A big portion of the work that data scientists do is handling and manipulating real world data, and depending on what that data represents it is going to have an associated **data distribution.**

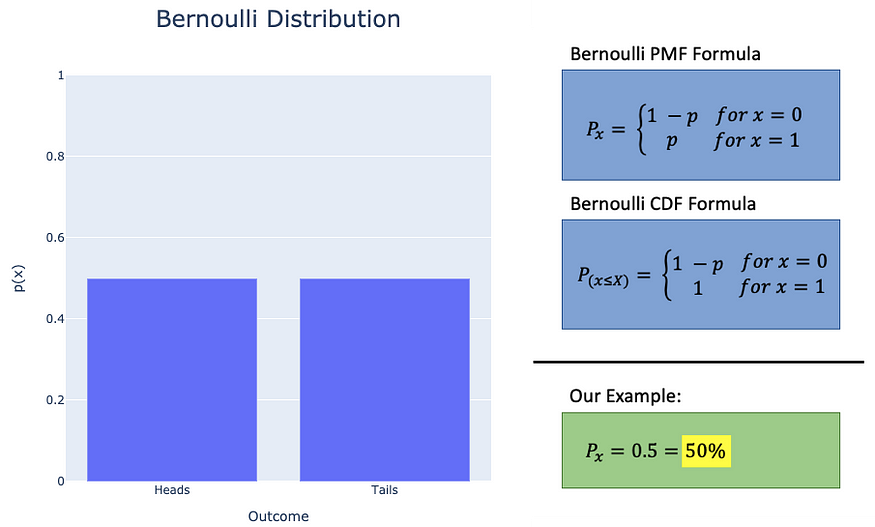
**Discrete Data Distributions**

**1. Bernoulli Distribution**

The Bernoulli Distribution captures the probability of receiving one of two outcomes (often called success or failure) given a single trial. It is actually just a special case of the Binomial distribution where *n=1*.

Example:

* You are out with your friend, and you pull a coin from your pocket to determine who is buying the next round of drinks 🍻 The outcome of this coin flip can be modelled using the Bernoulli distribution.



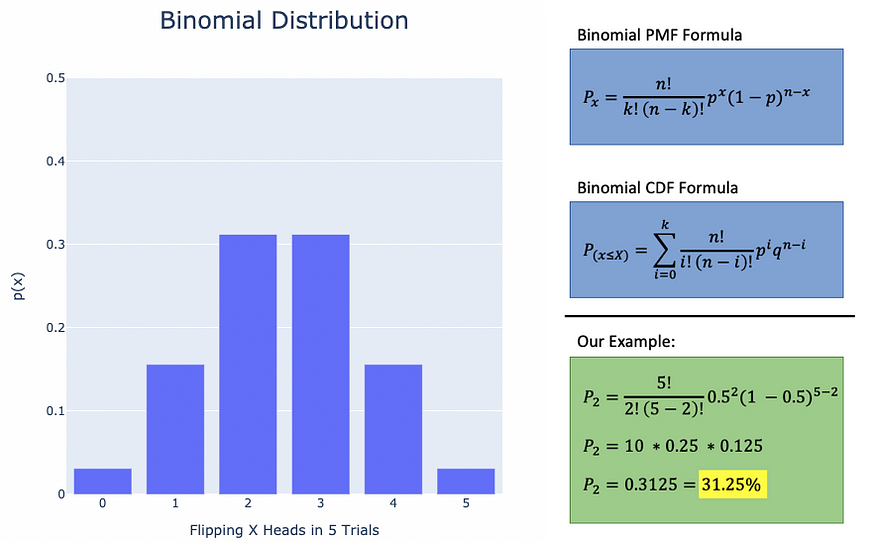
Bernoulli distribution — [learn more here](https://www.statisticshowto.com/bernoulli-distribution/)

**2. Binomial Distribution**

The binomial distribution is just taking Bernoulli one step further. We still have trials that result in one of two outcomes (success or failure), but now we are looking at the **probability that a specific number of outcomes (*x*)occur in *n trials*instead of a single trial.**

Example:

* You decide to try and trick your friend (with statistics, of course) and you say that you will buy the next round of drinks if you flip a coin 5 times and heads comes up *exactly* 2 times 🪙



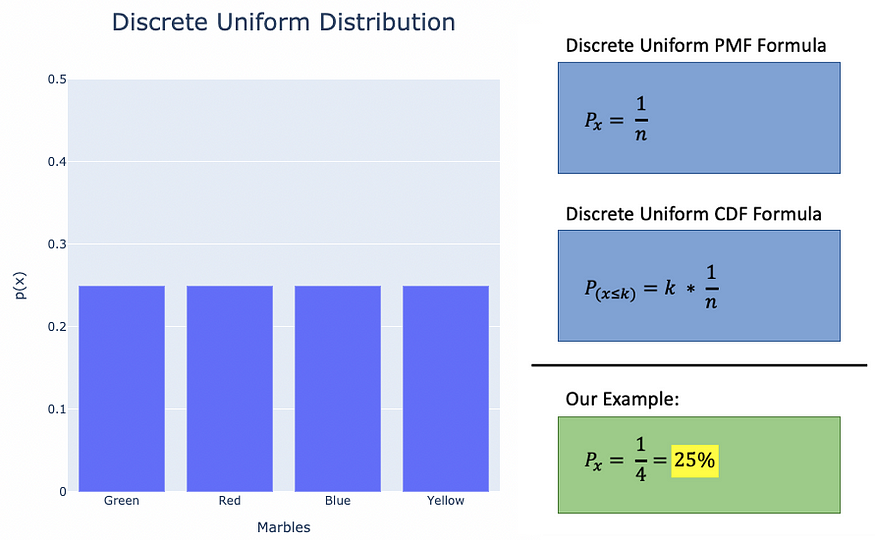
Binomial distribution — [learn more here](https://www.statisticshowto.com/probability-and-statistics/binomial-theorem/binomial-distribution-formula/)

**3. Discrete Uniform Distribution**

A discrete uniform distribution is a simple distribution where we have a set of potential outcomes (*n*), each of which has an equal likelihood of occurring.

Example:

* You blindly reach into a bag of marbles that contains a green marble, a red marble, a blue marble, and a yellow marble. What are the chances of picking the yellow marble? 🟡



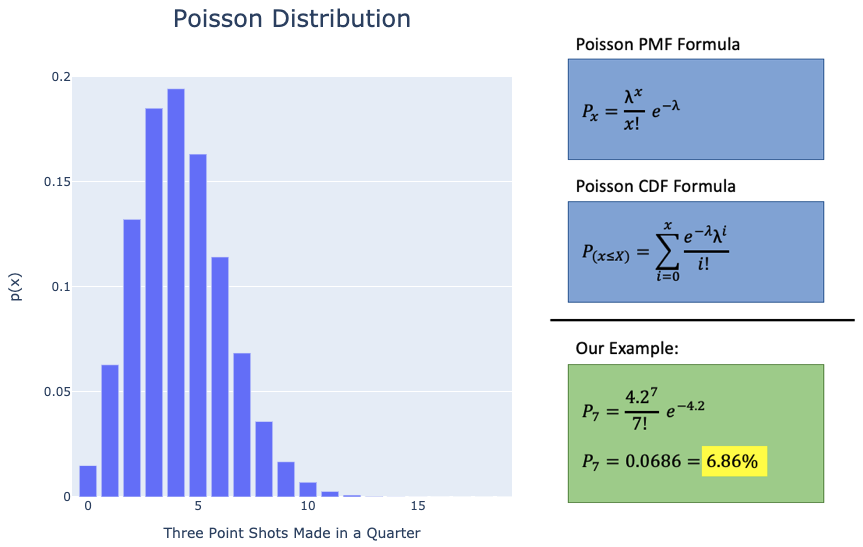
discrete uniform distribution— [learn more here](https://mathworld.wolfram.com/DiscreteUniformDistribution.html)

**4. Poisson Distribution**

The Poisson distribution is used to answer the question **how many times is an event likely to occur over a given period of time?**The Poisson distribution is defined by a *rate parameter* (λ), which is the mean number of occurrences of that event in a single unit of the observed time.

Example:

* Let’s say that a basketball team scores an average of 4.2 three point shots per quarter 🏀 If that is true, then what is the likelihood that this team will score exactly 7 three point shots in a quarter?



Poisson distribution — [learn more here](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/probability-distribution/poisson-distribution/)

**Continuous Data Distributions**

**1. Gaussian (Normal) Distribution**

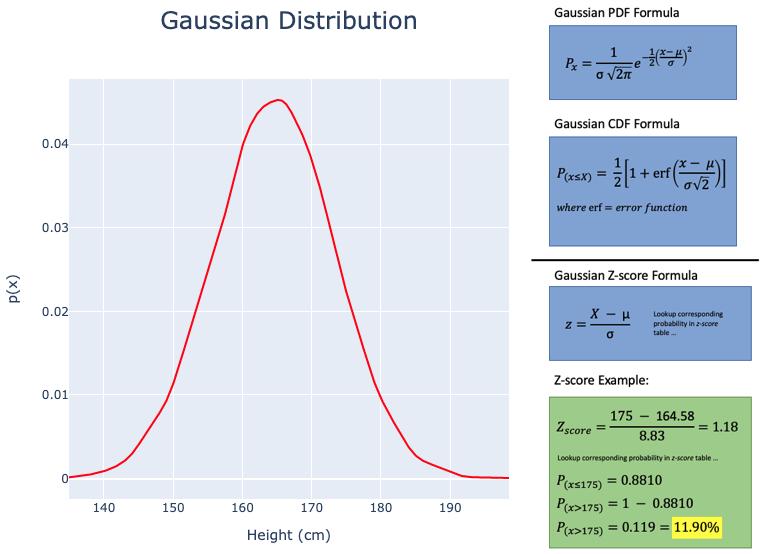
The Gaussian distribution is probably the **most widely known and observed distribution in nature.** It is defined by a *mean value* (μ) and a *standard deviation (*σ).

Also now that we are talking about continuous values, we can no longer say “what is the likelihood of this *exact* value occurring” because technically there is no *exact* values in a continuous space. Instead we ask the question “what is the likelihood of a sample *falling within a given range of values*?”

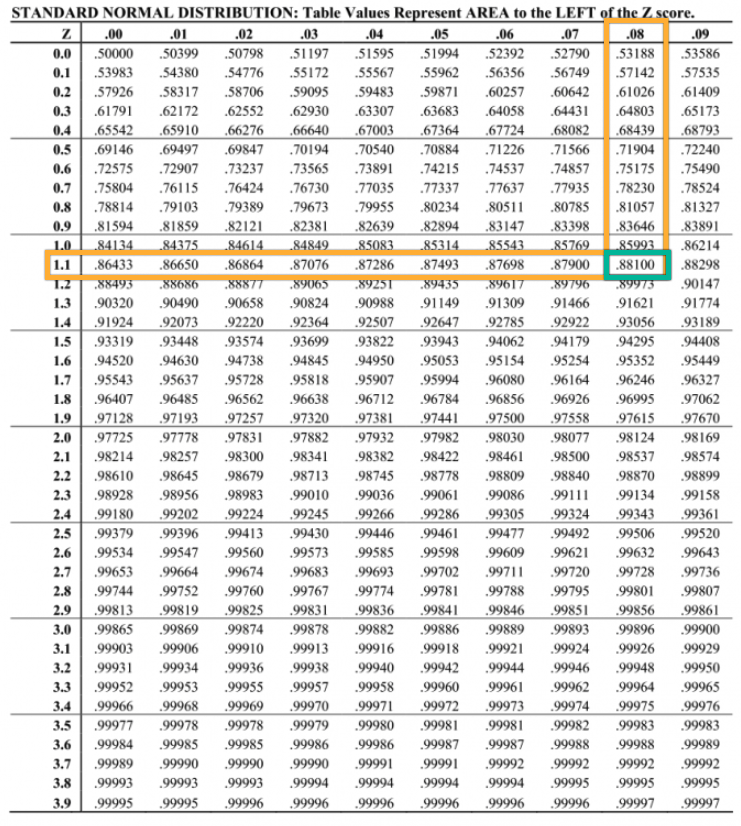
Example:

* Let’s say I am friends with every single person in the world (😎) and everyone volunteers their height information to me. The mean height of the population turns out to be 164.58cm with a standard deviation of 8.83cm.
* Given the information above, what is the probability of someone being taller than 175cm?

*\*\*NOTE\*\* —*for calculations in a continuous space, we sometimes approximate probabilities by calculating *scores* and finding their associated probabilities in a *lookup table*. For Gaussian distribution we use what is called the***z-score*.**



Gaussian distribution — [learn more here](https://www.scribbr.com/statistics/normal-distribution/)



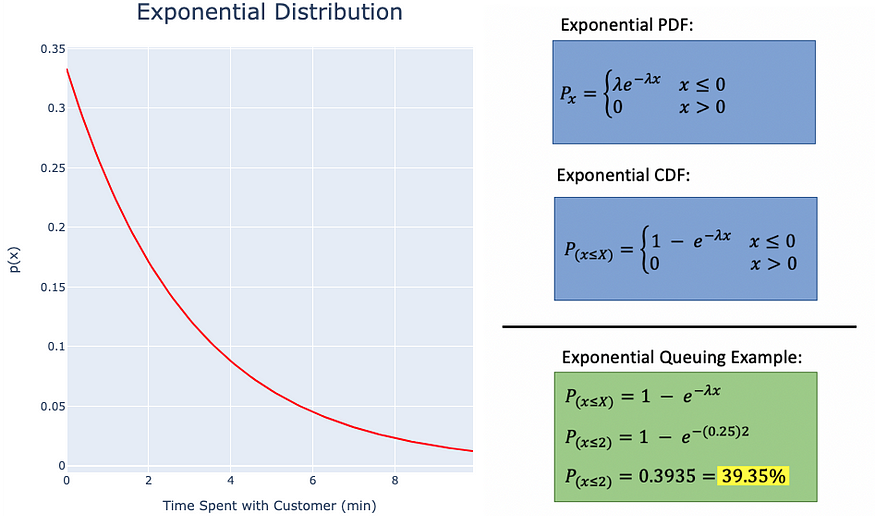
z-score table (positive z-scores)

**2. Exponential Distribution**

The exponential distribution is a distribution who’s random values follow an exponential pattern. **The exponential function is often used in determining the waiting time until the next event occurs.**The important parameter for this distribution is the *rate parameter* (λ), which is the rate of events done/per unit time.

Example:

* A postal worker spends an average of 4 minutes with each customer, and the time they spend with customers can be represented as an exponential function (0.25 customers/min)
* Given a random customer, what is the probability the postal worker will spend less than 2 minutes with them?



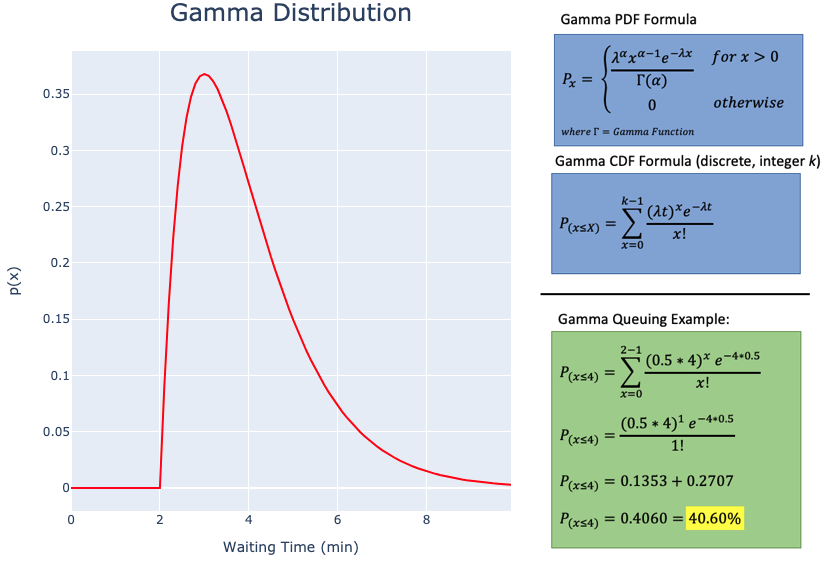
Exponential distribution — [learn more here](https://opentextbc.ca/introstatopenstax/chapter/the-exponential-distribution/)

**3. Gamma Distribution**

Similar to the exponential function, Gamma distributions are often used for waiting time problems. The difference is that Gamma distributions **are used for the finding the probability associated with waiting for *k* number of events**, instead of just one event as is the case with exponential. The important parameters for Gamma are the *number of events to wait* (*k*) and the *rate parameter* (λ).

Example:

* You get to your favourite restaurant and there is 2 people in line ordering ahead of you. The mean order time at this restaurant is 2 minutes (or a *rate*of 0.5 customers per minute).
* How likely is it that you will get to begin placing your order in the next 4 minutes? 🍔



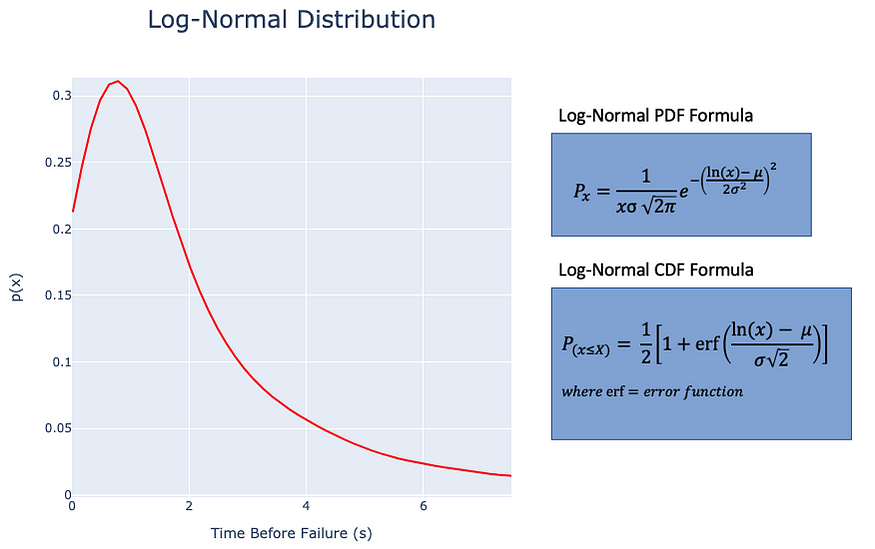
Gamma distribution — [learn more here](https://towardsdatascience.com/gamma-distribution-intuition-derivation-and-examples-55f407423840)

**4. Log-normal Distribution**

A log-normal distribution is a continuous distribution of random variables whose logarithms are distributed normally. **In other words, the lognormal distribution is generated by the function of eˣ, where x (random variable) is supposed to be normally distributed.**

Example:

* Let’s say you work as a quality control engineer ⚙️ You have to perform 10,000 stress tests of your product underneath a hydraulic press, to see how long it takes before it breaks/fails. The results from all the samples may look like the lognormal distribution below.



Log-normal distribution — [learn more here](https://brilliant.org/wiki/log-normal-distribution/)

Just knowing that your data is log-normal distributed is valuable, because we can easily translate log-normal data to a normal distribution using the *log(x)* function!

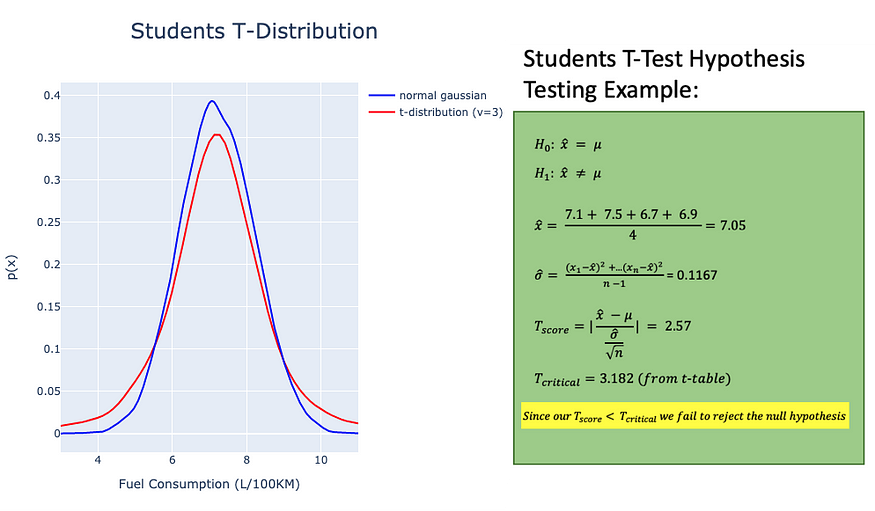
**5. Students T-Distribution**

The Students t-test distribution holds a similar shape to a Gaussian distribution, but is slightly shorter and wider. **It is useful in helping to make a comparison between groups when our sample size is small and/or the population standard deviation is unknown.**The important parameters for a Students t-distribution are the population mean (*μ*), sample mean (x̂), sample standard deviation (σ̂), and degrees of freedom (*n-1*).

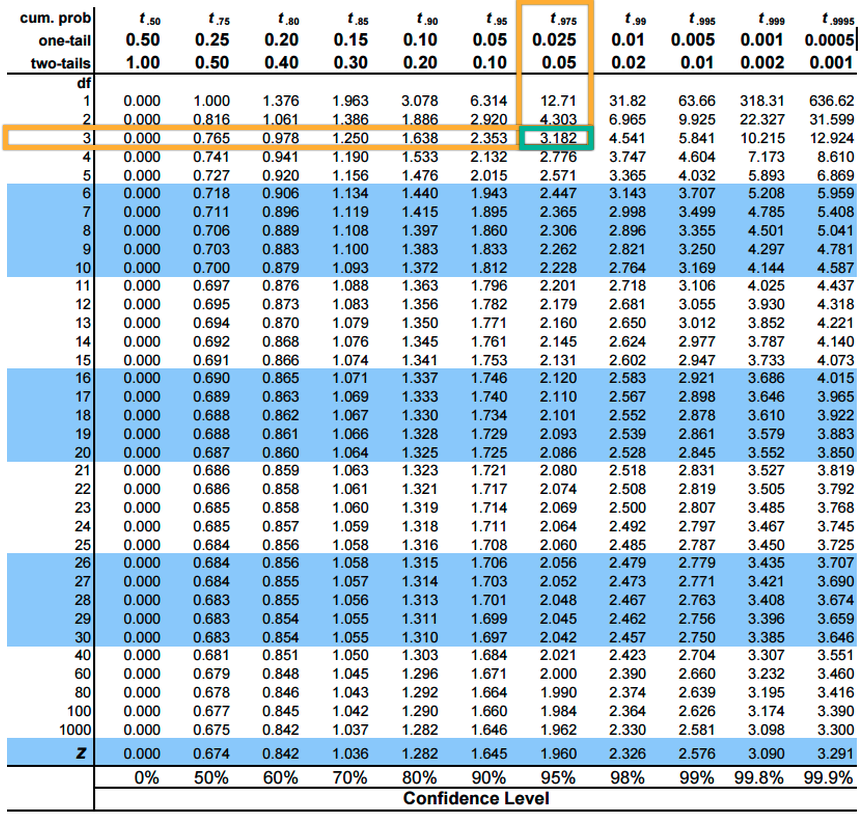
The t-distribution is most often used for calculations within the realm of [*hypothesis testing*](https://www.scribbr.com/statistics/t-test/#:~:text=A%20t%2Dtest%20is%20a%20statistical%20test%20that%20compares%20the,means%20is%20different%20from%20zero.)*,*as we will see in the example below!

Example:

* Suppose that a new car is introduced into the market and the manufacturer claims it has an average fuel consumption rating of 7.2L/100KM 🚗 You decide to go out and test 4 of these cars yourself, and you record their fuel ratings:  
  [7.1, 7.5, 6.7, 6.9]
* With the data you’ve collected, and using a 95% confidence interval, can you confirm the manufacturers claim?



Students t-distribution — [learn more here](https://stattrek.com/probability-distributions/t-distribution.aspx)



t-test table

Since we fail to reject the null hypothesis, we can state that from the findings in our data that we are 95% confident that the car manufacturers claim is true!

**Conclusion**

Understanding the distributions of data are important because they can give us insights and open the door to perform further statistical analysis. This article covers some of the most common data distributions, but it is by no means a comprehensive list.