GTI

Randomness and Probability

A. Ada, K. Sutner

Carnegie Mellon University

Spring 2018

1. Randomness

- Probability Theory

As you may remember fondly, you already saw an introduction to probability in 15-151:

> C. Newstead & J. Mackey
> An Infinite Descent Into Pure Mathematics
> Chap. 7

This is the wilted stack of notes under your pillow . . .

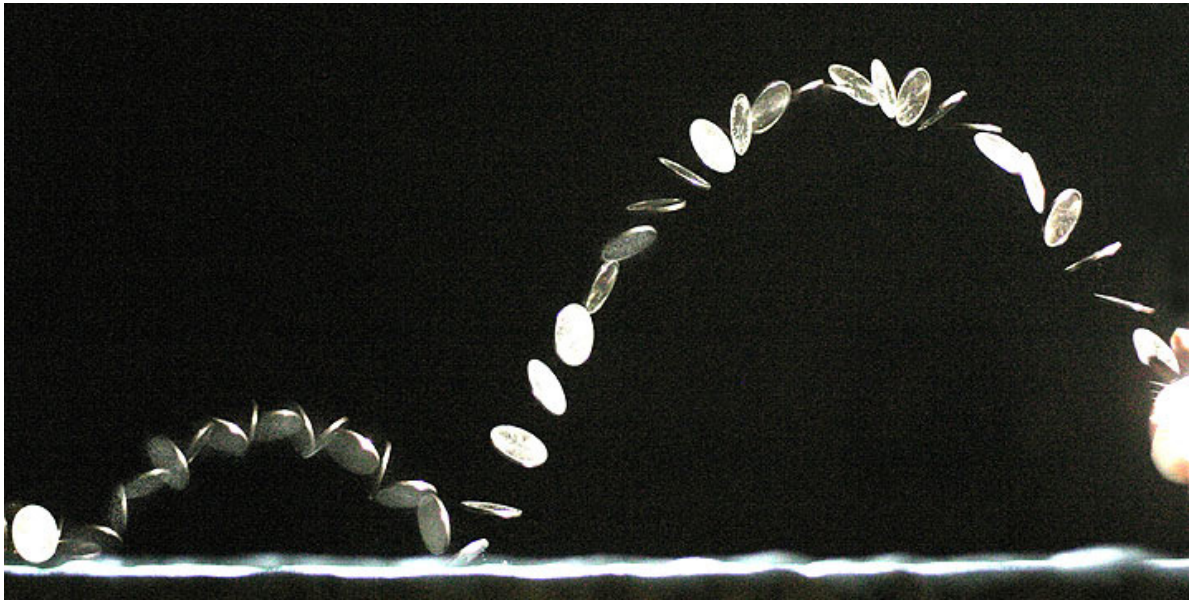
Next week we will discuss the use of randomness to speed up algorithms, one of the most important ideas in the theory of algorithms.

In preparation, this lecture is a

- gentle reminder to go back an re-read Chap. 7, if need be, and
- an attempt to explain some of the more foundational issues;

Randomness is one of the most perplexing ideas in ToC: defining randomness in any mathematically correct way is very, very hard.

Yet, any 6-year old knows very well from experience what randomness is: flipping a coin or rolling a die is a perfect example.

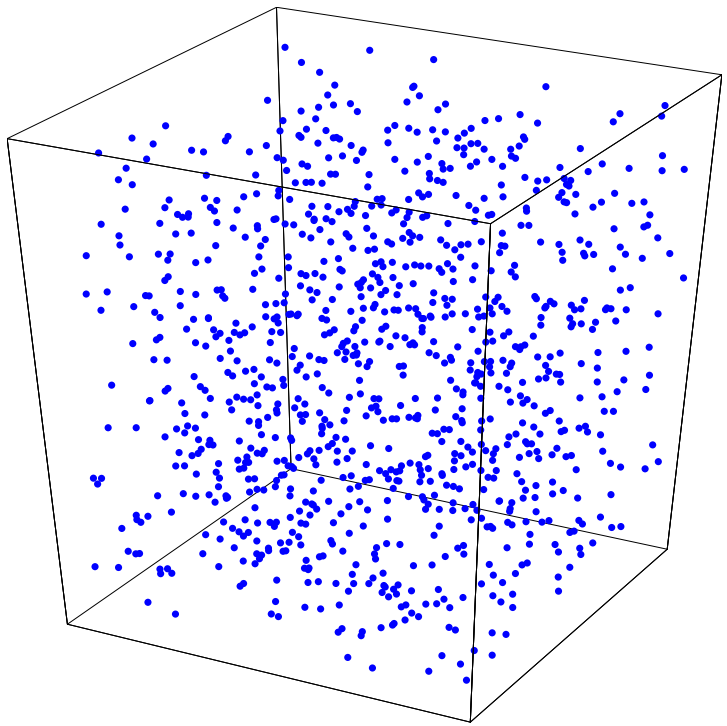


How Random Is It?

Is the randomness in a coin-toss real or is it actually confined to just the initial conditions?

Persi Diaconis, a Stanford mathematician and highly accomplished professional magician, supposedly can consistently produce ten consecutive heads flipping a coin – by carefully controlling the initial conditions.

Radioactivity is another great source of randomness – except that no one likes to keep a lump of radioactive material and a Geiger-Müller counter on their desk. Solution: keep the radioactive stuff someplace else and get the random bits over the web.



True random bits from `www.fourmilab.ch`.

The last system (and also the lava lamps, see below) is very different from the others: if our current understanding of physics is halfway correct, there is no way to predict certain events in quantum physics, like radioactive decay. It is fundamentally impossible (even if we could establish initial conditions correctly, which we cannot thanks to Herr Heisenberg).

The other, purely mechanical systems such as dice and coins, we encounter deterministic chaos: given sufficiently precise descriptions of the initial conditions, and sufficient compute power, one could in principle compute the outcomes (if we think of them as classical systems).

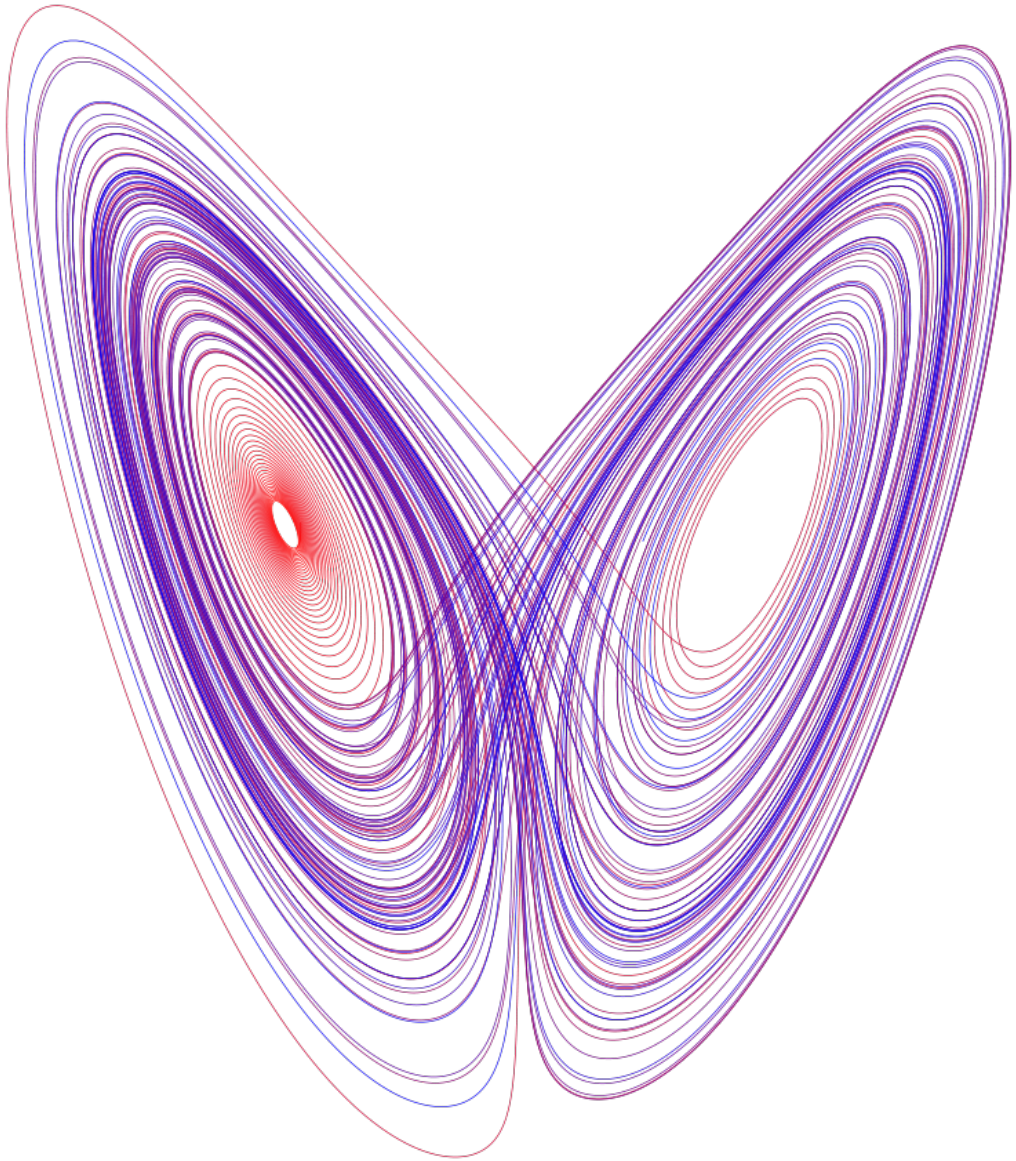
In principle only, not in practice.

Here is a famous example discovered by Lorenz in the 1963, in an attempt to study a hugely simplified model of heat convection in the atmosphere.

$$x' = \sigma(y - x)$$
$$y' = rx - y - xz$$
$$z' = xy - bz$$

These are not spatial coordinates, x stands for the amplitude of convective motion, y for temperature difference between rising and falling air currents, and z between temperature in the model and a simple linear approximation.

For certain values of the parameters we get the following behavior.



In the olden days, the RAND Corporation used a kind of electronic roulette wheel to generate a million random digits (rate: one per second).

In 1955 the data were published under the title:

A Million Random Digits With 100,000 Normal Deviates

"Normal deviates" simply means that the distribution of the random numbers is bell-shaped rather than uniform. But the New York Public Library shelved the book in the psychology section.

The RAND guys were surprised to find that their original sequence had several defects and required quite a bit of post-processing before it could pass muster as a random sequence. This took years to do.

Available at RAND.

Incidentally, Noll and Cooper at Silicon Graphics discovered one day that the pretty lava lamps were completely irrelevant: they could get even better random bits with the lens cap on (there is enough noise in the circuits to get good randomness).

Another way to use light, very much unlike the original lava lamp system, is to exploit an elementary quantum optical process: a photon hitting a semi-transparent mirror either passes or is reflected.

The Quantis systems was developed at the University of Geneva, the first practical model was released in 1998.

Note that quantum physics is the only part of physics that claims that the outcome of certain processes is fundamentally random (which is why Einstein was never very fond of quantum physics).

See Idquantique.

The Magic Device

A true random number generator.

Quantis TRNG

Features

- True quantum randomness
- High bit rate, up to 16Mbits/sec
- Low-cost device (1000+ Euros)
- Compact and reliable
- USB or PCI, drivers for Windows and Linux

Applications

- Numerical Simulations
- Statistical Research
- Lotteries and gambling
- Cryptography

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.

Kolmogorov axiomatized probability, but there is no hope for an axiomatic treatment of all of physics anywhere in the near future. It's all poetry.

Is should be noted that even today not everyone participates in the quest for absolute Hilbertian precision.

For example, physics super-star Steven Weinberg writes in a book on quantum field theory

> . . . there are parts of this book that will bring tears to the eyes of the mathematically inclined reader.

In physics, this attitude may be a good thing that helps the field along. In ToC, it would more likely be an unmitigated disaster.

It is somewhat easier to define what one means by an infinite random bit sequence rather than dealing with random finite sequences:

$$\alpha = a_0, a_1, a_2, \ldots, a_n, \ldots \in \mathbf{2}^\omega$$

Intuitively, what properties would we expect from a random α?

Always think of α as being generated by infinitely many coin tosses. Of course, we want the coin to be fair.

It is a really obnoxiuous question to ask what it means for a coin to be fair.

It is easy to define the density of a finite binary word x of length n:

$$D(x) = 1/n \sum_i x_i$$

But how about an infinite sequence α?

Definition (Density)

Let $\alpha \in \mathbf{2}^\omega$ and define the density of α up to n to be $D(\alpha, n) = D(\alpha[n])$.

The limiting density of α is

$$D(\alpha) = \lim_{n \to \infty} D(\alpha, n)$$

Note that there is a huge problem with this definition: limits are precisely defined in analysis.

But α is a wild-and-woolly object of our imagination, and there is not much reason to assume that this particular limit should exist.

In fact, it does not always exist, but we will take the patented Weinberg Approach: fuggedaboudit.

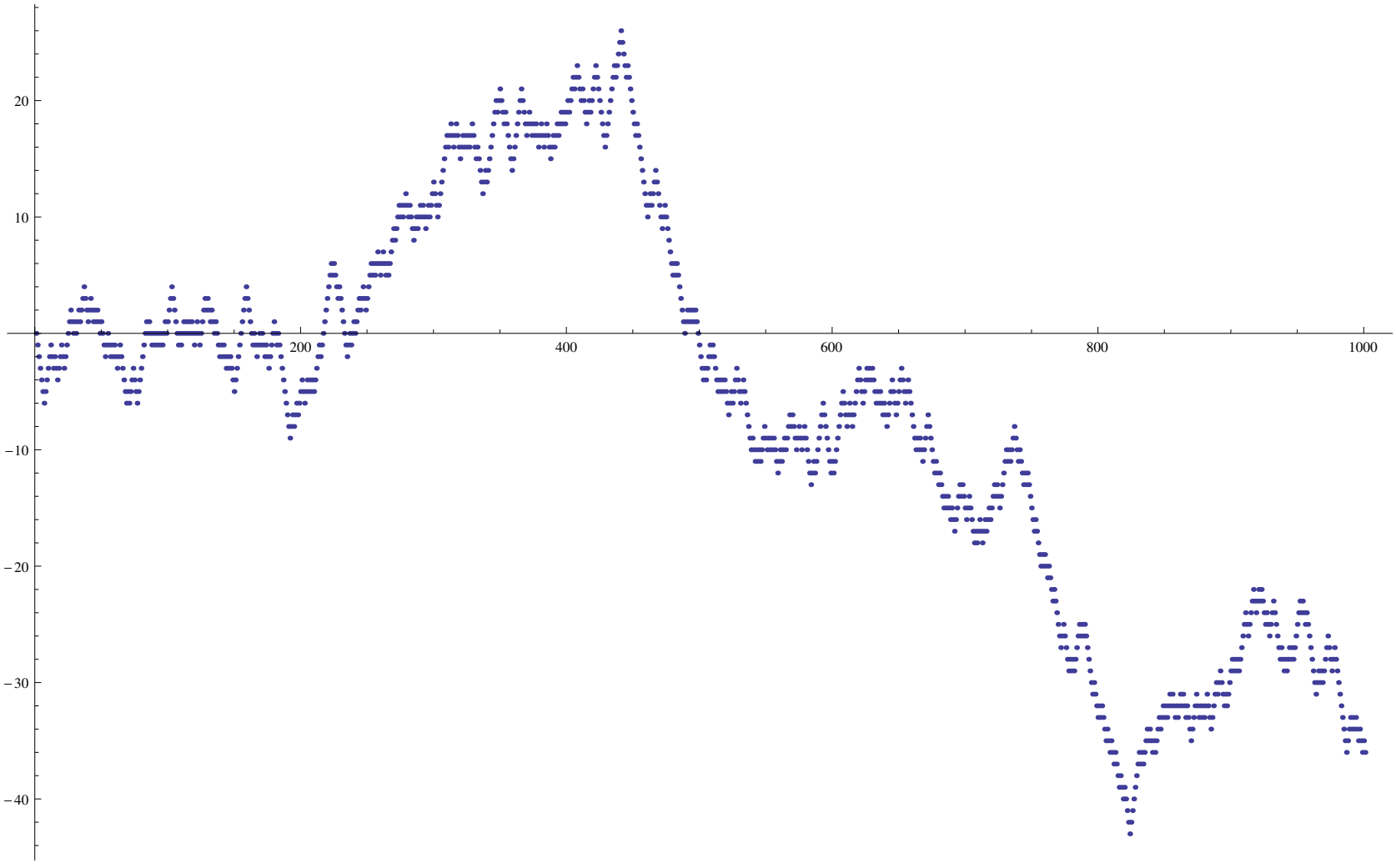
The LoLN says that if we repeat an experiment often, the observed average does in fact converge to the expected value; almost certainly.

For example, for an unbiased coin we should expect to approach the limiting density of $1/2$, almost always.

Also note that we should not expect the averages to be exactly equal to the expectation.

For example, performing a one-dimensional random walk with steps ± 1 we should expect to be up to $O(\sqrt{n})$ from the origin after n steps.

A Random Walk



Decimation

How about using Roman military traditions to define randomness?

In 1919 Richard von Mises suggested a notion of randomness based on the limiting density of the sequence itself and various decimations of it.

The idea is that "reasonable" subsequences of the given sequence should also have limiting density $1/2$.

Definition

An infinite sequence $\alpha \in \mathbf{2}^{\omega}$ is Mises random if the limiting density of any subsequence (a_{i_j}) is $1/2$ where the subsequence is selected by a Auswahlregel.

So what on earth is a Auswahlregel, a selection rule?

Intuitively, the following decimations all should have limiting density $1/2$:

$$a_0, a_1, a_2, \ldots, a_n, \ldots$$
$$a_0, a_2, a_4, \ldots, a_{2n}, \ldots$$
$$a_1, a_4, a_7, \ldots, a_{3n+1}, \ldots$$
$$a_0, a_1, a_4, \ldots, a_{n^2}, \ldots$$
$$a_2, a_3, a_5, \ldots, a_{15485863}, \ldots$$

In fact, we might want for any reasonable strictly monotonic function $f : \mathbb{N} \to \mathbb{N}$ that

$$\alpha_f = a_{f(0)}, a_{f(1)}, a_{f(2)}, \ldots, a_{f(n)}, \ldots$$

has limiting density $1/2$.

However, there is one big caveat: the selector function f must be defined without any knowledge of α: otherwise we can simply pick a subsequence of all 0's or all 1's.

Now suppose we have a countable system of Auswahlregeln and our sequence passes all these tests. In other words, for all f we have

$$D(\alpha_f) = 1/2.$$

Then α is Mises-random. One can show that for any countable collection of Auswahlregeln there are always uncountably many sequences that are random in this sense.

Sounds all eminently reasonable.

Unfortunately, in 1939 J. Ville showed that for any countable system of Auswahlregeln there is always a sequence α that passes all the tests (i.e., the limiting density is $1/2$ for all these subsequences) but that is nonetheless biased towards 1.

More precisely, it was known that a random sequence should have

$$\limsup_n \sqrt{\frac{2n}{\log\log n}}\left(D(\alpha,n)-1/2\right)=1$$

$$\liminf_n \sqrt{\frac{2n}{\log\log n}}\left(D(\alpha,n)-1/2\right)=-1$$

and Ville's example violated the second condition.

There are excellent definitions of randomness based on better tests. Instead of AuswahlregeIn one uses tests with foundations in

- computability theory, and
- topology.

The most famous one is due to Per Martin-Löf (who has also done groundbreaking work in type theory).

Following Weinberg's proud example, we will forgo this opportunity to inflict mental pain and anguish on the student body, and skip over the definition.

Unfortunately, all definitions of randomness have one unpleasant side-effect: random sequences are not computable.

This is not a big surprise: computable means predictable, and we want exactly he opposite.

> Anyone attempting to produce random numbers by purely arithmetic means is, of course, in a state of sin.
>
> John von Neumann

Mike Pence will object, but we have no problem with sin.

- Randomness

2. Probability Theory

- Discrete Probability
  - Finite spaces: Really a matter of combinatorial counting, though sometimes it is easier to argue in terms of probability.
  - Countably infinite spaces: Deals with infinite spaces and infinite summations, but everything remains civilized.
- Continuous Probability
  Uncountable spaces, really a part of measure theory, and annoyingly dependent on set theory. All hell breaks loose.

The collection of all possible outcomes of an experiment is called a sample space Ω and its elements are the elementary events or atomic events. A (compound) event is a subset of Ω.

We want to associate a probability for the occurrence of each event, a map

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

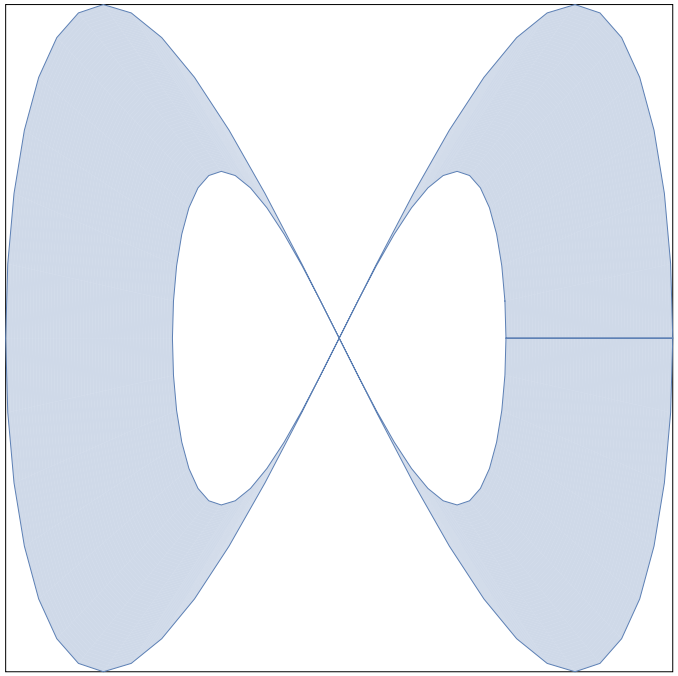
- $0 \leq \Pr[A]$
- $\Pr[\Omega] = 1$
- $A \cap B = \emptyset$ implies $\Pr[A \cup B] = \Pr[A] + \Pr[B]$

What Could Go Wrong?

Discrete probability is pretty safe, but consider the following continuous problem: we would like to measure the area of regions in Euclidean space, something like $\mu(A)$ where $A \subseteq \mathbb{R}^d$.

This is closely related to probability theory: think about an experiment like throwing a dart at the unit square. What is the probability that the dart ends up in the region A below?

Clearly we need to determine $\mu(A)$.



A Wishlist

For simplicity, consider $d = 2$.

Norm We want $\mu(R) = ab$ whenever R is an $a \times b$ rectangle.

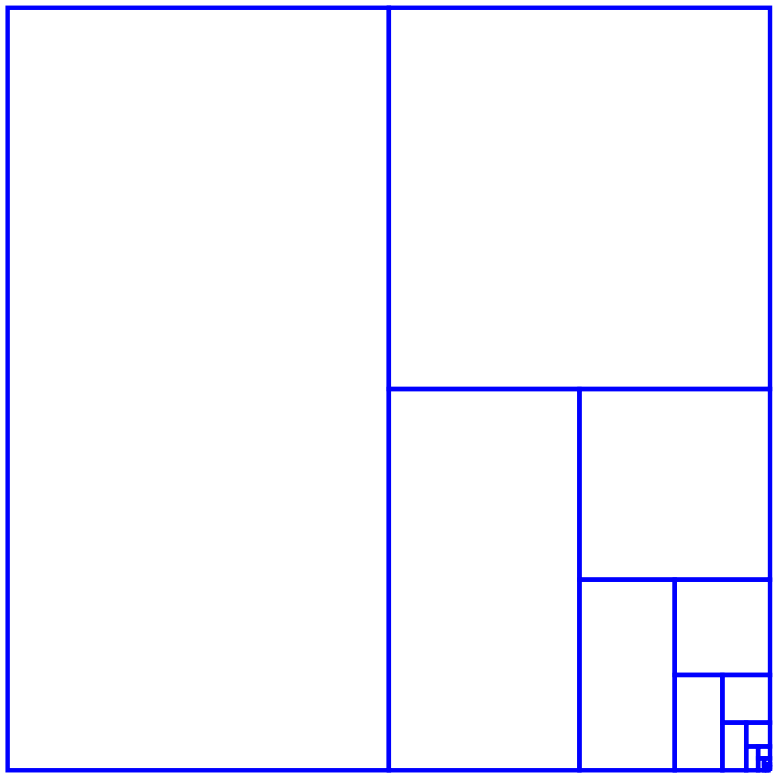
Additivity (finite) If $A = A_1 \cup \ldots \cup A_n$ then $\mu(A) \leq \sum \mu(A_i)$.
We have equality when the A_i are disjoint.

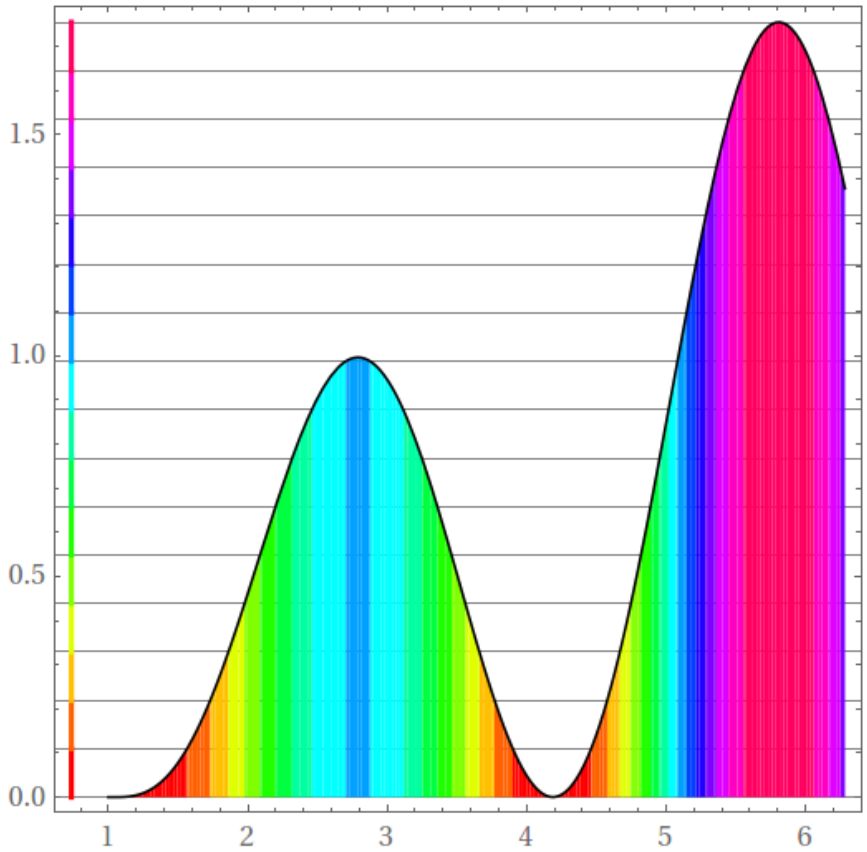
Additivity (countable) If $A = \bigcup_{i \geq 0} A_i$ then $\mu(A) \leq \sum \mu(A_i)$.
We have equality when the A_i are disjoint.

Invariance If B is congruent to A, then $\mu(A) = \mu(B)$.

The condition $A_i \cap A_j$ for $i < j$ means that the events are mutually exclusive.

No one doubts finite additivity, but countable additivity is also really quite natural: think about dividing the unit square into rectangles of size 2^{-n}.





The now standard answer to the design of such a measure was given by Henri Lebesgue in 1902 in his dissertation: to measure a region A, approximate it by lots of rectangles (but in a non-obvious way).

Theorem (Vitali 1905)

There are non-measurable sets of reals.

This requires a little group theory and the Axiom of Choice. On the other hand, Solovay has constructed universes where all sets of reals are measurable.

On the upside, there are finitely additive measures for $d = 1, 2$. Unfortunately, they fail to be unique.

Theorem (Hausdorff)

No finitely additive measures exist for $d > 2$.

The solution is quite natural: who cares about $\mathfrak{P}(\mathbb{R}^d)$?

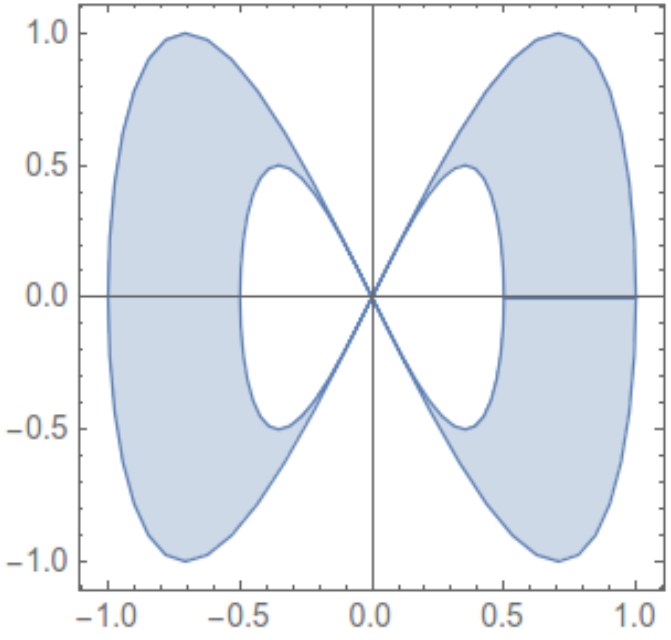
The full power set is a weird monstrosity anyway, so why not restrict the measure to civilized subsets?

The standard choice is Borel sets, sets that can be constructed from open sets by complements and countably unions. The Lebesgue measure works fine for Borel sets.

As a practical matter, you will never encounter a subset of $\mathbb{R}^d$ that is not Borel (unless you are a logician and thrive on other people's misery).

```
In[1]:= ParametricPlot[{r Cos[x], r Sin[2 x]}, {x, 0, 2 π}, {r, 1 / 2, 1}]
```

Out[1]=

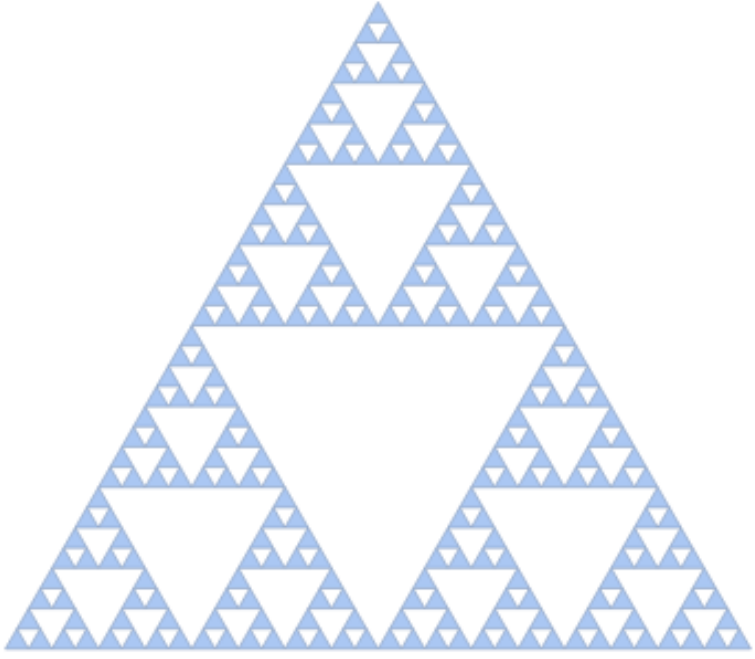



```
In[2]:= RegionMeasure[{r Cos[x], r Sin[2 x]}, {{x, 0, 2 π}, {r, 1 / 2, 1}}]
Out[2]= 2
```

Nowadays, CAS can automatically compute fairly complicated measures.

In[1083]:=
```
SierpinskiMesh[5]
```

Out[1083]=



In[1084]:=
```
Table[Area[SierpinskiMesh[n]], {n, 3}] // RootApproximant
```

Out[1084]=
$$\left\{\frac{3\sqrt{3}}{16}, \frac{9\sqrt{3}}{64}, \frac{27\sqrt{3}}{256}\right\}$$

In the countably infinite case we have basic probabilities $(p_a)_{a\in\Omega}$ such that $\sum_a p_a = 1$.

As a consequence, we can can compute $\sum_{a\in A} p_a$ for any $A \subseteq \Omega$.

In fact, we can decompose everything into atomic events:

$$\Pr[A] = \sum_{a\in A} \Pr[\{a\}]$$

This fails miserably for uncountable spaces where $\Pr[\{a\}] = 0$.

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

The axioms have several easy consequences.

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

The third axiom states additivity for two sets. By induction we immediately get full finite additivity: for any finite family of mutually exclusive events

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

How about general unions?

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

The last equation generalizes to unions of more than two terms, but is a bit clumsy to state (see the inclusion-exclusion principle in combinatorics).

Bode's Inequality:

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

Bonferroni's Inequality:

$$\Pr[A_1 \cap A_2 \cap \ldots \cap A_k] \geq \Pr[A_1] + \Pr[A_2] + \ldots + \Pr[A_k] - (k-1).$$

In the discrete case, we only need to determine the elementary probabilities $\Pr[\{a\}]$ for all $a \in \Omega$.

One possibility is to axiomatically claim certain values. For example, for a finite space Ω we might declare uniform probabilities

$$\Pr[\{a\}] = 1/|\Omega|$$

Or we could try to measure them by repeating an experiment (often) and determining frequencies:

$$\Pr[\{a\}] = \frac{\#\text{ successes}}{\#\text{ trials}}$$

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

$$\Pr[A \mid B] = \Pr[A \cap B]/\Pr[B]$$

Sometimes one can partition Ω into exclusive events $B_1, \ldots, B_k$. Then we have

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

The case $k = 2$ is often useful.

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

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

Exercise

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

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

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

Technically, in the continuous case we also need measurability of $X^{-1}(\mathbb{R}_{\geq r})$, but we won't worry.

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

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

Why Reals?

Sometimes one would like to associate outcomes other than reals with an experiment, our definition of a random variable does not allow that. The reason is that reals are just too convenient. For example, nothing stops us from computing

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

Or we could add RVs $X(a) + Y(a)$ and so on.

This will come in very handy.

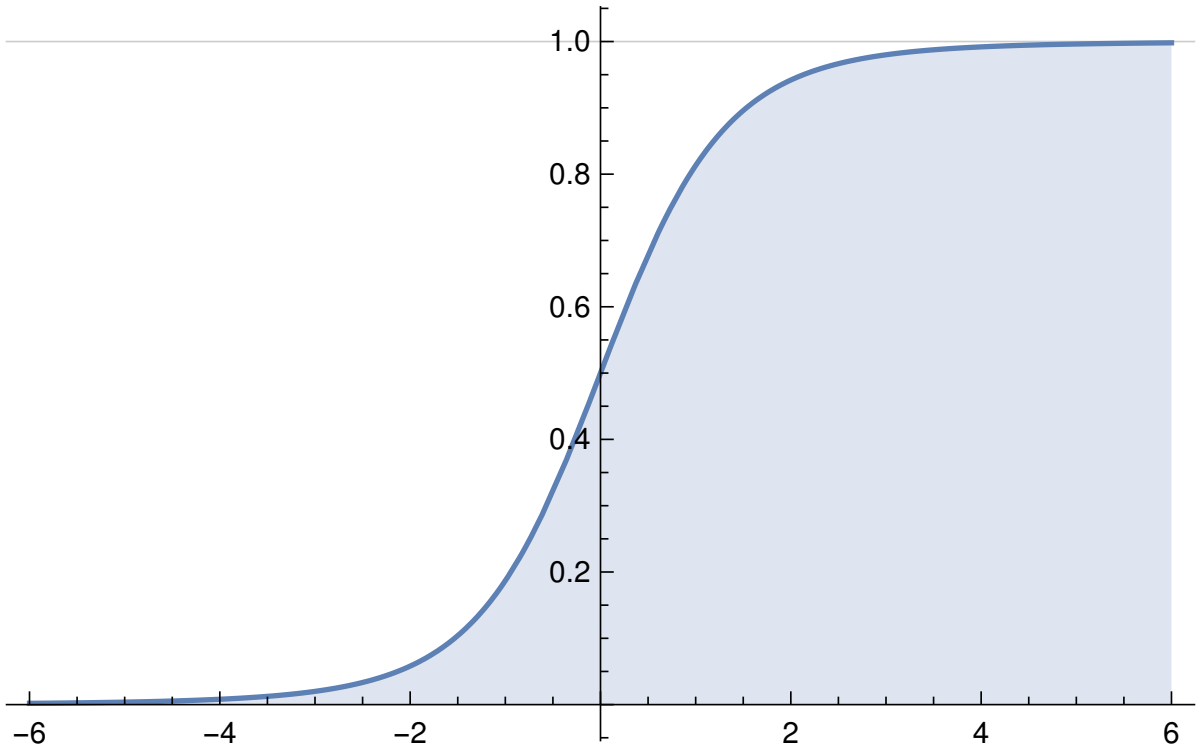
We can fake other properties by choosing our random variables appropriately.

For example, suppose we only care whether $a \in A$. This can be captured by an indicator variable

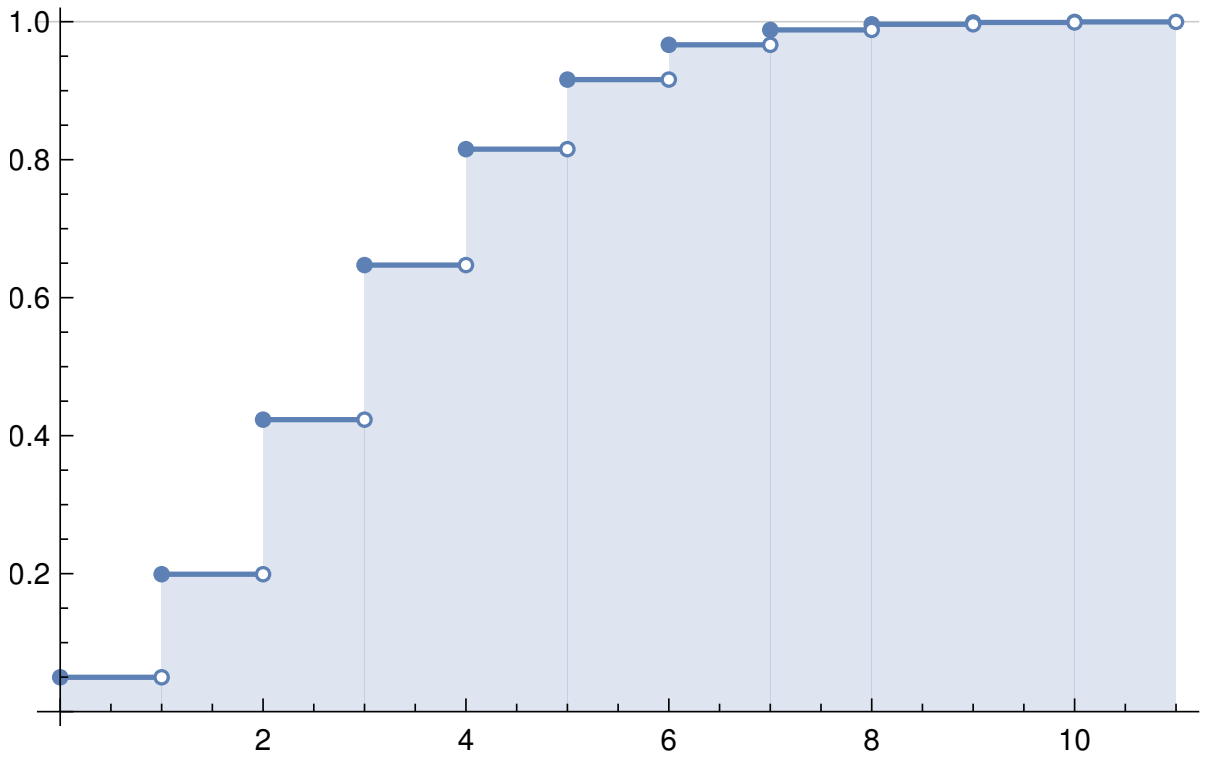
$$X(a) = \begin{cases} 1 & \text{if } a \in A, \\ 0 & \text{otherwise.} \end{cases}$$

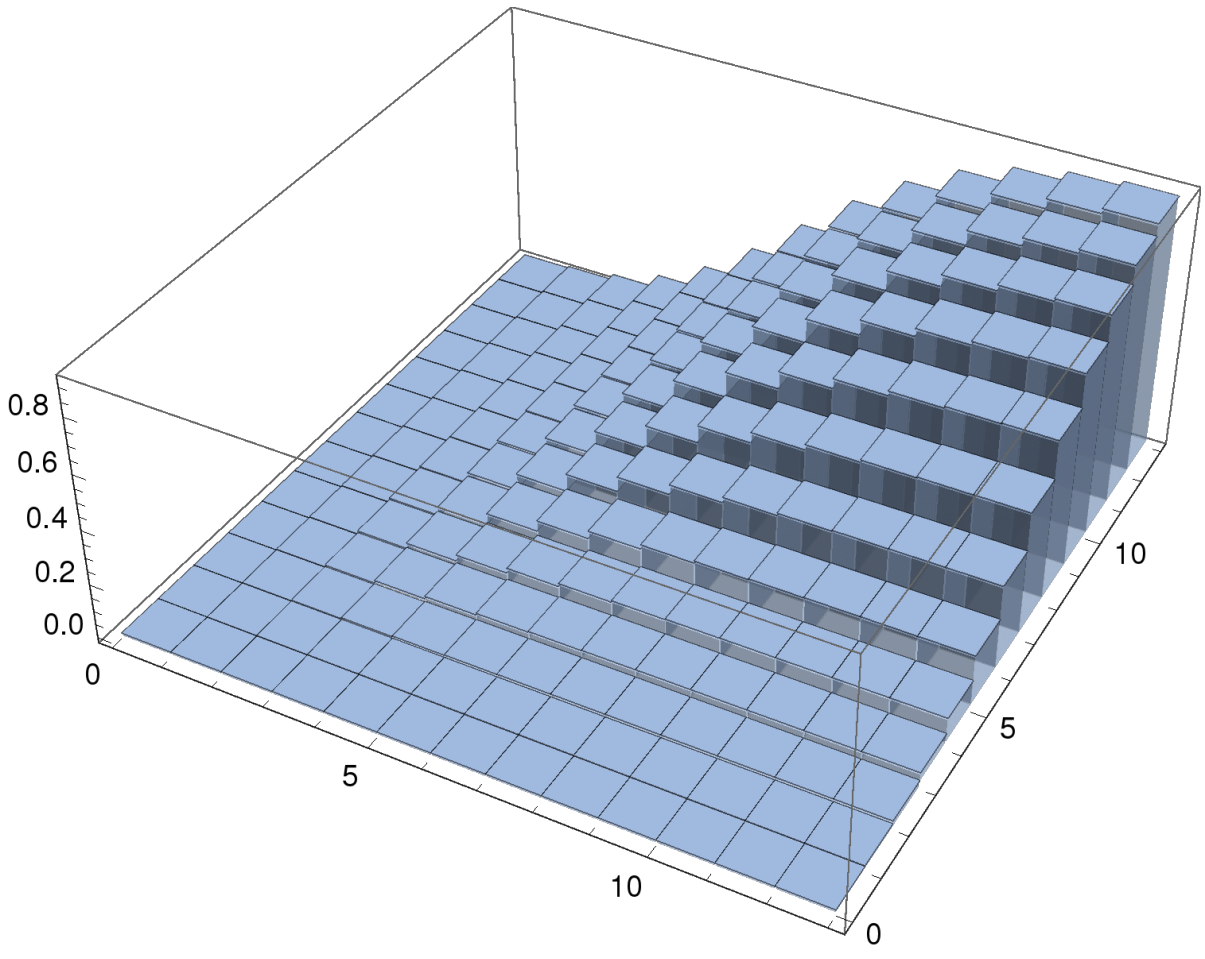
For continuous spaces we can still talk about the cumulative distribution function:

$$p(a) = \mathsf{Pr}[X \leq a]$$



In the discrete case, the cdfs increase in steps.




0.8
0.6
0.4
0.2
0.0
0
5
10
0
5
10

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

The expected value or expectation of X is

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

This is often abbreviated in slightly criminal manner to μ.

So expectation is a weighted sum, and corresponds to the intuitive notion of average.

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.

Other than the average it is also useful to now how far off the values of a random variable might be, on average.

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).

In other words,

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

Lemma

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

Lemma

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

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

Incidentally, for independent variables we have
$\mathsf{Var}[XY] = \mathsf{Var}[X] \cdot \mathsf{Var}[Y]$.

Lemma (Chebyshev's Inequality)

Suppose X has finite expectation μ and non-zero variance σ^2. Then

$$\Pr[|X - \mu| \geq c] \leq \sigma^2/c^2.$$

A continuous variable X is uniformly distributed if for some interval $[a, b] \subseteq \mathbb{R}$ we have the pdf

$$f(x) = \begin{cases} 1/(b-a) & \text{if } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathsf{E}[X] = (a+b)/2$$

$$\mathsf{Var}[X] = (b-a)^2/12$$

For a finite space Ω we similarly have

$$\Pr[X = a] = 1/|\Omega|$$

Just think about coins or dice.

Dire Warning: this does not work for countably infinite spaces.

This is the distribution of an indicator variable with

$$\mathsf{Pr}[X = 1] = p$$

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

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

Define an indicator variable X_i that is 1 if the ith repetition produces the event, and 0 otherwise and consider $X = X_1 + X_2 + \ldots + X_n$. If the probability of X_i is p then

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

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

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

If we count the number of times till heads appear we get a random variable X such that

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

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

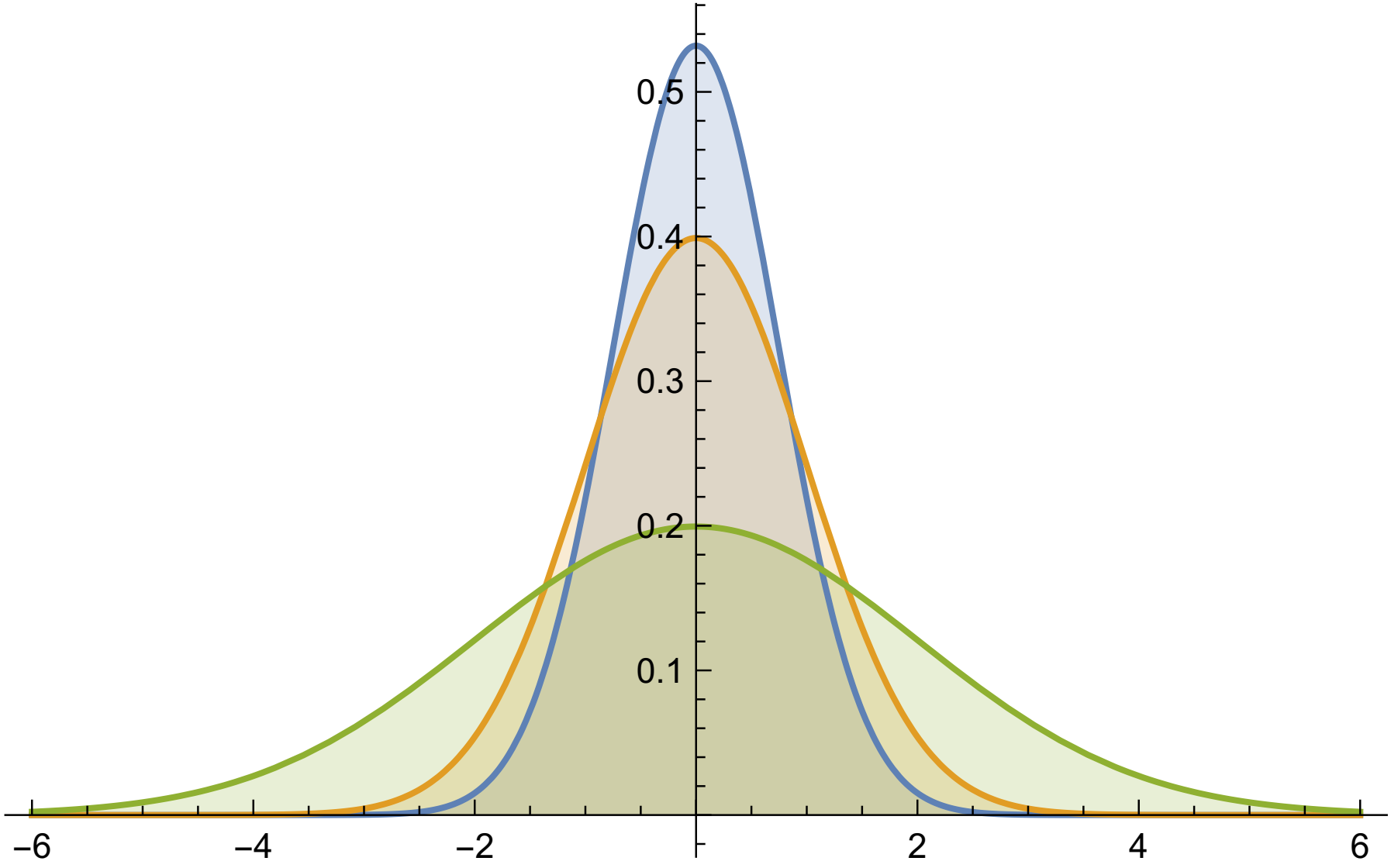
$$\mathsf{Var}[X] = (1-p)/p^2$$

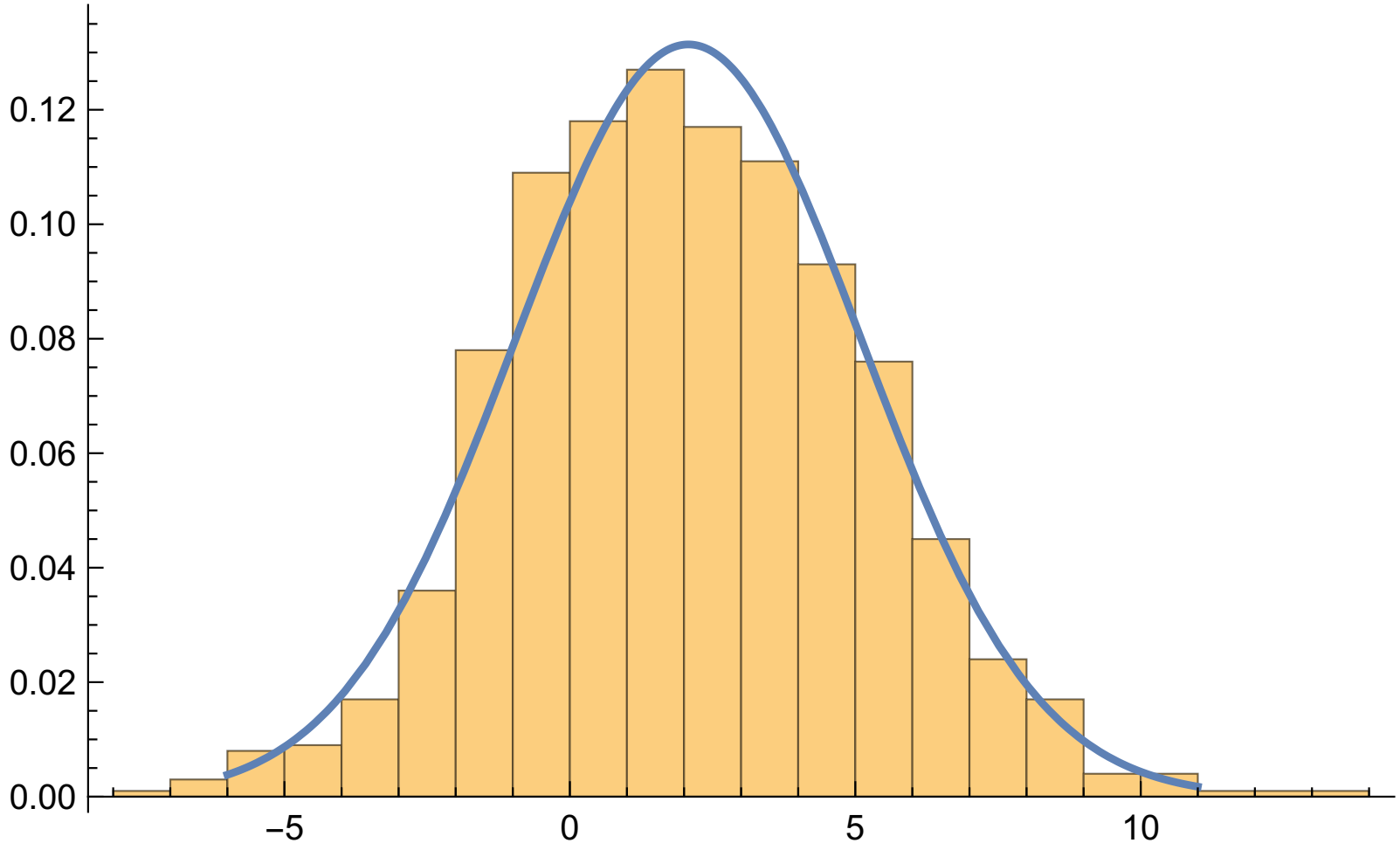
Parameters μ and σ.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}$$

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

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


0.5
0.4
0.3
0.2
0.1
−6
−4
−2
2
4
6



A lot of people spend a lot of time trying to match outcomes of experiments against a normal distribution.