MFCS

Probablility

Klaus Sutner

Carnegie Mellon University

Fall 2022

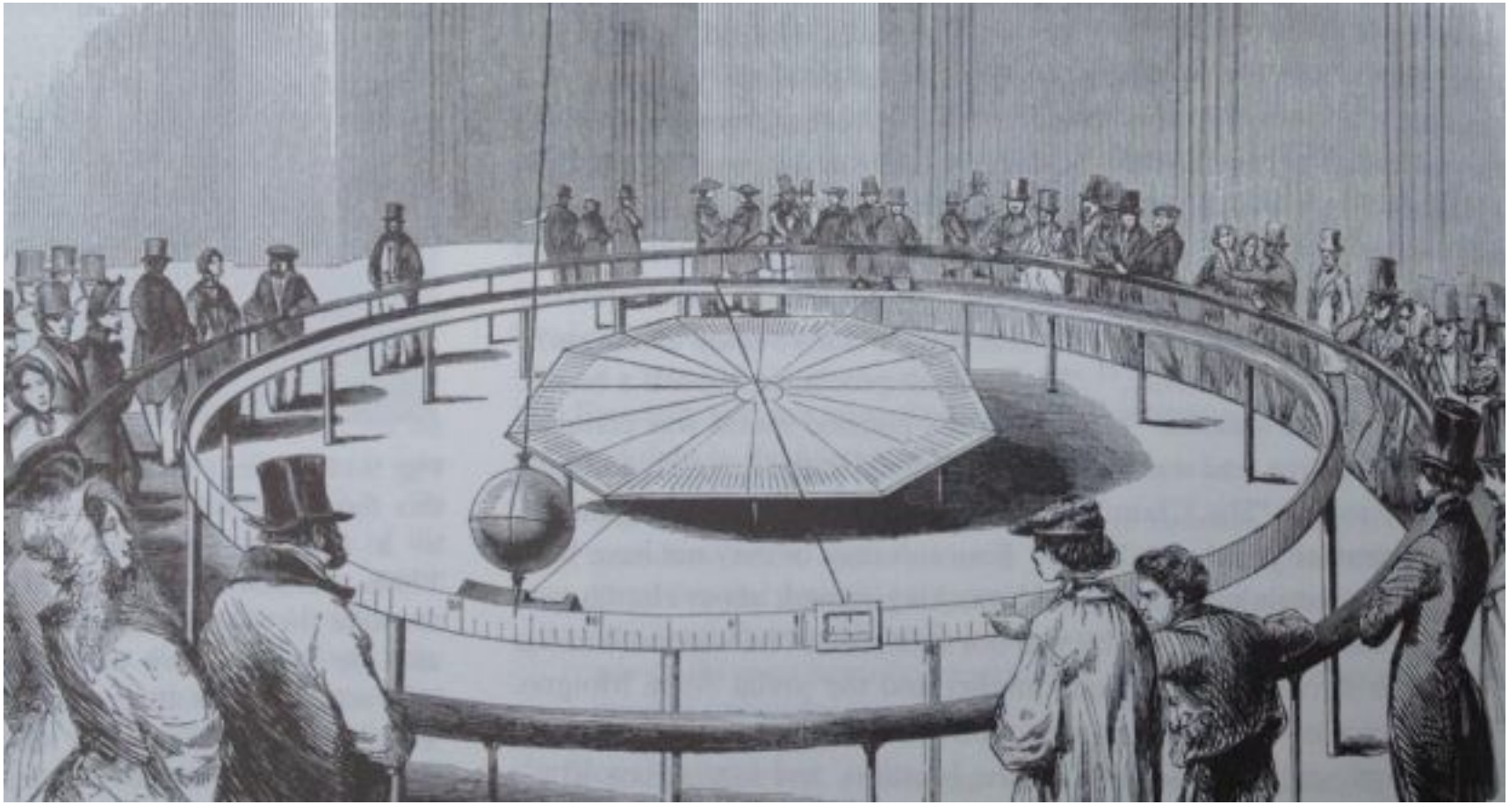
The Roots: Gambling

Probability theory is actually one of the older theories in math, inspired by the pressing need to develop sound gambling strategies.

The first major contribution appears to be by Gerolamo Cardano, the *Liber de Ludo Aleae*, around 1564.

Later, luminaries such as Galileo, Fermat and Pascal developed matters further.

Foucault



Dice

Randomness

What is the **huge** difference between Foucault's pendulum and rolling dice?

- Randomness:
  One cannot predict the outcome of the next roll, never.
- Determinism:
  The pendulum behaves in the exact same way, always.

Another way to think about this is computational incompressibility: there is no shortcut to computing the result for the dice, for the pendulum it is trivial (modulo knowledge of physics).

Hold It . . .

So far, all the concepts we talked about (sets, functions, relations, natural numbers, integers, rationals, reals, counting, ordered fields, vector spaces, cardinality) are founded entirely within mathematics.

To be sure, many of them are motivated by observations of the RealWorld$^{\text{TM}}$, but the definitions require no reference to that world, none whatsoever.

This is patently false for randomness.

The very concept of randomness is motivated by physics, it is quite difficult to come up with a purely mathematical definition.

How Random is It?

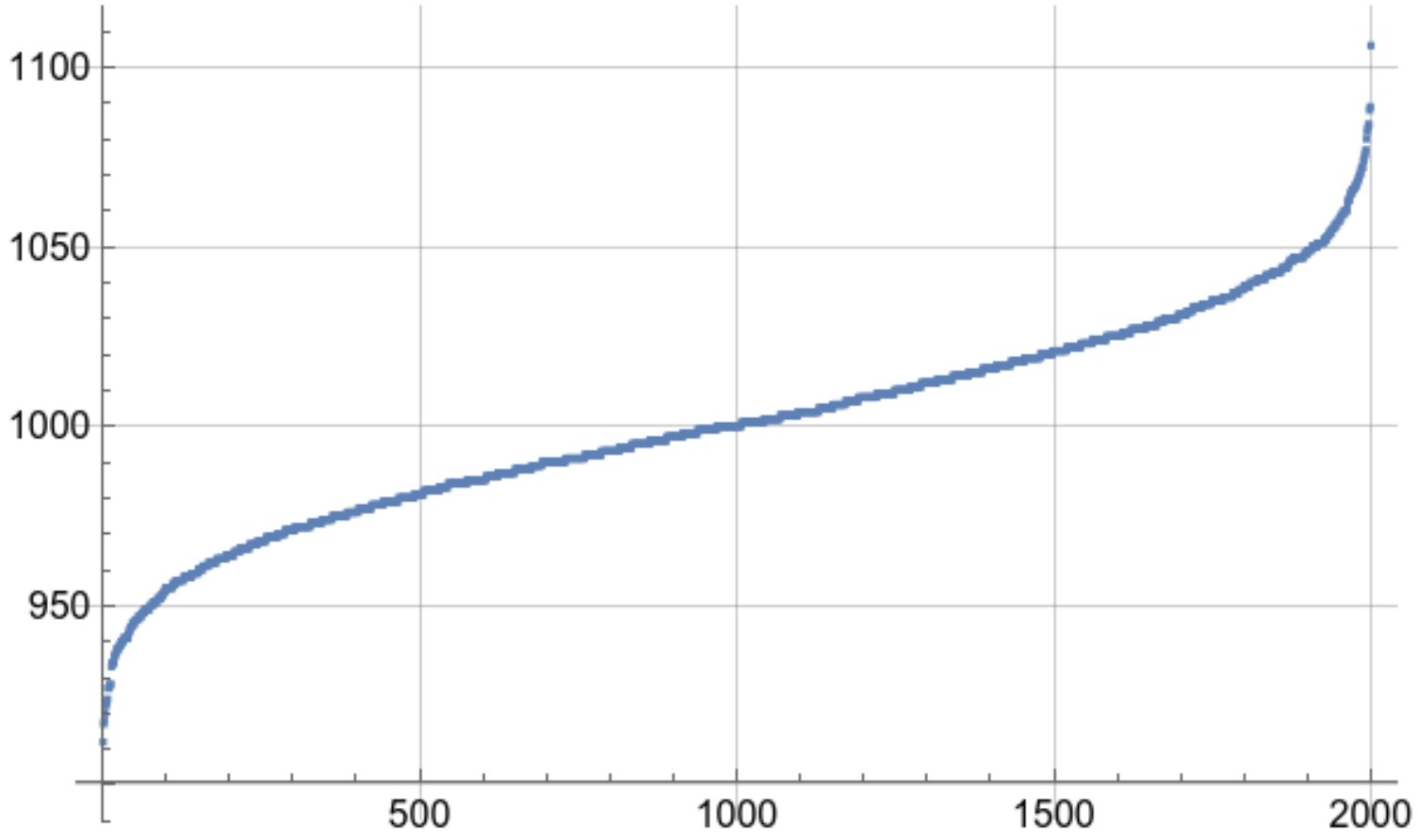
Anyone who has ever rolled dice or flipped coins knows from experience that outcomes are indeed unpredictable.

And yet, there is method to the madness: rolling a die 6000 times, one would expect the number of 6s to be somewhere around 1000. In fact, one might guess that the count will wind up in the interval $[950, 1050]$.

With more effort, one could even calculate that, if we expand the interval to $[913, 1087]$, then the likelihood of hitting it is 0.997.

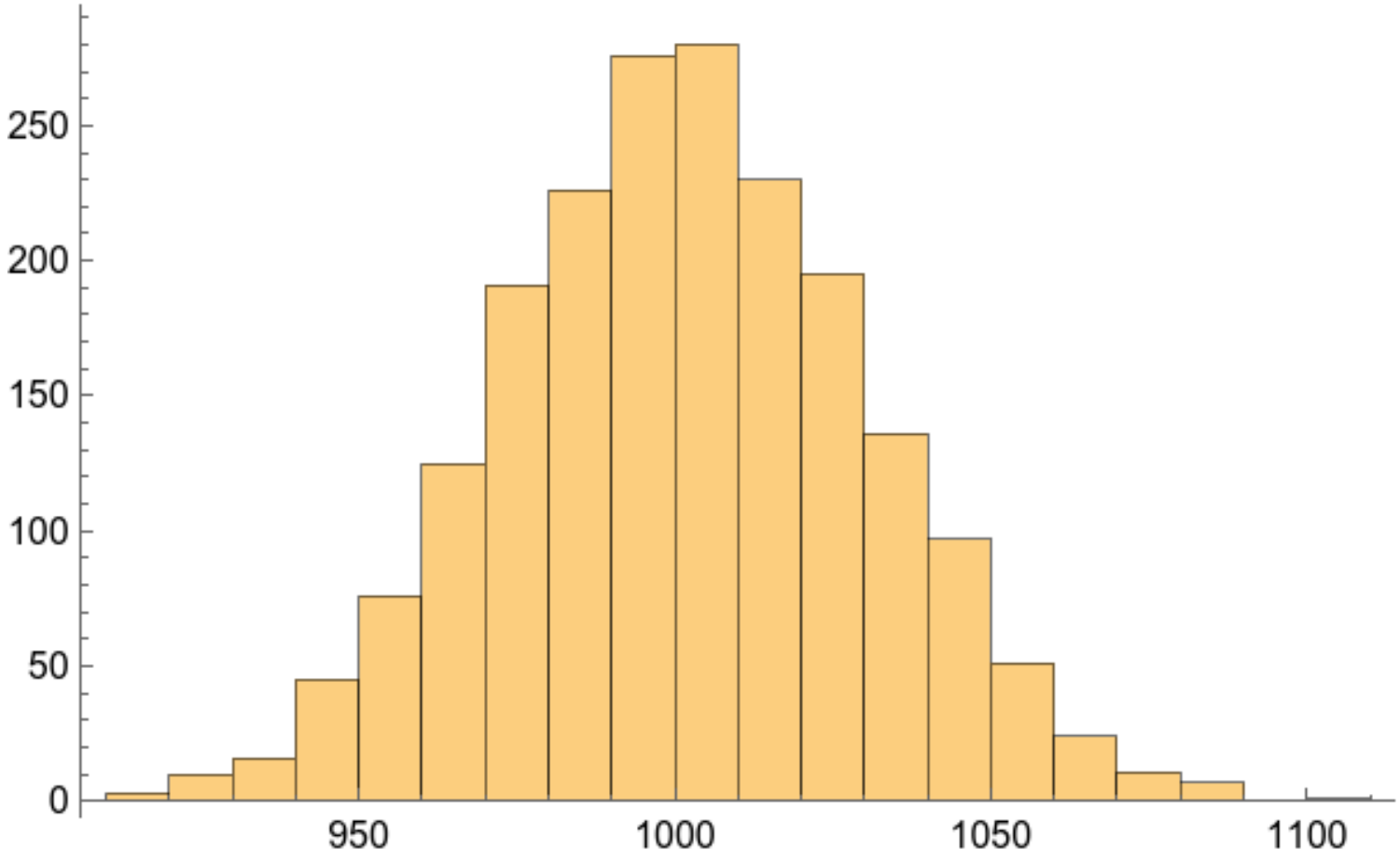
This is the kind of result we are after here, not a profound analysis of what randomness really means, or how one can define it mathematically.

Actual Experiment



Repeat the “roll 6000-times” experiment 2000 times.

Histogram



There is a brilliant application of randomness to combinatorics due to Paul Erdős. Suppose you want to show that there is a graph that has the infamous *foobag* property.

The honest-labor approach is to sit down and construct a *foobag* graph. Alas, that may be very hard.

The alternative is to do the following instead:

- consider a random graph, and
- show that the probability that is *foobag* is larger than zero.

Done!

One big surprise of the last half-century of algorithms research is that randomness is a priceless resource in the design of algorithms.

This is a bit weird, one usually thinks of an algorithm as a well-organized, strictly logical sequence of simple, mechanical steps. We always know exactly what happens next. Doing things at random sounds more like a big monkey wrench.

Wrong. Most complexity theorists as well as applied algorithms people would agree with the following statement:

> Any halfway reasonable concept of an "efficient algorithm" must allow for randomness.

Prime Example

All practical general primality testing algorithms use randomness (Solovay-Strassen, Miller-Rabin).

There is a polynomial time primality test due to Agrawal, Kayal and Saxena (2003) that essentially uses only high school arithmetic that runs in time (n^6) (just to be clear: n is the number of bits of the input).

Unfortunately, it is totally impractical.

Hilbert's 6th Problem

Hilbert was the first to point out that an axiomatization of probability was needed, a task solved by Kolmogorov in 1933 (note that this took three decades).

> Mathematical Treatment of the Axioms of Physics.
>
> The investigations on the foundations of geometry suggest the problem: To treat in the same manner, by means of axioms, those physical sciences in which already today mathematics plays an important part; in the first rank are the theory of probabilities and mechanics.

This is in reference to Hilbert's seminal "Grundlagen der Geometrie" from 1899, the first modern axiomatization of a mathematical area. Well worth reading even today.

Kolmogorov's Framework

Kolmogorov gave an almost unreasonably simple axiomatization of probability. Here are the key concepts[†]:

- There is a set Ω of all possible outcomes of an experiment. Ω is called the sample space.
- The elements of Ω are the elementary events.
- Subsets of Ω form compound events.

So an event is just any collection of basic, elementary events.

[†]Note how Kolmogorov is pushing probability in the direction of set theory.

Simple Examples

Experiment: roll two dice.

Event "doubles": both dice show the same number of spots.

Event "four": the sum of spots on both dice is 4.

Experiment: flip a coin 10 times.

Event: the coin lands Heads up exactly 5 times.

Event: the coin lands Heads up at least 7 times.

Kolmogorov Axioms

We want to associate a probability with each event that conforms more or less to our intuition and aligns with physical reality[†]. Technically, we need a probability measure or a probability distribution, a map

$$\Pr : \mathfrak{P}(\Omega) \to \mathbb{R}$$

subject to the following constraints:

- $0 \leq \Pr[A]$
- $\Pr[\Omega] = 1$
- $A \cap B = \emptyset$ implies $\Pr[A \cup B] = \Pr[A] + \Pr[B]$

$A \cap B = \emptyset$ means that the events are mutually exclusive. The last axiom expresses additivity of probability.

[†]If you are a gambler, the match better be really good

Old Hat

This is yet another chapter in our quest to measure things, and in particular sets.

We already have one hugely important measure for arbitrary sets, namely cardinality.

Probability measures are another, substantially different example of a measure.

Consequences

In the Kolmogorov setup, $\emptyset$ is the impossible event, and Ω the certain event, with probabilities 0 and 1, respectively.

- $\Pr[\,\overline{A}\,] = 1 - \Pr[A]$
- $0 \leq \Pr[A] \leq 1$
- $A \subseteq B$ implies $\Pr[A] \leq \Pr[B]$

For example, from the axioms we can directly compute

$$1 = \Pr[\Omega] = \Pr[A \cup \overline{A}] = \Pr[A] + \Pr[\overline{A}]$$

Additivity

From the additivity axiom and induction we immediately get full finite additivity: for any finite family of mutually exclusive events

$$\Pr[A_1 \cup A_2 \cup \ldots \cup A_k] = \Pr[A_1] + \Pr[A_2] + \ldots + \Pr[A_k].$$

For infinite spaces we often want more: countable additivity

$$\Pr[\bigcup_n A_n] = \sum_n \Pr[A_n]$$

for mutually exclusive events A_n; but, this is not required by the axioms.

A Glitch in the Matrix

Note well: Kolmogorov's axioms only describe probability, the likelihood of a certain event occurring.

The axioms are perfect in the sense that they are simple and yet enormously useful. But, they do **not** address the more basic question of randomness at all[†].

As a matter of experience, if we think of probability as some sort of limiting frequency, the axioms describe certain physical systems like dice very accurately—but they do not explain where the randomness comes from. That's physics.

[†]This is a smart move. Defining randomness is brutally hard and requires computability theory and measure theory

Discrete Spaces

We will focus on the case when Ω is not too large:

- finite spaces
- countably infinite spaces

Dealing with finite probability spaces often comes down to a lot of combinatorics, a lot of counting. Surprisingly, thinking about a problem probabilistically rather than in terms of direct counting can make life easier.

For countable spaces we need to deal with infinite sums, things look a bit more like analysis.

Uniform Probabilities

The easiest scenario for a finite space Ω occurs when all elementary events are equiprobable, their probabilities are all equal. For all $\omega \in \Omega$

$$\Pr[\omega] = 1/|\Omega|$$

Just think about fair coins or dice[†].

What does fair mean? Well, the probability of getting Heads is $1/2$.

Yup, the last explanation is completely useless. What we really need is a physical description of a fair coin. It is an excellent and difficult exercise to come up with one. Then try a fair die.

[†]We have written ω rather than the pedantic $\{\omega\}$

Frequencies

How do we determine the elementary probabilities $\Pr[\omega]$ for $\omega \in \Omega$?

How about the probability for a particular coin to show heads?

Well, we flip the coin N times, where N is "sufficiently large," and we count the number of heads. Then

$$p \approx \frac{\#\text{heads}}{N}$$

is a reasonable approximation for the probability of heads. The approximation gets better if we make N larger

Sounds plausible, right?

Workhorse

Uniform probabilities are often the default assumption when no better information is available.

For example, in the analysis of average case performance of algorithms, this is often the weapon of choice. It is relatively easy to deal with this case, and often the assumption appears not to be too far off the mark[†].

Typical example: average case analysis of sorting algorithms.

Similarly we may tackle randomized algorithms this way (quicksort, quickselect).

[†] 40 years ago, Smale's analysis of the simplex algorithm caused a bit of an uproar.

Dire Warning

Uniform probabilities do not work for countably infinite spaces.

The reason is simple: suppose we try to assign $\Pr[\omega] = p$ for some positive real p.

But then $\Pr[\Omega]$ diverges and certainly is not equal to 1.

For example, if we wanted to assign probabilities to natural numbers we would need to do something like

$$\Pr[n] = 2^{-(n+1)}$$

Typical Example: Rolling 2 Dice

The experiment is: roll one die, then roll another, report the spots.

It is critical here that the second die is not supposed to know anything about the first; the two outcomes are completely independent. No entanglement here.

How do we model this scenario:

- $\Omega = [6] \times [6]$
- uniform probability $1/36$ for all elementary events

Of course, this will not work if the dice are loaded.

Events

Rolling doubles: $A = \{ (i,i) \mid i \in [6] \}$.

$\Pr[A] = \sum_{\omega \in A} \Pr[\omega] = |A|/36 = 1/6$.

Come up with an argument why this result sounds right.

Counting spots: $A_k = \{ (i,j) \mid i + j = k \}$, $2 \leq k \leq 12$.
Let $C_k = |A_k|$, the number of elementary events in A_k.

k	2	3	4	5	6	7	8	9	10	11	12
C_k	1	2	3	4	5	6	5	4	3	2	1

The probabilities are then $\Pr[A_k] = C_k/36$.

Exercise

Question:
Could there be weird dice that produce the same probabilities?

More precisely, we want to assign positive natural numbers to the 2×6 faces of the dice. Each face is supposed to have probability $1/6$ and we want the same frequencies for totals as in the last table.

Unlike with ordinary dice, we are allowed to repeat numbers, and we are not constrained to $\{1, 2, \ldots, 6\}$.

Wurzelbrunft's Model

Back to standard dice. Wurzelbrunft is very fond of parallel computation and prefers to think of the two dice as being thrown at the same time, without any distinction between a first and second die.

The result would then be an unordered pair $\{a, b\}$ (we allow $a = b$ here). In this case, there are only 21 possible outcomes.

Here is the analogous table from above:

k	2	3	4	5	6	7	8	9	10	11	12
C_k	1	1	2	2	3	3	3	2	2	1	1

Quoi? What is wrong here? How do we fix it?

Finite Spaces

Dice are a typical example of the easiest setup: finite probability spaces

$$\langle \Omega, \mathsf{Pr} \rangle \qquad \Omega = \{\omega_1, \ldots, \omega_n\}$$

In this scenario, all event probabilities are obtained by summation from the elementary ones, $p_\omega = \mathsf{Pr}[\omega]$ for $\omega \in \Omega$: for all events $A \subseteq \Omega$ we have

$$\mathsf{Pr}[A] = \sum_{\omega \in A} p_\omega$$

This may look utterly straightforward, but for large spaces and complicated events it can be exceedingly difficult to evaluate these sums (recall finite combinatorics).

Countable Spaces

For countable spaces we still get away with summations

$$\Pr[A] = \sum_{\omega \in A} p_\omega$$

but this time the sums are (usually) infinite.

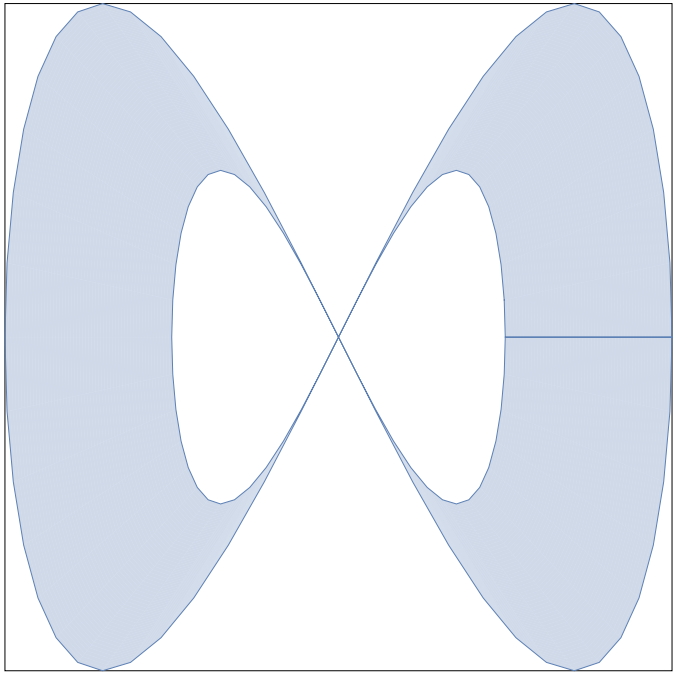
The good news is that we need not worry about convergence since

- $0 \leq p_\omega$
- $\sum_\omega p_\omega = 1$

So this is a lot easier than the standard scenario in calculus where one has to deal with convergence issues all the time.

And Continuous Spaces?

Here things get messy. Think about throwing a dart at the unit square. What is the probability that the dart ends up in the blue region below?



We would need to calculate the area of the blue region, a problem handled in measure theory.

The Measure Problem

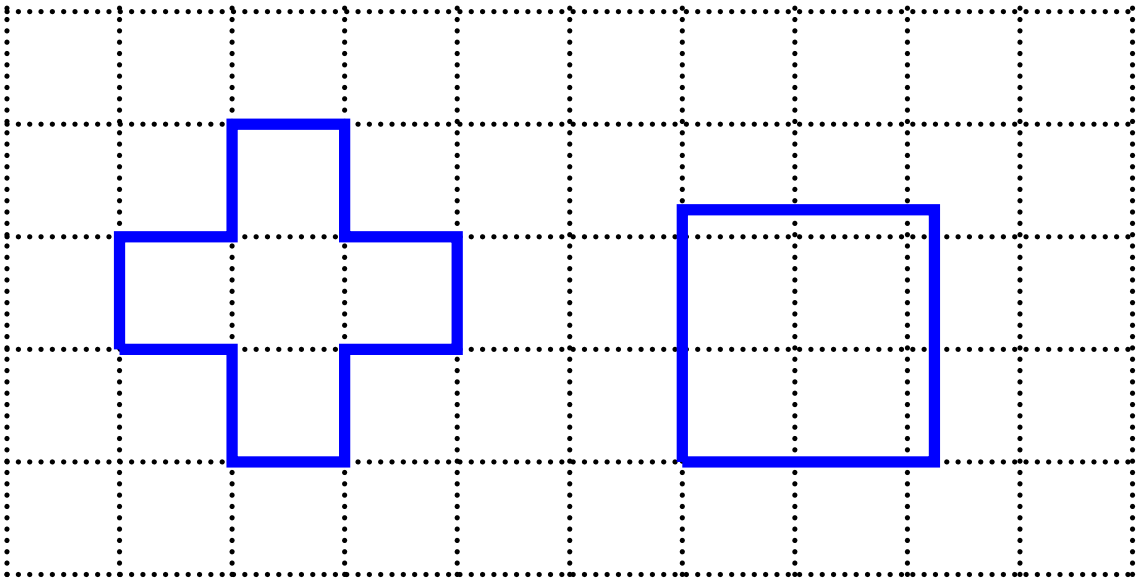
How do we measure area, volume and so on in a Euclidean space $\mathbb{R}^n$? Well, we need a d-dimensional measure, a map $\mu : \mathfrak{P}(\mathbb{R}^d) \to \mathbb{R}_{\geq 0}$ such that

- Normalization:
  $\mu([0,1]^d) = 1$.
- Finite additivity:
  $A \cap B = \emptyset$ implies $\mu(A \cup B) = \mu(A) + \mu(B)$.
- Invariance:
  for any two equidecomposable sets $A, B \subseteq \mathbb{R}^d$: $\mu(A) = \mu(B)$.

By equidecomposable we mean the following: we can partition A and B into finitely many pieces, $A = \bigcup_{i \in [n]} A_i$, $B = \bigcup_{i \in [n]} B_i$, such that A_i is congruent to B_i.

Totally reasonable requirements, or so it seems.

A Challenge



The square on the right is $\sqrt{5}$ by $\sqrt{5}$, so both polygons have area 5.

Show that the two polygons can be chopped up into congruent triangles. What is the least possible number of triangles?

Theorem (Banach-Tarski Paradox, 1924, (AC))

The unit sphere and the sphere of radius 2 are equidecomposable.

This sounds utterly insane: how can we make a big sphere out of a little sphere?

The reason the theorem works is that the pieces are very strange and cannot be visualized. In particular, we cannot assign qualities such as "volume" or "probability" to these pieces, they are not measurable. So, there is no contradiction.

Now What?

Just give up on the idea of finding measures on the whole powerset

$$\mu : \mathfrak{P}(\mathbb{R}^d) \longrightarrow \mathbb{R}$$

make do with a map defined only on "reasonable" subsets of $\mathbb{R}^d$. If you are familiar with the Lebesgue measure, that works just fine, no problem[†].

Back to Reality.

[†] If you are more daring, switch to a different universe where all sets of reals are Lebesgue measurable; R. Solovay, 1970.

General Unions

How about general unions, without mutual exclusiveness? For 2 events, there is no problem: we just have to avoid double-counting.

$$\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$$

Alas, for more than 2 events, things get complicated. If an *upper bound* is sufficient we may get away with

Union Bound (aka Boole's Inequality): [†]

$$\Pr[\bigcup A_i] \leq \Pr[A_1] + \Pr[A_2] + \ldots + \Pr[A_n].$$

[†] If this sounds utterly obvious, note that Boole worked almost a century before Kolmogorov. There were no axioms at the time.

And Three?

$$\begin{aligned}\Pr[A \cup B \cup C] = \Pr[A] + \Pr[B] + \Pr[C] + \\ - \Pr[A \cap B] - \Pr[A \cap C] - \Pr[B \cap C] + \\ + \Pr[A \cap B \cap C]\end{aligned}$$

Exercise

Check that every elementary event is counted exactly once.

Inclusion-Exclusion

For unions of n events, there is a group of inequalities that start at Boole's inequality, and end with full inclusion-exclusion. Let's assume that $A = \bigcup_{i \in [n]} A_i = \bigcap_{i \in \emptyset} A_i$. Then

$$\sum_{I \subseteq [n]} (-1)^{|I|} \Pr[\bigcap_{i \in I} A_i] = 0$$

This may seem rather messy, but for $n = 2$ we just get

$$\Pr[A_1 \cup A_2] - \Pr[A_1] - \Pr[A_2] + \Pr[A_1 \cap A_2] = 0$$

for $I = \emptyset, \{1\}, \{2\}, \{1, 2\}$, respectively. Nothing new here. For $n = 3$ we would get the result from the previous slide.

Bonferroni Inequalities

To lighten notation, define $\mathcal{I}_k = \{\, I \subseteq [n] \mid 1 \leq |I| \leq k \,\}$. Then for odd k

$$\Pr[A] \leq \sum_{I \in \mathcal{I}_k} (-1)^{|I|+1} \Pr[\bigcap\nolimits_{i \in I} A_i\,]$$

But note that for even k the inequality flips:

$$\Pr[A] \geq \sum_{I \in \mathcal{I}_k} (-1)^{|I|+1} \Pr[\bigcap\nolimits_{i \in I} A_i\,]$$

For $k = n$ we're back at full inclusion-exclusion, and equality holds.

Conditional Probability

Often one has additional information about the state of affairs that can affect the probability of some event A. This is captured by the notion of conditional probability: suppose $\Pr[B] > 0$ and set

$$\Pr[\, A \mid B \,] = \frac{\Pr[A \cap B]}{\Pr[B]}$$

For example, if we roll 2 dice, the probability of getting a 4 is $1/12$.

But, if we know that the result is even, the probability changes to $1/6$.

On the other hand, if the result is odd, it changes to 0.

Law of Total Probability

The last example suggests that it may help to partition Ω into mutually exclusive events $B_1, \ldots, B_k$.

We can then compute probabilities in slices:

$$\Pr[A] = \sum \Pr[A \cap B_i] = \sum \Pr[\, A \mid B_i \,] \Pr[B_i]$$

Of course, the trick is to choose the B_i so that the conditional probabilities are easy to compute.

Urns

It is an age-old principle that every discussion of basic probability involves urns.

Example: We have 3 urns each containing red or green balls: $\{G, G\}$, $\{G, R, R\}$, $\{G, R, R, R\}$. Experiment: pick an urn with probabilities $1/2, 1/4, 1/4$, then a ball in the chosen urn, uniformly at random.

What is the probability that the ball is red?

Solution

Events: let B_i "urn i is chosen" and A "ball is red."

We have $\Pr[B_1] = 1/2$, $\Pr[B_{2/3}] = 1/4$.

$\Pr[\, A \mid B_1 \,] = 0$

$\Pr[\, A \mid B_2 \,] = 2/3$

$\Pr[\, A \mid B_3 \,] = 3/4$

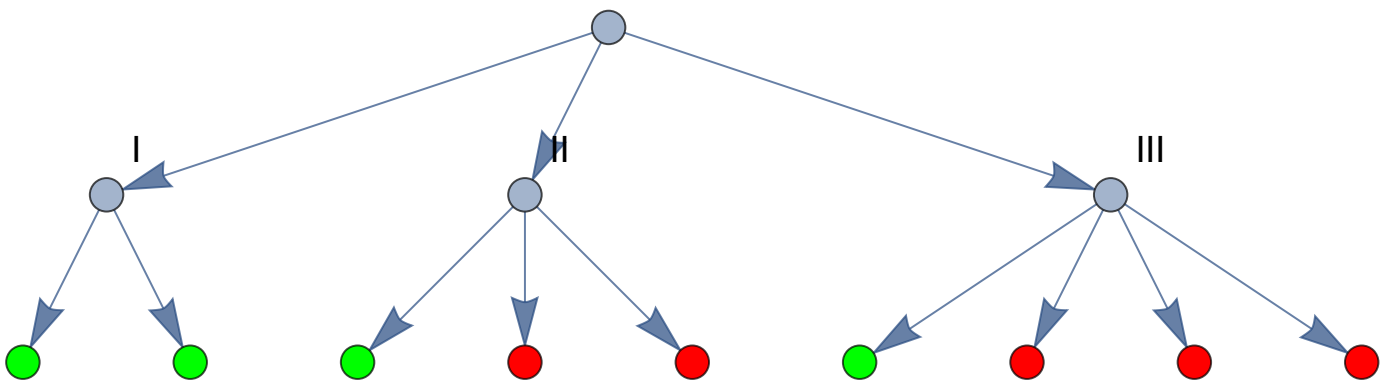
So the answer is

$$\Pr[A] = 0 \cdot 1/2 + 2/3 \cdot 1/4 + 3/4 \cdot 1/4 = 17/48 \approx 0.35$$

And Ω?

Note that we never specified the sample space Ω.

This is typical of probability texts, everyone states the Kolmogorov axioms, and then forgets what they actually say.



Independence

Here is the opposite idea: two events A and B are independent if knowledge of one provides no information about the other.

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$$

Independence is critical in randomized algorithms: we have to assume that the (pseudo-)random bits know nothing about each other, otherwise the analysis collapses.

Since randomized algorithms may produce false answers, one usually runs them repeatedly and then takes a majority vote. Again, independence is critical for this to work (think about using the same "random bits" in each run).

Independent Dice

Again roll 2 dice and consider the following events

- A "even"
- B "sum is in $\{4, 5\}$"
- C "sum is in $\{4, 5, 8, 9\}$"

Then A and B fail to be independent, but A and C are independent.

Make sure to do the calculations so you can see why.

Exercise

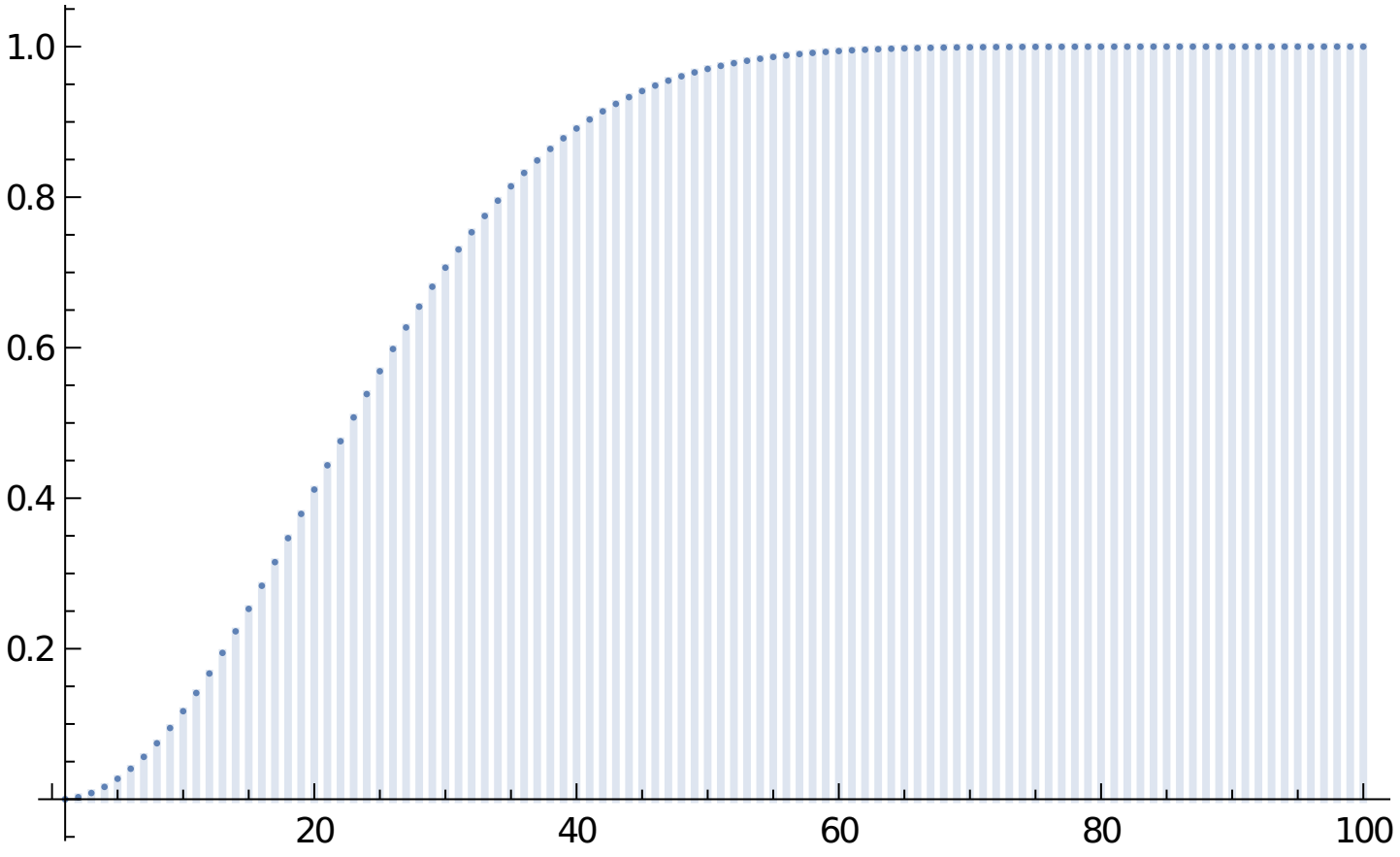
Suppose A and B are independent. Show that $\overline{A}$, B; A, $\overline{B}$ and $\overline{A}$, $\overline{B}$ are all independent.

Birthday Problem

If there are n people in a room, what is the likelihood that at least two share a birthday?

Tacit assumptions: there are 365 days in a year, birthdays are equiprobable, birthdays of people in the room are independent (no twins, triplets and the like), $n \leq 365$.

$$\Pr[E] = 1 - \frac{364}{365}\frac{363}{365}\cdots\frac{365-n+1}{365}$$


1.0
0.8
0.6
0.4
0.2
20
40
60
80
100

Generalize

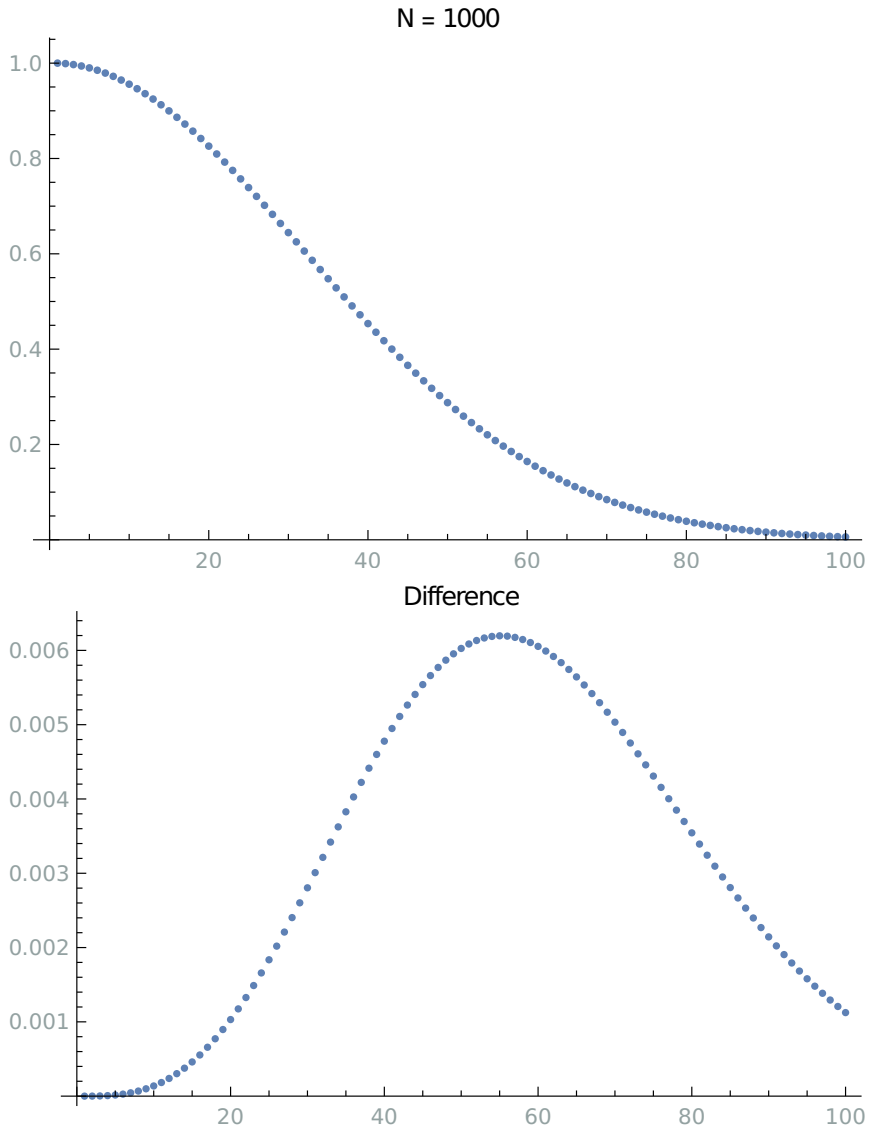
More generally, we could have n objects that all have one of N properties.

Under the usual assumptions, the probability that no two share the same property is

$$P(n, N) = (1-1/N)(1-2/N)\ldots(1-(n-2)/N)(1-(n-1)/N)$$

Using logarithms and the approximation $\log(1-x) \approx -x$ for small x we get the approximation

$$P(n, N) \approx e^{-n(n-1)/2N}$$


N = 1000
1.0
0.8
0.6
0.4
0.2
20
40
60
80
100
Difference
0.006
0.005
0.004
0.003
0.002
0.001
20
40
60
80
100

Bayes' Law

The following is an immediate consequence of our definitions. Suppose $\Pr[B] > 0$.

$$\Pr[\, A \mid B \,] = \frac{\Pr[\, B \mid A \,] \Pr[A]}{\Pr[B]}$$

Here is a useful generalization: suppose we have a partition $B_1, \ldots, B_n$ of Ω where $\Pr[B_i] > 0$ for all i.

$$\Pr[\, B_k \mid A \,] = \frac{\Pr[\, A \mid B_k \,] \Pr[B_k]}{\sum_i \Pr[\, A \mid B_i \,] \Pr[B_i]}$$

Some find this result rather puzzling: it seems to reverse the relationship between cause and effect.

More Urns

Suppose we have three urns with red and green balls. Let's assume the counts are $(3, 7)$, $(5, 5)$ and $(4, 6)$, respectively. We pick one of the urns and then select one of the balls in the urn, both uniformly at random. Suppose the result is a red ball.

What is the likelihood that urn 1 was chosen?

Events: B_i "urn i was chosen", A "red ball chosen" (as the final result).

Then Pr$[B_i] = 1/3$

Pr$[\, A \mid B_1 \,] = 3/10$, Pr$[\, A \mid B_2 \,] = 1/2$ and Pr$[\, A \mid B_3 \,] = 2/5$.

Hence Pr$[\, B_1 \mid A \,] = 1/4$.

Random Variables

Experiments are often associated with some numerical quantity, a measurement of sorts, which depends on random outcomes: these thingies are random variables and defined as maps

$$X : \Omega \to \mathbb{R}$$

So we are assigning a real value to each elementary event.

The discrete case is always fine and it makes sense to talk about the probability distribution or probability mass function (pmf)

$$p(a) = \mathsf{Pr}[X = a] = \mathsf{Pr}[X^{-1}(a)]$$

Why Reals?

Sometimes one would like to associate outcomes other than reals with an experiment. For example, we might want a Yes/No answer in a randomized decision algorithm (important example: primality checker).

This is no problem: as usual, we can fake Yes/No answers by using 1/0 instead. In probability theory, this device is called an indicator variable

$$X(a) = \begin{cases} 1 & \text{if } a \in A, \\ 0 & \text{otherwise.} \end{cases}$$

Yup, this is just the characteristic function of $A \subseteq \Omega$. Different field, different terminology, same idea.

And More

In addition, reals are just too convenient. For example, nothing stops us from computing

$$5X^2(a) + 3X(a) - 17$$

Or we could add or multiply random variables as in $X(a) + Y(a)$, and so on and so forth.

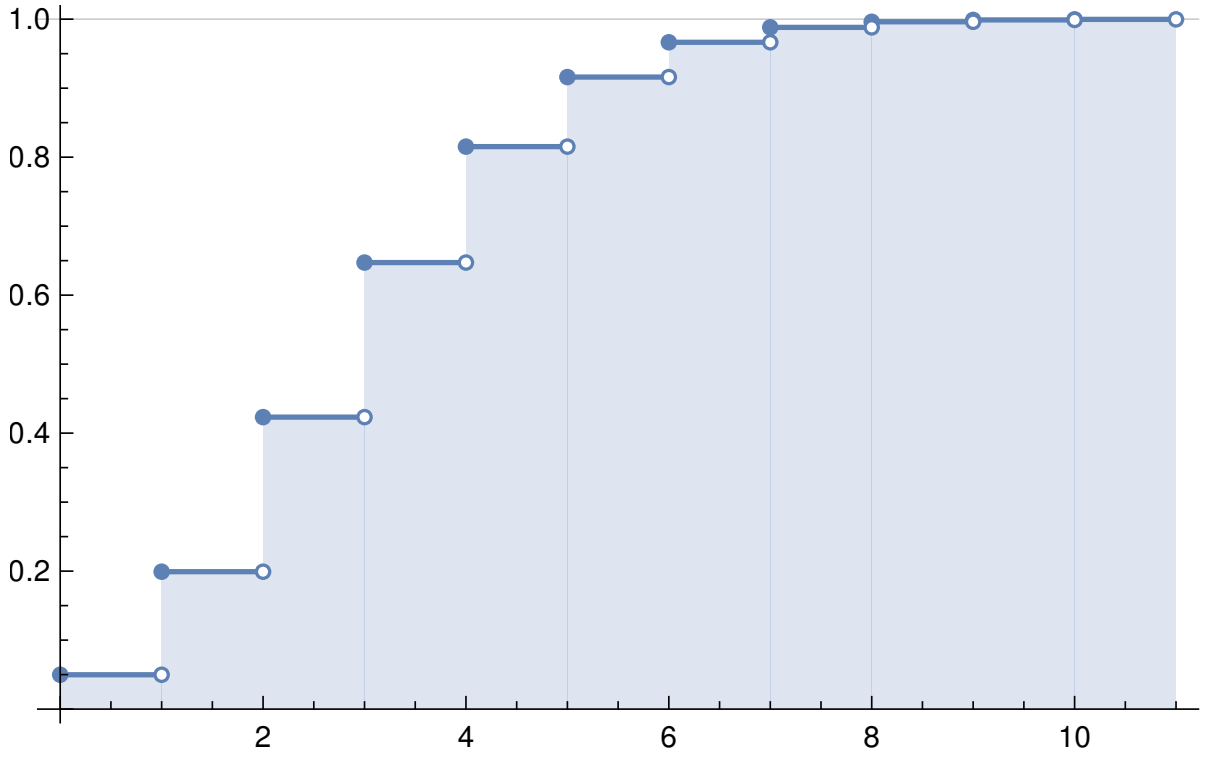
This allows us to exploit numerous tools from analysis.

Cumulative Distribution Function

The cumulative distribution function considers all smaller values of a random variable:

$$\Pr[X \leq a]$$

In the discrete case, the cdfs increase in steps.



Averages

The "average" value of a random variable is usually a good first step in any attempt to understand what is going on.

There are several reasonable ways to think about averages:

Mean:	sum of values over total number of values
Median:	middle numerical value
Mode:	most frequent value

Expectation

Suppose we have a discrete random variable X with pmf $p(a)$.

The expected value or expectation of X is

$$\mathsf{E}[X] = \sum X(a) \cdot p(a)$$

So expectation is the mean, a weighted sum and arguably the most intuitive notion of average. This is often abbreviated in slightly criminal manner to μ.

Lemma

Expectation is linear in the sense that

$$\mathsf{E}[aX + bY] = a\,\mathsf{E}[X] + b\,\mathsf{E}[Y]$$

where a and b are real constants.

Variance

Other than the average value itself, it is also useful to now how far off the values of a random variable might be, on average. In other words, we try to measure the dispersion of the values around the mean.

Let μ be the expectation of X. The variance of X is

$$\mathsf{Var}[X] = \mathsf{E}[(X-\mu)^2]$$

This is often written as σ^2 (where σ is the standard deviation).

One might be tempted to study $\mathsf{E}[|X-\mu|]$ instead, but using squares is slightly easier (no cases).

Proposition

$\mathsf{Var}[X] = \mathsf{E}[X^2] - \mathsf{E}[X]^2$

Proof.

$$\begin{aligned}
\mathsf{Var}[X] &= \mathsf{E}[(X-\mu)^2] \\
&= \mathsf{E}[X^2 - 2X\mu + \mu^2] \\
&= \mathsf{E}[X^2] - \mathsf{E}[2X\mu] + \mathsf{E}[\mu^2] \\
&= \mathsf{E}[X^2] - 2\mu^2 + \mathsf{E}[\mu^2] \\
&= \mathsf{E}[X^2] - \mu^2
\end{aligned}$$

□

Rules for Variance

Lemma

$$\mathsf{Var}[aX + b] = a^2 \, \mathsf{Var}[X].$$

Lemma

Assume that X and Y are independent. Then variance is additive and multiplicative in the sense that

$$\mathsf{Var}[X + Y] = \mathsf{Var}[X] + \mathsf{Var}[Y]$$

$$\mathsf{Var}[X \cdot Y] = \mathsf{Var}[X] \cdot \mathsf{Var}[Y]$$

Recall the probability mass function associated with a random variable $X : \Omega \to \mathbb{R}$.

Given any value $a \in \mathbb{R}$, the fiber of a constitutes an event: all elementary events for which X assumes value a. Hence we have the probabilities

$$\mathsf{Pr}[X = a] = \mathsf{Pr}[X^{-1}(a)]$$

This is really interesting only for $a \in \operatorname{rng} X$. It can be useful to think of X as a map $\Omega \to \operatorname{rng} X$, in particular in the discrete case.

Discrete Uniform Distribution

A random variable with finite range is uniformly distributed if the variable assumes all possible values with equal probability:

$$\Pr[X = a] = 1/|\text{rng}\, X|$$

We could think of rolling a die, with the number of spots as a random variable. If the die is fair, this variable is uniformly distributed.

Dire Warning: this does not work for countably infinite ranges. It does work for continuous distributions.

Bernoulli

Indicator variables with range $\{0, 1\}$ model the behavior of a (possibly biased) coin.

$$\Pr[X = 1] = p$$

$$\mathsf{E}[X] = p$$

$$\mathsf{Var}[X] = p(1 - p)$$

You can think of a Bernoulli distribution as modeling a (pseudo-)random bit source for a randomized algorithm (hopefully with $p = 1/2$).

Binomial

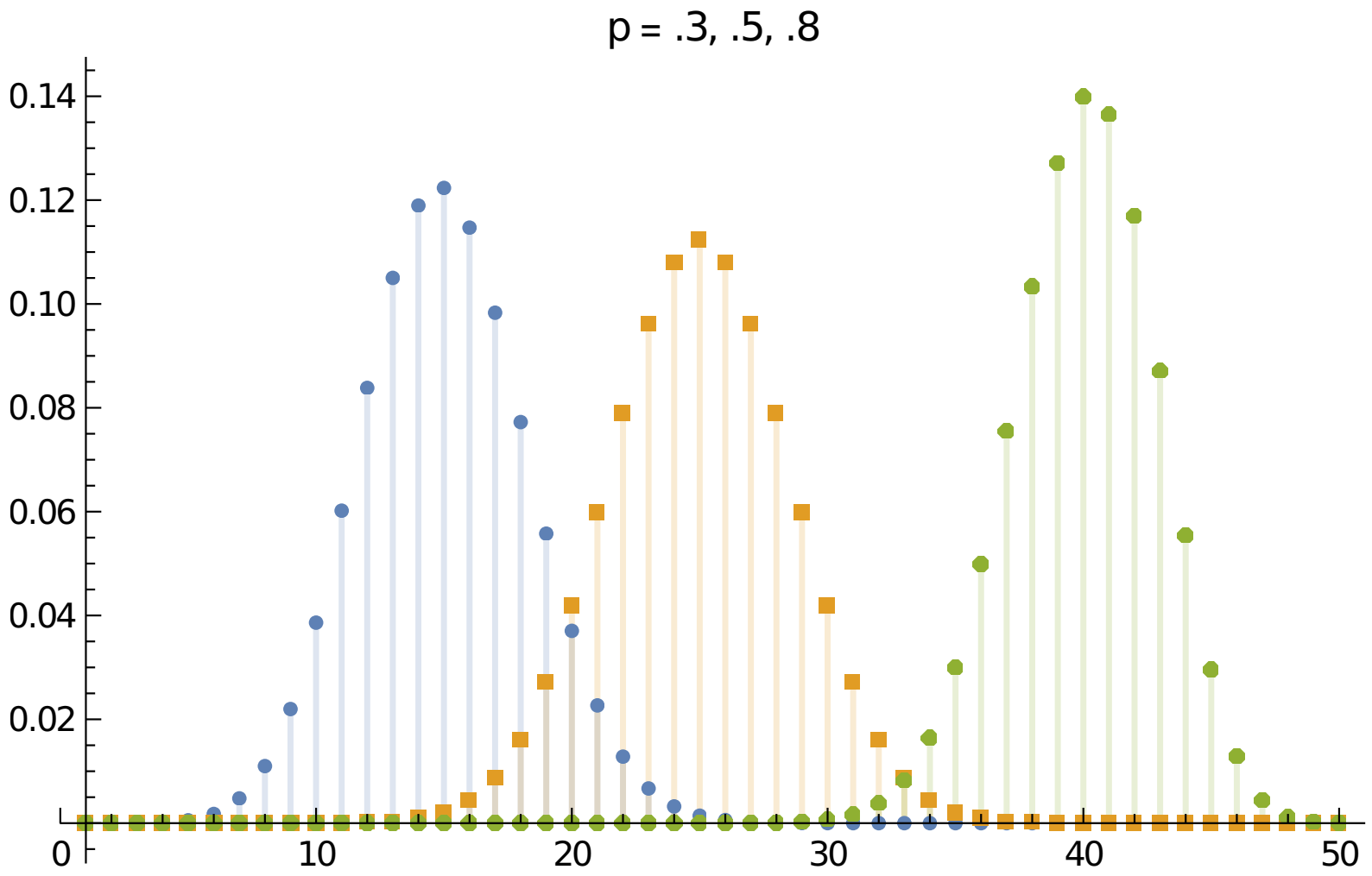
Single bits are not that interesting, so it is natural to ask what happens if we repeat a Bernoulli experiment n times, independently?

Define indicator variables X_i describing the ith repetition, and let $X = X_1 + X_2 + \ldots + X_n$. If the probability of all the X_i is p then

$$\Pr[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\mathrm{E}[X] = np$$

$$\mathrm{Var}[X] = np(1-p)$$


p = .3, .5, .8
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0
10
20
30
40
50

Geometric

Here is another variation on Bernoulli trials: suppose we count the number of times until the experiment succeeds. We get a random variable X such that

$$\Pr[X = k] = p(1-p)^{k-1}$$

$$\mathsf{E}[X] = 1/p$$

$$\mathsf{Var}[X] = (1-p)/p^2$$

Killer App: Randomized Algorithms

Several cryptographic methods require large primes (say, 1000 binary digits). And the primes have to be new, we cannot look them up in a table.

The good news is that one can prove that "a lot" of primes with a given number of bits exist.

The bad news is that we don't know how to simply generate a k-bit prime at will. Instead, we generate a random k-bit number and check whether it is prime. Since there are lots of primes, we will find one fairly soon.

How do we check primality? Well, the only feasible algorithms are themselves randomized. There is a deterministic polynomial time algorithm (which fact is utterly amazing) but it is entirely useless in the RealWorld$^{\mathsf{TM}}$.

A Dirty Secret

All randomized algorithms make mistakes.

Why? If a "randomized algorithm" never makes mistakes, regardless of the random bits used, we can simply use $000\ldots000$. Or $010101\ldots01010$ if you find all-zeros too depressing.

With luck, the errors are one-sided. E.g., there is a primality test due to Solovay and Strassen that will always return "Yes?" given a prime number as input, but will say "No" on composite numbers only with probability $1/2$.

Bounds

Algorithms like that primality tester usually have unacceptably large error probability. In applications, we have to run them multiple times and then take a majority vote to figure out what the answer should be.

For Solovay-Strassen the expected time to get a "No" given a composite number is only 2.

But note: that is not enough. We need to understand the likelihood that the actual outcome will deviate far from the expected value. Running the algorithm only twice is a very bad idea.

As usual, there is trade-off: reliability versus running time.

Markov's Inequality

Lemma (Markov's Inequality)

Let X be a non-negative random variable with finite expectation μ. Then for any $a > 0$: $\mathsf{Pr}[X \geq a] \leq \mu/a$.

Proof.

$$\begin{aligned}
\mathsf{E}[X] &= \sum_{x<a} x \,\mathsf{Pr}[X = x] + \sum_{x \geq a} x \,\mathsf{Pr}[X = x] \\
&\geq \sum_{x \geq a} x \,\mathsf{Pr}[X = x] \\
&\geq a \sum_{x \geq a} \mathsf{Pr}[X = x] \\
&= a \,\mathsf{Pr}[X \geq a]
\end{aligned}$$

□

Chebyshev's Inequality

Lemma (Chebyshev's Inequality)

Let X be a random variable with finite expectation μ and standard deviation σ. Then for any $a > 0$: $\mathsf{Pr}[|X - \mu| \geq a\,\sigma] \leq 1/a^2$.

Proof.

Use Markov's inequality for the variable $Y = (X - \mu)^2$ (so $\mathsf{E}[Y] = \mathsf{Var}[X]$) and the fact that $x \mapsto x^2$ is a strictly monotonic function on $\mathbb{R}_{\geq 0}$.

$$\begin{aligned}\mathsf{Pr}[|X - \mu| \geq a\,\sigma] &= \mathsf{Pr}[(X - \mu)^2 \geq a^2\sigma^2] \\ &\leq \frac{\mathsf{E}[(X - \mu)^2]}{a^2\sigma^2} = 1/a^2\end{aligned}$$

□

More Urns

Suppose we throw n balls into n urns, independently and uniformly at random. Let X be the random variable: number of balls in urn #1. Define indicator variables $X_i = 1$ iff ball i lands in urn #1. So $X = \sum X_i$, $\mathsf{Pr}[X_i = 1] = 1/n$ and $\mu = \mathsf{E}[X] = 1$. Also $\mathsf{Var}[X_i] = 1/n\,(1 - 1/n)$ and $\mathsf{Var}[X] = \sum \mathsf{Var}[X_i] = 1 - 1/n$.

Markov	$\mathsf{Pr}[X \geq 7] \leq 1/7 \approx 0.142857$
Chebyshev	$\mathsf{Pr}[X \geq 7] = \mathsf{Pr}[\lvert X - 1\rvert \geq 6]$
	$\leq \dfrac{1 - 1/n}{36} \leq 1/36 \approx 0.0277778$

A much more complicated family of bounds due to Chernoff produces a much better bound of 0.00048987.

Experimental Verification

```
Table[
       Count[ Table[ RandomChoice[Range[100]], {100} ], 1 ],
       {100000} ] // Frequencies
```

produces the output

0	36485
1	37200
2	18576
3	5945
4	1486
5	254
6	50
7	3
8	1

Another Bound

Suppose we want a bound for $\Pr[X \geq 1 + 10 \ln n]$. This is a more interesting situation.

Markov $\quad \Pr[X \geq 1 + 10 \ln n] \leq 1/(1 + 10 \ln n)$

Chebyshev $\quad \Pr[X \geq 1 + 10 \ln n] = \Pr[|X - 1| \geq 10 \ln n]$

$$\leq \frac{1 - 1/n}{10^2 \ln^2 n} \leq \frac{1}{100 \ln^2 n}$$

Again, Chernoff produces much stronger results: we get a bound of $1/2\, n^{-6.931}$, and, with more skulduggery, even n^{-10}.