Probability and Statistics
Section 3: Random Variables and Probability Distribution

1. Random Variable

- A random variable is a function that associates a real number with each element in the sample space.
- The outcome probabilities must be between 0 and 1 and have sum 1.

1. a. Discrete random variables

A discrete random variable X is some uncertain quantity that can take on a countable number of values in $(x_1, x_2, \ldots, x_n)$. $P(X = x_i)$ or $P(x_i)$ is the probability that the random variable X takes on value x_i

A random variable X can be the event outcome of throwing a dice, which can take 6 different values: $\{1,2,3,4,5,6\}$

This random variable has to mantain the previous event identities:

$$0 \leq P(X = x_i) \leq 1$$

$$\sum_{x_i} P(X = x_i) = 1$$

1.b. Continuous random variables

A continuous random variable $\boldsymbol{X}$ is some uncertain quantity that can take on an uncountable range of values, and assumes each value in this range with probability zero.
A random variable $\boldsymbol{X}$ can be a range of temperatures between the interval $[\mathbf{0}, \mathbf{100}]$ for example.

It is characterised by a **probability density function** $\boldsymbol{p}(\boldsymbol{x})$or $\boldsymbol{p}(\boldsymbol{X} = \boldsymbol{x})$ which can help us calculate, for example, the probability that the temperatures lie between $\mathbf{34.0}$ and $\mathbf{39.0}$ degrees:

$$P(34.0 \leq X \leq 39.0) = \int_{34.0}^{39.0} p(x)dx$$

Initially, we don't know the shape of this function $p(x)$, if it's linear, quadratic, trigonometric, etc. But in the next section, we will see some examples of the typical ones.

Like in the discrete variables, the sum of all the possible outcomes must be one. In this case, the sum is calculated by an integral over all possible values of x:

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

There's also the **cumulative distribution function $\boldsymbol{F(x)}$**, which can tell us if a random variable is at most a certain value:

$$F(x) = P(X \leq x)$$
$$P(a < X \leq b) = F(b) - F(a)$$

There are two ways of characterizing a probability distribution, discrete or continuous: the mean or expected value, and its variance.

1.c. Expected values

The **expected value** of a variable $\boldsymbol{X}$ can be seen as its mean or average value.

For discrete random variables:

$$E[X] = \sum x\left(p(x)\right)$$

and for continuous random variables:

$$E[X] = \int_{-\infty}^{\infty} x\, p(x) dx$$

This value is a weighted sum of the values that X can take on, where the weights are the probabilities of those respective values. It can also be called the centre of mass of our values

For the dice example, we've already seen that each outcome's probability equals $1/6$.

The value we expect would be halfway between the possible values the dice can take. Let's check it with the formula

$$E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{21}{6} = 3.5$$

So the expectation is a value in the middle between 1 and 6.

1.d. Variance

If expected value is the center of mass of our values of X, the **variance** $Var(X)$ is the spread around that center:

$$Var(X) = E[(X - E[X])^2]$$

This expression evaluates the distance from each outcome X from the average value $E[X]$. That's why it uses the difference $\boldsymbol{X - E[X]}$ to compute the variance.

Using algebraic properties, it can be simplified as

$$Var(X) = E[X^2] - E[X]^2$$

For the dice example, we need to calculate two terms:

$$E[X]^2 = (3.5^2) = 12.25$$

$$E[X^2] = 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6} = \frac{91}{6}$$
$$= 15.17$$

Then, the variance

$$Var(X) = 15.17 - 12.25 = 2.92$$

Standard deviation is another powerful value. It represents the same property as the variance, but with the same units as the variable per se:

$$\sqrt{Var(X)} = \sqrt{2.92} = 1.71$$

1.e. Variance vs Standard Deviation

Variance is the average squared deviations from the mean, while standard deviation is the square root of this number. Both measures reflect variability in a distribution, but their units differ:

- Standard deviation is expressed in the same units as the original values (e.g., minutes or meters).
- Variance is expressed in much larger units (e.g., meters squared).

Although the units of variance are harder to intuitively understand, variance is important in statistical tests.

1..f. Covariance

Let's go one step further with random variables. Let's imagine we have two random variables, X and Y, and we want to know how related they are. **Covariance** is a measure of the linear relationship between X and Y:

$$\begin{aligned} CoV(X,Y) &= E[\ \ (X - E[X])(Y - E[Y])\ \ ] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

It is easy to see that covariance of a variable with itself is $CoV(X,X) = Var(X)$

Correlation:

A good indicator of the relationship between X and Y is **correlation**, which is nothing more than covariance normalized:

$$\rho(X,Y) = \frac{CoV(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}}$$

Then, correlation gives a number between 1 and -1. Two variables are said to be uncorrelated if

$$CoV(X,Y) = 0 \ \rightarrow \ \rho(X,Y) = 0$$

2. Probability distributions

2.a. The Uniform Distribution

This distribution gives the same probability p to all values of a random variable X, between a range $[a, b]$. It's like we brought the dice example to an uncountable number of possible values of X. Then,

- For $a \leq x \leq b$: $P(X = x) = p$
- For $a < x \; and \; x > b$: $P(X = x) = 0$

For example, for a random value between range $[2,6]$ that has to mantain that the sum of all the probabilities of the posible values has to be 1:

$$\int_{2}^{6} p(x)dx = \; 1$$

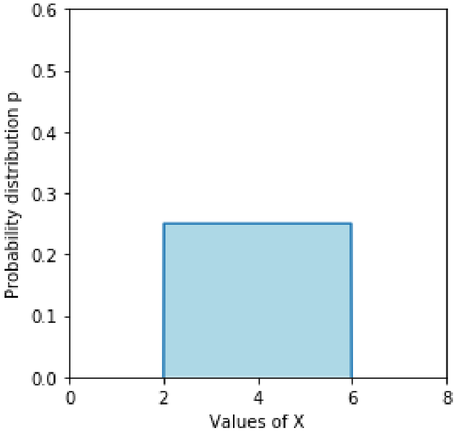
and that probability distribution is a constant $p(x) = p$, then

$$\int_{2}^{6} p \; dx = px|_{2}^{6} = p.(6 - 2) = 1$$

Solving the equation,

$$p = \frac{1}{6 - 2}$$

Plotting these probability distributions is very useful because it allows us to calculate probabilities graphically.



We already know that the probability that X takes a value between 2 and 6 equals 1:

$$P(2 \leq X \leq 6) = 1$$

If we would want to know the probability that X takes a value between 4 and 6. we would just have to calculate the area of the rectangle delimited in x by 4 and 6 (Area of a rectangle):

$$\begin{aligned} P(4 \leq X \leq 6) &= \text{rectangle width} \cdot \text{rectangle height} \\ &= (6-4) \cdot (p-0) = 2 \cdot 0.25 = 0.5 \end{aligned}$$

Then, there's a 50% chance that a value of X lies within the range $[4{,}6]$

2.b. The Cumulative Distribution

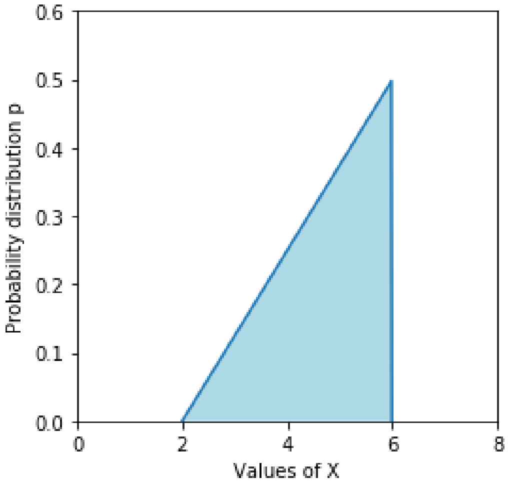
his distribution gives more probability to the highest values of X. The probability increases linearly with X, between the range $[a, b]$

For example, for a random value between the range [2,6] that has to mantain that the sum of all the probabilities of the possible values has to be 1:

$$\int_{2}^{6} p(x)dx = 1$$

$$\int_{2}^{6} \mathrm{m}(x\text{-}2)dx \;=\; m\; \cdot \frac{(x-2)^2}{2} |_2^6 \;=\; 1$$

$$m = \frac{1}{\frac{(6-2)^2}{2}}$$



With this function, if we would want to know the probability that X takes a value between 2 and 4 we would have to calculate the area of the triangle (Area of the triangle):

$$\begin{aligned} P(2 \leq X \leq 4) &= \frac{1}{2} \cdot \text{triangle base} \cdot \text{triangle height} \\ &= \frac{1}{2} \cdot (4-2) \cdot \big(p(x=4) - p(x=2)\big) \\ &= \frac{1}{2} \cdot 2 \cdot (0.125 * (4-2) - 0) = 0.25 \end{aligned}$$

Then, there's a 25%chance that a value of X lies within the range [2,4].

2.c. The Gaussian Distribution

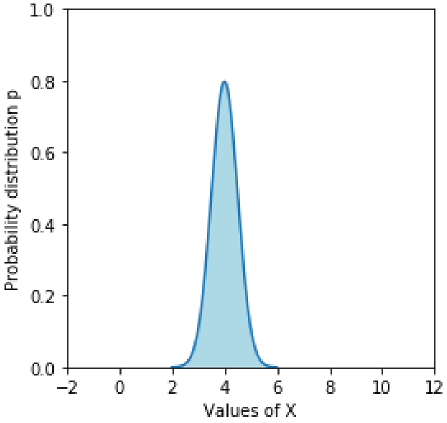
The Gaussian distribution is also known as the Normal Distribution. It is a continuous distribution, with a high peak in the middle, and lower values on the extrema. A random variable X following this distribution is often expressed as

$$X \sim \mathcal{N}(\mu, \sigma)$$

where μ is the mean of the X values, and σ is the standard deviation of the X values

The density probability distribution p(x) can be written as:

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



This distribution also maintains that the sum of probabilities is 1. As you can see on the above plot, a normal distribution with $\mu = 4$ has the highest probability at that point, $X = 4$, meaning that it is the most likely, followed by the points around it. With $\sigma = 0.5$, it is highly unlikely that any point X gets a value lower than $X = 2$ or higher than $X = 6$. As σ increases, the sparser the distribution is, and it is more likely to find values far away from the mean $X = 4$.