# Probability
# (Devore Chapter Two)

1016-345-01: Probability and Statistics for Engineers*

Spring 2013

## Contents


**0 Preliminaries** **3**
- 0.1 Motivation . . . 3
- 0.2 Administrata . . . 4
- 0.3 Outline . . . 5

**1 The Mathematics of Probability** **5**
- 1.1 Events . . . 5
  - 1.1.1 Logical Definition . . . 6
  - 1.1.2 Set Theory Definition . . . 6
- 1.2 The Probability of an Event . . . 7
- 1.3 Rules of Probability . . . 8
- 1.4 Venn Diagrams . . . 9
  - 1.4.1 Example . . . 12
- 1.5 Assigning Probabilities . . . 12

**2 Counting Techniques** **13**
- 2.1 Ordered Sequences . . . 13
- 2.2 Permutations and Combinations . . . 14

**3 Conditional Probabilities and Tree Diagrams** **17**
- 3.1 Example: Odds of Winning at Craps . . . 17
- 3.2 Definition of Conditional Probability . . . 19
- 3.3 Independence . . . 20

**4 Bayes's Theorem** **20**
4.1 Approach Considering a Hypothetical Population . . . . . . . . . . . . . . . . 20
4.2 Approach Using Axiomatic Probability . . . . . . . . . . . . . . . . . . . . . . 22

**Tuesday 5 March 2013**

# 0 Preliminaries

## 0.1 Motivation

Probability and Statistics are powerful tools, because they let us talk in quantitative ways about things that are uncertain or random. We may not be able to say whether the first person listed on page 47 of next year's phone book will be male or female, left-handed, right-handed or ambidextrous, but we can assign probabilities to those alternatives. And if we choose a large number of people at random we can predict that the number who are left-handed will fall within a certain range. Conversely, if we make many observations, we can use those to model the underlying probabilities of alternatives, and to predict the likely results of future observations. All of this can be done without deterministic predictions that one particular observation will definitely have a given result. (More than ever in this era of "Big Data", these tools are the key to interpreting the information to which we have access.)

With the concepts and methods we learn this quarter, we'll be able to answer questions like:

- One-tenth of one percent of the members of my ethnic group have a particular condition. I take a test for the condition which has a two percent false positive rate and a one percent false negatice rate, and it comes back positive. How likely is it that I have the disease? (Conditional Probabilities and Bayes's Theorem, week 1)
- If I flip a fair coin 100 times, what are the odds that it will come up heads 60 times or more? (Binomial Distribution, week 2)
- Historically, major earthquakes have occurred at an average rate of 13 per decade. What are the odds that there will be three or more in 2014? (Poisson Distribution, week 3)
- An experiment measures three quantities with specified uncertainties associated with each measurement. I construct a normalized "squared error distance" by scaling the error in each quantity by the measurement uncertainty, squaring these three numbers, and adding them. How likely is it that this number will be 6 or more? (Chi-Square Distribution, week 4)
- If I take ten independent measurements of the same quantity, each with the same uncertainty, and average them together, what will be the uncertainty associated with this combined measurement? (Random Samples, week 8)
- Some unknown fraction of the overall voting population prefers Candidate A to Candidate B. We survey 1000 of them at random, and 542 of them prefer Candidate A. What is the range of values for the unknown overall fraction of Candidate A supporters for which this is a reasonable result? (Interval Estimation, week 9)

Different branches of probability and statistics describe different parts of the problem:

- The rules of **Probability** allow us to make predictions about the relative likelihood of different possible samples, given properties of the underlying (real or hypothetical) population. This is the subject of Chapters Two to Five of Devore, and makes up much of this course.
- **Descriptive statistics** provides ways of describing properties of a sample to better understand it. This is the subject of Chapter One of Devore, which we'll turn to in the second half of the course.
- The field of **Inferential Statistics** is concerned with deducing the properties of an underlying population from the properties of a sample. This is the subject of Chapters Six and beyond of Devore, including Chapter Seven, which we will cover at the end of this course.

## 0.2 Administrata

- Syllabus
- Instructor's name (Whelan) rhymes with "wailin'".
- Text: Devore, *Probability and Statistics for Engineering and the Sciences*. The official version is the 8th edition, but the 7th edition is nearly identical. However, a few of the problems from the 7th edition have been replaced with different ones in the 8th edition, so make sure you're doing the correct homework problems. The 7th edition is on reserve at the Library.
- Course website: `http://ccrg.rit.edu/~whelan/1016-345/`
  - Contains links to quizzes and exams from previous sections of Probability and Statistics for Engineers (1016-345) and Probability (1016-351), which had a similar curriculum. (The corresponding course calendars will be useful for lining up the quizzes.)
- Course calendar: *tentative* timetable for course.
- Structure:
  - Read relevant sections of textbook before class
  - Lectures to reinforce and complement the textbook
  - Practice problems (odd numbers; answers in back but more useful if you try them before looking!).
  - Problem sets to hand in: practice at writing up your own work neatly & coherently. Note: doing the problems is *very* important step in mastering the material.
  - Quizzes: closed book, closed notes, use *scientific* calculator (not graphing calculator, *not* your phone!)
  - Prelim exam (think midterm, but there are two of them) in class at end of each half of course: closed book, one handwritten formula sheet, use scientific calculator (*not* your phone!)
  - Final exam to cover both halves of course

- Grading:

  5% Problem Sets
  10% Quizzes
  25% First Prelim Exam
  25% Second Prelim Exam
  35% Final Exam

  You'll get a separate grade on the "quality point" scale (e.g., 2.5–3.5 is the B range) for each of these five components; course grade is weighted average.

### 0.3 Outline

Part One (Probability):

1. Probability (Chapter Two)
2. Discrete Random Variables (Chapter Three)
3. Continuous Random Variables (Chapter Four)

Part Two (Statistics):

1. Descriptive Statistics (Chapter One)
2. Probability Plots (Section 4.6)
3. Joint Probability Distributions & Random Samples (Chapter Five)
4. Interval Estimation (Chapter Seven)

Warning: this class will move pretty fast, and in particular we'll rely on your knowledge of calculus.

## 1 The Mathematics of Probability

Many of the rules of probability appear to be self-evident, but it's useful to have a precise language in which they can be described. To that end, Devore develops with some care a mathematical theory of probability. Here we'll mostly summarize the key definitions and results, and try to make contact with their practical uses.

Ultimately, the whole game is about assigning a *probability*, i.e., a number between 0 and 1, to each *event* associated with a problem.

### 1.1 Events

"Event" is a technical term; there are two equivalent definitions, one related to how we define events in practice, and one related to the set theory manipulations that help us calculate probabilities. An event can be thought of as either

- a statement which could be either true or false, or
- a set of possible outcomes to an experiment.

### 1.1.1 Logical Definition

In practice, when we approach a problem, we will define events to which we need to associate probabilities, like "this roll of a pair of six-sided dice will total seven" or "four or more of the thousand circuit boards which I test will fail quality control" or "the patient being tested has cancer". These are all declarative statements which could be true or false. The most interesting events are those which are not definitely true or definitely false. We usually consider this uncertainty to be a result of some randomness in the problem. Ideally, we should be considering an experiment which can be repeated under identical circumstances, and in some repetitions the statement associated with the event will be true, and in others it will be false. Alternatively we could have some large population of individuals, and the statement will be true for some individuals and false for others; picking an individual at random means we have some probability of the statement being true. The uncertainty can also arise from our incomplete knowledge of the system. Even though each molecule in a chamber of gas obeys deterministic laws of physics, and we could determine its trajectory from its initial conditions, we treat thermodynamics as a statistical science, since we only can really keep track of the bulk description of the system.

If we think about events in terms of statements about a system or the world as a whole, we can combine the events according to some familiar rules of logic:

- The event $A'$ ("not $A$") corresponds to a statement which is true if $A$'s statement is false, and false if $A$'s statement is true. This is also known as the **complement** of $A$.
- The event $A \cup B$ ("$A$ or $B$") is defined by an inclusive or. Its statement is true if $A$'s statement, or $B$'s, or both, are true. This is also known as the **union** of $A$ and $B$.
- The event $A \cap B$ ("$A$ and $B$") is defined by a logical and. Its statement is true if $A$'s statement and $B$'s are both true. This is also known as the **intersection** of $A$ and $B$.

Finally, events $A$ and $B$ are called **mutually exclusive** if they corresponding statements cannot both be true.

### 1.1.2 Set Theory Definition

The notation for these combined events comes from the other interpretation, using set theory, which is Devore's starting point. Devore defines probability in terms of an **experiment** which can have one of a set of possible **outcomes**.

- The **sample space** of an experiment, written $\mathcal{S}$, is the set of all possible outcomes.
- An **event** is a subset of $\mathcal{S}$, a set of possible outcomes to the experiment. Special cases are:
  - The **null event** $\varnothing$ is an event consisting of no outcomes (the empty set)
  - A **simple event** consists of exactly one outcome
  - A **compound event** consists of more than one outcome

The sample space $\mathcal{S}$ itself an event, of course.

The way in which the two definitions are equivalent is that the event $A$ is the set of all outcomes for which the corresponding statement is true.

One example of an experiment is flipping a coin three times. The outcomes in that case are $HHH$, $HHT$, $HTH$, $HTT$, $THH$, $THT$, $TTH$, and $TTT$. Possible events include:

- Exactly two heads: $\{HHT, HTH, THH\}$
- The first flip is heads: $\{HHH, HHT, HTH, HTT\}$
- The second and third flips are the same: $\{HHH, HTT, THH, TTT\}$

Note that each of these events is associated with a statement, as described in the logical approach.

The various logical operations for combining events can be phrased in terms of set theory operations:

- The **complement** $A'$ ("not $A$") of an event $A$, is the set of all outcomes in $\mathcal{S}$ which are *not* in $A$.
- The **union** $A \cup B$ ("$A$ or $B$") of two events $A$ and $B$, is the set of all outcomes which are in $A$ *or* $B$, including those which are in both.
- The **intersection** $A \cap B$ ("$A$ and $B$") is the set of all outcomes which are in both $A$ *and* $B$.

In the case of coin flips, if the events are $A = \{HHT, HTH, THH\}$ (exactly two heads) and $B = \{HHH, HHT, HTH, HTT\}$ (first flip heads), we can construct, among other things,

$$
\begin{aligned}
A' &= \{HHH, HTT, THT, TTH, TTT\} \\
A \cup B &= \{HHH, HHT, HTH, HTT, THH\} \\
A \cap B &= \{HHT, HTH\}
\end{aligned}
$$

Another useful definition is that $A$ and $B$ are **disjoint** or **mutually exclusive** events if $A \cap B = \varnothing$. In the logical picture, two disjoint events correspond to statements which cannot both be true (e.g., "an individual is under five feet tall" and "an individual is over six feet tall").

**Note that the trickiest part of many problems is actually keeping straight what the events are to which you're assigning probabilities!**

## 1.2 The Probability of an Event

Mathematically speaking, probability is a number between 0 ans 1 which is assigned to each event. I.e., the event $A$ has probability $P(A)$. If we think about the logical definition of events, then we have

- $P(A) = 1$ means the statement corresponding to $A$ is definitely true.

- $P(A) = 0$ means the statement corresponding to $A$ is definitely false.
- $0 < P(A) < 1$ means the statement corresponding to $A$ could be true or false.

The standard numerical interpretation of the probability $P(A)$ is in terms of a repeatable experiment with some random element. Imagine that we repeat the same experiment over and over again many times under identical conditions. In each iteration of the experiment (each game of craps, sequence of coin flips, opinion survey, etc), a given outcome or event will represent a statement that is either true or false. Over the long run, the fraction of experiments in which the statement is true will be approximately given by the probability of the corresponding outcome or event. If we write the number of repetitions of the experiment as $N$, and the number of experiments out of those $N$ in which $A$ is true as $N_A(N)$, then

$$\lim_{N\to\infty} \frac{N_A(N)}{N} = P(A) \tag{1.1}$$

You can test this proposition on the optional numerical exercise on this week's problem set. This interpretation of probability is sometimes called the "frequentist" interpretation, since it involves the relative frequency of outcomes in repeated experiements. It's actually a somewhat more limited interpretation than the "Bayesian" interpretation, in which the probability of an event corresponds to a quantitative degree of certainty that the corresponding statement is true. (Devore somewhat pejoratively calls this "subjective probability".) These finer points are beyond the scope of this course, but if you're interested, you may want to look up e.g., *Probability Theory: The Logic of Science* by E. T. Jaynes.

## 1.3 Rules of Probability

Devore develops a formal theory of probability starting from a few axioms, and derives other sensible results from those. This is an interesting intellectual exercise, but for our purposes, it's enough to note certain simple properties which make sense for our understanding of probability as the likelihood that a statement is true:

1. For any event $A$, $0 \leq P(A) \leq 1$
2. $P(\mathcal{S}) = 1$ and $P(\varnothing) = 0$ (something always happens)
3. $P(A') = 1 - P(A)$ (the probability that a statement is false is one minus the probability that it's true.
4. If $A$ and $B$ are disjoint events, $P(A \cup B) = P(A) + P(B)$

One useful non-trivial result concerns the probability of the union of any two events. Since $A \cup B = (A \cap B') \cup (A \cap B) \cup (A' \cap B)$, the union of three disjoint events,

$$P(A \cup B) = P(A \cap B') + P(A \cap B) + P(A' \cap B) \tag{1.2}$$

On the other hand, $A = (A \cap B') \cup (A \cap B)$ and $B = (A \cap B) \cup (A' \cap B)$, so

$$P(A) = P(A \cap B') + P(A \cap B) \tag{1.3a}$$

$$P(B) = P(A \cap B) + P(A' \cap B) \tag{1.3b}$$

which means that

$$P(A) + P(B) = P(A \cap B') + 2P(A \cap B) + P(A' \cap B) = P(A \cup B) + P(A \cap B) \quad (1.4)$$

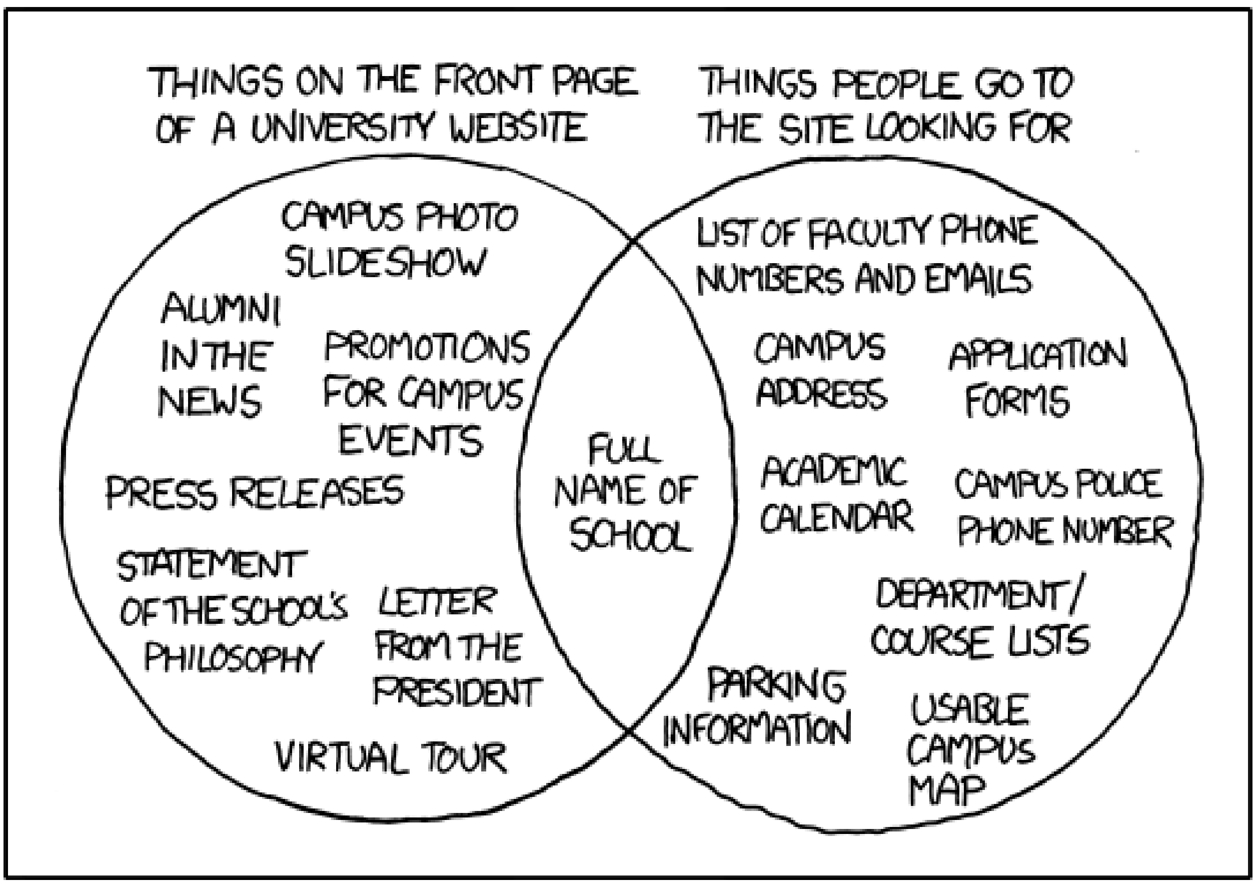
so

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (1.5)$$

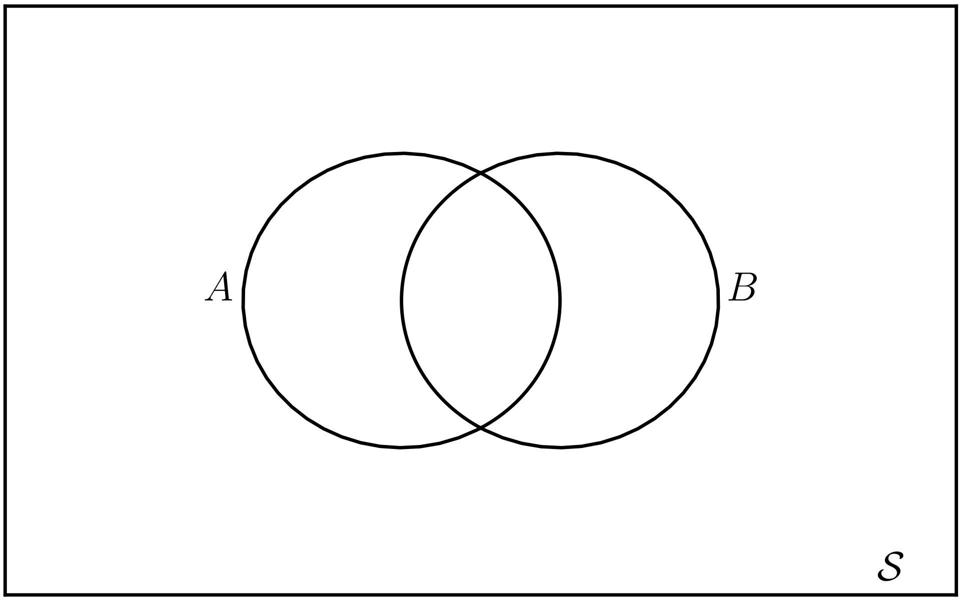
## 1.4 Venn Diagrams

"University Website" http://xkcd.com/773/

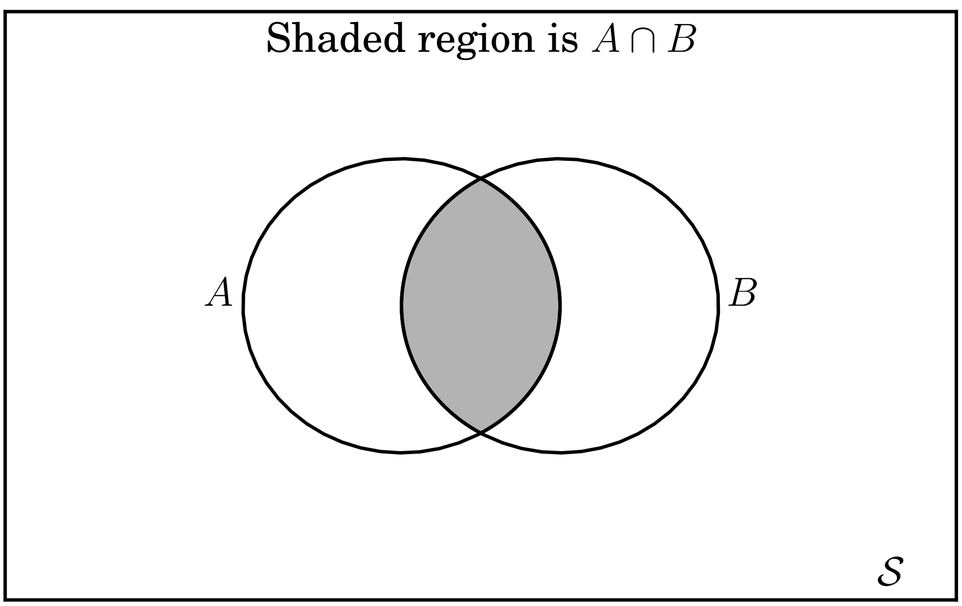
*Note that this Venn diagram illustrates the relationship between two sets of items, so it's a more general set theory application rather than one specific to outcomes and events.*



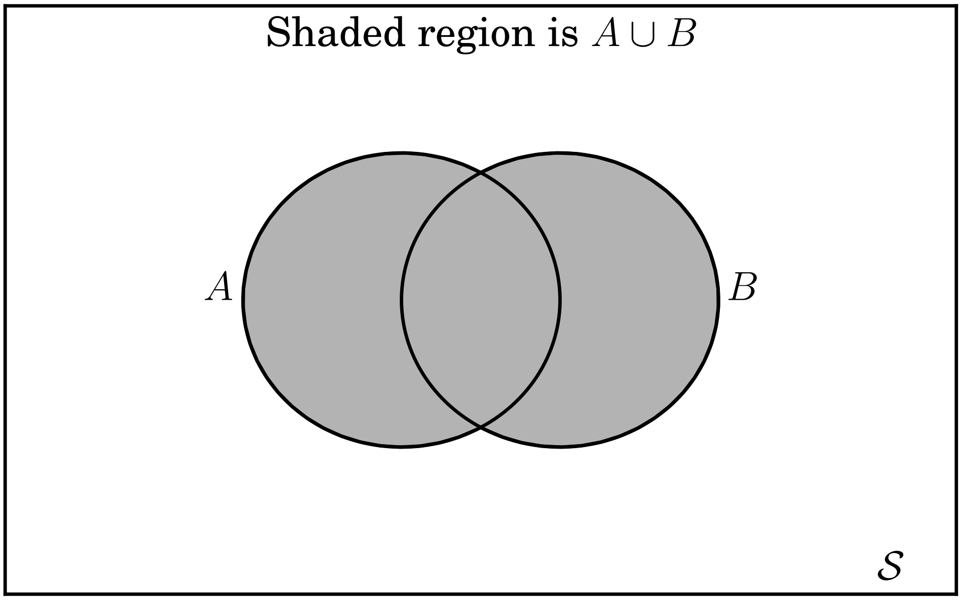
We can often gain insight into addition of probabilities with Venn diagrams, a tool from set theory in which sets, in this case events, are represented pictorially as regions in a plane. For example, here the two overlapping circles represent the events $A$ and $B$:




The intersection of those two events is shaded here:

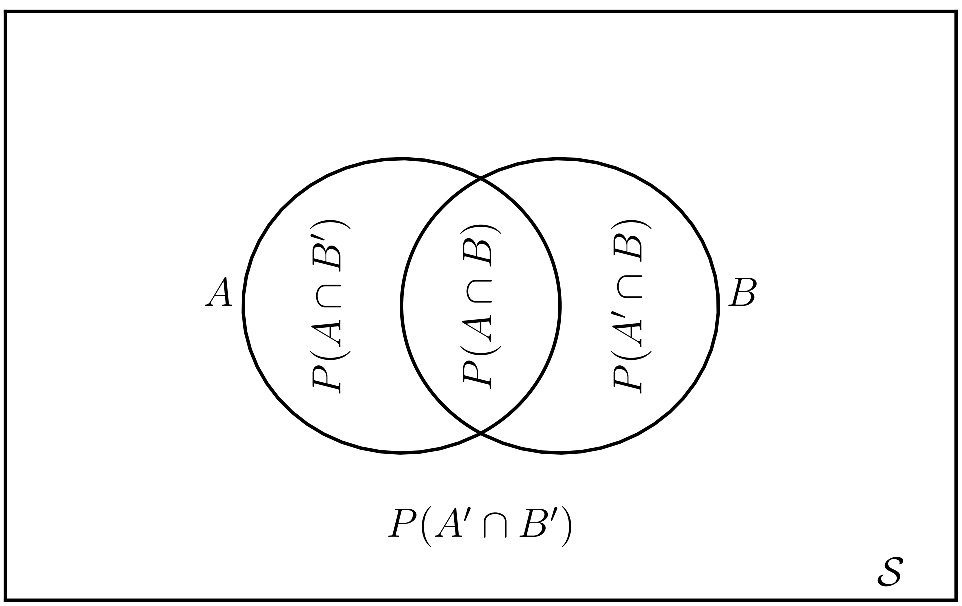


The union the two events is shaded here:



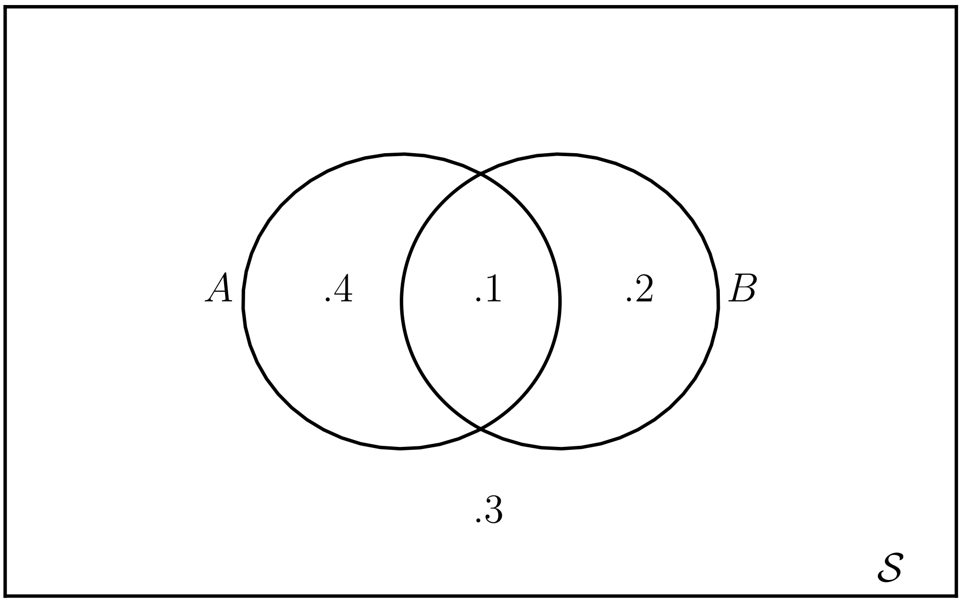

Thus we see that if we add $P(A)$ and $P(B)$ by counting all of the outcomes in each circle, we've double-counted the outcomes in the overlap, which is why we have to subtract $P(A \cap B)$ in (1.5)

It's sometimes useful to keep track of the probabilities of various intersections of events by writing those probabilities in a Venn diagram like this:



### 1.4.1 Example

Suppose we're told that $P(A) = .5$, $P(B) = .6$, and $P(A \cap B) = .4$ and asked to calculate something like $P(A \cup B)$. We could just use the formula, of course, or we could fill in the probabilities on a Venn diagram. Since $P(A \cap B) = .4$, we must have $P(A \cap B') = .1$ in order that the two will add up to $P(A) = .5$, and likewise we can deduce that $P(A' \cap B) = .2$ and fill in the probabilities like so:




Note that the four regions of the Venn diagram represent an exhaustive set of mutually exclusive events, so their probabilities have to add up to 1.

## 1.5 Assigning Probabilities

If we have a way of assigning probabilities to each outcome, and therefore each simple event, then we can use the sum rule for disjoint events to write the probability of any event as the sum of the probabilities of the simple events which make it up. I.e.,

$$P(A) = \sum_{E_i \text{ in } A} P(E_i) \tag{1.6}$$

One possibility is that each outcome, i.e., each simple event, might be equally likely. In that case, if there are $N$ outcomes total, the probability of each of the simple events is $P(E_i) = 1/N$ (so that $\sum_{i=1}^{N} P(E_i) = P(\mathcal{S}) = 1$), and in that case

$$P(A) = \sum_{E_i \text{ in } A} \frac{1}{N} = \frac{N(A)}{N} \tag{1.7}$$

where $N(A)$ is the number of outcomes which make up the event $A$.

Note, however, that one has to consider whether it's appropriate to take all of the outcomes to be equally likely. For instance, in our craps example, we considered each roll, e.g., 2 and 4 to be its own outcome. But you can also consider the rolls of the individual dice, and then the two dice totalling 4 would be a composite event consisting of the outcomes $(1, 3)$, $(2, 2)$, and $(3, 1)$. For a pair of fair dice, the 36 possible outcomes defined by the numbers on the two dice taken in order (suppose one die is green and the other red) are equally likely outcomes.

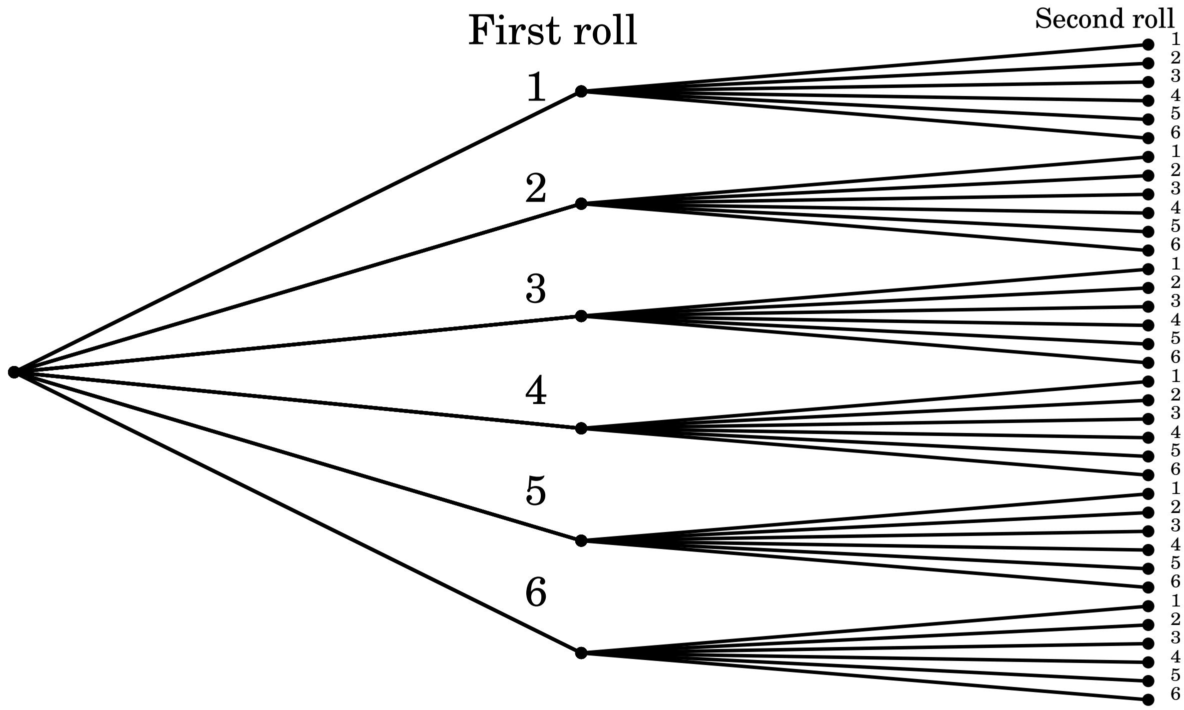
# 2 Counting Techniques

## 2.1 Ordered Sequences

We can come up with 36 as the number of possible results on a pair of fair dice in a couple of ways. We could make a table

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 6 | 7 | 8 | 9 | 10 | 11 | 12 |

which is also useful for counting the number of occurrences of each total. Or we could use something called a tree diagram:




This works well for counting a small number of possible outcomes, but already with 36 outcomes it is becoming unwieldy. So instead of literally counting the possible outcomes, we should calculate how many there will be. In this case, where the outcome is an ordered pair of numbers from 1 to 6, there are 6 possibilities for the first number, and corresponding to each of those there are 6 possibilities for the second number. So the total is $6 \times 6 = 36$.

More generally, if we have an ordered set of $k$ objects, with $n_1$ possibilities for the first, $n_2$ for the second, etc, the number of possible ordered $k$-tuples is $n_1 n_2 \ldots n_k$, which we can also write as

$$\prod_{i=1}^{k} n_i \; . \tag{2.1}$$

## 2.2 Permutations and Combinations

Consider the probability of getting a poker hand (5 cards out of the 52-card deck) which consists entirely of hearts.[1] Since there are four different suits, you might think the odds are $(1/4)(1/4)(1/4)(1/4)(1/4) = (1/4)^5 = 1/4^5$. However, once a heart has been drawn on the first card, there are only 12 hearts left in the deck out of 51; after two hearts there are 11 out of 50, etc., so the actual odds are

$$P(\heartsuit\heartsuit\heartsuit\heartsuit\heartsuit) = \left(\frac{13}{52}\right)\left(\frac{12}{51}\right)\left(\frac{11}{50}\right)\left(\frac{10}{49}\right)\left(\frac{9}{48}\right) \tag{2.2}$$

[1]This is, hopefully self-apparently, one-quarter of the probability of getting a flush of any kind.

This turns out not to be the most effective way to calculate the odds of poker hands, though. (For instance, it's basically impossible to do a card-by-card accounting of the probability of getting a full house.) Instead we'd like to take the approach of counting the total number of possible five-card hands (outcomes) and then counting up how many fall into a particular category (event). The terms for the quantities we will be interested in are **permutation** and **combination**.

First, let's consider the number of possible sequences of five cards drawn out of a deck of 52. This is the permutation number of permutations of 5 objects out of 52, called $P_{5,52}$. The first card can be any of the 52; the second can be any of the remaining 51; the third can be any of the remaining 50, etc. The number of permutations is

$$P_{5,52} = 52 \times 51 \times 50 \times 49 \times 48 \tag{2.3}$$

In general

$$P_{k,n} = n(n-1)(n-2)\cdots(n-k+1) = \prod_{\ell=0}^{k-1}(n-\ell) \; . \tag{2.4}$$

Now, there is a handy way to write this in terms of the factorial function. Remember that the factorial is defined as

$$n! = n(n-1)(n-1)\cdots(2)(1) = \prod_{\ell=1}^{n}\ell \tag{2.5}$$

with the special case that $0! = 1$. Then we can see that

$$\begin{aligned}\frac{n!}{(n-k)!} &= \frac{n(n-1)(n-2)\cdots(n-k+1)\cancel{(n-k)}\cancel{(n-k-1)}\cdots\cancel{(2)}\cancel{(1)}}{\cancel{(n-k)}\cancel{(n-k-1)}\cdots\cancel{(2)}\cancel{(1)}} \\ &= P_{k,n}\end{aligned} \tag{2.6}$$

Note in particular that the number of ways of arranging $n$ items is

$$P_{n,n} = \frac{n!}{(n-n)!} = \frac{n!}{0!} = n! \tag{2.7}$$

Now, when we think about the number of different poker hands, actually we don't consider the cards in a hand to be ordered. So in fact all we care about is the number of ways of choosing 5 objects out of a set of 52, without regard to order. This is the number of **combinations**, which is sometimes written $C_{5,52}$, but which we'll write as $\binom{52}{5}$, pronounced "52 choose 5". When we counted the number of different permutations of 5 cards out of 52, we actually counted each possible hand a bunch of times, once for each of the ways of arranging the cards. There are $P_{5,5} = 5!$ different ways of arranging the five cards of a poker hand, so the number of permutations of 5 cards out of 52 is the number of combinations times the number of permutations of the 5 cards among themselves:

$$P_{5,52} = \binom{52}{5} P_{5,5} \tag{2.8}$$

The factor of $P_{5,5} = 5!$ is the factor by which we overcounted, so we divide by it to get

$$\binom{52}{5} = \frac{P_{5,52}}{P_{5,5}} = \frac{52!}{47!5!} = 2598960 \tag{2.9}$$

or in general

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} \tag{2.10}$$

So to return to the question of the odds of getting five hearts, there are $\binom{52}{5}$ different poker hands, and $\binom{13}{5}$ different hands of all hearts (since there are 13 hearts in the deck), which means the probability of the event $A = \heartsuit\heartsuit\heartsuit\heartsuit\heartsuit$ is

$$P(A) = \frac{N(A)}{N} = \frac{\binom{13}{5}}{\binom{52}{5}} = \frac{\frac{13!}{8!5!}}{\frac{52!}{47!5!}} = \frac{13!47!}{8!52!} = \frac{(13)(12)(11)(10)(9)}{(52)(51)(50)(49)(48)} \tag{2.11}$$

which is of course what we calculated before. Numerically, $P(A) \approx 4.95 \times 10^{-4}$, while $1/4^5 \approx 9.77 \times 10^{-4}$. The odds of getting any flush are four times the odds of getting an all heart flush, i.e., $1.98 \times 10^{-3}$.

Actually, if we want to calculate the odds of getting a flush, we have over-counted somewhat, since we have also included straight flushes, e.g., 4$\heartsuit$-5$\heartsuit$-6$\heartsuit$-7$\heartsuit$-8$\heartsuit$. If we want to count only hands which are flushes, we need to subtract those. Since aces can count as either high or low, there are ten different all-heart straight flushes, which means the number of different all-heart flushes which are not straight flushes is

$$\binom{13}{5} - 10 = \frac{13!}{8!5!} - 10 = 1287 - 10 = 1277 \tag{2.12}$$

and the probability of getting an all-heart flush is $4.92 \times 10^{-4}$, or $1.97 \times 10^{-3}$ for any flush.

Exercise: work out the number of possible straights and therefore the odds of getting a straight.

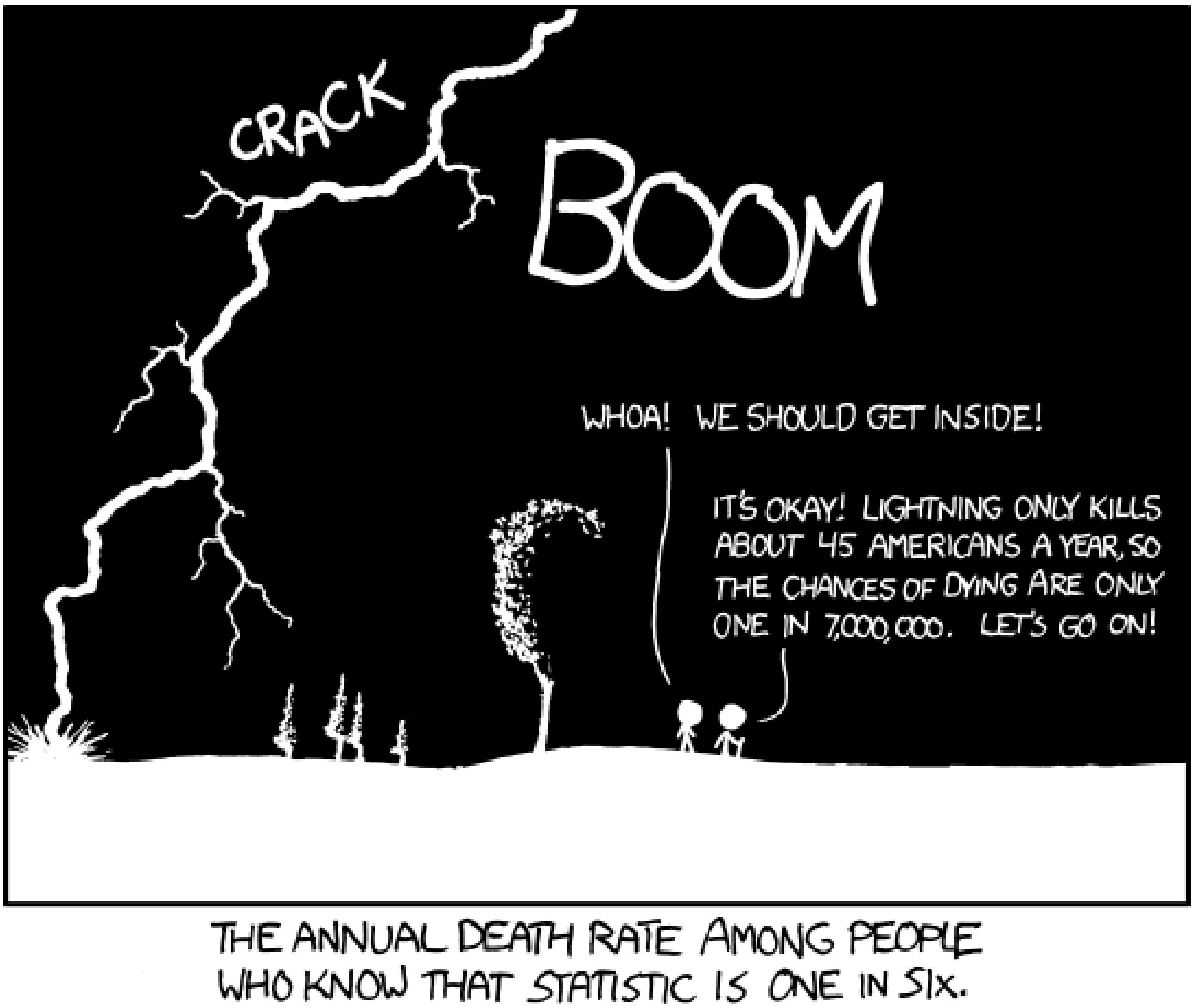
## Practice Problems

2.5, 2.9, 2.13, 2.17, 2.29, 2.33, 2.43

**Thursday 7 March 2013**

# 3 Conditional Probabilities and Tree Diagrams

"Conditional Risk" http://xkcd.com/795/




## 3.1 Example: Odds of Winning at Craps

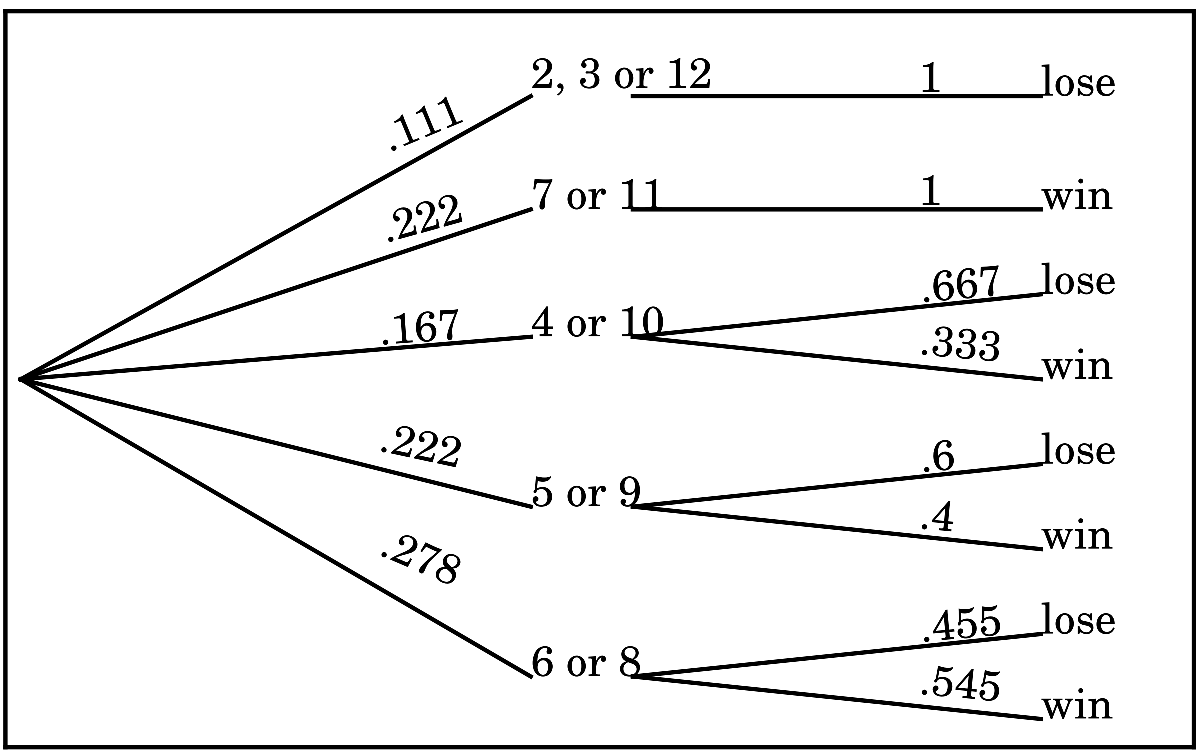
Although there are an infinite number of possible outcomes to a craps game, we can still calculate the probability of winning.

First, the sample space can be divided up into mutually exclusive events based on the result of the first roll:

| Event | Probability | Result of game |
|---|---|---|
| 2, 3 or 12 on 1st roll | $\frac{1+2+1}{36} = \frac{4}{36} \approx 11.1\%$ | lose |
| 7 or 11 on 1st roll | $\frac{6+2}{36} = \frac{8}{36} \approx 22.2\%$ | win |
| 4 or 10 on 1st roll | $\frac{3+3}{36} = \frac{6}{36} \approx 16.7\%$ | ??? |
| 5 or 9 on 1st roll | $\frac{4+4}{36} = \frac{8}{36} \approx 22.2\%$ | ??? |
| 6 or 8 on 1st roll | $\frac{5+5}{36} = \frac{10}{36} \approx 27.8\%$ | ??? |

The last three events each contain some outcomes that correspond to winning, and some that correspond to losing. We can figure out the probability of winning if, for example, you roll a 4 initially. Then you will win if another 4 comes up before a 7, and lose if a 7 comes up before a 4. On any given roll, a 7 is twice as likely to come up as a 4 (6/36 vs 3/36), so the odds are $6/9 = 2/3 \approx 66.7\%$ that you will roll a 7 before a 4 and lose. Thus the odds of losing after starting with a 4 are 66.7%, while the odds of winning after starting with a 4 are 33.3%. The same calculation applies if you get a 10 on the first roll. This means that the $6/36 \approx 16.7\%$ probability of rolling a 4 or 10 initially can be divided up into a $4/36 \approx 11.1\%$ probability to start with a 4 or 10 and eventually lose, and a $2/36 \approx 5.6\%$ probability to start with a 4 or 10 and eventually win.

We can summarize this branching of probabilities with a tree diagram:




The probability of winning given that you've rolled a 4 or 10 initially is an example of a conditional probability. If $A$ is the event "roll a 4 or 10 initially" and $B$ is the event "win the game", we write the conditional probability for event $B$ given that $A$ occurs as $P(B|A)$. We have argued that the probability for both $A$ and $B$ to occur, $P(A \cap B)$, should be the probability of $A$ times the conditional probability of $B$ given $A$, i.e.,

$$P(A \cap B) = P(B|A)P(A) \tag{3.1}$$

We can use this to fill out a table of probabilities for different sets of outcomes of a craps game, analogous to the tree diagram.

| $A$ | $P(A)$ | $B$ | $P(B\|A)$ | $P(A \cap B) = P(B\|A)P(A)$ |
|---|---|---|---|---|
| 2, 3 or 12 on 1st roll | .111 | lose | 1 | .111 |
| 7 or 11 on 1st roll | .222 | win | 1 | .222 |
| 4 or 10 on 1st roll | .167 | lose | .667 | .111 |
| | | win | .333 | .056 |
| 5 or 9 on 1st roll | .222 | lose | .6 | .133 |
| | | win | .4 | .089 |
| 6 or 8 on 1st roll | .278 | lose | .545 | .152 |
| | | win | .455 | .126 |

Since the rows all describe disjoint events whose union is the sample space $\mathcal{S}$, we can add the probabilities of winning and find that

$$P(\text{win}) \approx .222 + .056 + .089 + .126 \approx .493 \tag{3.2}$$

and

$$P(\text{lose}) \approx .111 + .111 + .133 + .152 \approx .507 \tag{3.3}$$

## 3.2 Definition of Conditional Probability

We've motivated the concept of conditional probability and applied it via (3.1). In fact, from a formal point of view, conditional probability is *defined* as

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \ . \tag{3.4}$$

We actually used that definition in another context above without realizing it, when we were calculating the probability of rolling a 7 before rolling a 4. We know that $P(7) = 6/36$ and $P(4) = 3/36$ on any given roll. The probability of rolling a 7 given that the game ends on that throw is

$$P(7|7 \cup 4) = \frac{P(7)}{P(7 \cup 4)} = \frac{P(7)}{P(7) + P(4)} = \frac{6/36}{9/36} = \frac{6}{9} \tag{3.5}$$

We calculated that using the definition of conditional probability.

## 3.3 Independence

We will often say that two events $A$ and $B$ are independent, which means that the probability of $A$ happening is the same whether or not $B$ happens. This can be stated mathematically in several equivalent ways:

1. $P(A|B) = P(A)$
2. $P(B|A) = P(B)$
3. $P(A \cap B) = P(A)P(B)$

The first two tie most obviously to the notion of independence (the probability of $A$ given $B$ is the same as the probability of $A$), but the third is more convenient. To show that they're equivalent, suppose that $P(A \cap B) = P(A)P(B)$. In that case, it's easy to show

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)\cancel{P(B)}}{\cancel{P(B)}} = P(A) \qquad \text{if } P(A \cap B) = P(A)P(B) \tag{3.6}$$

The symmetrical definition also works even if $P(A)$ or $P(B)$ are zero, and it extends obviously to more than two events:

$$A\text{, } B\text{, and } C \text{ are } \textit{mutually independent} \qquad \equiv \qquad P(A \cap B \cap C) = P(A)P(B)P(C) \tag{3.7}$$

# 4 Bayes's Theorem

*Some of the arguments in this section are adapted from*
`http://yudkowsky.net/rational/bayes`
*which gives a nice explanation of Bayes's theorem.*

The laws of probability are pretty good at predicting how likely something is to happen given certain underlying circumstances. But often what you really want to know is the opposite: given that some thing happened, what were the circumstances? The classic example of this is a test for a disease.

Suppose that one one-thousandth of the population has a disease. There is a test that can detect the disease, but it has a 2% false positive rate (on average one out of fifty healthy people will test positive) and as 1% false negative rate (on average one out of one hundred sick people will test negative). The question we ultimately want to answer is: if someone gets a positive test result, what is the probability that they actually have the disease. Note, it is not 98%!

## 4.1 Approach Considering a Hypothetical Population

The standard treatment of Bayes's Theorem and the Law of Total Probability can be sort of abstract, so it's useful to keep track of what's going on by considering a hypothetical population which tracks the various probabilities. So, assume the probabilities arise from a population of 100,000 individuals. Of those, one one-one-thousandth, or 100, have the disease. The other 99,900 do not. The 2% false positive rate means that of the 99,900 healthy individuals, 2% of them, or 1,998, will test positive. The other 97,902 will test negative. The 1% false negative rate means that of the 100 sick individuals, one will test negative and the other 99 will test positive. So let's collect this into a table:

| | Positive | Negative | Total |
|---|---|---|---|
| Sick | 99 | 1 | 100 |
| Healthy | 1,998 | 97,902 | 99,900 |
| Total | 2,097 | 97,903 | 100,000 |

(As a reminder, if we choose a *sample* of 100,000 individuals out of a larger population, we won't expect to get exactly this number of results, but the 100,000-member population is a useful conceptual construct.)

Translating from numbers in this hypothetical population, we can confirm that it captures the input information:

$$P(\text{sick}) = \frac{100}{100,000} = .001 \tag{4.1a}$$

$$P(\text{positive}|\text{healthy}) = \frac{1,998}{99,900} = .02 \tag{4.1b}$$

$$P(\text{negative}|\text{sick}) = \frac{1}{100} = .01 \tag{4.1c}$$

But now we can also calculate what we want, the conditional probability of being sick given a positive result. That is the fraction of the total number of individuals with positive test results that are in the "sick *and* positive" category:

$$P(\text{sick}|\text{positive}) = \frac{99}{2,097} \approx .04721 \tag{4.2}$$

or about 4.7%.

Note that we can forego the artificial construct of a 100,000-member hypothetical population. If we divide all the numbers in the table by 100,000, they become probabilities for the corresponding events. For example,

$$P(\text{sick} \cap \text{positive}) = \frac{99}{100,000} = .00099$$

That is the approach of the slightly more axiomatic (and general) method described in the next section.

## 4.2 Approach Using Axiomatic Probability

The quantity we're looking for (the probability of being sick, given a positive test result) is a conditional probability. To evaluate it, we need to define some events about which we'll discuss the probability. First, consider the events

$$A_1 \equiv \text{“individual has the disease”} \tag{4.3a}$$
$$A_2 \equiv \text{“individual does not have the disease”} \tag{4.3b}$$

These make a mutually exclusive, exhaustive set of events, i.e., $A_1 \cap A_2 = \varnothing$ and $A_1 \cap A_2 = \mathcal{S}$. (We call them $A_1$ and $A_2$ because in a more general case there might be more than two events in the mutually exclusive, exhaustive set.) We are told in the statement of the problem that one person in 1000 has the disease, which means that

$$P(A_1) = .001 \tag{4.4a}$$
$$P(A_2) = 1 - P(A_1) = .999 \tag{4.4b}$$
$$\tag{4.4c}$$

Now consider the events associated with the test:

$$B \equiv \text{“individual tests positive”} \tag{4.5a}$$
$$B' \equiv \text{“individual tests negative”} \tag{4.5b}$$

(We call them $B$ and $B'$ because we will focus more on $B$.) The 2% false positive and 1% false negative rates tell us

$$P(B|A_2) = .02 \tag{4.6a}$$
$$P(B'|A_1) = .01 \tag{4.6b}$$

Note that if we want to talk about probabilities involving $B$, we should use the fact that

$$P(B|A_1) + P(B'|A_1) = 1 \tag{4.7}$$

to state that

$$P(B|A_1) = 1 - P(B'|A_1) = .99 \tag{4.8}$$

Now we can write down the quantity we actually want, using the definition of conditional probability:

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} \tag{4.9}$$

Now, we don't actually have an expression for $P(A_1 \cap B)$ or $P(B)$ yet. The fundamental things we know are

$$P(A_1) = .001 \tag{4.10a}$$
$$P(A_2) = .999 \tag{4.10b}$$
$$P(B|A_2) = .02 \tag{4.10c}$$
$$P(B|A_1) = .99 \tag{4.10d}$$

However, we know how to calculate $P(A_1 \cap B)$ and $P(B)$ from the things we do know. First, to get $P(A_1 \cap B)$, we can notice that we know $P(B|A_1)$ and $P(A_1)$, so we can solve

$$P(B|A_1) = \frac{P(A_1 \cap B)}{P(A_1)} \tag{4.11}$$

for

$$P(A_1 \cap B) = P(B|A_1)P(A_1) = (.99)(.001) = .00099 \tag{4.12}$$

Logically, the probability of having the disease *and* testing positive for it is the probability of having the disease in the first place times the probability of testing positive, given that you have the disease.

Since we know $P(B|A_2)$ and $P(A_2)$ we can similarly calculate the probability of not having the disease but testing positive anyway:

$$P(A_2 \cap B) = P(B|A_2)P(A_2) = (.02)(.999) = .01998 \tag{4.13}$$

But now we have enough information to calculate $P(B)$, since the overall probability of testing positive for the disease has to be the probability of having the disease and testing positive plus the probability of not having the disease and testing positive:

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) = .00099 + .01998 = .02097 \tag{4.14}$$

This is an application of the *Law of total probability*, which says, in the more general case of $k$ mutually exclusive, exhaustive events $A_1, \ldots, A_k$,

$$P(B) = P(B \cap A_1) + \cdots + P(B \cap A_k) = P(B|A_1)P(A_1) + \cdots + P(B|A_k)P(A_k) \tag{4.15}$$

Now we are ready to calculate

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{.0099}{.02097} \approx .04721 \tag{4.16}$$

So only about 4.7% of people who test positive have the disease. It's a lot more than one in a thousand, but a lot less than 99%.

This is an application of Bayes's theorem, which says

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2)} \tag{4.17}$$

or, for $k$ mutually exclusive, exhaustive alternatives,

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{\sum_{i=1}^{k} P(B|A_i)P(A_i)} \tag{4.18}$$

## Practice Problems

2.45, 2.59, 2.63, 2.67, 2.71, 2.105 parts a & b