# CO7093/CO3093 - Big Data & Predictive Analytics
# Lecture Notes on Basic Probability & Statistics

School of Informatics
University of Leicester

In computing, we need probability to, for example, analyse average case complexities or to perform inferential statistics. In these notes, we discuss a certain number of probability concepts relevant to predictive analytics. Probability enables us to account for uncertainties in modeling the data, analysing it, inferring something or making decision.

## 1 Fundamentals of Statistics

**Population:** Conceptually, this represents all members of a group about which we want to reach a conclusion. For example, the entire set of user accounts at UoL.

**Sample:** represents the part of the population selected for analysis. For example, 50 users who accessed the Student Records System per day.

**Variable** : this represents a characteristic of an item or an individual that will analysed using statistics. For example, the role of users who access the Student Record System. A variable can be *categorical* or *numerical*. A *categorical* variable take its values from an established list of categories. For example, the variable **gender** takes the values Male or Female. A *numerical* variable takes its values from counting or measuring. For example, the number of students present in the class or the time it takes for a student to complete their assignment.

**A statistic** is a numerical measure that describes a variable from a population. For example, the mean time that students spend in the CO3093 homework.

**Event:** Conceptually, an *event* is an outcome of an experiment or a survey. For example, rolling a die and turning up three dots or someone using a mobile phone. An *elementary* event is an outcome that satisfies only one criterion while an *joint* event satisfies two or more criteria.

**Random variable:** Conceptually, this represents a variable whose numerical values represent the events of an experiment. For example, the student marks in an exam or the number of email messages received by a person in a day.

**Probability:** This is then defined as the number that represents the chance that a particular event will occur for a random variable. For example, the chance that an individual will own a Ferrari, the likelihood of a serving PM to win the election or the chance of getting Head in a fair coin flip. Informally, *a probability p of an event E is defined as the number of favourable outcomes divided by the number of all possible outcomes.*

$$p(E) = \frac{\text{number of favourable outcomes}}{\text{number of all possible outcomes}}$$

Consequently, a probability is a decimal number in the range of 0.0 to 1.0. The *sample space* is the set of all possible events. The sum of the individual probabilities associated with a set of all possible events is always 1.

## 1.1 Some Rules of Probability

A set of rules govern the calculation of the probabilities of elementary and joint events.

**Rule 1:** The probability of an event must be between 0 and 1 inclusive.

**Rule 2:** The event that $A$ does not occur is called **A complement** or simply **not A**, and is denoted by the symbol $\bar{A}$. If $P(A)$ represents the probability of event $A$ occurring, then $1 - P(A)$ represents the probability of $\bar{A}$.

**Rule 3:** If two events $A$ and $B$ are mutually exclusive, then the probability of both events $A$ and $B$ occurring at the same time is 0.

**Rule 4:** If two events $A$ and $B$ are mutually exclusive, the probability of either event $A$ or event $B$ occurring is the sum of their separate probabilities.

**Rule 5:** If the events in a set are mutually exclusive and form the sample space, then the sum of their probabilities must add up to 1.

**Rule 6:** If two events $A$ and $B$ are not mutually exclusive, the probability of either event $A$ or event $B$ occurring is the sum of their separate probabilities minus the probability of their simultaneous occurrence (the joint probability).

**Rule 7:** If two events $A$ and $B$ are independent, the probability of both events $A$ and $B$ occurring is equal to the product of their individual probabilities. In this case, the occurrence of one event, in no way, affects the probability of the other event.

**Rule 8:** If two events $A$ and $B$ are not independent, the probability of both events $A$ and $B$ occurring is the product of the probability of event $A$ multiplied by the probability $P(B|A)$ of event $B$ occurring, given that event $A$ has occurred. This leads to the so-called **Bayes formula**, which we will review a bit later with the Bayesian statistics.

### 1.2 The Bayes formula

The Bayes' theorem enables us to compute the probability $P(A|B)$ of observing an outcome A given the event B as follows:

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

This Bayesian approach allows us to account for a prior knowledge but this is also a key weakness in the sense that we need to have that prior knowledge.

**A simple example:** Assume you have a fair 6-sided die. Use Bayes's theorem to calculate the probability the first roll is 5 given the second roll is 1 results in $P$(First is 5| Second is 1).

$P$(First is 5) $= 1/6,\ P$(Second is 1) $= 1/6,\ P$(Second is 1| First is 5) $= 1/6$.

Applying the Bayes formula, we get $P$(First is 5| Second is 1) $= 1/6$. As the events are independent, the probability of rolling a specific number on any roll is $1/6$.

### 1.3 Assigning Probabilities

To assign probabilities to the events of a random variable, we can use the following three approaches:

1. **Classical approach:** This assumes that all elementary events are equally likely to occur based on prior knowledge of the process involved. For example, rolling a die and assigning the probability of turning up the face with three dots.

2. **Empirical approach:** This does not rely on prior knowledge and assigns probabilities based on frequencies obtained from empirically observed data. For example, probabilities determined by marketing surveys.

3. **Subjective approach:** This assigns probabilities based on expert opinions or "gut" feelings. For example, a commentator stating that a particular team will win the championship.

## 2 Probability distributions

A probability distribution for a random variable provides the probabilities associated with the events for that variable. The distribution can take different forms depending on whether the variable is discrete or continuous.

### 2.1 Discrete Random Variables

Discrete variables take integer values. In this case, we would like to determine the probability distribution associated with such a variable. Let us start with a work-out problem.

**Problem 1:** You want to determine the probability of getting 0, 1, 2, or 3 heads when you toss a fair coin (one with an equal probability of a head or a tail) three times in a row.

Counting the number of favourable outcomes lead us to the following table:

Indeed the sum of the probabilities of the exhaustive events is 1.0.

| Number of Heads | Number of favourable outcomes | Probability |
|---|---|---|
| 0 | 1 | 1/8=0.125 |
| 1 | 3 | 3/8=0.375 |
| 2 | 3 | 3/8=0.375 |
| 3 | 1 | 1/8=0.125 |

Table 1: Probability Distribution for Tossing a Fair Coin Three Times

## 2.2 Summarizing Distributions

To summarise a distribution, we usually compute a certain number of statistics for its associated random variable. These statistics include the expected value, the variance, or the standard deviation of the given random variable.

### The Expected Value of a Variable

This is defined as the sum of the products formed by multiplying each possible event in a discrete probability distribution by its corresponding probability. Denoting by $\mu$ the expected value, we have the following formula:

$$\mu = \text{Expected or Mean Value } = \text{ Sum of [each value} \times \text{ the probability of each value]}$$

Referring to the example summarised in Table 1, we can calculate the mean as follows:

$$\mu = (0)(0.125) + (1)(0.375) + (2)(0.375) + (3)(0.125) = 1.50$$

### Standard Deviation

The standard deviation measures the variation around the expected value of a random variable. Denoting by $\sigma$ the standard deviation, we can write the following formula:

$$\sigma = \begin{array}{l}\text{Square root of [ Sum of (Squared differences between each value} \\ \text{and the mean)} \times \text{(Probability of the value)]}\end{array}$$

Referring to the example summarised in Table 1, we can calculate the standard deviation as follows:

$$\begin{aligned}\sigma &= \sqrt{(0-1.5)^2(0.125) + (1-1.5)^2(0.375) + (2-1.5)^2(0.375) + (3-1.5)^2(0.125)} \\ &= \sqrt{0.75} \\ &= 0.866\end{aligned}$$

**Problem 2:** Suppose that you are deciding between two alternative investments. Investment A is a mutual fund whose portfolio consists of a combination of stocks that make up the Dow Jones Industrial Average. Investment B consists of shares of a growth stock. You estimate the returns (per \$1,000 investment) for each investment alternative under three economic condition events (recession, stable economy, and expanding economy), and also provide your subjective probability of the occurrence of each economic condition as follows. Based

| Probability | Economic Event | Dow Jones Fund (A) | Growth Stock B |
|---|---|---|---|
| 0.2 | Recession | -100 | -200 |
| 0.5 | Stable economy | +100 | +50 |
| 0.3 | Expanding economy | +250 | +350 |

Table 2: Estimated Return for Two Investments Under Three Economic Conditions

on the mean and standard deviation calculations, can you recommend which one of the two is worth investing to?

Calculating the mean values and standard deviations for both investments A and B, we obtain

$$\mu_A = 105 \qquad \mu_B = 90$$
$$\sigma_A = 121.35 \quad \sigma_B = 194.68$$

Having a higher mean return with less variation makes the Dow Jones fund (A) a more desirable investment than the growth fund (B).

### 2.2.1 The Bernouilli distribution

If we toss a coin, we have a random variable $X$ with two mutually exclusive events (outcomes) that are Head or Tail. These two outcomes can be viewed Success (1) or Failure (0). We get the following distribution:

| Outcome x | Probability p |
|---|---|
| 1 | p |
| 0 | 1-p |

The variable $X$ follows a Bernouilli distribution of probability $p$, which we denote as $X \sim \text{Bernouilli}(p)$.

### 2.2.2 The Binomial distribution

The binomial distribution is used for random variables consisting of a fixed number $n$ of Bernoulli trials. If $p$ is the probability of success for a Bernoulli trial, we can use a mathematical equation to calculate the probability of obtaining $k$ successes out of $n$ trials. Denoting by $X$ the random variable, we can calculate the probability $P(X = k)$ as follows:

$$P(X = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

We denote $X \sim \text{Binomial}(n, p)$ to mean the variable $X$ represents a binomial distribution made of $n$ Bernoulli trials with each having the probability of success $p$. The notation $n!$ reads factorial n and represents the number $n \times (n-1) \times (n-2) \times \ldots \times 1$.

**Work-out example:** Consider the example in Problem 1 and re-compute the distribution shown in Table 1. It is easy to see that the variable in Problem 1 represents a binomial distribution with $p = 0.5$ and $n = 3$. The number of successes $k$ takes the values $0, 1, 2, 3$.

The characteristics of the binomial distribution are:

- Mean $\mu = np$
- Variance $\sigma^2 = np(1-p)$
- Standard deviation $\sigma = \sqrt{np(1-p)}$

### 2.2.3 The Poisson distribution

The Poisson distribution is a limiting version of the Binomial distribution $X \sim \text{Binomial}(n, p)$ when

$$np \to \lambda$$

It is a probability distribution for a discrete random variable with the following criteria:

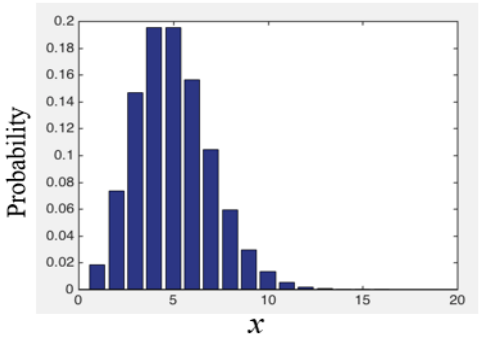
- We are interested in the number of times a particular event occurs in a unit while the probability that an event occurs in a particular unit is the same for all other units.
- The number of events that occur in a unit is independent of the number of events that occur in other units.
- As the unit gets smaller, the probability that two or more events will occur in that unit approaches zero.

**Examples:** Number of computer network failures per day, number of surface defects per square meter of floor covering, or the number of customers arriving at a bank during the 12 noon to 1 p.m. hour.

If $X \sim$ Poisson($\lambda$), then the probability of obtaining $X = x$ for some integer value $x$ is calculated as follows:

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

For $\lambda = 4$, we obtain the following distribution:




$\lambda = 4$

**Problem 3:** You want to determine the probabilities that a specific number of customers will arrive at a bank branch in a one-minute interval during the lunch hour: Will zero customers arrive, one customer, two customers, and so on?

To solve this problem, you can proceed as follows. First, we need to convince ourselves that the criteria of a Poisson distribution are met in the case. Then, we need to find out the

mean (expected value) by using historical data. Assume that the historical data shows that the average number of customers per minute during the lunch time hour is 3. We denote by $\lambda$ the mean so $\lambda = 3$.Then, we can use a mathematical formula to estimate the probabilities of getting $0, 1, 2, \ldots$ arrivals. Below is the Poisson distribution formula:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

The characteristics of the Poisson distribution are:

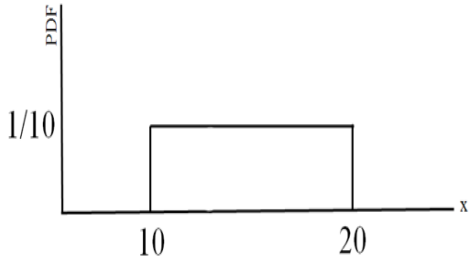
- Mean is the population mean $\mu = \lambda$
- Variance is also the population mean $\sigma^2 = \lambda$
- Standard deviation $\sigma = \sqrt{\lambda}$

## 2.3 Continuous probability distributions

Unlike discrete variables, *continuous* variables can take any value in the range and not only an integer value. Probabilities are expressed as area under a curve representing the distribution and the probability of getting exactly any specific value is zero. The most important probability distribution is the Gaussian or Normal distribution as it can model different continuous variables describing frequent phenomena.

### Uniform distribution

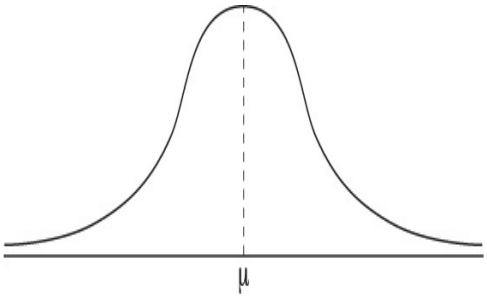
For a continuous random variable, probabilities are expressed as an area under a curve representing the distribution. The curve is simply given by a function, which specifies the probability of the random variable falling within a particular range of values. Such a function is called the PDF (probability density function).



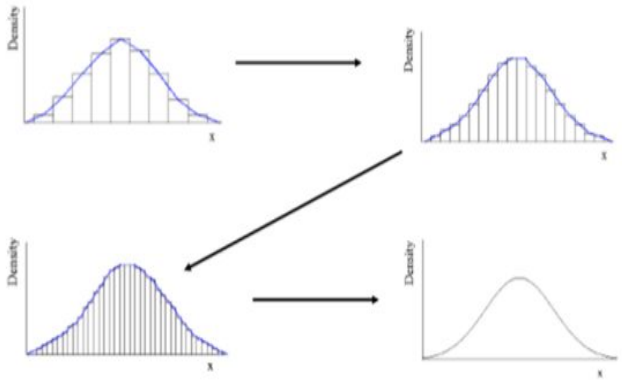

This figure illustrates the uniform distribution in the interval $[10, 20]$. The corresponding function is defined as $f : x \mapsto \frac{1}{10}$. In general, the PDF of uniform distribution in the interval $[a, b]$ is the function that associates $x \in [a, b]$ to $\frac{1}{b-a}$ with $a \neq b$.

### Normal distribution

This distribution extends from negative to positive infinity and displays a curve that is bell-shaped and symmetrical around the mean $\mu$.




Examples of normal distribution include the physical characteristics such as height/weight of a population or the students' scores on a standardized exam. The normal distribution can represent an approximation of various discrete probability distributions such as the binomial and Poisson distributions.



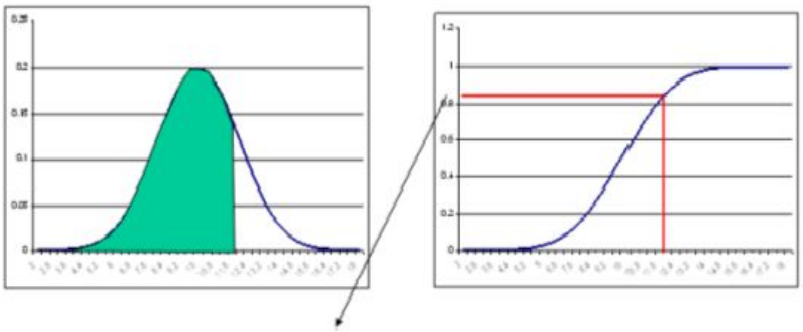

Probabilities are always cumulative and expressed as inequalities, such as $P(X \leq x)$ or $P(X \geq x)$, where $x$ is a value for the variable $X$. The following notations are used:

- $F(x) = P(X \leq x)$ represents the **CDF (cumulative distribution function)**.
- $N(\mu, \sigma)$ represents the normal distribution of mean $\mu$ and standard deviation $\sigma$.
- $\Phi(x)$ is the probability density function and represents the bell-shaped curve. It is used in calculating the probability as an integral but we do not need to manipulate such a complex formula. Just for information, $\Phi$ is defined as

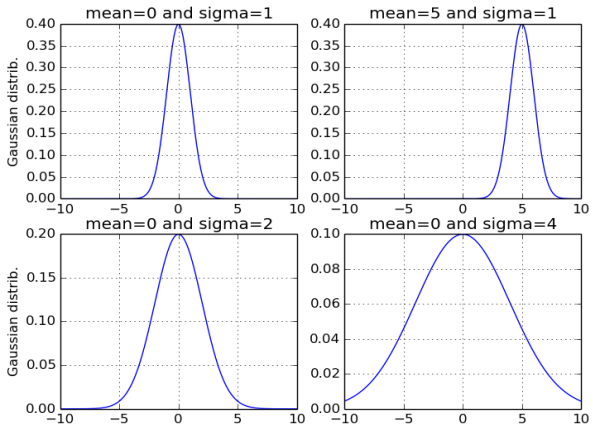
$$\Phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}.$$

The probabilities are defined as **area under the distribution curve** and this gives us the **cumulative distribution** as follows.

For example $F(-1.96) = P(X \leq -1.96) = 0.025$ or $F(1.96) = P(X \leq 1.96) = 0.975$ for the normal distribution $N(0, 1)$. We can get normal probabilities by using a table of normal probabilities or by using software functions from within a package such as Excel, MATLAB or Python.




```
>>> import scipy.stats
>>> scipy.stats.norm(0,1).pdf(0)
0.3989422804014327
>>> scipy.stats.norm(0,1).cdf(0)
0.5
>>> scipy.stats.norm(0,1).cdf(-1.96)
0.024997895148220435
```




We can convert a random variable $X$ (of mean $\mu$ and standard deviation $\sigma$) to its corresponding $Z$ score (of mean 0 and standard deviation 1) by using the following formula:

$$Z = \frac{X - \mu}{\sigma}$$

The mean ($\mu$) and standard deviation ($\sigma$) represent respectively the center and the 'width' of the bell-shaped curve. A higher value for $\sigma$ indicates a 'fatter' bell-shaped curve meaning a high spread of the data around the mean.