ASSIGNMENT 1

THE ROLE OF PROBABILITY AND STATISTICS

MAS291

TRAN THANH DUONG
SE160185

Table of Contents

1. What are Probability and Statistics?

Statistics is a branch of mathematics that deals with collecting, analyzing, interpreting and presenting data. It helps us in making decisions and drawing conclusions in the presence of uncertainty. In short, statistics is a **science of data**. Statistics is an interdisciplinary scientific field because it has an important role in the development of new systems, as well as the design, development, and improvement of production processes.

Probability is discipline of mathematics concerned with numerical explanations of the likelihood of an event occurring of the truth of a claim. The probability of an occurrence is a number between 0 and 1, with 0 indicating impossibility and 1 indicating certainty. Rolling unbiased 6-face dice is a basic illustration. There are 6 equally likely possibilities. All six faces have the same probability ($\frac{1}{6}$ or 0.1666 ...).

The fundamental idea in the field of statistics is describing and understanding the **variability**. It refers to the fact that subsequent observations of a system or phenomena may not always provide the same result. In our daily lives, we all experience unpredictability and **statistical thinking** may help us incorporate this variety into our decision-making processes.

For example, when you compile or run some lines of code, is the runtime always the same for every time you run those same lines of code? Of course not, the runtime changes for each run of the code. Its variation can depend on many factors, such as the change in performance of the computer from time to time (maybe the computer has to perform many other tasks at that time such as downloading a file on the internet or running an application), the difference in hardware will also affect the running speed, or the power will also affect the variability of the runtime. Just because the variability, we can consider the runtime to be a **random variable**. The following simple model is a useful way to conceive about a random variable, say X, that reflects a measurement:

$$X = \mu + \epsilon$$

Where μ is a **constant** and ϵ is a **random disturbance**. The constant μ is always the same for every measurement, ϵ changes when there are impacts cannot be controlled (a difference in performance of the computer or the power supply is unstable)

Probability is now like a mathematical language for discussing that uncertain events, and it is an important key role of statistics. There are several reasons of variance in every measurement or data gathering activity. This means that if the same measurement was taken again, the result would most likely vary. In each scenario, we have to try to understand and manage the causes of variance.

2. Importance of Probability and Statistics in AI

Artificial intelligence (AI) is concerned with making predictions and identifying patterns in data structures in order to make such predictions. This enables the machine to do a variety of analytical activities without the need for human involvement.

Probability and Statistics play roles of a collection of rules for obtaining data and making decisions based on it. They establish a link between numerous pieces of data as well as between itself. As a result, they play a key part in AI, and anybody working in the field should be familiar with probability and statistics. Finding out how data is distributed, knowledge about dependent and independent variables, and so on are all necessary steps in solving AI challenges.

At the very least, **probability** provides us with some rules for thinking logically about uncertain situations. Probability theory may be a highly useful tool for making predictions and decisions that relate to the actual world if our probability model has some relationship with it.

Whether your predictions and decisions are accurate is determined by the model you use. **Statistics** is a field whose goal is to complement probability theory by using data to develop good models.

Some cases using Probability and Statistics

A/B Tests or Experiments

To comprehend an AB test, one must first grasp the idea of p-value, a probability concept that indicates the likelihood that the null hypothesis is incorrect when it is accurate. Strong probability principles are required to have an intuitive knowledge of p-value.

The sample size for the experiment is another step in AB testing. A proper sampling plan is required in order to generalize the experiment's results to the entire population. To determine the needed sample size, one must first grasp the concepts of power and the many sorts of mistakes that might occur in an experiment, both of which are taught in statistics.

Machine Learning Model

Many machine learning models are based on the assumption that the data follows a specific pattern. An engineer who doesn't understand probability won't be able to tell if the data meets the machine learning model's assumptions, and hence won't be able to make the optimal decision in the machine learning model.

A linear regression model, for example, assumes that each data point's noise follows a normal distribution. To comprehend what this implies, a data scientist must first grasp what a normal distribution is, which is something you learn in probability class.

You can't avoid statistics whether you're running a regression, classification, or clustering model with traditional machine learning methods or deep learning approaches.

Some ideas using Probability and Statistics in AI

1. **Metrics** (such as *accuracy, precision, recall, F1-score, RMSE,* etc.) are used to monitor and measure the performance of a model.
2. **Metric summaries** such as median, mean, and standard deviation (std).
3. **Data visualization and exploration** is a concept that aided in the discovery of new and unexpected data insights. With this knowledge, the widespread concept that statistics are employed to confirm what we already know was refuted, and discoveries in other disciplines of AI.
4. **Regression** is a machine learning technique in which a collection of inputs is used to predict an output variable. To predict results, the algorithm needed a lot of inputs and interactions, which made it statistically unstable.
5. Pie charts, histograms, and scatter plots are common components of a **statistical graphics framework**. However, a new framework was created to investigate the relationship between data and visualization in a more abstract way. It's a significant step toward incorporating exploratory data and model analysis into data science and AI workflows.

3. Probability and Statistics Models

Most engineering problems need the use of **models** for analysis. Engineers spend a lot of their formal education studying about the models that are important to certain professions, as well as the strategies for using these models in problem formulation and solving.

Statistics Models

Mechanistic Model

A mechanistic model is one that use a theory to predict what will occur in the real world. A simple example for this type of model is Ohm's law which is used to calculate the current in a copper wire:

$$Current\ (I) = \frac{Voltage\ (E)}{Resistance\ (R)}$$

This model is built from knowledge of basic physics mechanism. But in practice, the measurements are not equal to the result of the above model, in fact it varies slightly at different times we measure. Consequently, a more realistic model of the observed current should be:

$$Current\ (I) = \frac{Voltage\ (E)}{Resistance\ (R)} + \epsilon$$

Empirical Model

Empirical model, on the other hand, studies real-world events in order to develop a theory. Engineers are often faced with issues for which there is no clear or well-understood mechanistic model that can explain the phenomenon. In this case, the effective solution is building a model from multiple experimental measurements or **data**.

Let's give an example through one of the most famous basic problems in Machine Learning of all time - **House Pricing Prediction**. Suppose we want to calculate the price of a house (P_{house}) which

is related to the coordinate of the house – *longitude* (Lg) and *latitude* (Lt), *housing median age* (A), *total rooms* (R), *total bedrooms* (B) and *population* (p) at the area that the house is located.

$$P_{house} = f(Lg, Lt, R, B, p)$$

A brief table of **California house pricing dataset**[1]:

Longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	**median_house_value**
-122.23	37.88	41.0	880.0	129.0	322.0	**452600.0**
-122.22	37.86	21.0	7099.0	1106.0	2401.0	**358500.0**
-122.24	37.85	52.0	1467.0	190.0	496.0	**352100.0**
-122.25	37.85	52.0	1274.0	235.0	558.0	**341300.0**
-122.25	37.85	52.0	1627.0	280.0	565.0	**342200.0**
-122.25	37.85	52.0	919.0	213.0	413.0	**269700.0**
-122.25	37.84	52.0	2535.0	489.0	1094.0	**299200.0**
-122.25	37.84	52.0	3104.0	687.0	1157.0	**241400.0**
-122.26	37.84	42.0	2555.0	665.0	1206.0	**226700.0**
-122.25	37.84	52.0	3549.0	707.0	1551.0	**261100.0**
-122.26	37.85	52.0	2202.0	434.0	910.0	**281500.0**
-122.26	37.85	52.0	3503.0	752.0	1504.0	**241800.0**
…	…	…	…	…	…	**…**

The function f in this case is unknown. In fact, there's no any underlying mechanisms can be used to measure the house pricing. Perhaps the model could be developed from a first-order Taylor series expansion added a random disturbance.

$$P_{house} = \theta_0 + \theta_1 \times Lg + \theta_2 \times Lt + \theta_3 \times R + \theta_4 \times B + \theta_5 \times p + \epsilon$$

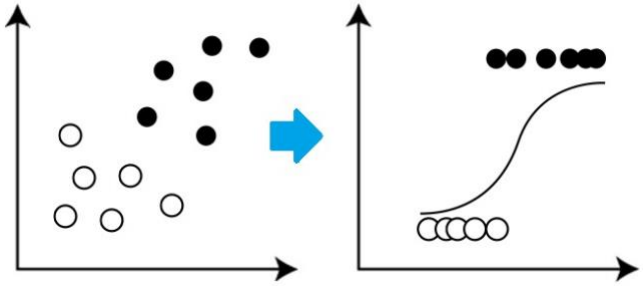
$\theta_i, i = \{0,1,2,3,4,5\}$ is unknown and mutable. Now we need to find θ_i to as fit as possible we can. This empirical model is called **regression model**.

In addition, there are some other popular statistics models such as: **logistic regression**, **clustering**, and **decision trees**.

Logistic regression

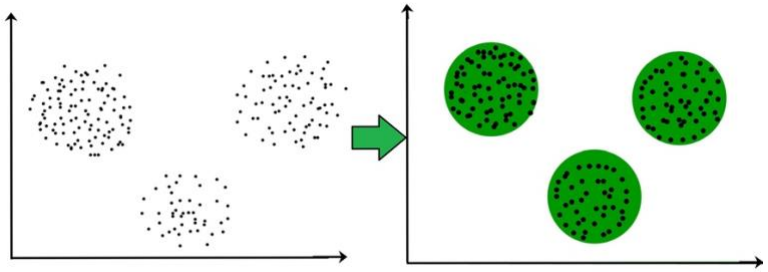
The **logistic model** is used in statistics to represent the likelihood of a given class or event, such as pass/fail, win/lose, alive/dead, or healthy/sick. This may be used to represent a variety of occurrences, such as identifying whether a picture contains a cat, dog, lion, or other animal.

[1] https://github.com/ageron/handson-ml2/tree/master/datasets/housing



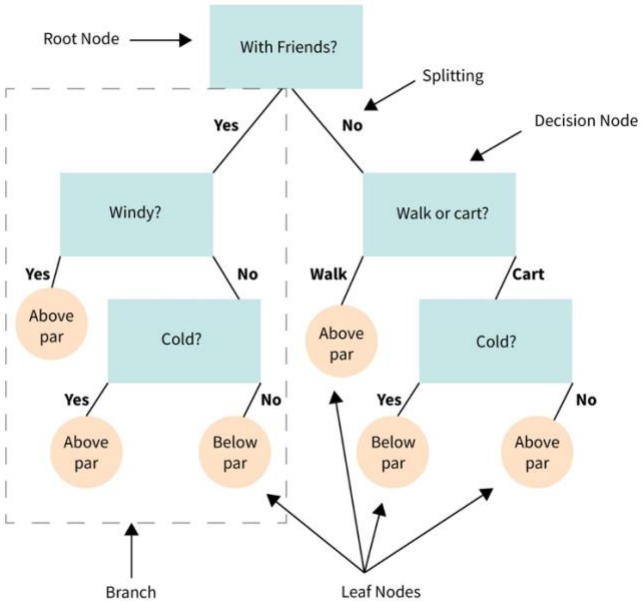
Clustering

Clustering is the process of dividing a population or set of data points into many groups so that data points in the same group are more similar than data points in other groups.



Decision Trees

For classification and regression, **Decision Trees** (DTs) are a non-parametric supervised learning approach. The objective is to learn basic decision rules from data features to develop a model that predicts the value of a target variable. A tree is an approximation to a piecewise constant.



Probability Models

A mathematical representation of a random phenomenon is known as a **probability model.** It is defined by its **sample space, events** within the sample space, and **probabilities** associated with each event. Probability models can be used to answer interesting questions about uncertain real-world systems and help in quantifying the risks associated with **statistical inference,** or the risks associated with making decisions.

To build a probability model:

- First, identify every outcome
- Second, determine the total number of possible outcomes
- Third, compare each outcome to the total number of possible outcomes to get probabilities.

For example, to construct a probability model for rolling a single, fair dice, with the event being the number shown on the die.

Make a list of all possible outcomes for the experiment to get started. The numbers $1, 2, 3, 4, 5,$ and 6 are the results that can be rolled. The sample space is made up of 6 possible outcomes.

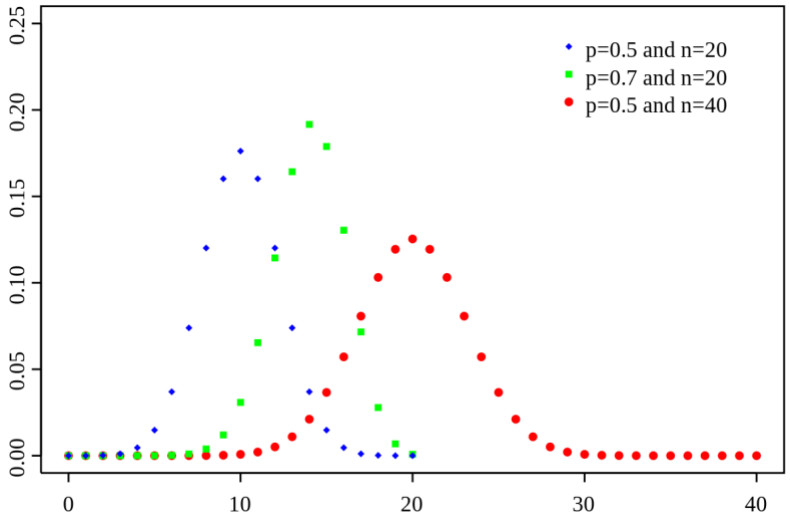
Determine a ratio of the outcome to the number of possibilities to assign probabilities to each outcome in the sample space. Because there is one of each of the six numbers on the cube and no reason to believe that any one face is more likely to appear than the others, the probabilities of rolling any number is $\frac{1}{6}$.

Outcome	Roll of 1	Roll of 2	Roll of 3	Roll of 4	Roll of 5	Roll of 6
Probability	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

There are many types of probability models: **the binomial distribution, the Poisson distribution, the normal distribution** and **hypergeometric distribution.**

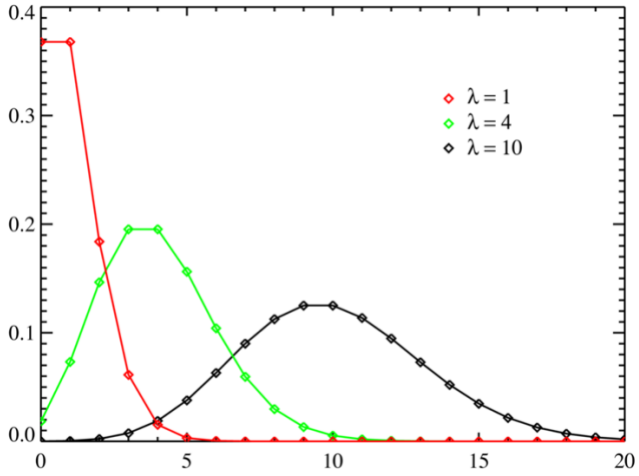
Binomial Distribution

The discrete probability distribution of the number of successes in a series of n separate experiments, each asking a yes–no question and each with its own Boolean-valued outcome: success or failure, is known as the **binomial distribution** with parameters n and p.



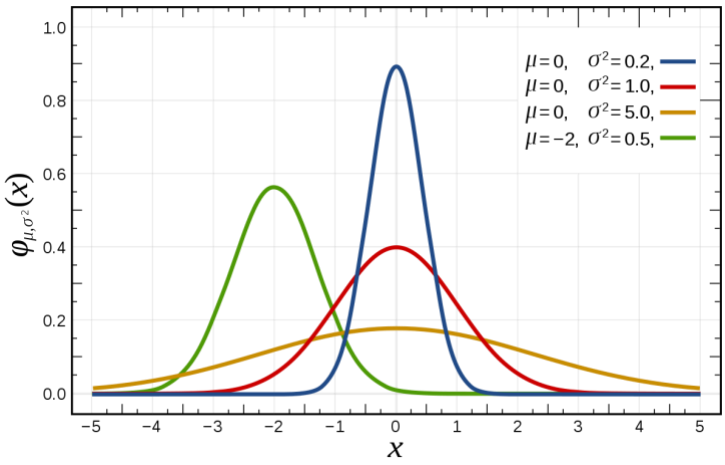
Poisson distribution

The **Poisson distribution** is a discrete probability distribution that expresses the probability of a specific number of events occurring in a fixed interval of time or space at a known constant mean rate (λ), regardless of the time since the last event.



Normal distribution

The **normal distribution,** also known as the **Gaussian distribution,** is a symmetric probability distribution centered on the mean (μ), indicating that data around the mean (σ^2) occur more frequently than data far from it. The normal distribution will show as a bell curve on a graph.



Hypergeometric distribution

The **hypergeometric distribution** is a discrete probability distribution that describes the probability of k successes (random draws for which the drawn object has a specified feature) in n draws without replacement from a finite population of size N containing exactly K objects with that feature, where each draw is either a success or a failure.