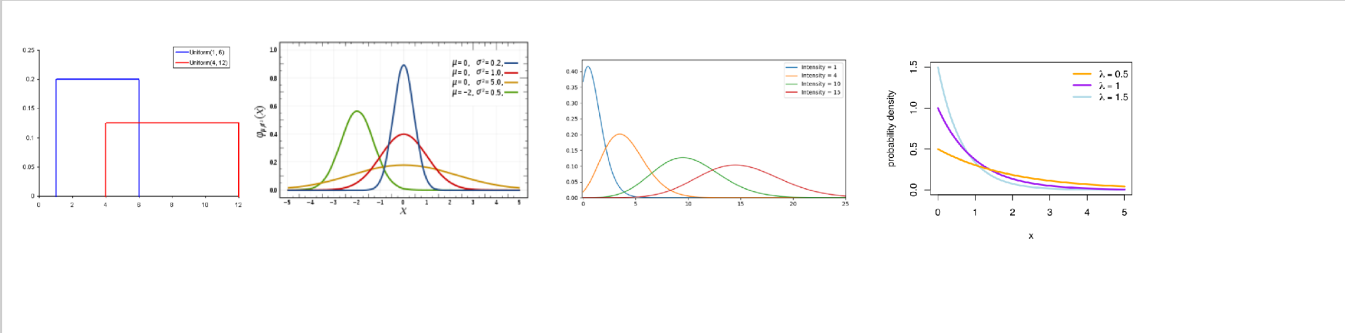
Probability Distributions Recap

We'll cover the following

Recap With Hints

We have seen six major probability distributions. Let's recap the relations among them so that we can get a better understanding of what to use and when to use it.



From left to right: Uniform, Normal, Poisson, Exponential

The first step is to identify whether we are dealing with a continuous or discrete random variable. Once that's done, here are some helpful hints to proceed from there:

- If we see a Normal (Gaussian) Distribution, we should go for it because there are many algorithms that, by default, will perform well specifically with this distribution; it is the most widely applicable distribution — ***it's called "Normal" for a reason!***
- A Poisson Distribution is similar to the Normal distribution, but with an added factor of skewedness. When the skewedness is low, a Poisson distribution will have a relatively uniform spread in all directions just like the Normal distribution.

- Binomial distribution when dealing with the number of successes in n trials.
- Bernoulli distribution is a special case of Binomial distribution with a single trial.
- Poisson is about the number of events in an interval of space or time.
- Exponential distribution is handy when dealing with time between events.
- Categorical variables can be easily interpreted using a Uniform distribution.

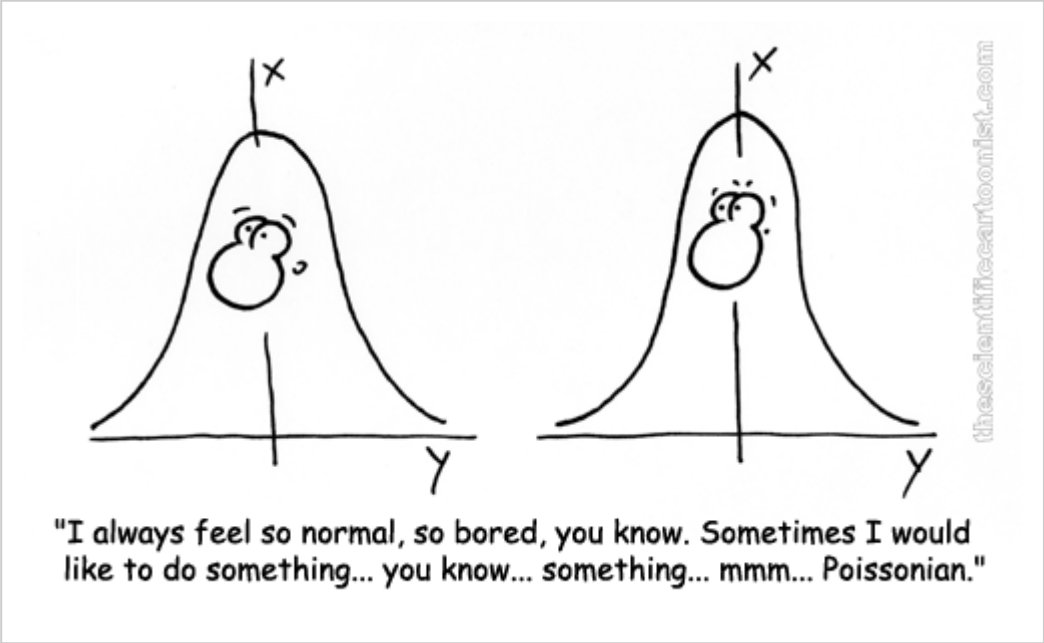


Image Credits: http://www.thescientificcartoonist.com/

- Out of the six distributions that we have seen, **make sure to have a firm grasp over at least the Uniform and Normal/Gaussian distributions**.

Before we stop talking about distributions, let's see some code examples for plotting distributions in Python. We can easily use the `distplot()` method from the *seaborn* library as follows:

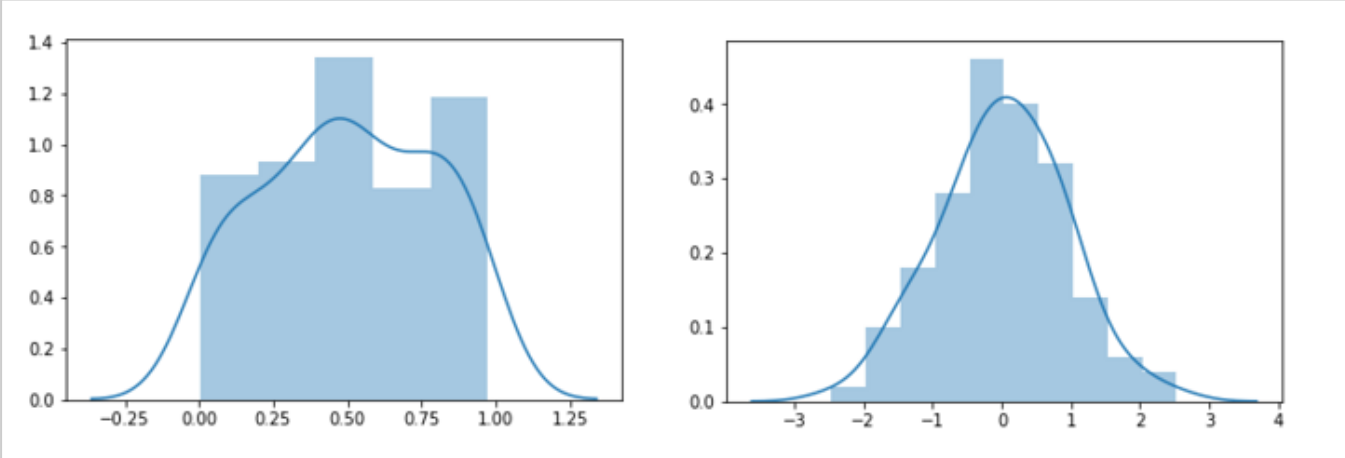
```
# Import libraries
import numpy as np
import seaborn as sns

# Create some random fake data
```

```
# Create some random fake data
x = np.random.random(size=100)


# Plot the distribution
sns.distplot(x);
```

Distribution plot for randomly generated data

```
# Create some data that follows a Normal Distribution
x = np.random.normal(size=100)

# Plot the distribution
sns.distplot(x);
```

Distribution plot for data that follows a Normal distribution



Distribution plots: Randomly distributed data (left) and Normally distributed data (right)

Final Thoughts

Probability distributions are a tool that you must have in your data scientist's toolbox, you will need them at one point or another! In these lessons we have done a deep dive in six major distributions and learned about their applications. Now, **you do not need to memorize their functions and all the nitty-gritty details**. However, it is important to be able to **identify, relate and differentiate among these distributions**.

Statistical significance is a term that is often thrown around without actually understanding what it means. In the next lesson we are going to learn about ***Statistical Significance, a wildly misunderstood concept.*** So, keep going! We are almost at the end of this section on Statistics! 👊