Probability Review

Intelligent Systems and Control

2019

Sepehr Maleki

University of Lincoln
School of Engineering

What is Probability?

Consider the famous coin flip example. We know the probability that a fair coin will land heads is 0.5 (actually it is proven to be 51%, but ignore this for simplicity).

There are actually at least two different interpretations of probability.

- **The Frequentist Interpretation**: Probabilities represent long run frequencies of events. In the example above, this means if we flip the coin many times, we expect it to land heads about half the time.

- **The Bayesian Interpretation**: Probability is used to quantify the uncertainty about something; hence it is fundamentally related to information rather than repeated trials. In the example above, this interpretation means we believe the coin is equally likely to land heads or tails on the next toss.

Some Definitions in Probability Theory

Sample Space (Ω): The set of all possible outcomes of an experiment.

Event: A subset of the sample space.
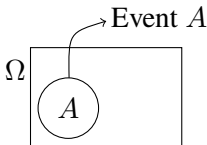
The expression $p(A)$ denotes the probability that the *event A* is true.

$$0 \leq p(A) \leq 1 .$$

- $p(A) = 1$: The event definitely will happen.
- $p(A) = 0$: The event is false.

Probability

A function p that assigns a real number $p(A)$ to each event A is called a *probability distribution* or a *probability measure* if it satisfies the following three axioms:

- **Axiom 1**: $p(A) \geq 0$ for every A ;

- **Axiom 2**: $\sum_{i=1}^{\infty} p(A_i) = 1$;

- **Axiom 3**: If A_1, A_2, ... are disjoint, then:
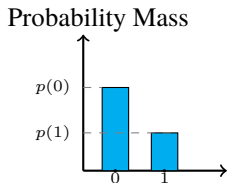
$$p\left(\bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} p(A_i) \ .$$

Random Variables

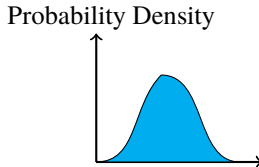A random variable is a mapping $X : \Omega \to \mathbb{R}$ that assigns a real number $X(\omega)$ to each outcome ω.

Example:

Tomorrow's weather $\to X = \begin{cases} 1 & sunny \\ 0 & otherwise \end{cases}$

Probability Mass

A Person's weight $\to X$ (Continuous)

Probability Density

Discrete Probability Distributions

Distributions that are used to describe events that only take on discrete values (e.g., count of some event).

More formally, the distribution of a random variable X is discrete, and X is called a discrete random variable, if:

- $p(X = x) \geq 0$;

- $\sum_x (X = x) = 1$;

as x runs through the set of all possible values of X.

Continuous Probability Distributions

Continuous probability distributions are used to represent variables which are measured on a continuous scale (e.g., temperature, stock values, etc.).

More formally, for the continuous random variable X with a probability density function $f(x)$:

$$p[a \leq X \leq b] = \int_a^b f(x)dx \ .$$

Probability laws require:

- $p(X) \geq 0$;

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

In particular, the probability for X to take any single value a (that is $a \leq X \leq a$) is zero.

Discrete Joint Probability Distributions

Given a pair of discrete random variables X, Y, the joint probability mass function is defined by:

$$f(x, y) = p(X = x, Y = y) \, .$$

Probability laws require:

- $p(X, Y) \geq 0$;

- $\sum_x \sum_y p(X, Y) = 1$;

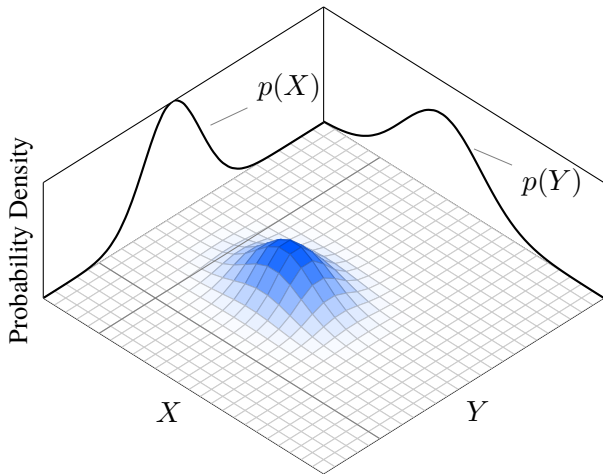Consider two random variables X, Y, each taking values 0 or 1:

$f(0, 0)$
$p(X = 0, Y = 0)$

	$Y = 0$	$Y = 1$	
$X = 0$	1/9	2/9	1/3
$X = 1$	2/9	4/9	2/3
	1/3	2/3	1

Continuous Joint Probability Distributions

- $p(X, Y) \geq 0$;
- $\int_0^\infty \int_0^\infty p(X, Y) dX \, dY$;

Discrete Marginal Probability Distributions

If X, Y have joint distribution with mass function $f(x, y)$, then the marginal mass function for X is defined by:

$$f(x) = p(X = x) = \sum_y p(X = x, Y = y) = \sum_y f(x, y) \,,$$

and the marginal mass function for Y is defined by:

$$f(y) = p(Y = y) = \sum_x p(X = x, Y = y) = \sum_x f(x, y) \,.$$

	$Y = 0$	$Y = 1$	
$X = 0$	$1/9$	$2/9$	$1/3$
$X = 1$	$2/9$	$4/9$	$2/3$
	$1/3$	$2/3$	1

Continuous Marginal Probability Distributions

If X, Y have joint distribution with mass function $f(x, y)$, then the marginal mass function for X is defined by:

$$f(x) = \int f(x, y) dy,$$

and the marginal mass function for Y is defined by:

$$f(y) \int f(x, y) dx.$$

Example: $f(x, y) = \begin{cases} x + y & \text{if } 0 \le x \le 1, \ 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$.

$$f(y) = \int_0^1 (x + y) dx = \int_0^1 x \, dx + y \int_0^1 dx = \frac{1}{2} + y.$$

Discrete Conditional Probability Distributions

If X and Y are discrete, then the conditional distribution of X given that we have observed $Y = y$ (i.e., $p(y) > 0$), is given by the following probability mass function:

$$f(x|y) = p(X = x|Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} = \frac{f(x, y)}{f(y)}$$

	$Y = 0$	$Y = 1$	
$X = 0$	$1/9$	$2/9$	$1/3$
$X = 1$	$2/9$	$4/9$	$2/3$
	$1/3$	$2/3$	1

Continuous Conditional Probability Distributions

For continuous random variables X and Y, the conditional probability density function (assuming $f(y) > 0$) is given by:

$$f(x|y) = \frac{f(x,y)}{f(y)} .$$

Then:

$$p(X \in A|y) = \int_A f(x|y)dx .$$

Example: Find $p(X < 1/4|Y = 1/3)$, given:

$$f(x,y) = \begin{cases} x + y & \text{if } 0 \le x \le 1, \ 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases} .$$

Independence of Random Variables

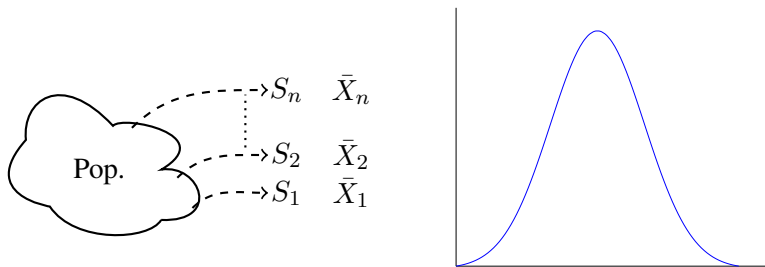Two random variables X and Y are *independent* if:

$$p(X, Y) = p(X)p(Y) .$$

Independence of X and Y is denoted by $X \perp\!\!\!\perp Y$.

Example:

|	$Y = 0$	$Y = 1$	
$X = 0$	$1/4$	$1/4$	$1/2$
$X = 1$	$1/4$	$1/4$	$1/2$
	$1/2$	$1/2$	1

Central Limit Theorems

The central limit theorem says that the distribution of averages of sufficiently large samples approaches a Normal distribution regardless of the underlying distribution of the population.

The Law of Total Probability

Let $A_1, ..., A_k$ be a partition of a sample space Ω. Then for any event B:

$$p(B) = \sum_{i=1}^{k} p(B|A_i)p(A_i) \ .$$

- $p(B, A_1)$
- $p(B, A_2)$
- $p(B, A_3)$
- $p(B, A_4)$
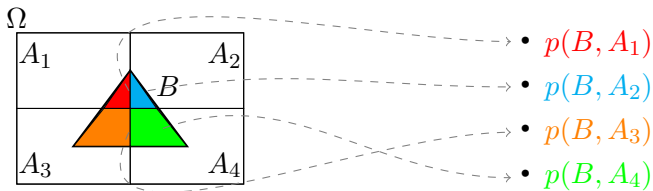
$$p(B) = \sum_{i=1}^{4} p(B, A_i) = \sum_{i=1}^{4} p(B|A_i)p(A_i) \ .$$

Bayes' Theorem for Point Probabilities

Let $A_1, ..., A_k$ be a partition of a sample space Ω such that $p(A_i) > 0$ for each i. If $p(B) > 0$, for event B, then:

$$p(A_i|B) = \frac{p(B|A_i)p(A_i)}{\sum_j p(B|A_j)p(A_j)} \ .$$

- $p(A_i)$ is called the prior probability of A_i.

- $p(B|A_i)$ is called the likelihood of event B, given A_i.

- $p(A_i|B)$ is called the posterior probability of A_i.

Bayes' Theorem – Example 1

I divide my email into three categories:

I. A_1: Spam ; from experience $\rightarrow p(A_1) = .7$,

II. A_2: Low priority ; from experience $\rightarrow p(A_2) = .2$,

III. A_3 High priority ; from experience $\rightarrow p(A_3) = .1$.

Note that $p(A_1) + p(A_2) + p(A_3) = 1$. Let B be the event that an email contains the word "free". Then from experience:

• $p(B|A_1) = .9$,   • $p(B|A_2) = .01$,   • $p(B|A_3) = .01$.

What is the probability that the email is spam?

$$p(A_1|B) = \frac{0.9 \times 0.7}{(0.9 \times 0.7) + (0.01 \times 0.2) + (0.01 \times 0.1)} = 0.995$$

Bayes' Theorem – Example 2

Consider the mammogram test for diagnosis of breast cancer. Patient A who is in her 40s takes the test. What is the probability that A has cancer, if the test is positive?

We know the test has a sensitivity of 80%, which means, if A has cancer, the test will be positive with probability 0.8. That is:

$$p(x = 1|y = 1) = 0.8 \ ,$$

where $x = 1$ is the event the mammogram is positive and $y = 1$ is the event that A has breast cancer.

From research, we also know that the probability that a person has breast cancer is 0.004. That is, $p(y = 1) = 0.004$.

Bayes' Theorem – Example 2

Like many other tests, breast cancer tests may return a false positive result. Unfortunately, such false positives are quite likely (with recent screening technology):

$$p(x = 1|y = 0) = 0.1 .$$

So:

$$p(y = 1|x = 1) = \frac{p(x = 1|y = 1)p(y = 1)}{p(x = 1|y = 1)p(y = 1) + p(x = 1|y = 0)p(y = 0)}$$
$$= \frac{0.8 \times 0.004}{0.8 \times 0.004 + 0.1 \times 0.996} = 0.031 ,$$

where $p(y = 0) = 1 - p(y = 1) = 0.996$. In other words, if you test positive, you only have about a 3% chance of actually having breast cancer!

Some Important Distributions

I. Binomial Distribution

Consider the following assumptions:

- A set of n experiments or trials are conducted.

- Each trial could result in either a success or a failure.

- The probability p of success is the same for all trials.

- The outcomes of different trials are independent.

- We are interested in the total number of successes in these n trials.

Let X be the total number of successes. Then, X is called a binomial random variable, and the probability distribution of X is called the binomial distribution.

I. Binomial Distribution

Examples of binomial distribution include:

- The number of heads/tails in a sequence of coin flips.
- Vote counts for two different candidates in an election.
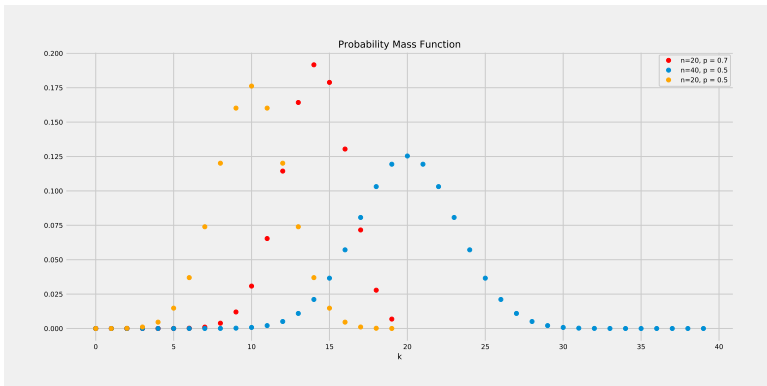- The number of successful sales calls.

Notation: $B(n, p)$.

Parameters $\begin{cases} n \in \{0, 1, 2, ...\} - \text{number of trials} \\ p \in [0, 1] - \text{success probability for each trial} \end{cases}$.

I. Binomial Distribution

PMF: $p(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, for $k = 0, 1, 2, ..., n$ and
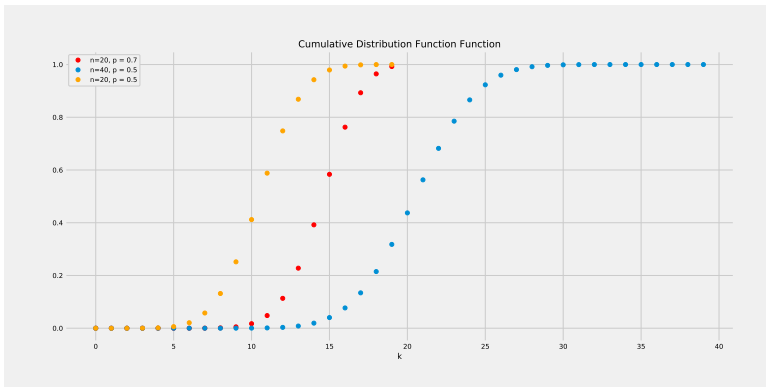
$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \ .$$

Probability Mass Function

I. Binomial Distribution

CDF:

$$p(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \, .$$

Cumulative Distribution Function Function

II. Bernoulli Distribution

The Bernoulli distribution is a special case of the binomial distribution where a single trial is conducted (so n would be 1 for such a binomial distribution).

Examples:

- Toss of a coin.
- Success of medical treatment.
- Student passes an exam.

Notation: $Bernoulli(p)$.

Parameters $\begin{cases} 0 \le p \le 1 \\ q = 1 - p \end{cases}$.

III. Uniform Distribution

The uniform distribution is the simplest example of a continuous probability distribution. A random variable X is said to be uniformly distributed if its density function is given by:
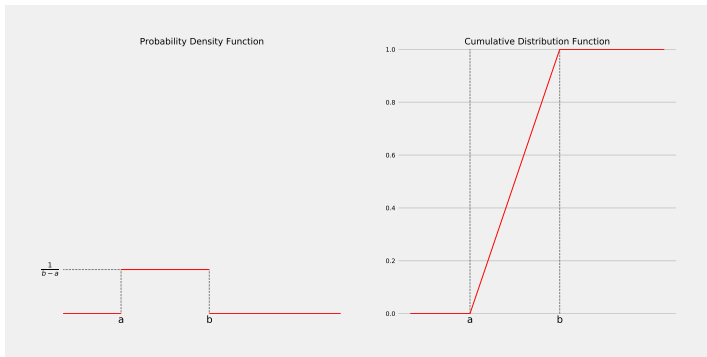
$$f(x) = \frac{1}{b-1} \ .$$

Notation: $\mathcal{U}(a,b)$ or $\mathrm{unif}(a,b)$.

Parameters: $-\infty < a < b < \infty$.

Mean: $\frac{1}{2}(a+b)$.

Variance: $\frac{1}{12}(b-a)^2$.

III. Uniform Distribution

$$\textbf{PDF} \begin{cases} \frac{1}{b-a} & for\, x \in [a.b] \\ 0 & otherwise \end{cases},$$

$$\textbf{CDF} \begin{cases} 0 & for\, x < a \\ \frac{x-a}{b-a} & for\, x \in [a.b) \\ 1 & for\, x \geq b \end{cases}.$$

IV. Normal Distribution

The normal distribution is the most important distribution in statistics, since it arises naturally in numerous applications.

A random variable X is said to have the normal distribution with parameters μ (mean) and σ^2 (variance) if its density function is given by:
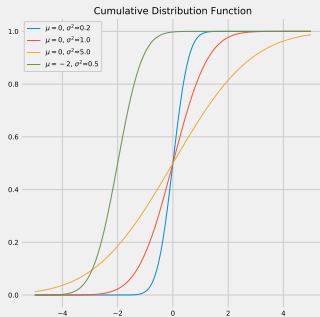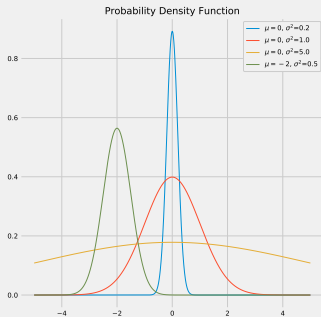
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} ,$$

for $-\infty < x < \infty$.

Notation: $\mathcal{N}(\mu, \sigma^2)$.

Parameters: $-\infty < a < b < \infty$.

IV. Normal Distribution

PDF: $\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, $\qquad$ **CDF**: $\frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$.

V. Poisson Distribution

The Poisson distribution is another family of distributions that arises in a great number of business situations. It usually is applicable in situations where random "events" occur at a certain rate over a period of time.
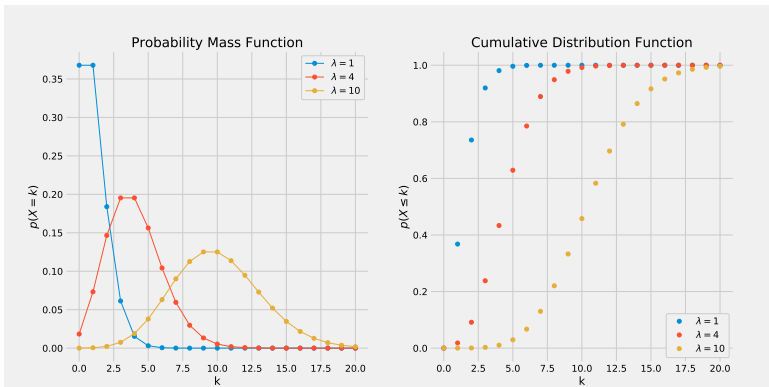
Examples:

- The hourly number of customers arriving at a bank.
- Monthly demands for a particular product.
- The daily number of emergency calls in Lincoln.

Parameters: $\lambda > 0 \ (real)$ - rate.

Mean: λ.

Variance: λ.

V. Poisson Distribution

PMF: $\frac{\lambda^k e^{-\lambda}}{k!}$,

CDF: $e^{-\lambda} \sum_{i=0}^{k} \frac{\lambda^i}{i!}$.

VI. Exponential Distribution

This family of continuous distributions is characterised by a single parameter λ, which is called the rate. Intuitively, λ can be thought of as the instantaneous "failure rate" of a "device" at any time t, given that the device has survived up to t.

The exponential distribution is typically used to model time intervals between "random events".

Examples:

- The length of time between telephone calls.
- The life time of electronic components.
- The length of time between arrivals at a service station.
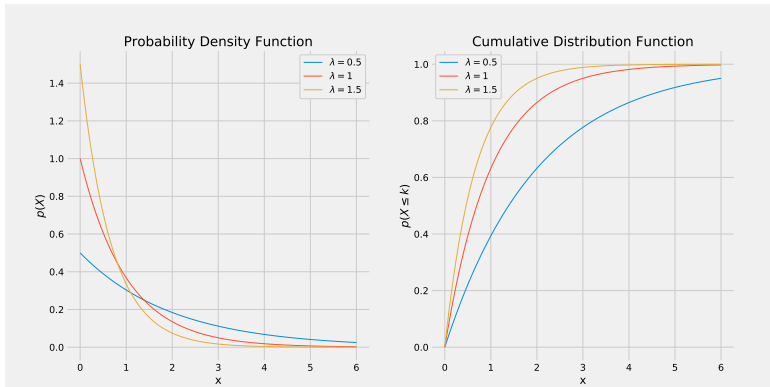
VI. Exponential Distribution

when times between random "events" follow the exponential distribution with rate λ, then the total number of events in a time period of length t follows the Poisson distribution with parameter λt.

Parameters: $\lambda > 0$ - rate.

Mean: λ^{-1}.

Variance: λ^{-2}.

VI. Exponential Distribution

PMF: $\lambda e^{-\lambda x}$,

CDF: $1 - e^{-\lambda x}$.

VII. Beta Distribution

Beta distribution is a family of continuous probability distributions defined on the interval $[0, 1]$ parametrised by two positive shape parameters, denoted by α and β, that appear as exponents of the random variable and control the shape of the distribution.
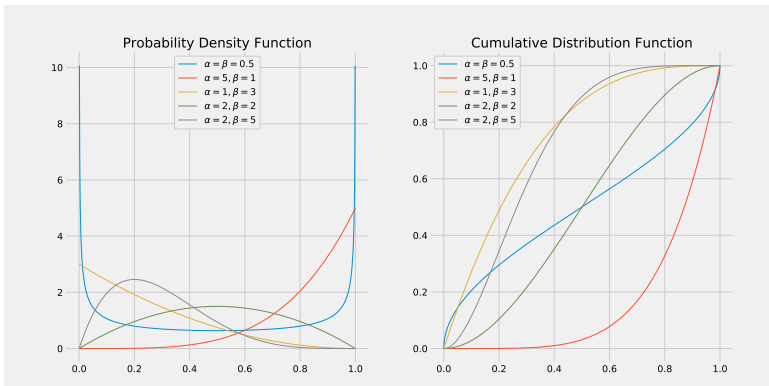
Notation: $Beta(\alpha, \beta)$.

Parameters: $\alpha > 0, \beta > 0$ (shape - real).

Mean: $\frac{\alpha}{\alpha+\beta}$.

Variance: $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

VII. Beta Distribution

PMF: $\dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$,

CDF: $I_x(\alpha,\beta)$.

Some References

- Probability Theory – E. T. Jaynes, 2003.

- Introduction to Probability – Bertsekas and Tsitsiklis, 2008.

- All of Statistics: A Concise Course in Statistical Inference – Larry A. Wasserman, 2004.