

Introduction to Bayesian Methods - 1

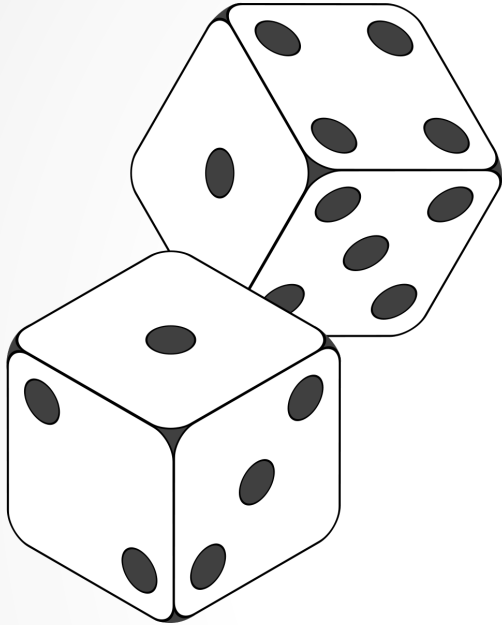
Edoardo Milotti

Università di Trieste and INFN-Sezione di Trieste

Point your browser to

<http://wwwusers.ts.infn.it/~milotti/Didattica/Bayes/2014-MiBi/Bayes.html>

The nature of probabilities



In a dice throwing game one defines probabilities of different events by counting the outcomes

Examples:

- with one die, the probability of getting a 4 is $1/6$
- with two dice, the probability of getting two 4's is $1/36$
- with two dice, the probability of getting one 4 AND one 5 is $1/18$
- with two dice, the probability of getting one 4 OR one 5 is ??

Die 1	Die 2	# of outcomes
4	Not 4 or 5	4
5	Not 4 or 5	4
Not 4 or 5	4	4
Not 4 or 5	5	4
4	4	1
4	5	1
5	4	1
5	5	1
		Total: 20

NB, if 4 and 5 were independent, we would have


$$P(4 \text{ OR } 5) = P(4) + P(5) = 1/3 + 1/3 = 2/3$$




$$p = \frac{20}{36} = \frac{5}{9} < \frac{2}{3}$$

Die 1	Die 2	# of outcomes
4	Not 4 or 5	4
5	Not 4 or 5	4
Not 4 or 5	4	4
Not 4 or 5	5	4
4	4	1
4	5	1
5	4	1
5	5	1
		Total: 20

Outcomes = elementary events



Composite events contain many elementary events



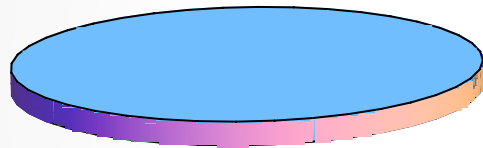
We usually assume that elementary events are all equally likely.

This is not true for biased dice.

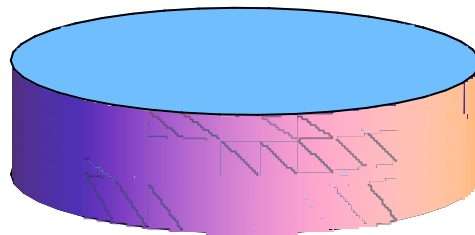


Now, consider “coins” with different aspect ratio r

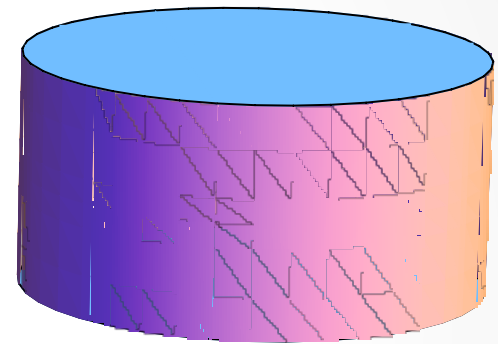
(aspect ratio = thickness/diameter)



$r = 0.05$



$r=0.25$



$r = 0.5$

How do these coins land on heads, tails, sides? When is the probability of landing on the side equal to the probability of landing on heads or tails?

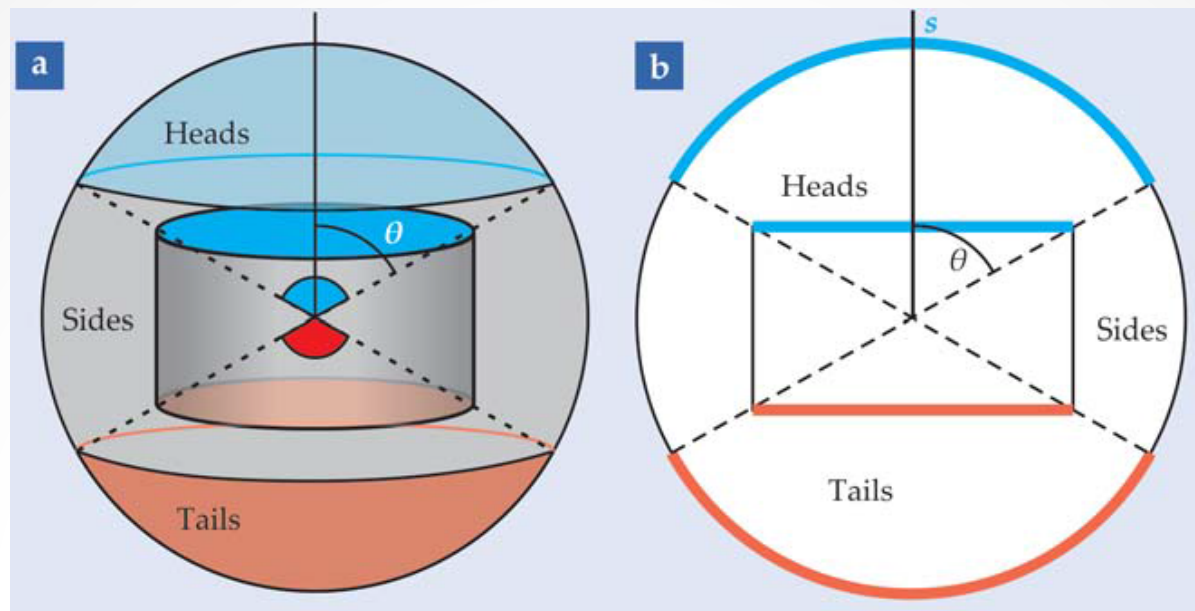
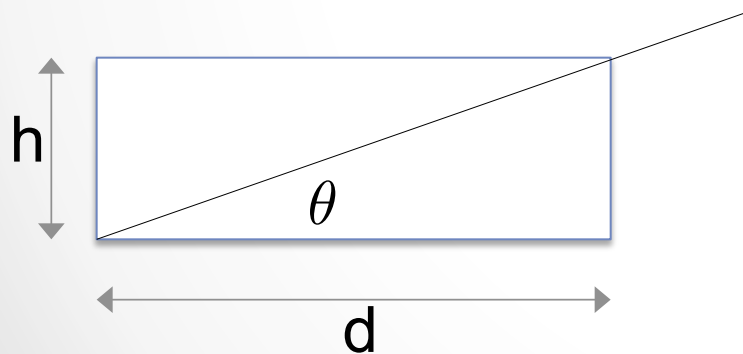


figure from Mahadevan and Yong,
"Probability, physics, and the coin toss",
Phys. Today, July 2011, pp. 66-67

a. Von Neumann's answer: consider solid angles subtended by heads, tails, sides



$$\Omega_{\text{heads}} = \Omega_{\text{tails}} = \Omega_{\text{sides}} = 4\pi/3$$

$$\Rightarrow 2\pi(1 - \cos \theta_0) = 4\pi/3$$

$$\Rightarrow \frac{h}{\sqrt{h^2 + d^2}} = \frac{r}{\sqrt{r^2 + 1}} = 2/3$$

$$\Rightarrow r = 1/2\sqrt{2}$$

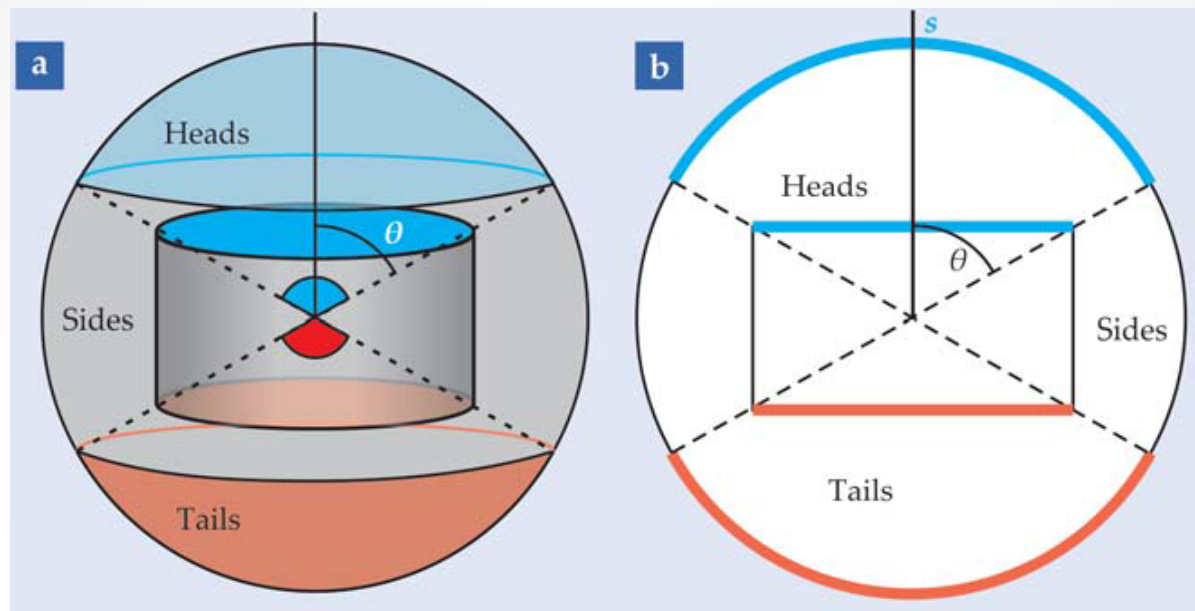


figure from Mahadevan and Yong,
 “Probability, physics, and the coin toss”,
 Phys. Today, July 2011, pp. 66-67

b. alternative answer: consider *angles* subtended by heads, tails, sides (rotation about axis through center of coin, and parallel to faces)

$$\theta_{\text{heads}} = \theta_{\text{tails}} = \theta_{\text{sides}} = \pi/3$$

$$\Rightarrow \cos \theta_0 = 1/2$$

$$\Rightarrow \frac{h}{\sqrt{h^2 + d^2}} = \frac{r}{\sqrt{r^2 + 1}} = 1/2$$

$$\Rightarrow r = 1/\sqrt{3}$$

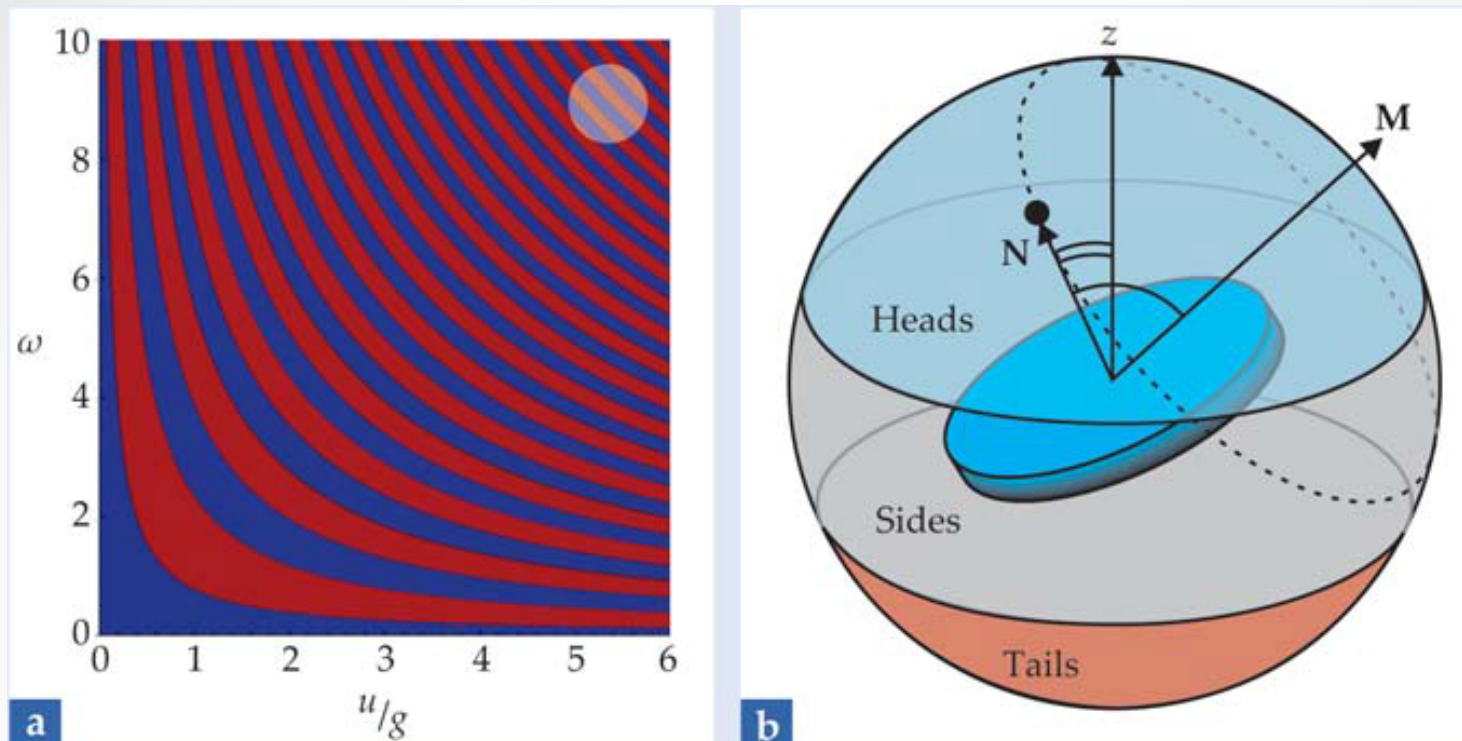
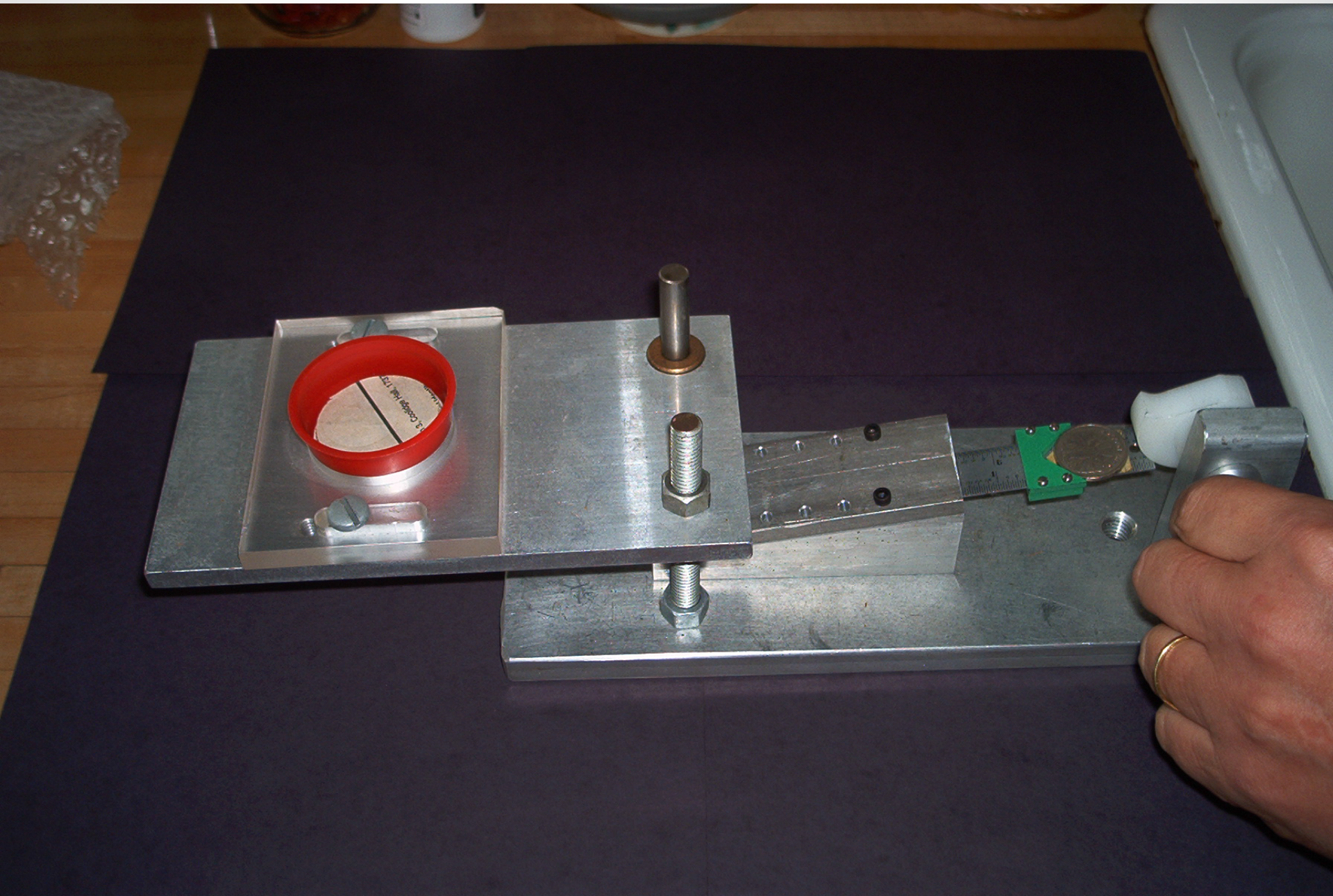


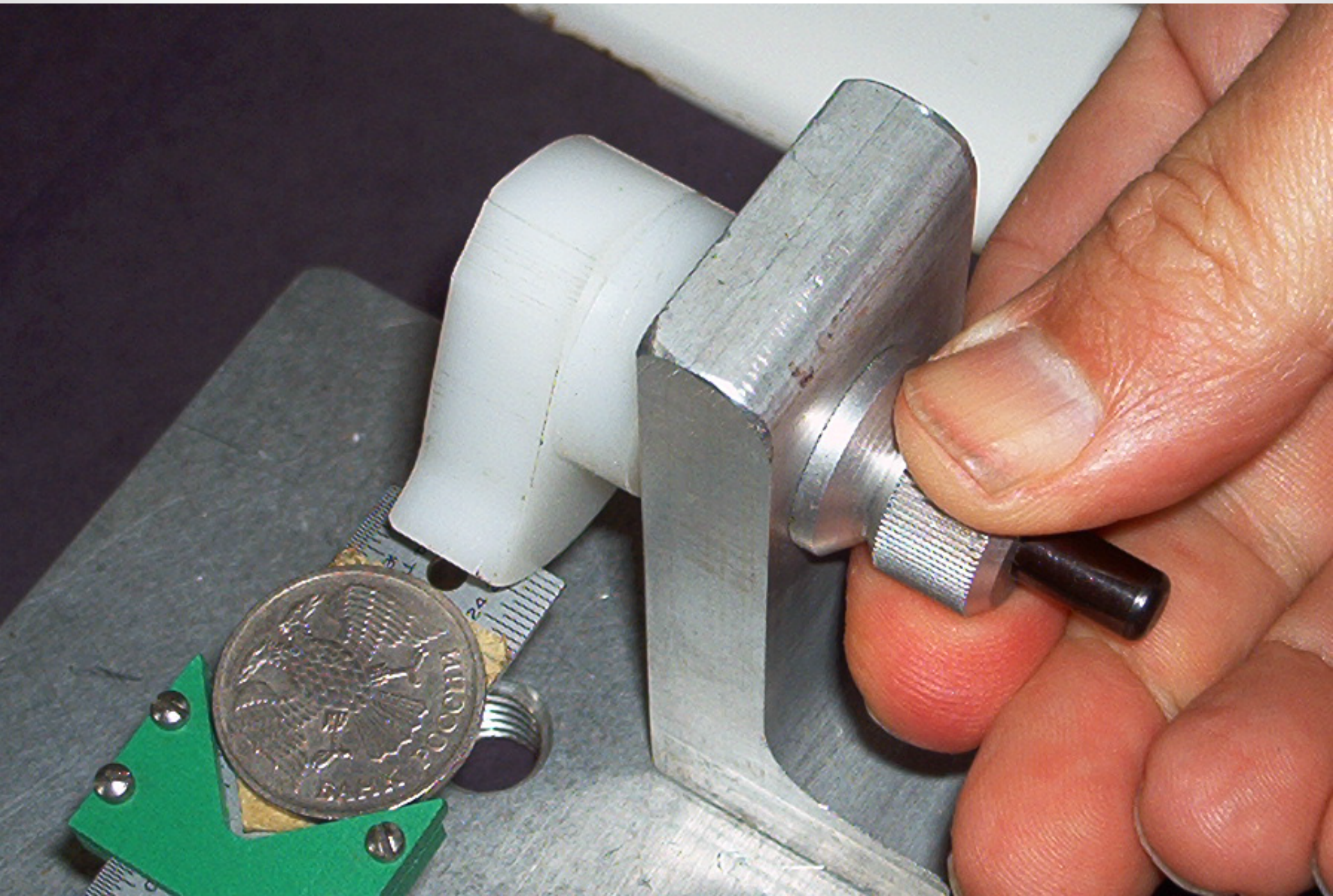
figure from Mahadevan and Yong,
 "Probability, physics, and the coin toss",
 Phys. Today, July 2011, pp. 66-67

In 1986 J. B. Keller analyzed the infinitely thin coin and found that coin toss is not random for finite rotation speed and vertical speed (rotation axis as in previous case b)

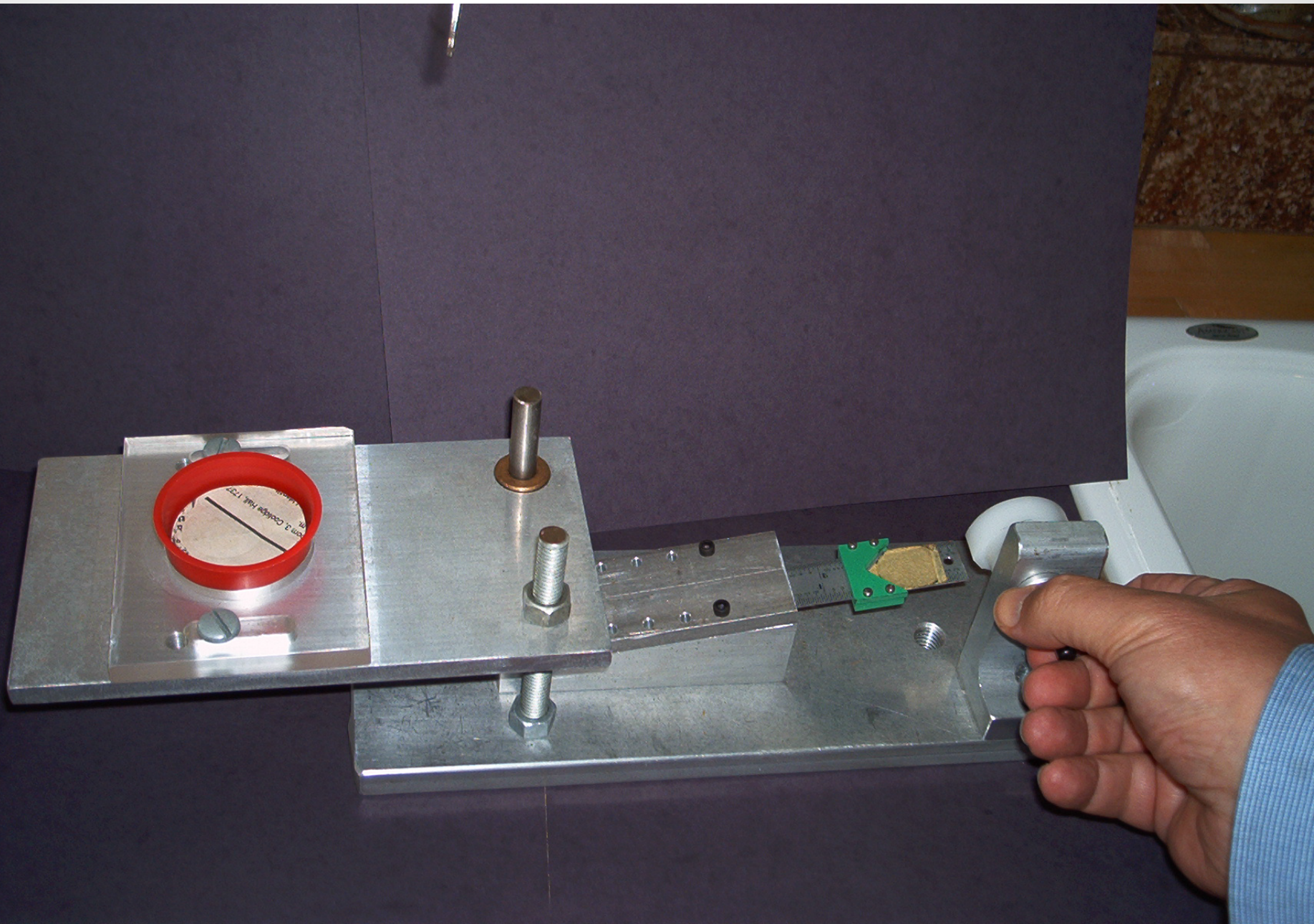
Coin tossing machine (Diaconis, Holmes and Montgomery 2007)



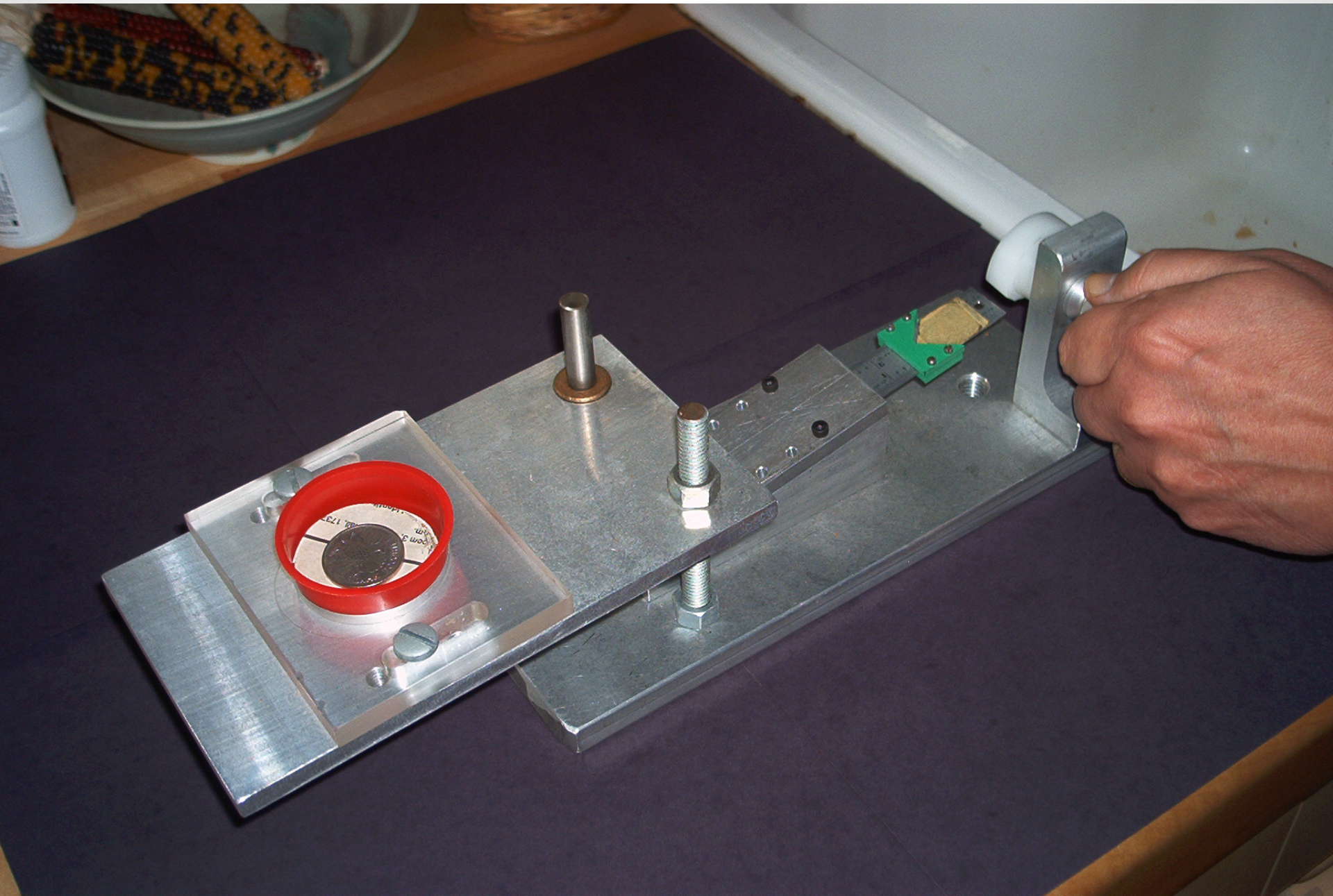
Coin tossing machine (Diaconis, Holmes and Montgomery 2007)



Coin tossing machine (P. Diaconis, S. Holmes and R. Montgomery 2007)



Coin tossing machine (Diaconis, Holmes and Montgomery 2007)



... Coin-tossing is a basic example of a random phenomenon. However, naturally tossed coins obey the laws of mechanics (we neglect air resistance) and their flight is determined by their initial conditions. Figure 1 a-d shows a coin-tossing machine. The coin is placed on a spring, the spring released by a ratchet, the coin flips up doing a natural spin and lands in the cup. **With careful adjustment, the coin started heads up always lands heads up – one hundred percent of the time.** We conclude that coin-tossing is ‘physics’ not ‘random’. ...

(Diaconis, Holmes and Montgomery, “Dynamical bias in the coin toss”, *SIAM Rev.* **49** (2007) 211)

Therefore, the assumed randomness of coin toss – and in general, of complex mechanical processes – is related to the difficulty in determining the outcome, both because of the complex and often unknown dynamics, and because of the uncertain initial conditions.

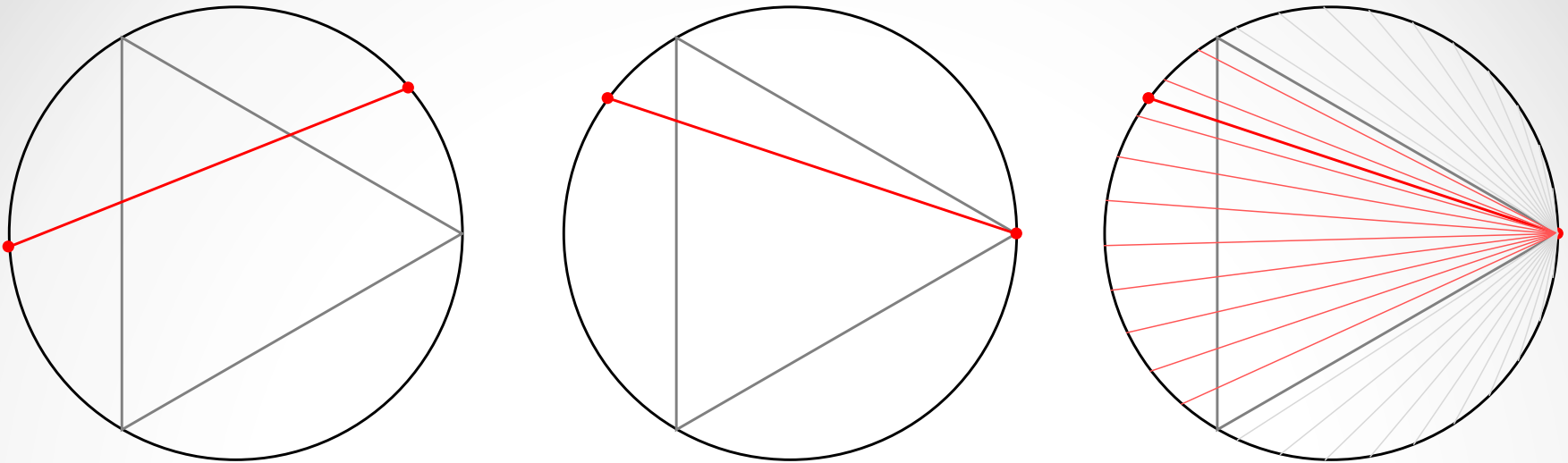
Thus – at least in this case – probabilities are a measure of our own ignorance rather than an intrinsic property of the physical system.

Bertrand's paradox and the ambiguities of probability models

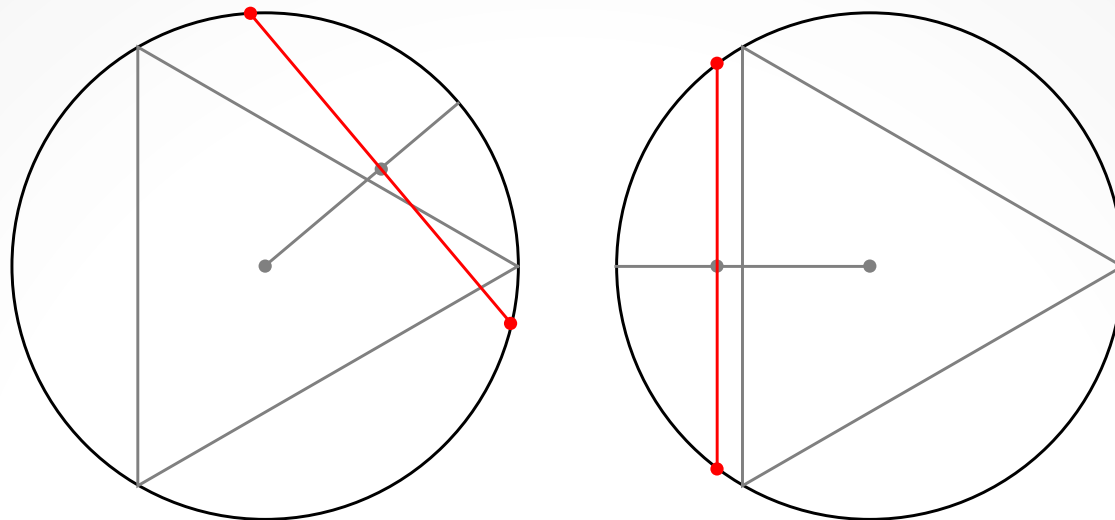
Bertrand's paradox goes as follows:

“consider an equilateral triangle inscribed inside a circle, and suppose that a chord is chosen at random. What is the probability that the chord is longer than a side of the triangle?”

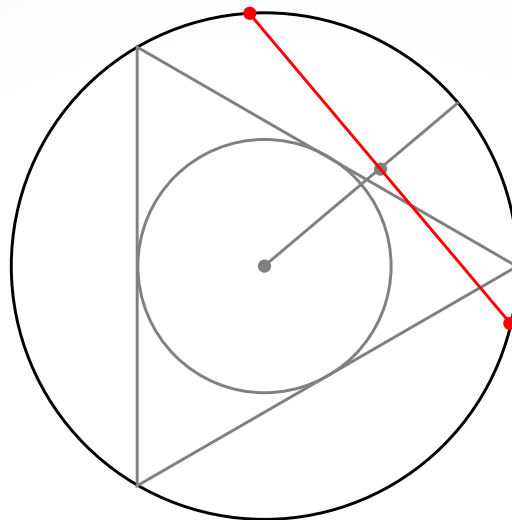
(Bertrand, 1889)



Solution: we take two random points on the circle (radius R), then we rotate the circle so that one of the two points coincides with one of the vertices of the inscribed triangle. Thus a random chord is equivalent to taking the first point that defines the chord as one vertex of the triangle while the other is taken “at random” on the circle. Here “at random” means that it is uniformly distributed on the circumference. Then only those chords that cross the opposite side of the triangle are actually longer than each side. Since the subtended arc is $1/3$ of the circumference, **the probability of drawing a random chord that is longer than one side of the triangle is $1/3$.**



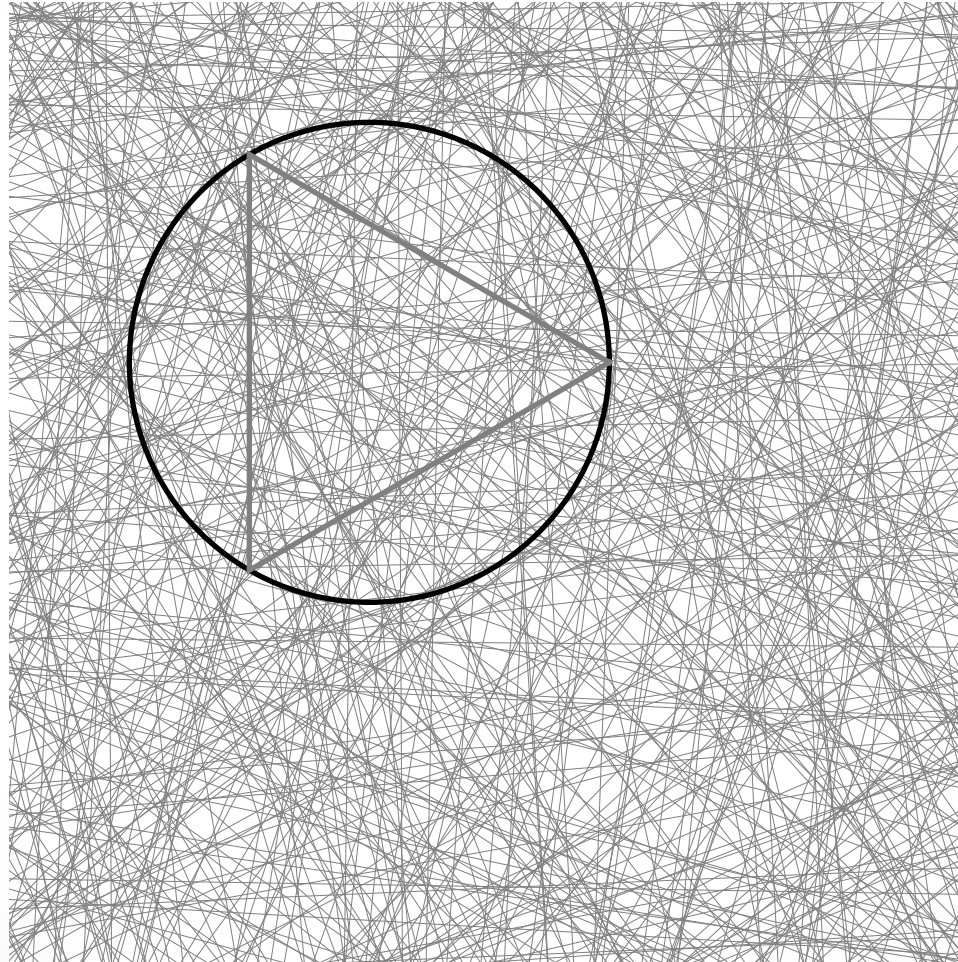
Solution 2: we take first a random radius, and next we choose a random point on this random radius. Then, we take the chord through this point and perpendicular to the radius. When we rotate the triangle so that the radius is perpendicular to one of the sides, we see that half of the points give chords longer than one side of the triangle, therefore **the probability is $1/2$** .

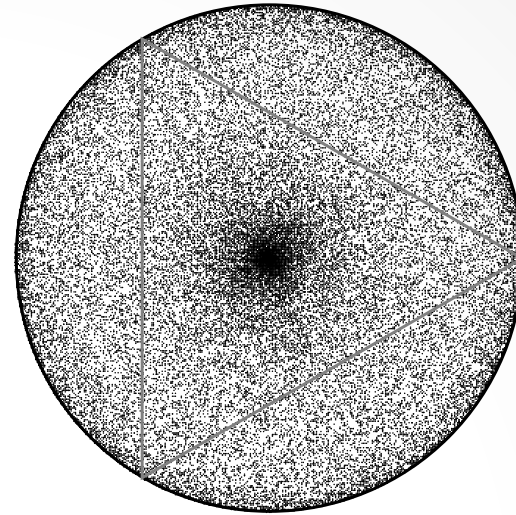
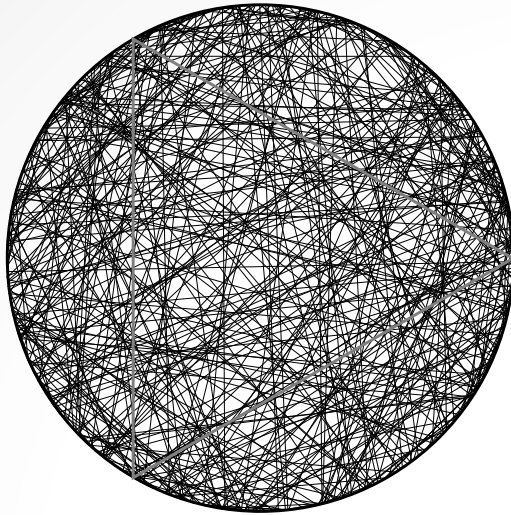


Solution 3: we take the chord midpoints located inside the circle inscribed in the triangle, and we obtain chords that are longer than one side of the triangle. Since the ratio of the areas of the two circles is $1/4$, we find that now **the probability of drawing a long chord is just $1/4$.**

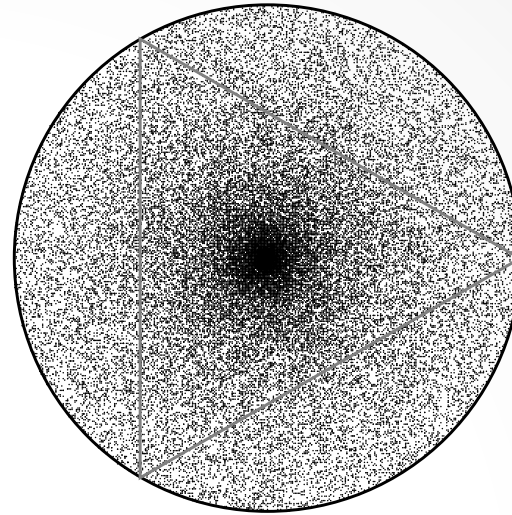
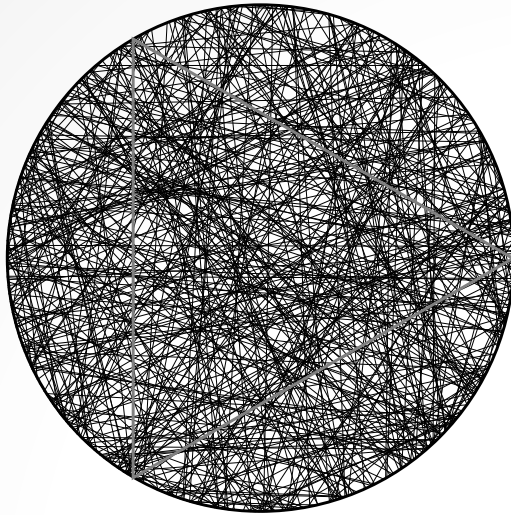
At least 3 different “solutions”: which one is correct, and why?

Now we widen the scope of the problem and we consider the distribution of chords in the plane





Distribution 1: distribution of chords (left panel) and of midpoints (right panel) in the first solution of Bertrand's paradox (the left panel shows 400 chords, the right panel shows 100000 midpoints).



Distribution 2: Distribution of chords (left panel) and of midpoints (right panel) in the second solution of Bertrand's paradox (the left panel shows 400 chords, the right panel shows 100000 midpoints).

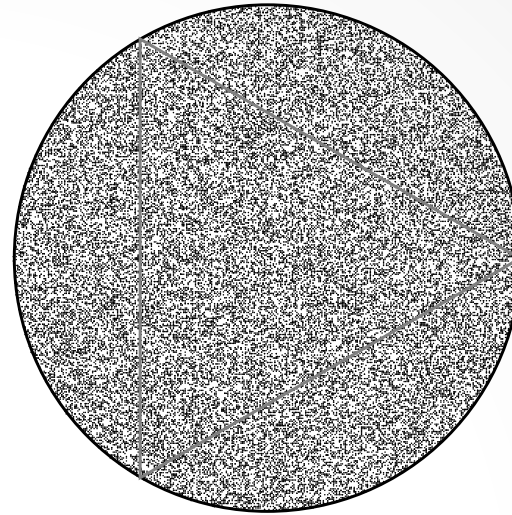
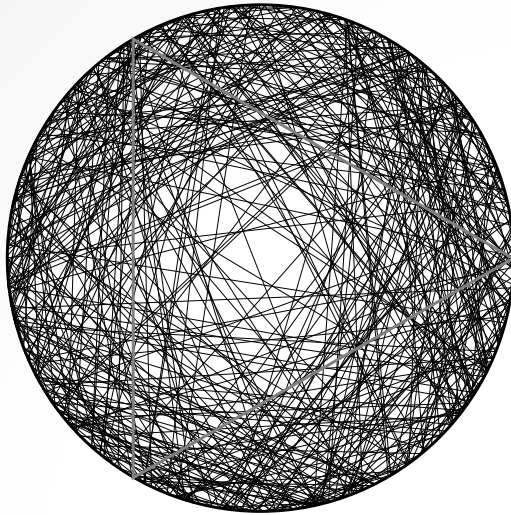
In this case it is very easy to find the radial density function of chord centers, since here we take first a random radius, and next we choose a random point (the center) on this random radius.

Therefore

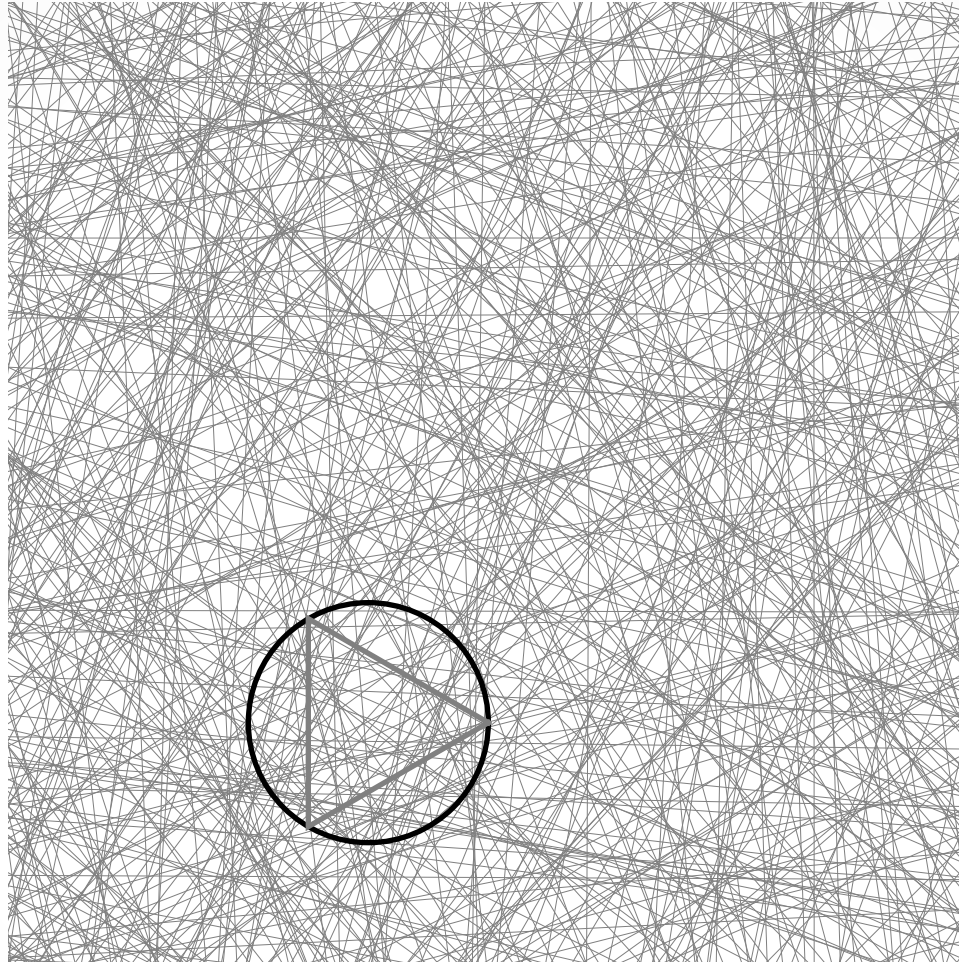
$$f(r, \theta) = f(r) = C/r$$

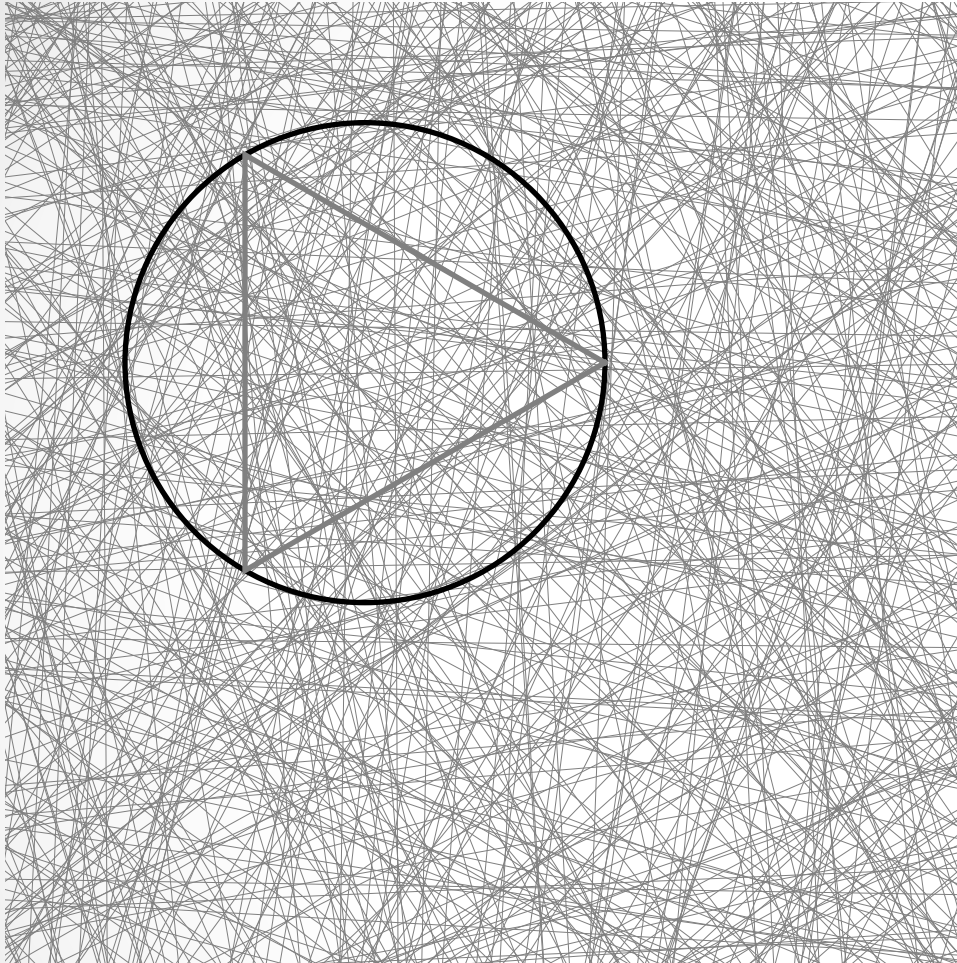
$$\Rightarrow \quad (\text{normalization}) \quad 1 = \int_C f(r) 2\pi r dr = 2\pi C R$$

$$\Rightarrow \quad f(r) = \frac{1}{2\pi r R}$$



Distribution 3: Distribution of chords (left panel) and of midpoints (right panel) in the third solution of Bertrand's paradox (the left panel shows 400 chords, the right panel shows 100000 midpoints). Notice that while the distribution of midpoints is uniform, the distribution of the resulting chords is distinctly non-uniform.





Hidden assumptions (Jaynes):

- rotational invariance
- scale invariance
- translational invariance

Now let

$$f(r, \theta)$$

be the probability density
of chord centers

Rotational invariance

In a reference frame which is at an angle α with respect to the original frame, i.e., the new angle $\theta' = \theta - \alpha$, the distribution of centers is given by a different distribution function $g(r, \theta') = g(r, \theta - \alpha)$. Since we require rotational invariance

$$f(r, \theta) = g(r, \theta - \alpha)$$

with the condition $g(r, \theta)|_{\alpha=0} = f(r, \theta)$, and this must hold for every angle α , so the only possibility is that there is no dependence on θ , and $f(r, \theta) = g(r, \theta) = f(r)$.

Scale invariance

When we consider a circle with radius R , the normalization of the distribution $f(r)$ is given by the integral

$$\int_0^{2\pi} \int_0^R f(r) r dr d\theta = 2\pi \int_0^R f(r) r dr = 1$$

The same distribution induces a similar distribution $h(r)$ on a smaller concentric circle with radius aR ($0 < a < 1$), such that $h(r)$ is proportional to $f(r)$, i.e., $h(r) = Kf(r)$, and

$$1 = 2\pi \int_0^{aR} h(u) u du = 2\pi \int_0^{aR} Kf(u) u du = 2\pi K \int_0^{aR} f(u) u du$$

i.e.,

$$K^{-1} = 2\pi \int_0^{aR} f(u) u du$$

and

$$f(r) = 2\pi h(r) \int_0^{aR} f(u) u du$$

inside the smaller circle.

Now we invoke the assumed scale invariance: the probability of finding a center in an annulus with radii r and $r + dr$ in the original circle, must be equal to the probability of finding a center in the scaled down annulus,

$$h(ar)(ar)d(ar) = f(r)rdr$$

and therefore

$$a^2 h(ar) = f(r)$$

Equation

$$a^2 h(ar) = f(r)$$

can also be rewritten in the form

$$h(r) = \frac{1}{a^2} f\left(\frac{r}{a}\right) \quad (1)$$

and inserting this into equation

$$f(r) = 2\pi h(r) \int_0^{aR} f(u) u du$$

we find

$$a^2 f(ar) = 2\pi f(r) \int_0^{aR} f(u) u du \quad (2)$$

We solve equation

$$a^2 f(ar) = 2\pi f(r) \int_0^{aR} f(u) u du$$

taking first its derivative with respect to a : the relation that we find must hold for all a 's, and therefore also for $a = 1$ (no scaling), and we find the differential equation

$$rf'(r) = \left(2\pi R^2 f(R) - 2\right) f(r)$$

i.e.,

$$rf'(r) = (q - 2)f(r)$$

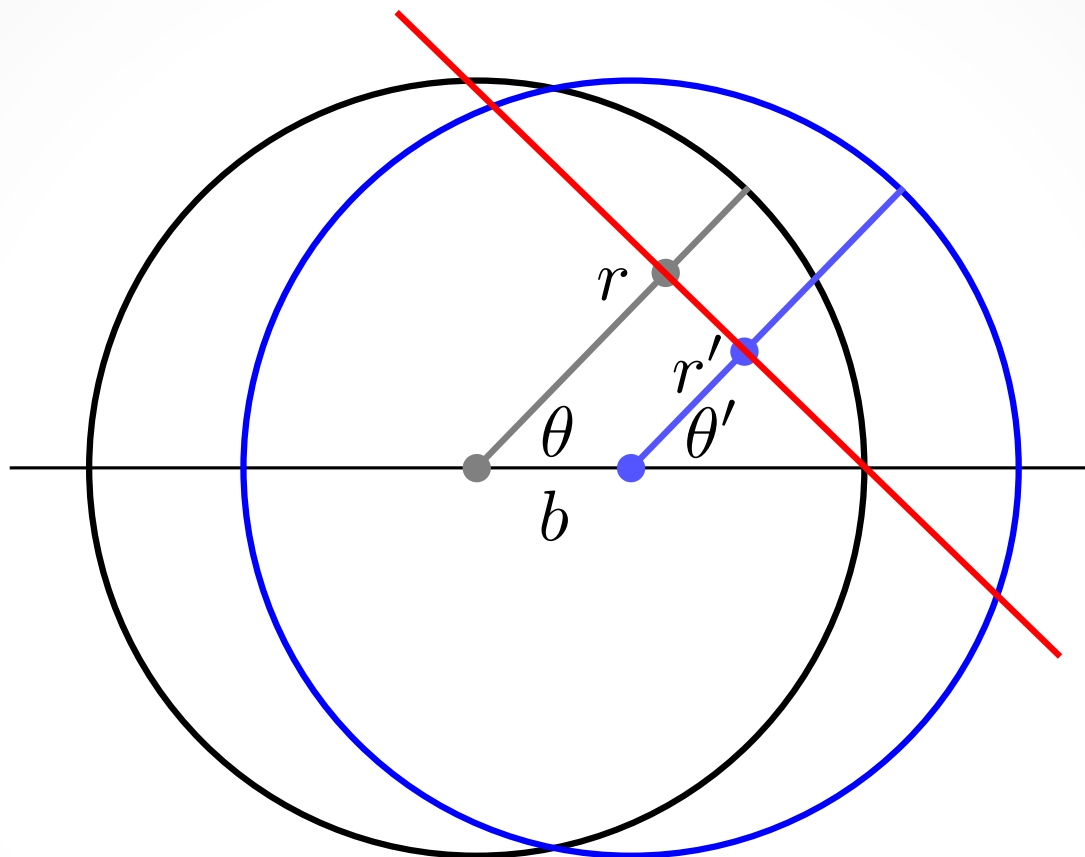
where the constant $q = 2\pi R^2 f(R)$ is unknown. However, we can still solve the equation and find

$$f(r) = Ar^{q-2}$$

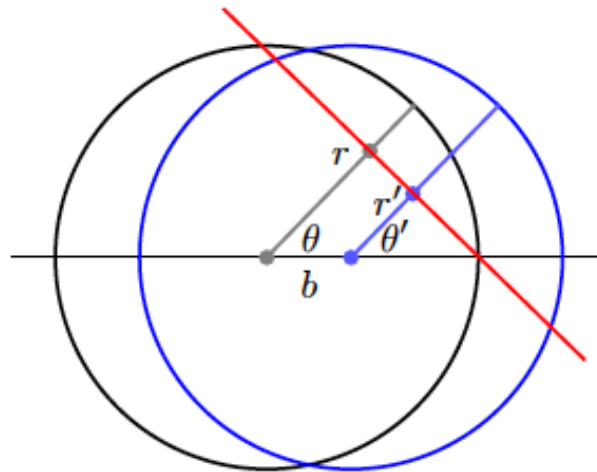
The constant A is easy to find from the normalization condition: $A = q/2\pi R^q$, and therefore

$$f(r) = \frac{qr^{q-2}}{2\pi R^q}$$

Translational invariance



Geometrical construction for the discussion of translational invariance. The original circle (black) is crossed by a straight line (red) which defines the chord. The translated circle is shown in blue.



This circle is displaced by the amount b , and the new radius and angle that define the midpoint of the chord are

$$r' = |r - b \cos \theta|$$

$$\theta' = \theta \quad (\text{if } r \geq b \cos \theta) \quad \text{or} \quad \theta' = \theta + \pi \quad (\text{if } r < b \cos \theta)$$

Now consider a region Γ surrounding the midpoint in the original circle, which is transformed into a region Γ' by the translation. The probability of finding a chord with the midpoint in the region Γ is

$$\int_{\Gamma} f(r) r dr d\theta = \int_{\Gamma} \frac{q r^{q-1}}{2\pi R^q} dr d\theta = \frac{q}{2\pi R^q} \int_{\Gamma} r^{q-1} dr d\theta$$

Likewise, the same probability for the translated circle is

$$\frac{q}{2\pi R^q} \int_{\Gamma'} (r')^{q-1} dr' d\theta' = \frac{q}{2\pi R^q} \int_{\Gamma} |r - b \cos \theta|^{q-1} dr d\theta \quad (3)$$

where the Jacobian of the transformation is 1. Equating these expressions, we see that the integrand must be a constant, and therefore $q = 1$, and

$$f(r, \theta) = \frac{1}{2\pi R r} \quad (r \leq R; \quad 0 \leq \theta < 2\pi)$$

Using this distribution, we find that the probability of finding a midpoint inside the circle with radius $R/2$ – i.e., the probability of finding a chord longer than the side of the triangle in Bertrand's paradox – is

$$\int_0^{2\pi} d\theta \int_0^{R/2} f(r, \theta) r dr = 2\pi \int_0^{R/2} \frac{1}{2\pi R r} r dr = \frac{1}{2}$$

which corresponds to the second alternative in the previous discussion of Bertrand's paradox.

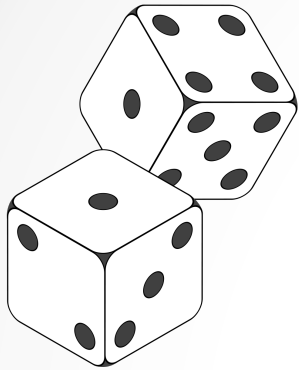
Lesson drawn from Bertrand's paradox:

probability models depend on physical assumptions, they are not God-given. We define the elementary events on the basis of real-world constraints, derived from our own experience.

Probabilities as a measure of “reasonable expectation”, and their relationship with statistical inference. (Cox, 1946)

- We construct – explicitly or implicitly – probabilistic theoretical models to understand measurements (the most common such model is the Gaussian model)
- We utilize the empirical probability distributions to infer the parameter values of physical models

Example: a population of dice



- We have a bag full of dice: we extract one of them, we throw it and we record the result.
- We replace the die in the bag, we mix it and we extract another die.
- We repeat the procedure again and again, and we count the number of times each face shows up.
- We find that in N throws each face shows up about $N/6$ times.

Can we conclude that the population of dice is “honest”?

No, we could have obtained the same result, e.g., with a population of dice where each die has the same value on every face, and such that there are $N/6$ dice of each kind (a different probability model)

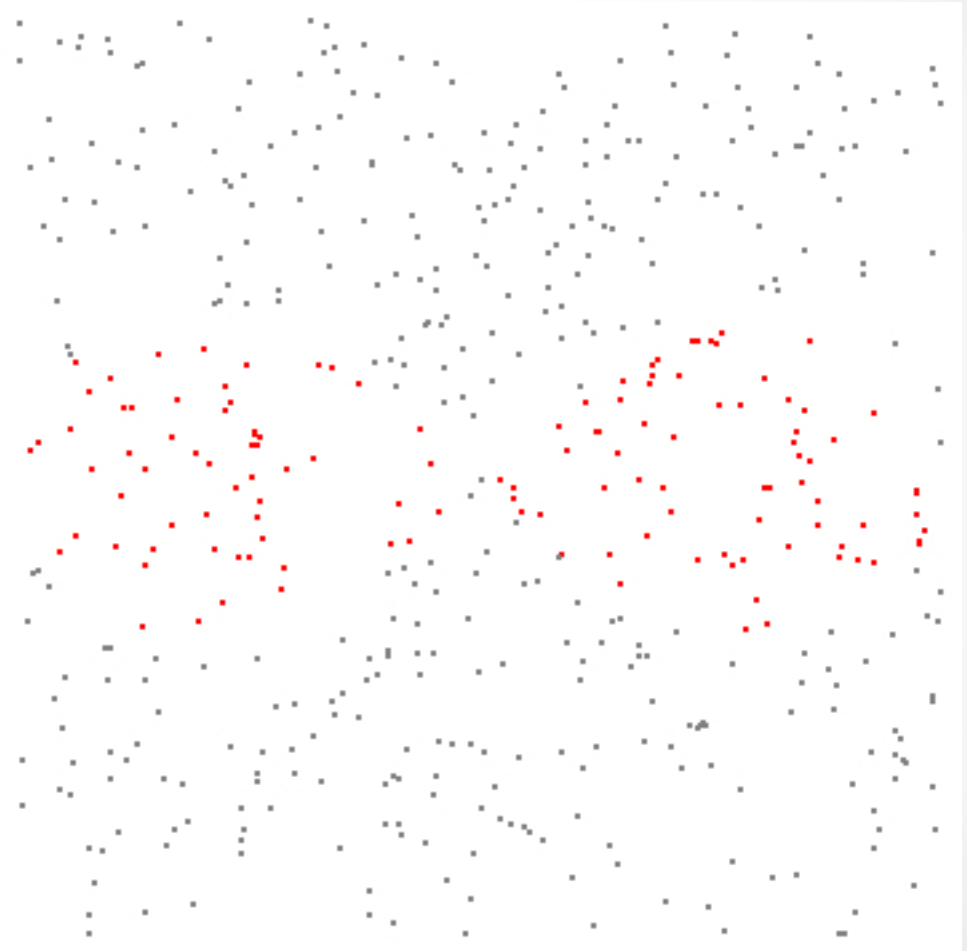
Here it certainly makes sense to consider the distribution of outcomes as a potential physical property of the population of dice, and the averages as properties of the population rather than of the individual dice.

However, could the same hold true when we measure a physical parameter, say the alpha constant of QED? *Usually* the probability models used in physics are uniquely determined by their transformation properties, just like in Jaynes' solution of Bertrand's paradox. Does it make sense to consider a “distribution” of the values of alpha?

What if we “measure” a mathematical constant instead of a physical parameter?

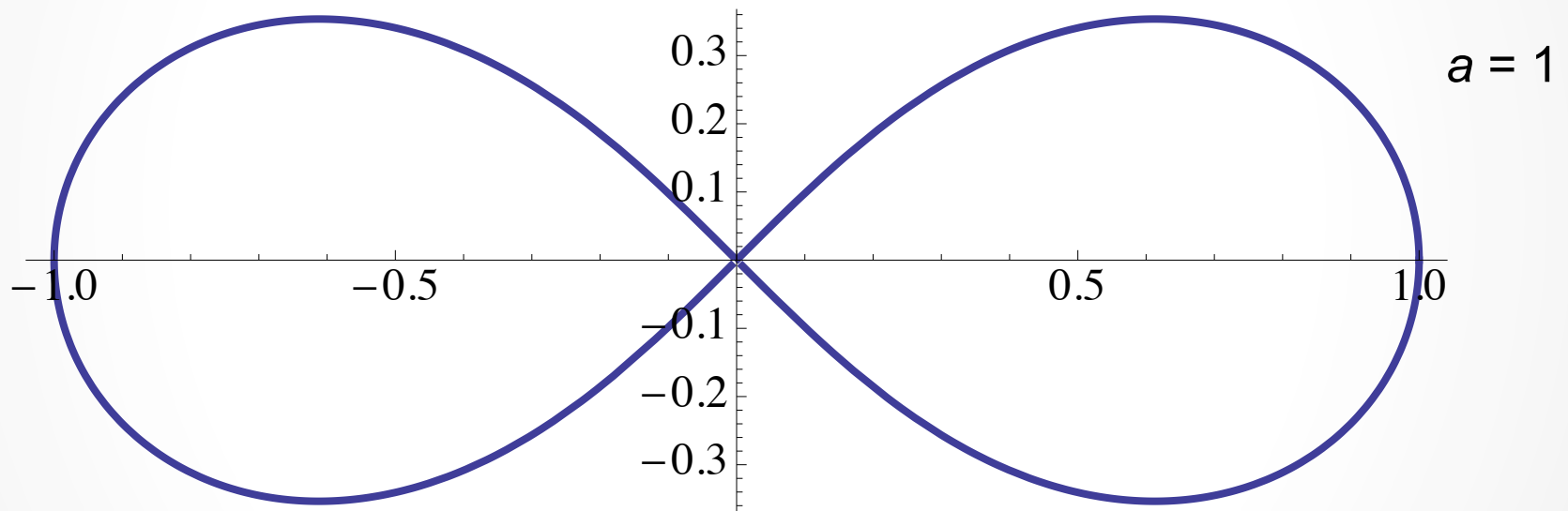
Example:

area of Bernoulli's lemniscate obtained with a Monte Carlo simulation.



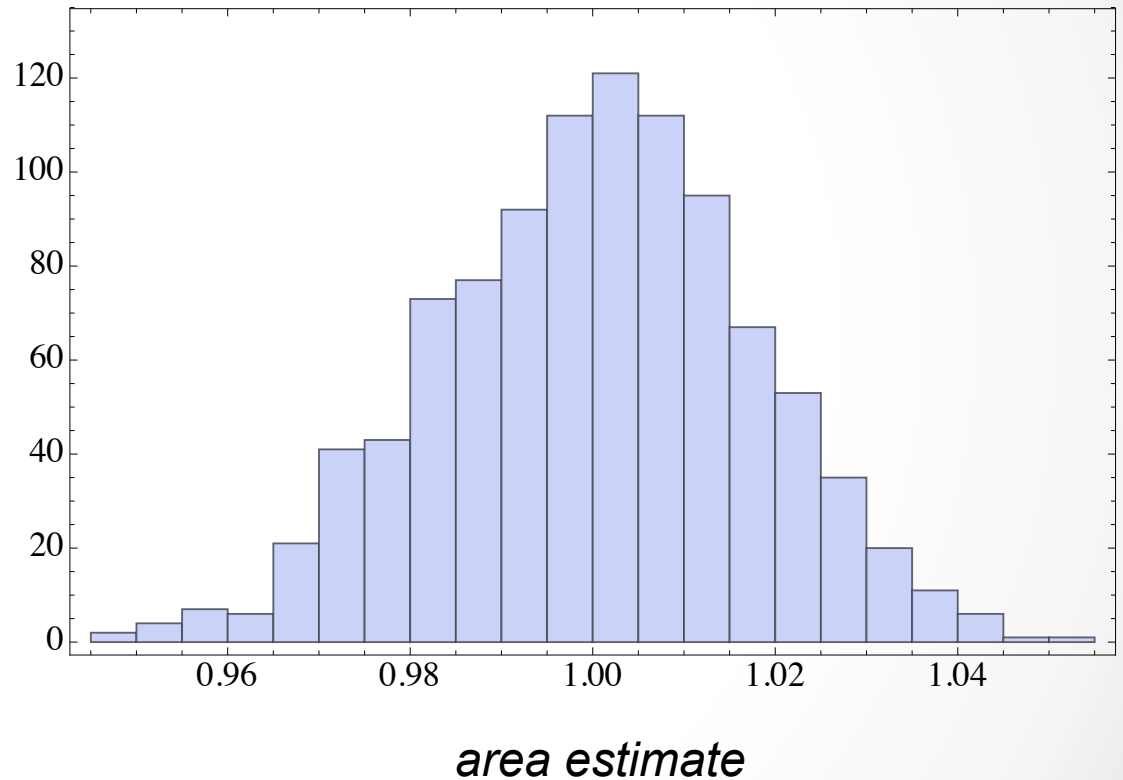
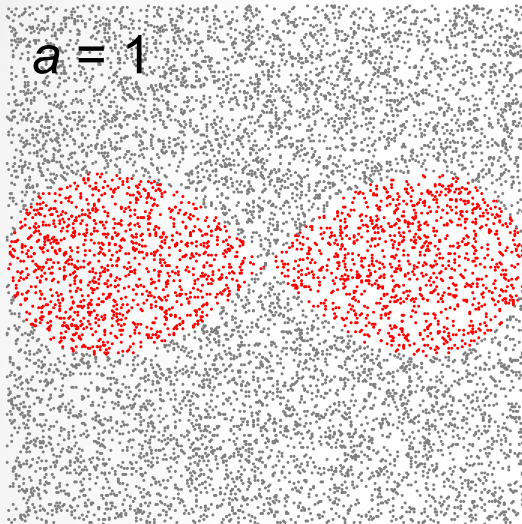
Parametric equation of Bernoulli's lemniscate

$$r = a\sqrt{\cos 2\theta}$$

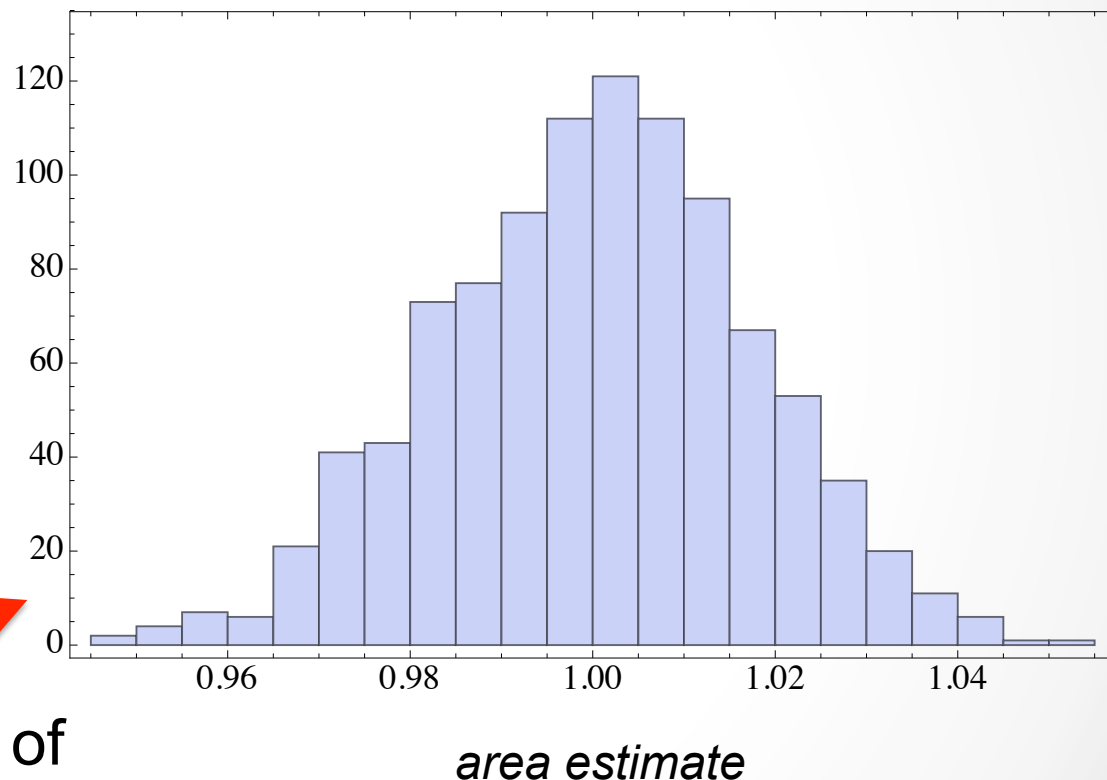
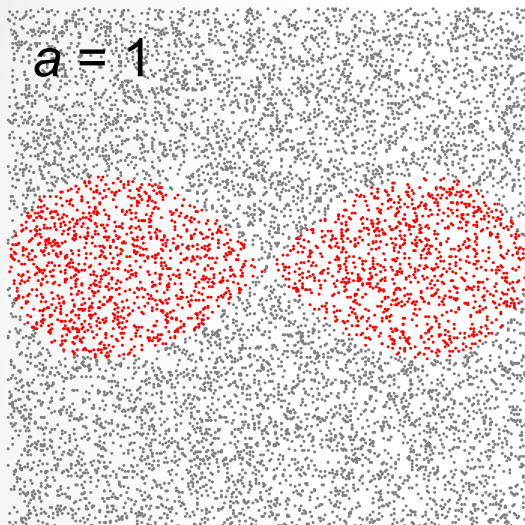


What is its area?

Empirical Monte Carlo distribution of the area estimate



Empirical Monte Carlo distribution of the area estimate



a probability distribution of
a mathematical constant???

Frequentist view: this is the distribution of an estimate, it does not make sense to talk of the distribution of a constant. However, while in this case the value to be estimated is unmistakably “true”, the physical model of the random process is not unlike the model in Bertrand’s paradox, and there is some “observer-related” indefiniteness.

Bayesian view: probability in inference should not be mistaken for probability in probability models, as it describes the state of uncertainty of the observer.

We can start from the Bayesian “reasonable expectation” and use it unambiguously as probability: indeed Cox showed that any reasonable measure of “reasonable expectation” must behave just like common probability.

AMERICAN JOURNAL *of* PHYSICS

A Journal Devoted to the Instructional and Cultural Aspects of Physical Science

VOLUME 14, NUMBER 1

JANUARY-FEBRUARY, 1946

Probability, Frequency and Reasonable Expectation

R. T. COX

The Johns Hopkins University, Baltimore 18, Maryland

Boolean algebra (symbolic logic)

a, b, c ... propositions (true or false)

Basic operations

OR: **$a \vee b$**

AND: **$a \cdot b$**

NOT: **$\sim a$**

Truth tables

a	b	$a \vee b$
T	T	T
T	F	T
F	T	T
F	F	F

a	b	$a \cdot b$
T	T	T
T	F	F
F	T	F
F	F	F

a	$\sim a$
T	F
F	T

Combinations of propositions

$$\sim \sim a = a, \quad (1)$$

$$a \cdot b = b \cdot a, \quad (2) \quad a \vee b = b \vee a, \quad (2')$$

$$a \cdot a = a, \quad (3) \quad a \vee a = a, \quad (3')$$

$$a \cdot (b \cdot c) = (a \cdot b) \cdot c = a \cdot b \cdot c, \quad (4)$$

$$a \vee (b \vee c) = (a \vee b) \vee c = a \vee b \vee c, \quad (4')$$

$$\sim (a \cdot b) = \sim a \vee \sim b, \quad (5)$$

$$\sim (a \vee b) = \sim a \cdot \sim b, \quad (5')$$

$$a \cdot (a \vee b) = a, \quad (6) \quad a \vee (a \cdot b) = a. \quad (6')$$

The combination rules are not all independent, e.g., consider

$$\sim(\mathbf{a} \cdot \mathbf{b}) = \sim\mathbf{a} \vee \sim\mathbf{b} \quad \text{and} \quad \sim(\mathbf{a} \vee \mathbf{b}) = \sim\mathbf{a} \cdot \sim\mathbf{b}$$

When we assume the first, and utilizing $\sim\sim\mathbf{a} = \mathbf{a}$, we can deduce the second:

$$\sim(\mathbf{a} \vee \mathbf{b}) = \sim(\sim\sim\mathbf{a} \vee \sim\sim\mathbf{b}) = \sim\sim(\sim\mathbf{a} \cdot \sim\mathbf{b}) = \sim\mathbf{a} \cdot \sim\mathbf{b}$$

Now let

$$p(\mathbf{b}|\mathbf{a})$$

denote any *measure of reasonable credibility (credibility for short)* of proposition \mathbf{b} when \mathbf{a} is known to be true, and let F be a function that combines credibilities

$$p(\mathbf{c} \cdot \mathbf{b}|\mathbf{a}) = F[p(\mathbf{c}|\mathbf{b} \cdot \mathbf{a}), p(\mathbf{b}|\mathbf{a})]$$

While p is still quite arbitrary, F is constrained by the algebra of propositions.

Now we derive a functional equation for F from

$$\begin{aligned} p(\mathbf{d} \cdot \mathbf{c} \cdot \mathbf{b}|\mathbf{a}) &= p((\mathbf{d} \cdot \mathbf{c}) \cdot \mathbf{b}|\mathbf{a}) \\ &= F[p(\mathbf{d} \cdot \mathbf{c}|\mathbf{b} \cdot \mathbf{a}), p(\mathbf{b}|\mathbf{a})] \\ &= F[F[p(\mathbf{d}|\mathbf{c} \cdot \mathbf{b} \cdot \mathbf{a}), p(\mathbf{c}|\mathbf{b} \cdot \mathbf{a})], p(\mathbf{b}|\mathbf{a})] \end{aligned}$$

and also

$$\begin{aligned} p(\mathbf{d} \cdot \mathbf{c} \cdot \mathbf{b}|\mathbf{a}) &= p(\mathbf{d} \cdot (\mathbf{c} \cdot \mathbf{b})|\mathbf{a}) \\ &= F[p(\mathbf{d}|\mathbf{c} \cdot \mathbf{b} \cdot \mathbf{a}), p(\mathbf{c} \cdot \mathbf{b}|\mathbf{a})] \\ &= F[p(\mathbf{d}|\mathbf{c} \cdot \mathbf{b} \cdot \mathbf{a}), F[p(\mathbf{c}|\mathbf{b} \cdot \mathbf{a}), p(\mathbf{b}|\mathbf{a})]] \end{aligned}$$

Therefore, setting

$$x = p(\mathbf{d}|\mathbf{c} \cdot \mathbf{b} \cdot \mathbf{a})$$

$$y = p(\mathbf{c}|\mathbf{b} \cdot \mathbf{a})$$

$$z = p(\mathbf{b}|\mathbf{a})$$

we find the functional equation

$$F[x, F[y, z]] = F[F[x, y], z]$$

It is easy to see by substitution that the equation

$$F[x, F[y, z]] = F[F[x, y], z]$$

has the solution

$$C f(F[p, q]) = f(p)f(q)$$

where C is an arbitrary constant and f is an arbitrary single-variable function. It can be shown that this is also the general solution if F has continuous second derivatives.

(**homework!**)

Given the arbitrariness of f , we take the identity function, so that

$$\begin{aligned} C p(\mathbf{c} \cdot \mathbf{b}|\mathbf{a}) &= C F[p(\mathbf{c}|\mathbf{b} \cdot \mathbf{a}), p(\mathbf{b}|\mathbf{a})] \\ &= p(\mathbf{c}|\mathbf{b} \cdot \mathbf{a})p(\mathbf{b}|\mathbf{a}) \end{aligned}$$

Then, when we let $\mathbf{c} = \mathbf{b}$, and we assume that credibility ranges from 0 (no credibility) to 1 (certainty), and therefore

$$p(\mathbf{a}|\mathbf{a}) = p(\text{certainty}) = 1$$

we find

$$C p(\mathbf{b} \cdot \mathbf{b}|\mathbf{a}) = C p(\mathbf{b}|\mathbf{a}) = p(\mathbf{b}|\mathbf{b} \cdot \mathbf{a})p(\mathbf{b}|\mathbf{a}) = p(\mathbf{b}|\mathbf{a})$$

and therefore $C = 1$.

Thus we have found that credibility satisfies the condition

$$p(\mathbf{c} \cdot \mathbf{b} | \mathbf{a}) = p(\mathbf{c} | \mathbf{b} \cdot \mathbf{a}) p(\mathbf{b} | \mathbf{a})$$

however this is not yet enough, because if we took a power law instead of the identity, we could still satisfy all the conditions and find, e.g., a condition like

$$p(\mathbf{c} \cdot \mathbf{b} | \mathbf{a})^m = p(\mathbf{c} | \mathbf{b} \cdot \mathbf{a})^m p(\mathbf{b} | \mathbf{a})^m$$

Can we do better?

We have used the properties of logical AND, but not yet those of logical NOT and OR ...

Taking a negated proposition we expect to find the relationship

$$p(\sim \mathbf{b}|\mathbf{a}) = S[p(\mathbf{b}|\mathbf{a})]$$

and therefore we find a functional equation

$$p(\mathbf{b}|\mathbf{a}) = p(\sim \sim \mathbf{b}|\mathbf{a}) = S[S[p(\mathbf{b}|\mathbf{a})]]$$

which, however, is not restrictive enough ...

Now we note that

$$\begin{aligned} S[p(\mathbf{c} \vee \mathbf{b}|\mathbf{a})] &= p(\sim(\mathbf{c} \vee \mathbf{b}|\mathbf{a})) = p(\sim\mathbf{c} \cdot \sim\mathbf{b}|\mathbf{a}) \\ &= p(\sim\mathbf{c}|\sim\mathbf{b} \cdot \mathbf{a})p(\sim\mathbf{b}|\mathbf{a}) \\ &= S[p(\mathbf{c}|\sim\mathbf{b} \cdot \mathbf{a})]S[p(\mathbf{b}|\mathbf{a})] \end{aligned}$$

and also that

$$\begin{aligned} p(\mathbf{c}|\sim\mathbf{b} \cdot \mathbf{a}) &= \frac{p(\mathbf{c} \cdot \sim\mathbf{b}|\mathbf{a})}{p(\sim\mathbf{b}|\mathbf{a})} = \frac{p(\sim\mathbf{b} \cdot \mathbf{c}|\mathbf{a})}{p(\sim\mathbf{b}|\mathbf{a})} \\ &= \frac{p(\sim\mathbf{b}|\mathbf{c} \cdot \mathbf{a}) p(\mathbf{c}|\mathbf{a})}{p(\sim\mathbf{b}|\mathbf{a})} \\ &= \frac{S[p(\mathbf{b}|\mathbf{c} \cdot \mathbf{a})] p(\mathbf{c}|\mathbf{a})}{S[p(\mathbf{b}|\mathbf{a})]} \end{aligned}$$

And finally we find

$$p(\mathbf{c} | \sim \mathbf{b} \cdot \mathbf{a}) = \frac{S[p(\mathbf{b} | \mathbf{c} \cdot \mathbf{a})] p(\mathbf{c} | \mathbf{a})}{S[p(\mathbf{b} | \mathbf{a})]} = S \left[\frac{S[p(\mathbf{c} \vee \mathbf{b} | \mathbf{a})]}{S[p(\mathbf{b} | \mathbf{a})]} \right]$$

or alternatively

$$S \left[\frac{p(\mathbf{c} \cdot \mathbf{b} | \mathbf{a})}{p(\mathbf{c} | \mathbf{a})} \right] \frac{p(\mathbf{c} | \mathbf{a})}{S[p(\mathbf{b} | \mathbf{a})]} = S \left[\frac{S[p(\mathbf{c} \vee \mathbf{b} | \mathbf{a})]}{S[p(\mathbf{b} | \mathbf{a})]} \right]$$

This results hold for all propositions, and if we let $\mathbf{b} = \mathbf{c} \cdot \mathbf{d}$ we find

$$S \left[\frac{p(\mathbf{c} \cdot \mathbf{d} | \mathbf{a})}{p(\mathbf{c} | \mathbf{a})} \right] \frac{p(\mathbf{c} | \mathbf{a})}{S[p(\mathbf{c} \cdot \mathbf{d} | \mathbf{a})]} = S \left[\frac{S[p(\mathbf{c} | \mathbf{a})]}{S[p(\mathbf{c} \cdot \mathbf{d} | \mathbf{a})]} \right]$$

Introducing the auxiliary variables

$$x = p(\mathbf{c} | \mathbf{a}); \quad y = S[p(\mathbf{c} \cdot \mathbf{d} | \mathbf{a})]$$

we obtain a compact form for the functional equation for S

$$x \ S \left[\frac{S[y]}{x} \right] = y \ S \left[\frac{S[x]}{y} \right]$$

It is easy to see by substitution that the equation

$$x \, S \left[\frac{S[y]}{x} \right] = y \, S \left[\frac{S[x]}{y} \right]$$

has the solution

$$S[p] = (1 - p^m)^{1/m}$$

It can be shown that this is also the general solution if S is twice differentiable.

(**homework!**)

Again, this means that

$$p(\sim \mathbf{b}|\mathbf{a}) = S[p(\mathbf{b}|\mathbf{a})] = (1 - p(\mathbf{b}|\mathbf{a})^m)^{1/m}$$
$$\Rightarrow p(\mathbf{b}|\mathbf{a})^m + p(\sim \mathbf{b}|\mathbf{a})^m = 1$$

and again, **whatever the value of m , credibility satisfies the usual probability rule**. Since the choice of m is conventional we take $m = 1$.

Summarizing, we have the following collection of assumptions and rules:

$$p(\text{certainty}) = 1$$

$$p(\text{impossibility}) = 0$$

$$p(\mathbf{b}|\mathbf{a}) + p(\sim \mathbf{b}|\mathbf{a}) = 1$$

$$p(\mathbf{c} \cdot \mathbf{b}|\mathbf{a}) = p(\mathbf{c}|\mathbf{b} \cdot \mathbf{a})p(\mathbf{b}|\mathbf{a})$$

and from these all the usual rules of probability follow.

Therefore we can take probabilities as measures of credibility.

The algebra of probabilities

Let A and B be statements that can be either true or false, and such that we can assign probabilities. Then the following rules apply:

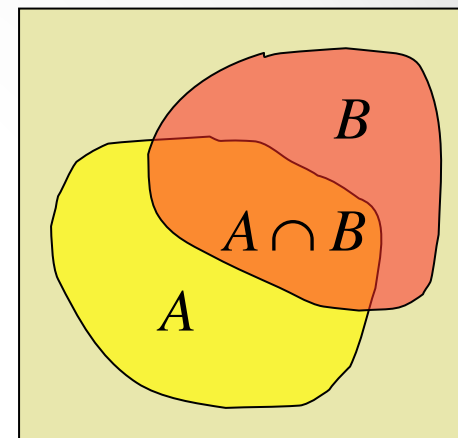
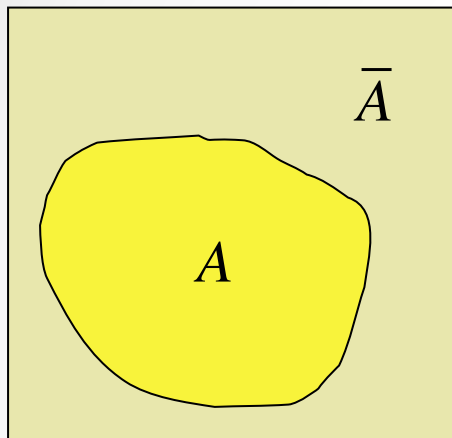
$$0 \leq P(A) \leq 1$$

$$P(\Omega) = 1$$

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

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

Probability space and measure theory



$$0 \leq P(A) \leq 1$$

$$P(\Omega) = 1; \quad P(A) + P(\bar{A}) = 1$$

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

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

Ω

Bayes' Theorem

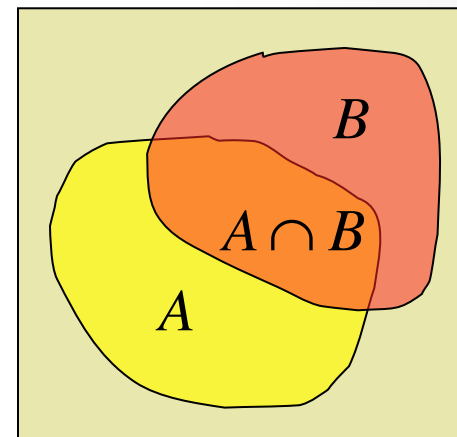
$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

Independent events:

$$P(A \text{ e } B) = P(A) \cdot P(B)$$

Dependent events:

$$P(A \text{ e } B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$



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

Bayes' theorem



rev. Thomas Bayes (1702-1761)

Thomas Bayes was the son of a London Presbyterian minister, Joshua Bayes born perhaps in Hertfordshire. In 1719 he enrolled at the University of Edinburgh to study logic and theology.

He is known to have published two works in his lifetime: *Divine Benevolence, or an Attempt to Prove That the Principal End of the Divine Providence and Government is the Happiness of His Creatures* (1731), and *An Introduction to the Doctrine of Fluxions, and a Defence of the Mathematicians Against the Objections of the Author of the Analyst* (published anonymously in 1736), in which he defended the logical foundation of Isaac Newton's calculus against the criticism of George Berkeley, author of *The Analyst*.

It is speculated that Bayes was elected as a Fellow of the Royal Society in 1742 on the strength of *the Introduction to the Doctrine of Fluxions*, as he is not known to have published any other mathematical works during his lifetime.

Some feel that he became interested in probability while reviewing a work written in 1755 by Thomas Simpson, but others think he learned mathematics and probability from a book by de Moivre. Bayes died in Tunbridge Wells, Kent. He is buried in Bunhill Fields Cemetery in London where many Nonconformists are buried.

(from Wikipedia)

The ideas of Bayes were clarified, extended and put to good use by Pierre Simon, Marquis de Laplace

“... In order to give some interesting applications of it I have profited by the immense work which M. Bouvard has just finished on the movements of Jupiter and Saturn ... His calculations give him the mass of Saturn equal to 3512th part of that of the sun. Applying to them my formulae of probability, I find that it is a bet of 11,000 against one that the error of this result is not 1/100th of its value ...”

(from the *Philosophical essay on probabilities*)



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

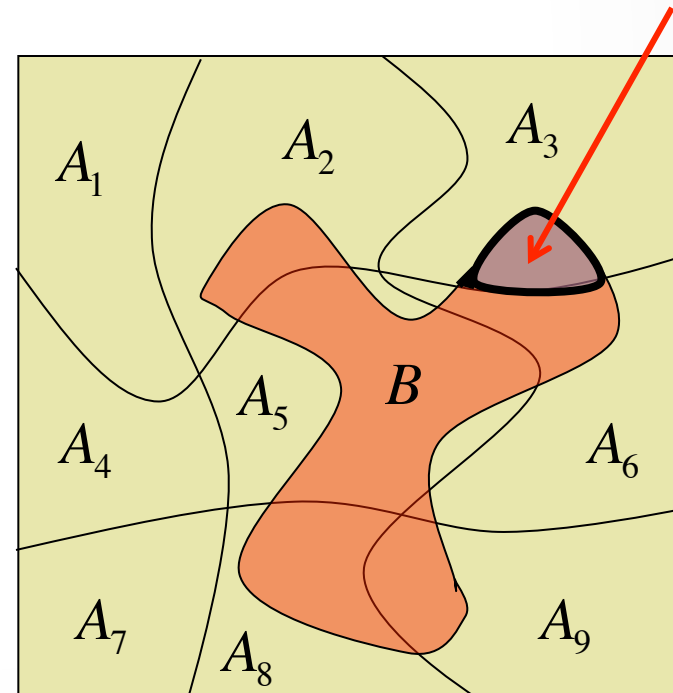
$$P(A_k | B) = \frac{P(B | A_k) \cdot P(A_k)}{P(B)}$$

$$k = 1, \dots, N$$

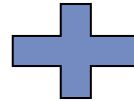
$$P(B | A_3) \cdot P(A_3)$$

if the events A_k are mutually exclusive, and they fill the universe

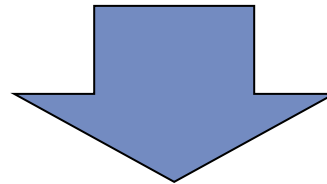
$$P(B) = \sum_{k=1}^N P(B | A_k) \cdot P(A_k)$$



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

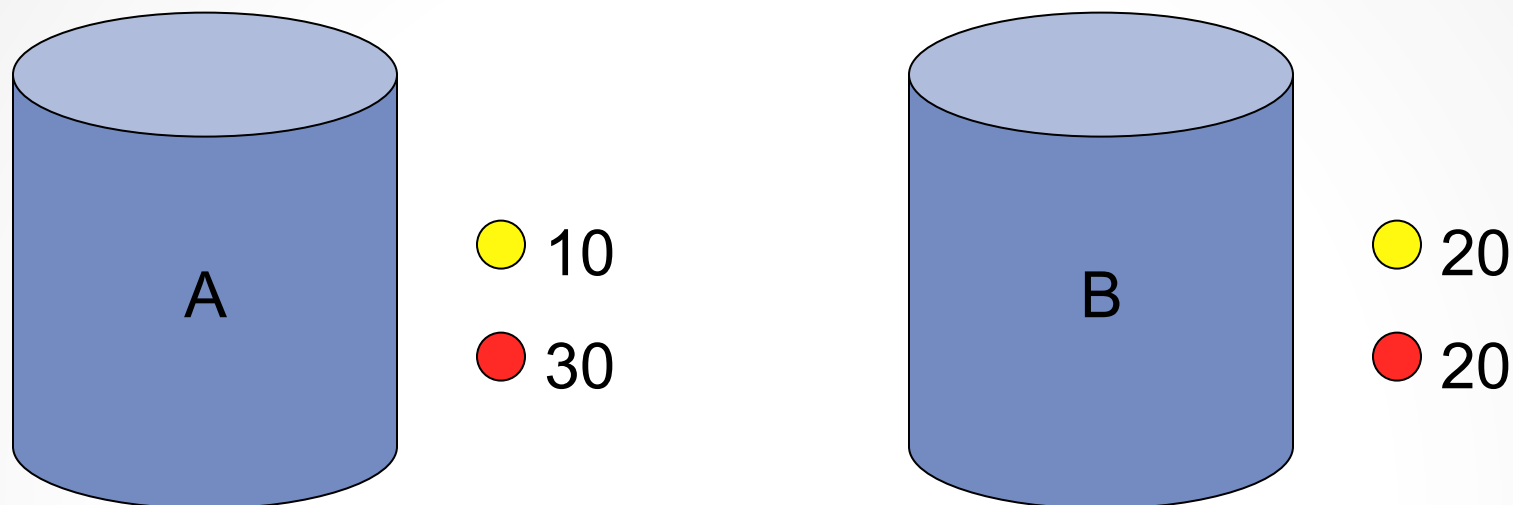


$$P(B) = \sum_{k=1}^N P(B|A_k) \cdot P(A_k)$$



$$P(A_k|B) = \frac{P(B|A_k) \cdot P(A_k)}{\sum_{k=1}^N P(B|A_k) \cdot P(A_k)}$$

A simple probability model based on conditional probabilities



Here we choose a ball as follows:

1. We choose the urn first
2. We draw a ball from that urn

What is the probability of drawing one red ball?

$P(A) = P(B) = 1/2$ (probability of choosing either A or B)

$P(G|A) = 1/4$ (probability of drawing a yellow ball from A)

$P(R|A) = 3/4$ (probability of drawing a red ball from A)

$P(G|B) = 1/2$ (probability of drawing a yellow ball from B)

$P(R|B) = 1/2$ (probability of drawing a red ball from A)

and therefore

$$\begin{aligned} P(R) &= P(R|A) \cdot P(A) + P(R|B) \cdot P(B) \\ &= (3/4) \cdot (1/2) + (1/2) \cdot (1/2) = 5/8 = 0.625 \end{aligned}$$

Inverse problem: if we drew a red ball, what is the probability that we drew it from urn A?

(NB: here we assume that the “physical model” is known, i.e., we assume we know how many red and yellow balls are in each urn)

“a priori” probability: $P(A) = 1/2$

Now we apply Bayes’ theorem

$$P(A | R) = \frac{P(R | A) \cdot P(A)}{P(R)} = \frac{(3/4) \cdot (1/2)}{(5/8)} = \frac{3}{5} = 0.6$$



posterior probability

This is a simple example of Bayesian inference

We draw another red ball, still from the same urn, (however we do not know whether this is A or B). Since now

$$P(R) = P(R|A) \cdot P(A) + P(R|B) \cdot P(B) = 0.65$$

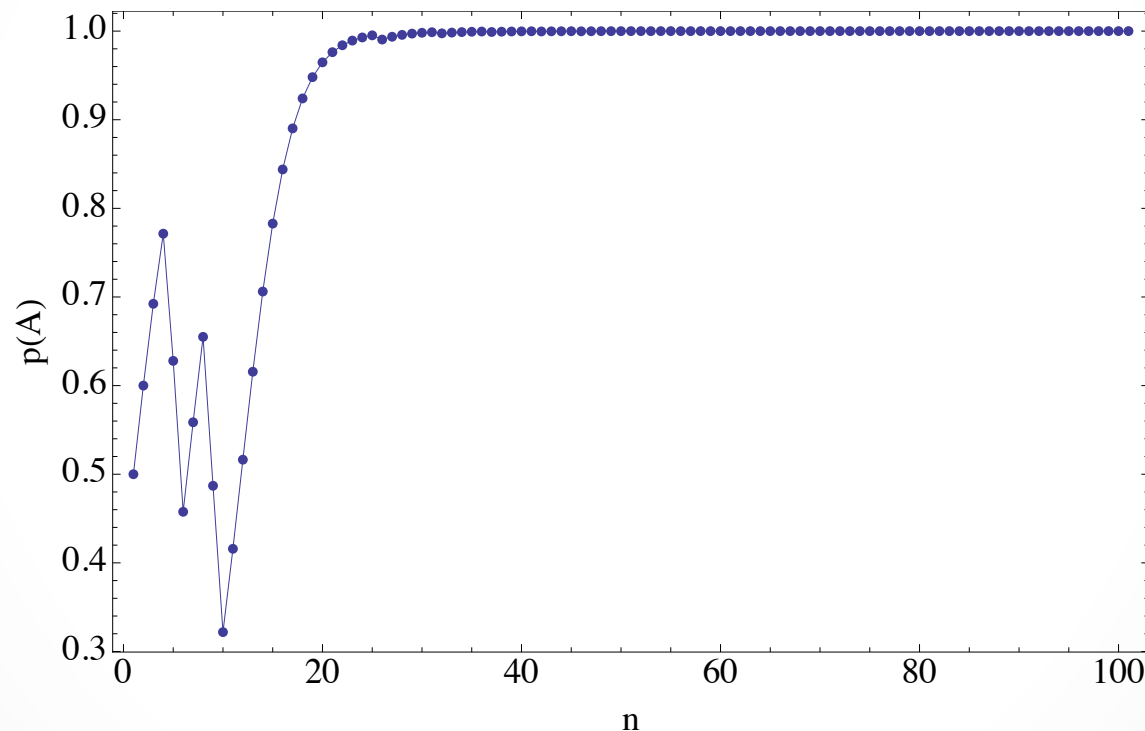
we find

$$P(A|\{R,R\},I) = \frac{P(R|A,I) \cdot P(A|R,I)}{P(R,I)} \approx 0.692308$$

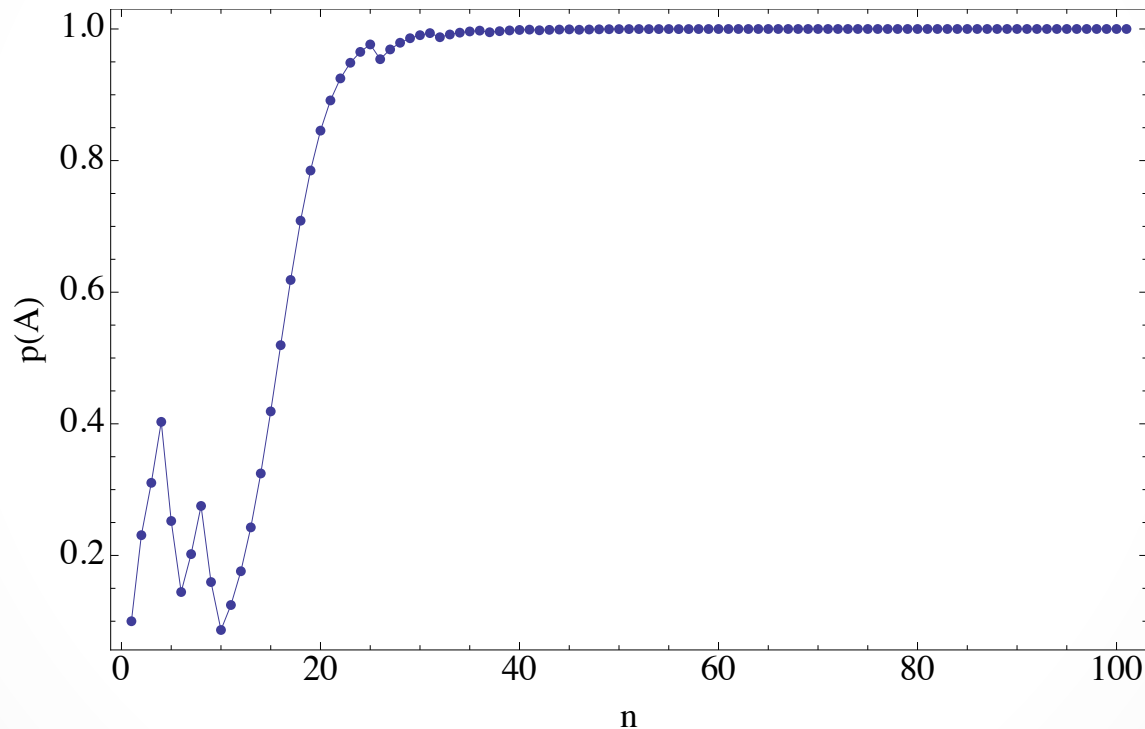
Notice that data can be inserted one by one!

100 successive draws ...

R, R, R, Y, Y, R, R, Y, Y, R, R, R, R, R, R, R, R, R, R, R, R, R, R, Y,
R, R, R, R, R, Y, R, R, R, R, Y, R, R, R, R, Y, R, R, Y, R, R, R, R, R,
R, Y, R, R, Y, R, R, R, R, R, Y, R, R, R, Y, R, R, Y, R, Y, R, R, Y,
Y, R, R, Y, R, R, R, Y, R, R, Y, R, R, R, R, R, R, R, R, R, Y, Y, R, R



... a different starting point: here the initial prior probability is 0.05 instead of 0.5.



A simple application to medical tests (example of HIV test)

$$P(\text{positive} | \text{infect}) = 1$$

$$P(\text{positive} | \text{not infect}) = 1.5\%$$

what is the probability $P(\text{infect} | \text{positive})$?

A common answer is 98.5% ... and it is wrong!

Let's use Bayes' theorem ...

$$P(A_k | B) = \frac{P(B | A_k) \cdot P(A_k)}{\sum_{k=1}^N P(B | A_k) \cdot P(A_k)}$$

$$\begin{aligned} P(\text{infect} | \text{positive}) &= \frac{P(\text{positive} | \text{infect}) \cdot P(\text{infect})}{P(\text{positive} | \text{infect}) \cdot P(\text{infect}) + P(\text{positive} | \text{not infect}) \cdot P(\text{non infect})} \\ &= \frac{P(\text{positive} | \text{infect})}{P(\text{positive} | \text{infect}) \cdot P(\text{infect}) + P(\text{positive} | \text{not infect}) \cdot P(\text{non infect})} \cdot P(\text{infect}) \end{aligned}$$

The estimate depends on the size of the infect population
i.e., on the probabilities

$P(\text{infect})$ $P(\text{not infect})$

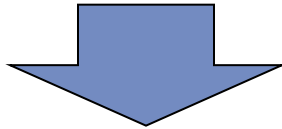
$P(\text{infect} \mid \text{positive})$

$$= \frac{P(\text{positive} \mid \text{infect})}{P(\text{positive} \mid \text{infect}) \cdot P(\text{infect}) + P(\text{positive} \mid \text{not infect}) \cdot P(\text{non infect})} \cdot P(\text{infect})$$

The posterior estimate strongly depends on the prior probability

Example: AIDS frequency in Italy 0.4 %

AIDS frequency in South Africa 18.1%



$$P(\text{infect} \mid \text{positive}) = \frac{1}{1 \cdot 0.004 + 0.015 \cdot 0.996} \cdot 0.004 \approx 21.1\%$$

Italy

$$P(\text{infect} \mid \text{positive}) = \frac{1}{1 \cdot 0.181 + 0.015 \cdot 0.819} \cdot 0.181 \approx 93.6\%$$

South Africa

the large number of false positives and the small probability of finding a sick person mean that the probability of being infected if positive is not actually very high.

If we find a positive result in a repeated measurement:

$$P(\textit{infect} | \{\textit{positive}, \textit{positive}\}) = 94.7\% \quad \text{Italy}$$

$$P(\textit{infect} | \{\textit{positive}, \textit{positive}\}) = 99.9\% \quad \text{South Africa}$$

The first test changes the reference population, and the second test, if positive, gives a significant result.


Prosecutor's fallacy & Defendant's fallacy

Two common mistakes, associated to the wrong reference population

$P(\text{DNA compatible} \mid \text{innocent})$

$P(\text{innocent} \mid \text{DNA compatible})$

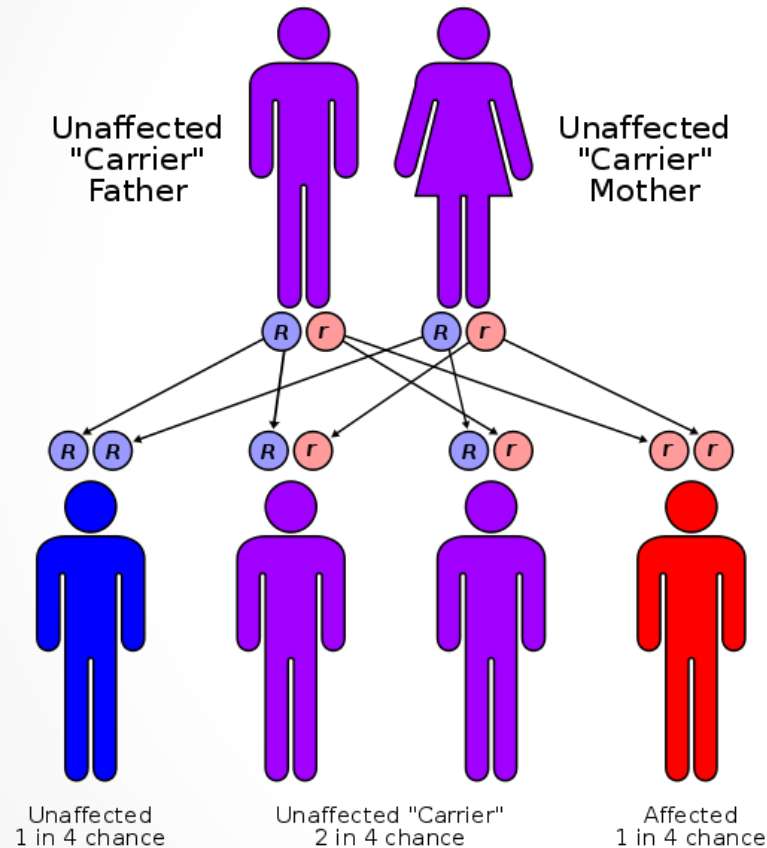
this is
what we
want!



$$P(\text{innocent} \mid \text{DNA compatible}, I) = \frac{P(\text{DNA compatible} \mid \text{innocent}, I)}{P(\text{DNA compatible}, I)} P(\text{innocent} \mid I)$$

DNA classification - 1: alleles

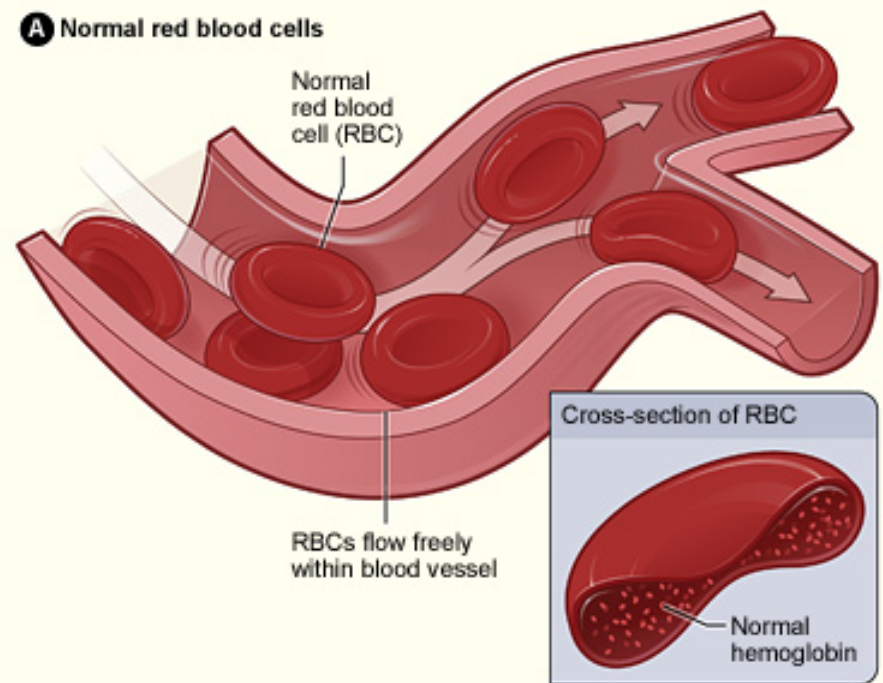
allele: one of two or more alternative forms of the same gene, at the same position in a chromosome.



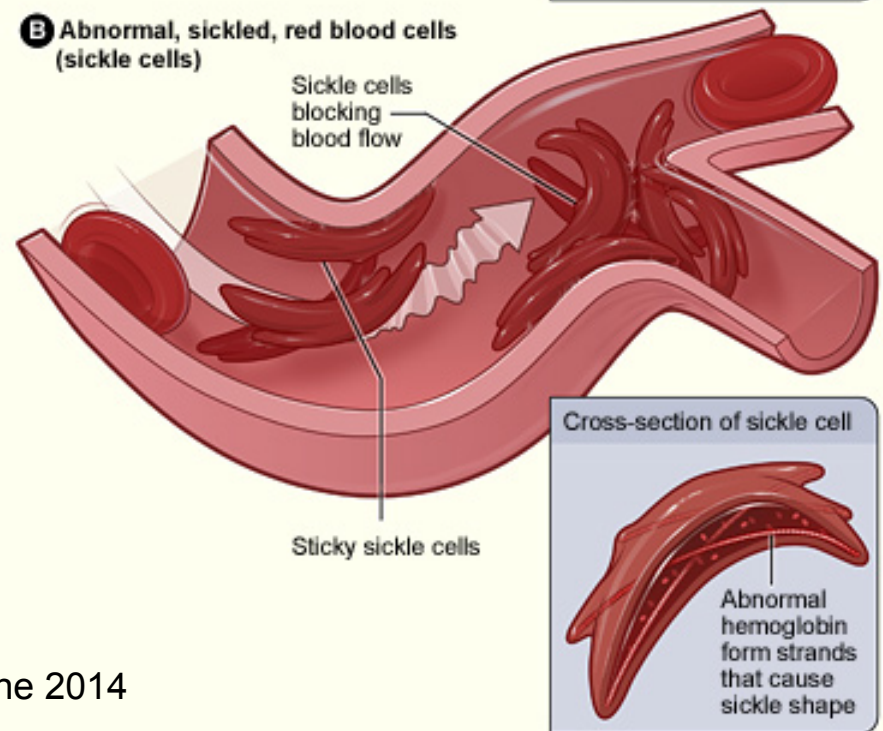
example: sickle
cell anemia

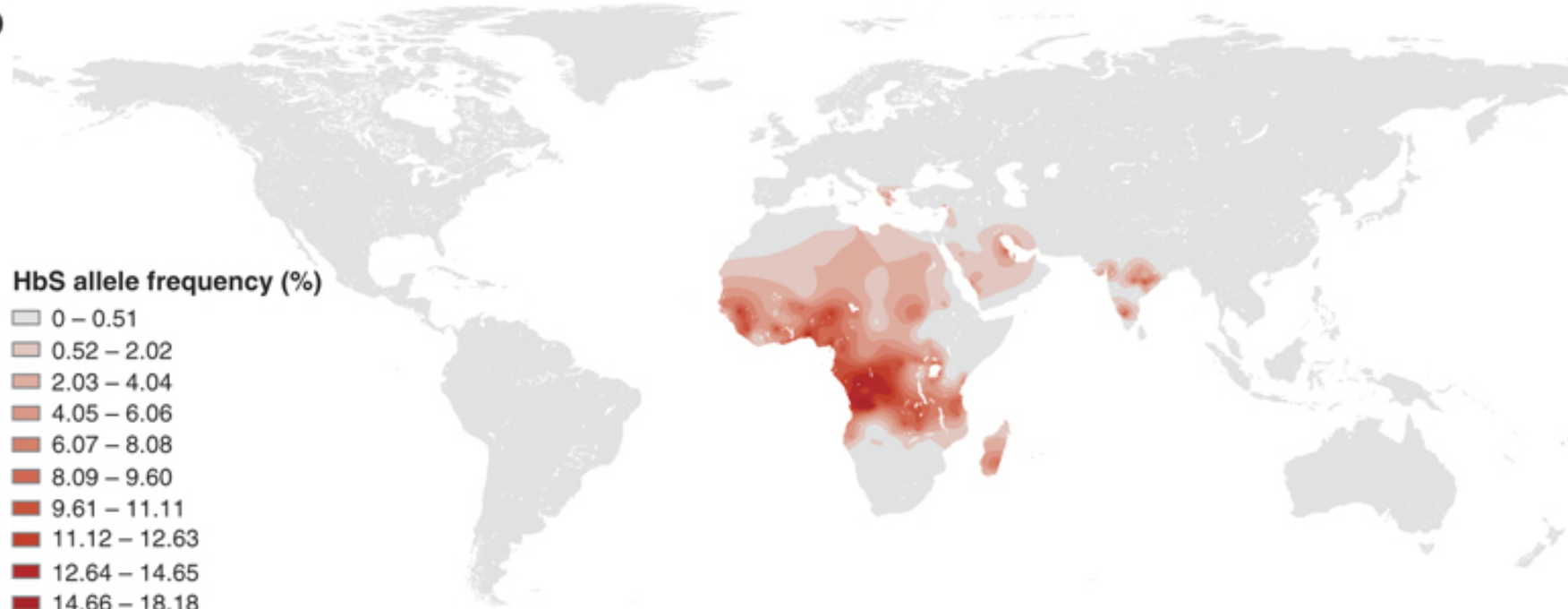
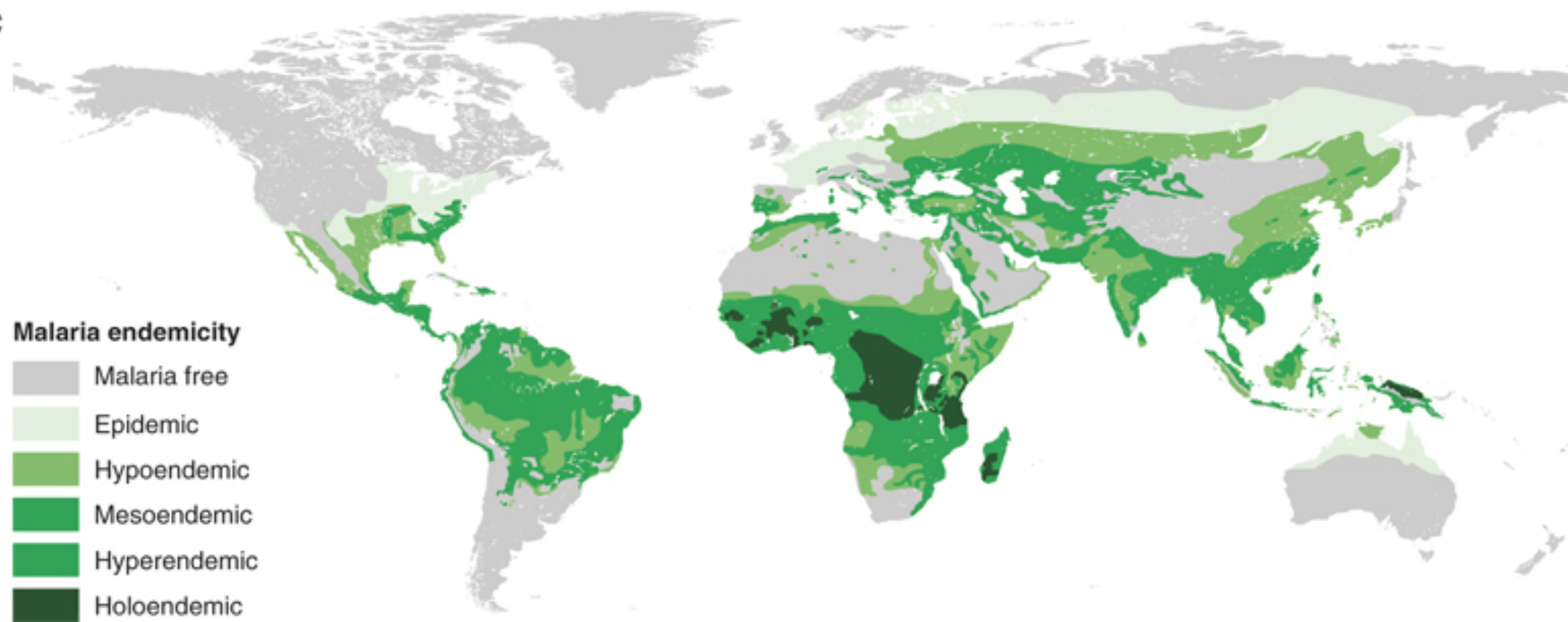


A Normal red blood cells



B Abnormal, sickled, red blood cells (sickle cells)



b**c**

DNA classification - 2: allele frequency

A copy

B copy

DNA Profile		Allele frequency from database				Genotype frequency for locus	
Locus	Alleles	times allele observed	size of database	Frequency		formula	number
CSF1PO	10	109	432	$p=$	0.25	$2pq$	0.16
	11	134		$q=$	0.31		
TPOX	8	229	432	$p=$	0.53	p^2	0.28
	8						
THO1	6	102	428	$p=$	0.24	$2pq$	0.07
	7	64		$q=$	0.15		
vWA	16	91	428	$p=$	0.21	p^2	0.05
	16						
			profile frequency=				0.00014

taken from <http://www.dna-view.com/profile.htm>

Database of human alleles (ALeLe FREquency Database:
<http://alfred.med.yale.edu/alfred/index.asp>

≈ 1/7000, frequency of
 profile in reference
 population

$$P(\text{innocent}|\text{given allele sequence}, I) = \frac{P(\text{given allele sequence}|\text{innocent}, I)}{P(\text{given allele sequence}, I)}P(\text{innocent}|I)$$

where

$$P(\text{given allele sequence}, I) = P(\text{given allele sequence}|\text{innocent}, I)P(\text{innocent}|I) + P(\text{given allele sequence}|\text{guilty}, I)P(\text{guilty}|I)$$

Since the test has a very low error probability, i.e.,

$$P(\text{given allele sequence}|\text{guilty}, I) \approx 1$$

we find

$$P(\text{given allele sequence}, I) = 0.00014 \times P(\text{innocent}|I) + 1 \times P(\text{guilty}|I)$$

Once again, just like in the previous example, we see that it is all-important to determine the prior probabilities $P(\text{innocent}|I)$ and $P(\text{guilty}|I)$. For instance, if we pick a suspect at random in a large population, e.g., in a city with 1 million inhabitants, then

$$P(\text{innocent}|I) = 1 - 10^{-6} = 0.999999; \quad P(\text{guilty}|I) = 10^{-6} = 0.000001$$

$$P(\text{given allele sequence}, I) = 0.00014 \times (1 - 10^{-6}) + 1 \times 10^{-6} \approx 0.000141$$

and finally

$$P(\text{innocent}|\text{given allele sequence}, I) = \frac{0.00014}{0.000141}(1 - 10^{-6}) \approx 0.992908$$

This last result shows that the DNA test is quite inconclusive in this case, because it decreases the probability that the suspect is innocent from 0.999999 to 0.992908, only. How can it be? The reason is that in this case the number of random matches is not small, indeed in this city there are on average $1000000/7000 \approx 143$ people that randomly match the given allele sequence.

The argument can be turned upside down by a cunning lawyer, who might claim that since there are so many random matches, the DNA test is not relevant. However it is not so, and this claim is the “defendant’s fallacy”. Indeed, the problem that we met above was that the starting population was far too large. Other evidence might considerably reduce the number of possible suspects, for instance a surveillance camera might help identify all the people who entered a building and who had a chance to commit the crime, and thus reduce the starting population to, say, 100 people. When we repeat the relevant calculations, we find

$$P(\text{innocent}|I) = 1 - 1/100 = 0.99; \quad P(\text{guilty}|I) = 1/100 = 0.01$$

$$P(\text{given allele sequence}, I) = 0.00014 \times 0.99 + 1 \times 0.01 \approx 0.01014$$

and finally

$$P(\text{innocent}|\text{given allele sequence}, I) = \frac{0.00014}{0.000141}(1 - 10^{-2}) \approx 0.0137$$

Check the webpage:

<http://wwwusers.ts.infn.it/~milotti/Didattica/Bayes/2014-MiBi/Bayes.html>

References:

- L. Mahadevan and E. H. Yong, “Probability, physics, and the coin toss”, *Phys. Today*, July 2011, pp. 66-67
- P. Diaconis, S. Holmes and R. Montgomery, “Dynamical bias in the coin toss”, *SIAM Rev.* **49** (2007) 211
- G. D’Agostini, *Rep. Prog. Phys.* **66** (2003) 1383
- V. Dose, *Rep. Prog. Phys.* **66** (2003) 1421
- W. C. Thompson and E. L. Schulman, *Law and Human Behavior* **11** (1987) 167
- M. Botje, lecture notes: <http://www.nikhef.nl/~h24/bayes/>