Advanced Statistics

1. Probability Theory

Christian Aßmann

Chair of Survey Statistics and Data Analysis – Otto-Friedrich-Universität Bamberg

1.1 Sample Space and Events

The probability theory is the foundation upon which all of statistics and econometrics is built.

The objective of probability theory is to quantify the level of uncertainty associated with observing various outcomes of a chance situation (i.e. the possible outcomes of a random experiment or a random phenomenon).

For example, we might be interested in

- the level of uncertainty associated with the event that an ideal coin will turn up heads;
- the level of uncertainty associated with the event that the German GDP (gross domestic product) increases this year by more than 3%.

One of the fundamental tools to measure uncertainty is probability.

In the following, we will start to consider some important definitions (sample space and events) that are used to discuss probability.

DEFINITION (SAMPLE SPACE): A set, S, that contains all possible outcomes of a given experiment is called sample space.

EXAMPLE: If the experiment consists of tossing a die, the sample space contains six possible outcomes given by:

$$S = \{\boxed{\bullet}, \boxed{\begin{smallmatrix}\bullet & \\ & \bullet\end{smallmatrix}}, \boxed{\begin{smallmatrix}\bullet & \\ \bullet & \bullet\end{smallmatrix}}, \boxed{\begin{smallmatrix}\bullet & \bullet \\ \bullet & \bullet\end{smallmatrix}}, \boxed{\begin{smallmatrix}\bullet & \bullet \\ \bullet & \bullet \\ \bullet & \end{smallmatrix}}, \boxed{\begin{smallmatrix}\bullet & \bullet \\ \bullet & \bullet \\ \bullet & \bullet\end{smallmatrix}}\}. \quad ||$$

EXAMPLE: If the experiment consists of recording the number of traffic deaths in Germany next year, the sample space would contain all positive integers

$$S = \{0, 1, 2, ...\}. \quad ||$$

EXAMPLE: If the experiment consists of observing the length of life of a light bulb, the sample space would contain all positive numbers

$$S = (0, \infty). ||$$

REMARK: The sample space need not be identically equal to the set of all possible outcomes (could be larger). The only concern of practical importance is that the sample space is specified large enough to contain the set of all possible outcomes as a subset. $\diamondsuit$

The sample space S, as all sets, can be classified according to whether the number of elements in the set are

- finite (discrete sample space), e.g., $S = \{0, 1, 2, .., 6\}$
- countably infinite (discrete sample space), e.g., $S = \mathbb{N} = \{0, 1, 2, ...\}$
- uncountably infinite (continuous sample space), e.g., $S = \mathbb{R}$.

I come to the fundamental entities to which probabilities will be assigned, namely, the events.

DEFINITION (EVENT): An event, say A, is a subset of the sample space S (including S itself).

- Let A be an event, a subset of S. We say the event A occurs if the outcome of the experiment is in the set A.
- An event consisting of a single element or outcome is called elementary event.
- The event S is called the sure or certain event.

EXAMPLE: The experiment consists of tossing a die and counting the number of dots facing up. The sample space is defined to be $S = \{1, 2, ..., 6\}$. Consider the following subsets of S:

$$A_1 = \{1, 2, 3\}, \qquad A_2 = \{2, 4, 6\}, \qquad A_3 = \{6\}.$$

A_1 is an event whose occurrence means that the number of dots is less than four. A_2 is an event whose occurrence means that the number of dots is even. A_3 is an elementary event.
Note that the intersection of A_1 and A_2 and of A_1 and A_3 are

$$A_1 \cap A_2 = \{2\}, \qquad A_1 \cap A_3 = \emptyset$$

This means A_1 and A_3 can not, in contrast to A_1 and A_2, occur simultaneously. They are called mutually exclusive events. ||

REMARK: For each event A in S we want to associate a number between 0 and 1 that will be called the probability of A. For this purpose we will use an appropriate set function, say $P(\cdot)$, with the set of all events as the domain of $P(\cdot)$.

Hence, it would seem natural to define the domain of P (and hence the collection of all events) as the collection of all subsets of S. However, a technical problem arises for uncountable sample spaces such that certain subsets of S will not be considered as events because it will be impossible to assign probability to them in a consistent manner[1].

This issue is addressed in the definition of an event; it implies that every event is a subset of S but does not say that every subset of S is an event! $\diamondsuit$

[1] Subsets of S that can not be an event are so complicated that they are irrelevant for all practical purposes.

DEFINITION (EVENT SPACE): The set of all events in the sample space S is called the event space Y.

REMARK: The definitions of events and event space as the set of all events introduced above do not indicate which subset of S is an event belonging to the event space Y to which we want to assign probability. $\diamondsuit$

In the following we will use a collection of subsets of S which represents a sigma algebra in S as our event space Y. A sigma algebra is as defined as follows:

DEFINITION (SIGMA ALGEBRA): A collection of subsets of S is called a sigma algebra, denoted by $\mathcal{B}$, if it satisfies the following conditions:

(i) $\emptyset \in \mathcal{B}$ (empty set is an element of $\mathcal{B}$);

(ii) If $A \in \mathcal{B}$, then $\overline{A} \in \mathcal{B}$ ($\mathcal{B}$ is closed under complementation);

(iii) If $A_1, A_2, ... \in \mathcal{B}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$ ($\mathcal{B}$ is closed under countable unions).

REMARK: Property (i) states that the empty set is always in a sigma algebra. Since $S = \overline{\emptyset}$, property (i) and (ii) imply that S is always in a sigma algebra also. Hence, by using a sigma algebra in S as our events space we make sure that it contains the certain event. $\diamondsuit$

REMARK: An event space with property (ii) ensures the following: If A is an event (to which we can assign a certain probability), then $\overline{A}$ is also an event so that we can assign a probability that A does not occur. $\diamondsuit$

REMARK: An event space with property (iii) ensures the following: If $A_1, A_2, ...$ are events, then $\cup_{i=1}^{\infty} A_i$ is also an event. $\diamondsuit$

REMARK: Finally, by using DeMorgan's Law we obtain from properties (ii) and (iii) that $\cap_{i=1}^{\infty} A_i \in \mathcal{B}$ ($\mathcal{B}$ is also closed under countable intersections). This result obtains as follows:

If $A_1, A_2, ... \in \mathcal{B}$, then it follows from property (ii) that $\overline{A}_1, \overline{A}_2, ... \in \mathcal{B}$ and from property (iii) that $\cup_{i=1}^{\infty} \overline{A}_i \in \mathcal{B}$. According to property (ii) it follows that $\overline{\cup_{i=1}^{\infty} \overline{A}_i} \in \mathcal{B}$ and by DeMorgan's Law ($\overline{A \cup B} = \overline{A} \cap \overline{B}$) we have $\overline{\cup_{i=1}^{\infty} \overline{A}_i} = \cap_{i=1}^{\infty} \overline{\overline{A}}_i = \cap_{i=1}^{\infty} A_i \in \mathcal{B}$. $\diamondsuit$

Associated with the sample space S we can have many different sigma algebras. For example, the collection of the two sets $\{\emptyset, S\}$ is a sigma algebra $\mathcal{B}$ in S.

A typical sigma algebra used as event space Y if the sample space S is finite or countable is

$$\mathcal{B} = \{\text{all subsets of } S, \text{ including } S\}.$$

Note that if S has n elements there are 2^n sets in $\mathcal{B}$.

EXAMPLE: If $S = \{1, 2, 3\}$, then the sigma algebra consisting of all subsets of S is the following collection of $2^3 = 8$ sets:

$$\{1\},\ \{2\},\ \{3\},\ \{1, 2\},\ \{1, 3\},\ \{2, 3\},\ \{1, 2, 3\},\ \{\emptyset\}.||$$

As mentioned above, if S is uncountable a sigma algebra containing all subsets of S cannot be used as an event space.

A typical sigma algebra used as event space Y if the sample space is an interval on the real line, i.e., $S \subset \mathbb{R}$ is a

$\mathcal{B}$ containing all sets of all closed, open and half-open intervals:

$$[a, b],\ (a, b],\ [a, b),\ [a, b], \quad \forall a, b \in S,$$

as well as all sets that can be formed by taking (possibly countably infinite) unions and intersections of these intervals[2].

[2] This special sigma algebra is usually referred to as a collection of Borel sets (see, e.g., Mittelhammer, 1996, p.21).

1.2 Probability

Having defined the sample space S and the event space Y of an experiment, we are now in a position to define probability.

Before we discuss the axiomatic probability definitions used in probability theory, we consider important non-axiomatic probability definitions.

There are three major non-axiomatic definitions in the course of the development of probability, the classical probability, the relative frequency probability and the subjective probability.

DEFINITION (CLASSICAL PROBABILITY): Let S be the finite sample space for an experiment having $N(S)$[3] equally likely outcomes, and let $A \subset S$ be an event containing $N(A)$ elements. Then the probability of the event A, denoted by $P(A)$, is given by $P(A) = N(A)/N(S)$ (relative size of the event set A).

This probability concept was introduced by the French mathematician Pierre-Simon Laplace. According to this definition probabilities are images of sets generated by a set function, P, with a domain consisting of all subsets of a finite sample space S and with a range given by the interval $[0, 1]$.

[3] $N(\cdot)$ denotes the size-of-set function assigning to a set A the number of elements that are in set A.

Fig. 1: Pierre-Simon Laplace (1749-1827)
(Source: `http://en.wikipedia.org/wiki/laplace`*)*

EXAMPLE: The experiment consists of tossing a fair die and counting the number of dots facing up. The sample space with equally likely outcomes is $S = \{1, 2, ..., 6\}$. We have $N(S) = 6$. Let E_i $(i = 1, ..., 6)$ denote the elementary events in S. According to the classical definition we have

$$P(E_i) = \frac{N(E_i)}{N(S)} = \frac{1}{6}, \quad \text{and} \quad P(S) = \frac{N(S)}{N(S)} = 1 \text{ (probability of the certain event).}$$

For the event $A = \{1, 2, 3\}$ we obtain

$$P(A) = \frac{N(A)}{N(S)} = \frac{1}{2}.||$$

REMARK: The classical definition has two major drawbacks.

- First, its use requires that the sample space is finite. Note that for infinite sample spaces we have $N(S) = \infty$ and possibly $N(A) = \infty$.
- Second, its use requires that the outcomes of an experiment must be equally likely. Hence, the classical definition cannot be used in an experiment consisting, e.g., of tossing an unfair die. $\diamondsuit$

A probability definition which does not suffer from these limitations is the relative frequency probability which is defined as follows:

DEFINITION (RELATIVE FREQUENCY PROBABILITY): Let n be the number of times that an experiment is repeatedly performed under identical conditions. Let A be an event in the sample space S, and define n_A to be the number of times in n repetitions of the experiment that the event A occurs. Then the probability of the event A is given by the limit of the relative frequency n_A/n, as $P(A) = \lim_{n\to\infty} \frac{n_A}{n}$.

According to this definition the probability of an event A is the image of A generated by a set function P, where the image is defined as the limiting fraction of the number of outcomes in a sequence of experiments that are observed to be elements in A.

Note that the range of the set function is the interval $[0, 1]$, since $0 \leq n_A \leq n$.

EXAMPLE: The experiment consists of tossing a coin with $S = \{\text{head, tail}\}$. The coin was tossed various numbers of times, with the following results:

n (No. of tosses)	n_{head} (No. of heads)	n_{head}/n (Rel. freq.)
100	48	.4800
500	259	.5180
5000	2,509	.5018

It would appear that $\lim_{n\to\infty}(n_{\text{head}}/n) = 1/2$. ||

REMARK: The frequency definition allows – in contrast to the classical definition – for an infinite sample space S as well as for outcomes which are not equally likely. However, the frequency definition has the following drawbacks:

- First, while for many types of experiments n_A/n will converge to a limit value (such as in our coin-tossing example) we can not exclude situations where the limit of n_A/n does not exist.
- Second, how could we ever observe the limiting value if an infinite number of repetitions of the experiment is required? $\diamondsuit$

A third approach to defining probability is the subjective probability which is defined as follows:

DEFINITION (SUBJECTIVE PROBABILITY): The subjective probability of an event A is a real number, $P(A)$, in the interval $[0, 1]$, chosen to express the degree of personal belief in the likelihood of occurrence or validity of event A, the number 1 being associated with certainty.

Like the classical and frequency definition of probability, subjective probabilities can be interpreted as images of set functions. However, $P(A)$ as the image of A can obviously vary depending on who is assigning the probability [4].

REMARK: The subjective probability definition has the following properties:

- Unlike the relative frequency approach, subjective probabilities can be defined for experiments that cannot be repeated. For example, consider the event that the social democrats will win the next election. This does not fit into the frequency definition of probability, since one can observe the outcome of that election once.

(CONTINUES)

[4] The subjective probability plays a crucial role in Bayesian statistics and Decision theory.

Remark (continued):

- Often we are interested in the 'true' likelihood of an event and not in our personal perceptions. For example, if we consider some game of chance (such as a lottery or roulette) we are typically interested in the loss probability as a result of the particular construction of the game. $\diamondsuit$

1.4 Axiomatic Probability Definition

In this subsection we consider the axiomatic definition of probability which is the fundament of modern probability theory and of modern statistics. This axiomatic definition was introduced by the Russian mathematician Andrey N. Kolmogorov[5].

It consists of a set of axioms defining desirable mathematical properties of the measure (P) which we use in order to assign probabilities to events.

As we shall see, the axiomatic definition of probability is general enough to accommodate all the non-axiomatic concepts discussed above.

[5] See Andrey N. Kolmogorov (1956), Foundations of the Theory of Probability, New York: Chelsea; the original German version (Grundbegriffe der Wahrscheinlichkeitsrechnung) appeared in 1933.

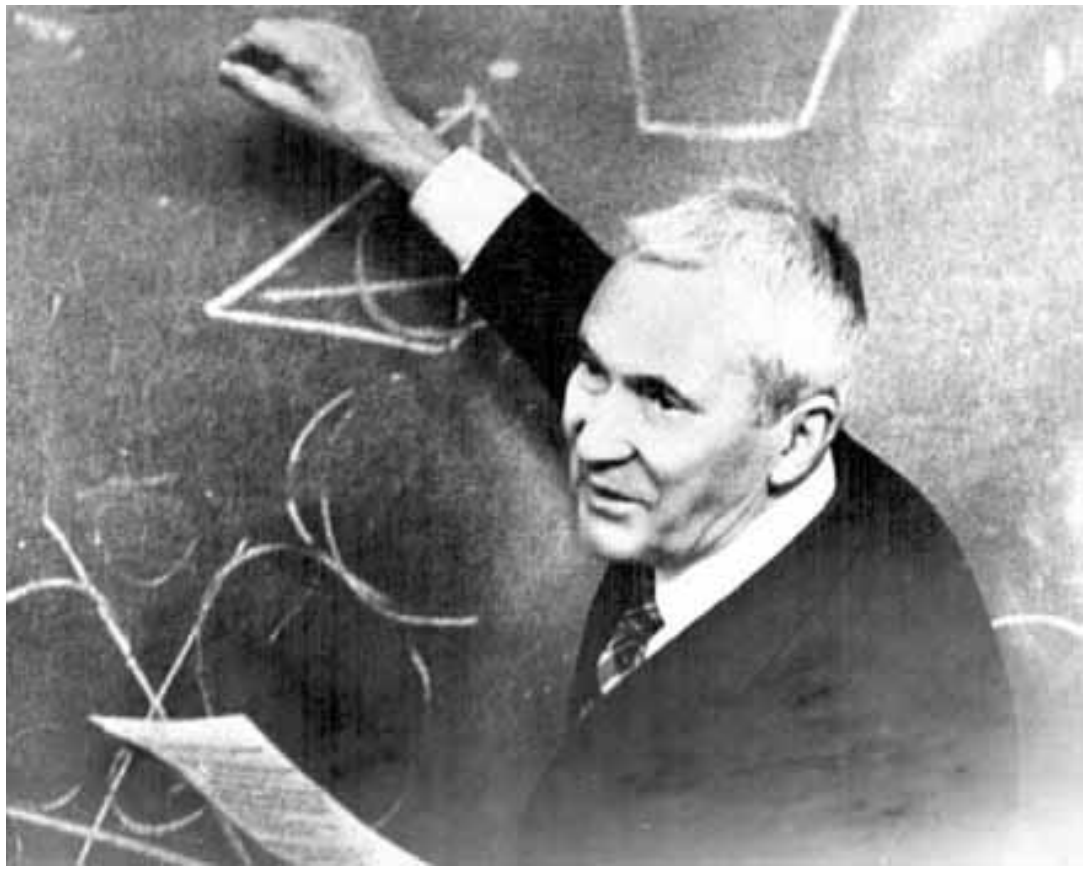


Fig. 2: Andrey Nikolaevich Kolmogorov (April 25, 1903 – October 20, 1987)
(Source: `http://en.wikipedia.org/wiki/Andrey Nikolaevich Kolmogorov`*)*

DEFINITION (PROBABILITY FUNCTION): Given a sample space S and an associated event space Y (a sigma algebra on S), a probability (set) function is a set function P with domain Y that satisfies the following axioms:

(Axiom 1) $P(A) \geq 0$ for all $A \in Y$ (non-negativity).

(Axiom 2) $P(S) = 1$ (standardization).

(Axiom 3) If $A_1, A_2, ... \in Y$ is a sequence of disjoint events (that is, $A_i \cap A_j = \emptyset$ for $i \neq j$; $i, j = 1, 2, ...$), then $P\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P\left(A_i\right)$ (additivity).

REMARK: This definition tells us which set functions can be used as probability set functions to assign probabilities; it does not tell us what value the probability set function P assigns to a given event and it makes no attempt to tell what particular set function P to choose. For any sample space S many different probability functions can be defined. $\diamondsuit$

EXAMPLE: Let $S = \{1, 2, ..., 6\}$ be the sample space for rolling a fair die and observing the number of dots facing up. The set function

$$P(A) = N(A)/6 \quad \text{for} \quad A \subset S$$

(where $N(A)$ is the size of set A) represents a probability set function on the events of S. We can verify this by noting that

- the value of the function $P(A) \geq 0$ for all $A \subset S$ (non-negativity);
- the value of the function for the set S is $P(S) = N(S)/6 = 1$ (standardization);
- for any collection of disjoint sets $A_1, A_2, ..., A_n$ we have

$$P(\cup_{i=1}^{n} A_i) = \frac{N(\cup_{i=1}^{n} A_i)}{6} = \frac{\sum_{i=1}^{n} N(A_i)}{6} = \sum_{i=1}^{n} P(A_i) \quad \text{(additivity)}.||$$

EXAMPLE: Let the sample space be $S = \{1, 2, ...\} = \mathbb{N}$ and consider the set function

$$P(A) = \sum_{x \in A} \left(\frac{1}{2}\right)^x \quad \text{for} \quad A \subset S.$$

This set function represents a probability set function since

- the value of the function $P(A) \geq 0$ for all $A \subset S$, because P is defined as the sum of non-negative numbers (non-negativity);
- the value of the function for the set S is

$$P(S) = \sum_{x \in S} \left(\frac{1}{2}\right)^x = \sum_{x=1}^{\infty} \left(\frac{1}{2}\right)^x = \underbrace{\sum_{x=0}^{\infty} \left(\frac{1}{2}\right)^x}_{\text{infinte geom. series}} - 1 = \frac{1}{1 - \frac{1}{2}} - 1 = 1$$

  (standardization);
- for any collection of disjoint sets $A_1, A_2, ..., A_n$ we have

$$P(\cup_{i=1}^n A_i) = \sum_{x \in (\cup_{i=1}^n A_i)} \left(\frac{1}{2}\right)^x = \sum_{i=1}^n \left[\sum_{x \in A_i} \left(\frac{1}{2}\right)^x\right] = \sum_{i=1}^n P(A_i) \quad \text{(additivity)}. ||$$

EXAMPLE: Let $S = [0, \infty)$ be the sample space for an experiment consisting of observing the length of life of a light bulb and consider the set function

$$P(A) = \int_{x \in A} \frac{1}{2} e^{-\frac{x}{2}} dx \quad \text{for} \quad A \in Y.$$

This set function represents a probability set function since

- the value of the function $P(A) \geq 0$ for all $A \subset S$, because P is defined as an integral with a non-negative integrand (non-negativity);
- the value of the function for the set S is

$$P(S) = \int_{x \in S} \frac{1}{2} e^{-\frac{x}{2}} dx = \int_0^{\infty} \frac{1}{2} e^{-\frac{x}{2}} dx = 1 \quad \text{(standardization)};$$

- for any collection of disjoint sets $A_1, A_2, ..., A_n$ we have

$$P(\cup_{i=1}^n A_i) = \underbrace{\int_{x \in (\cup_{i=1}^n A_i)} \frac{1}{2} e^{-\frac{x}{2}} dx = \sum_{i=1}^n \left[\int_{x \in A_i} \frac{1}{2} e^{-\frac{x}{2}} dx \right]}_{A_i's \text{ are non-overlapping intervals: addititivity property of Riemann integrals}} = \sum_{i=1}^n P(A_i)$$

(additivity). ||

REMARK: Once we have defined the 3-tuple $\{S, Y, P\}$ (called Probability space) for an experiment under consideration all of the information is established that is needed to assign probabilities to the various events of interest.

It is the discovery of the appropriate probability set function P that represents the major challenge in statistical real-life applications. This is the objective of statistical inference procedures (inferential statistics) to be discussed in the course next semester. $\diamondsuit$

The three axioms governing the behavior of a probability function entail many properties of the probability function. Some of these properties will be discussed in the next subsection.

THEOREM 1.1 *Let A be an event in S. Then $P(A) = 1 - P\left(\bar{A}\right)$.*

PROOF: The sets A and $\bar{A}$ form a partition of S, that is, $A \cup \bar{A} = S$ with $A \cap \bar{A} = \emptyset$. Therefore

$$P(S) = P(A \cup \bar{A}) \overset{(Ax.3)}{=} P(A) + P(\bar{A}) \overset{(Ax.2)}{=} 1$$

Solving the last Equation for $P(A)$ obtains the result. □

THEOREM 1.2 $P(\emptyset) = 0$.

PROOF: Since $\bar{\emptyset} = S$ we immediately have

$$P(\emptyset) \overset{(Th.1.1)}{=} 1 - P(S) \overset{(Ax.2)}{=} 1 - 1 = 0. \square$$

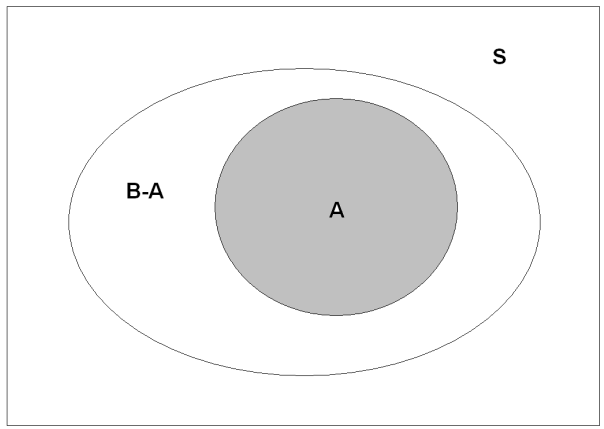
THEOREM 1.3 *Let A and B be events in S such that $A \subset B$. Then $P(A) \leq P(B)$ and $P(B - A) = P(B) - P(A)$.*

PROOF: Since $A \subset B$, we have $A \cap (B - A) = \emptyset$ and $A \cup (B - A) = B$ (see Fig. 3) and thus

$$P(B) \overset{(Ax.3)}{=} P(A) + P(B - A).$$

The second result of the theorem follows immediately. Since $P(B - A) \geq 0$ by Axiom 1, we also have $P(A) \leq P(B)$. $\square$

Fig. 3.



THEOREM 1.4 *Let A and B be events in S. Then $P(A) = P(A \cap B) + P(A \cap \bar{B})$.*

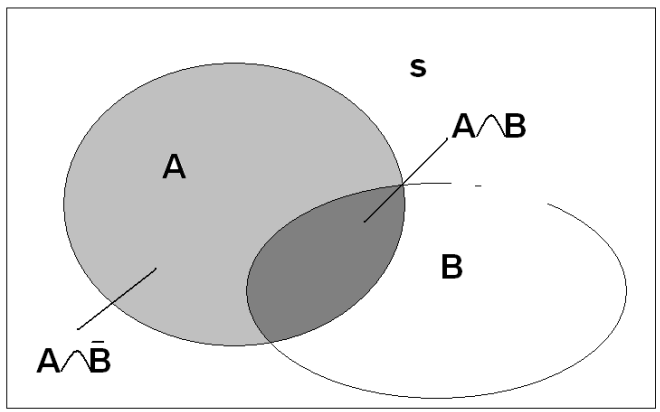
PROOF: The set A can be written as

$$\begin{aligned} A = A \cap S &= A \cap (B \cup \bar{B}) \\ &= (A \cap B) \cup (A \cap \bar{B}) \quad \text{(intersection is distributive).} \end{aligned}$$

Since $(A \cap B) \cap (A \cap \bar{B}) = \emptyset$ (see Fig. 4), we have by Axiom 3

$$P(A) = P[(A \cap B) \cup (A \cap \bar{B})] \overset{(Ax.3)}{=} P(A \cap B) + P(A \cap \bar{B}). \square$$

Fig. 4.



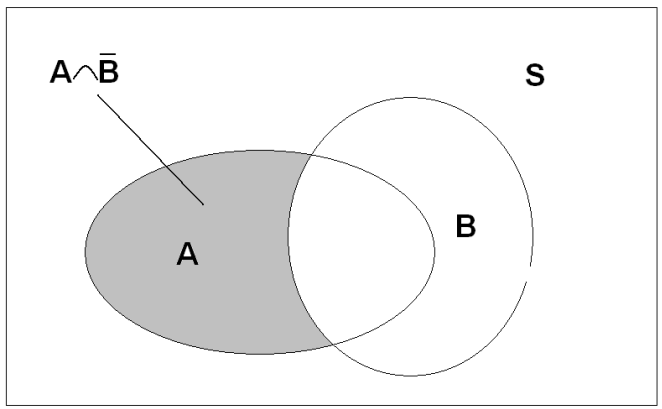
THEOREM 1.5 *Let A and B be events in S. Then* $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

PROOF: The union of A and B can be written as $A \cup B = B \cup (A \cap \bar{B})$, where $B \cap (A \cap \bar{B}) = \emptyset$ (see Fig. 5). Therefore

$$\begin{aligned} P(A \cup B) &= P[B \cup (A \cap \bar{B})] \\ &\overset{(Ax.3)}{=} P(B) + P(A \cap \bar{B}) \\ &\overset{(Th.1.4)}{=} P(B) + P(A) - P(A \cap B). \square \end{aligned}$$

COROLLARY 1.1 (BOOLE'S INEQUALITY) $P(A \cup B) \leq P(A) + P(B)$. *(Follows from Theorem 1.5 since* $P(A \cap B) \geq 0$*.)*

Fig. 5.



THEOREM 1.6 *Let A be an event in S. Then $P(A) \in [0,1]$.*

PROOF: The fact that $\emptyset \subset A$ implies that $P(\emptyset) \leq P(A)$ (Theorem 1.3) and the fact that $A \subset S$ implies that $P(A) \leq P(S)$ (Theorem 1.3). Since $P(S) = 1$ and $P(\emptyset) = 0$, we have $0 \leq P(A) \leq 1$. $\square$

THEOREM 1.7 (BONFERRONI'S INEQUALITY) *Let A and B be events in S. Then $P(A \cap B) \geq 1 - P(\bar{A}) - P(\bar{B})$.*

PROOF: By Theorem 1.1 we have $P(A \cap B) = 1 - P(\overline{A \cap B})$. DeMorgan's law states that $\overline{A \cap B} = \bar{A} \cup \bar{B}$. Therefore

$$\begin{aligned} P(A \cap B) &= 1 - P(\bar{A} \cup \bar{B}) \\ &\overset{(Th.1.5)}{=} 1 - P(\bar{A}) - P(\bar{B}) + P(\bar{A} \cap \bar{B}). \end{aligned}$$

Since $P(\bar{A} \cap \bar{B}) \geq 0$ (Axiom 1) we have $P(A \cap B) \geq 1 - P(\bar{A}) - P(\bar{B})$. $\square$

THEOREM 1.8 *Let $A_1, \ldots, A_n$ be events in S. Then $P\left(\cap_{i=1}^{n} A_i\right) \geq 1 - \sum_{i=1}^{n} P(\bar{A}_i)$.*

PROOF: The proposition can be proven by induction using the base case $n = 2$, for which the statement holds according to Theorem 1.7 (for further details, see Mittelhammer, 1996, p.17). □

THEOREM 1.9 (CLASSICAL PROBABILITY) *Let S be the finite sample space for an experiment having $n = N(S)$ equally likely outcomes, say $E_1, ..., E_n$, and let $A \subset S$ be an event containing $N(A)$ elements. Then the probability of the event A is given by $N(A)/N(S)$.*

PROOF: Since all outcomes are equally likely with $P(E_1) = ... = P(E_n) \overset{(say)}{=} k$, and since $S = \cup_{i=1}^{n} E_i$ with $E_i \cap E_j = \emptyset \ \forall i \neq j$, we have by Axioms 2 and 3 that

$$P(S) = P(\cup_{i=1}^{n} E_i) \overset{(Ax.3)}{=} \sum_{i=1}^{n} P(E_i) = nk \overset{(Ax.2)}{=} 1.$$

It follows that $P(E_i) = 1/n$. Let $I \subset \{1, ..., n\}$ be the index set identifying the $N(A)$ number of outcomes that define A, that is, $A = \cup_{i \in I} E_i$. Then by Axiom 3 we have

$$P(A) = P(\cup_{i \in I} E_i) \overset{(Ax.3)}{=} \sum_{i \in I} P(E_i) = \sum_{i \in I} \frac{1}{n} = \frac{N(A)}{N(S)}. \square$$

REMARK: Theorem 1.9 states that the classical probability definition is implied by the axiomatic definition. Thus whenever the conditions of the classical definition (finite S with equally likely outcomes) apply, we can use the classical definition to assign probabilities to events. $\diamondsuit$

1.6 Conditional Probability

So far, we have considered probabilities of events on the assumption that no information was available about the experiment other than the sample space S.

Sometimes, however, it is known that an event B has happened. The question is then, how can we use this information in making a statement concerning the outcome of another event A, that is, how can we update the probability calculation for the event A based on the information that B has happened.

EXAMPLE: The experiment consists of tossing two fair coins. The sample space is $S = \{(\text{H,H}), (\text{H,T}), (\text{T,H}), (\text{T,T})\}$ (H= Head, T=Tail). Consider the events

$$A = \{\text{both coins show same face}\}, \quad B = \{\text{at least one coin shows H}\}.$$

Then $P(A) = 2/4 = 1/2$. If B is known to have happened, we know for sure that the outcome (T,T) cannot happen. This suggest that

$$P(A \text{ conditional on the information that } B \text{ has happened}) = 1/3.||$$

This update of the probability calculation is the calculation of conditional probability which is defined as follows.

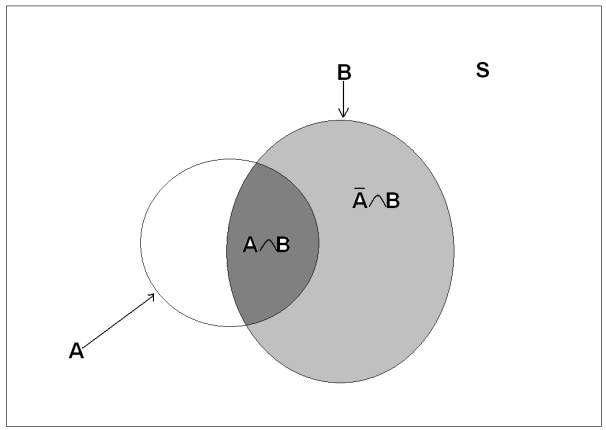
Definition (Conditional probability): Let A and B be any two events in a sample space S. If $P(B) \neq 0$, then the conditional probability of event A, given event B, is given by $P(A \mid B) = P(A \cap B)/P(B)$.

Note that what happens in the conditional probability calculation is that B becomes the sample space:

$$P(B|B) \stackrel{Def.}{=} P(B \cap B)/P(B) = P(B)/P(B) = 1.$$

The intuition is that the original sample space S, has been updated to B (since it is given that B occurs). Note further that since B will occur, it is clear that A occurs iff A occurs concurrently with B, that is iff $A \cap B$ occurs. Hence, $P(A \mid B) \propto P(A \cap B)$ (see Fig. 6). The division by $P(B)$ ensures that $P(A|B)$, as defined, represents a probability function.

Fig. 6.



REMARK: It is clear that the conditional probabilities as defined above are values of a set function. That these are values of a probability set function satisfying the Axioms 1–3 is established in the following theorem. ♢

THEOREM *Given a probability space* $\{S, Y, P\}$ *and an event* B *for which* $P(B) \neq 0$, $P(A \mid B) = P(A \cap B)/P(B)$ *defines a probability set function with domain* Y.

PROOF: To proof the theorem we need to show that the set function $P(A \mid B) = P(A \cap B)/P(B)$ adheres to the Axioms 1–3 of probability on the domain Y.

- Clearly, $P(A|B) \geq 0$ for all $A \in Y$, since $P(A \cap B) \geq 0$ and $P(B) \geq 0$.
- Also, $P(S|B) = 1$, since $P(S|B) = P(S \cap B)/P(B) = P(B)/P(B)$.
- If $A_1, A_2, \ldots$ is a disjoint sequence of sets in Y, then

$$
\begin{aligned}
P(\cup_{i=1}^{\infty} A_i|B) &= P[(\cup_{i=1}^{\infty} A_i) \cap B]/P(B) \quad \text{(by def. of conditional probability)} \\
&= P[\cup_{i=1}^{\infty} (A_i \cap B)]/P(B) \quad \text{(since } \cap \text{ is distributive)} \\
&= \sum_{i=1}^{\infty} P(A_i \cap B)/P(B) \\
&\qquad \text{(since } (A_i \cap B) \cap (A_j \cap B) = \emptyset \text{ for } i \neq j\text{)} \\
&= \sum_{i=1}^{\infty} P(A_i|B) \quad \text{(by def. of conditional probability).} \quad \square
\end{aligned}
$$

REMARK: Since $P(A|B)$ adheres to the probability axioms, all of the properties that we have discussed for unconditional probabilities apply to conditional probabilities (see Mittelhammer, 1996 Theorem 1.1^c-1.8^c). $\diamondsuit$

EXAMPLE: The experiment consists of tossing two fair coins. The sample space is $S = \{(\text{H,H}), (\text{H,T}), (\text{T,H}), (\text{T,T})\}$. The conditional probability of the event obtaining two heads

$$A = \{(\text{H,H})\},$$

given the first coin toss results in heads

$$B = \{(\text{H,H}), (\text{H,T})\}$$

is

$$P(A|B) = P(A \cap B)/P(B) \overset{(\textit{classical def.})}{=} (1/4)/(1/2) = 1/2.||$$

The definition of conditional probability can be transformed to obtain the multiplication rule. It allows one to factorize the joint probability for the events A and B into the conditional probability for event A, given event B and the unconditional probability of B.

THEOREM (MULTIPLICATION RULE) *Let A and B be any two events in S for which $P(B) \neq 0$. Then $P(A \cap B) = P(A \mid B)P(B)$.*

PROOF: The proof follows from the definition of conditional probability. □

EXAMPLE: A computer manufacturer has quality-control inspectors examine every produced computer. A computer is shipped to a retailer only if it passes inspection.

- The probability that a computer is defective, say event D, is $P(D) = 0.02$.
- The probability that an inspector assigns a 'pass' (event A) to a defective computer (that is given event D) is $P(A|D) = 0.05$.

The joint probability that a computer is defective (D) and shipped to the retailer (A) is $P(A \cap D) = P(A \mid D)P(D) = 0.05 \cdot 0.02 = 0.001$. ||

The multiplication rule can be extended to more than two events as follows:

THEOREM (EXTENDED MULTIPLICATION RULE) *Let $A_1, A_2, \ldots, A_n$, $n \geq 2$, be events in S. Then if all of the conditional probabilities exist,*

$$\begin{aligned} P\left(\cap_{i=1}^{n} A_i\right) &= P(A_1) \cdot P(A_2|A_1) \cdot \ldots \cdot P(A_n|A_{n-1} \cap A_{n-2} \cap \ldots \cap A_1) \\ &= P(A_1) \prod_{i=2}^{n} P\left(A_i \mid \cap_{j=1}^{i-1} A_j\right). \end{aligned}$$

PROOF: Let $B = \cap_{i=1}^{n-1} A_i$ such that $P\left(\cap_{i=1}^{n} A_i\right) = P(A_n \cap B)$. Hence we have by the multiplication rule for $n = 2$

$$P\left(\cap_{i=1}^{n} A_i\right) = P(A_n|B)P(B) = P(A_n| \cap_{i=1}^{n-1} A_i) \cdot P(\cap_{i=1}^{n-1} A_i).$$

Now consider the last factor of the last equation and let $C = \cap_{i=1}^{n-2} A_i$ such that $P\left(\cap_{i=1}^{n-1} A_i\right) = P(A_{n-1} \cap C) = P(A_{n-1}|C)P(C)$. Hence we have

$$P\left(\cap_{i=1}^{n} A_i\right) = P(A_n| \cap_{i=1}^{n-1} A_i) \cdot P(A_{n-1}| \cap_{i=1}^{n-2} A_i) \cdot P(\cap_{i=1}^{n-2} A_i).$$

Sequentially repeating this factorization of the last factor obtains the result. □

Definition (Independence of events (2–event case)): Let A and B be two events in S. Then A and B are independent iff $P(A \cap B) = P(A)P(B)$. If A and B are not independent, A and B are said to be dependent events.

An intuitively appealing interpretation of independence obtains by considering its implication for conditional probabilities. In particular, independence of A and B implies

$$\begin{aligned} P(A|B) &= P(A \cap B)/P(B) = P(A)P(B)/P(B) = P(A), \text{ as long as } P(B) > 0 \\ P(B|A) &= P(B \cap A)/P(A) = P(B)P(A)/P(A) = P(B), \text{ as long as } P(A) > 0. \end{aligned}$$

Thus the probability of event A occurring is unaffected by the occurrence of event B, and vice versa.

Independence of A and B implies independence of the complements also. In fact we have the following theorem:

THEOREM *If events A and B are independent, then events A and $\bar{B}$, $\bar{A}$ and B, and $\bar{A}$ and $\bar{B}$ are also independent.*

PROOF: To establish the independence of A and $\bar{B}$ note that

$$\begin{aligned} P(A \cap \bar{B}) &= P(A) - P(A \cap B) \quad \text{(Theorem 1.4)} \\ &= P(A) - P(A)P(B) \quad \text{(independence of } A \text{ and } B\text{)} \\ &= P(A)[1 - P(B)] \\ &= P(A)P(\bar{B}) \quad \text{(Theorem 1.1).} \end{aligned}$$

The independence of $\bar{A}$ and B obtains analogously. To establish the independence of $\bar{A}$ and $\bar{B}$ note that

$$\begin{aligned} P(\bar{A} \cap \bar{B}) &= P(\overline{A \cup B}) \quad \text{(DeMorgans law)} \\ &= 1 - P(A \cup B) \quad \text{(Theorem 1.1)} \\ &= 1 - [P(A) + P(B) - P(A \cap B)] \quad \text{(Theorem 1.5)} \\ &= 1 - P(A) - P(B) + P(A)P(B) \quad \text{(independence of } A \text{ and } B\text{)} \\ &= 1 - P(A) - P(B)[1 - P(A)] \\ &= P(\bar{A}) - P(B)P(\bar{A}) \quad \text{(Theorem 1.1)} \\ &= P(\bar{A})[1 - P(B)] = P(\bar{A})P(\bar{B}) \quad \text{(Theorem 1.1).} \square \end{aligned}$$

The following example illustrates the concept of independence.

EXAMPLE: The work force of a company has the following distribution among type and gender of workers:

	Type of worker			
Sex	Sales	Clerical	Production	Total
Male	825	675	750	2,250
Female	1,675	825	250	2,750
Total	2,500	1,500	1,000	5000

The experiment consists of randomly choosing a worker and observing type and sex. Is the event of observing a female (A) and the event of observing a clerical worker (B) independent?

According to the classical probability we obtain $P(A) = 2,750/5000 = 0.55$ and $P(B) = 1,500/5000 = 0.30$ with $P(A)P(B) = 0.165$. Also, $P(A \cap B) = 825/5000 = 0.165$. Hence A and B are independent. ||

The concept of independent events can be generalized to more than two events as follows:

Definition (Independence of events (n–event case)):
Let $A_1, A_2, \ldots, A_n$, be events in the sample space S. The events $A_1, A_2, \ldots, A_n$ are (jointly) independent iff

$$P(\cap_{j\in J} A_j) = \prod_{j\in J} P\left(A_j\right), \quad \text{for all subsets} \quad J \subset \{1, 2, \ldots, n\}$$

for which $N(J) \geq 2$. If the events $A_1, A_2, \ldots, A_n$ are not independent, they are said to be dependent events.

Remark: Note that this definition requires that the joint probability of any subcollection of events from $A_1, A_2, \ldots, A_n$ can be factorized. In the case of $n = 3$ events, joint independence requires:

$$P(A_1 \cap A_2) = P(A_1)P(A_2),\ P(A_1 \cap A_3) = P(A_1)P(A_3),\ P(A_3 \cap A_2) = P(A_3)P(A_2),$$

(pairwise independence) and

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3).\diamondsuit$$

The following example illustrates that pairwise independence between all pairs of events (A_i, A_j) is not sufficient for joint independence of $A_1, A_2, \ldots, A_n$.

EXAMPLE: Let the sample space S consists of all permutations of the letters a, b, c along with three triples of each letter, that is,

$$S = \{\text{aaa}, \text{bbb}, \text{ccc}, \text{abc}, \text{bca}, \text{cba}, \text{acb}, \text{bac}, \text{cab}\}.$$

Furthermore, let each element of S have probability $1/9$. Consider the events

$$A_i = \{i\text{th place in the triple is occupied by a}\}.$$

According to the classical probability we obtain for all $i = 1, 2, 3$

$$P(A_i) = 3/9 = 1/3, \quad \text{and} \quad P(A_1 \cap A_2) = P(A_1 \cap A_3) = P(A_2 \cap A_3) = 1/9,$$

so A_1, A_2, A_3 are pairwise independent. But they are not jointly independent since

$$P(A_1 \cap A_2 \cap A_3) = 1/9 \neq P(A_1)P(A_2)P(A_3) = 1/27.||$$

1.8 Total Probability and Bayes's Rule

Bayes's rule provides an alternative representation of conditional probabilities. This representation provides the means for reevaluating the probability of an event B, given the additional information that event A occurs. By this rule the probability of the event B is, in effect, updated in light of the additional information provided by the occurrence of event A.

This rule was discovered by the English clergyman and mathematician Thomas Bayes.

Bayes's rule is a simple consequence of the total probability rule established in the following theorem:

THEOREM (THEOREM OF TOTAL PROBABILITY) *Let the events $B_i, i \in I$, be a finite or countably infinite partition of S, so that $B_j \cap B_k = \emptyset$ for $j \neq k$, and $\cup_{i \in I} B_i = S$. Let $P(B_i) > 0 \; \forall \; i$. Then total probability of event A is*

$$P(A) = \sum_{i \in I} P(A \mid B_i) P(B_i).$$

PROOF: First note that

$$A = A \cap S = A \cap (\cup_{i \in I} B_i) = \cup_{i \in I} (A \cap B_i) \quad \text{(since } \cap \text{ is distributive)}.$$

From $B_j \cap B_k = \emptyset$ it follows that $(A \cap B_j) \cap (A \cap B_k) = \emptyset$ for all $j \neq k$. By Axiom 3 and the multiplication rule we have

$$\begin{aligned} P(A) &= P[\cup_{i \in I} (A \cap B_i)] \\ &= \sum_{i \in I} P(A \cap B_i) \quad \text{(by Axiom 3)} \\ &= \sum_{i \in I} P(A|B_i) P(B_i) \quad \text{(by multiplication rule).} \square \end{aligned}$$

Note that the total probability rule states that the total (unconditional) probability of an event A can be represented by the sum of conditional probabilities given the events B_i weighted by the unconditional probabilities $P(B_i)$.

The Bayes's rule obtains as the following corallary to the Theorem of Total Probability:

COROLLARY (BAYES'S RULE) *Let the events $B_i, i \in I$, be a finite or countably infinite partition of S, so that $B_j \cap B_k = \emptyset$ for $j \neq k$ and $\cup_{i \in I} B_i = S$. Let $P(B_i) > 0 \; \forall \; i \in I$. Then, provided $P(A) \neq 0$,*

$$P(B_j \mid A) = \frac{P(A \mid B_j)P(B_j)}{\sum_{i \in I} P(A \mid B_i)P(B_i)}, \quad \forall \, j \in I.$$

PROOF: By the definition of the conditional probability for $B_j|A$, the multiplication rule and the total probability rule we immediately have

$$P(B_j \mid A) = \frac{P(B_j \cap A)}{P(A)} = \frac{P(A \mid B_j)P(B_j)}{\sum_{i \in I} P(A \mid B_i)P(B_i)}.\square$$

REMARK: In the case of two events with $S = B \cup \bar{B}$ Bayes's rule implies

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid \bar{B})P(\bar{B})},$$

which obtains by setting $I = \{1, 2\}$. $\diamondsuit$

EXAMPLE: Consider a blood test for detecting a certain disease. Let A be the event that the test result is positive and B be the event that the individual has the disease. The test detects with probability 0.98 the disease given that the disease is, in fact, in the individual being tested, that is, $P(A|B) = 0.98$. The test yields a 'false positive' result for 1 percent, that is, $P(A|\bar{B}) = 0.01$.

0.1 percent of the population has the disease which implies that $P(B) = 0.001$. What is the probability that a randomly chosen person to be tested actually has the disease if the test result is positive?

The application of Bayes's rule yields

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid \bar{B}) \underbrace{P(\bar{B})}_{1-P(B)}} = \frac{.98 \cdot .001}{.98 \cdot .001 + .01 \cdot .999} = .089.$$

Hence, Bayes's rule provides the means for updating the (unconditional) probability of the event B, given the information signal that the event A occurs. ||