Caltech Department of Mathematics

Ma 3/103
 KC Border

Introduction to Probability and Statistics — Winter 2017

Lecture 1: Probability: Intuition, Examples, Formalism

Relevant textbook passages:

Pitman [28]: Sections 1.1, 1.2, first part of 1.3, pp. 1–26.

Larsen–Marx [25]: Sections 1.3, 2.1, 2.2, pp. 7–26.

1.1 Uncertainty, randomness, and probability

Karl Orff's *O Fortuna* is a musical tribute to **Fortune**. The lyrics are from an irreverent 13th century poem attributed to student monks (Wikipedia). The poem paints a picture of Fortune as "*variabilis, semper crescis aut decrescis* [changeable, ever waxing and waning]." Fortune is associated with "*Sors immanis et inanis, rota tu volubilis, status malus* [Fate—monstrous and empty, you whirling wheel, you are malevolent]."

This view of Fortune, or randomness, or uncertainty, as monstrous and subject to no law save its own malevolence is an ancient view of randomness. See. e.g.. Larsen–Marx [25, §1.4]. Indeed some have gone so far as to suggest that it was this view of luck that kept the ancient Greeks from developing the insurance and financial infrastructure needed to conquer the world. Peter Bernstein [3, p. 1] writes (emphasis mine):

> What is it that distinguishes the thousands of years of history from what we think of as modern times? The answer goes way beyond the progress of science, technology, capitalism, and democracy.
>
> [...]
>
> *The revolutionary idea that defines the boundary between modern times and the past is the mastery of risk: the notion that the future is more than a whim of the gods and that men and women are not passive before nature. Until human beings discovered a way across that boundary, the future was a mirror of the past or the murky domain of oracles and soothsayers who held a monopoly over knowledge of anticipated events.*

But traces of the ancient view remain. It is perhaps this view of randomness as chaos, anarchy, and malevolence, that led Albert Einstein (in a December 4, 1926 letter to Max Born) to insist that

> *Gott würfelt nicht mit dem Universum.*
> [God does not play dice with the universe.]

Except that is not what Einstein actually wrote. The correct quote[1] according to Born [4, pp. 129–130] is, "Jedenfalls bin ich überzeugt, daß *der* nicht würfelt." [Anyway, I am convinced that *he* does not play dice.]

One of my colleagues in applied math, [redacted] , suggested that mixing probability and data analysis in a single course was dangerous because students might "believe that things are probabilistic." (I disagree that this is dangerous. In fact I encourage you to think that the world is full of randomness.) This view was also expressed by a Ma 2b student as, "But earthquakes don't happen at random. They happen for a reason."

[1] I thank Lindsay Cleary, the HSS librarian for tracking this down for me.

Our view of luck and fortune began to change in the 17th century when Blaise Pascal (1623–1662) and Pierre de Fermat (1601?–1665) began systematic investigations into games of chance. We now understand that

Randomness is not simply anarchy.
It obeys mathematical laws.

It is these laws that we shall begin to study in this course.

1.2 Probability and its interpretations

The great mathematician Henri Poincaré[2] (1854–1912) wrote as the first sentence of the first chapter of his *Calcul des Probabilités* [29, p. 24], the following:

On ne peut guère donner une définition satisfaisante de la *Probabilité*.

[One can hardly give a satisfactory definition of *Probability*.]

Probability is our way of quantifying or measuring our uncertainty. We normalize it to be a number between 0 and 1 inclusive. The Institute has an entire course, (**HPS/Pl 122. Probability, Evidence, and Belief**) devoted to the interpretation of these numbers, but I shall briefly discuss the major views as I see them. But for a more thorough job by a professional philosopher, I recommend Alan Hájek's survey [16].

"Classical" probability as a ratio of possible cases: The earliest notion of probability is due perhaps to James Bernoulli, and set out by Pierre Simon, Marquis de Laplace [24, pp. 6–7]:

> The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all cases possible is the measure of this probability, which is thus simply a fraction whose number is the number of favorable cases and whose denominator is the number of all cases possible.

The idea that absent any reason to believe otherwise, we should treat cases as "equally possible" is known as the **Principle of Insufficient Reason** or the **Principle of Indifference**. The problem with this as a definition of probability is that it does not explain which cases are "equally possible." Presumably this means that they have the same probability, but then the definition is circular. Nevertheless the Bernoulli–Laplace notion is useful for many of the problems that we shall encounter, but it is not always obvious how to apply it.

Pitman [28]: § 1.2

Frequentist school: The frequentist school views probabilities as long-run average frequencies. Joseph Hodges and Erich Lehmann [17, pp. 4, 9–10] put it this way:

> We shall refer to experiments that are not deterministic, and thus do not always yield the same result when repeated under the same conditions, as *random experiments*. Probability theory and statistics are the branches of mathematics that have been developed to deal with random experiments.

[2] According to [20, p. 224], while testifying for the defense in the Affaire Dreyfus, "Poincaré had identified himself on the stand as the greatest living expert on probability, a tactical error which he later justified to his friends by pointing out that he was under oath." (Part of the prosecution's case was a statistical argument by Monsieur Bertillon, a handwriting expert for the Paris police, who claimed that Dreyfus had forged his own handwriting so that he could claim that an incriminating document was a forgery. Poincaré pointed out numerous problems with Bertillon's analysis.)

> […]
>
> Data … gathered from many sources over a long period of time, indicate the following *stability property of frequencies:* for sequences of sufficient length the value of [the frequency] f will be practically constant; that is, if we observed f in several such sequences, we would find it to have practically the same value in each of them. …
>
> It is essential for the stability of long-run frequencies that the conditions of the experiment be kept constant. … Actually, in reality, it is of course never possible to keep the conditions of the experiment exactly constant. There is in fact a circularity in the argument here: we consider that the conditions are *essentially* constant as long as the frequency is observed to be stable. …
>
> The stability property of frequencies … is not a consequence of logical deduction. It is quite possible to conceive of a world in which frequencies would not stabilize as the number of repetitions of the experiment become large. That frequencies actually do possess this property is an empirical or observational fact based on literally millions of observations. This fact is the experimental basis for the concept of probability …

Putative examples:

- Coin tossing: the fraction of heads in repeated tosses tends to 1/2 (or does it?)
- The fraction of times two dice total 7 tends to 1/6.
- Games of chance, such as poker, roulette, and bridge, are full of examples of that frequentists would accepts as probabilities.

There are many problems with the frequentist approach. One is the above noted circularity in the definition. As a practical matter, we often do not get enough observations to figure out long-run averages. Moreover, one of the things we shall prove in this course is that if the probability that a coin toss results in Heads is 1/2, then the probability of getting exactly n Heads in $2n$ tosses of a coin actually tends to zero as n tends to infinity. So how could we ever figure out the frequentist probability? Do we just have to settle for statements like "the probability that a coin toss results in Heads is probably about 1/2?" [The answer, I believe, is yes.] For a vicious dissection of the frequentist approach see the papers by my former colleague, Alan Hájek [14, 15].

Empirical Probability: Empirical probabilities are observed frequencies in large samples, and a conceptually close to long-run frequencies. For example:

- The probability that a child is a boy.

Pitman [28]: § 1.5

In the U.S. from 2000 through 2008, 51.2% of all live births were boys, so the probability of a child being born a boy is 0.512. (Source: U.S. Census Bureau, *Statistical Abstract of the United States, 2012*, Table 80. http://www.census.gov/compendia/statab/2012/tables/12s0080.pdf)

- Life Tables.

According to the U.S. Centers for Disease Control, National Vital Statistics Report, vol. 61, no. 3 (Sep. 24, 2012), http://www.cdc.gov/nchs/data/nvsr/nvsr61/nvsr61_03.pdf, Table 5, pp. 18–19:[3]

A U.S. white male has an 86.2% chance of surviving to age 60; and an 80.9% chance of living to age 65. Does that mean that a 60-year old white male has only a 80.9% chance of living to 65?

[3] There are two types of life tables: the cohort (or generation) life table and the period (or current) life table. The cohort life table presents the mortality experience of a particular birth cohort—all persons born in a particular year from the moment of birth through consecutive ages in successive calendar years. The drawback of a cohort table is doesn't lend itself to projecting the future mortality of those currently alive. The period life table tries to circumvent this problem by looking at a particular reference year, and finding the death rate for each age in that year. (What fraction of those born in that year, died in their first year; what fraction of one year olds in that year died before age two, etc.) It then calculates what would happen to a cohort if the death rate at each age for the cohort is the same as the death rate for that age in the reference year. The table in this report is a period life table.

No. Since he has already lived to 60, his chance of making to 65 is actually $80.9/86.2 = 93.9\%$. This is an example of **conditional probability** that we shall discuss in just a bit.

[You might ask, why did I look at the tables for white males? When I was a 60-year old white male, I had to decide whether to renew my term life insurance policy.]

Physical Probability and Initial Conditions: In this view, the probability of an event is derived from an analysis of the laws of physics. For example, consider coin tossing. We know the physics of rotating and falling objects, so the only uncertainty stems from not observing the initial conditions.

Example: Coin tossing:

- Karl Menger [27] provides a simple model of coin tossing in which the height h from which the coin was dropped and its angular velocity ω determined whether it turns up as Heads or Tails. The key point is the set of initial conditions (h, ω) contains an equal area of conditions that lead to Heads as Tails.

Here are the initial conditions that lead to hitting on edge after k half-turns.

$$h = c\frac{k^2\pi^2}{\omega^2} + 1, \qquad k = 1, 2, \dots$$

where the coin has radius 1, and c depends on units and the acceleration of gravity. These loci are graphed for various k in Figure 1. The regions between these curves alternately produce Heads and Tails. See Figure 1.1.

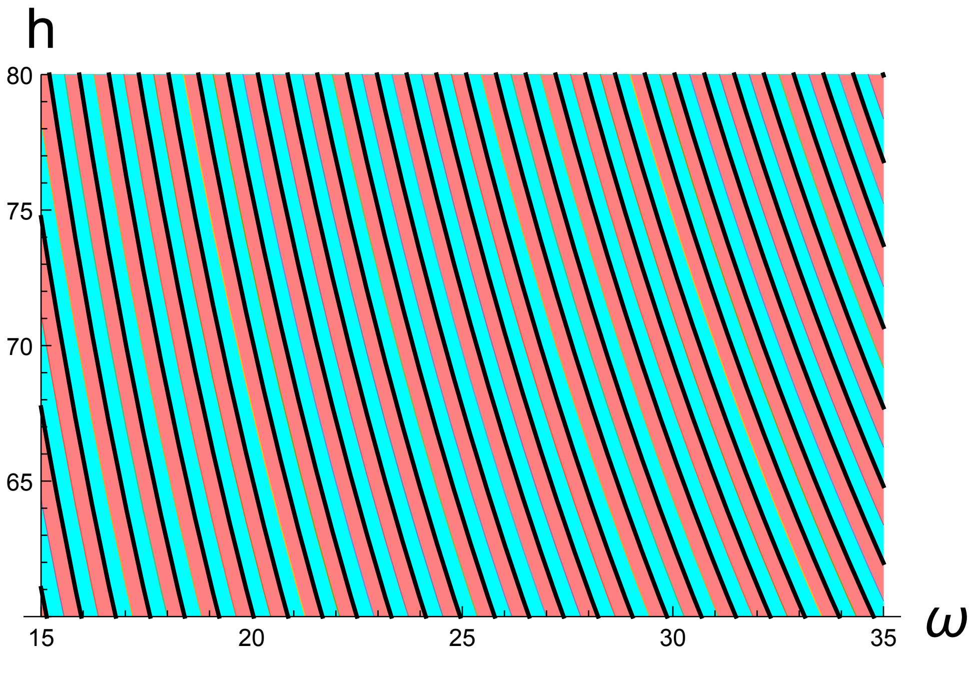


Figure 1.1. Initial conditions (ω, h) that land vertically in the Menger model are indicated by the thin black lines. Shaded areas in between these curves alternate in producing Head and Tails.

- A more sophisticated model of the physics of tossing and catching a coin, due to Persi Diaconis, Susan Holmes, and Richard Montgomery [9] takes into account wobbling and precession, and a calibrated version of their model suggests that the probability a coin comes up in the same position it started is about 51%!

That is why your first assignment will be to toss coins, but more on that later.

• Andrzej Lasota and Michael Mackey's book [26], *Chaos, Fractals, and Noise* (1994), formerly known as *Probabilistic Properties of Deterministic Systems* (1985),[4] make a persuasive case that *chaotic* dynamics are best described in terms of probability.

• Some phenomena at very small scales appear to be truly random and can only be described in terms of quantum mechanical randomness. The impossibility of predicting through which slit a photon will pass is one of them. But macroscopic random phenomena cannot be explained on the basis of quantum mechanical randomness.

Pitman [28]: pp. 16–17

Subjective Probability: The subjective school of probability treats probabilities as statements about the **degree of belief** of a decision maker. The label "Bayesian" is commonly attached to the subjectivist school. Bruno de Finetti, a first rate mathematician, takes the extreme view [11, p. x] that

"in order to avoid becoming involved in a philosophical controversy,"
we should simply agree that
"Probability does not exist."

By that he means it has no independent existence outside of our minds. De Finetti takes this point of view seriously enough to invent a new term, *prevision*, to replace probability.

Examples:

• Horse racing. There is an old saying that it takes a difference of opinion to make a horse race. Different bettors have different beliefs about which horse will win. These beliefs may be based on a variety of evidence, but it is unlikely to come from a physical model of the horses.

• Weather forecasting is partially subjective: This is why "skill scoring rules" were invented. The practice of expressing weather forecasts in terms of rough probabilities was initiated in Western Australia by W. E. Cooke in 1905 [6]. Interestingly, his idea was criticized by E. B. Garriot [12] of the U.S. because "the bewildering complication of uncertainties it involves would confuse even the patient interpolator" and "our public insist upon having our forecasts expressed concisely and in unequivocal terms."

• One might question why purely subjective degrees of belief would obey the rules of probability that we are about to lay out. An answer was given by de Finetti [10]. He showed that if beliefs are not subject to the laws of probability, then they are *incoherent*. That is, if your subjective beliefs are not probabilistic, then you can be forced to lose money in a gambling situation. De Finetti the deduces many of the properties of probability (such as additivity and monotonicity) from the principle of coherence.

Probability as a branch of logic: John Maynard Keynes in his 1921 *Treatise on Probability* [21] argued that probability was the branch of logic concerned with the plausibility (as opposed to truth) of propositions. His ideas influenced a number of others, including the physicist R. T. Cox [7, 8] and through him, the physicist Edwin T. Jaynes. The late Jaynes may be the most outspoken proponent of this view. His posthumous treatise, *Probability Theory: The Logic of Science* makes for some interesting reading. He sets out three "desiderata" (he eschews the term axiom, on the grounds that "they do not assert anything is 'true' but only state what appear to be desirable goals." [19, p. 16]) for a robot to do scientific plausible reasoning. They are (i) the degrees of plausibility are represented by real numbers, (ii) they exhibit qualitative correspondence with common sense, and (iii) the robot reasons consistently [19, pp. 17–19]. Naturally, there is a bit more to it than these assertions.

Where does Jeffreys enter into this?

[4] The new title is a lot sexier and more marketable.

Jaynes considers himself to be an “objective Bayesian,” and points out some similarities between his approach and de Finetti’s notion of coherence. But he rejects using coherence arguments on three grounds.

1. The first is aesthetic: “it seems to us inelegant to base the principles of logic on such a vulgar thing as expectation of profit.” [19, p. 655]

2. The second is strategic: “If probabilities are defined in terms of betting preferences,” it “belongs to the field of psychology.” Moreover, his robot does not have preferences over gambles. [19, p. 655] Jaynes takes the position that there are objectively correct beliefs.

3. The third is that de Finetti did not articulate Cox’s principle of consistency. [19, p. 656] Consistency is related to Bayes’ Law, but I will not go into that here.

- Laplace’s “Principle of Insufficient Reason” is often invoked to assign equal probabilities to events, and it is sometimes regarded as a form of subjective belief. It can also be viewed as a requirement of invariance under certain kinds of transformations.

- The maximum entropy principle is a more sophisticated version of the principle of insufficient reason for assigning probabilities. See, e.g., Jaynes [18] for a persuasive argument in favor of the maximum entropy principle. It is usually not considered to be subjective, especially by its most ardent practitioners. They would argue that probability can and **must be deduced on logical grounds**.

1.2.1 An observation on random sequence generation

But before we go further, indulge me, and let me make the following outrageous claim.

The following statement represents the opinion of the author, and does not necessarily reflect that of the California Institute of Technology or its Mathematics Department.

The digits of π are as random as coin tossing is.

By this I mean that it you cannot predict how a sequence of digits of π will continue, unless you know the starting point—just as you cannot predict how a coin land without knowing its initial position, momentum, and angular momentum. For instance,

- What digit follows the following sequence:

$$3 \quad 1 \quad 4 \quad 1 \quad 5 \quad \ldots$$

I hope most of you would say 9, because 9 is the fifth digit after the decimal point in the decimal expansion of π. But that is not necessarily the case. Let’s see why, by examining the first billion (thousand million, for you Anglophiles) digits of π. I asked Mathematica 10 to compute π to a billion places, and it did so in 41 and a half minutes on my early 2009 Mac Pro. (By the way, Mathematica 8 would only compute about 200 million digits before crashing.) I then asked it to count the number of occurrences of each digit. This took another 16 minutes plus change. Here is a table of the digit counts:

digit	number	deviation
0	99,997,333	-2,667
1	100,002,411	2,411
2	99,986,912	-13,088
3	100,011,958	11,958
4	99,998,885	-1,115
5	100,010,387	10,387
6	99,996,061	-3,939
7	100,001,839	1,839
8	100,000,272	272
9	99,993,942	-6,058
	1,000,000,000	0

If the digits were evenly distributed you would expect about 100 million of each. The deviation from 100 million is listed in the last column of the table. You can see that we are very close. (The largest deviation is 0.013%.) We can treat the list of frequencies as a vector in $\boldsymbol{R}^{10}$ and compute its distance from the theoretical vector of 100 million in each component. We shall learn later on about the marvelous chi-square test for uniformity, and see that if the digits were randomly generated, the distance from perfect uniformity due simply to randomness would be at least this great about 84% of the time. The first billion digits of π pass this simple test for randomness.

But now let's get back to the question of what comes after 31415? By my count, the sequence 31415 occurs 10,010 times in the first billion and one digits of π.[5] There are slightly less than a billion starting points for sequences of five consecutive digits in a billion and one digits. There are 100,000 different 5-digit sequences. If each were equally likely, there would be about 10,000 of each in a billion, so 10,010 is uncannily close, and each digit should occur about 1001 times. (Note that two sequences of 31415 cannot overlap, so each occurrence wipes out 4 more starting points. But that effect is negligible. The actual distribution is somewhat complicated.) Here are the number of occurrences 31415x:

string	occurrences	deviation
314150	1015	14
314151	1043	42
314152	946	-55
314153	958	-43
314154	1018	17
314155	978	-23
314156	1012	11
314157	1037	36
314158	1000	-1
314159	1003	2

There are 100 different digit-pairs that can follow 31415. With 10,010 such pairs we would expect about 100 occurrences of each. See Table 1.1 for the results. A natural question is whether the deviations observed are large or small. We shall answer this question in Lecture 23, where we derive what is called the χ^2 test. But the answer is that these deviations are small, and are very consistent with the hypothesis that the digits of π are a random sequence.

There are other tests for randomness that we can perform. For instance, we could look at each digit string of length n and compare its frequency to what we would expect if they were evenly distributed. I have done this for $n = 1, \dots, 6$. I am not going to list all the frequencies

[5] When I asked Mathematica 10 to write out the billon digits it actually wrote out about a billion and forty past the decimal point. I don't know why. So I kept the initial digit 3, threw out the decimal point and took the next billion digits. It took Mathematica 10 fifteen minutes to write the file to disk. But it took my Perl script a mere 6 seconds to read the file and count the occurrences of 31415.

string	occurrences	deviation
3141500	103	3
3141501	100	0
3141502	95	-5
3141503	87	-13
3141504	116	16
3141505	108	8
3141506	101	1
3141507	102	2
3141508	107	7
3141509	96	-4
3141510	102	2
3141511	104	4
3141512	106	6
3141513	99	-1
3141514	103	3
3141515	104	4
3141516	114	14
3141517	113	13
3141518	101	1
3141519	97	-3
3141520	61	-39
3141521	84	-16
3141522	86	-14
3141523	99	-1
3141524	101	1
3141525	115	15
3141526	98	-2
3141527	105	5
3141528	107	7
3141529	90	-10
3141530	95	-5
3141531	92	-8
3141532	89	-11
3141533	93	-7
3141534	95	-5
3141535	105	5
3141536	84	-16
3141537	86	-14
3141538	105	5
3141539	114	14
3141540	105	5
3141541	105	5
3141542	89	-11
3141543	96	-4
3141544	131	31
3141545	106	6
3141546	87	-13
3141547	99	-1
3141548	99	-1
3141549	101	1
3141550	105	5
3141551	101	1
3141552	86	-14
3141553	87	-13
3141554	105	5
3141555	99	-1
3141556	104	4
3141557	107	7
3141558	106	6
3141559	78	-22
3141560	99	-1
3141561	97	-3
3141562	100	0
3141563	98	-2
3141564	107	7
3141565	107	7
3141566	105	5
3141567	95	-5
3141568	95	-5
3141569	109	9
3141570	105	5
3141571	99	-1
3141572	99	-1
3141573	102	2
3141574	113	13
3141575	106	6
3141576	97	-3
3141577	115	15
3141578	95	-5
3141579	106	6
3141580	103	3
3141581	89	-11
3141582	89	-11
3141583	94	-6
3141584	103	3
3141585	117	17
3141586	97	-3
3141587	90	-10
3141588	104	4
3141589	114	14
3141590	98	-2
3141591	99	-1
3141592	100	0
3141593	97	-3
3141594	98	-2
3141595	93	-7
3141596	87	-13
3141597	102	2
3141598	110	10
3141599	119	19

Table 1.1. Occurrences of 31414xy.

(think of how much paper that would take), but here are what are called the p-values (rounded to nearest hundredth) for the chi-square test. The p-value is a number between 0 and 1 that tells you the "goodness of fit" of the digits to the model that they are randomly distributed.

string length	p-value
1	0.84
2	0.92
3	0.99
4	0.86
5	1.00
6	1.00

The point of all this is that even though the sequence of digits in the decimal expansion of π is completely deterministic, it still makes a good random number generator, in the sense that if I do not tell you where I start in the sequence, you cannot tell what is coming next—the next digit behaves as if it were random. In this sense, the digits of π are as random as a sequence of coin tosses.

But I only checked the first billion digits of π. Will this result hold up for the first $10^1$00 digits? Will it hold up as an infinite limit? The answer is that we don't know yet. A number with the property that each sequence of n digits is equally likely in the long run is called **normal**. See the very readable paper by Bailey and Borwein [2] for a recent survey of what we know about π.

(Actually, the digits of π are a terrible random sequence generator, because computing the sequence of digits of π is very time-consuming. It takes 41 minutes to generate a billion digits of π, but only 20 seconds to generate a billion pseudorandom digits using Mathematica's built-in `RandomInteger` function. If you are interested in algorithms to generate the digits of π you might want to start with this nice paper by Borwein, Borwein, and Bailey [5]. At the time it was an impressive accomplishment to generate the first billion digits of π. You may also want to visit the GMP page and perhaps download Hanhong Xue's C program, which uses $8n$ bytes of memory to compute n digits.)

The idea that completely deterministic algorithms can mimic random processes is at the heart of "pseudorandom" number generation. My view is that good pseudorandom numbers are as random as coin tosses. But the great John von Neumann (1903–1957) quipped, "Anyone who attempts to generate random numbers by deterministic means is, of course, living in a state of sin." (brainyquote.com)

1.2.2 The advantage of a formal model

My own view leans toward's de Finetti's, as I really want to avoid becoming embroiled in metaphysical controversies, so I am willing to just treat probability as a mathematical construct. But I am also impressed by all that empirical evidence that Lehmann and Hodges cite, and others cite. It turns out that real physical phenomena are well modeled by the mathematical construct. Perhaps we should adopt the approach of Robert Ash [1, p. 14], who suggests

> "[I]n probability theory we are faced with situations in which our intuition or some physical experiments we have carried out suggest certain results. Intuition and experience lead us to an *assignment* of probabilities to events. As far as the mathematics is concerned, any assignment of probabilities will do, subject to the rules of mathematical consistency. However, our hope is to develop mathematical results that, when interpreted and related to physical experience, will help to make precise such notions as "the ratio of the number of heads to the total number of

> observations in a very large sample of independent tosses of an unbiased coin is very likely to be close to 1/2."
>
> We emphasize that the insights gained by the early workers in probability are not to be discarded, but instead cast in a more precise form.

We shall take a "formal approach" to probability. That is, we shall introduce "primitive terms" and be careful with our reasoning. The advantage of this is that you don't have to grok the interpretation.[6]

1.3 Administrative Details

Make sure to update this every year.

Now that everyone has had a chance to arrive and settle in, it is a good time to go over administrative details. Everything can be found on the course website at

http://www.math.caltech.edu/~2016-17/2term/ma003/

- Ma 3/103
- Kim Border, 205 Baxter, x4218, kcb@caltech.edu
- Office Hours: Fridays, 1:30–3:00 p.m.
- Lead TA: William Chan, wcchan@caltech.edu
- **All questions regarding grading, extensions, late work, etc., should be directed to the Lead TA.**
- Course administrator: Elsa Echegaray, 253 Sloan, x 4203, elsa.echegaray@caltech.edu
- **All requests for changes of section should be addressed to the Course Administrator.**
- Collaboration: See the course web page for details, but collaboration is encouraged on the homework, and not allowed on exams.
- Assignment 0: Coin tossing

1.4 A formal approach to probability

John von Neumann once said,

> "There's no sense in being precise when you don't even know what you're talking about." [a]
>
> [a] Quoted by, among others, professional gambler Barry Greenstein in his autobiography *Ace on the River* [13, p. 157].

Yet even though I am not sure about what the correct interpretation of probability is, I am going to give a precise mathematical framework for working with it. I can at least understand the mathematical framework. Or maybe not, for as von Neumann [31, p. 208] also said,

> "In mathematics you don't understand things. You just get used to them."

[6] Perhaps that puts me among those to whom Jaynes was referring when he wrote that "those who lay the greatest stress on mathematical rigor are just the ones who, lacking a sure sense of the real world, tie their arguments to unrealistic premises and thus destroy their relevance." [19, p. xxvii], a point he attributes to Harold Jeffreys.

The dominant contemporary model of probability was developed in the early 20th century by a number of mostly French, Italian, and Russian mathematicians and was finally codified by Andrey Nikolaevich Kolmogorov (Андрей Николаевич Колмогоров 1903–1987) in his slim *Grundbegriffe Der Wahrscheinlichkeitsrechnung* [22, 23] in 1933. Glenn Shafer and Vladimir Vovk [30] give an account of the history of probability theory preceding Kolmogorov and how he synthesized the contributions of his predecessors.

1.4.1 Experiments and sample spaces

A **random experiment** is something we observe that generates an **outcome** that will not be known or precisely predictable in advance. (For the time being I will leave the term random as a primitive.) The set of all possible outcomes is called the **sample space** or the **outcome space** of the experiment. The sample space is sometimes denoted S, or sometimes (as in Pitman) Ω. In these notes I will try to use S to refer to a finite or countably infinite sample space, and I will try to use Ω to denote an uncountable sample space. The mathematics of uncountable sample spaces is more complicated.

Pitman [28]: § 1.3, p. 19

1.4.1 Example (Coin tossing) Consider the results of tossing a coin. The outcome of the toss could be either Heads, denoted H or tails, T, so we could take as our sample space the set:

$$S = \{H, \quad T\}.$$

Or perhaps we are willing to accept the possibility that the coin could land on edge, E. Then the sample space would be

$$S = \{H, \quad T, \quad E\}.$$

Or I might wish to include the possibility that my crazed Labrador Retriever might see this as an opportunity to demonstrate her talent for retrieving flying objects and snatch the coin out of the air, outcome L, so maybe the sample space should be

$$S = \{H, \quad T, \quad E, \quad L\}.$$

Or maybe the FBI would confiscate the coin in a counterfeiting investigation. (This is rather unlikely, as the Secret Service investigates counterfeiting.)

The point is, the sample space is a *mathematical model* chosen by you the analyst, to represent the outcomes worthy of consideration. And for most uses, that means the sample space for a coin toss has two points,

$$S = \{H, \quad T\}.$$

□

1.4.2 Example (Repeated coin tossing) Now consider the results of tossing a coin three times. The outcome of each toss could be either Heads, denoted H or tails, T. (We won't consider Labradors, or coins on edge, or intervention by aliens or the FBI.) With three tosses there are eight possible outcomes to the experiment, so we take as our sample space the set:

$$S = \{HHH, \quad HHT, \quad HTH, \quad HTT, \quad THH, \quad THT, \quad TTH, \quad TTT\}.$$

Clearly, if we toss a coin n times the sample space will contain 2^n outcomes. □

1.4.3 Example (Repeated coin tossing with a stopping rule) In this experiment, we toss a coin repeatedly until it comes up heads. The sample space for this experiment is quite large. In fact it is infinite, but denumerably infinite. It includes every finite sequence of n Tails followed by a single Head, for $n = 0, 1, 2, \ldots$, and it includes the infinite sequence of only Tails.

$$S = \{H, \quad TH, \quad TTH, \quad \cdots, \quad \underbrace{TT\cdots T}_{n}H, \quad \cdots, \quad \overline{TTTT\cdots}\}.$$

□

1.4.2 Events

The next element in our formal approach is the notion of an **event**. An event is simply an "observable" subset of the sample space. I use the word observable here as a primitive, and I will come back to that later. 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).

At this point, let me digress and discuss some set-theoretic notation. Many probabilists, Pitman [28] included, use the symbol EF to denote these intersection of the sets E and F, so I will do likewise in the notes. Also $E \setminus F$ denotes the set of elements of E that do not belong to F, and E^{c} denotes the complement of E.

Pitman [28]: § 1.3, pp. 19–21
Larsen–Marx [25]: §2.2, pp. 18–27

The set of all 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*. They correspond to union, intersection, and complementation.

1.4.4 Example (Coin tossing events) For the sample space in Example 1.4.3, Coin Tossing until Heads, let $\mathcal{E}$ be the set of all subsets of S. We can consider events such as

$$E = \text{the first Head occurs on an odd-numbered toss} = \{H, TTH, TTTTH, \cdots\}$$
$$F = \text{the first Head occurs on an even-numbered toss} = \{TH, TTTH, TTTTTH, \cdots\}$$
$$G = \text{Heads never occur} = \{\overline{TTTT\cdots}\}.$$

Note that $E \cup F \neq S$, but $(E \cup F)^{\mathrm{c}} = G$, and $EF = \varnothing$. □

Aside: The notion of the set of events as a set of subsets of S may seem unwieldy. You may be used to thinking of sets of points, not sets of sets. But you have used such collections for years. Think of the set of intervals on a line, or the set of triangles in a plane. These are all sets of sets.

In most real applications I can think of, for each outcome $s \in S$, the singleton set $\{s\}$ is an event, that is, $\{s\} \in \mathcal{E}$. [Note that $s \notin \mathcal{E}$!]

1.4.3 Probability measures

Larsen–Marx [25]: §2.3, pp. 27–32
Pitman [28]: §1.3, pp. 19–32

A **probability measure** or **probability distribution** (as in Pitman) 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 $P(S) = 1$.

Nonnegativity For each event E, we have $P(E) \geqslant 0$.

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 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.

To cut down on the number of delimiters in our notation, when a set is delimited with braces or with statistician's notation, we may omit the parentheses surrounding it and simply write something like $P\,(f = 1)$ 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.

What is statistician's notation?

1.4.5 Definition *For any event E, let $|E|$ denote the number of elements of E.*

The next result is Laplace's assertion that probability is the ration of "favorable" cases to the number of "equally possible" cases.

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

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

1.4.7 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 simply $S = \{H, T\}$ and $\mathcal{E}$ is all four subsets of S, $\mathcal{E} = \big\{\varnothing, S, \{H\}, \{T\}\big\}$, then

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

so additivity implies

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

so $P\{H\} = P\{T\}$ implies

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

□

1.5 Analogies

The additivity property of probability makes it analogous to many other kinds of measurements, such as length, area, or mass. Indeed sometimes these measurements (when normalized) are actually the same as probabilities.

For instance, with a well balanced spinner with a very fine pointer, the outcome essentially gives an angle, which corresponds to a point in the real interval $[0, 2\pi)$. For a good spinner the probability of coming to rest in any sector is proportional to the angle subtended by the sector, which is just the length of the corresponding interval. The total length of two disjoint segments is just the sum of their lengths. This is the additivity property.

Likewise, if I throw a very fine dart at a dart board, and if my aim is sufficiently poor that any part of the dart board the probability of an region of the board is equally likely to be hit, then the area of the region is proportional to its probability. If my aim is better, so that regions near the center of the board are more likely, then we may need to weight the area by some *probability density*. But again the probability of the unions of two disjoint regions should be their sum.

We shall explicitly like probability to mass in Lecture 5, when we discuss the expectation of a random variable in terms of a balance beam. The total mass of two distinct objects is just the sum of their masses.

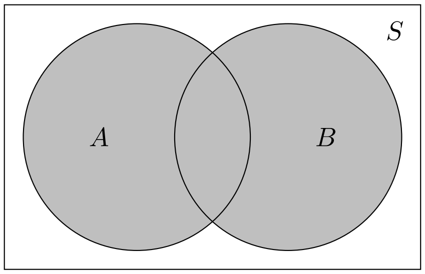
1.6 Appendix: Review of Set Operations

A quick review of set theory can be found in Ash [1], section 1.2. We shall follow Pitman [28], and use the notation AB rather than $A \cap B$ to denote the intersection of A and B.

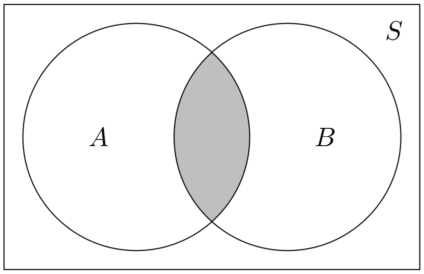
Pitman [28]: pp. 19–20

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

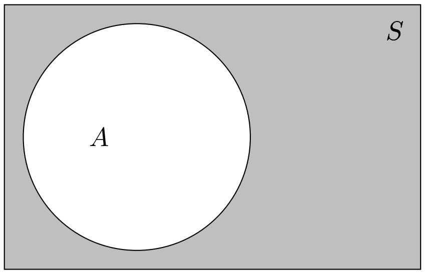
$A \cup B$:



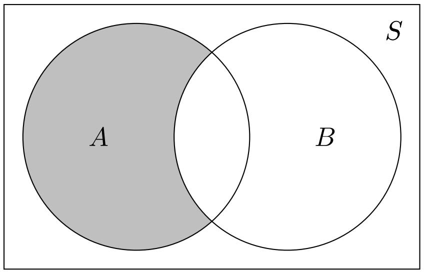
$A \cap B$ or AB :



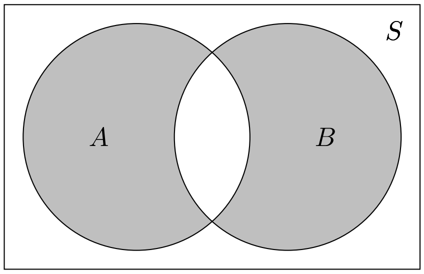
A^{c} :



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



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

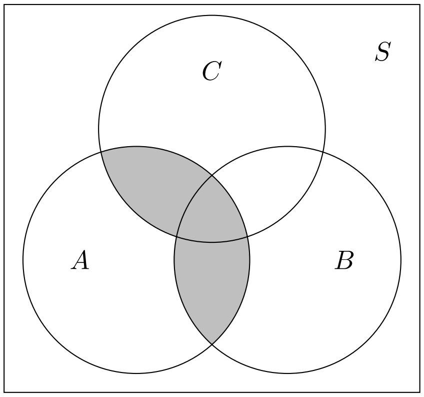


1.6.1 Definition *For any set E, let $|E|$ denote the* ***cardinality****, or number of elements, of E. We use this notation primarily with finite sets.*

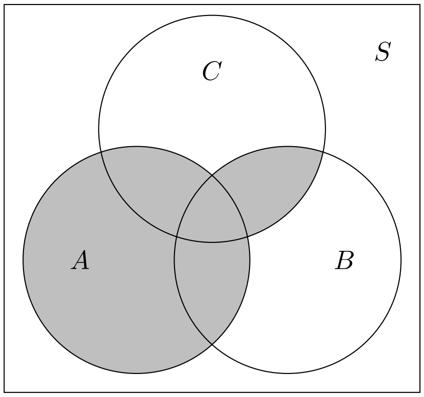
1.6.2 Definition *A* ***partition*** *of a set E is a collection $\mathcal{A}$ of subsets of E such that every point in E belongs to exactly one of the sets in $\mathcal{A}$*

Here are some useful identities.

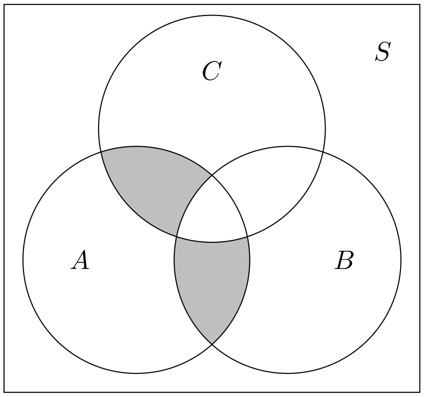
$A(B \cup C) = (AB) \cup (AC)$:



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



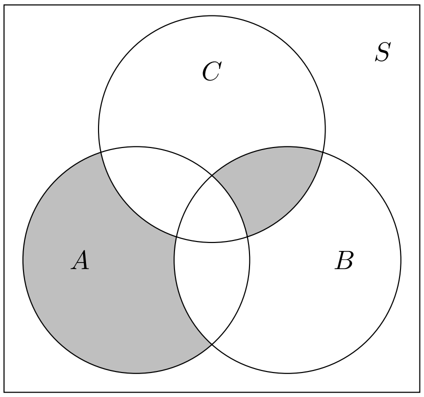
$A(B \triangle C) = (AB) \triangle (AC)$:



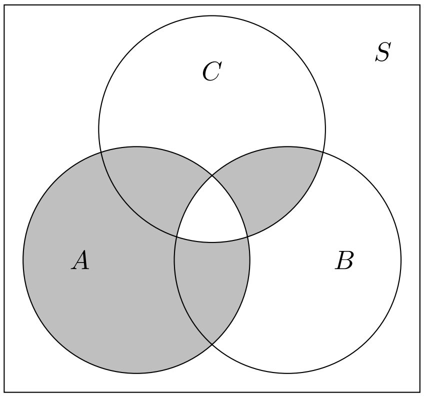
Note that

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

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



$A \triangle (BC)$



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. (Of course, math is one of the options that does not require this course.)

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] D. H. Bailey and J. Borwein. 2014. Pi day is upon us again and we still do not know if pi is normal. *American Mathematical Monthly* 121(3):191–206. http://www.jstor.org/stable/10.4169/amer.math.monthly.121.03.191

[3] P. L. Bernstein. 1996. *Against the gods: The remarkable story of risk*. New York: Wiley.

[4] M. Born, ed. 1969. *Briefwechsel 1916–1955 [von] Albert Einstein [und] Hedwig und Max Born*. München: Nymphenberger Verlagshandlung.

[5] J. M. Borwein, P. B. Borwein, and D. H. Bailey. 1989. Ramanujan, modular equations, and approximations to pi or how to compute one billion digits of pi. *American Mathematical Monthly* 96(3):201–219. http://www.jstor.org/stable/2325206

[6] W. E. Cooke. 1906. Forecasts and verifications in Western Australia. *Monthly Weather Review* 34(1):23–24. http://docs.lib.noaa.gov/rescue/mwr/034/mwr-034-01-0023.pdf

[7] R. T. Cox. 1946. Probability, frequency, and reasonable expectation. *American Journal of Physics* 14(1):1–13. DOI: 10.1119/1.1990764

[8] ——— . 1961. *The algebra of probable inference.* Baltimore, Maryland: Johns Hopkin University Press.

[9] P. Diaconis, S. Holmes, and R. Montgomery. 2007. Dynamical bias in the coin toss. *SIAM Review* 49(2):211–235. DOI: 10.1137/S0036144504446436

[10] B. de Finetti. 1937. La prevision: Ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré* 7(1):1–68. Translated as "Foresight: Its Logical Laws, Its Subjective Sources" in *Studies in Subjective Probability*, H. E. Kyburg, Jr. and H. E. Smokler, eds., Robert E. Krieger Publishing, Huntington, New York, 1980, pages 53–118. http://www.numdam.org/item?id=AIHP_1937__7_1_1_0

[11] ——— . 1974. *Theory of probability*, volume 1. London: Wiley.

[12] E. B. Garriott. 1906. Note by Prof. E. B. Garriott. *Monthly Weather Review* 34(1):24. http://docs.lib.noaa.gov/rescue/mwr/034/mwr-034-01-0023.pdf

[13] B. Greenstein. 2005. *Ace on the river: An advanced poker guide.* Fort Collins, Colorado: Last Knight Publishing Company.

[14] A. Hájek. 1996. "Mises redux"—redux: Fifteen arguments against finite frequentism. *Erkenntnis* 45(2–3):209–227. DOI: 10.1007/BF00276791

[15] ——— . 2009. Fifteen arguments against hypothetical frequentism. *Erkenntnis* 70(2):211–235. DOI: 10.1007/s10670-009-9154-1

[16] ——— . 2012. Interpretations of probability. In E. N. Zalta, ed., *The Stanford Encyclopedia of Philosophy*. Winter 2012 edition ed. http://plato.stanford.edu/archives/win2012/entries/probability-interpret/

[17] J. L. Hodges, Jr. and E. L. Lehmann. 2005. *Basic concepts of probability and statistics*, 2d. ed. Number 48 in Classics in Applied Mathematics. Philadelphia: SIAM.

[18] E. T. Jaynes. 1968. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics* 4(3):227–241. DOI: 10.1109/TSSC.1968.300117

[19] ——— . 2003. *Probability theory: The language of science.* Cambridge: Cambridge University Press.

[20] A. Kaplan. 1998. *The conduct of inquiry: Methodology for behavioral science*, 2d. ed. New Brunswick, New Jersey: Transaction Publishers.

[21] J. M. Keynes. 1921. *A treatise on probability.* London: Macmillan and Co.

[22] A. N. Kolmogorov. 1933. *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Berlin: Springer. A Russian translation by G. M. Bavli, appeared in 1936, with a second edition, slightly expanded by Kolmogorov with the assistance of A. N. Shiryaev, in 1974, and a third edition in 1998. An English translation by N. Morrison appeared under the title *Foundations of the Theory of Probability* (Chelsea, New York) in 1950, with a second edition in 1956.

[23] ——— . 1956. *Foundations of the theory of probability.* New York: Chelsea. Translated from the 1933 German edition by N. Morrison.

[24] P. S. Laplace. 1995. *A philosophical essay on probability*. New York: Dover Publications. The Dover edition, first published in 1995, is an unaltered and unabridged republication of the work originally published by John Wiley and Sons in 1902, and previously reprinted by Dover in 1952. The English translation by F. W. Truscott and F. L. Emory, is from the the sixth French edition of the work titled *Essai philosophique sur les probabilités*, published by Gauthier–Villars (Paris) as part of the 15-volume series of Laplace's collected works. The original French edition was published in 1814 (Paris).

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

[26] A. Lasota and M. C. Mackey. 1994. *Chaos, fractals, and noise*, 2d. ed. New York: Springer–Verlag. Second edition of *Probabilistic Properties of Deterministic Systems*, published by Cambridge University Press, 1985.

[27] K. Menger. 1954. Tossing a coin. *American Mathematical Monthly* 61(9):634–636. http://www.jstor.org/stable/2307682

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

[29] H. Poincaré. 1912. *Calcul des probabilités*, 2d. ed. Paris: Gauthier-Villars.

[30] G. Shafer and V. Vovk. 2006. The sources of Kolmogorov's *Grundbegriffe*. *Statistical Science* 21(1):70–98. DOI: 10.1214/088342305000000467

[31] G. Zukav. 1979. *The dancing Wu Li masters: An overview of the new physics*. New York: Morrow.