Fundamental Concepts of Probability

ECE275A – Lecture Supplement – Fall 2008

Ken Kreutz-Delgado
Electrical and Computer Engineering
Jacobs School of Engineering
University of California, San Diego

VERSION ECE275A.F08.LSProb.v1.0

September 24, 2008

1. Modeling Random Events

Mathematical Model. A mathematical model is intended to capture important properties of a real-world phenomenon of interest without necessarily catching all of the fine details. It is important to distinguish between a model and the reality which it is supposed to represent.

Models are justified after-the-fact, by how well they explain and predict the measured real-world behavior of the phenomenon. Models can be either *deterministic*, in which case they attempt to provide perfect predictions of a phenomenon, or *random* (or *stochastic*), in which case they attempt to model phenomena which have an intrinsic unpredictable variability. Phenomena of the latter type are referred to as random or stochastic phenomena.

Ockham's Razor.[1] This principle, also known as the *Principle of Parsimony*, says that given two equally good models one should choose the simpler one. Think of the "razor"'s job as one of shaving off unnecessary complications in a theory.

[1] Named after William of Ockham, a medieval English scholastic philosopher of the fourteenth century. He said that "entities are not to be multiplied beyond necessity," meaning that philosophical explanations should be kept as simple as possible. A good introduction to Ockham and his philosophical contributions can be found in *A History of Philosophy, Volume III: Late Medieval and Renaissance Philosphy; Ockham, Francis Bacon, and the beginning of the Modern World,* Frederick Copleston, Image Books/Doubleday, 1953/1993. The Principle of Parsimony has also been articulated by many others, from Aristotle to Einstein.

Theory of Probability. A mathematical theory which enables us to make predictions about the likelihood and frequency of occurrence of *outcomes* of a *random event.* Note that this theory requires clear definitions of the terms "outcome" and "random event."

Random Trial or Experiment. An experimental measurement of some random phenomenon of interest whose outcome cannot be predicted exactly in advance. It is usually assumed that the trial/experiment is repeatable under similar conditions, in which case lack of predictability is reflected in variations of the measurements from experiment-to-experiment.

Experimental Outcome or Sample Point, ω. The measurement value of a single trial/experiment. This value is usually denoted by the Greek letter ω, η, or ζ.

Sample Space, Ω or S. Also known as the *Universal Set* of possible experimental outcomes. The Sample Space Ω (also often denoted as S). is the set of *all* possible sample points (experimental outcomes) ω of an experiment/trial of interest. Any outcome of the experiment *must* be one of the points of Ω and *one and only one* of the outcomes in Ω *must* occur.

For example, if a toss of a 6-sided die is our experiment/trial and we assume the sample space to be comprised of outcomes corresponding to each of the possible face-up positions,

$$\Omega_1 = \left\{ \boxed{1}, \boxed{2}, \boxed{3}, \boxed{4}, \boxed{5}, \boxed{6} \right\},$$

then we do *not* admit the possibility of the die landing on an edge or corner. If these are to be admitted as experimental possibilities, then the sample space *must* be modified to reflect this fact,

$$\Omega_2 = \left\{ \boxed{1}, \boxed{2}, \boxed{3}, \boxed{4}, \boxed{5}, \boxed{6}, \boxed{\text{E}}, \boxed{\text{C}} \right\}.$$

Thus in order to construct the sample space, Ω, it is assumed in advance of performing an experiment that we know *all* possible outcomes. (Of course, because the experiment is random, we don't know *which* of the known possible outcomes in Ω will arise when the experiment is actually performed.) Note that the sample space is not uniquely defined as it can be larger than strictly needed. For example, both of the sample spaces Ω_1 and Ω_2 will suffice for an experiment for which one, and only one, of the six sides of a die will be observed.[2]

The set Ω is a *finite sample space* when it contains a finite number of outcomes. If Ω is not a finite space, then Ω is said to be an *infinite sample space.* When Ω contains a finite or denumerable (countably infinite) number of outcomes, we say that Ω is a *discrete sample space.* If Ω is not discrete, we say that it is a *continuous sample space.*

[2]Of course, if we know this to be the case, Ockham's razor would lead us to prefer a model base on the use of Ω_1.

Realization. Given an experiment with a sample space (universal set of outcomes) Ω, according to our theory *a single outcome* in Ω *must* occur when an experiment is actually performed. Let the experiment be performed; the *actual outcome,* $\omega \in \Omega$, which is observed to occur is called a *realization* of the experiment.

Elementary Random Event. When an experiment results in a realization ω we say that the *elementary event* $\{\omega\}$ has occurred. Am elementary event, then, is a singleton set $\{\omega\} \subset \Omega$, where ω is an experimental outcome (sample point) in the Sample Space, $\omega \in \Omega$. A common abuse of notation is to refer to the *sample point* ω as an 'elementary event," when, strictly speaking, it is $\{\omega\}$ which is the object being referred to.

General Random Event, $A \subset \Omega$. A *Random Event, A*, is associated with a random experiment with Sample Space Ω. An event A corresponds to a proposition,[3] $\alpha(\omega)$, about an experimental outcome, $\omega \in \Omega$, that has a 'true' or 'false' answer depending on the outcome. It is assumed that the truth value of the proposition $\alpha(\omega)$ can be ascertained for *every* experimental outcome in Ω. The event A occurs precisely when the proposition is true, so that A is a *subset* of the sample space Ω,

$$A = \{\omega \,|\, \alpha(\omega)\} = \{\omega \,|\, \text{the proposition } \alpha(\omega) \text{ is true}\} \subset \Omega\,.$$

Note that by definition $\omega \in A$ if and only if $\alpha(\omega)$ is true.[4]

For example, in the case of the 6-sided die discussed above, consider the event A_{even} defined by the logical proposition,

$$\alpha(\omega) = \text{"the number of face-up dots for the outcome } \omega \text{ is even"}.$$

In this case the event is obviously a subset of Ω given by the following set of outcomes,

$$A_{\text{even}} = \left\{ \boxed{2}, \boxed{4}, \boxed{6} \right\} \subset \Omega\,.$$

Note that for a subset A to be an event, there must be an equivalent logical proposition which (i) can be applied to *every* outcome in A; and (ii) is *true* for every outcome in A. As a consequence the sample space, Ω, and the empty set, $\emptyset$, are events,

$$\Omega = \{\omega \,|\, \alpha(\omega) = \text{"}\omega \text{ is an admissible experimental outcome"}\}\,,$$

$$\emptyset = \{\omega \,|\, \neg\,\alpha(\omega) = \text{"}\omega \text{ is not an admissible experimental outcome"}\}\,.$$

Note that the truth value of these propositions can be ascertained for every outcome. For every outcome $\alpha(\omega)$ evaluates as true (and hence Ω is the set of all outcomes) and $\neg\,\alpha(\omega)$ ('*not* $\alpha(\omega)$') evaluates as false (and hence $\emptyset$ has no elements). We can think of the event Ω as the event "something happens"[5] and $\emptyset$ as the event "nothing happens."[6]

[3] I.e., a logical fact or statement.

[4] An Elementary Event $\{\zeta\}$, as defined previously, is the random event associated with the elementary proposition, $\alpha(\omega) =$ "the experimental outcome ω is equal to the sample point ζ."

[5] That is, Ω is the event "one of the possible experimental outcomes occurs," which is true.

[6] That is, $\emptyset$ is the event "none of the possible experimental outcomes occurs," which is false.

Boolean Algebra, σ-Algebra of Events. For two logical propositions about an experimental outcome ω, $\alpha(\omega)$ and $\beta(\omega)$, let '$\alpha(\omega) \vee \beta(\omega)$' correspond to logical 'or' (logical disjunction) and let '$\alpha(\omega) \wedge \beta(\omega)$' correspond to logical 'and' (logical conjunction). Let the proposition 'not $\alpha(\omega)$' (logical negation) be denoted by '$\neg\, \alpha(\omega)$.' The logical propositions obey the rules of boolean logic and form a mathematical structure known as a boolean algebra.[7] Note that if $\alpha(\omega)$ and $\beta(\omega)$ are logical statements (propositions) which can be applied to every outcome in Ω, then so are the propositions $\alpha(\omega) \vee \beta(\omega)$, $\alpha(\omega) \wedge \beta(\omega)$, and $\neg\, \alpha(\omega)$. Thus if there are events which correspond to $\alpha(\omega)$ and $\beta(\omega)$, there must also be events corresponding to the conjunction, disjunction, and negation of these events.

Let $\mathcal{A}$ be a nonempty class of subsets of a Sample Space, Ω, and assume that *every* element of $\mathcal{A}$ is an event (and hence must each correspond to a logical proposition which can be applied to all outcomes in Ω). Now let A and B be events in $\mathcal{A}$ corresponding to the logical propositions $\alpha(\omega)$ and $\beta(\omega)$ resepectively. From our discussion above, it must be the case that the sets,

$$A \cup B = \{\omega \,|\, \alpha(\omega)\} \cup \{\omega \,|\, \beta(\omega)\} = \{\omega \,|\, \alpha(\omega) \vee \beta(\omega)\} \ ,$$

$$A \cap B = \{\omega \,|\, \alpha(\omega)\} \cap \{\omega \,|\, \beta(\omega)\} = \{\omega \,|\, \alpha(\omega) \wedge \beta(\omega)\} \ ,$$

and[8]

$$A' = \{\omega \,|\, \alpha(\omega)\}' = \{\omega \,|\, \neg\, \alpha(\omega)\} \ ,$$

are also events (subsets of Ω). However, in general these sets might not belong to $\mathcal{A}$. *Henceforth, we make the assumption that, in fact, they do belong to $\mathcal{A}$ and that this is true for all sets A and B in $\mathcal{A}$.*

This assumption corresponds to requiring that $\mathcal{A}$ be closed under a finite number of set operations (intersection, union, and complementation), and makes $\mathcal{A}$ a boolean algebra, just like the set of underlying logical propositions. In fact, one can view the two boolean algebras (one, an algebra of propositions, the other, an algebra of subsets) as essentially *equivalent* because of the one-to-one relationship between an event-set and an event-proposition.[9] To reiterate, closure ensures that if A and B are events in a boolean algebra, $\mathcal{A}$, of event subsets of Ω, then so are $A \cup B$, $A \cap B$, and A'. Also, if A belongs to $\mathcal{A}$, then so must $A \cup A' = \Omega$ and $A \cap A' = \emptyset$, showing that every nonempty boolean algebra of subsets of Ω must contain both Ω and the empty set, $\emptyset$.

Because of the equivalence between events, A, and logical propositions, $\alpha(\cdot)$, usually no care is taken to distinguish them and we treat A as the logical proposition $\alpha(\cdot)$ itself—we can

[7] Which you should know either from courses in digital logic or CSE20.

[8] For A a subset of Ω, A' denotes the set complement of A in Ω, $A' = \Omega \setminus A$.

[9] This correspondence is known as *Stone's Theorem* and is the reason Venn diagrams work—areas in the plane are point sets which we interpret as events corresponding to logical propositions. To emphasize the *set*-theoretic properties of closure under finite unions and intersections one can speak of a *field* of sets. However, we prefer to emphasize the (equivalent) *boolean algebra* interpetation as is done in many presentations. (See, for example *Artificial Intelligence: A Modern Approach*, 2nd Edition, S. Russell and P. Norvig, Prentice-Hall, 2002, or *Introduction to Probability Theory,* K. Ito, Cambridge University Press, 1978).

think of the event A as an "event-proposition." Note, that if A and B are event-propositions in a boolean algebra $\mathcal{A}$ then $A \cup B$ is the event-proposition "A or B", $A \cap B$ is the event-proposition "A and B", and A' is the event-proposition "not A', and these event-propositions must also be in the boolean algebra $\mathcal{A}$ because of closure under countable set operations. In summary:

> Boolean Algebra of Events. A *boolean algebra of events*, $\mathcal{A}$, is (i) a class of subsets of the sample space, Ω, which (ii) is closed under a *finite* number of set operations. Assuming that it is nonempty, it must contain the sample space, Ω, and the empty set, $\emptyset$.

More generalize, we need $\mathcal{A}$ to be closed under a *countable* number of set operations. It can be shown that this corresponds to requiring that the conditions for a boolean algebra be strengthened by requiring closure under countable set unions, $A \cup B \cup C \cup \cdots$. Because an alternative notation for the required set union property is closure of $A + B + C + \cdots$, a boolean algebra $\mathcal{A}$ which has this additional property is known as a *σ-algebra* (as the Greek letter σ corresponds to the roman letter s, which in turn stands for *summable*).[10]

> σ-Algebra of Events. A *σ-algebra of events,* $\mathcal{A}$, is (i) a class of subsets of the sample space, Ω, which (ii) is closed under a *countable* number of set operations. Assuming that it is nonempty, it must contain the sample space, Ω, and the empty set, $\emptyset$.

2. Probability Space

Kolmogorov Probability Axioms.[11] In the *axiomatic approach to probability theory* (see the discussion given below), we *assume* the existence of a probability function and justify its correctness *after the fact* via systematic testing and validation. A probability function (also known as a probability measure) is a set-function, $P(A)$, which maps events, $A \subset \Omega$, to the nonnegative real numbers. It is assume that the probability measure satisfies the following three axioms.[12]

[10] In our course we will take $A + B$ to be synonymous with the standard set-union operation so that $A \cup B$, $A + B = A \cup B$ for all A and B. However the "+" notation can be confusing. This is particularly true because some people make the addition assumption that $A + B$ means that A and B are disjoint. However, we *don't* make any such assumption and $A + B$ for us is just the standard set-union operation.

[11] Andrei N. Kolmogorov (1903–1987), was a brilliant Russian mathematician and physicist who made important, original contributions to probability theory, random processes, information theory, complexity theory, mechanics, fluid dynamics, and nonlinear dynamical systems theory. He proposed the axiomatic approach to probability in 1933, when he has 30 years old. He is a giant of the 20th century.

[12] Instead of applying the axioms to the event sets, A, (as is the usual case in engineering and mathematical analysis) one instead can equivalently apply them to the underlying logical propositions, $\alpha(\omega)$. This is done, for instance in the CSE150 textbook, *Artificial Intelligence: A Modern Approach*, 2nd Edition, S. Russell and P. Norvig, Prentice-Hall, 2002.

Kolmogorov Probability Axioms

1. For every event $A \subset \Omega$, $P(A) \geq 0$.
2. $P(\Omega) = 1$.
3. If the countable sequence of events A_ℓ are mutually exclusive, $A_i \cap A_j = \emptyset$ for $i \neq j$, then

$$P(\bigcup_\ell A_\ell) = \sum_\ell P(A_\ell)\,.$$

Property (1) is the property of *nonnegativity* of $P(\cdot)$; Property (2), *normalization* of $P(\cdot)$; and Property (3), *countable additivity* of $P(\cdot)$. That $P(\cdot)$ obeys $P(\emptyset) = 0$ and is finitely additivity is entailed by Properties (2) and (3).

(Kolmogorov) Probability Space. A Probability Space is a mathematical model used to describe the behavior of the measured outcomes of a real-world phenomenon of interest. Because it is based on the framework first proposed by Kolomogorov in the 1930's, it would perhaps be most informative to call it a *Kolmogorov Probability Model*. However, it is commonly known simply as a *Probability Space.*[13]

(Kolmogorov) Probability Space

A Probability Space is a triple $(\Omega, \mathcal{A}, P)$ where

1. Ω is a sample space of outcomes.
2. $\mathcal{A}$ is a nonempty σ-algebra of Ω-events (subsets of Ω).
3. $P(\cdot)$ is a probability measure on $\mathcal{A}$ which satisfies the Kolmogorov probability axioms.

3. Determination of Probabilities.

A nontrivial question, of course, is how to determine the actual numerical values, $P(A)$, of the probability measure as a function of the events $A \in \mathcal{A}$. Before the advent of the axiomatic approach, one would try to *derive* a probability model from *a priori* arguments (e.g., "outcomes are equally likely" or "probabilities are relative frequencies"), but this could never be put on a rigorous footing which was universally applicable.[14] In the axiomatic framework, one instead *postulates* that there exists a well-defined probability function, even

[13] It is important, however, to recognize that it is just a model and that, in fact, there are domains where it is inadequate, or breaks down, and must be augmented or replaced. This is the case, for example, in quantum mechanics. See, e.g., *The Structure and Interpretation of Quantum Mechanics,* R.I.G. Hughes, Harvard University Press, 1989.

[14] For instance, the assumption of equally likely outcomes could never handle the problem of a weighted coin or loaded die.

if we don't know exactly what it is, and constructs a mathematical model based on that assumption. We then make principled choices of the probability values to be assigned to specific events. To do so, we use intuition; experience; mathematical and physical reasoning; symmetry arguments; and engineering custom. However the assignments are made, the model must then be justified after the fact via testing and experimentation. Often, a few iterations are required before an acceptable model is determined.[15]

Classical Probability Theory. A probability space is finite if its sample space is finite. Classical probability theory assumes a finite probability space and equiprobable outcomes, $P(\omega) = \text{constant}, \forall \omega \in \Omega$. Using the probability axioms, it is easy to show that the probability of a single outcome is $\frac{1}{N}$, where $N = N(\Omega) = \#(\Omega) =$ cardinality of Ω. An outcome, ω, is said to be *favorable* to A if $\omega \in A$. The number of outcomes favorable to the event A is $N(A) = \#(A) =$ cardinality of A. Using the probability axioms it is readily shown that the probability of the event A is

$$P(A) = \frac{N(A)}{N(\Omega)} = \frac{\#(A)}{\#(\Omega)} .$$

Thus, we see that the name of the game in classical probability theory is *counting*. One needs to count the number of possible outcomes to determine $N(\Omega) = \#(\Omega)$ (eg, how many total five-card deals are possible) and $N(A) = \#(A)$ (eg, how may ways can a royal flush be dealt) before one can determine the probability of A (eg, $A = \{\text{royal flush}\}$). Not surprisingly, then, *combinatorics* (the theory of counting) is an important topic in classical probability and much time is spent developing proficiency in computing permutations and combinations.

> Permutation, Combination. A *permutation* is an *ordered arrangement* of distinct objects. Synonymous terms are *ordered sample* and *linear arrangement.* Note that the order of the distinct objects *matters* in this definition. The number of distinct permutations (each one comprised of an ordered arrangement of distinct objects) that one can form by selecting any possible subgroup of k distinct objects from a larger group of n distinct objects is
>
> $$P_k^n = (n)_k = \frac{n!}{(n-k)!} .$$
>
> A *combination* is a *collection* of distinct objects *without regard to order.* Synonymous terms are *group, set, unordered sample, population,* and *subpopulation.* Note that the order of the distinct objects *does not matter* in this definition. The

[15] This is particularly the case in communications theory. Common questions about the random behavior of a wireless channel are "is it Gaussian?"; "is it Rayleigh?", "is it Rician?", " is it multipath?", etc, etc, etc.

number of distinct collections (or subpopulations) that one can form by selecting k distinct objects from a larger group (population) of n distinct objects is

$$C_k^n = \binom{n}{k} = \frac{n!}{k!\,(n-k)!} = \frac{(n)_k}{k!}\,.$$

Relative Frequency Approach to Probability.[16] Suppose that we have repeated trials so that an experiment whose sample space is Ω is repeatedly performed under exactly the same conditions. After n trials (repetitions of the experiment) have been performed, for any event, $A \subset \Omega$, we define $n(A)$ to be the number of times the event A occurred. The *Relative Frequency of the Event* A, $f_n(A)$, is defined to be the proportion of times that A occurred in the n trials,

$$f_n(A) \triangleq \frac{n(A)}{n}\,.$$

The 'old-fashioned' approach to probability is based on assuming that the limit of the measured (experimentally determined) relatively frequency of an event A exits, and can therefore be used to *define* the probability of the event as

$$P(A) \triangleq \lim_{n\to\infty} f_n(A)\,.$$

A theory of probability is then built-up around these probabilities. Note that this approach starts from observed relative frequencies and *only then* moves to the development of a mathematical model. This is opposite to the Axiomatic Approach, which starts with an abstract model and then proceeds to see if it can explain observed relative frequencies. The Relative Frequency Approach fell into disfavor because of various logical difficulties associated with it (e.g., how do we know that the relative frequencies will converge?) which do not bedevil the Axiomatic Approach.

Axiomatic Approach to Probability. In this approach, the one which we have been following, a self-consistent mathematical model of probability and events is constructed based on the assumption of a few fundamental axioms. (E.g., in our case we are working with the Kolmogorov probability space axiomatic model.) Only *after* the axiomatic model has been constructed, are mathematical consequences and predictions of the model then compared to the measured behavior of a real-world situation or system of interest. If there is a "reasonably good" match between the model's predicted behavior and the corresponding measured behavior of the real-world situation, then the axiomatic model is deemed to be an acceptable mathematical model of that situation. If the match is poor, we go back to the drawing board and attempt to revise our model. Thus, *axiomatic probability models are justified after-the-fact,* by how well they explain and predict measured real-world behavior. Self-consistent axiomatic models themselves are neither true or false, rather they are 'better or worse' in their degree of correspondence to a measured phenomenon of interest.

[16] Not to be confused with the 'Relative Frequency *Interpretation* of Probability' discussed subsequently.

Relative Frequency Interpretation of Probability. Although there is little, or no, controversy about the use of an axiomatic model of probability, there is controversy about its interpretation when used to explain real-world phenomena of interest. So-called *Frequentists* are perhaps the most conservative and accept only the *Relative Frequency Interpretation of Probability.* This interpretation says that probability models can only be used to model situations where an (potentially) unlimited number of repeated trials is possible. Frequentists do not admit any probabilistic interpretations of so-called "one-off" events (events which only occur one-time). The Frequency Interpretation of Probability assumes that repeated trials can be performed so that relative frequencies can be computed. In this case, a model is acceptable if it can be determined that whenever relative frequencies of an event A are empirically measured we have that,

$$P(A) \approx f_n(A) = \frac{n(A)}{n} \quad \text{for } n \text{ "large."}$$

In this case, we accept the model and go on to interpret the probability, $P(B)$, of any other event as the likely relative frequency of occurrence in n trials for n "large enough."[17] Equivalently, for n "large enough" we expect to find that,

$$n(B) \approx n \cdot P(B)\,.$$

For instance, suppose a patient has been told that base on the positive outcome of a medical test he has a 10% probability of contracting a certain genetic blood disorder after age 60. When asked what this means, he is told by his Frequentist doctor that, based on data amassed in clinical trials, it means that of 1000 men who test positive on this test, one can expect about 100 of them to contract the disorder after age 60. Note that this interpretation assumes that sufficient data exists to back up such an assertion. The Frequency Interpretation is sometimes called an *Objectivist Interpretation* as one tries to justify it using objective, measured data collected from repeated trials.

Subjectivist Interpretation of Probability. So-called *subjectivists* go beyond the objective Frequency Interpretation of Probability and are willing to interpret probabilities in one-off situations where there isn't, perhaps never has been and never will be, data sufficient to construct relative frequencies. In this case, a subjectivist interprets the probability of an event A as a measure of his or her *personal belief* that the event A will occur. Not surprisingly, the subjectivist interpretation is controversial, even though in practice it is used extensively.

For example, suppose in the medical situation described above there is little or no data available to make a frequency interpretation (e.g., perhaps only 5 people in the world have ever even had the disease throughout history!). And suppose the doctor *still* tells the patient that, in his opinion, he has a 10% probability of contracting the disease. When the patient

[17] This approach can be partially justified by appealing to theoretical results known as *Strong Laws of Large Numbers.*

asks what that means, the doctor says that based on his professional judgment built up from years of looking at related ailments, *it is his personal belief* that the patient will likely get the ailment is 10%. (Whatever that means!—which is why this interpretation is considered *subjective* rather than *objective*.)

4. Manipulation of Probabilities

Inclusion-Exclusion Formula. The two-event *inclusion-exclusion* formula is just

$$P(A+B) = P(A) + P(B) - P(AB)\,.$$

Note that we "include" single events and "exclude" double events on the right-hand side of this formula. The three-event inclusion-exclusion formula is

$$P(A+B+C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC)\,.$$

Note that here we "include" single events, "exclude" double events and "include" triple events.

By induction one can prove the general n-event inclusion-exclusion formula,

$$P(A_1+\cdots+A_n) = \sum_{1\le j\le n} P(A_j) - \sum_{1\le i<j\le n} P(A_iA_j) + \sum_{1\le i<j<k\le n} P(A_iA_jA_k) - \cdots + (-1)^{n+1}P(A_1\cdots A_n)\,.$$

Note that in this general formula we "include" odd-number events and "exclude" even-number events. Each summation shown in the right-hand side of the general n-event inclusion-exclusion formula is a sum over all distinct ways, without regard to order, that one can select $\ell \ge 1$ events from the set of n-events under consideration.[18] Therefore, the number of terms in each such summation is given by $\binom{n}{\ell}$.

Almost Sure Equality of Two Events. Recall that two events (sets), A and B, are defined to be equal if and only if they both contain exactly the same elements. A weaker form of equality is based on treating two events as *equal almost surely* (or with *Probability 1*) if outcomes in A or B which are not in the intersection of A and B have zero probability of occurring. We can present this formally as

Almost Sure Equality of Two Events

> Two events A and B are said to be equal *almost surely* (a.s.) or with *Probability 1* (P–1), designated symbolically by
>
> $$A = B \text{ a.s.}\,,$$
>
> if and only if
>
> $$P(A) = P(B) = P(AB)\,.$$

[18] Thus in the first term we are interested in the single ($\ell = 1$) event indexed by j, in the second term the two events ($\ell = 2$) indexed by i and j, etc.

The condition of almost sure equality can actually be said in a number of equivalent ways, as shown by the following theorem.[19]

Equivalent Conditions for Almost Sure Equality

$A = B$ a.s. if and only if any of the following conditions holds:

1) $P(A \triangle B) = 0$.[20]

2) $P(A + B) = P(AB)$.

3) $A' = B'$ a.s.

Conditional Probability. If A and B are both events in $(\Omega, \mathcal{A}, P)$ and $P(B) > 0$, we define the *Conditional Probability of the Event A given the Event B*, $P(A\,|B)$, by

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

The conditional probability satisfies the 3 Kolmogorov Axioms required of a probability measure on $(\Omega, \mathcal{A})$, so that $P(A\,|B)$ is a probability in its own right (and hence the triple $(\Omega, \mathcal{A}, P(A\,|B))$ is a probability space in its own right). Henceforth, when writing expressions like $P(G\,|H)$ it will always be tacitly assumed that $P(H) > 0$.

The conditional probability has the following intuitively nice properties:

a) $AB = \emptyset \Rightarrow P(A\,|B) = 0$; b) $A \subset B \Rightarrow P(A\,|B) \geq P(A)$; and c) $A \supset B \Rightarrow P(A\,|B) = 1$.

Product Rule for Conditional Probabilities. Note from the definition of conditional probabilities that,

$$P(A_2 A_1) = P(A_2\,|A_1)\; P(A_1)\,.$$

Similarly,

$$P(A_3 A_2 A_1) = P(A_3\,|A_2 A_1) P(A_2 A_1) = P(A_3\,|A_2 A_1) P(A_2\,|A_1)\; P(A_1)\,.$$

Proceeding inductively in this manner, we have that

$$P(A_n \cdots A_1) = P(A_n\,|A_{n-1} \cdots A_1) \cdots P(A_3\,|A_2 A_1) P(A_2\,|A_1)\; P(A_1)\,.$$

This general result is known as the *Product Rule for Conditional Probabilities.*

[19] This material has been drawn from *Concepts of Probability Theory,* 2nd Revised Edition, Paul Pfeiffer, Dover Publications, 1978.

[20] $A \triangle B = A \setminus B + B \setminus A$ is called the *symmetric difference* between A and B. It is also known as *disjoint union*, *disjunctive union*, or *exclusive-or*. This operation is also commonly referred to as *exclusive or* and denoted by $A \oplus B$. Thus, using the exclusive-or notation, we have that $A = B$ a.s. iff and only if $P(A \oplus B) = 0$.

Partition of an Event, Almost Sure Partition of a Set. Let B_i, $i = 1, 2, \cdots$, be a countable collection of mutually disjoint sets. This collection is defined to be a *partition* of A if and only if $A \subset B_1 + B_2 + \cdots$. Note that this is equivalent to demanding that $A(B_1 + B_2 + \cdots)' = \emptyset$. If the *weaker* condition that $P\left(A(B_1 + B_2 + \cdots)'\right) = 0$ holds, then the collection is said to be *an almost sure partition* of A.[21] *Note that a partition must be an almost sure partition*, so that any condition which holds for an almost sure partition holds for a partition. Also note that in particular we can take $A = \Omega$, in which case we have respectively a partition or an almost sure partition of the sample space Ω.

Theorem of Total Probability (TTP). Given a *disjoint,* countable sequence, A_i, $i = 1, 2, \cdots$, under suitable conditions (to be discussed below) the *Theorem of Total Probability* (TTP) of an event B can be invoked to yield,

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots .$$

At the level of presentation given in most textbooks at the undergraduate level, it is assumed that $P(A_i) > 0$, for all $i = 1, 2, \cdots$. If this is the case, the TTP holds for a *particular* event B if the disjoint collection, A_i, is an almost sure partition for B, and the TTP holds for *every* event B if the disjoint collection is an almost sure partition of the sample space Ω. Note that the result must therefore hold for the stronger condition that the collection A_i is a partition.

Bayes' Rule. Note that using the Product Rule for Conditional Probabilities, we can expand $P(AB)$ either by conditioning on A or by conditioning on B,

$$P(A\,|B)\;P(B) = P(AB) = P(B\,|A)\;P(A)\,.$$

This symmetric expansion immediately yields *Bayes' Rule,*

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

$P(B)$ is called the *a priori*, or prior, probability of the B, while $P(B\,|A)$ is called the *a posteriori*, or posterior, probability of B given the measured event A. $P(A\,|B)$ is called the *likelihood* of B given the measured event A, and provides "evidence" that B is the case given the occurrence of an event A. Bayes' rule, then, provides an evidentiary procedure for updating one's prior belief that B is the case (as measured by the prior probability $P(B)$) using "evidence" obtained from measuring a related phenomenon A. One's updated belief that B is the case is measured by the posterior probability $P(B\,|A)$.

If the disjoint collection, B_j, $j = 1, 2, \cdots$, is an a.s. partition of A, then we can invoke the TTP to write Bayes' rule as,

$$P(B_i\,|A) = \frac{P(A\,|B_i)\;P(B_i)}{\sum_j P(A\,|B_j)\;P(B_j)} \quad \text{for } i = 1, 2, \cdots .$$

[21] Or almost surely a partition, or a partition almost surely, etc..

Independence of Events. The event A is said to be independent of the event B if and only if,

$$P(A\,|B) = P(A)\,.$$

If A is independent of B, it is easy to show that,

$$P(B\,|A) = P(B)\,,$$

showing that then B must be independent of A. Indeed, it is also easy to show that independence of events A and B is equivalent to the symmetric condition,

$$P(AB) = P(A)P(B)\,,$$

which clearly indicates that independence is a property which holds mutually.

It is easy to show that Ω and $\emptyset$ are independent of any event A,

$$P(A\Omega) = P(A) = P(A)\cdot 1 = P(A)P(\Omega)\,,$$

$$P(A\emptyset) = P(\emptyset) = 0 = P(A)\cdot 0 = P(A)P(\emptyset)\,.$$

We define three events A, B, and C to be an *independent collection* if they we satisfy the *four* conditions,

$$P(AB) = P(A)P(B)\,,$$
$$P(AC) = P(A)P(C)\,,$$
$$P(BC) = P(B)P(C)\,,$$

and

$$P(ABC) = P(A)P(B)P(C)\,.$$

The first three conditions holding is called *pair-wise independence*. The last condition is required in addition to the first three to ensure, for example, independence of the event A and the event BC,

$$P(ABC) = P(A)P(B)P(C) = P(A)P(BC)\,.$$

More generally, an collection of events A_j, $j = 1, 2, \cdots$, is an independent collection if and only if for any sub-collection, A_{α_j}, $j = 1, \cdots, r$, $r \geq 2$, we have that

$$P(A_{\alpha_1}\cdots A_{\alpha_r}) = P(A_{\alpha_1})\cdots P(A_{\alpha_r})\,.$$

An equivalent statement is we that have an independent collection of events if and only if each subcollection is itself an independent collection.

The following theorem is extremely useful for computing probabilities of compound events.[22]

[22] A proof is given in Pfeiffer, op. cit., page 62.

Subcollection Independence Property

> Given a collection of events A_j, $j = 1, 2, \cdots$, let $\overline{A}_{\alpha_i}$ denote one of the admissible event-choice possibilities $\overline{A}_{\alpha_i} = \emptyset$, Ω, A_{α_i}, or A'_{α_i}, for $i = 1, \cdots, r$. Then A_j, $j = 1, 2, \cdots$, is an independent collection if and only $\overline{A}_{\alpha_i}$, $i = 1, \cdots r$, is an independent collection for any admissible event-choice for each $\overline{A}_{\alpha_i}$ and for every $r \geq 2$.

As a simple example of the utility of this theorem, note that independence of the collection A, B, C, is equivalent to independence of the collection A', B', C', so that

$$\begin{aligned} P(A+B+C) &= 1 - P\left((A+B+C)'\right) \\ &= 1 - P\left(A'B'C'\right) \\ &= 1 - P(A')P(B')P(C') \\ &= 1 - (1-P(A))\,(1-P(B))\,(1-P(C)) \ , \end{aligned}$$

allowing the probability of the compound event $A + B + C$ to be determined from single event probabilities.

Let C_1 and C_2 both be subcollections of an independent collection assembled in the manner described in the Subcollection Independence Property stated immediately above. Assume that they share no events in common. Let $F(\mathsf{C}_1)$ and $G(\mathsf{C}_2)$ each denote an event determined by a finite number of set operations (set unions, intersections, and complements) on the elements of the subcollections in their arguments. Then it is a most useful fact that $F(\mathsf{C}_1)$ and $G(\mathsf{C}_2)$ are independent,[23]

$$P\left(F\left(\mathsf{C}_1\right)\, G\left(\mathsf{C}_2\right)\right) = P\left(F(\mathsf{C}_1)\right) \cdot P\left(G(\mathsf{C}_2)\right) \ .$$

For example, if A, B, C, D, form an independent collection then,

$$P((A+D')\,B \triangle C) = P(A+D') \cdot P(B \triangle C) \, .$$

[23] Pfeiffer, op. cit., page 83.