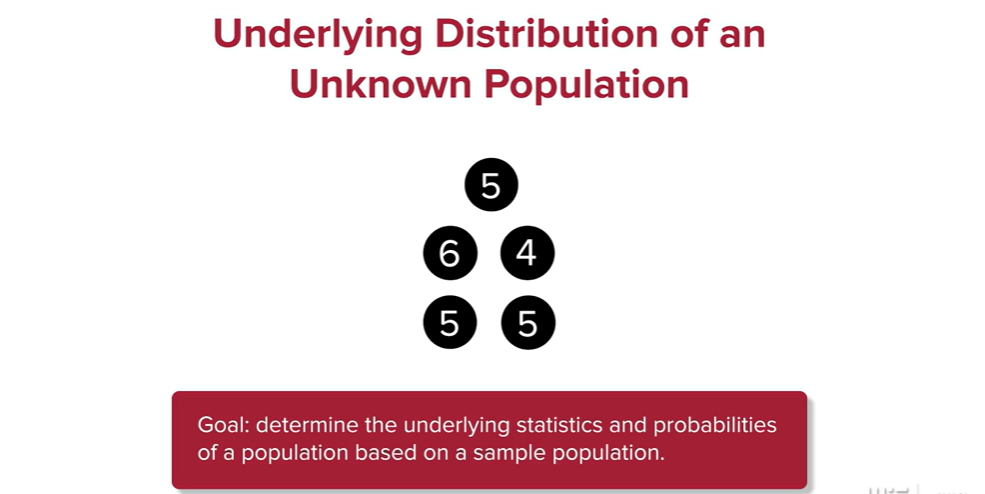
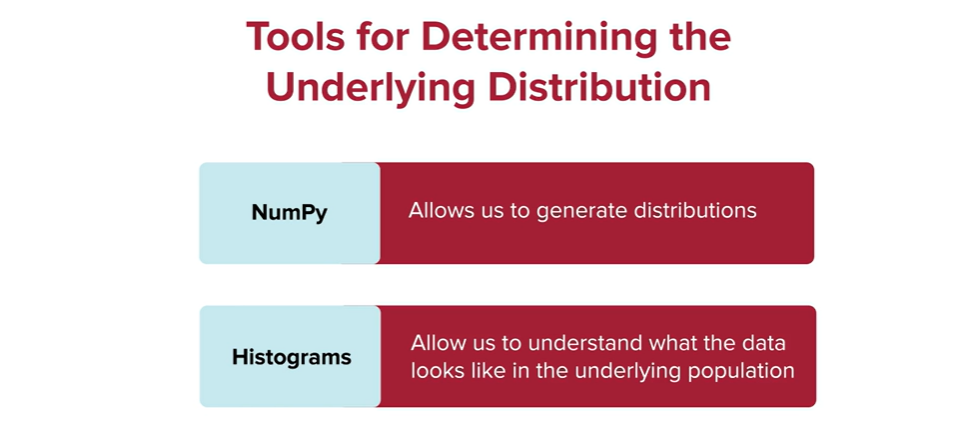
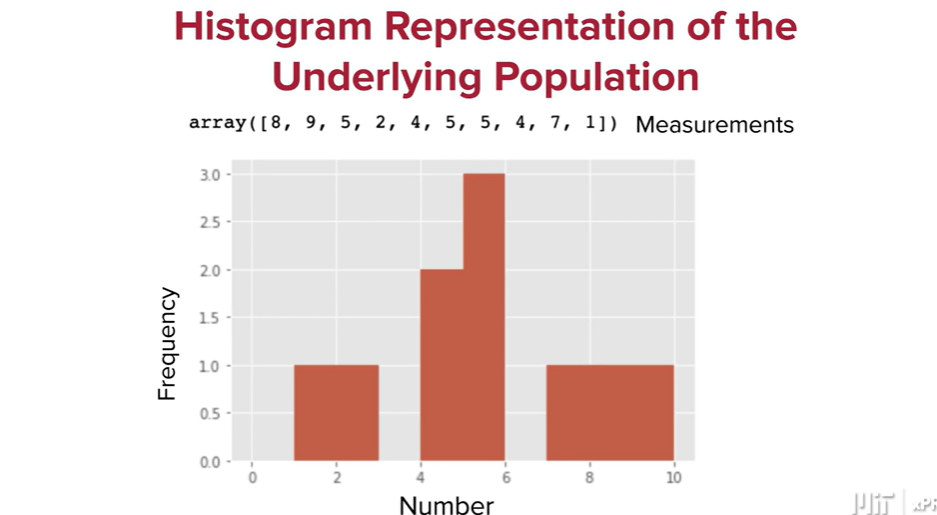
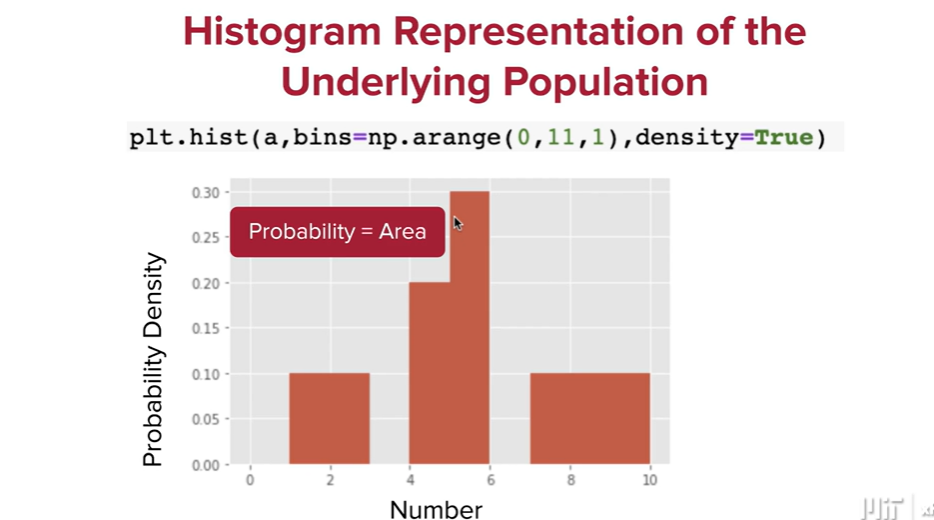
**MIT xPRO Data Engineering Certificate**

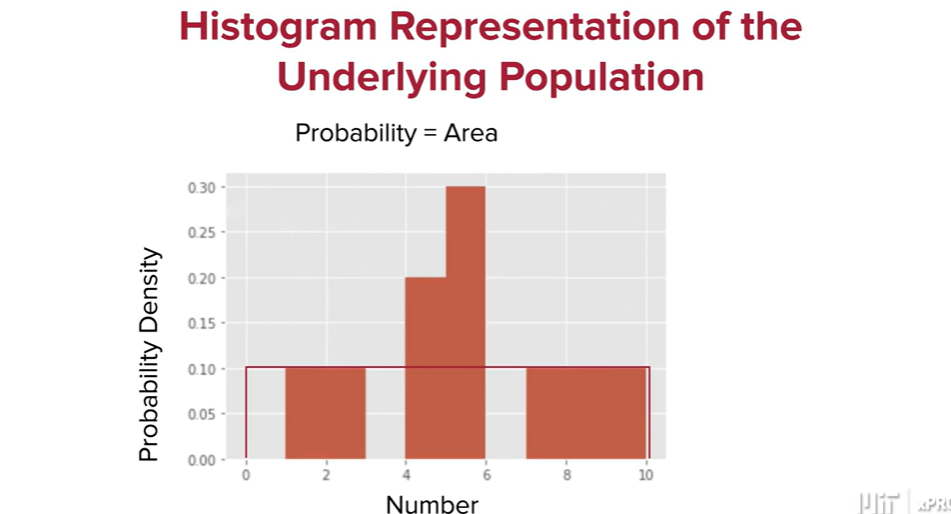
**Probability**

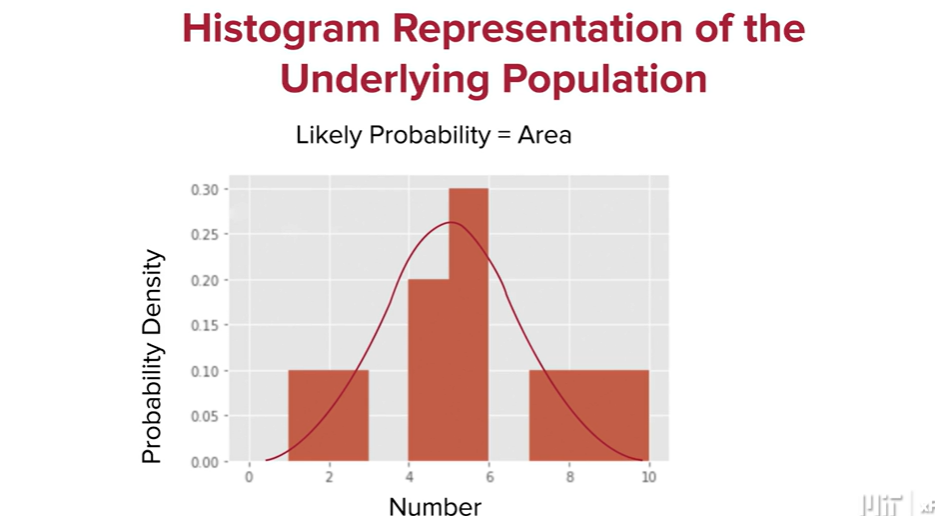


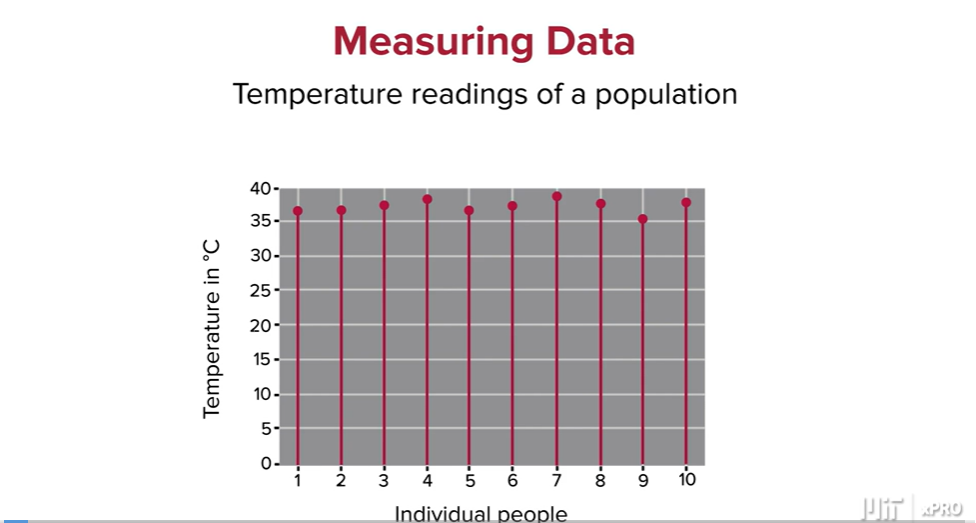


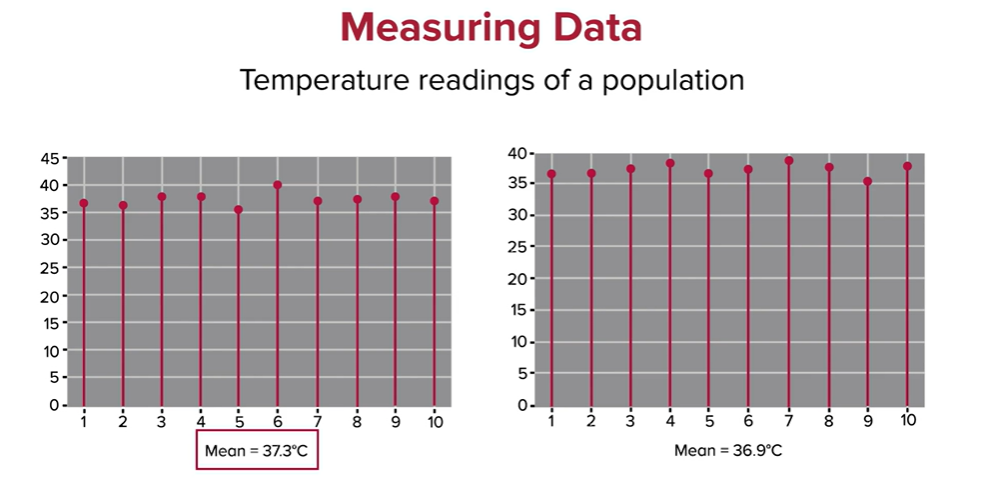


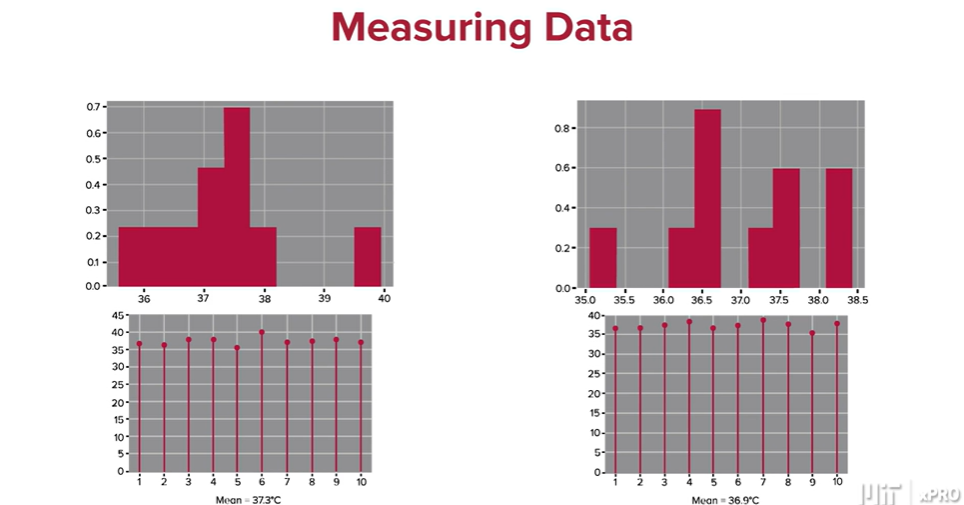


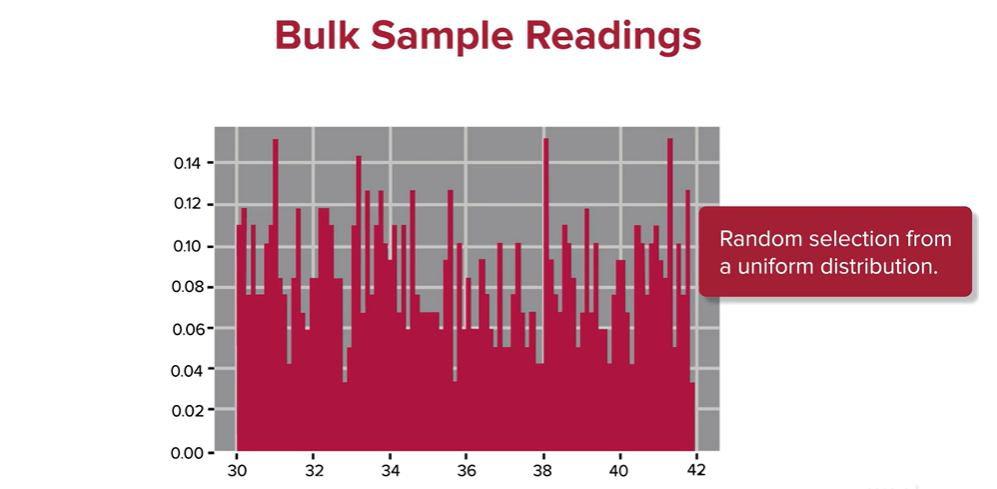


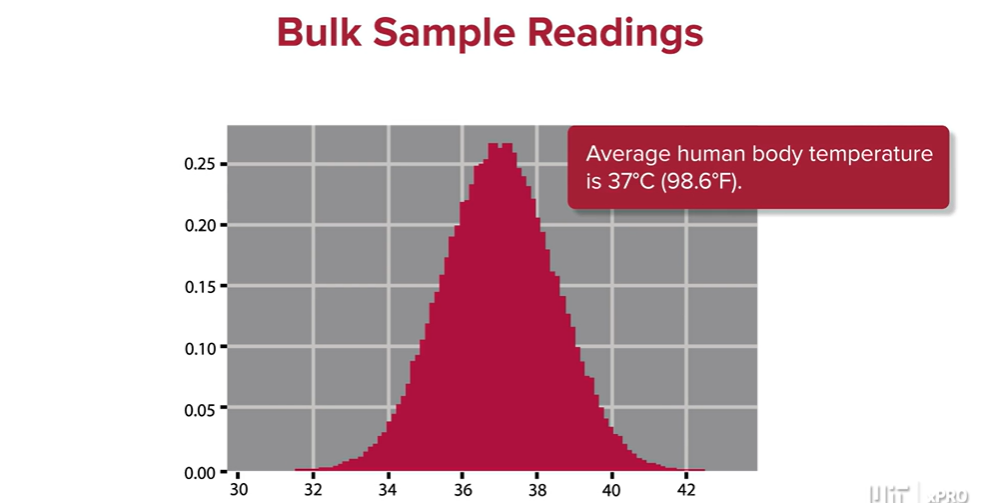




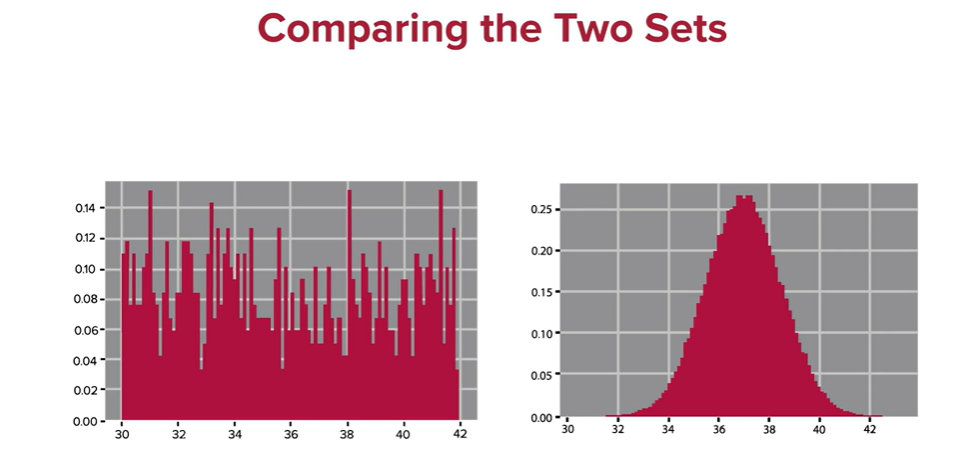




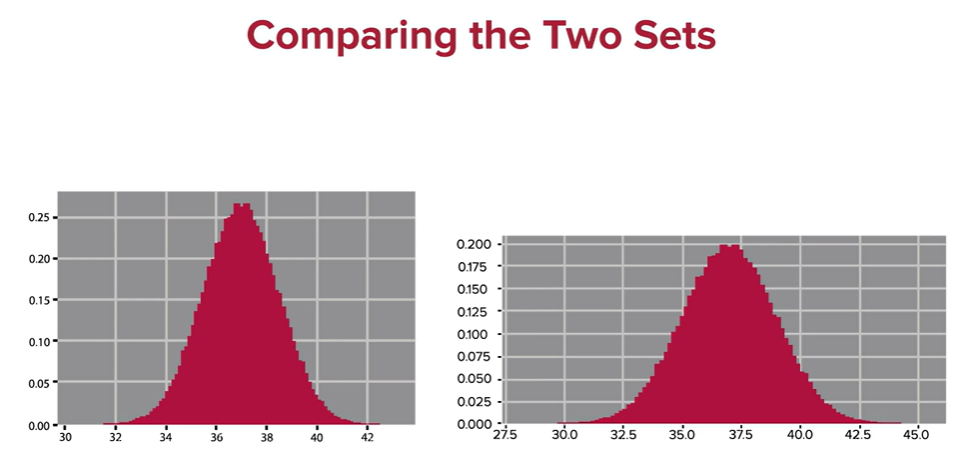




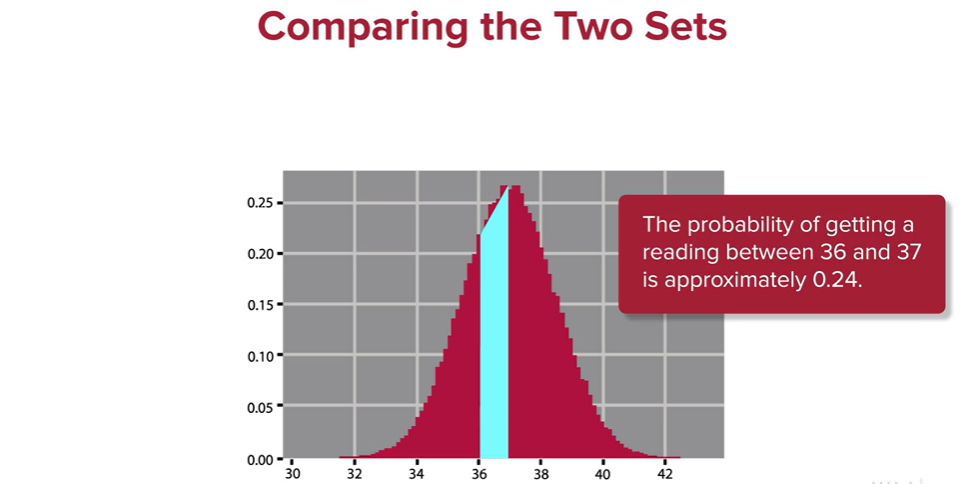


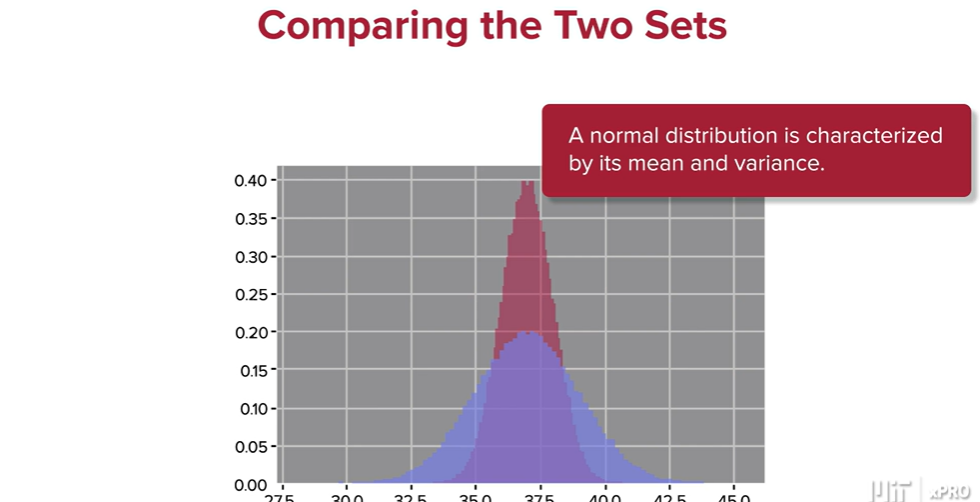


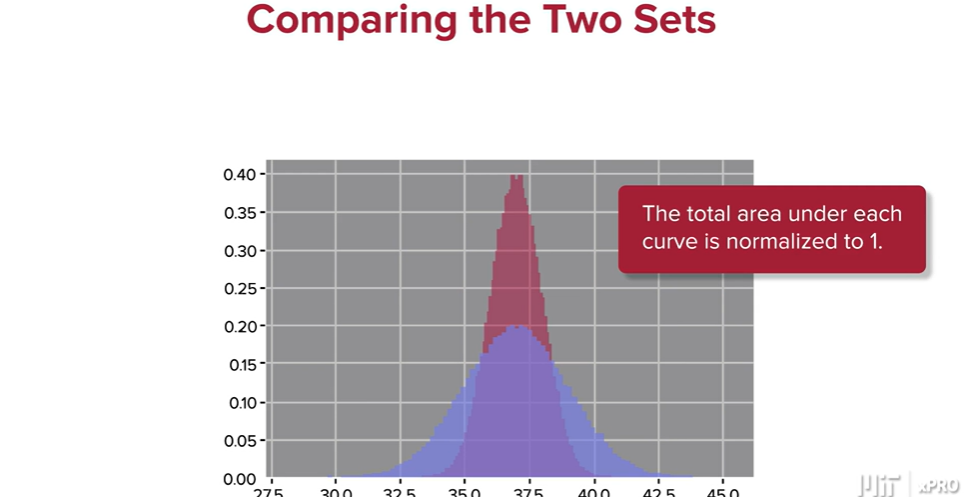
Variance is 1 (LHS) and variance is 2 (RHS):

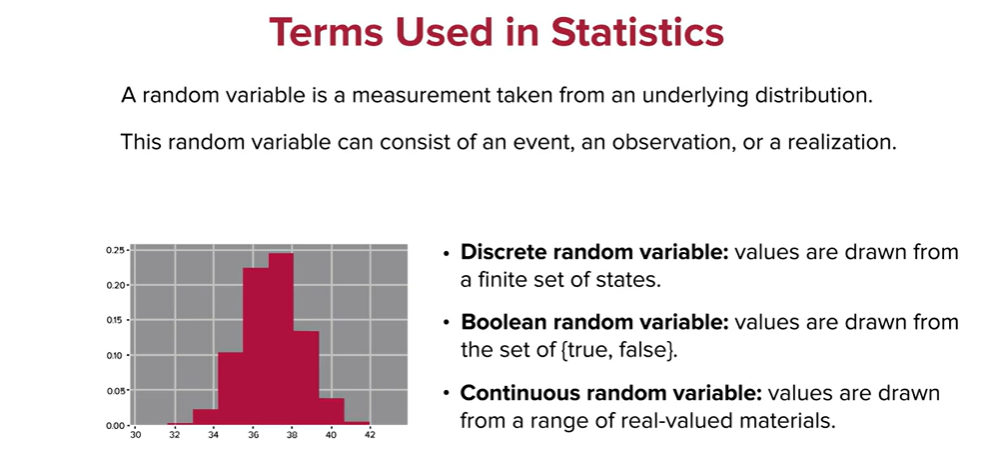


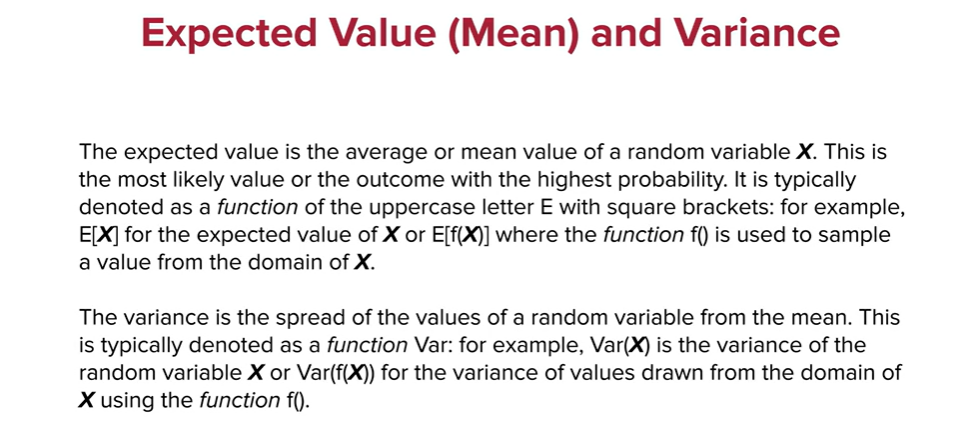




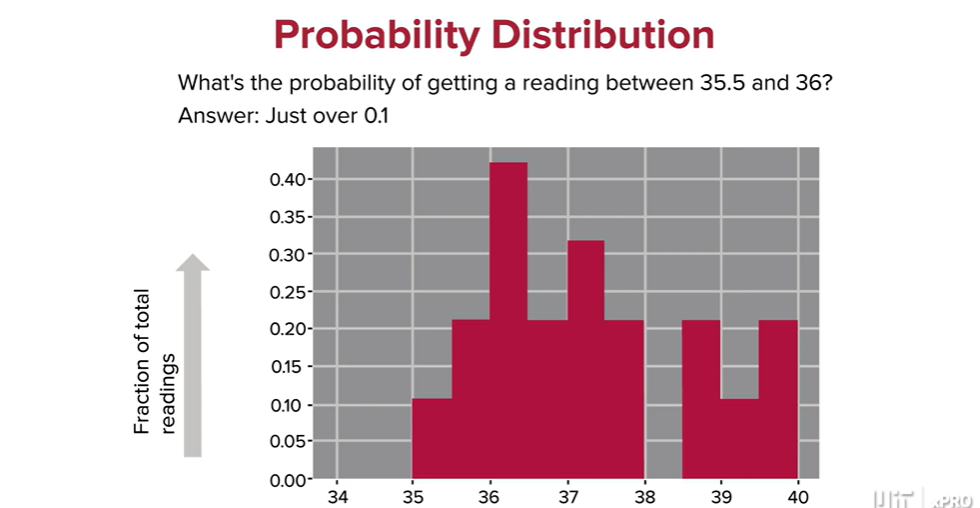




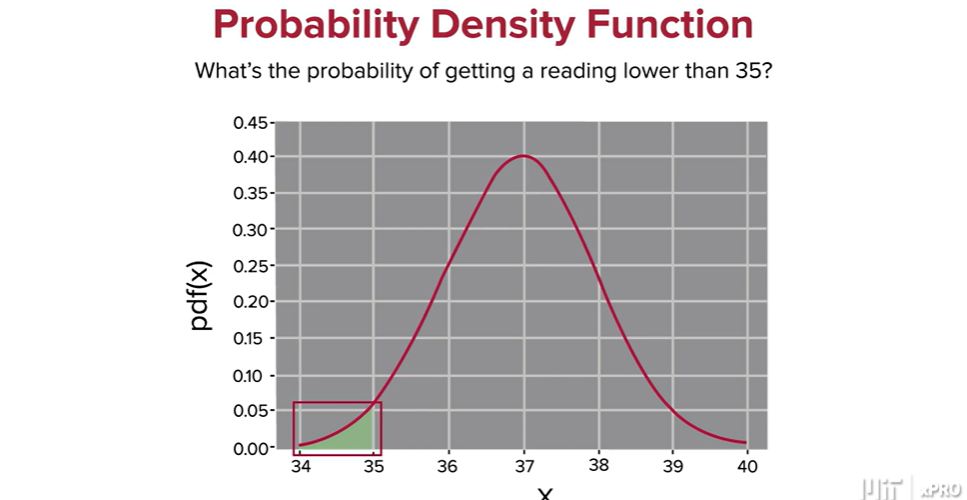


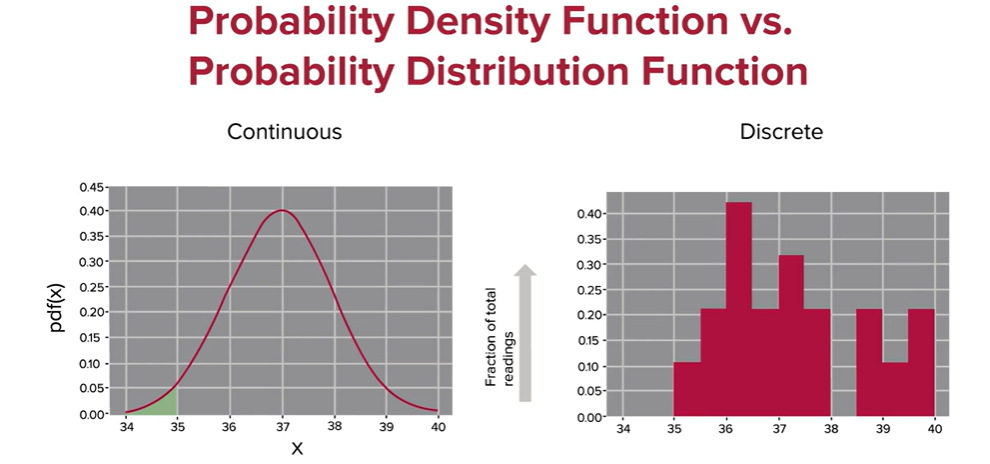


Discrete **probability distribution function**. Answer below is area of 0.5 (bin width) x 0.21 height of normalised frequency:

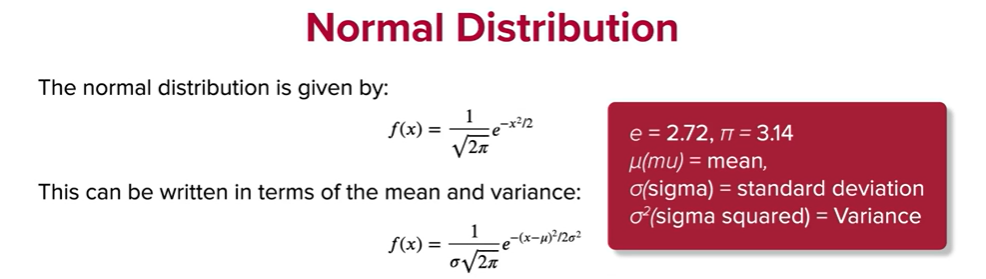


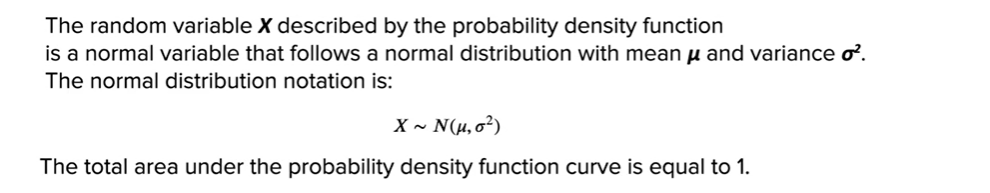
A continuous **probability density function**. You need to calculate the area under the line in green using the integral of the function:

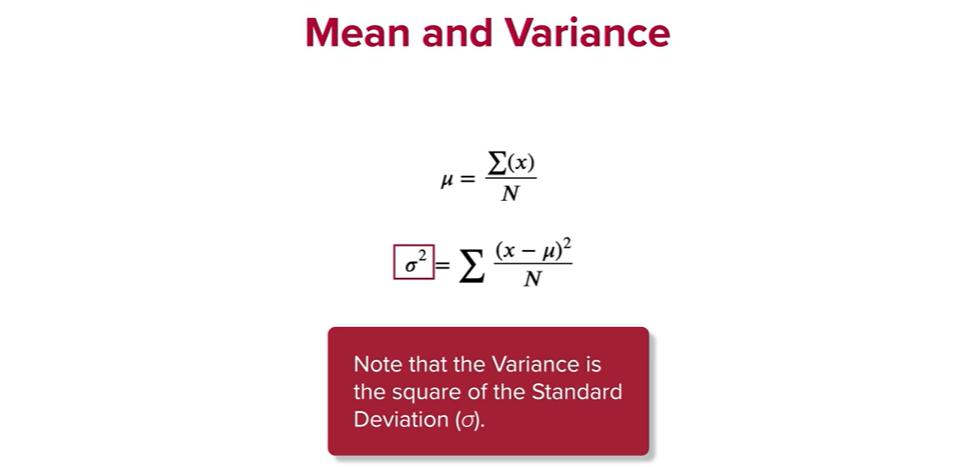


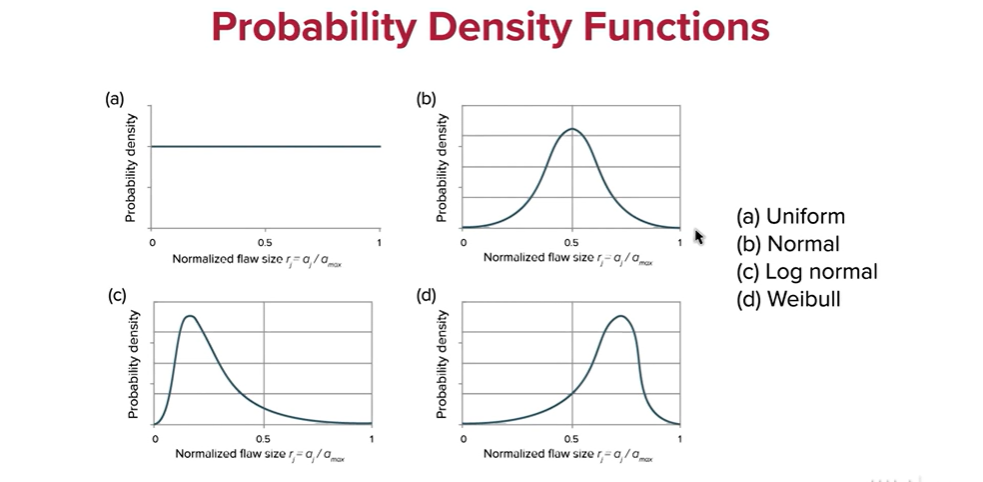


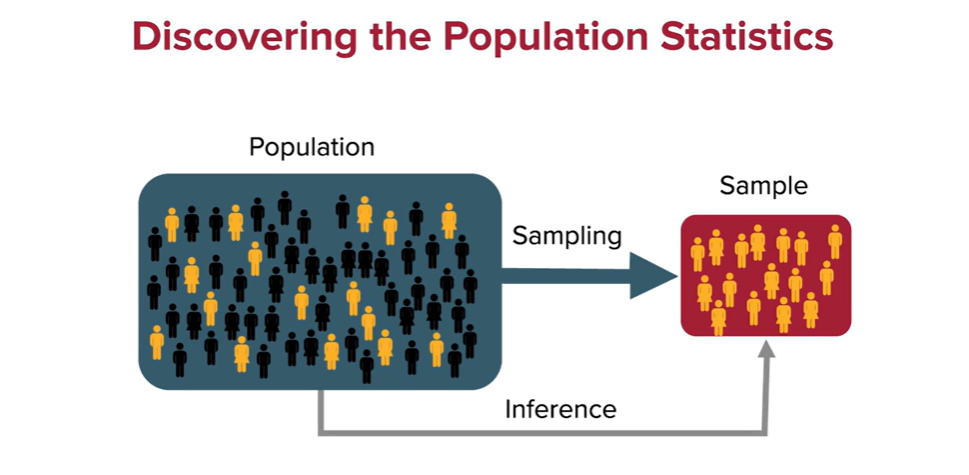
The top formula is the case when the distribution is centred on the origin (u=0) and var/SD=1. The bottom formulas is the full formula:

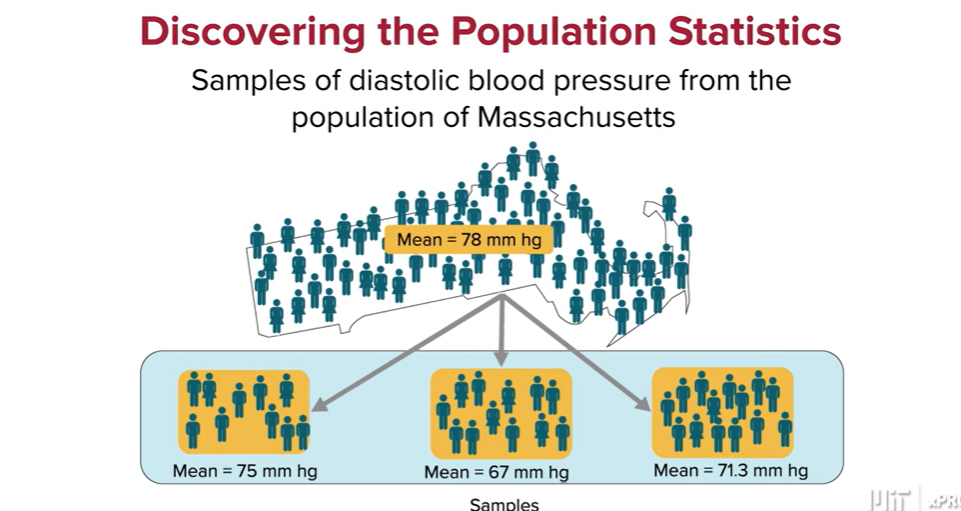


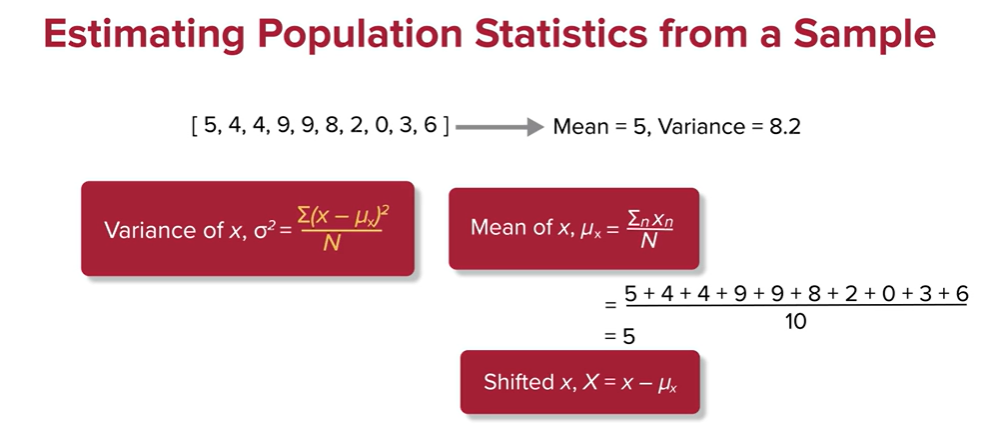




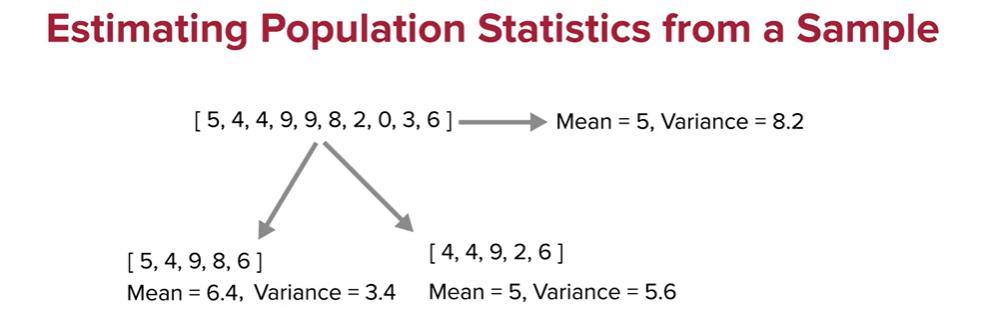


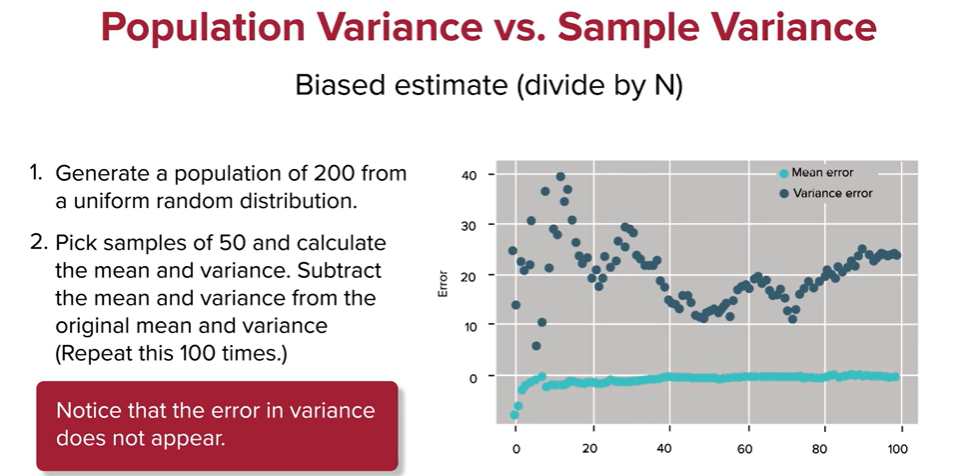




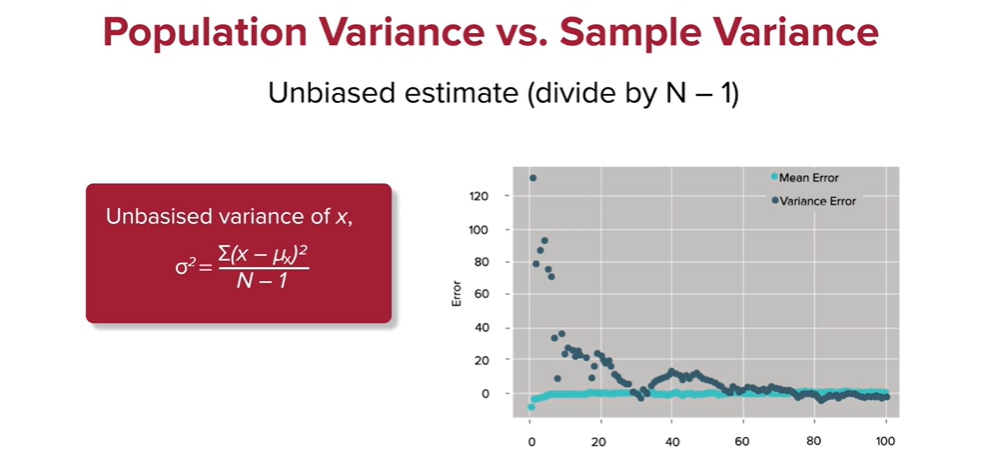


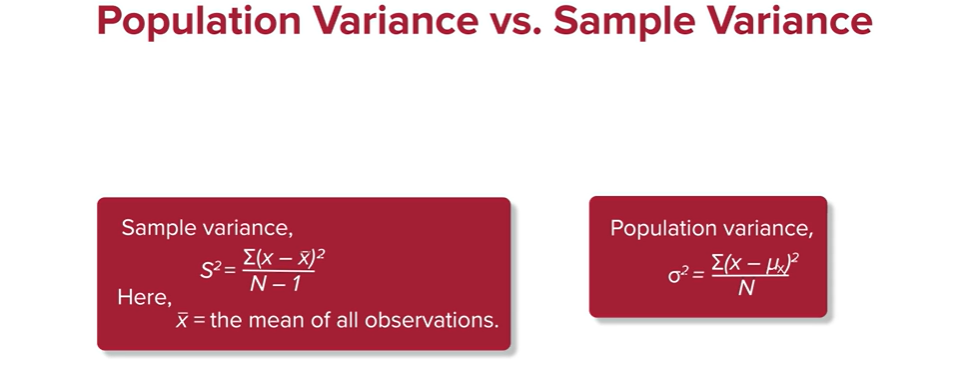
When we take a sample of these, the mean changes and the variance is smaller:

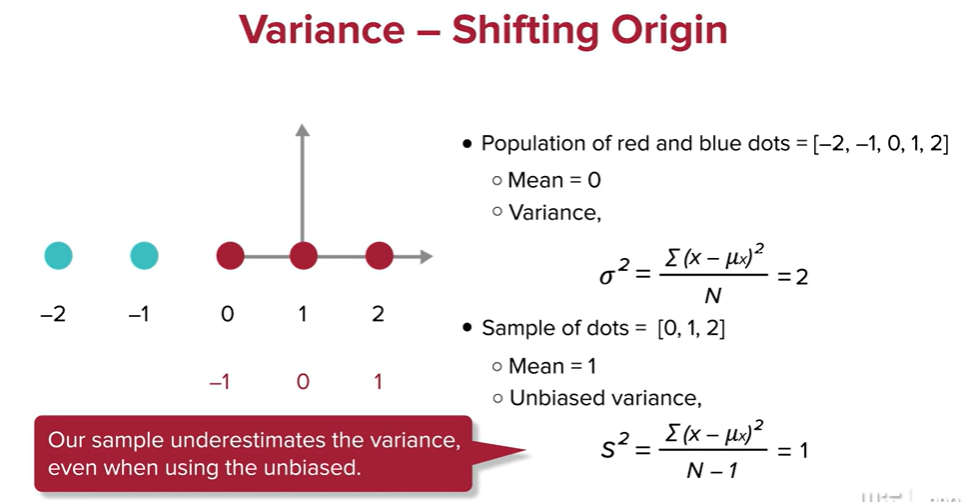


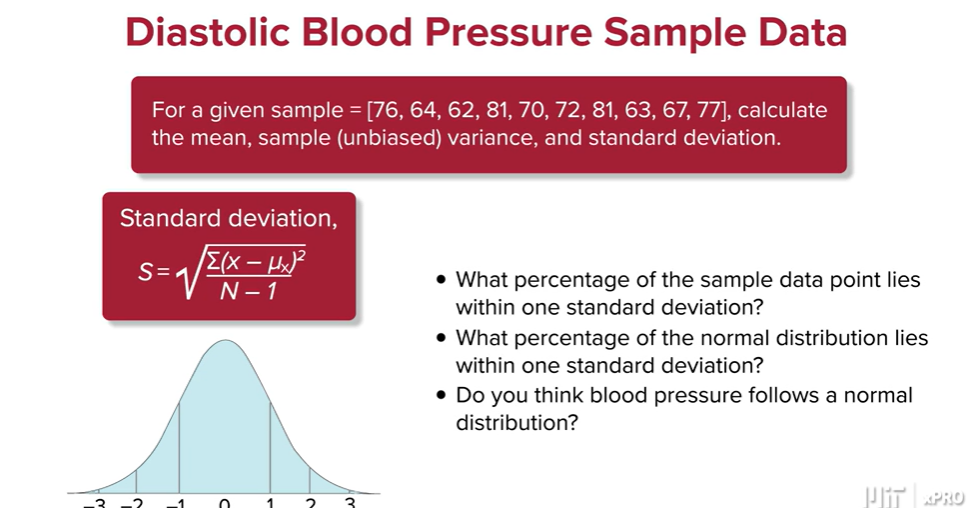


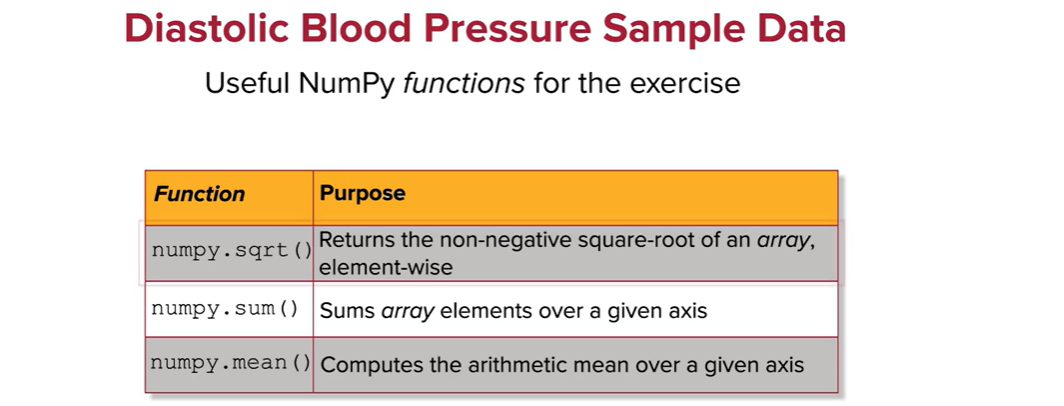
Dividing by N-1 is known as the Bessel correction. It improves he estimate, but it is still not necessarily mathematically accurate:



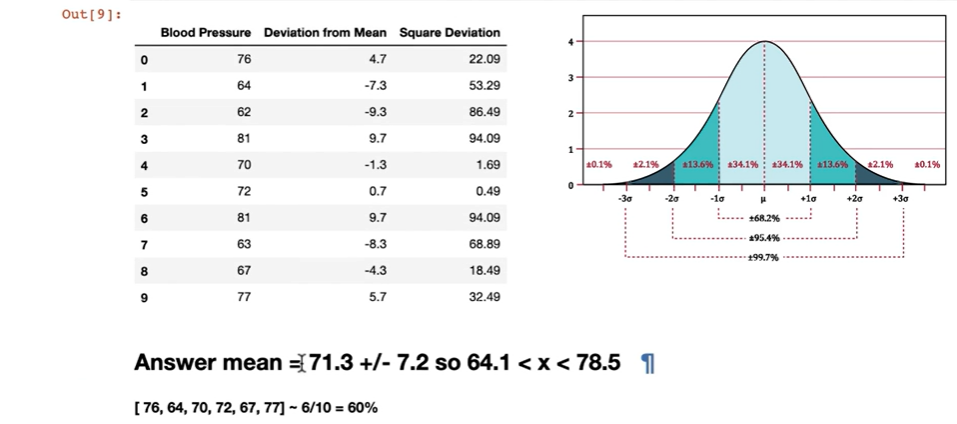


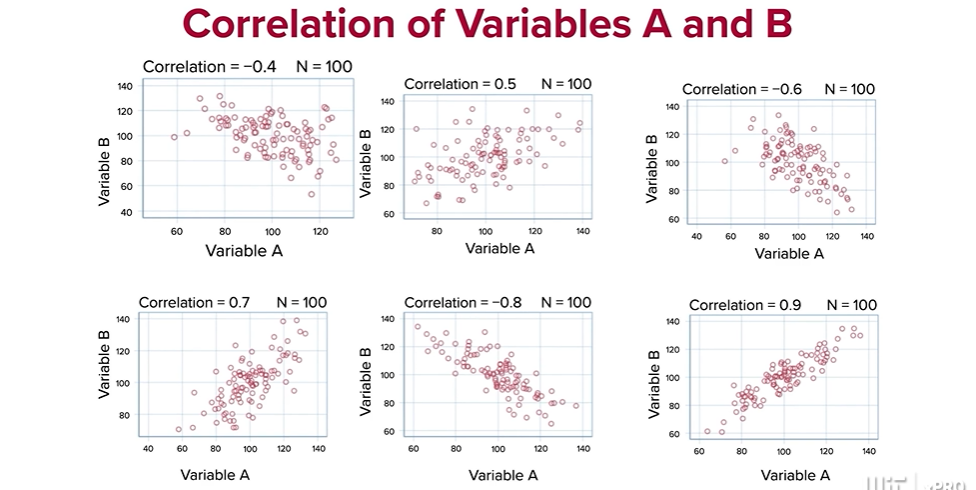


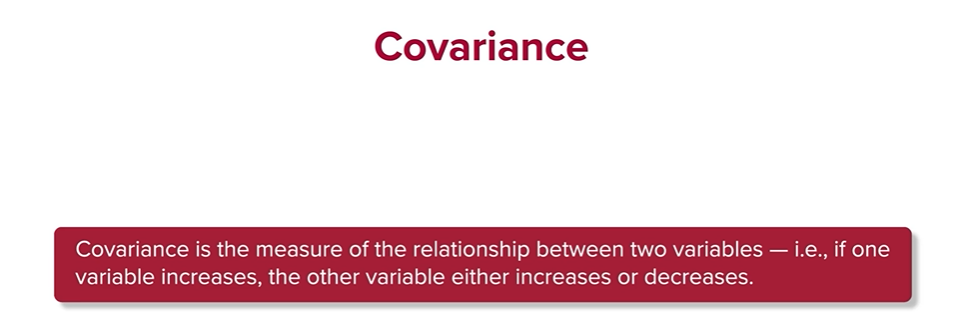


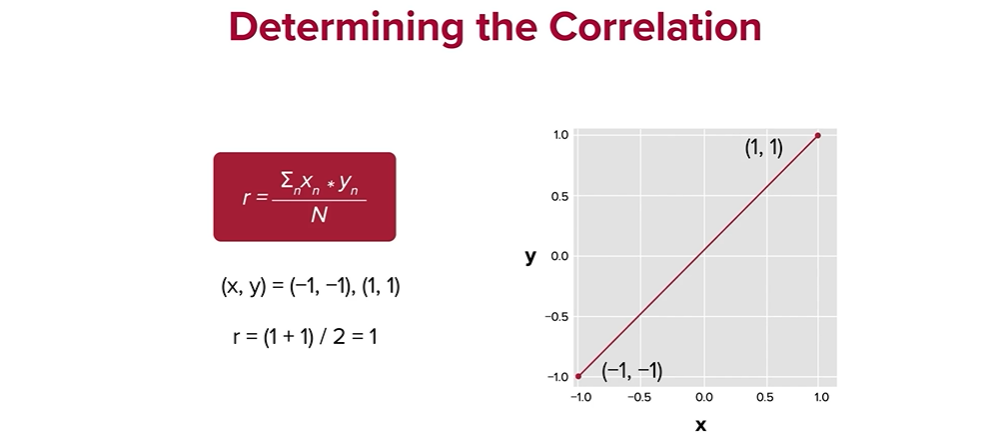


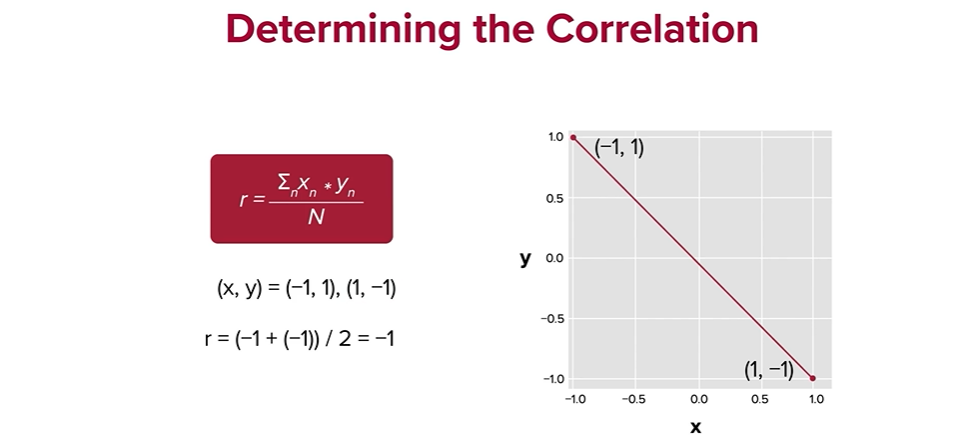
In this example, 60% of the data is within 1 SD of the mean, which is similar to the normal distribution (68%), which indicates blood pressure readings follow a normal distribution:

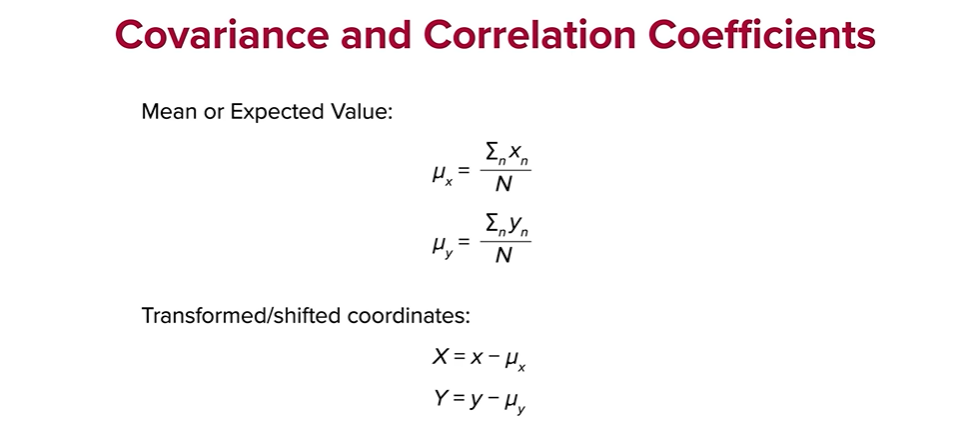


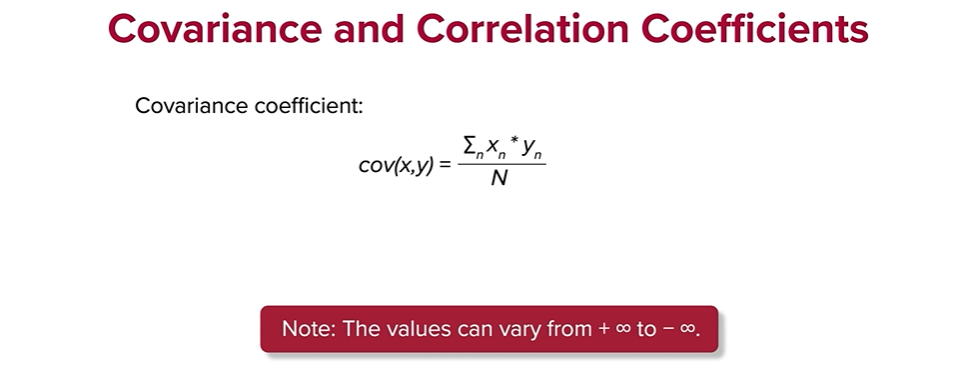




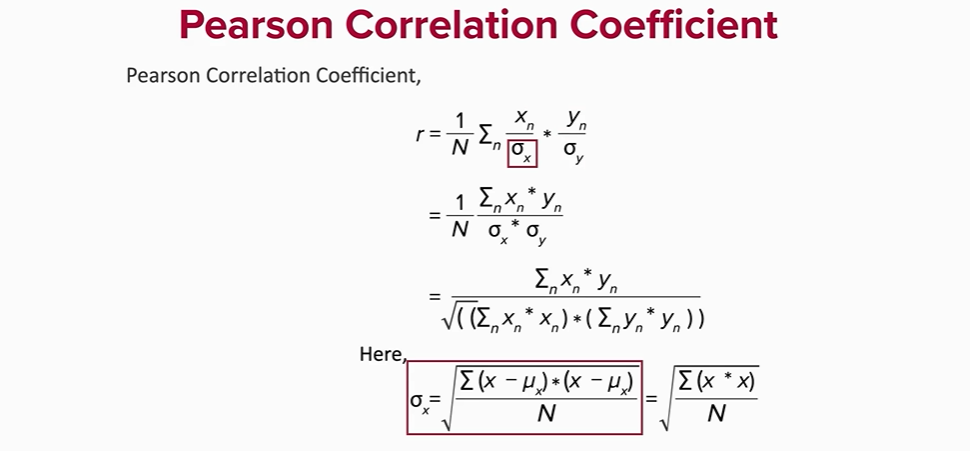




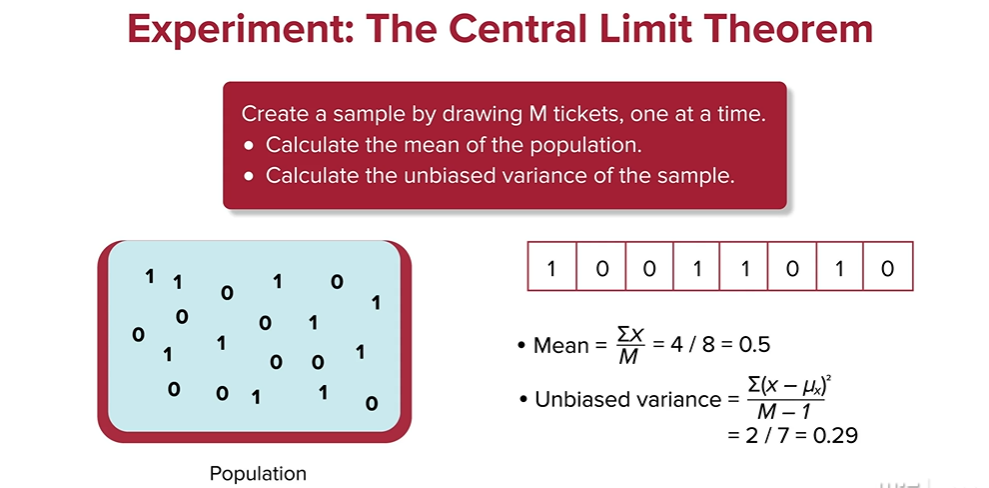




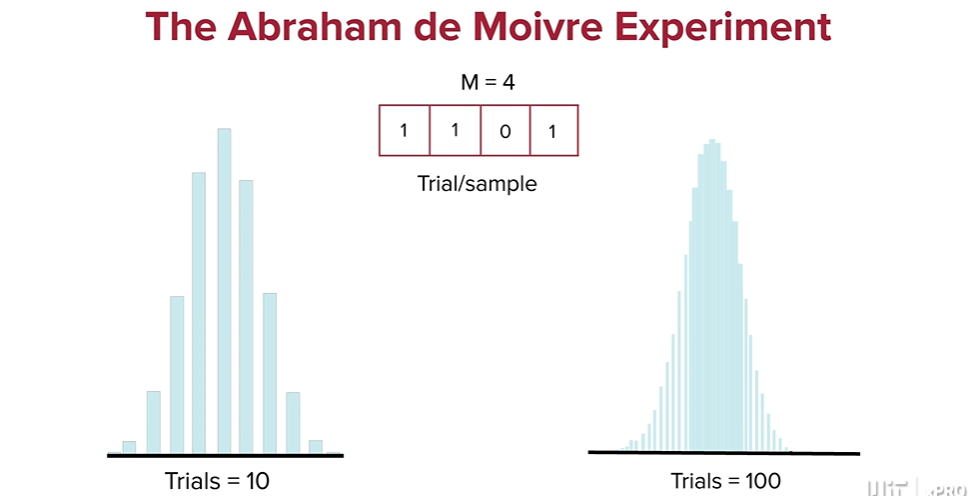
The correlation coefficient is normalised so it ranges from -1 to 1:



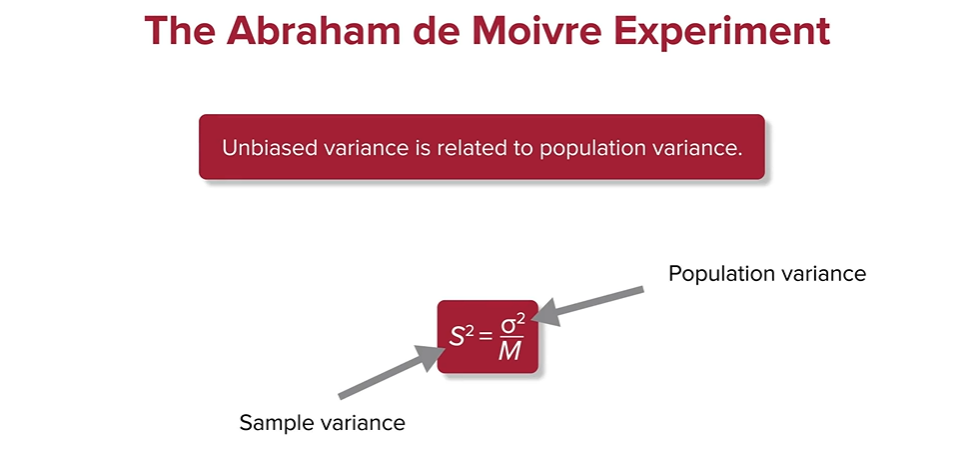
Draw samples with replacement:



The mean of trials has a normal distribution, even if the underlying distribution is not normal:



The more trials, the more the average sample variance decreases towards the variance of the population, which means the mean becomes more accurate:



**Probability**

Probability theory is “a branch of mathematics concerned with the analysis of random phenomena” (Britannica 2021). Probability focuses on the numerical descriptions of how likely an event is to occur. The result of probability is based on theoretical a priori propositions about events that have yet to occur. In contrast, statistics is a branch of applied mathematics that focuses on the results of events that have already happened (a posteriori).

Probability is measured on a scale from 0 to 1, or from 0% to 100%, where 0 describes the likelihood of an impossible event and 1 refers to the probability of an event that will definitely happen.

In mathematical terms, the probability of an event ***X*** can be expressed as

P(***X***) = (Numbers of ways ***X*** can happen) ∕ (Total number of outcomes for the experiment)

**Example 1: The Probability of a Coin Toss**

Suppose you want to compute the probability of getting heads when tossing a fair coin. Because a coin can land on either heads or tails, the possible number of outcomes for this experiment is two, and the number of ways to get heads is one.

Therefore, the probability of the desired outcome is computed with

P(heads) = 1 ∕ 2 = 0.5 = 50%

**Example 2: The Probability of Drawing a Spade from a Standard Deck of Cards**

A standard deck of cards has 52 cards equally grouped in four suits: hearts, diamonds, spades, and clubs. Therefore, there are exactly 13 cards in each suit.

The probability of drawing spades is therefore

P(spades) = 13 ∕ 52 = 0.25 = 25%

**What Is a Random Variable?**

Arandom variable ***X*** is a variable whose possible values are numerical outcomes of a random phenomenon.

For example, in the coin toss experiment, if the outcomes of tails and heads are mapped as 0 and 1, respectively, then therandom variable ***X*** for this experiment will have possible values of 0 and 1.

On the other hand, when rolling a fair die, the possible values for therandom variable are 1, 2, 3, 4, 5, and 6, because these are the possible values that can be obtained in the experiment.

There are two types of random variable*s*: discrete and continuous.

If arandom variable can take only a finite number of distinct values, then it must be discrete. The variables described above are examples of discrete random variable*s*.

“A continuousrandom variable is one that takes an infinite number of possible values. Continuousrandom variable*s* are usually measurements” (Yale University n.d.). Examples include depth or weight because these measurements can assume any value within a specified range.

A continuousrandom variable is arandom variable with a set of possible values (known as the range) that is infinite and uncountable. Continuous random variables are usually measurements, such as height and weight, because they can assume any value within a specified range.

**What Is a Probability Distribution *Function*?**

The probability distribution *function* for a random variable***X***maps the values that ***X*** can take to the probability value for that outcome.

For example, consider again the experiment of rolling a fair die. As described above, the random variable***X*** can assume any integer value between 1 and 6. If the die is fair, then any outcome has an equal probability of 1/6.

Therefore, the probability distribution for this experiment can be summarized as follows:

| Outcome | 1 | 2 | 3 | 4 | 5 | 6 |
| --- | --- | --- | --- | --- | --- | --- |
| Probability | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 |

As you will see next, probability distribution *functions* can be discrete or continuous.

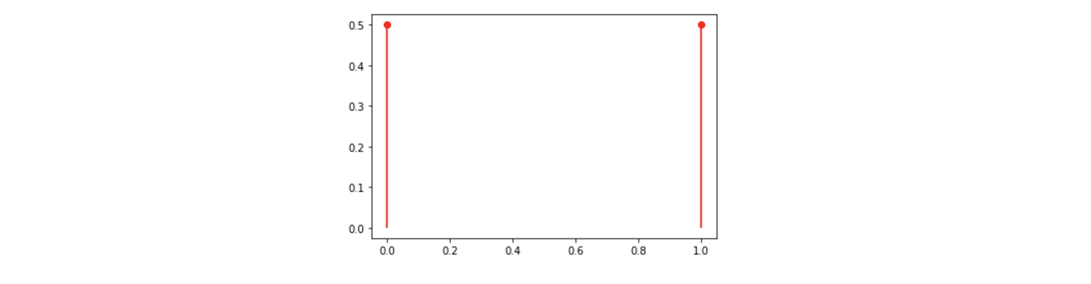
**Common Discrete Probability Distribution *Functions***

**Binomial Distribution**

One common discrete probability distribution *function* is the binomial distribution. “A binomial distribution describes the probability of success or failure of an experiment that is repeated multiple times” (Glen 2021). Therefore, the binomial is a type of distribution that only displays two possible outcomes.

A coin toss has only two possible outcomes: heads or tails. Therefore, a coin-tossing experiment can be described by the binomial distribution.

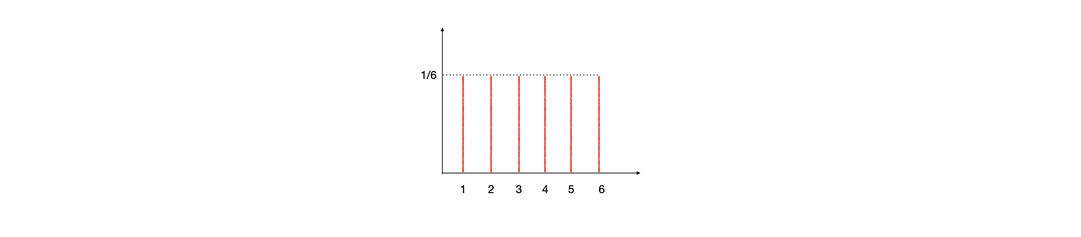
The graph below shows the distribution for a single toss of a fair coin.



**Uniform Distribution**

A uniform distribution, also called a rectangular distribution, is a probability distribution that has constant probability.

Rolling a single die is one example of a discrete uniform distribution. A die roll has six possible outcomes: 1, 2, 3, 4, 5, or 6. There is a 1/6 probability of rolling each number.



**Poisson Distribution**

Another type of discrete probability distribution *function* is the Poisson distribution. The Poisson distribution can be used to predict the probability of events occurring within a fixed interval based on how often they have occurred in the past.

The Poisson distribution is given by



where 𝜆 represents the average number of occurrences in an interval. This parameter also dictates the shape of the Poisson distribution.



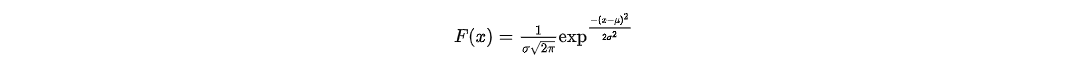
Here is an example of how to use the Poisson distribution. “A textbook store rents an average of 200 books every Saturday night. Using this data, it is possible to predict the probability that more books will sell (perhaps 220 or 240) on the following Saturday nights” (Glen 2013).

**Common Continuous Distribution *Functions***

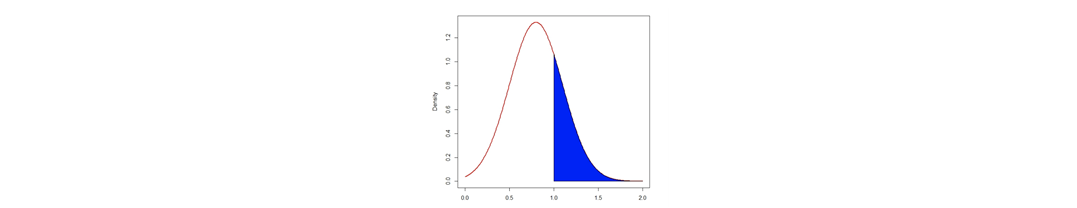
**Normal Probability Distribution**

The normal probability distribution, a widely used continuous distribution, is actually a family of distributions with differing means (indicated by the Greek letter μ) and standard deviations (indicated by the Greek letter σ).

The normal distribution is given by



The normal distribution is symmetric and centered on the mean (which is the same as the median and the mode).“Because there are infinite values that the random variable ***X*** could assume, the probability of ***X*** taking on anyone specific value is 0” (Nicholas School of the Environment 2021). Therefore, values are often expressed in ranges. For example, the probability that a random variable ***X*** is greater than 1 is given by the area shaded in blue in the image below:

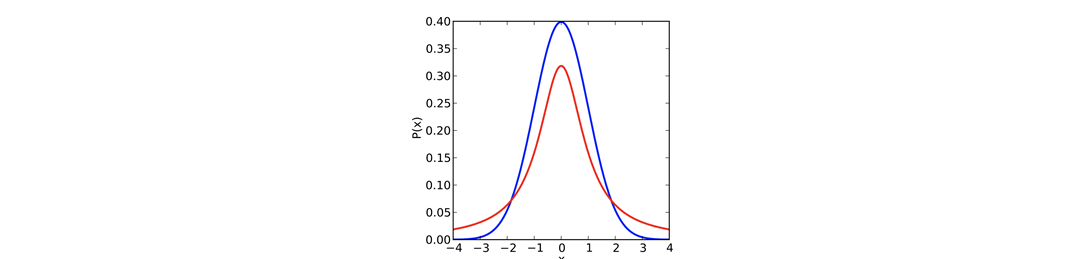


**Student’s *t-Distribution***

Another type of discrete probability distribution *function* is the *t-distribution*. The *t-distribution*, “also known as the Student’s *t-distribution*, is a type of probability distribution that is similar to the normal distribution with its bell shape but has heavier tails” (Hayes 2021). It is, therefore, more prone to producing values that fall far from its mean. “*T-distributions* have a greater chance for extreme values than normal distributions, hence the fatter tails” (Hayes 2021).

The shape of a *t-distribution* is determined by a parameter called degrees of freedom. The concept of degrees of freedom indicates that the value may shift depending on the factors that influence the calculation of the statistic (Frost 2021). If the value of the degrees of freedom is small, the distributions will have heavier tails. On the other hand, a higher value for the degrees of freedom will make the *t-distribution* resemble a standard normal distribution.

The probability density *function* is symmetric, and its overall shape resembles the bell shape of a normally distributed variable with mean 0 and variance 1, except that it is a bit lower and wider. If the value of the degrees of freedom is small, the distribution will have a heavier tail. On the other hand, a higher value for the degrees of freedom will make the *t-distribution* resemble a standard normal distribution.



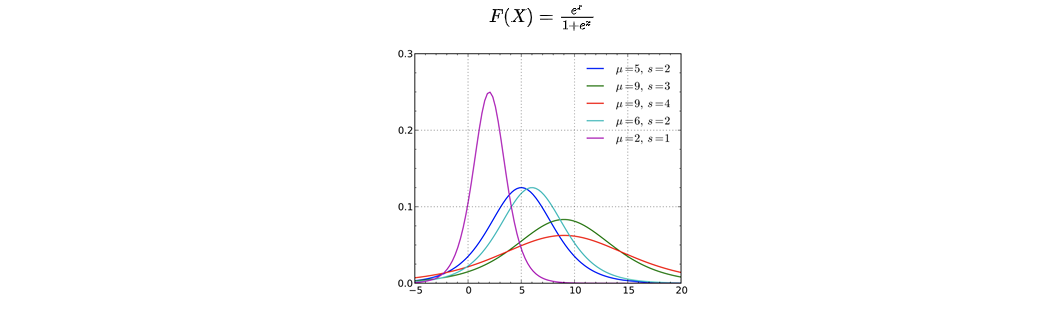
The *t-distribution* plays a role in many widely used statistical analyses, including the Student's t-test, which assesses the statistical significance of the difference between two sample means (Frost 2021).

**Logistic Distribution**

The logistic distribution is used for various growth models and a type of regression known, appropriately, as logistic regression (Wikipedia 2021).

The standard logistic distribution is a [continuous distribution](https://www.randomservices.org/random/dist/Continuous.html) on the set of real numbers with [distribution *function*](https://www.randomservices.org/random/dist/CDF.html)

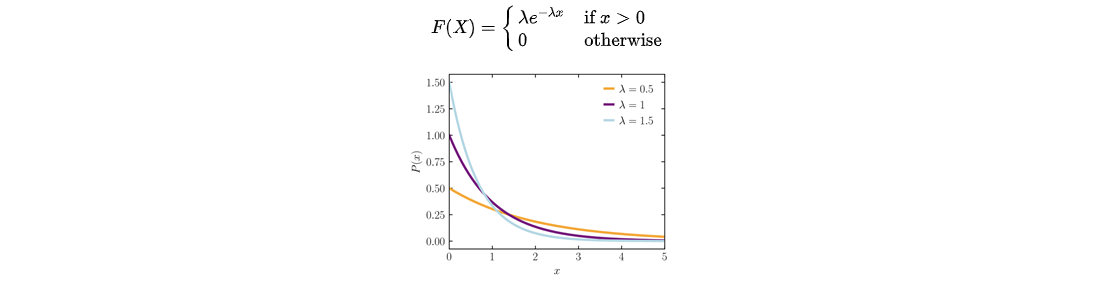
*F* given by



**Exponential Distribution**

The exponential distribution is a widely used continuous distribution. It is often used to model the time elapsed between events.

A continuous random variable***X*** is said to have an exponential distribution with a parameter λ > 0 if its probability density *function* is given by



To get an intuitive feel for this interpretation of the exponential distribution, suppose you are waiting for an event to happen. For example, you are at a store and are waiting for the next customer. In each millisecond, the probability that a new customer will enter the store is very small. Imagine that a coin is tossed each millisecond. If it lands on heads, a new customer enters. The time until a new customer arrives will approximately follow an exponential distribution.