Introduction to Bayesian Methods- 2

Edoardo Milotti
Università di Trieste and INFN-Sezione di Trieste

Likelihood

Prior distribution

Posterior distribution

$$P(H|D) = \frac{P(D|H)}{P(D)} P(H)$$

Evidence

$$P(H_k|D) = \frac{P(D|H_k)}{\sum_j P(D|H_j)P(H_j)} P(H_k)$$

MAP estimates

Example: a decision problem (Skilling 1998)

Let T be the temperature of a liquid which can be either water or ethanol.

1. We suppose first that the liquid is water: then we take a uniform prior distribution for T, between $0\ ^{\circ}\mathrm{C}$ and $100\ ^{\circ}\mathrm{C}$
2. The experimental apparatus and the measurement process is defined by the likelihood function $\mathbf{P(D|T,water,I)}$. We assume that measurements are uniformly distributed within a range $\pm 5\ ^{\circ}\mathrm{C}$. Therefore $\mathbf{P(D|T,water,I) = 0.1\ (°C)^{-1}}$ in the interval $\mathbf{[T-5°C, T+5°C]}$, and zero elsewhere.
3. We take a single measurement $\mathbf{D = -3°C}$.

4. The evidence P(D) is

$$P\left(D\middle|water,I\right)=\int_{T}P\left(D\middle|T,water,I\right)P\left(T\right)dT$$

$$=\int_{0°C}^{2°C}\frac{\left(°C\right)^{-1}}{10}\cdot\frac{\left(°C\right)^{-1}}{100}dT\left(°C\right)=0.002\left(°C\right)^{-1}$$

5. Using Bayes' theorem we find

$$P\left(T\middle|D,water,I\right)=\frac{P\left(D\middle|T,water,I\right)}{P\left(D,water,I\right)}P\left(T\right)=\frac{0.1\left(°C\right)^{-1}}{0.002\left(°C\right)^{-1}}0.01\left(°C\right)^{-1}$$

$$=0.5\left(°C\right)^{-1}\qquad\left(0°C<T<2°C\right)$$

Now suppose that the liquid is ethanol, so that the temperature range is -80°C<T<80°C

1. P(T) = $(160°C)^{-1}$ in -80°C<T<80°C.
2. P(D|T,ethanol,I) = 0.1 $(°C)^{-1}$ in [T-5°C, T+5°C], and zero elsewhere.
3. We take a single measurement D = -3°C.
4. The evidence P(D,ethanol,I) is

$$P(D,ethanol,I)=\int_T P(D|T,ethanol,I)P(T)dT=\int_{-8°C}^{2°C}\frac{(°C)^{-1}}{10}\cdot\frac{(°C)^{-1}}{160}dT\,(°C)=0.00625(°C)^{-1}$$

5. Using Bayes' theorem we find

$$P(T|D,ethanol,I)=\frac{P(D|T,ethanol,I)}{P(D,ethanol,I)}P(T)=\frac{0.1(°C)^{-1}}{0.00625(°C)^{-1}}\frac{1}{160}(°C)^{-1}=0.1(°C)^{-1}$$

$$(-8°C<T<2°C)$$

Assuming a prior for the water-ethanol choice, we can discriminate between water and ethanol

$$P_{water} = P_{ethanol} = 0.5$$

Indeed,

$$P(water|D,I) = \frac{P(D|water,I)}{P(D|water,I)P(water,I) + P(D|ethanol,I)P(ethanol,I)} P(water,I)$$

$$= \frac{P(D|water,I)}{P(D|water,I) + P(D|ethanol,I)}$$

and therefore the ratio of the posteriors is given by the Bayes' factor

$$\frac{P(water|D,I)}{P(ethanol|D,I)} = \frac{P(D|water,I)}{P(D|ethanol,I)}$$

We have found earlier that

$$P\left(D\middle|water\right)=0.002\left(°C\right)^{-1}$$

$$P\left(D\middle|ethanol\right)=0.00625\left(°C\right)^{-1}$$

therefore

$$\frac{P\left(ethanol\middle|D,I\right)}{P\left(water\middle|D,I\right)}=\frac{P\left(D\middle|ethanol,I\right)}{P\left(D\middle|water,I\right)}=3.125$$

and we conclude that the observation favors the hypothesis of liquid ethanol.

$\log_{10}(B)$	B	Evidence support
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

Interpretation of the Bayes factor B as evidence support according to Jeffreys.

In the case of the water-ethanol problem, and according to Jeffreys' categories, the preference for ethanol is "not worth more than a bare mention", although it happens to be in the upper part of the range.

Example: analytical straight-line fit

$$y_i = ax_i + b + \varepsilon_i$$

y_i measured value

x_i independent variable (“exactly” known)

a, b fit parametes: eventually we expect to find pdf’s for these parameters

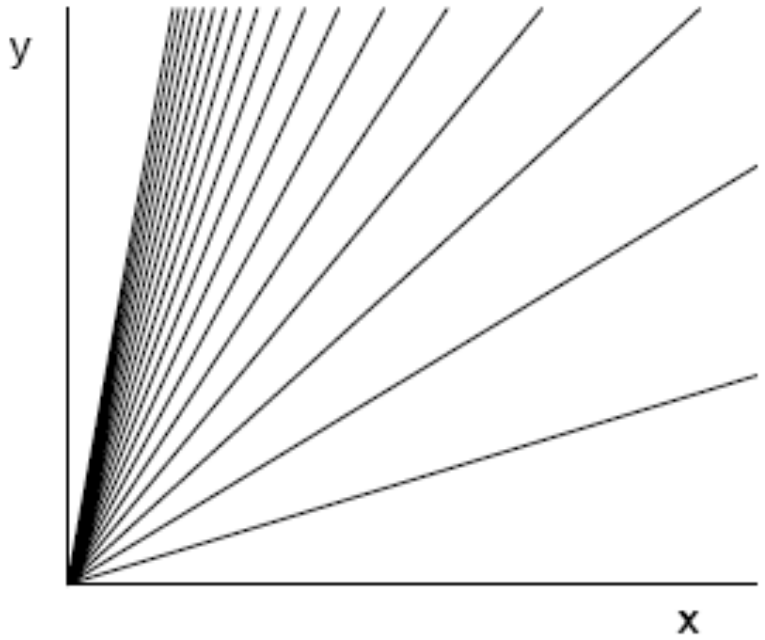
ε_i statistical error

$\langle \varepsilon_i \rangle = 0; \quad \langle \varepsilon_i^2 \rangle = \sigma^2 \quad \Rightarrow$ the statistical measurement error has a Gaussian distribution

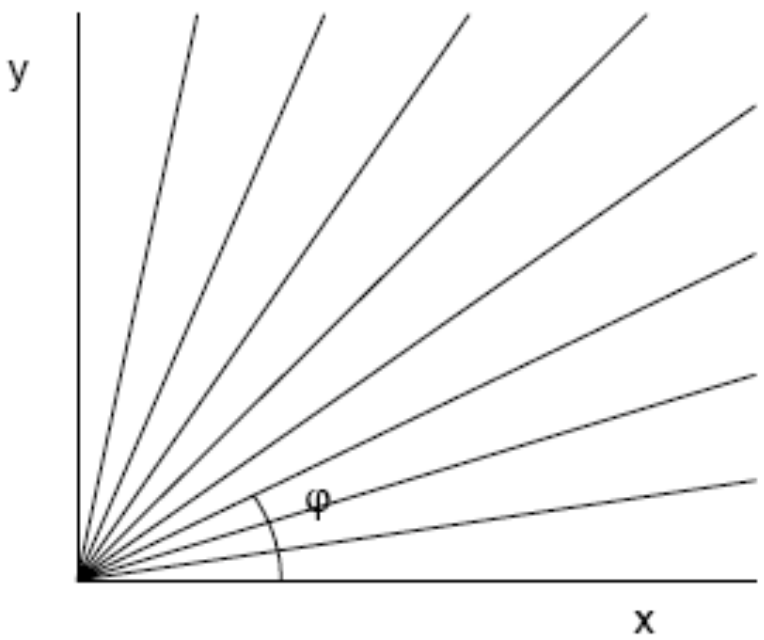
setting up the likelihood

$$p(\mathbf{y} \mid a, b, \mathbf{x}, \sigma) = \left(2\pi\sigma^2\right)^{-N/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - ax_i - b\right)^2\right]$$

prior angular distribution



uniform *a*



uniform angle

The uniform distribution of a introduces an angular bias. The least informative choice corresponds to a uniform angular distribution

$$p_\varphi(\varphi) = \frac{1}{\pi}; \quad -\frac{\pi}{2} \le \varphi < \frac{\pi}{2}$$

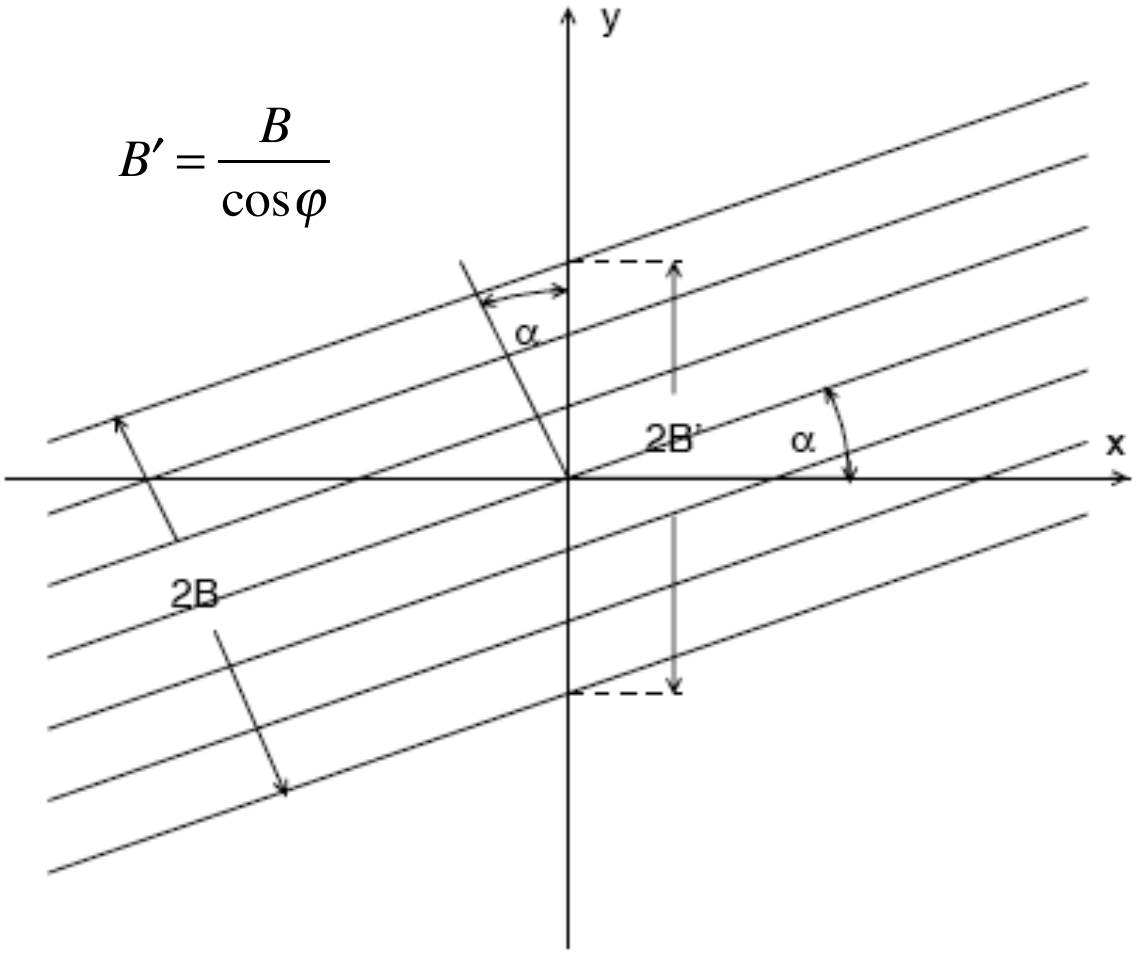
and we obtain the distribution of *a* with the transformation method:

$$a = \tan\varphi$$

$$\Rightarrow \quad p_\varphi(\varphi)d\varphi = p_a(a)da = p_a(a)d(\tan\varphi) = p_a(a)\sec^2\varphi d\varphi$$

$$\Rightarrow \quad p_a(a) = \frac{1}{\pi\sec^2\varphi} = \frac{1}{\pi\left(1+\tan^2\varphi\right)} = \frac{1}{\pi\left(1+a^2\right)}$$

prior distribution of *b*: improper uniform distribution, related to the distribution of *a*



$$B' = \frac{B}{\cos\varphi}$$

$$p(b \mid a = 0) = \frac{1}{2B}; \quad p(b \mid a) = \frac{1}{2B'} = \frac{\cos\varphi}{2B} = \frac{1}{2B} \cdot \frac{1}{\sqrt{1 + a^2}}$$

we obtain the posterior from Bayes' theorem

$$p(a,b \mid \mathbf{y},\mathbf{x},\sigma) = \frac{p(\mathbf{y} \mid a,b,\mathbf{x},\sigma)}{\int_{-\infty}^{+\infty} da \int_{-B/\cos\varphi}^{B/\cos\varphi} db\; p(\mathbf{y} \mid a,b,\mathbf{x},\sigma)\cdot p(a,b)} \cdot p(a,b)$$

where the prior is

$$p(a,b) = p(b \mid a)\cdot p(a) = \left(\frac{1}{2B}\cdot\frac{1}{\sqrt{1+a^2}}\right)\left(\frac{1}{\pi\left(1+a^2\right)}\right)$$

$$\propto \frac{1}{\left(1+a^2\right)^{3/2}}$$

finally we find

$$p(a,b \mid \mathbf{y},\mathbf{x},\sigma) = \frac{\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - ax_i - b\right)^2\right]}{\left\{\int_{-\infty}^{+\infty} da \int_{-B/\cos\varphi}^{B/\cos\varphi} db \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - ax_i - b\right)^2\right]\cdot\frac{1}{\left(1+a^2\right)^{3/2}}\right\}}\cdot\frac{1}{\left(1+a^2\right)^{3/2}}$$

$$\approx \frac{\frac{1}{\left(1+a^2\right)^{3/2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - ax_i - b\right)^2\right]}{\left\{\int_{-\infty}^{+\infty}\frac{da}{\left(1+a^2\right)^{3/2}}\int_{-\infty}^{+\infty} db \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - ax_i - b\right)^2\right]\right\}}$$

This expression has a partly Gaussian structure, and now we rearrange the quadratic expression in the exponential

$$\sum_{i=1}^{N}\left(y_i - ax_i - b\right)^2 = \sum_{i=1}^{N}\left[\left(y_i - ax_i\right)^2 - 2b\left(y_i - ax_i\right) + b^2\right]$$

$$= \sum_{i=1}^{N}\left(y_i - ax_i\right)^2 - 2b\sum_{i=1}^{N}\left(y_i - ax_i\right) + Nb^2$$

$$= N\left\{\left[b^2 - 2b\frac{1}{N}\sum_{i=1}^{N}\left(y_i - ax_i\right) + \left(\frac{1}{N}\sum_{i=1}^{N}\left(y_i - ax_i\right)\right)^2\right] + \frac{1}{N}\sum_{i=1}^{N}\left(y_i - ax_i\right)^2 - \left(\frac{1}{N}\sum_{i=1}^{N}\left(y_i - ax_i\right)\right)^2\right\}$$

$$= N\left\{\left(b - \frac{1}{N}\sum_{i=1}^{N}\left(y_i - ax_i\right)\right)^2 + \frac{1}{N}\sum_{i=1}^{N}\left(y_i - ax_i\right)^2 - \left(\frac{1}{N}\sum_{i=1}^{N}\left(y_i - ax_i\right)\right)^2\right\}$$

$$= N\left(b - \frac{1}{N}\sum_{i=1}^{N}\left(y_i - ax_i\right)\right)^2 + N\left(\frac{1}{N}\sum_{i=1}^{N}y_i^2 - 2a\frac{1}{N}\sum_{i=1}^{N}x_iy_i + a^2\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - N\left(\frac{1}{N}\sum_{i=1}^{N}y_i - a\frac{1}{N}\sum_{i=1}^{N}x_i\right)^2$$

$$= N\left(b - \frac{1}{N}\sum_{i=1}^{N}\left(y_i - ax_i\right)\right)^2 + N\left(\operatorname{var} y - 2a\operatorname{cov}(x,y) + a^2\operatorname{var} x\right)$$

therefore the normalization integral becomes

$$\int_{-\infty}^{+\infty}\frac{da}{\left(1+a^2\right)^{3/2}}\exp\left[-\frac{N}{2\sigma^2}\left(\operatorname{var} y - 2a\operatorname{cov}(x,y) + a^2\operatorname{var} x\right)\right]\int_{-\infty}^{+\infty}db\,\exp\left[-\frac{N}{2\sigma^2}\left(b - \frac{1}{N}\sum_{i=1}^{N}\left(y_i - ax_i\right)\right)^2\right]$$

$$= \sqrt{\frac{2\pi\sigma^2}{N}}\int_{-\infty}^{+\infty}\frac{da}{\left(1+a^2\right)^{3/2}}\exp\left[-\frac{N}{2\sigma^2}\left(\operatorname{var} y - 2a\operatorname{cov}(x,y) + a^2\operatorname{var} x\right)\right]$$

Approximate integration of the remaining integral

$$\int_{-\infty}^{+\infty} \frac{da}{\left(1+a^2\right)^{3/2}} \exp\left[-\frac{N}{2\sigma^2}\left(\operatorname{var} y - 2a\operatorname{cov}(x,y) + a^2 \operatorname{var} x\right)\right]$$

We evaluate this integral by integrating about the peak of the integrand, assuming that the peak is narrow.

We start with the logarithm of the integrand, we find its maximum and we Taylor expand about the maximum

$$\Phi(a) = -\frac{3}{2}\ln\left(1+a^2\right) - \frac{N}{2\sigma^2}\left(\operatorname{var} y - 2a\operatorname{cov}(x,y) + a^2 \operatorname{var} x\right)$$

$$\Phi(a) = -\frac{3}{2}\ln\left(1+a^2\right) - \frac{N}{2\sigma^2}\left(\operatorname{var} y - 2a\operatorname{cov}(x,y) + a^2 \operatorname{var} x\right)$$

$$\frac{d\Phi}{da} = -\frac{3a}{1+a^2} + \frac{N}{\sigma^2}\left(\operatorname{cov}(x,y) - a\operatorname{var} x\right) = 0$$

we find *a* from this cubic equation

note that when N>>1 the peak is at position $a_0 \approx \dfrac{\operatorname{cov}(x,y)}{\operatorname{var} x}$

We use the Newton-Raphson method for the solution of the cubic equation:

$$f(a_0) = -\frac{3a_0}{1+a_0^2}$$

$$f'(a_0) = -3\frac{1-a_0^2}{(1+a_0^2)^2} - \frac{N}{\sigma^2}\operatorname{var} x \approx -\frac{N}{\sigma^2}\operatorname{var} x$$

then

$$\delta a_1 = -\frac{3a_0}{1+a_0^2}\frac{\sigma^2}{N\,\mathrm{var}\,x} \qquad a_1 = a_0 - \frac{3a_0}{1+a_0^2}\frac{\sigma^2}{N\,\mathrm{var}\,x} \qquad (1)$$

Now, to complete the expansion, we must evaluate the second derivative at a_1:

$$\frac{d^2\Phi}{da^2} = -3\frac{1-a_1^2}{(1+a_1^2)^2} - \frac{N}{\sigma^2}\mathrm{var}\ x = -\frac{1}{\sigma_1^2} \qquad (2)$$

$$\Phi(a) \approx \Phi(a_1) + \frac{1}{2}\frac{d^2\Phi}{da^2}\bigg|_{a_1}(a-a_1)^2 = \Phi(a_1) - \frac{(a-a_1)^2}{2\sigma_1^2}$$

we find this by using equations (1) and (2)

Now we complete the evaluation of the integral

$$\int_{-\infty}^{+\infty} \frac{da}{\left(1+a^2\right)^{3/2}} \exp\left[-\frac{N}{2\sigma^2}\left(\operatorname{var} y - 2a\operatorname{cov}(x,y) + a^2 \operatorname{var} x\right)\right]$$

$$= \int_{-\infty}^{+\infty} \exp\left[\Phi(a)\right] da$$

$$\approx \int_{-\infty}^{+\infty} \exp\left[\Phi(a_1) - \frac{(a-a_1)^2}{2\sigma_1^2}\right] da = \sqrt{2\pi\sigma_1^2}\exp\left[\Phi(a_1)\right]$$

and finally we find the posterior distribution.

Moreover

$$p(a,b \mid \mathbf{y},\mathbf{x},\sigma) \propto \frac{1}{\left(1+a^2\right)^{3/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - ax_i - b\right)^2\right]$$

$$\approx \exp\left[-\Phi(a_1) - \frac{\left(a-a_1\right)^2}{2\sigma_1^2}\right] \exp\left[-\frac{N}{2\sigma^2}\left(b - \frac{1}{N}\sum_{i=1}^{N}\left(y_i - a_1 x_i\right)\right)^2\right]$$

and thus we see that:

$$\langle a \rangle = a_1; \quad \operatorname{var} a = \sigma_1^2;$$

$$\langle b \rangle = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - a_1 x_i\right); \quad \operatorname{var} b = \frac{\sigma^2}{N}$$

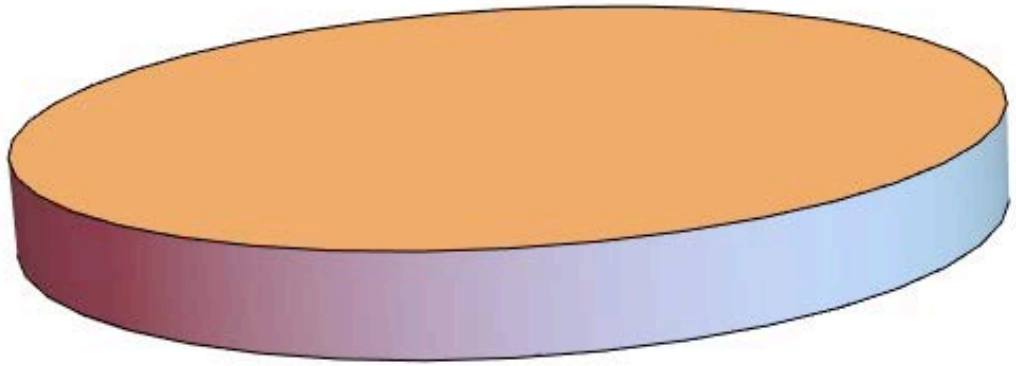
Variable transformations and prior distributions

$$p_x(x)dx = p_x\left(x(y)\right)\left|\frac{dx}{dy}\right|dy = p_y(y)dy$$

$$\Rightarrow \quad p_y(y) = p_x\left(x(y)\right)\left|\frac{dx}{dy}\right|$$

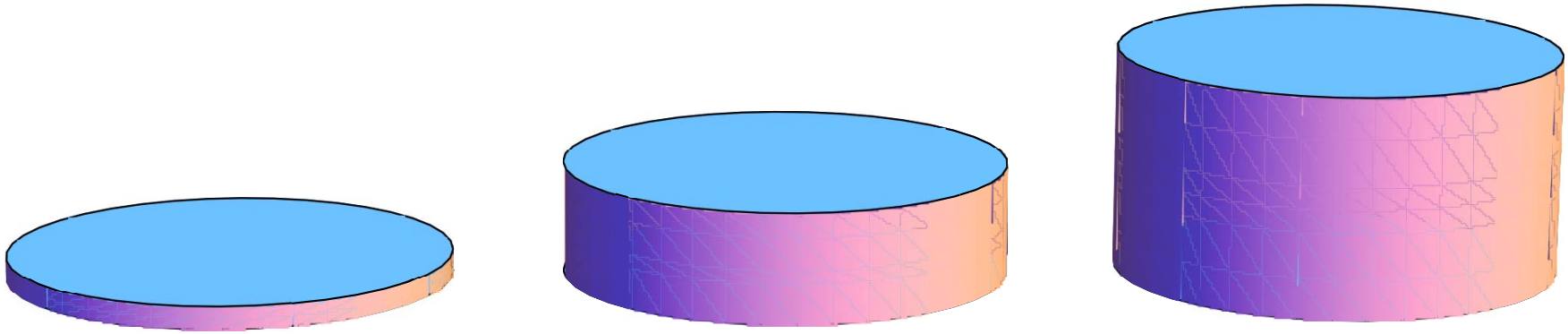
In general, if the first pdf is uniform, the other one is not.

How can we "objectively" choose a prior distribution???



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?

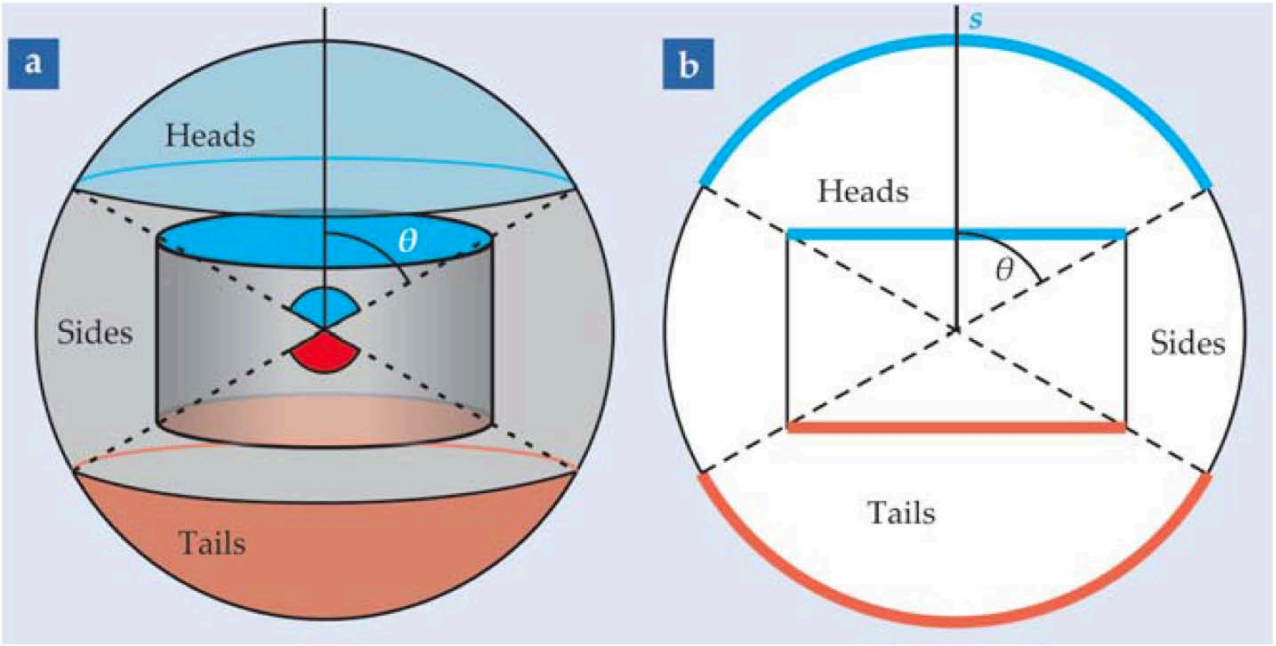
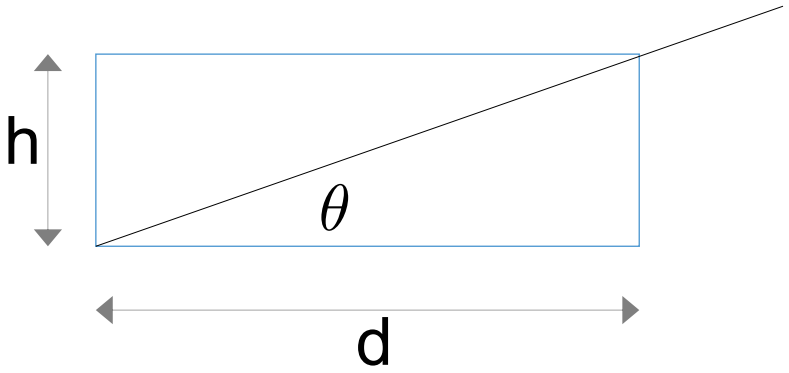


figure from Mahadevan and Yong, "Probability, physics, and the coin toss", Phys. Today, July 2011, pp. 66-67

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

$$2\pi \times \int_0^{\theta_0} \sin\theta d\theta = 2\pi(1-\cos\theta_0)$$



$$\begin{aligned}
\Omega_{\text{heads}} = \Omega_{\text{tails}} = \Omega_{\text{sides}} &= 4\pi/3 \\
\Rightarrow\ & 2\pi(1-\cos\theta_0) = 4\pi/3 \\
\Rightarrow\ & \frac{h}{\sqrt{h^2+d^2}} = \frac{r}{\sqrt{r^2+1}} = 1/3 \\
\Rightarrow\ & r = 1/2\sqrt{2}
\end{aligned}$$

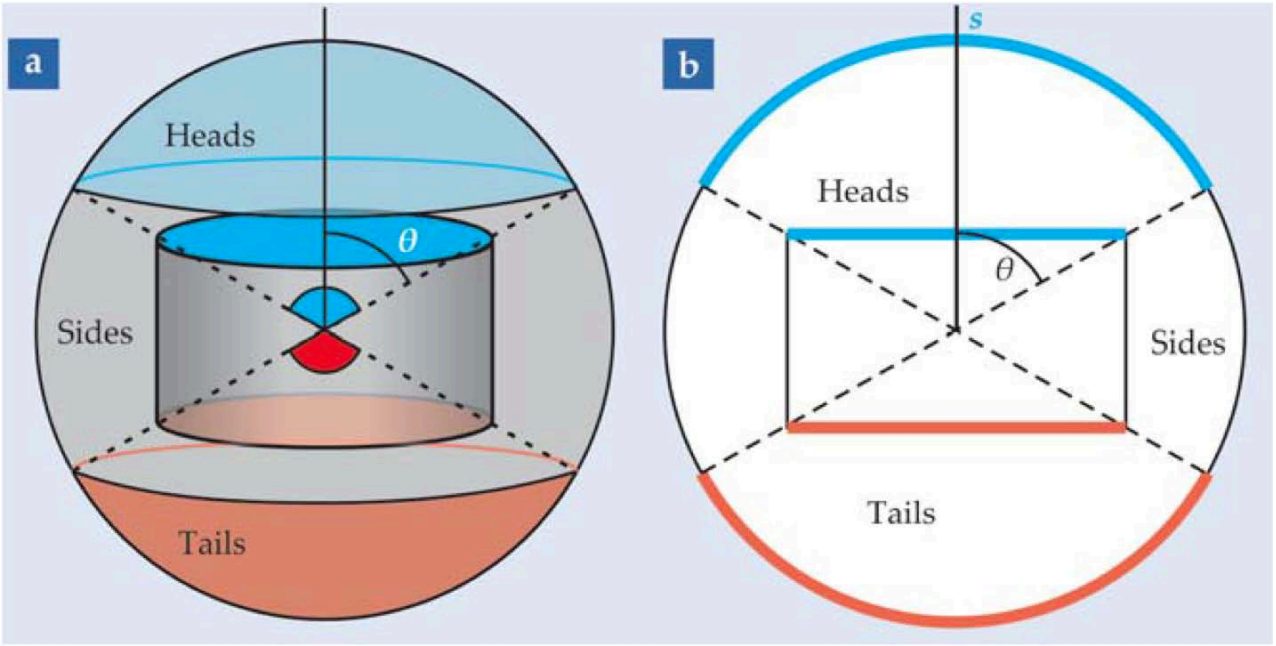


figure from Mahadevan and Yong, "Probability, physics, and the coin toss", Phys. Today, July 2011, pp. 66-67

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

$$\theta_{\text{heads}} = \theta_{\text{tails}} = \theta_{\text{sides}} = \pi/3$$
$$\Rightarrow \cos\theta_0 = 1/2$$
$$\Rightarrow \frac{h}{\sqrt{h^2+d^2}} = \frac{r}{\sqrt{r^2+1}} = 1/2$$
$$\Rightarrow r = 1/\sqrt{3}$$

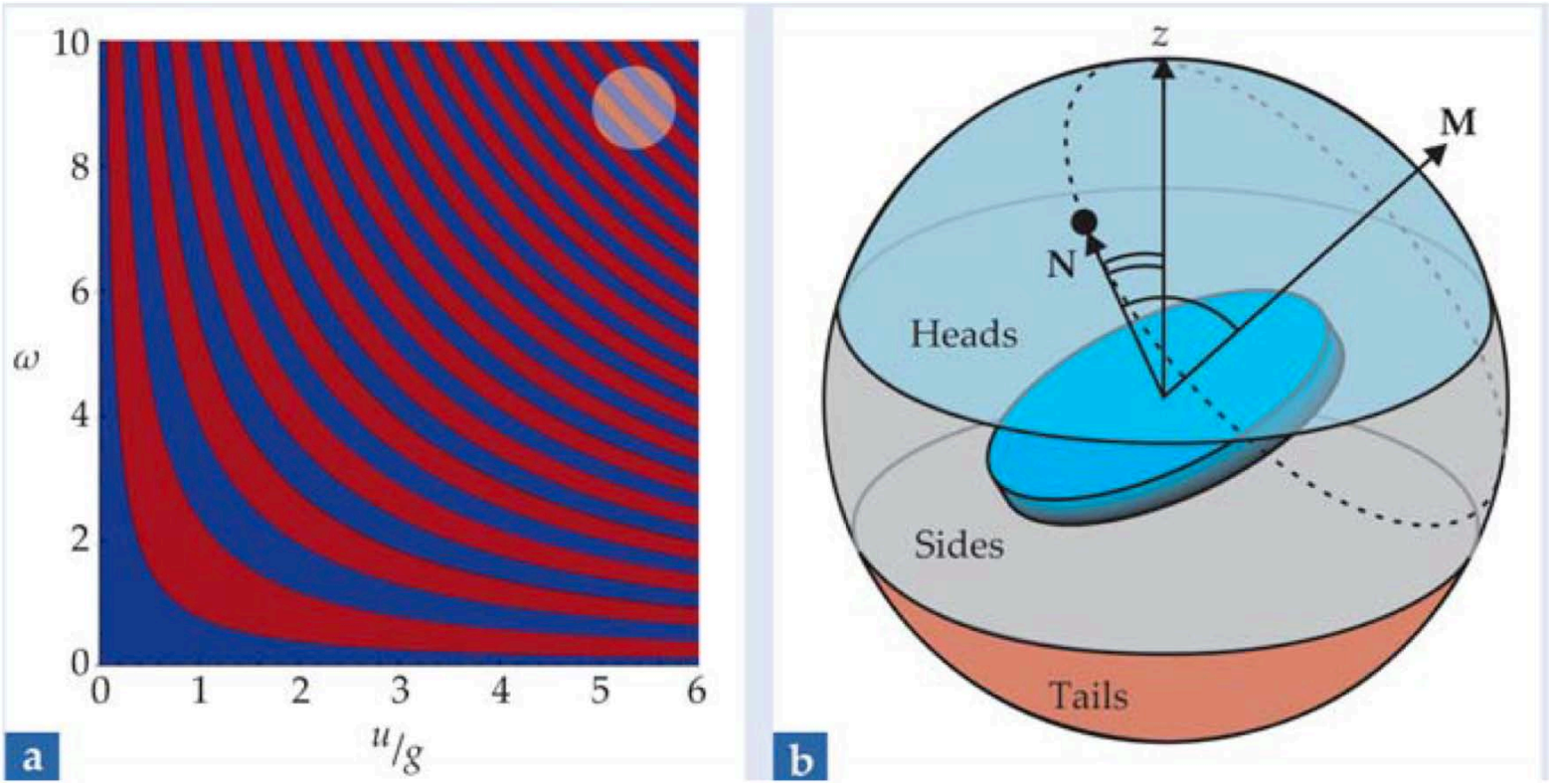
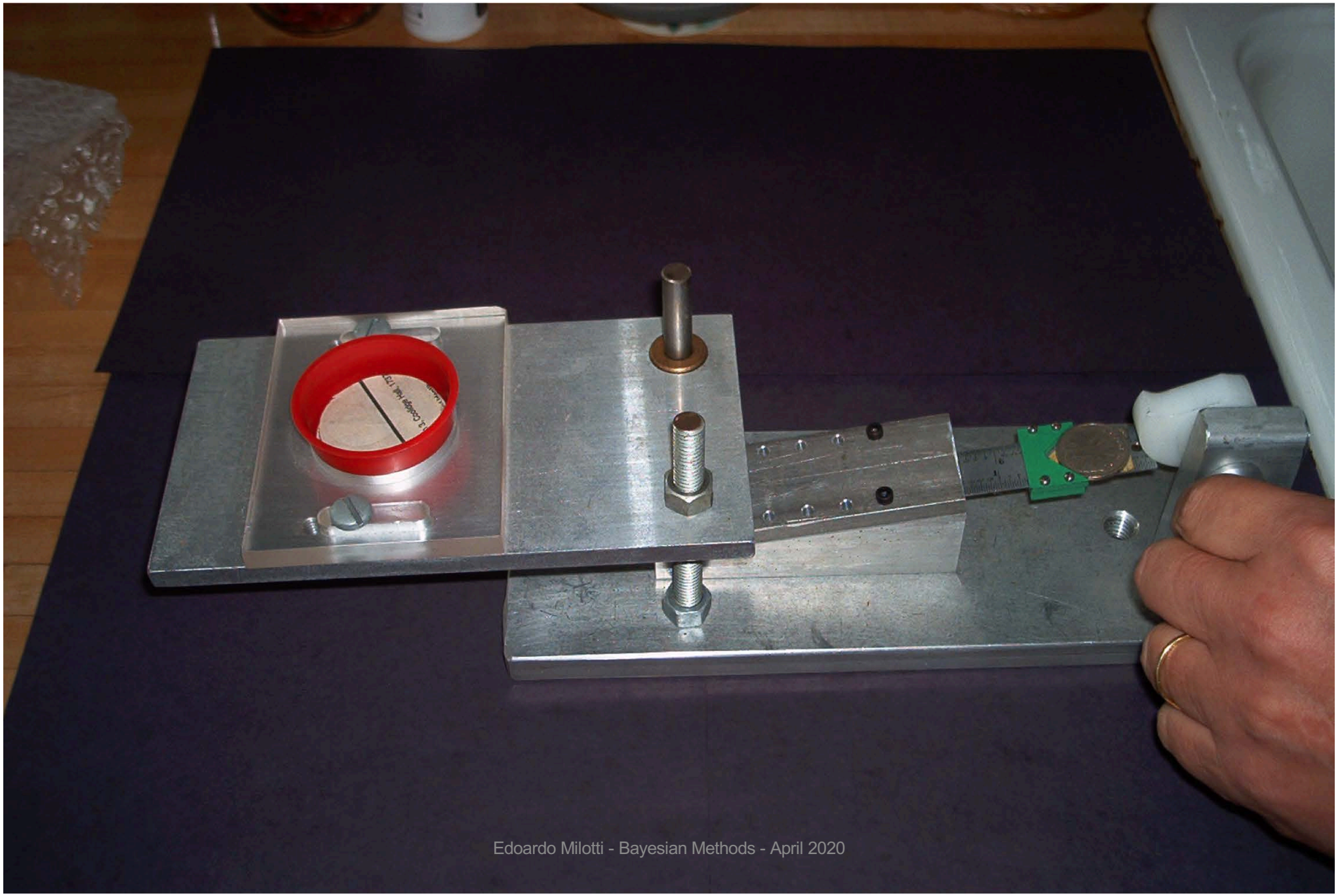


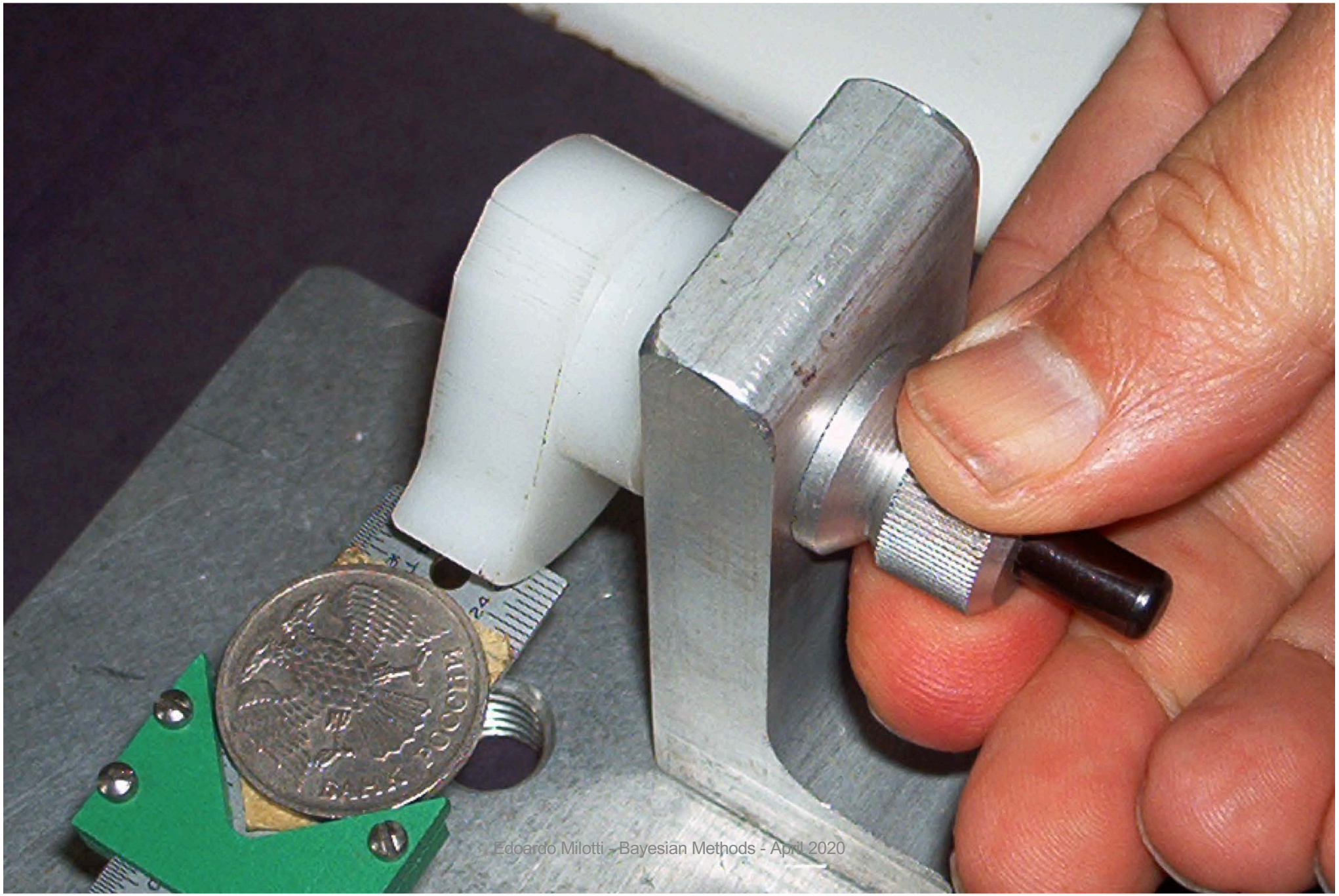
figure from Mahadevan and Yong,
"Probability, physics, and the coin toss",
Phys. Today, July 2011, pp. 66-67

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)

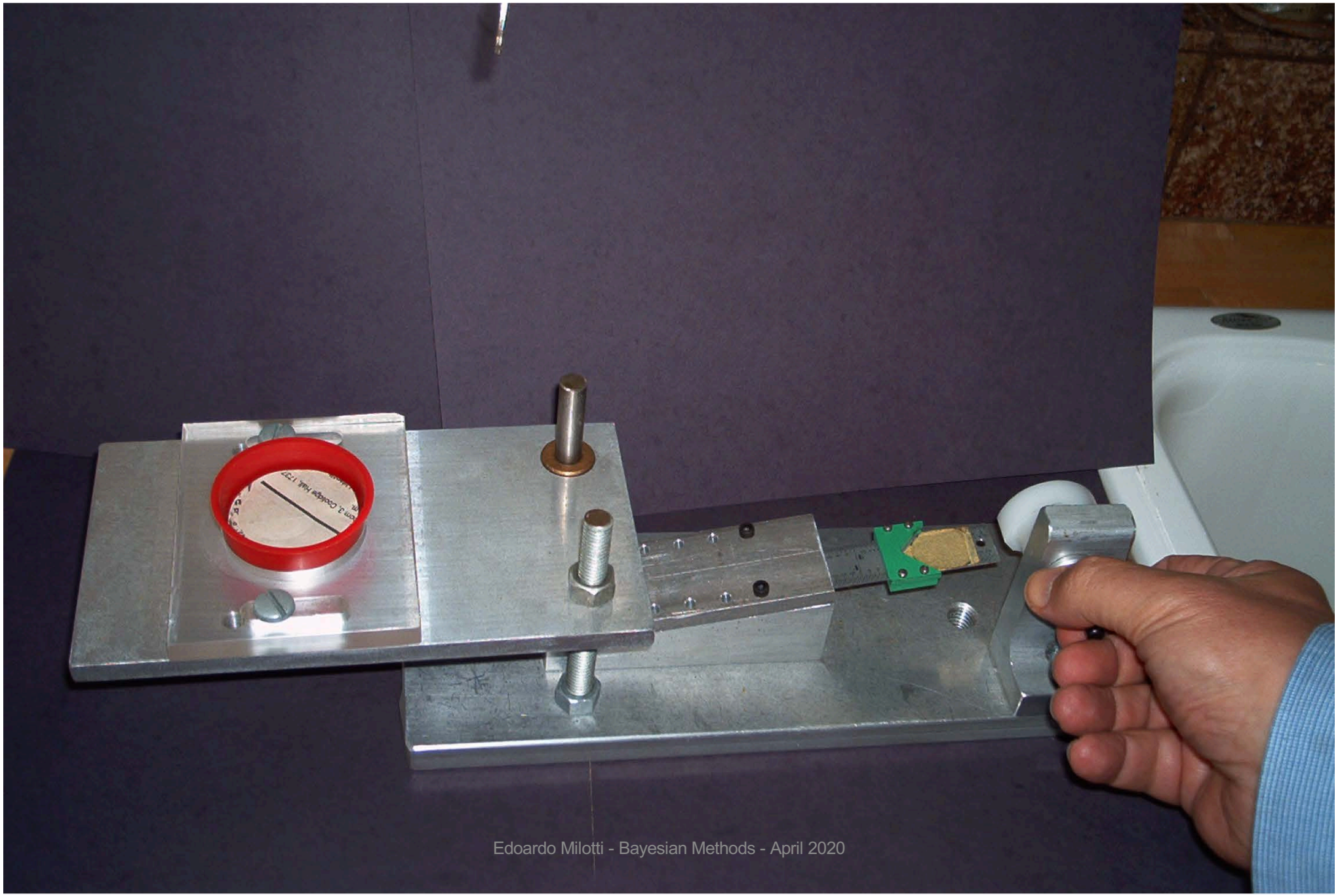
Coin tossing machine (Diaconis, Holmes and Montgomery 2007)



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.

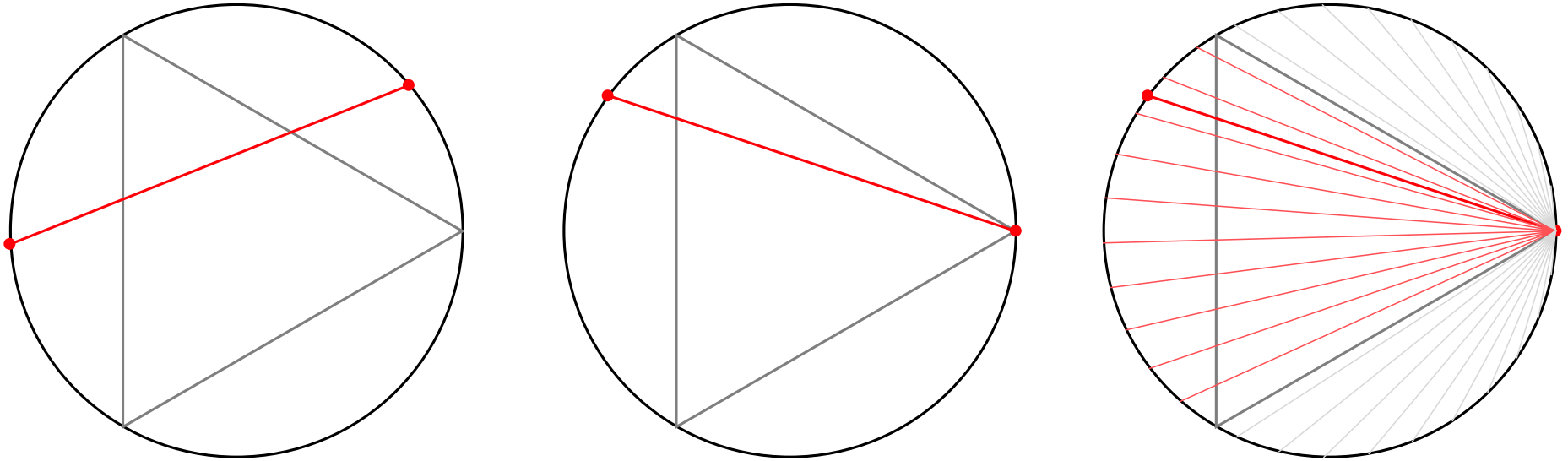
Thus – at least in this case – probabilities are a measure of our own ignorance rather than an intrinsic property of the physical system.

Bertrand's paradox and the ambiguities of probability models

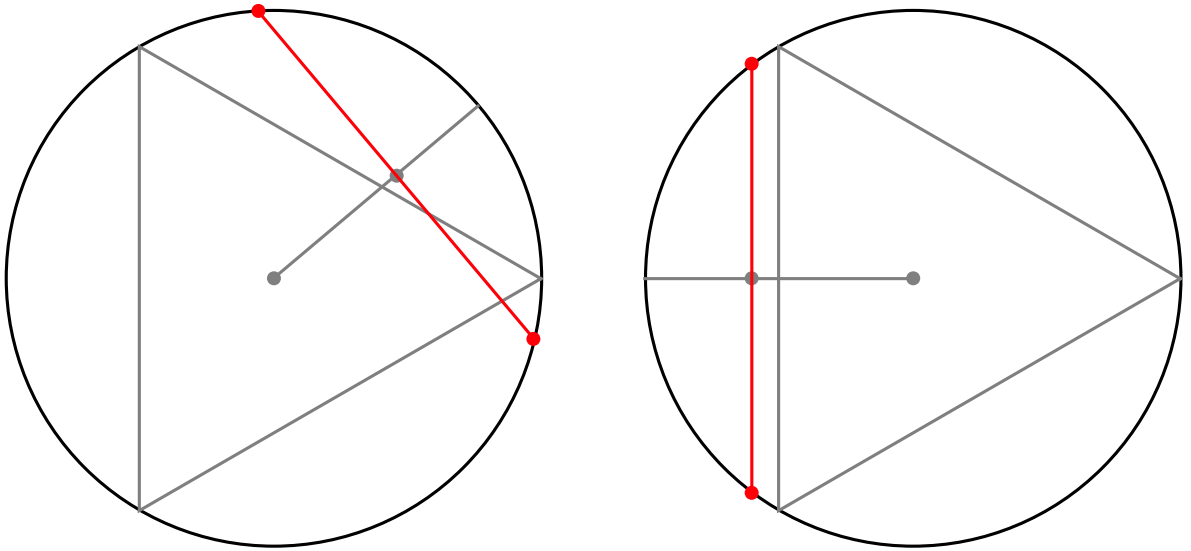
Bertrand's paradox goes as follows:

"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?"

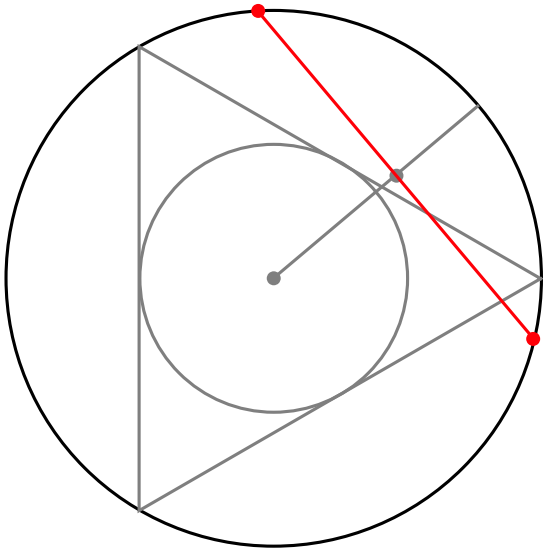
(Bertrand, 1889)



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.



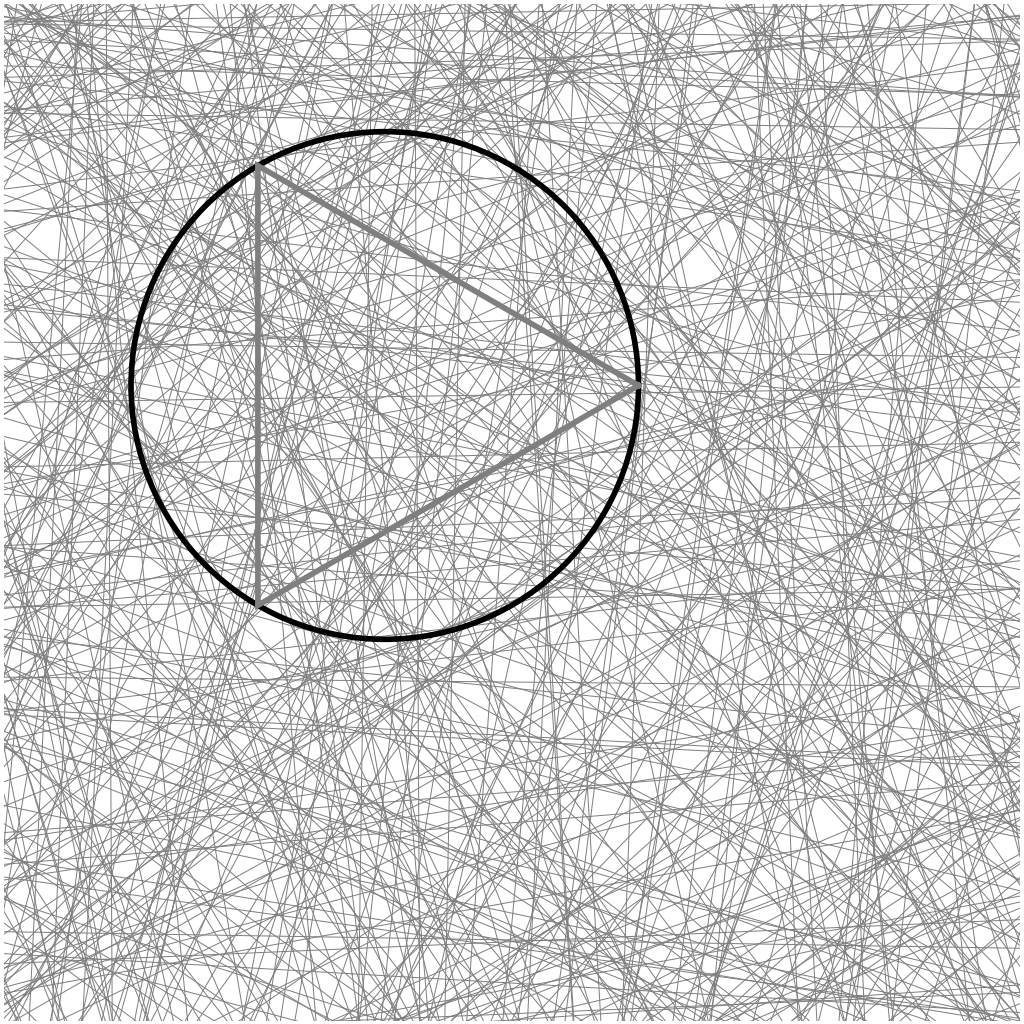
Solution 2: we take first a random radius, and next we choose a random point on this random radius. Then, we take the chord through this point and perpendicular to the radius. When we rotate the triangle so that the radius is perpendicular to one of the sides, we see that half of the points give chords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

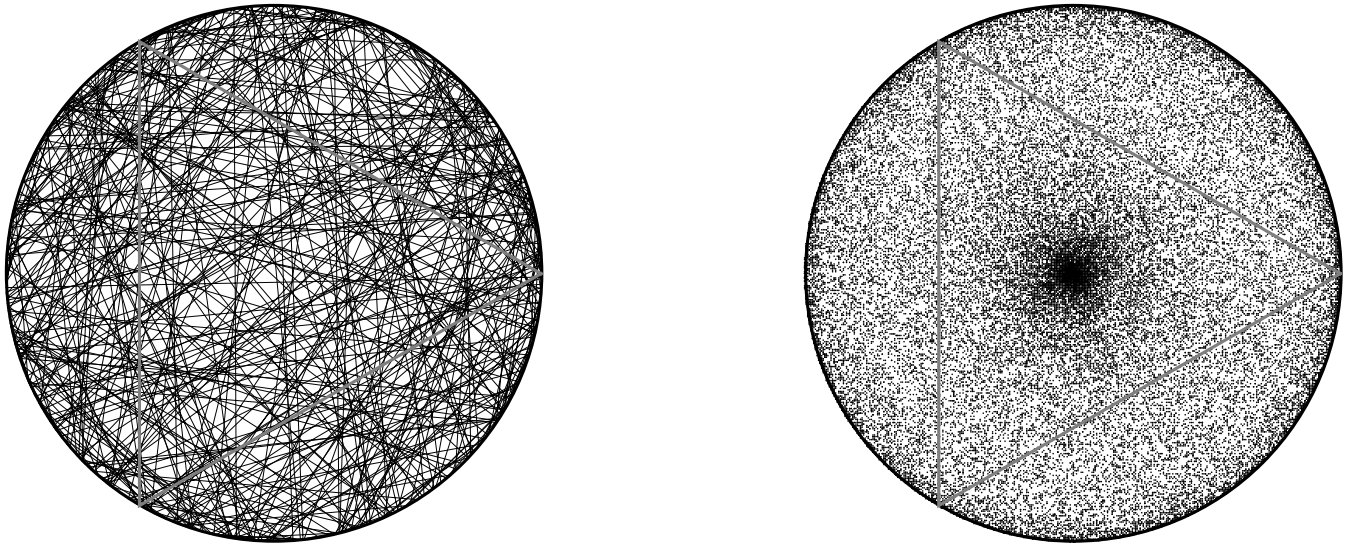


Solution 3: we take the chord midpoints located inside the circle inscribed in the triangle, and we obtain chords that are longer than one side of the triangle. Since the ratio of the areas of the two circles is 1/4, we find that now the probability of drawing a long chord is just 1/4.

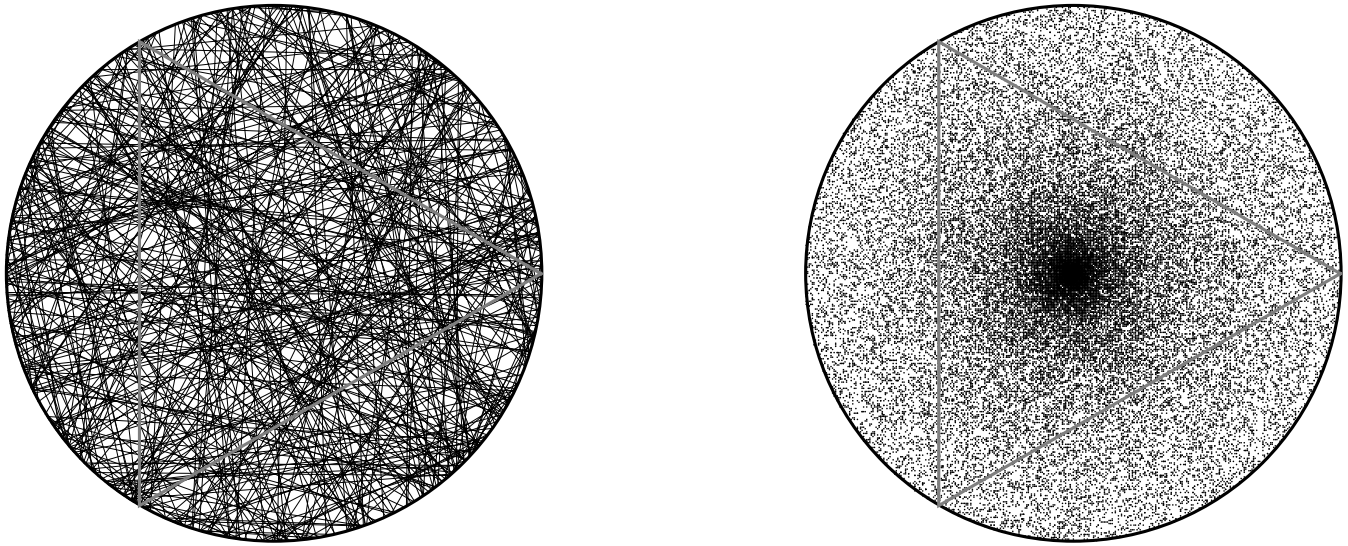
At least 3 different "solutions": which one is correct, and why?

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



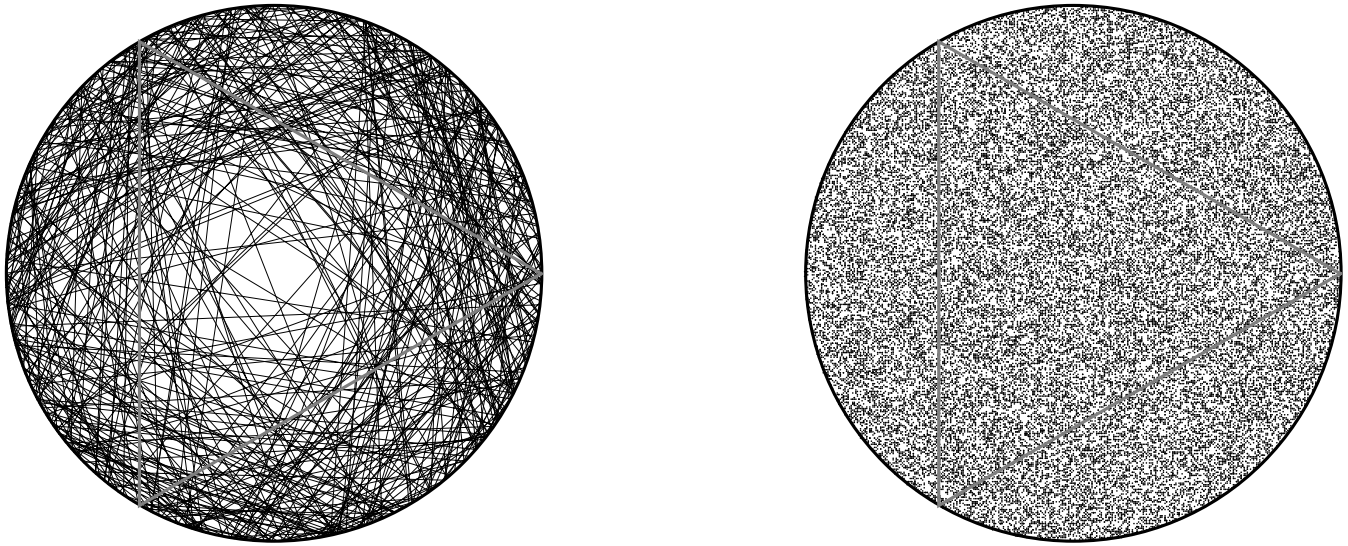


Distribution 1: distribution of chords (left panel) and of midpoints (right panel) in the first solution of Bertrand's paradox (the left panel shows 400 chords, the right panel shows 100000 midpoints).

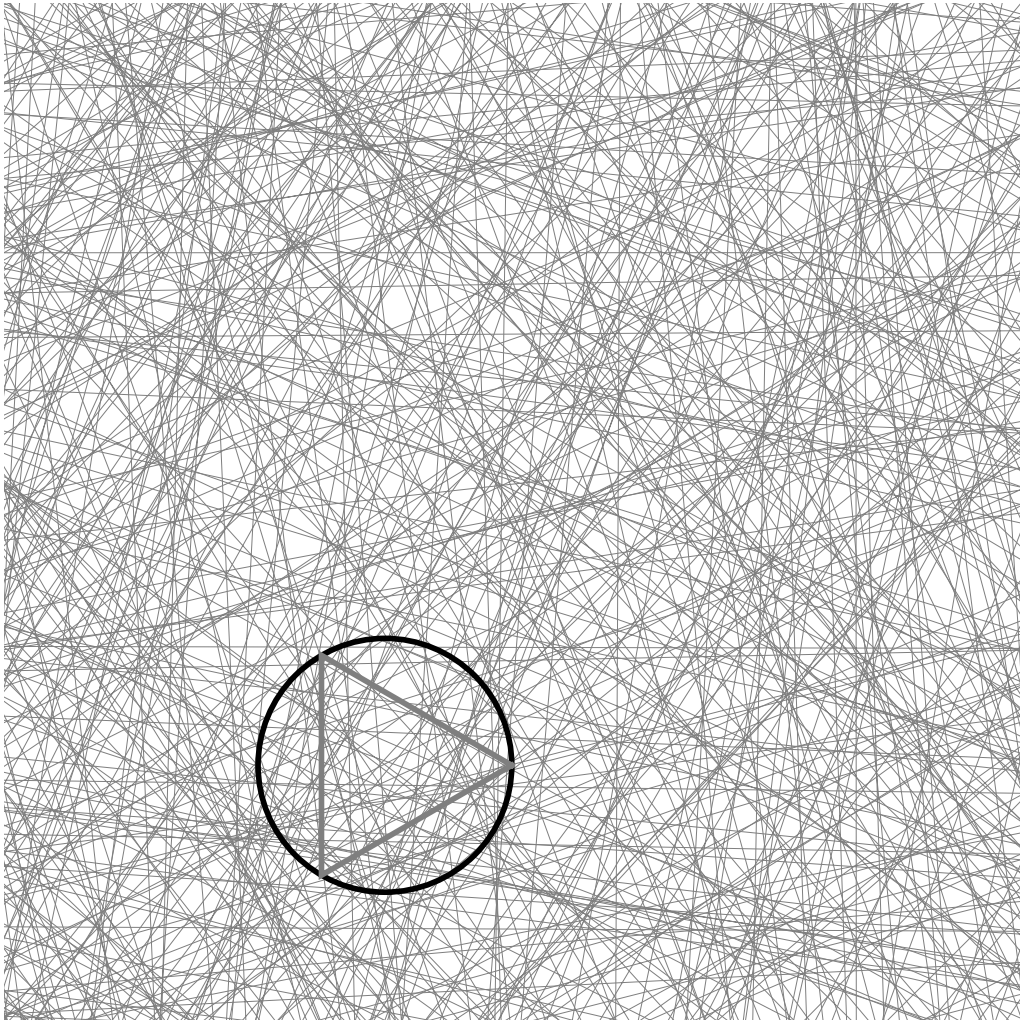


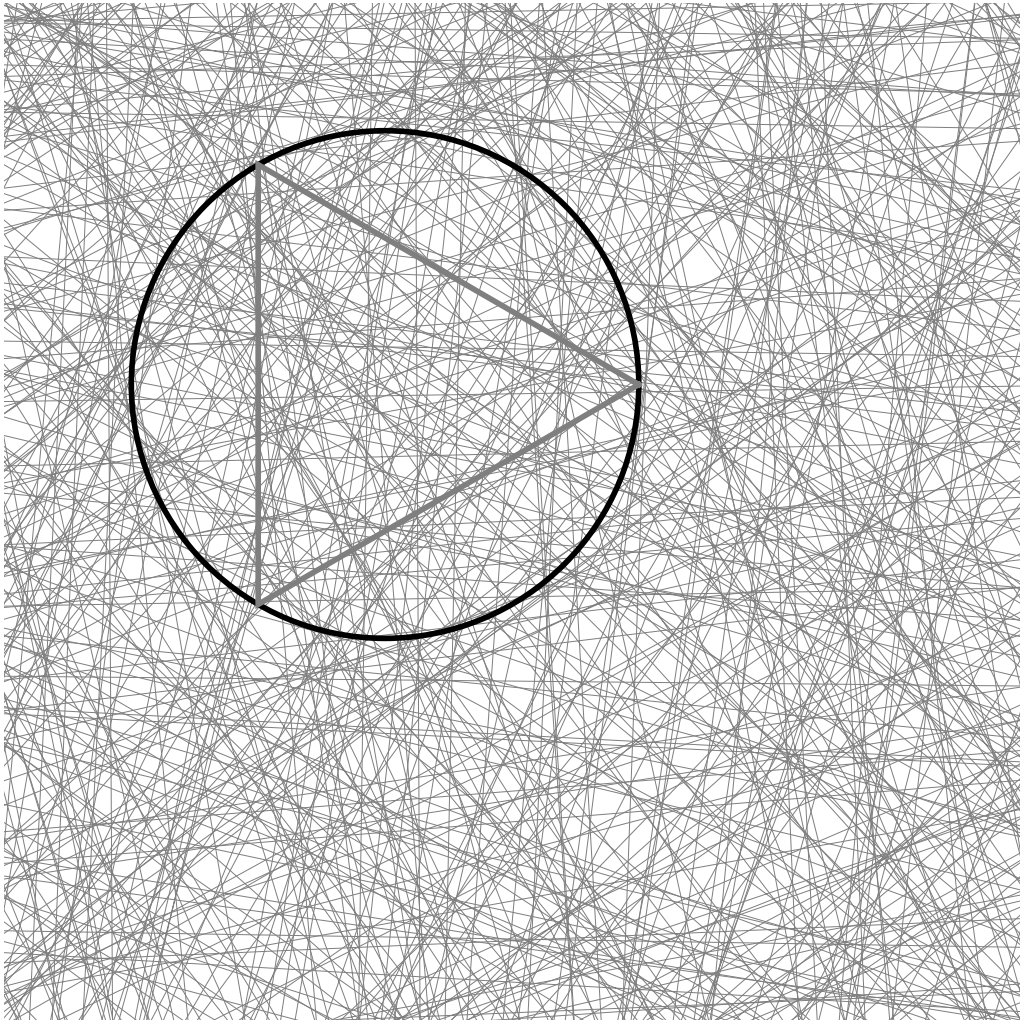
Distribution 2: Distribution of chords (left panel) and of midpoints (right panel) in the second solution of Bertrand's paradox (the left panel shows 400 chords, the right panel shows 100000 midpoints).

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.



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

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

and inserting this into equation

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

we find

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

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

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

i.e.,

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

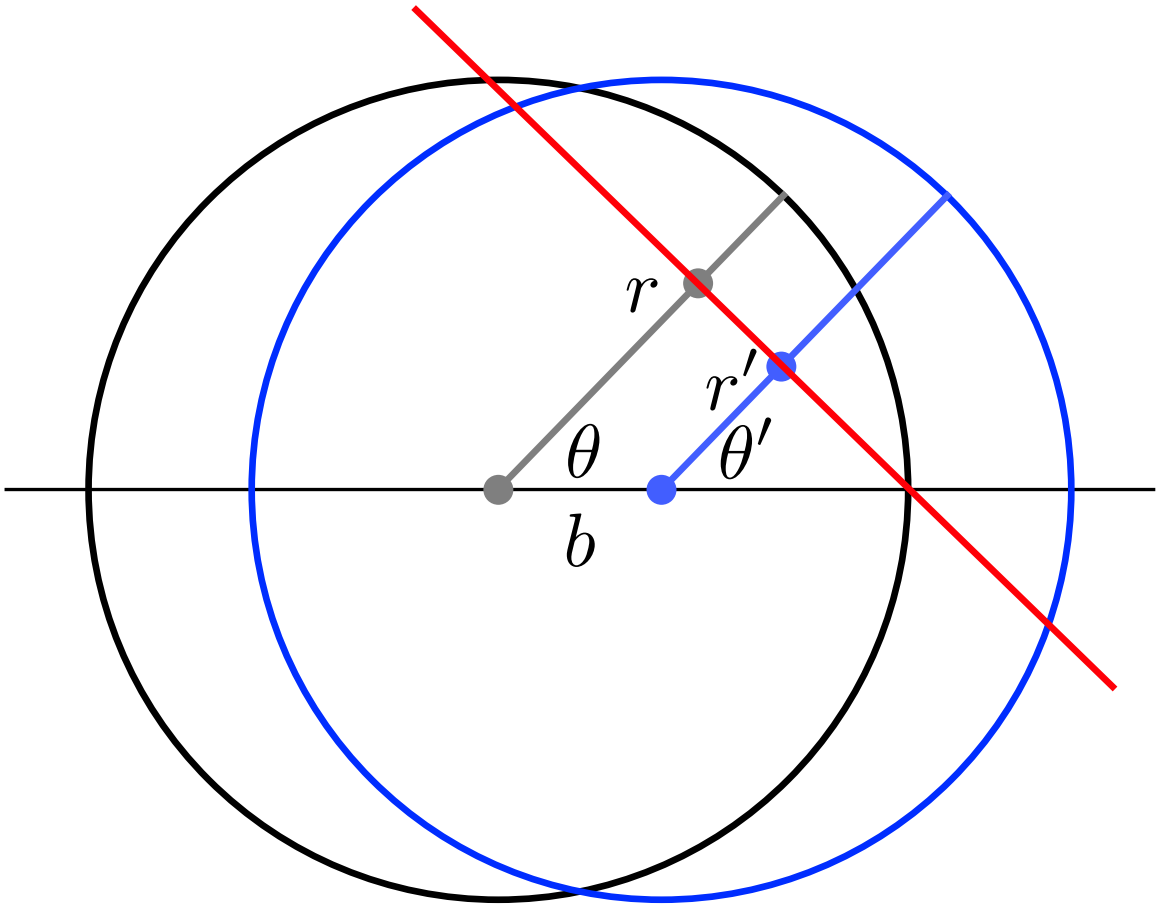
where the constant $q = 2\pi R^2 f(R)$ is unknown. However, we can still solve the equation and find

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

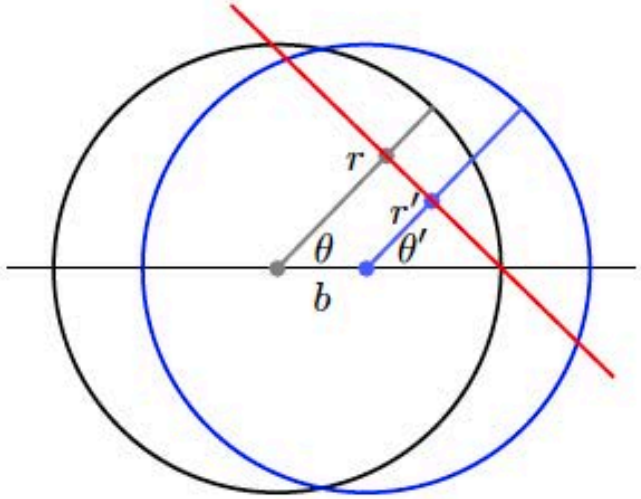
The constant A is easy to find from the normalization condition: $A = q/2\pi R^q$, and therefore

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

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

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

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

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

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

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

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

Therefore

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

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

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

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.

A way forward to "objective" priors: Jeffreys' priors

An invariant form for the prior probability in estimation problems

By Harold Jeffreys, F.R.S.

(*Received* 23 *November* 1945)

It is shown that a certain differential form depending on the values of the parameters in a law of chance is invariant for all transformations of the parameters when the law is differentiable with regard to all parameters. For laws containing a location and a scale parameter a form with a somewhat restricted type of invariance is found even when the law is not everywhere differentiable with regard to the parameters. This form has the properties required to give a general rule for stating the prior probability in a large class of estimation problems.

Starting remark: here we concentrate on a problem of parametric statistics.

We seek the parameter that best adapts our theory to data. This means that we search among incompatible hypotheses that are defined by different parameter values.

The different hypotheses (and therefore, the different parameters) correspond to different pdf's

$$p(x|\theta)$$

Step 1: Bartlett identities for a parametric pdf family

$$\mathbf{E}\left[\frac{\partial \ln p(x|\theta)}{\partial \theta}\right] = 0$$

$$\mathbf{E}\left[\frac{\partial^2 \ln p(x|\theta)}{\partial \theta^2}\right] = -\mathbf{E}\left[\left(\frac{\partial \ln p(x|\theta)}{\partial \theta}\right)^2\right]$$

Step 2: a parameter-dependent Likelihood is a family of pdf's that represent the distribution of the data, given the value of the parameter(s).

It can be shown that the following inequality holds

$$\mathrm{var}[\hat{\theta}(D)] \geq = \frac{1}{\mathbf{E}\left[\left(\frac{\partial \ln L(D,\theta_0)}{\partial \theta_0}\right)^2\right]} = \frac{1}{-\mathbf{E}\frac{\partial^2 \ln L(D,\theta_0)}{\partial \theta_0^2}}$$

where θ_0 is the "true" value of the parameter, and $\hat{\theta}(D)$ is the ML estimator (Cramer-Rao-Fisher bound).

Step 3: definition of Fisher Information. A very concentrated pdf is very informative. Therefore, the smaller the variance, the greater the "information".

Thus, from the Cramer-Rao-Fisher bound

$$\mathrm{var}[\hat{\theta}(D)] \geq= \frac{1}{\mathbf{E}\left[\left(\frac{\partial \ln L(D,\theta_0)}{\partial \theta_0}\right)^2\right]} = \frac{1}{-\mathbf{E}\frac{\partial^2 \ln L(D,\theta_0)}{\partial \theta_0^2}}$$

one is led to the Fisher Information

$$I(\theta) = \mathbf{E}\left[\left(\frac{\partial \ln p(x,\theta)}{\partial \theta}\right)^2\right] = -\mathbf{E}\frac{\partial^2 \ln p(x,\theta)}{\partial \theta^2}$$

Step 4: it can be shown that the Fisher Information is a local (and symmetrical) form of the Kullback-Leibler divergence.

$$I_{KL}\left(p(x|\theta), p(x|\theta+\epsilon)\right) = -\frac{1}{2}\mathbf{E}\left[\frac{\partial^2 \ln p(x|\theta)}{\partial \theta^2}\right]\epsilon^2 = \frac{1}{2}I(\theta)\epsilon^2$$

From this, and from the properties of the KL divergence, we see that the Fisher Information behaves like a (squared) distance between distributions.

Step 5: the KL divergence is invariant with respect to parameter transformations. From the definition of KL divergence, and from the transformation formula for pdf's we find

$$\int_{-\infty}^{+\infty} p_y(y) \ln\left(\frac{p_y(y)}{q_y(y)}\right) dy = \int_{-\infty}^{+\infty} p_x(x) \ln\left(\frac{p_x(x)\left|\frac{dx}{dy}\right|}{q_x(x)\left|\frac{dx}{dy}\right|}\right) dy$$

$$= \int_{-\infty}^{+\infty} p_x(x) \ln\left(\frac{p_x(x)}{q_x(x)}\right) dy$$

Therefore, the Fisher Information is also invariant with respect to parameter transformations.

Step 6: from the equation that relates KL divergence and Fisher Information, we find a corresponding pdf:

$$I_{KL}\left(p(x|\theta), p(x|\theta+\epsilon)\right) = -\frac{1}{2}\mathbf{E}\left[\frac{\partial^2 \ln p(x|\theta)}{\partial\theta^2}\right]\epsilon^2 = \frac{1}{2}I(\theta)\epsilon^2$$

$$f(\theta) \sim \sqrt{I(\theta)}$$

This pdf is invariant with respect to parameter transformations. Taking this pdf for the parameter amounts to taking a uniform distribution in the space of parameterized pdf's.

Example: a simple Gaussian Likelihood for *n* datapoints

$$L(D|\mu) = \prod_n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_n-\mu)^2}{2\sigma^2}\right)$$

$$\Rightarrow \quad \ln L(D|\mu) \sim \sum_n \left(-\ln\sigma - \frac{(x_n-\mu)^2}{2\sigma^2}\right)$$ fixed sigma

$$\Rightarrow \quad I(\mu) = \mathbf{E}\left[-\frac{\partial^2 \ln L(D|\mu)}{\partial\mu^2}\right] \sim \text{constant}$$

This points to a uniform prior for mu. In general, this uniform prior is an improper prior.

Example: a simple Gaussian Likelihood for *n* datapoints (ctd.)

$$L(D|\mu) = \prod_n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_n-\mu)^2}{2\sigma^2}\right)$$

$$I(\sigma) = \mathbf{E}\left[-\frac{\partial^2 \ln L(D|\sigma)}{\partial\sigma^2}\right] \sim \frac{1}{\sigma^2}$$

fixed mu

$$\sqrt{I(\sigma)} \sim \frac{1}{\sigma}$$

This power-law prior is another improper prior.