his clinical trial. He then randomly assigns the 20 patients into 2 groups of 10; one group will take the drug for 6 months while the other group takes the placebo (no drug) for 6 months. At the beginning of the 6 months, all 20 patients have their ventricular volume measured with MRI. This is repeated at the end of the 6 months. For each patient, the volume at the end of the 6 months minus the volume at the beginning gives the change in ventricular size.

The question of whether the drug slows the progress of Alzheimer's can then be reformulated as whether the ventricular volume for patients using the drug grew less than the patients on placebo.

Dataset #1

changes in ml.	Drug	Placebo
	4.5	10.5
	3.5	9.6
	7.8	7.4
	-1.1	7.6
	5.8	10.6
	7.2	6.4
	6.7	11.6
	6.2	11.0
	4.6	6.8
	6.5	8.5
	+	+
	51.7	90

Sample mean $= \frac{51.7}{10} = 5.17$

Sample mean $= \frac{90}{10} = 9.0$

Sample mean is the center of mass.

Defn: Suppose that the observations in a sample are $x_1, x_2, \ldots, x_n$. The sample mean denoted by $\bar{x}$ is

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Dataset #2

Drug	Placebo
8.6	7.3
5.6	6.7
-0.4	12.3
6.7	18.2
5.9	6.0
11.9	4.2
10.2	7.8
1.0	13.3
-1.9	9.1
4.1	5.1

Sample mean = 5.17

Sample mean = 9.0

Defn: The sample variance, denoted by s^2, is given by

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

The sample standard deviation, denoted by s, is the positive square root of s^2.

---

* $n-1$ is the degrees of freedom associated with the variance
* Can not compute variance of sample with sample size 1 ($n=1$).
* Using the defn of sample mean, derive

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

So if we know $x_i - \bar{x}$ for the first $n-1$ observations, we can uniquely determine it for the n'th. Therefore, there are $n-1$ pieces of information that produce s^2 not n pieces.

* Variance is a measure of variability.

Sample Variances in our example (Units: ml^2)

	Drug	Placebo
Dataset 1	6.61	3.59
Dataset 2	20.55	18.99

Sample standard deviations (Units: ml)

	Drug	Placebo
Dataset 1	2.57	1.89
Dataset 2	4.53	4.36

* An alternative (faster) way to compute s^2

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} (x_i^2 + \bar{x}^2 - 2x_i\bar{x})$$

$$= \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} \bar{x}^2 - \sum_{i=1}^{n} 2x_i\bar{x} \right)$$

$$= \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2\bar{x} \sum_{i=1}^{n} x_i \right)$$

$$= \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2\bar{x} \cdot n\bar{x} \right)$$

$$s^2 = \frac{1}{n-1} \left[\left(\sum_{i=1}^{n} x_i^2 \right) - n\bar{x}^2 \right]$$

Avoids n subtractions

* Observational study vs. experimental design.

* Discrete vs. Continuous data.

* Random sampling
  Biased vs unbiased sample
  Sample size

* For a statistical problem, the sample along with inferential statistics allow us to draw conclusions about the ~~data~~ population.

* Problems in probability allow us to draw conclusions about characteristics of hypothetical data taken from the population based on known features of the population.