Notes on Probability Theory

George Baker

Probability Spaces

The Sample Space Ω

Ω (capital omega) is the the set of all possible outcomes of the experiment.

E.g. two successive coin tosses:

$\Omega = \{hh, tt, ht, th\}$

The number of possible outcomes in this scenario is (obviously) four. It's the length of the set Ω.

Formally, this is defined by the cardinality of Ω which is denoted as $|\Omega|$. I.e. $|\Omega| = 4$.

The Event Space A

A (capital alpha) is the set of possible events.

E.g. if we roll a six-sided die, what is the probability the number on the die is odd?

$\Omega = \{1, 2, 3, 4, 5, 6\}$

$A = \{1, 3, 5\}$

Probability

It follows that for any set of possible events A, $P(A) = \frac{|A|}{|\Omega|}$.

E.g. following on from the example above:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$
$$|\Omega| = 6$$
$$A = \{1, 3, 5\}$$
$$|A| = 3$$

$$P(A) = \frac{|A|}{|\Omega|}$$
$$= \frac{3}{6}$$

The Kolmogorov Axioms

The Kolmogorov axioms are the foundations of probability theory introduced by Andrey Kolmogorov in 1933.

First Axiom

The probability of an event E is a non-negative real number:

$$P(E)\, \epsilon\, \mathbb{R},\ P(E) \geq 0$$
$$\forall E\, \epsilon\, A$$

Second Axiom

The probability that at least one of the events in the sample space Ω will occur is 1.

$$P(\Omega) = 1$$

Third Axiom

If events $\{E_1, E_2, ...\}$ are mutually exclusive (cannot occur at the same time), then $P(E_1 \cup E_2 \cup ...)$ is $P(E_1) + P(E_2) + ... + P(E_n)$

$$P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$$

Where $\bigcup$ is the big capital version of $\cup$.

Mutual Exclusivity

Two events are mutually exclusive (disjoint) if they cannot both occur at the same time. E.g. if we toss a coin, we cannot get heads and tails at the same time.

Types of Probability Spaces

Discrete Probability Spaces

Discrete data can only take certain values. E.g.

- the number of students in a class;
- the results of rolling two dice.

In a discrete probability space, Ω is **finite**.

This makes sense because there are always a finite number of students in a class, and a finite number of results you can get from rolling two dice.

Calculations on a discrete probability space involve summations i.e. $\sum$.

Continuous Probability Spaces

Continuous data can take any value (within a range) e.g.

- a person's height could be any value (within the range of human heights);
- a dog's weight (within the range of dog weights);
- the length of a leaf;
- a person's IQ.

In a continuous probability space, Ω is **infinite**.

This makes sense because there are an infinite number of heights a human could have, and an infinite number of lengths a leaf could have.

Calculations on a continuous probability space involve integrals i.e. $\int$.

Conditional Probability

In probability theory, conditional probability is a measure of the probability of an event E_2 occurring, given that another event E_1 has already occurred.

Given two events A and B, the conditional probability of A given B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Join Probability

Given two events A and B, the joint probability of A and B is

$$P(A, B) = P(A \cup B) = P(A)P(A|B)$$

Law of Total Probability

Let $\{E_1, E_2, ...\}$ be finite events in Ω. Let A be any event.

$$P(A) = \sum_{i=1}^{n} P(A \cup E_i) = \sum_{i=1}^{n} P(E_i)P(A|E_i)$$

Independence

Two events A and B are independent if

$$P(A \cup B) = P(A)P(B)$$

Independence and mutual exclusion are not the same thing. That's to say, two events might not be able to happen at the same time, but that does not necessarily mean those events are independent.

Random Variables

What is a Random Variable?

A random variable can be described *informally* as a variable whose values depend on outcomes of a **random** phenomenon (hence the name).

Formally, in probability theory, a random variable is a function defined on a probability space that maps from the sample space Ω to $\mathbb{R}$.

By way of an example, let's flip a coin. We can use the greek letter ω (lower case omega) to denote outcomes: heads as ω_H, and tails as ω_T.

Let's say that the random variable X takes one of these outcomes, and outputs a real number. We'll choose $1 \, \epsilon \, \mathbb{R}$ for heads and $0 \, \epsilon \, \mathbb{R}$ for tails:

$$X(\omega_H) = 1$$
$$X(\omega_T) = 0$$

Discrete Random Variables

A discrete random variable is a random variable over a discrete probability space. I.e. its value is obtained by counting.

E.g. if we have a jar of marbles with 5 red marbles, and 7 blue marbles, and we pick 3 marbles out of the jar, we can use the discrete random variable X to map the outcomes to a numerical value.

In this case, let's say we're interested in the number of red marbles.

ω_0: zero red marbles recorded,

ω_1: one red marble recorded,

ω_2: two red marbles recorded,

ω_3: three red marbles recorded.

$$X(\omega_0) = 0$$
$$X(\omega_1) = 1$$
$$X(\omega_2) = 2$$
$$X(\omega_3) = 3$$

Continuous Random Variables

A continuous random variable is a random variable over a continuous probability space. I.e. its value is obtained by measuring.

E.g. let's measure the heights of students in a class. We can use the continuous random variable X to map the heights to a numerical value.

But unlike discrete random variables (which are countable), we cannot be as precise with our outcomes.

ω_0: height is between 150cm and 160cm,

ω_1: height is between 160 and 170cm.

$$X(\omega_0) = 0$$
$$X(\omega_1) = 1$$

Codomain

The codomain is the set of all **possible** output values of a function.

For all random variables, the codomain is the set of real numbers $\mathbb{R}$.

Range

The range is the set of **actual** output values of a function.

I.e. the range is the set of outputs $\{x_0, x_1, ...\}$ denoted as R_X.

If we again use the example of a coin flip with $X(\omega_H) = 0$, and $X(\omega_T) = 1$, the range $R_X = \{0, 1\}$.

NB: If X is a discrete random variable, then its range R_X is **countable**.

Probability Mass Function

A probability mass function (pmf) is a function that gives the probability that a discrete random variable X is exactly equal to some value x.

Definition

$$pX(x_i) = P(X = x)$$

The probability mass function is often the primary means of defining a discrete probability distribution.

A pmf satisfies the following three properties:

1. $\sum pX(x_i) = 1$
2. $pX(x_i) > 0$
3. $p(x) = 0$ for all other x

Thinking of probability as mass helps to avoid mistakes since the physical mass is conserved as is the total probability for all hypothetical outcomes x.

Discrete Probability Distributions

A probability distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment.

A discrete probability distribution is a probability distribution that can take on a countable number of values.

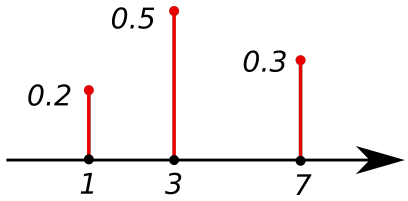


Figure 1: The probability mass function of a discrete probability distribution. The probabilities of the singletons {1}, {3}, and {7} are respectively 0.2, 0.5, 0.3. A set not containing any of these points has probability zero.

Examples of Discrete Probability Distributions

- Poisson Distribution
- Bernoulli Distribution
- Binomial Distribution
- Geometric Distribution
- Discrete Uniform Distribution

Cumulative Distribution Function

A cumulative distribution function (cdf) is a function that gives the probability that a random variable (discrete or continuous) X will take a value less than or equal to x.

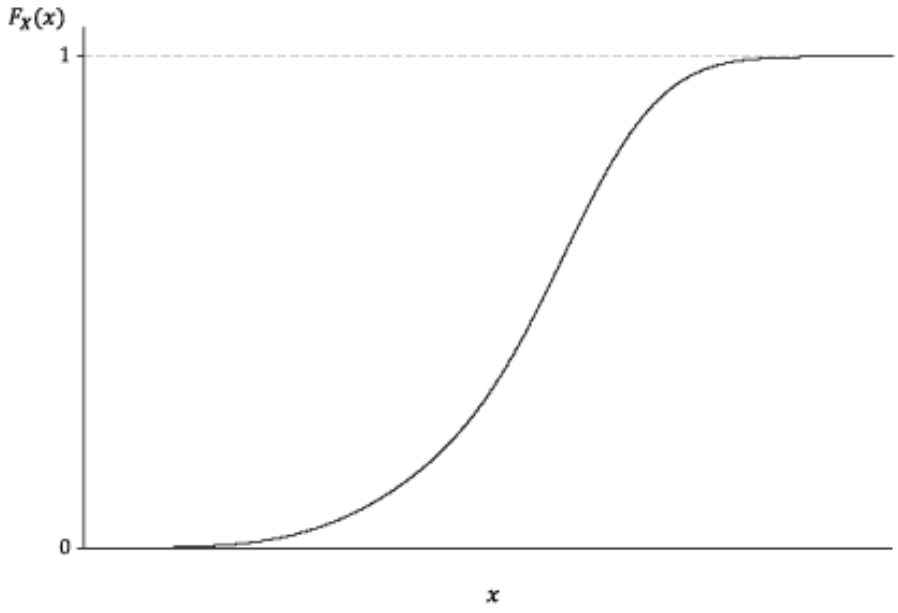
Definition

The cumulative distribution function of a random variable X is the function given by:

$$F_X(x) = P(X \leq x)$$

The probability that X lies in the interval (a, b), where $a < b$ is:

$$P(a < X \leq B) = F_X(b) - F_X(a)$$



Let's say that $0 \leq x \leq 100$.

As x increases, so does $F_X(x)$ i.e. the cumulative probability.

As x approaches one hundred, the cumulative probability approaches one. This makes sense because the sum of all probabilities over any probability space should equal one.

Probability Density Function

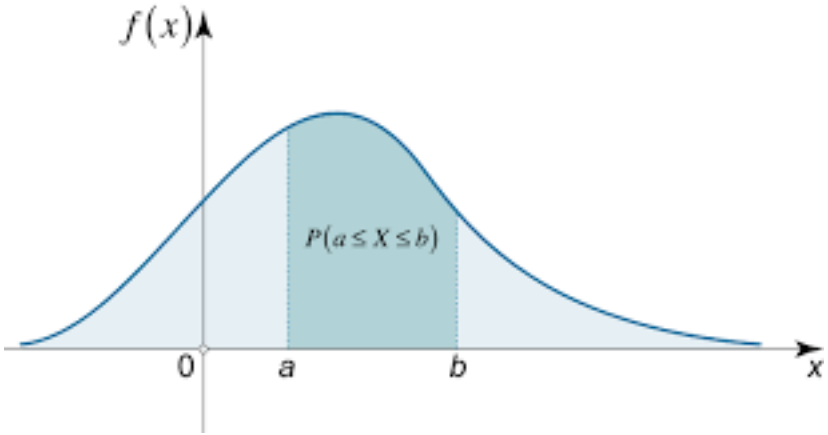
Note from author: pdf is probably the hardest probability function to understand. I recommend watching this video.

A probability mass function (pmf) gives the probability that a **discrete** random variable X is exactly equal to some value x. However, for a **continuous** random variable, there is no such function. That's because the probability that a continuous random variable is exactly equal to a given value is zero.

E.g. what's the probability that a UK male's height is exactly equal to 157? The probability of that event occurring is so infinitesimally small that it might as well be zero (because there are an infinite number of exact heights like 157.000000000001, etc.).

What we can calculate is the probability that X is **between** two values a and b.

E.g. see the following probability density function (pdf) for a continuous random variable X:



Definition

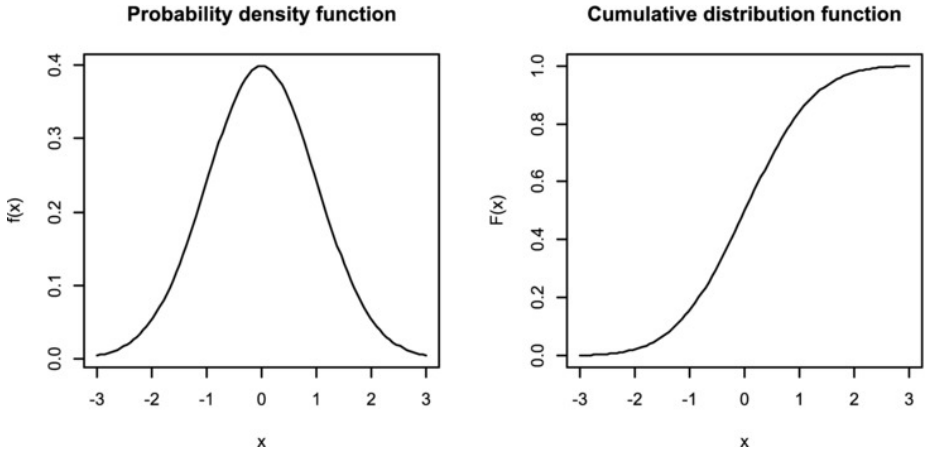
Consider a continuous random variable X with an absolutely continuous cdf $F_X(x)$.

The pdf of X is a function $f_X(x)$:

$$\begin{aligned} f_X(x) &= \frac{dF_X(x)}{dx} \\ &= F'_X(x) \end{aligned}$$

Cumulative Distribution Function to Probability Density Function

The pdf of a discrete random variable X is simply the derivative of its cdf as demonstrated in the following graphic:



Important Continuous Probability Distributions

Exponential Distribution

The exponential distribution is the probability distribution of the time between events in a poisson process which can predict the probability of a fixed number of events occurring in a fixed interval of time.

Examples

- The amount of time until an earthquake occurs;
- battery life.

Cumulative Distribution Function (cdf)

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & x < 0. \\ 0, & \text{otherwise.} \end{cases}$$

Probability Density Function (pdf)

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x < 0. \\ 0, & \text{otherwise.} \end{cases}$$

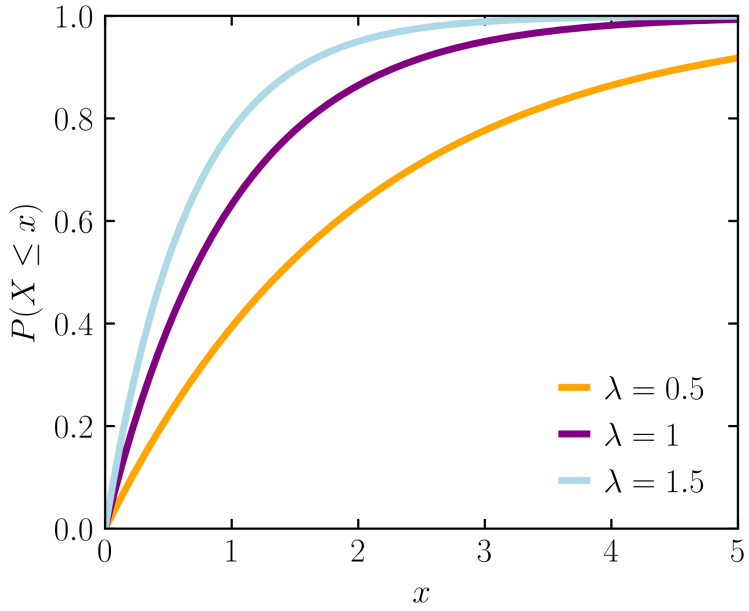


Figure 2: Exponential cdf

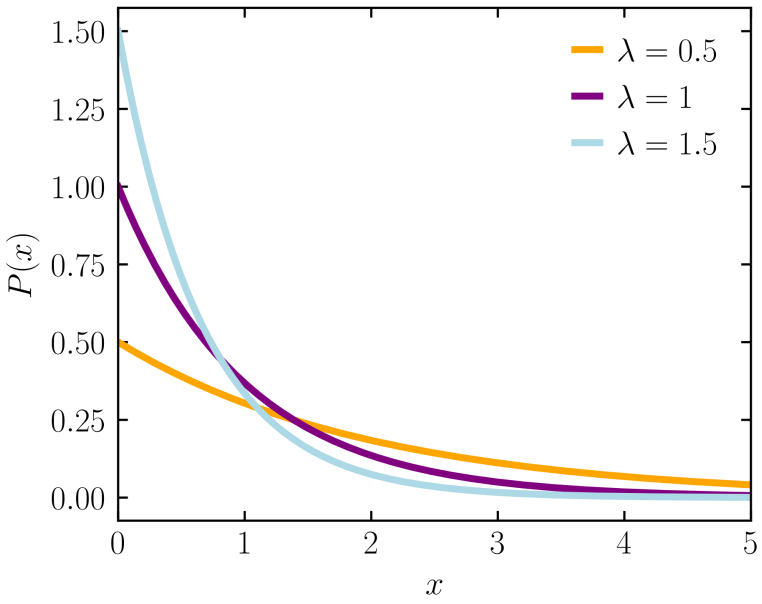


Figure 3: Exponential pdf

Normal Distribution

The normal (gaussian) distribution is a type of continuous probability distribution which is symmetrically distributed with a central peak.

Notation

The normal distribution is often referred to as $\mathcal{N}(\mu, \sigma^2)$. Thus, when a random variable X is normally distributed, with mean μ and standard deviation σ, one may write:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

Probability Density Function (pdf)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

The Standard Normal Distribution

The simplest case of a normal distribution is known as the *standard normal distribution*.

This is a special case where $\mu = 0$ and $\sigma = 1$, and it is described by this probability density function:

$$\varphi(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$$

Where φ is the greek letter 'phi'.

One can obtain the cumulative distribution function (cdf) of the standard normal distribution, often referred to as $\Phi(x)$ by integrating its pdf $\varphi(x)$:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt$$

Where Φ is the capital of the greek letter 'phi'.

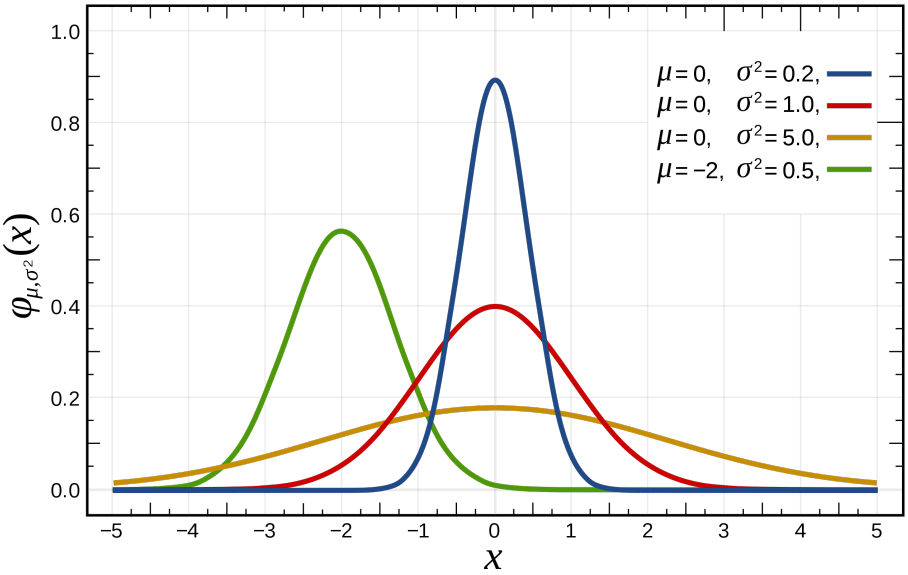


Figure 4: Normal pdf

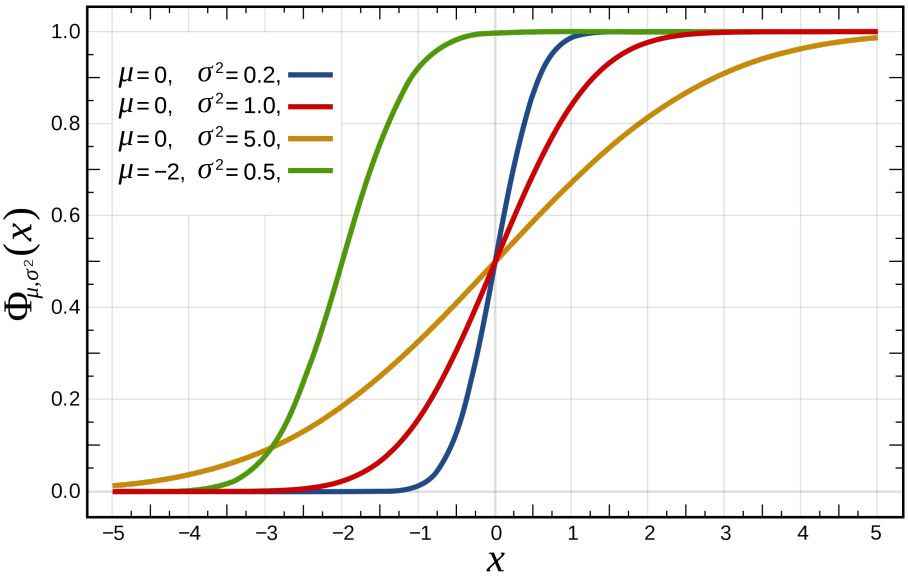


Figure 5: Normal cdf