Refresher on Statistics and Probability Distributions

Dr. Matias Valdenegro

March 15, 2022

Today's Agenda

1. Probability
2. Probability Distributions
3. Common Probability Distributions

Outline

Definition

A probability is defined for an event as the ratio of possible ways it can happen to the total number of outcomes.

$$P(E) = \frac{\text{\# of ways } E \text{ can happen}}{\text{\# of total outcomes}} \tag{1}$$

Probabilities are real numbers in the $[0, 1]$ range, and are interpreted as likelihood or chance that the event will happen.

A simple example can be made with a six-sided die, P(D is odd) = P(even) = 1/2, P(D = 1) = 1/6, P(D > 2) = 4/6, etc.

Event Properties

Mutual Exclusivity

Events A and B are mutually exclusive if one happens, the other cannot happen. This is equivalent to:

$$\mathbb{P}(A \text{ or } B) = \mathbb{P}(A) + \mathbb{P}(B) \tag{2}$$

Independence

Events A and B are independent if one happening does not affect the other. This is equivalent to:

$$\mathbb{P}(A \text{ and } B) = \mathbb{P}(A)\mathbb{P}(B) \tag{3}$$

Conditional Probability

It is the probability of an event A given that another event B has already occurred, which takes the assumption that both events have some sort of relationship.

This is denoted by $\mathbb{P}(A \mid B)$ and computed by definition as:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{4}$$

Note that in general $\mathbb{P}(A) \neq \mathbb{P}(A \mid B) \neq \mathbb{P}(B \mid A) \neq \mathbb{P}(B)$.

Only if $\mathbb{P}(A \mid B) = \mathbb{P}(A)$ and/or $\mathbb{P}(B \mid A) = \mathbb{P}(B)$, this indicates that A and B are independent.

Basic Properties

Event	Probability
E	$\mathbb{P}(E) \in [0, 1]$
not E	$\mathbb{P}(\text{not } E) = 1 - \mathbb{P}(E)$
A or B	$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
	$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ if A and B are mut. exclusive.
A and B	$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B) = \mathbb{P}(B \mid A)\mathbb{P}(A)$
	$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ if A and B are independent.
A given B	$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}$ (Bayes Rule)

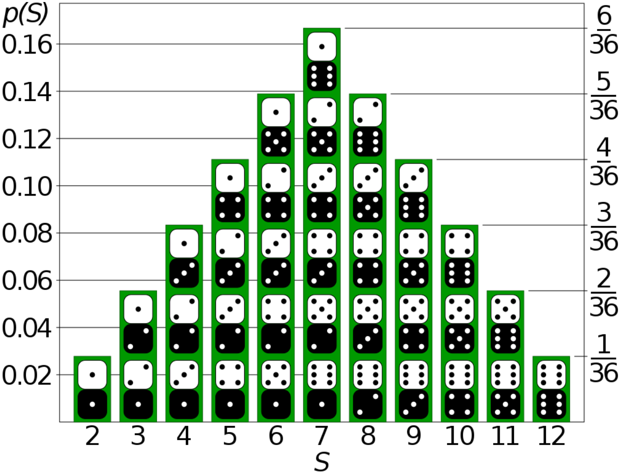
Outline

Concept

A probability distribution is a function that maps from some input variable to probability (discrete) or probability density (continuous) values, as a way to characterize how probability is distributed on the variable.

The input variable is called a *Random Variable*, because its value depends on a random outcome or randomness. The possible values of the variable are events, which are subset of a *sample space*, which is any predefined set (of any kind, real numbers, matrices, vectors, etc).

Concept - Distribution of Sum of 2 Six-sided Dies



Source:
`https://en.wikipedia.org/wiki/Probability_distribution`

Random Variables

A random variable is a mathematical concept used to model randomness, where it is a object which depend on a random event.

Sample Space
: A random value has a set of allowable values, usually called sample space and denoted by $\mathcal{A}$.

Mapping
: A random variable is a mapping $X : \mathcal{A} \to \mathbb{R}$, where $\mathcal{A}$ are possible outcomes in the sample space, mapped to real numbers (their probability).

Random variables can be discrete or continuous, which defines their sample space.

Probability Function (PFs)

The PF is a function that usually defines a probability distribution, and is defined as:

$$f_X(x) = \mathbb{P}(X = x) \tag{5}$$

This is very similar to the mapping X we defined previously in a random variable. This definition works well for discrete random variables, but it is problematic for continuous ones, because $\mathbb{P}(X = x) = 0$ for all x.

For discrete distributions, this is usually called the probability mass function.

Kolmogorov Axioms

All probability distributions need to follow these axioms:

Probability values are between 0 and 1

$$0 \leq \mathbb{P}(X \in E) \leq 1 \qquad \forall E \in \mathcal{A}$$

Sum of all events is 1

$$\textstyle\sum_{X \in \mathcal{A}} \mathbb{P}(()X) = 1 \qquad \int_{\mathcal{A}} P(X) dX = 1$$

Disjoint Family of Sets

$$\mathbb{P}(X \in \cup_i E_i) = \textstyle\sum_i \mathbb{P}(X \in E_i)$$

For any disjoint family of sets $E_i \in \mathcal{A}$

Probability Density Function (PDFs)

For continuous probability distributions, the PDF is a function that usually defines a probability distribution, and is defined as $f_X(x)$ that follows:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)dx \tag{6}$$

For any $a \leq b$. PDFs model probability density instead of plain probability, so they can (and will be) bigger than one. This does not violate Kolmogorov's Axioms.
Note that if $a = b$, then:

$$\mathbb{P}(a \leq X \leq a) = \int_a^a f_X(x)dx = 0 \tag{7}$$

Cumulative Distribution Function (CDFs)

The CDF is a function defined by:

$$F_X(x) = \mathbb{P}(X \leq x) \tag{8}$$

This function gives the probability that a random variable X is less than a value x. A special property is:

$$\mathbb{P}(a < X \leq b) = F(b) - F(a) \tag{9}$$

Continuous CDFs

For continuous random variables, there are some additional dualities:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du \tag{10}$$

$$f_X(x) = \frac{d}{dx}F_X(x) \tag{11}$$

The CDF and PDF are related through the integral/derivative of each other.

Inverse Cumulative Function or Quantile Function

$$F^{-1}(q) = \inf\{x : F(x) > q\} \tag{12}$$

Where $q \in [0, 1]$ is a specific quantile. The **inf** method can be conceptually thought as taking the minimum.

If $F(x)$ is a strictly increasing function, then $F^{-1}(q)$ is the unique value x so $F(x) = q$ holds.

Quantiles are equal divisions of the probability space, some have special names, for example percentiles divide in 100 divisions, or quartiles over 4 divisions. The percentile 50 is always the median ($F^{-1}(0.5)$).

Expectation

Expectation or expected value is a linear operation defined for continuous distributions with PDF $X \sim f_X$ as:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x) dx \tag{13}$$

And for discrete distributions as:

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} x_i\, f_X(x_i) \tag{14}$$

The expected value $\mathbb{E}[X]$ is associated with the mean, while the variance can be computed as $\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

Covariance

Covariance is a measure of the joint variation between two related variables X and Y, and is defined as:

$$\mathrm{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \tag{15}$$

Not to be confused with the covariance matrix. If covariance is normalized with the standard deviation of each variable (σ_X, σ_Y), you obtain the correlation:

$$\mathrm{CORR}(X, Y) = \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sigma_X \sigma_Y} \tag{16}$$

Transformations on Random Variables

Some important identities and transformations on random variables X and Y:

$$
\begin{aligned}
\mathrm{Var}[X+Y] &= \mathrm{Var}[X] + 2\mathrm{Cov}[X,Y] + \mathrm{Var}[Y] \\
\mathrm{Var}[X-Y] &= \mathrm{Var}[X] - 2\mathrm{Cov}[X,Y] + \mathrm{Var}[Y] \\
\mathrm{Var}[X] &= \mathrm{Cov}[X,X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2
\end{aligned}
$$

For a constant k:

$$
\begin{aligned}
\mathrm{Var}[X+k] &= \mathrm{Var}[X] \\
\mathrm{Var}[kX] &= k^2\mathrm{Var}[X]
\end{aligned}
$$

Some Special Transformations and Laws

Law of Total Variance

$$\mathrm{Var}[Y] = \mathbb{E}[\mathrm{Var}[Y \mid X]] + \mathrm{Var}[\mathbb{E}[X \mid Y]] \tag{17}$$

Law of Total Expectation

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]] \tag{18}$$

Addition of Probability Distributions

One important result in statistics is the addition/sum of two random variables, which is the convolution of their PDFs. If we have $X \sim f_X$, $Y \sim f_Y$, then $Z = X + Y$ has PDF:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx \tag{19}$$

Transformations on Random Variables

Passing random variables through a function is in general difficult to do, for example, if we have X a distribution with PDF f_X, and a function g that is increasing in the domain of X, then the PDF of $Y = g(X)$ is given by:

$$f_Y(y) = f_X(g_{-1}(y)) \left| \frac{d}{dy} g_{-1}(y) \right| \tag{20}$$

This is of course only valid for continuous distributions. For a discrete distribution X with probability mass function f_X, then the transformation is:

$$f_Y(y) = \sum_{x \in g_{-1}(y)} f_X(x) \tag{21}$$

Outline

Normal or Gaussian Distribution

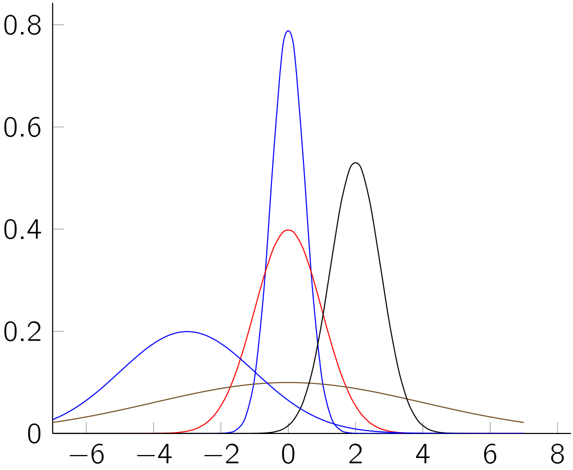
Is a continuous distribution defined for $x \in \mathbb{R}$ with two parameters, mean μ and variance σ^2, and with PDF given by:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-0.5(\frac{x-\mu}{\sigma})^2} \tag{22}$$

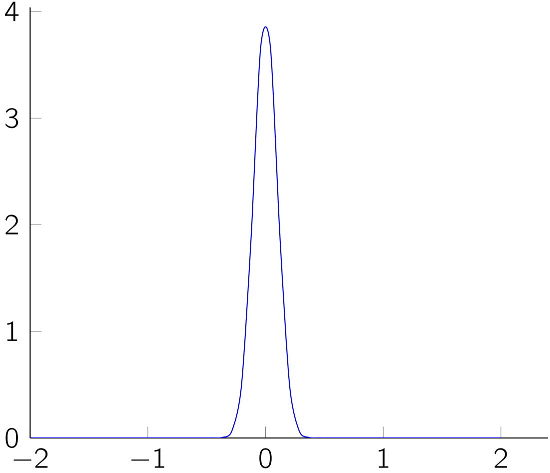
The expectation and variance of the Gaussian distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ is:

$$\mathbb{E}[X] = \mu$$
$$\mathrm{Var}[X] = \sigma^2$$

Gaussian Distribution PDF Examples



Gaussian - Something Wrong?



Uniform Distribution

Continuous or Discrete distribution where all values in a range $[a, b]$ are equally likely. It is parameterized by a and b and has PDF/mass fn as:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

It is usually denoted as $X \sim U(a, b)$.

The mean is $\mathbb{E}[X] = 0.5(a + b)$ and variance is $\mathrm{Var}[X] = \frac{1}{12}(b - a)^2$.

Question...

Why is the Gaussian Distribution so important and fundamental?

Central Limit Theorem

Given a sequence of random variables $X_1, X_2, ..., X_n$ that are independent and identically distributed with population mean μ and variance σ^2. First let's define the sample mean or average:

$$\bar{X}_n = \frac{X_1 + X_2 + ... + X_n}{n} \tag{24}$$

The central limit theorem states that as $n \to \infty$, then the distribution of $\bar{X}_n$ is a Gaussian distribution:

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \tag{25}$$

That is the variance of the sample mean decreases with $\frac{1}{n}$, while the standard deviation decreases with $\frac{1}{\sqrt{n}}$.

Random Numbers given a Probability Distribution

Sometimes it is needed to generate numbers not from a uniform distribution, but following another distribution with a given CDF or PDF.

Inversion Method

If we generate $U \sim \text{Uniform}(0, 1)$, then with the following computation using the target CDF F_X:

$$X = F_X^{-1}(U) \tag{26}$$

X will follow the distribution P defined by CDF F_X. This only works if F_X is invertible.

Random Numbers given a Probability Distribution

Rejection Sampling

This method works using a proposal distribution Y that has PDF g, to generate samples of distribution X with PDF f. It works like this:

1. Generate sample y from Y and a sample u from $\text{Uniform}(0, 1)$.
2. If $u < \frac{f(y)}{Mg(y)}$
   - Accept y as a sample drawn from X.
3. If not, reject y and return to step 2.

M is a constant that fulfills $f(x) < Mg(x)$, and $1 < M < \infty$. It is selected as a scale to make sure f and g are compatible, and also decides the accept rate $\frac{1}{M}$, so on average a sample is generated every M iterations of this algorithm.

A Note on Probability Distributions

In general probability distributions are a difficult concept to grasp, this is because:

- They are abstract mathematical objects that model (ideal) behavior in the world.
- The only way to get a number from a probability distribution is to generate a sample from it. The key concept is *sampling*.
- Their behavior can only be seen through obtaining multiple samples and making a histogram (to visualize the distribution).

Importance for the Course

- The concepts presented here are the base for the rest of the course.
- In particular Probability Distributions are key for Uncertainty Quantification in Machine Learning...
- Sampling s also a key concept as a lot of UQ methods rely on sampling to produce a distribution.

Questions?

Bibliography I