## LECTURE 1

### §1. BAYESIAN PRINCIPLE

Posterior information = prior information + data information

In the classical approach, the parameter $\theta$ is assumed to be an unknown, but fixed quantity. A random sample $X_1, X_2, \ldots, X_n$ is drawn from a population with probability density function $f(x, \theta)$ and based on observed values in the sample, knowledge about the value of $\theta$ is obtained.

In Bayesian approach $\theta$ is considered to be a quantity whose variation can be described by a probability distribution (known as a prior distribution). This is a subjective distribution, based on the experimenter's belief, and is formulated before the data are seen (and hence the name prior distribution). A sample is then taken from a population where $\theta$ is a parameter and the prior distribution is updated with this sample information. This updated prior is called the posterior distribution. The updating is done with the help of Bayes theorem and hence the name Bayesian method.

The frequentist and Bayesian approaches to statistics differ in the definition of Probability. For a Frequentist, probability is the relative frequency of the occurrence of an event in a large set of repetitions of the experiment (or in a large ensemble of identical systems) and is, as such, a property of a so-called random variable. In Bayesian statistics, on the other hand, probability is not defined as a frequency of occurrence but as the plausibility that a proportion is true, given the available information. Probabilities are then – in the Bayesian view – not properties of random variables but a quantitative encoding of our state of knowledge about these variables. This view has far-reaching consequences when it comes to data analysis since Bayesian can assign probabilities to proportions, or hypotheses, while Frequentists cannot.

The classical methods of estimation that you have studied are based solely on information provided by the random sample. These methods essentially interpret probabilities as relative frequencies. For example, in arriving at a $95\%$ confidence interval for mean, we interpret the statement

$$P(-1.96 < Z < 1.96) = 0.95$$

the mean that 95 percent of the time in repeated experiments $Z$ (standard normal random variable) will fall between -1.96 and 1.96. Probabilities of this type that can be interpreted in the frequency sense will be referred to as objective probabilities. The Bayesian approach to statistical methods of estimation combines sample information with other available prior information that may appear to be pertinent. The probabilities associated with this prior information are called subjective probabilities, in that they measure a person's degree of belief in a proportion. The person uses his own experience and knowledge as the basis for arriving at a subjective probability.

## §2. AXIOMS OF PROBABILITY

Let us consider an experiment and let $\Omega$ be the sample space of the experiment. A *probability* on $\Omega$ is a real function $P$ which is defined on events of $\Omega$ and satisfies the following three axioms.

**Axiom 1.** $P(A) \geq 0,$ for any event $A$.

**Axiom 2.** $P(\Omega) = 1.$

**Axiom 3.** For any sequence of mutually exclusive events $A_1$, $A_2$, ... (that is, events for which $A_i \cap A_j = \emptyset$ if $i \neq j$)

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n). \tag{2}$$

We call $P(A)$ the probability of event $A$.

Thus, Axiom 1 states that the probability that outcome of the experiment is contained in $A$ is some nonnegative number. Axiom 2 states that the probability of the certain event is always equal to one. Axiom 3 states that for any sequence of mutually exclusive events $\{A_n\}_{n=1}^{\infty}$ the probability that at least one of these events occurs is equal to the sum of their respective probabilities.

## PROPERTIES OF PROBABILITY

**Example 10.** In Example 2, if we assume that a head is equally likely to appear as a tail, then we would have

$$P(\{\omega_1\}) = P(\{\omega_2\}) = \frac{1}{2}.$$

On the other hand, if we had a biased coin and felt that a head was twice as likely to appear as a tail, then we would have

$$P(\{\omega_1\}) = \frac{2}{3} \qquad P(\{\omega_2\}) = \frac{1}{3}.$$

In Example 3, if we supposed that all six outcomes were equally likely to appear, then we would have

$$P(\{\omega_1\}) = P(\{\omega_2\}) = ... = P(\{\omega_6\}) = \frac{1}{6}.$$

From Axiom 3 follows that the probability of getting an even number would equal

$$P(\{\omega_2, \omega_4, \omega_6\}) = P(\{\omega_2\}) + P(\{\omega_4\}) + P(\{\omega_6\}) = \frac{1}{2}.$$

These axioms will now be used to prove the simplest properties concerning probabilities.

**Property 1.** $P(\emptyset) = 0.$

That is, the impossible event has probability $0$ of occurring.

*Proof of Property 1:* If we consider a sequence of events $A_1, A_2, ...$, where $A_1 = \Omega$, $A_k = \emptyset$ for $k > 1$ then, as the events are mutually exclusive and as $\Omega = \bigcup_{n=1}^{\infty} A_n$, we have from Axiom 3 that

$$P(\Omega) = \sum_{n=1}^{\infty} P(A_n) = P(\Omega) + \sum_{n=2}^{\infty} P(A_n)$$

and by Axiom 2 $P(\Omega) = 1$ we obtain

$$\sum_{n=2}^{\infty} P(\emptyset) = 0$$

implying that

$$P(\emptyset) = 0.$$

It should also be noted that it follows that for any finite number of mutually exclusive events $A_1, ..., A_n$

$$P\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} P(A_k). \tag{3}$$

In particular, for any two mutually exclusive events $A$ and $B$

$$P(A \cup B) = P(A) + P(B). \tag{4}$$

*Proof of (3):* It follows from Axiom 3 by defining $A_i$ to be the impossible event for all values of $i$ greater than $n$. Indeed,

$$P\left(\bigcup_{i=1}^{n} A_i \bigcup \left[\bigcup_{i=n+1}^{\infty} \emptyset\right]\right) = \sum_{i=1}^{n} P(A_i) + \sum_{i=n+1}^{\infty} P(\emptyset).$$

As $P(\emptyset) = 0$ we obtain (3).

Therefore Axiom 3 is valid both for finite number of events (see (3) and (4)) and for countable number of events.

**Property 2.** For any event $A$

$$P(\overline{A}) = 1 - P(A).$$

*Proof of Property 2:* We first note that $A$ and $\overline{A}$ are always mutually exclusive and since $A \cup \overline{A} = \Omega$ we have by Axiom 3 that

$$P(\Omega) = P(A \cup \overline{A}) = P(A) + P(\overline{A}) \quad \text{and by Axiom 2} \quad P(A) + P(\overline{A}) = 1.$$

The proof is complete.

As a special case we find that $P(\emptyset) = 1 - P(\Omega) = 0$, since the impossible event is the complement of $\Omega$.

**Property 3.** For any two events $A$ and $B$

$$P(B \setminus A) = P(B) - P(A \cap B). \tag{5}$$

*Proof:* The events $A \cap B$ and $B \cap \overline{A}$ are mutually exclusive, and their union is $B$. Therefore, by Axiom 3, $P(B) = P(A \cap B) + P(B \cap \overline{A})$, from which (5) follows immediately because $B \setminus A = B \cap \overline{A}$.

**Property 4.** If $A \subset B$ then

$$P(B \setminus A) = P(B) - P(A).$$

*Proof:* Property 4 is a corollary of Property 3.

**Property 5.** If $A \subset B$ then $P(A) \leq P(B)$, that is probability $P$ is nondecreasing function.

*Proof of Property 5:* As $P(B \setminus A) \geq 0$, then Property 5 implies from Property 4.

**Property 6.** For any event $A$

$$P(A) \leq 1.$$

Property 6 immediately follows from both Property 5 where we substitute $B = \Omega$ and from Axiom 2 (i. e. any event $A$ is a subevent of the certain event).

Therefore Axiom 1 and Property 6 state that the probability that the outcome of the experiment is contained in $A$ is some number between $0$ and $1$, i. e.

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

**Property 7.** For any two events $A$ and $B$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{6}$$

*Proof of Property 7:* It is not difficult to prove the following identity:

$$A \cup B = A \cup (B \cap \overline{A}),$$

where $A$ and $B \cap \overline{A}$ are mutually exclusive.

By Axiom 3 we get

$$P(A \cup B) = P(A) + P(B \cap \overline{A}). \tag{7}$$

Because $B \cap \overline{A} = B \setminus A$, by Property 3 we obtain

$$P(B \cap \overline{A}) = P(B) - P(A \cap B). \tag{8}$$

Substituting (8) into (7) we have (6). The Property 7 is proved.

**Example 11.** A card is selected at random from a deck of 52 playing cards. We will win if the card is either a club or a king. What is the probability that we will win?

*Solution:* Denote by $A$ the event that the card is clubs and by $B$ that it is a king. The desired probability is equal to $P(A \cup B)$. It follows from Property 7 that

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

As

$$P(A) = \frac{1}{4}, \quad P(B) = \frac{4}{52} \quad \text{and} \quad P(A \cap B) = \frac{1}{52}$$

we obtain

$$P(A \cup B) = \frac{1}{4} + \frac{4}{52} - \frac{1}{52} = \frac{4}{13}.$$

**Property 8 (Inclusion–Exclusion Principle).** For any events $A_1$, $A_2$, ..., $A_n$ we have

$$P\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} P(A_k) - \sum_{k<j} P(A_k \cap A_j) + \sum_{k<j<i} P(A_k \cap A_j \cap A_i) - ... + (-1)^{n-1} P\left(\bigcap_{k=1}^{n} A_k\right). \quad (9)$$

In words, formula (9) states that the probability of the union of $n$ events equals the sum of the probabilities of these events taken one at a time minus the sum of the probabilities of these events taken two at a time plus the sum of the probabilities of these events taken three at a time, and so on.

## §3. CONDITIONAL PROBABILITY

In this section we ask and answer the following question. Suppose we assign a probability to a sample space and then learn that an event $A$ has occurred. How should we change the probabilities of the remaining events? We call the new probability for an event $B$ the *conditional probability of* $B$ *given* $A$ and denote it by $P(B/A)$.

The probability of an event may depend on the occurrence (or nonoccurrence) of another event. If this dependence exists, the associated probability is a conditional probability. In the sample space $\Omega$, the conditional probability $P(B/A)$, means the likelihood of realizing an outcome in $B$ assuming that it belongs to $A$. In other words, we are interested in the event $B$ within the reconstituted sample space $A$. Hence, with the appropriate normalization, we obtain the conditional probability of $B$ given $A$.

**Definition 1.** Let $A$ and $B$ be two events on a sample space $\Omega$, on the subsets of which is defined a probability $P(\cdot)$. The conditional probability of the event $B$, given the event $A$, denoted by

$$P(B\,/A\,) = \frac{P(A \cap B)}{P(A)} \qquad \text{if} \quad P(A) \neq 0 \tag{1}$$

and if $P(A) = 0$, then $P(B\,/A\,)$ is undefined.
By multiplying both sides of equation (1) by $P(A)$ we obtain

$$P(A \cap B) = P(A) \cdot P(B\,/A\,) = P(B) \cdot P(A\,/B\,). \tag{2}$$

This formula is often useful as a tool to enable us in computing the desired probabilities more easily.

## §4. INDEPENDENCE AND DEPENDENCE

The notions of independent and dependent events play a central role in probability theory. If the events $A$ and $B$ have the property that the conditional probability of $B$, given $A$, is equal to the unconditional probability of $B$, one intuitively feels that event $B$ is statistically independent of $A$, in the sense that the probability of $B$ having occurred is not affected by the knowledge that $A$ has occurred.
Since $P(B\,/A\,) = \dfrac{P(A \cap B)}{P(A)}$ we see that $B$ is independent of $A$ if

$$P(A \cap B) = P(A) \cdot P(B). \tag{3}$$

As equation (3) is symmetric in $A$ and $B$, it shows that whenever $B$ is independent of $A$, $A$ is also independent of $B$. We thus have the following definition

**Definition 2.** Two events $A$ and $B$ are said to be independent if equation (3) holds. Two events $A$ and $B$ that are not independent are said to be dependent.

## §5. TOTAL PROBABILITY AND BAYES' FORMULAE

Sometimes the probability of an event $A$ cannot be determined directly. However, its occurrence is always accompanied by the occurrence of other events $B_i$, $i \geq 1$ such that the probability of $A$ will depend on which of the events $B_i$ has occurred. In such a case

the probability of $A$ will be an expected probability (that is, the average probability weighted by those of $B_i$). Such problems require the *Theorem of Total Probability.*
Let $A$ be a subevent of $\bigcup\limits_{n\geq 1} B_n$ (i. e. $A \subset \bigcup\limits_{n\geq 1} B_n$), $\{B_n\}$ be mutually exclusive and $P(B_n) \neq 0$ for any $n$. Then

$$P(A) = \sum_{n=1}^{\infty} P(B_n)\, P(A\,/B_n\,). \tag{4}$$

We will call formula (4) by *formula of Complete or Total Probability.*

**Proof of TOTAL PROBABILITY formula.**

Since $A$ is a subevent of the union of $B_n$, then

$$A = \bigcup_{n=1}^{\infty} (A \cap B_n),$$

where $\{A \cap B_n\}$ are mutually exclusive because $\{B_n\}$ are mutually exclusive. Therefore by a property of probability

$$P(A) = \sum_{n=1}^{\infty} P(A \cap B_n)$$

and by formula (2) $P(A \cap B_n) = P(B_n) \cdot P(A\,/B_n\,)$. The proof is complete.

**Example 1.** Urn I contains $6$ white and $4$ black balls. Urn II contains $5$ white and $2$ black balls. From urn I one ball is transferred to urn II. Then 2 balls are drawn without replacement from urn II. What is the probability that the two balls are white?

**Solution:** Denote by $B_1$ the event that from urn I a white ball is transferred to urn II, and by $B_2$ the event that from urn I is a black ball transferred to urn II. Denote by $A$ the event that from urn II are selected two white balls.
By formula (4)

$$P(A) = P(B_1)\, P(A\,/B_1\,) + P(B_2)\, P(A\,/B_2\,).$$

Since

$$P(B_1) = \frac{3}{5}, \quad P(B_2) = \frac{2}{5}, \quad P(A\,/B_1\,) = \frac{15}{28}, \quad P(A\,/B_2\,) = \frac{5}{14}$$

we obtain

$$P(A) = \frac{13}{28}.$$

Thus, for any event $A$ one may express the unconditional probability $P(A)$ of $A$ in terms of the conditional probabilities $P(A/B_1),...,P(A/B_n)...$ and the unconditional probabilities $P(B_1),...,P(B_n)....$

There is an interesting consequence of the formula of complete probability. Suppose that all conditions of the previous formula are satisfied. Then the following formula

$$P(B_i/A) = \frac{P(B_i)\cdot P(A/B_i)}{\sum\limits_{n=1}^{\infty} P(B_n)\,P(A/B_n)} \qquad i = 1, 2, .... \tag{5}$$

is known as *Bayes' formula*.

If we think of the events $B_n$ as being possible "hypotheses" about some subject matter, then Bayes' formula may be interpreted as showing us how opinions about these hypotheses held before the experiment [that is, the $P(B_n)$] should be modified by the evidence of the experiment.

Let us prove (5). Applying (2) to the events $A$ and $B_i$, we have

$$P(B_i)\cdot P(A/B_i) = P(A)\cdot P(B_i/A).$$

Therefore we obtain the desired probability

$$P(B_i/A) = \frac{P(B_i)\cdot P(A/B_i)}{P(A)}. \tag{6}$$

Using the total probability formula, (6) becomes (5).

**Example 2.** Suppose that there are $n$ balls in an urn. Every ball can be either white or black. The number of white and black balls in the urn is unknown. Let our aim be to find out that number. Determine the following hypotheses:

Denote by $B_i$ the event that the urn consists of exactly $i$ white balls (from $n$ balls), $i = 0, 1, 2, ..., n$.

As we do not have any additional information about the contents of the urn, therefore all hypotheses are equally likely, that is

$$P(B_i) = \frac{1}{n+1}, \quad \text{for any} \quad i = 0, 1, ..., n.$$

2-I. **Suppose that we have selected a ball and it is white (the event $A_1$).** It is obvious that we have to modify the probabilities of $B_i$. For example

$$P(B_0 / A_1) = 0.$$

What are the remained probabilities equal to? Let us use the Bayes' formula.

$$P(B_i / A_1) = \frac{P(B_i) \cdot P(A_1 / B_i)}{\sum\limits_{k=0}^{n} P(B_k) P(A_1 / B_k)} \qquad i = 0, 1, ...n.$$

Since

$$P(B_i) = \frac{1}{n+1}, \qquad P(A_1 / B_i) = \frac{i}{n}$$

we get

$$P(B_i / A_1) = \frac{\frac{i}{n}}{\sum\limits_{k=0}^{n} \frac{k}{n}} = \frac{2i}{n(n+1)}, \qquad i = 0, 1, ...n.$$

In particular, for $n = 3$ we obtain

$$P(B_0 / A_1) = 0, \quad P(B_1 / A_1) = \frac{1}{6}, \quad P(B_2 / A_1) = \frac{1}{3}, \quad P(B_3 / A_1) = \frac{1}{2}.$$

2-II. **Suppose that we have selected two balls and they are white (the event $A_2$).** It is obvious that we have to modify the probabilities of $B_i$. For example

$$P(B_0 / A_2) = 0, \qquad P(B_1 / A_2) = 0.$$

What are the remained probabilities equal to? Let us use the Bayes' formula.

$$P(B_i / A_2) = \frac{P(B_i) \cdot P(A_2 / B_i)}{\sum\limits_{k=0}^{n} P(B_k) P(A_2 / B_k)} \qquad i = 0, 1, ...n.$$

Since

$$P(B_i) = \frac{1}{n+1}, \qquad P(A_2 / B_i) = \frac{\binom{i}{2}}{\binom{n}{2}}$$

and

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$$

we get

$$P(B_i\,/A_2\,) = \frac{\binom{i}{2}\Big/\binom{n}{2}}{\sum\limits_{k=2}^{n}\binom{k}{2}\Big/\binom{n}{2}} = \frac{i(i-1)}{\sum\limits_{k=2}^{n}k^2 - \sum\limits_{k=2}^{n}k} = \frac{i(i-1)}{\sum\limits_{k=1}^{n}k^2 - \sum\limits_{k=1}^{n}k} = \frac{3i(i-1)}{(n-1)n(n+1)}, \qquad i = 0, 1, ...n.$$

In particular, for $n = 3$ we obtain

$$P(B_0\,/A_2\,) = 0, \quad P(B_1\,/A_2\,) = 0 \quad P(B_2\,/A_2\,) = \frac{1}{4}, \quad P(B_3\,/A_2\,) = \frac{3}{4}.$$

## APPENDIX–1.

### $P(\cdot\,/B\,)$ IS A PROBABILITY

**Conditional Probabilities satisfy all of the properties of ordinary probabilities. This is proved by Theorem 1 which shows that $P(\cdot\,/B\,)$ satisfies the three axioms of a probability.**

**Theorem 1.** *Conditional probability $P(A\,/B\,)$ as a function of event $A$ satisfies the following conditions:*

*a)* $P(A\,/B\,) \geq 0$ *for any $A$;*

*b)* $P(\Omega\,/B\,) = 1$*;*

*c) If events $A_i$ are mutually exclusive, then*

$$P\left(\bigcup_{i=1}^{\infty} A_i \Big/ B\right) = \sum_{i=1}^{\infty} P(A_i\,/B\,).$$

*Proof:* **Condition a) is obvious. Condition b) follows because**

$$P(\Omega\,/B\,) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1.$$

**Condition c) follows since**

$$P\left(\bigcup_{i=1}^{\infty} A_i \Big/ B\right) = \frac{P\left(\left(\bigcup_{i=1}^{\infty} A_i\right) \bigcap B\right)}{P(B)} = \frac{P\left(\bigcup_{i=1}^{\infty} (A_i \bigcap B)\right)}{P(B)} = \frac{\sum_{i=1}^{\infty} P(A_i \bigcap B)}{P(B)} = \sum_{i=1}^{\infty} P(A_i\,/B\,),$$

**where the next–to–last equality follows because $A_i \cap A_j = \emptyset$ implies that $(A_i \cap B) \cap (A_j \cap B) = \emptyset$. The proof is complete.**

**If we define $P_1(A) = P(A\,/B\,)$ (event $B$ is fixed and $P(B) \neq 0$), then it follows from Theorem 1 that $P_1(\cdot)$ may be regarded as a probability function on the events of the sample space $\Omega$. Hence all of the properties proved for probabilities apply to it.**

## APPENDIX-2.

Take a look at

https://www.youtube.com/watch?v=R13BD8qKeTg