Ma 3 KC Border
Introduction to Probability and Statistics Winter 2015

Chapter 2: Properties of Probability; Conditional Probability; Independence

Relevant textbook passages:

Pitman [4]: Sections 1.3–1.4., pp. 26–46.

Larsen–Marx [3]: Sections 2.2–2.5, pp. 18–66.

John von Neumann once said, "There's no sense in being precise when you don't even know what you're talking about." Yet even though I am not sure about what the correct interpretation of probability is, I am going to give a precise mathematical definition of it. I can at least understand the mathematical definition. Or maybe not, for as von Neumann also said, "In mathematics you don't understand things. You just get used to them."

2.1 Events

Pitman [4]: § 1.3

One of the key concepts in our formal model is an **event**. An event is simply an "observable" subset of the sample space. If the experiment produces an outcome $s \in S$ and s belongs to the event E, then we say that the event E **occurs** (or has occurred).

The **set of events** is denoted $\mathcal{E}$, (or sometimes, in keeping with a Greek theme, by Σ). Often, especially when the sample space is finite or denumerably infinite, $\mathcal{E}$ will consist of *all* subsets of S. [As you go on to study more mathematics, you will learn that there are problems with a nondenumerable sample space that force you to work with a smaller set of events.]

We require at a minimum that the set of events be an **algebra** or **field** of sets. That is, $\mathcal{E}$ satisfies:

1. $\varnothing \in \mathcal{E}$, $S \in \mathcal{E}$.
2. If $E \in \mathcal{E}$, then $E^{\mathrm{c}} \in \mathcal{E}$.
3. If E and F belong to $\mathcal{E}$, then EF and $E \cup F$ belong to $\mathcal{E}$.

Most probabilists assume further that $\mathcal{E}$ is a ***σ*-algebra** or ***σ*-field**, which requires in addition that

3′. If $E_1, E_2, \ldots$ belong to $\mathcal{E}$, then $\bigcap_{i=1}^{\infty} E_i$ and $\bigcup_{i=1}^{\infty} E_i$ belong to $\mathcal{E}$.

Note that if S is finite and $\mathcal{E}$ is an algebra then, it is automatically a σ-algebra. Why?

The reason for these properties is that we think of events as having a description in some language. Then we can think of the descriptions being joined by *or* or *and* or *not*. The correspond to union, intersection, and complementation.

2.2 Probability measures

Pitman [4]: § 1.3
Larsen–Marx [3]: § 2.3

A **probability measure** or **probability distribution** (as in Pitman [4]) or simply a **probability** (although this usage can be confusing) is a **set function** $P\colon \mathcal{E} \to [0,1]$ that satisfies:

Normalization $P(\varnothing) = 0$; and $P(S) = 1$.

Additivity If $E \cap F = \varnothing$, then $P(E \cup F) = P(E) + P(F)$.

Most probabilists require the following stronger property, called **countable additivity**:

Countable additivity $P\Big(\bigcup_{i=1}^{\infty} E_i\Big) = \sum_{i=1}^{\infty} P(E_i)$ provided $E_i \cap E_j = \varnothing$ for $i \neq j$.

Aside: You need to take an advanced analysis course to understand that for infinite sample spaces, there can be probability measures that are additive, but not countably additive. So don't worry too much about it.

Note that while the domain of P is technically $\mathcal{E}$, the set of events, we may also refer to P as a probability (measure) on S, the set of samples.

2.2.1 Remark To cut down on the number of delimiters in our notation, we may omit the some of them simply write something like $P\left(f(s) = 1\right)$ or $P\{s \in S : f(s) = 1\}$ instead of $P\Big(\{s \in S : f(s) = 1\}\Big)$ and we may write $P(s)$ instead of $P\Big(\{s\}\Big)$. You will come to appreciate this.

2.2.1 Odds

I find it exasperating that even generally linguistically reliable sources,such as *The New York Times*, confuse probabilities and odds.

Pitman [4]: pp. 6–8

2.2.2 Definition *The* ***odds against the event E*** *is the ratio*

$$\frac{P(E^{\mathrm{c}})}{P(E)}.$$

That is, it is a ratio of probabilities, not a probability. It is usually spoken as "the odds against the event E are $P(E^{\mathrm{c}})/P(E)$ to one," or as a ratio of integers. (That is, we typically say "3 to 2" instead of "1 1/2 to 1.")

*The **odds in favor of the event E** is*

$$\frac{P(E)}{P(E^{\mathrm{c}})}.$$

This is to be distinguished from the **payoff odds**. The payoff odds are the ratio of the amount won to the amount wagered for a simple **bet**. For instance, in roulette, if you bet \$1 on the number 2 and 2 comes up, you get a payoff of \$35, so the payoff odds are "35 to 1." But (assuming that all numbers on a roulette wheel are equally likely) the odds against 2 are 37 to one since a roulette wheel has the "numbers" 0 and 00 in addition to the numbers 1 through 36.[1] [2] [3] (Pitman [4, p. 7] describes the outcomes and bets for a roulette wheel.)

Unfortunately, you often run across statements such as, "the odds are one in ten that X will happen," when the author probably means, "the probability that X will happen is one-tenth," so that the odds in favor of X happening are one to nine.

2.3 Probability spaces

Our complete formal model of **random experiment** is what we call a probability space.

2.3.1 Definition *A **probability space** is a triple $(S, \mathcal{E}, P)$, where S is a nonempty set, the **sample space** or **outcome space** of the experiment, $\mathcal{E}$ is the set of **events**, which is a σ-field of subsets of S, and P is a countably additive probability measure on $\mathcal{E}$.*

2.3.1 An example: Uniform probability

2.3.2 Theorem (Uniform probability) *Consider the case where S is finite and $\mathcal{E}$ contains all subsets of S. Enumerate S as $S = \{s_1, \dots, s_n\}$, where $n = |S|$. Then $1 = P(S) = P(s_1) + \cdots + P(s_n)$. (Why?) If each outcome is equally likely (has the same probability), then $P(s_1) = \cdots = P(s_n) = 1/|S|$, and*

$$P(E) = \frac{|E|}{|S|}.$$

2.3.3 Example (Coin Tossing) We usually think of a coin as being equally likely to come up H as T. That is, $P\{H\} = P\{T\}$. If our sample space is the simple $S = \{H, T\}$ and $\mathcal{E}$ is all four subsets of S, $\mathcal{E} = \Big\{\varnothing, S, \{H\}, \{T\}\Big\}$, then

$$\{H\}\{T\} = \varnothing \text{ and } \{H\} \cup \{T\} = S$$

[1] Actually, there are (at least) two kinds of roulette wheels. In Las Vegas, roulette wheels have 0 and 00, but in Monte Carlo, the 00 is missing.

[2] The term roulette wheel is a pleonasm, since *roulette* is French for "little wheel."

[3] The word "pleonasm" is one of my favorites. Look it up.

so

$$1 = P(S) = P\Big(\{H\} \cup \{T\}\Big) = P\{H\} + P\{T\},$$

which implies

$$P\{H\} = P\{T\} = 1/2.$$

□

2.4 Random variables

For some reason, your textbooks put off the definition of random variables, even though they are a fundamental concept.

2.4.1 Definition *A **random variable** on a probability space $(S, \mathcal{E}, P)$ is a real-valued function on S which has the property that for every interval $I \subset \boldsymbol{R}$ the inverse image of I is an event.*

Note that when the collection $\mathcal{E}$ of events consists of all subsets of S, then the requirement that inverse images of intervals be events is automatically satisfied.

2.4.2 Remark An interpretation of random variables used by engineers is that they represent *measurements* on the state of a system. See, e.g., Robert Gray [2].

Traditionally, probabilists and statisticians use upper-case Latin letters near the end of the alphabet to denote random variables. This has confused generations of students, who have trouble thinking of random variables as functions. For the sake of tradition, and so that you get used to it, we follow suit. So a **random variable** X is a function

$$X\colon S \to \boldsymbol{R} \quad \text{such that for each interval } I, \quad \{s \in S : X(s) \in I\} \in \mathcal{E}.$$

We shall adopt the following notational convention, which I refer to as **statistician's notation**, that

$$(X \in I) \text{ means } \{s \in S : X(s) \in I\}.$$

Likewise $(X \leqslant t)$ means $\{s \in S : X(s) \leqslant t\}$, etc.

If E belongs to $\mathcal{E}$, then its **indicator function $\mathbf{1}_E$**, defined by

$$\mathbf{1}_E(s) = \begin{cases} 0 & s \notin E \\ 1 & s \in E, \end{cases}$$

is a random variable.

2.4.3 Definition *A property $Q(s)$ parametrized by states of the world is said to hold* ***almost surely****, abbreviated* ***Qa.s.****, if the set of states of the world for which it fails to hold is a subset of an event of probability zero. That is,*

$$Qa.s. \iff (\exists E \in \mathcal{E})\ \Big[P(E) = 0 \ \&\ \{s \in S : \neg Q(s)\} \subset E\Big].$$

2.5 Elementary Probability Identities

Larsen–Marx [3]: § 2.3

1.
$$P(A^{\mathrm{c}}) = 1 - P(A)$$

2. If $B \subset A$, then
$$P(A \setminus B) = P(A) - P(B)$$

3. If $A_1, \dots, A_n$ are **pairwise disjoint**, i.e., $i \neq j \implies A_i A_j = \varnothing$, then
$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i).$$

Proof of last: Let $\mathbb{P}(n)$ stand for the proposition for n. Then $\mathbb{P}(2)$ is just Additivity. Assume $\mathbb{P}(n-1)$. Write
$$\bigcup_{i=1}^{n} A_i = \underbrace{\bigcup_{i=1}^{n-1} A_i}_{=B} \cup A_n$$
Then $BA_n = \varnothing$, so
$$\begin{aligned} P\left(\bigcup_{i=1}^{n} A_i\right) &= P\left(B \cup A_n\right) \\ &= P(B) + P(A_n) \qquad \text{by Additivity} \\ &= \sum_{i=1}^{n-1} P(A_i) + P(A_n) \quad \text{by } \mathbb{P}(n-1) \\ &= \sum_{i=1}^{n} P(A_i). \end{aligned}$$

2.5.1 Boole's Inequality *Even if events $A_1, \dots, A_n$ are not pairwise disjoint,*
$$P\left(\bigcup_{i=1}^{n} A_i\right) \leqslant \sum_{i=1}^{n} P(A_i).$$

The next results may seem theoretical and of no practical relevance, but they are crucial to understanding the properties of cumulative distribution functions.

A sequence $E_1, \dots, E_n \dots$ of events is decreasing, written $E_n \downarrow$, if
$$E_1 \supset E_2 \supset \cdots \supset E_n \supset \cdots .$$

A sequence $E_1, \dots, E_n \dots$ of events is increasing, written $E_n \uparrow$, if

$$E_1 \subset E_2 \subset \cdots \subset E_n \subset \cdots .$$

2.5.2 Proposition (Continuity and countable additivity) *If P is an additive probability, then*

1. *P is countably additive if and only if $E_n \downarrow$ implies $P\left(\bigcap_n E_n\right) = \lim_n P(E_n)$.*

2. *P is countably additive if and only if $E_n \uparrow$ implies $P\left(\bigcup_n E_n\right) = \lim_n P(E_n)$.*

2.6 Conditional Probability

Suppose we acquire new information on the outcome of a random experiment that takes the form, "The sample outcome lies in a set A," or "The event A has occurred." Then we should **update** or **revise** our probabilities to take into account this new information.

Pitman [4]: § 1.4

Larsen–Marx [3]: § 2.4

For example, suppose an experiment can three have equally likely outcomes, a, b, c, and we find out that the result lies in the event $A = \{a, b\}$. What should we revise our probabilities to be based on this new information? I claim that we should assign probability 0 to c, and probability 1/2 each to a and b. In other words:

2.6.1 Definition *If $P(A) > 0$, the **conditional probability of B given A**, written $P(B|A)$, is defined by*

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

(This only makes sense if $P(A) > 0$.)

Note that $P(A|A) = 1$.

At this point this may seem a bit abstract, but we shall come back to this in Lecture 4 and make more sense of it.

2.7 Independence

Larsen–Marx [3]: § 2.5
Pitman [4]: § 1.4

2.7.1 Definition *Events A and B are* ***(stochastically) independent*** *if for $P(B) \neq 0$,*

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

or equivalently,

$$P(AB) = P(A) \cdot P(B).$$

That is, knowing that B has occurred has no impact on the probability of A. The second formulation works even if A or B has probability 0.

I need to relate this to product sets and product measures.

2.7.2 Lemma *If A and B are independent, then A and B^{c} are independent; and A^{c} and B^{c} are independent; and A^{c} and B are independent.*

Proof: It suffices to prove that if A and B are independent, then A^{c} and B are independent. The other conclusions follow by symmetry. So write

$$B = (AB) \cup (A^{\mathrm{c}}B),$$

so by additivity

$$P(B) = P(AB) + P(A^{\mathrm{c}}B) = P(A)P(B) + P(A^{\mathrm{c}}B),$$

where the second inequality follows from the independence of A and B. Now solve for $P(A^{\mathrm{c}}B)$ to get

$$P(A^{\mathrm{c}}B) = \Big(1 - P(A)\Big)P(B) = P(A^{\mathrm{c}})P(B).$$

But this is just the definition of independence of A^{c} and B. ∎

2.8 Additivity and the Inclusion–Exclusion Principle

The **Inclusion–Exclusion Principle** describes the full power of additivity of probability measures when applied to unions of not necessarily pairwise disjoint sets. Early on, we expect small children to understand the relation between sets and their cardinality—If Alex has three apples and Blair has two apples, then how many apples do they have together? The implicit assumption is that the two sets of apples are disjoint (since they belong to different children), then the measure (count) of the union is the sum of the counts. But what if Alex and Blair own some of their apples in common?

Pitman [4]: p. 22

2.8.1 Proposition (Inclusion–Exclusion Principle, I) *Even if $AB \neq \varnothing$,*

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

Proof: Now

$$A \cup B = (AB^{\mathrm{c}}) \cup (AB) \cup (BA^{\mathrm{c}}).$$

and

$$\begin{aligned}(A \cap B^{\mathrm{c}})(A \cap B) &= \varnothing \\ (A \cap B)(A^{\mathrm{c}} \cap B) &= \varnothing \\ (A \cap B^{\mathrm{c}})(A^{\mathrm{c}} \cap B) &= \varnothing.\end{aligned}$$

Therefore

$$P(A \cup B) = P(AB^{\mathrm{c}}) + P(AB) + P(BA^{\mathrm{c}}).$$

Now

$$\begin{aligned}P(A) &= P(AB^{\mathrm{c}}) + P(AB) \\ P(B) &= P(BA^{\mathrm{c}}) + P(AB).\end{aligned}$$

So

$$P(A) + P(B) = \underbrace{P(AB^{\mathrm{c}}) + P(AB) + P(BA^{\mathrm{c}})}\, + P(AB)$$

$$= P(A \cup B) + P(AB).$$

This implies

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

∎

Additionally,

$$\begin{aligned}P(A \cup B \cup C) = \;& P(A) + P(B) + P(C) \\ & - P(AB) - P(AC) - P(BC) \\ & + P(ABC).\end{aligned}$$

A more general version of the Inclusion–Exclusion Principle may be found in Pitman [4], Exercise 1.3.12, p. 31.

2.8.2 Proposition (General Inclusion–Exclusion Principle)

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} A_i\right) = & \sum_i p(A_i) \\ & - \sum_{i<j} P(A_i A_j) \\ & + \sum_{i<j<k} P(A_i A_j A_k) \\ & \vdots \\ & + (-1)^{n+1} P(A_1 A_2 \dots A_n). \end{aligned}$$

(Recall that intersection is denoted by placing sets next to each other. Note that the sign preceding a sum with the intersection of m sets is $(-1)^{m+1}$. The reason for summing over increasing indices is to avoid double counting.)

While it is possible top prove this result now using induction, I will put off a proof until we learn about the expectation of random variables, which will make the proof much easier.

2.9 Learning to count

The Uniform Probability (or counting) model was the earliest and hence one of the most pervasive probability models for that reason it is important to learn to count. This is the reason that probability are and combinatorics are closely related.

Robert Ash [1], section 1.4 has a really good discussion of counting principles. Also see Pitman [4, Section 1.6].

- How many different outcomes are there for the experiment of tossing a coin n times? 2^n.

- When order matters:

How many ways can a standard deck of 52 cards be arranged? 52! [Elaborate.]

- When order does not matter:

How many different 5-card poker hands are there? $\binom{52}{5}$ [Elaborate.]

2.10 Generally accepted counting principles

2.10.1 Lists versus sets

If you've had CS 1, you've studied Python which has both **lists** and **sets**. Both are collections of n objects, but two lists are different unless the same object appears

in the same *position* in both lists.

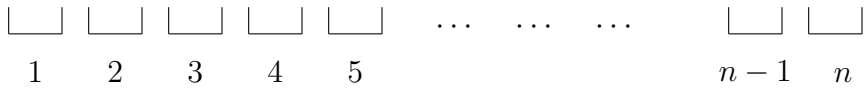
For instance,

123 and 213 are distinct lists of three elements, but the same set.

A list is sometimes referred to as a **permutation**.

2.10.2 Number of lists of length n

If I have n distinct objects, how many distinct ways can I can list them? Think of the object starting out in a bag and having to be distributed among n numbered boxes.



There are n choices for position (box) 1, $n-1$ for position 2, etc., so all together

there are $n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$ distinct lists of n objects.

The number $n!$ is read as ***n* factorial**.

By definition,

$$0! = 1.$$

2.10.3 Number of lists of length k of n objects

How many distinct lists of length k can I make with n objects? As before, there are n choices of the first position on the lists, and then $n-1$ choices for the second position, etc., down to $n-(k-1) = n-k+1$ choices for the k^{th} position on the list. Thus there are

$$\underbrace{n \times (n-1) \times \cdots \times (n-k+1)}_{k \text{ terms}}$$

distinct lists of k items chosen from n items. There is a more compact way to write this. Observe that

$$\begin{aligned}
&n \times (n-1) \times \cdots \times (n-k+1) \\
=&\frac{n \times (n-1) \times \cdots \times (n-k+1) \times (n-k) \times (n-k-1) \times \cdots \times 2 \times 1}{(n-k) \times (n-k-1) \times \cdots \times 2 \times 1} \\
=&\frac{n!}{(n-k)!}
\end{aligned}$$

Thus

there are $\dfrac{n!}{(n-k)!}$ distinct lists of length k chosen from n objects.

Note that when $k = n$ this reduces to $n!$ (since $0! = 1$), which agrees with the result in the previous section.

2.10.4 Number of subsets of size k of n objects

How many distinct subsets of size k can I make with n objects? (A subset is sometimes referred to as a **combination** of elements.) Well there are $\frac{n!}{(n-k)!}$ distinct lists of length k chosen from n objects. But when I have a set of k objects, I can write it $k!$ different ways as a list. Thus each set appears $k!$ in my listing of lists. So I have to take the number above and divide it by $k!$ to get the number of. Thus

there are $\dfrac{n!}{(n-k)! \cdot k!}$ distinct subsets of size k chosen from n objects.

2.10.1 Definition *For natural numbers* $0 \leqslant k \leqslant n$

$$\binom{n}{k} = \frac{n!}{(n-k)! \cdot k!},$$

is read as

"n choose k"

It is the number of distinct subsets of size k chosen from a set with n elements. It is also known as the **binomial coefficient**.

Other notations you may encounter include $C(n,k)$, ${}^nC_k$, and ${}_nC_k$. (These notations are easier to typeset in lines of text.)

2.10.5 Some useful identities

$$\binom{n}{n} = 1$$

$$\binom{n}{1} = n$$

$$\binom{n}{k} = \binom{n}{n-k}$$

$$\binom{n+1}{k+1} = \binom{n}{k+1} + \binom{n}{k} \tag{1}$$

You should do the calculations to verify (1). It gives rise to **Pascal's Triangle**, which gives $\binom{n}{k}$ as the k^{th} entry of the n^{th} row (where the numbering starts with $k = 0$). Each number is the sum of the two above it:

$$\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1 \\
1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1
\end{array}$$

etc.

Equation (1) also implies

$$\binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \cdots + (-1)^k \binom{n}{k} = (-1)^k \binom{n-1}{k}.$$

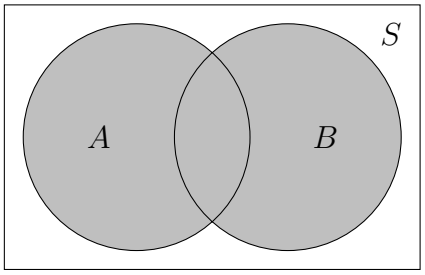
2.11 Appendix: Review of Set Operations

A quick review of set theory can be found in Ash [1], section 1.2. We shall follow Pitman [4], and use the notation AB rather than $A \cap B$ to denote the intersection of A and B.

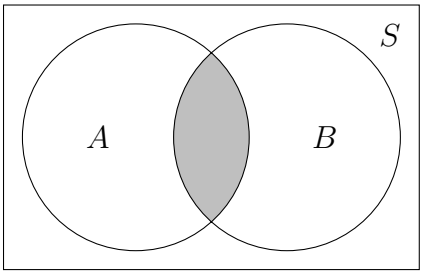
Pitman [4]: pp. 19–20

For subsets A and B of the set S we have the following **Venn diagrams**:

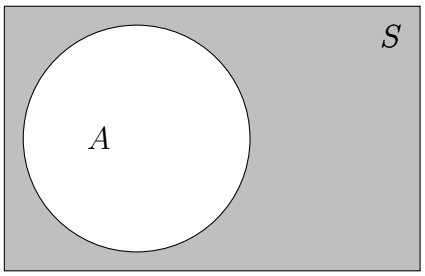
$A \cup B$:



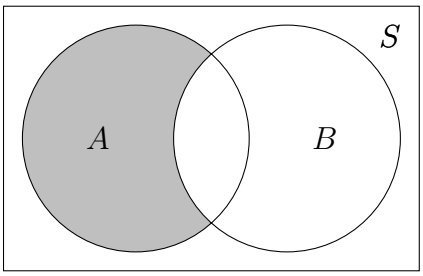
$A \cap B$ or AB :



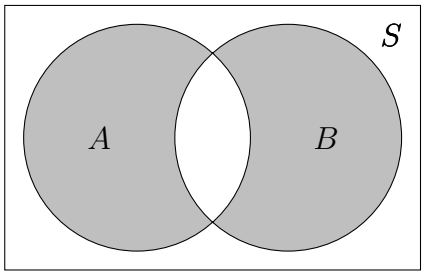
A^{c} :



$A \setminus B = AB^{\mathrm{c}}$:



$A \triangle B = (A \setminus B) \cup (B \setminus A) = (A \cup B) \setminus (AB)$:

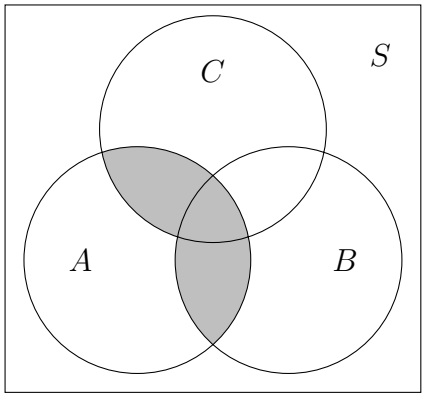


2.11.1 Definition *For any set E, let $|E|$ denote the* ***cardinality****, or number of elements, of E. We use this notation primarily with finite sets.*

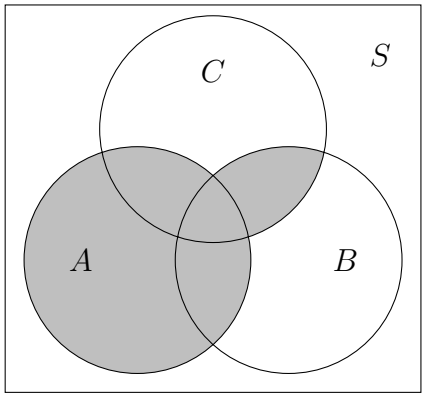
2.11.2 Definition *A* ***partition*** *of a set E is a collection $\mathcal{A}$ of subsets of E such that every point in E belongs to exactly one of the sets in $\mathcal{A}$*

Here are some useful identities.

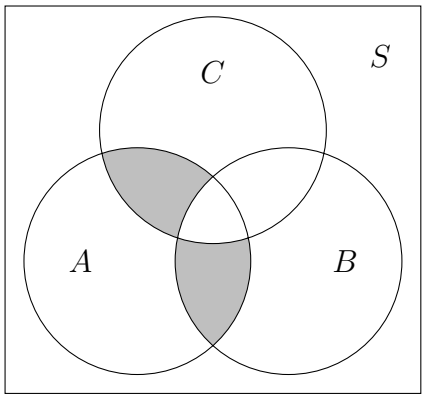
$A(B \cup C) = (AB) \cup (AC)$:



$A \cup (BC) = (A \cup B)(A \cup C)$:



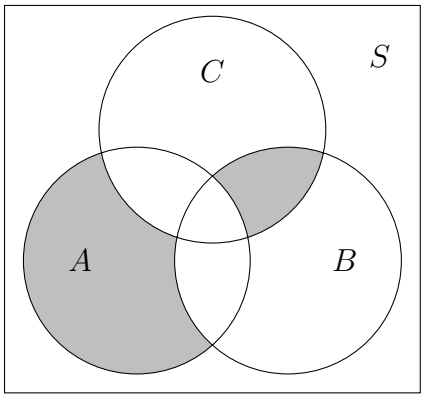
$A(B \triangle C) = (AB) \triangle (AC)$:



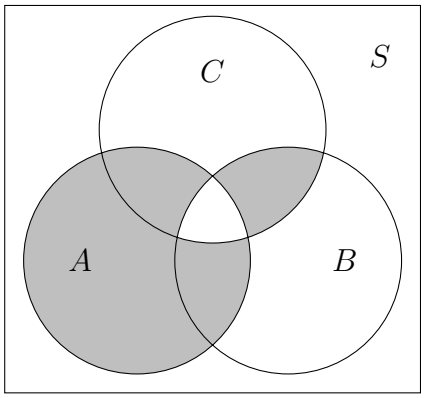
Note that

$$A \triangle (BC) \neq (A \triangle B)(A \triangle C).$$

$(A \triangle B)(A \triangle C)$



$A \triangle (BC)$



Aside: The use of the notation AB for the intersection of A and B suggests that intersection is a kind of multiplication operation for sets. In fact the set S acts as a multiplicative identity (unity or one). It also suggests that union may be a kind of addition with the empty set as the additive identity (or zero). A problem with this analogy is that there is then no additive inverse. That is, if A is nonempty, there is no set B such that $A \cup B = \varnothing$.

Aside: This is an aside to an aside, and should be ignored by everyone except math majors.

The integers under addition and multiplication form a **ring**: There is an additive identity, 0, and a multiplicative identity, 1, and every integer n has an additive inverse, $-n$, but not a multiplicative inverse. Moreover $0 \cdot n = 0$ for any integer n.

A similar algebraic structure exists for an algebra of subsets of S: Let intersection be multiplication, and let symmetric difference be addition. Both are commutative, and the distributive law $A(B \triangle C) = (AB) \triangle (AC)$ holds. The empty set $\varnothing$ is the additive identity, $A \triangle \varnothing = A$ and every set is its own additive inverse: $A \triangle A = \varnothing$. The multiplicative identity is S, $AS = A$. We also have $\varnothing A = \varnothing$ for any A.

Even cooler is the fact that the function d defined by $d(A, B) = P(A \triangle B)$ is a (semi-)metric.

Bibliography

[1] R. B. Ash. 2008. *Basic probability theory*. Mineola, New York: Dover. Reprint of the 1970 edition published by John Wiley and Sons.

[2] R. M. Gray. 1988. *Probability, random processes, and ergodic properties*. New York: Springer–Verlag.

[3] R. J. Larsen and M. L. Marx. 2012. *An introduction to mathematical statistics and its applications*, fifth ed. Boston: Prentice Hall.

[4] J. Pitman. 1993. *Probability*. Springer Texts in Statistics. New York, Berlin, and Heidelberg: Springer.