Statistics Workshop

Ivan Corneillet

Lead Data Science Instructor, Principal Data Scientist

Day 2: Basics of Statistics and Basics of Machine Learning

Day 2: Basics of Statistics and Basics of Machine Learning

- Random Variables
- Common Distributions
- Linear Regression
- Logistic Regression

Day 2: Basics of Statistics

"Statistics is the grammar of science." – Karl Pearson

Random Variables

Random Variables

A random variable X is an object that can be used to generate numbers, in a way that valid probabilistic statements about the generated numbers can be made.

E.g.,

$$P(X = 1) = .5$$

$$P(X = 1) = \frac{1}{12}$$

$$P(X > 6) = 0$$

are all probabilistic statements about a random variable X.

Random Variables

Random variables can be used to model real life events or measurements when there is some variation in their outcomes that we cannot account for. E.g.,

- Number of heads seen in ten flips of a quarter.
- Number of heads seen in ten flips of a dime.
- Number of buses that arrive late to a stop in San Francisco in a single day.
- Number of times my dog asks for food between 5 and 6 pm in a given day.
- Temperature on a mid-summers day in San Francisco.
- Rainfall in a mid-winter day in San Francisco.

Distribution

We naturally have a feeling that there is something **the same** about these two situations:

- Number of heads seen in ten flips of a quarter.
- Number of heads seen in ten flips of a dime.

What is it?

Distribution

We expect that the probabilities:

$$P(\text{5 heads in 10 flips of a quarter})$$

$$P(\text{5 heads in 10 flips of a dime})$$

are **equal**.

So are:

$$P(\text{2 heads in 10 flips of a quarter})$$

$$P(\text{2 heads in 10 flips of a dime})$$

and so on...

Distribution

The **sameness** that we sense between:

- Number of heads seen in ten flips of a quarter.
- Number of heads seen in ten flips of a dime.

is that the probabilities of all events like:

$$P(N \text{ heads in 10 flips of a quarter})$$

$$P(N \text{ heads in 10 flips of a dime})$$

are all equal.

Distribution

We summarize this by saying that the random variables:

- Number of heads seen in ten flips of a quarter.
- Number of heads seen in ten flips of a dime.

have the same distribution.

The **distribution** of a random variable is the pattern of all probabilities we assign to all outcomes of the random variable.

So two random variables have the same distribution if **they assign the same probabilities to all possible outcomes**.

In this case, we say that these random variables are **equally distributed**.

Common Distributions

Common Distributions

Some distributions are so common, they have been named and entered our shared statistical consciousness.

Discrete Distributions

Discrete Distributions

E.g.,

- Bernoulli
- Binomial
- Poisson

Bernoulli Distribution

Bernoulli Distribution

A single coin flip is an example of a Bernoulli distributed random variable.

The **Bernoulli distribution** describes any random event with only two possible outcomes.

The **probability mass function** (PMF) of the Bernoulli distribution

We can draw a picture of a distribution by letting the height of bars represent the probabilities of certain outcomes occurring:

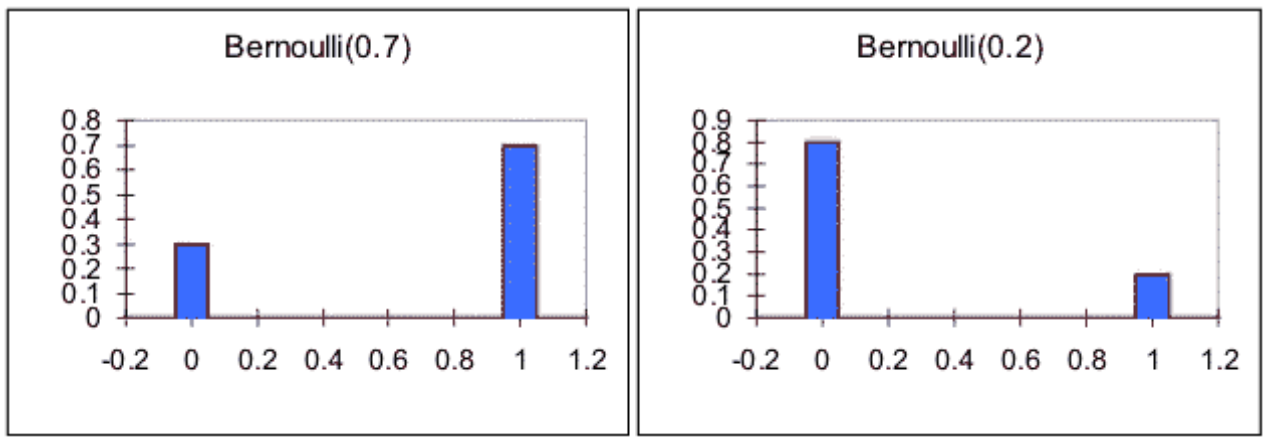

Figure 1:Bernoulli Distribution PMF

Notice how the first picture represents the occurrence of a **common event** and the second a **rare event**.

Bernoulli Distribution

Changing the probability that the event occurs changes the shape of the probability mass function. This is called **varying a parameter**.

Draw pictures of the probability mass functions of the following Bernoulli distributions:

- Flipping a fair coin.
- Rolling a six on a six-sided die.
- Rolling a twenty on a twenty-sided die.
- Rolling greater than a ten on a twenty-sided die.

Are any of these equally distributed?

Bernoulli Distribution

Binomial Distribution

Binomial Distribution

The two familiar random variables:

- Number of heads seen in ten flips of a quarter.
- Number of heads seen in ten flips of a dime.

Have a binomial distribution.

$$P(\text{2 heads in 10 flips of a quarter}) = \binom{10}{2} \times \left(\frac{1}{2}\right)^{10} = .044$$

The **binomial distribution** describes the number of events that happen in a fixed number of attempts when the events individually happen with the same probability.

Binomial Distribution

When we are flipping a fair coin, the heads happen with probability $\frac{1}{2}$, and

$$P(k \text{ heads in } n \text{ flips of a coin}) = \binom{n}{k} \times \left(\frac{1}{2}\right)^n$$

If the coin in unfair, so that the probability of an individual head is p, then

$$P(k \text{ heads in } n \text{ flips of an unfair coin}) = \binom{n}{k} \times p^k \times (1-p)^{n-k}$$

The binomial distribution has two parameters

- Number of attempts, usually called n.
- Probability the event occurs in a single attempt, usually called p.

Changing either n or p changes the shape of the binomial probability mass function.

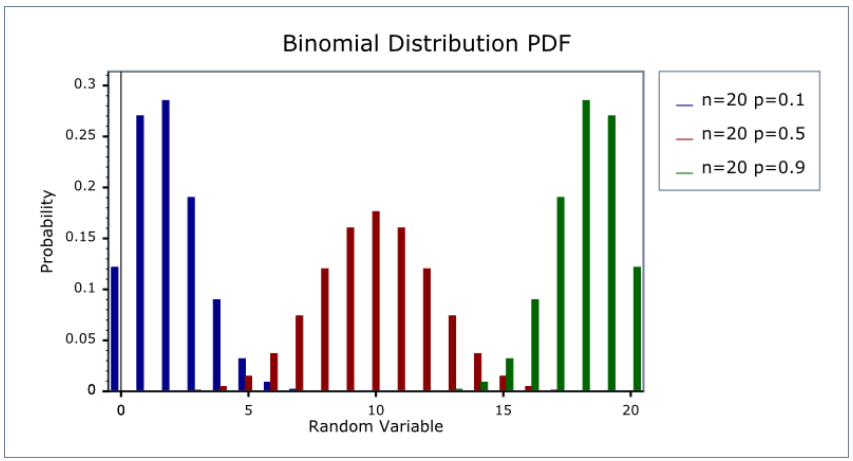

Figure 2:Binomial Distribution PMF

A critical hit is a roll of 20 on a 20-sided die. In a session of Dungeons and Dragons, you roll the die 20 times. What's the probability that you roll at least two critical hits?

A saving throw is a roll of at least 15 on a 20-sided die. In a session of Dungeons and Dragons you roll the die to attempt a saving throw 10 times. What's the probability you fail all of your saving throws?

Poisson Distribution

Poisson Distribution

- Number of buses that arrive late to a stop in San Francisco in a single day.
- Number of times my dog asks for food between 5 and 6 PM (when she is always fed) in a given day.

We see a similarity: they are both about the number of times an event happens in a given span of time (or space).

If we assume that the buses arrive at a fixed rate (but possibly unknown), and the dog barks at a fixed rate, then these are both examples of the **Poisson distribution**.

$$P(\text{dog barks } k \text{ times in one hour}) = e^{-\lambda}\frac{\lambda^k}{k!}$$

The λ above is the rate the event occurs.

Suppose we observe my dog bark 5 times in ten minutes. What's the probability that she will not bark at all in the next ten minutes?

Suppose we observe my dog bark 5 times in ten minutes. What's the probability that she will not bark at all in the next ten minutes?

The rate the dog barks is:

$$\lambda = \frac{5 \text{ barks}}{10 \text{ minutes}}$$

So using the Poisson equation

$$P(\text{dog barks zero times in ten minutes}) = e^{-5} \times \frac{5^0}{0!} = .007 = .7\%$$

What's the probability the dog barks zero times in the next hour?

What's the probability the dog barks zero times in the next hour?

$$\lambda = \frac{5 \text{ barks}}{10 \text{ minutes}} = \frac{30 \text{ barks}}{\text{hour}}$$

$$P(\text{dog barks zero times in the next hour}) = e^{-30} \times \frac{30^0}{0!} = 9.3 \times 10^{-14}$$

It's basically extremely improbable...

What's the probability the dog barks at least twice in the next twenty minutes?

In a batch of cookie batter, you dump in 100 chocolate chips and then mix the result thoroughly. You then portion out 20 equally sized cookies. What's the probability that at least one cookie has no chocolate chips in it?

General strategy for solving problems using probability distributions

- Use the information in the problem statement to determine
  - a likely distribution for the quantity of interest and
  - the values to use for the parameters of this distribution.
- Use the probability function of the distribution to compute the needed probability.

You may have to break the problem down into multiple steps to succeed. Practice and you'll start to see common patterns!

Continuous Distributions

Continuous Distributions

The distributions we discussed are **discrete**: the outcomes can only be individualized numbers (0, 1, 2, 3, . . .).

There are also **continuous distributions**. E.g.,

- Normal
- Uniform
- Exponential

Normal Distribution

Normal Distribution

The normal distribution usually shows up due to the **Central Limit Theorem**. (CLT)

Parameters: Mean μ and standard deviation σ.

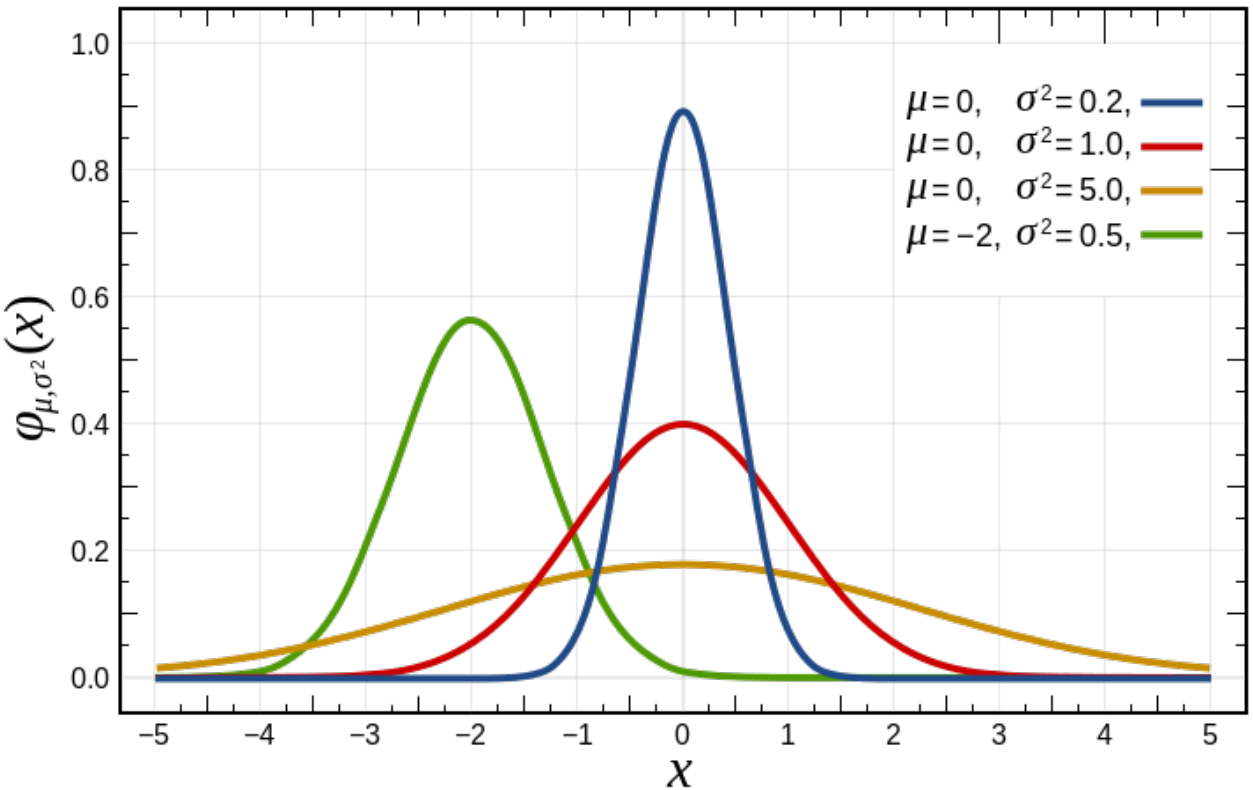

Figure 3:Normal Distribution PDF

Uniform Distribution

Uniform Distribution

The uniform distribution shows up when a random event can take any value in a range, each result being equally likely.

Parameters: Minimum a and maximum b.

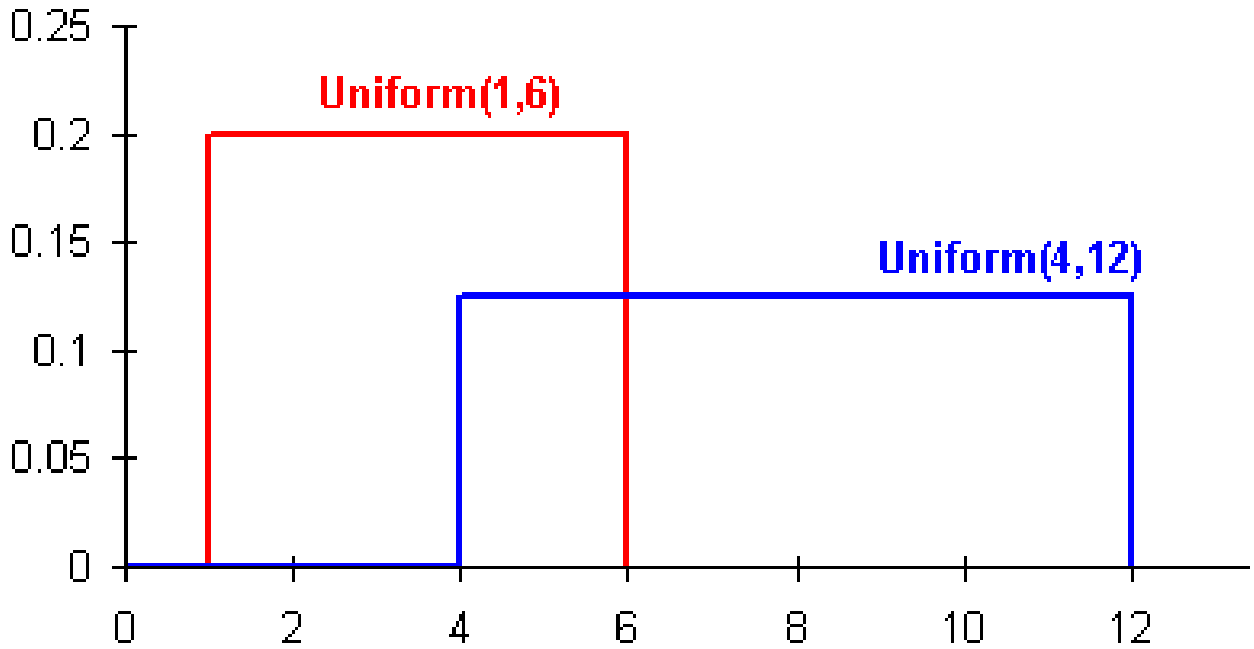

Figure 4:Uniform Distribution PDF

Exponential Distribution

Exponential Distribution

The exponential distribution describes the time you have to wait before observing an event when the events happen at a fixed rate.

Parameters: Rate α.

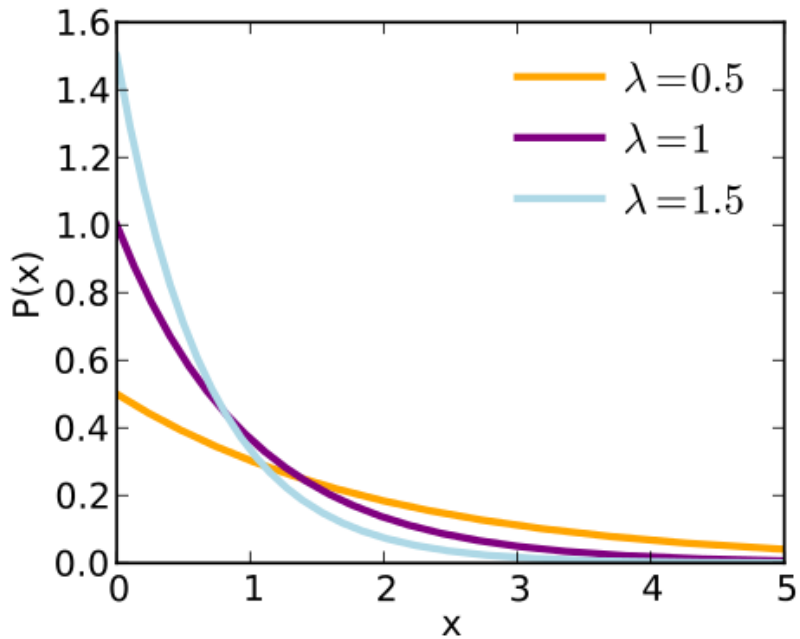

Figure 5:Exponential Distribution PDF