EPFL

Probability Essay

Eliot Dugelay and Kilian Wan

This paper is about two probabilistic paradoxes: Borel-Kolmogorov paradox and Bertrand's paradox.

I. Borel-Kolmogorov paradox

"The concepts of conditional probability and expected value with respect to a σ-field underlie much of modem probability theory. The difficulty in understanding these ideas has to do not with mathematical detail so much as with probabilistic meaning..." (Billingsley, 1995)

In general, the paradox is accredited to Émile Borel in 1909, but in reality it was originally formulated by Joseph Bertrand in 1889. Andreï Kolmogorov first provided a relevant answer in 1933; the same year he postulated his measure-theoric probability theory. It relates to conditional probability with respect to an event of probability zero. Indeed, conditional probability of an event A given the event B happens is:

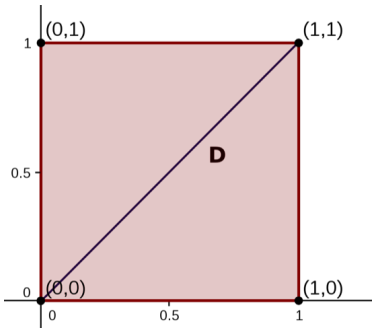
$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)},$$

this formula is well-defined if $\mathbb{P}(B) > 0$. In this sense, the Borel-Kolmogorov paradox suggests that in the case where $\mathbb{P}(B) = 0$, the conditional probability is indeterminate or ill-posed. We are going to present two examples manifesting this phenomena.

A. Choosing a random point uniformly on the unit square

Firstly, consider the unit square Ω. The vector $Z = (X, Y)$ composed of the coordinates X and Y on the square represents a uniform random vector on $[0, 1]^2$, in other terms X and Y are both uniform random variables on $[0, 1]$.

Secondly, consider the event that the point (X, Y) lies on the diagonal D, i.e. $X = Y$:



Notice that in reality this diagonal has a zero area and thus we have

$$\mathbb{P}((X,Y) \in D) = 0.$$

Now, let's use two similar reasonings to compute the conditional probability that the X coordinate of a point is greater than $\frac{1}{2}$ knowing the point is on the diagonal:

$$\mathbb{P}(X > \frac{1}{2} \mid (X,Y) \in D).$$

1) For the first one, consider a new random variable $U = X - Y$. Notice, then,

$$\mathbb{P}(X > \frac{1}{2} \mid (X,Y) \in D) = \mathbb{P}(X > \frac{1}{2} \mid U = 0)$$

Using Definition 4.19, we compute:

$$f_{X|U=0}(x) = \frac{f_{X,U}(x,0)}{f_U(0)}$$

Let's compute $f_{X,U}(x,0)$. To do this, first of all we define a diffeomorphism $\phi : U \to V$ with U, V open connected subsets of $\mathbb{R}^2$ such that, $\phi(X,Y) = (X,U)$ with $\phi(x,y) = (x, x-y)$ and $\phi^{-1}(x,u) = (x, x-u)$. Thus, by the Proposition 4.14,

$$f_{X,U}(x,u) = \frac{1}{|\mathbb{J}_\phi(\phi^{-1}(x,u))|} f_{X,Y}(\phi^{-1}(x,u))$$

However, here $|\mathbb{J}_\phi(\phi^{-1}(x,u))| = 1$ and

$$f_{X,Y}(x,y) = \begin{cases} 1 & \text{if } 0 \leq x,y \leq 1, \\ 0 & \text{otherwise} \end{cases}$$

We obtain thus,

$$f_{X,U}(x,u) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \text{ and } x-1 \leq u \leq x, \\ 0 & \text{otherwise.} \end{cases}$$

Next, let's compute $f_U(0)$ with the convolution formula (Corollary 4.16):

$$f_U(u) = \int_{-\infty}^{\infty} f_X(x) f_{-Y}(u-x)dx = 1 - |u|,$$

the distribution of U is called the triangular distribution. Finally, we find:

$$\mathbb{P}(X > \frac{1}{2} | U = 0) = \int_{\frac{1}{2}}^{1} f_{X|U=0}(x)dx = \int_{\frac{1}{2}}^{1} 1dx = \frac{1}{2}.$$

2) For the second manner, consider a new random variable $V = \frac{X}{Y}$. Then, now,

$$\mathbb{P}(X > \frac{1}{2} \mid (X,Y) \in D) = \mathbb{P}(X > \frac{1}{2} \mid V = 1)$$

We use the same reasoning as before to compute this probability by using the conditional density $f_{X|V=1}(x)$. Hence, here our diffeomorphism is defined as

$$\psi(X,Y) = (X,V)$$

with $\psi(x,y) = (x, \frac{y}{x})$ so

$$\psi^{-1}(x,v) = (x, xv)$$

By similar computations, we obtain that:

$$f_{X,V}(x,v) = \begin{cases} x & \text{if } 0 \leq x \leq 1 \text{ and } 0 \leq v \leq \frac{1}{x}, \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, we need to find the marginal density $f_V(v)$ by Lemma 4.12 we get:

$$f_V(v) = \int_{-\infty}^{\infty} f_{X,V}(x,v)dx = \begin{cases} \frac{1}{2} & \text{if } 0 \leq v \leq 1, \\ \frac{1}{2v^2} & \text{if } v \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Giving, $f_{X|V=1}(x) = \frac{f_{X,V}(x,1)}{f_V(1)} = 2x$ and then,

$$\mathbb{P}(X > \frac{1}{2} | V = 1) = \int_{\frac{1}{2}}^{1} f_{X|V=1}(x)dx = \int_{\frac{1}{2}}^{1} 2xdx = \frac{3}{4}.$$

We have found on one hand $\frac{1}{2}$ and on the other hand $\frac{3}{4}$, whereas both computations are totally correct.

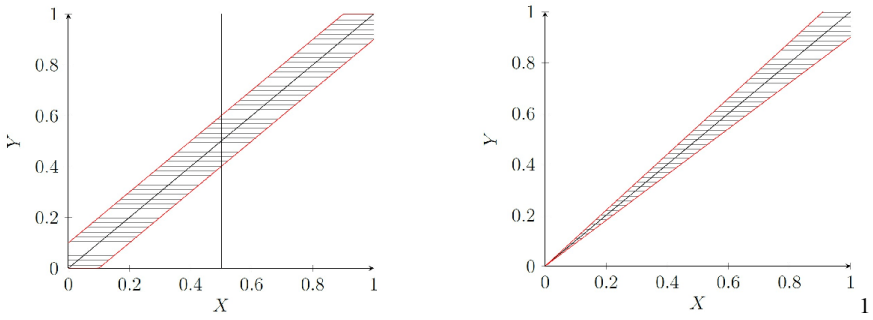
Let's try to find a visual explanation. First, notice that:

$$\mathbb{P}(X > \frac{1}{2}|U = 0) = \lim_{\epsilon \to 0} \mathbb{P}(X > \frac{1}{2} \ \big| |U| \leq \epsilon)$$

and

$$\mathbb{P}(X > \frac{1}{2}|V = 1) = \lim_{\epsilon \to 0} \mathbb{P}(X > \frac{1}{2} \ \big| |V - 1| \leq \epsilon)$$

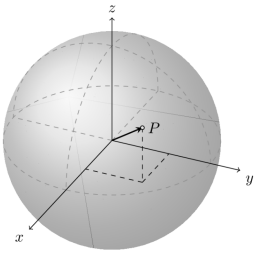
Hence, the two previous cases can be visualized as follows:



The figure A represents our first case. We observe that points with $X > \frac{1}{2}$ in the grey region occupy half of the total grey area giving our conditional probability of $\frac{1}{2}$. While, on the figure B, representing the second case, points with $X > \frac{1}{2}$ in the grey area occupy much more than half of the total grey area giving a conditional probability greater than $\frac{1}{2}$.

B. The paradox of the sphere

Suppose you pick randomly a point P on a perfect sphere. We are going to compute the conditional probability that the point lies on a subarc of a great circle knowing it is on such a great circle. Without loss of generality, we can suppose the great circle is the equator (latitude 0°).



[1]Alexander Meehan and Snow Zhang, "The Borel-Kolmogorov Paradox Is Your Paradox Too: A Puzzle for Conditional Physical Probability".

What we are going to do is define two different parametrizations on the sphere:

$$P_1 : \begin{cases} x = \cos\theta\cos\phi \\ y = \cos\theta\sin\phi \\ z = \sin\theta \end{cases} \qquad P_2 : \begin{cases} x = \sin\theta \\ y = \cos\theta\sin\phi \\ z = -\cos\theta\cos\phi \end{cases}$$

with $\theta \in [\frac{-\pi}{2}, \frac{\pi}{2}]$, $\phi \in [0, 2\pi)$. Notice that P_2 is P_1 but with a $90-$degree rotation around the y axis.

To simplify the computation, we will restrain us to the first octant of the sphere.

1) Let's see what happens for the first parametrization. We define two independent random variables:

$$\Theta = \arcsin(Z) \qquad \text{and} \qquad \Phi = \arctan_2\left(\tfrac{Y}{\sqrt{1-Z^2}}, \tfrac{X}{\sqrt{1-Z^2}}\right)$$

where Φ is a vector of $\mathbb{R}^2$ and the $\arctan_2$ function is the 2-argument arctangent[2].
The joint probability density function is given by:

$$f_{\Theta,\Phi}(\theta,\phi) = \frac{1}{4\pi}\cos\theta$$

Let's set

$$A = \left\{0 < \Phi < \tfrac{\pi}{4}\right\} = \{0 < X < 1, 0 < Y < X\} \qquad \text{and} \qquad B = \{\Theta = 0\} = \{Z = 0\}$$

In this way, we can see that the event "$A|B$" draws effectively the conditional probability that the point lies on a subarc of the equator. Hence, we compute the probability that the event A happens given B (noticing that $\mathbb{P}(B) = 0$):

$$\mathbb{P}(A|B) = \lim_{\epsilon\to 0}\frac{\mathbb{P}(A\cap B_\epsilon)}{\mathbb{P}(B_\epsilon)} \underbrace{=}_{\text{independent}} \lim_{\epsilon\to 0}\frac{\mathbb{P}(A)\cdot\mathbb{P}(B_\epsilon)}{\mathbb{P}(B_\epsilon)} = \mathbb{P}(0 < \Phi < \frac{\pi}{4}) = \frac{1}{8},$$

where $B_\epsilon = \{|\Theta| < \epsilon\} \to B$ when $\epsilon \to 0$, and the independence of A and B_ϵ comes from the independence of the two random variables.

2) Now, we take the second parametrization, and we do the same process: we redefine two independent random variables:

$$\Theta' = \arcsin(X) \qquad \text{and} \qquad \Phi' = \arctan_2\left(\tfrac{Y}{\sqrt{1-X^2}}, \tfrac{-Z}{\sqrt{1-X^2}}\right)$$

[2] $\arctan_2(y, x)$ is the argument of the complex number $x + iy$.

Furthermore, we have that the rotation that we have applied to get our second parametrization, is measure-preserving. Hence the joint density function is also given by:

$$f_{\Theta',\Phi'}(\theta,\phi) = \frac{1}{4\pi}\cos\theta$$

We redefine our two sets A and B with

$$A = \left\{0 < \Phi < \frac{\pi}{4}\right\} = \{0 < X < 1, 0 < Y < X\} = \left\{0 < \Phi' < \pi, 0 < \Theta' < \frac{\pi}{2}, \sin(\Phi') < \tan(\Theta')\right\},$$

and

$$B = \{\Theta = 0\} = \{Z = 0\} = \left\{\Phi' = -\frac{\pi}{2}\right\} \cup \left\{\Phi' = \frac{\pi}{2}\right\}$$

Hence we compute the probability that the event A happens given B (noticing that $\mathbb{P}(B) = 0$) by using again that $B_\epsilon = \{|\Phi' - \frac{\pi}{2}| < \epsilon\} \cup \{|\Phi' + \frac{\pi}{2}| < \epsilon\} \to B$ when $\epsilon \to 0$:

$$\begin{aligned}
\mathbb{P}(A\,|B) &= \lim_{\epsilon\to 0} \frac{\mathbb{P}(A \cap B_\epsilon)}{\mathbb{P}(B_\epsilon)} \\
&= \lim_{\epsilon\to 0} \frac{1}{\frac{4\epsilon}{2\pi}} \mathbb{P}\left(\frac{\pi}{2} - \epsilon < \Phi' < \frac{\pi}{2} + \epsilon, 0 < \Theta' < \frac{\pi}{2}, \sin(\Phi') < \tan(\Theta')\right) \\
&= \frac{\pi}{2} \lim_{\epsilon\to 0} \frac{\partial}{\partial\epsilon} \int_{\pi/2-\epsilon}^{\pi/2+\epsilon} \int_0^{\pi/2} \mathbb{1}_{\sin(\phi)<\tan(\theta)} f_{\Theta',\Phi'}(\theta,\phi)\, d\theta\, d\phi \\
&= \pi \int_0^{\pi/2} \mathbb{1}_{1<\tan(\theta)} f_{\Theta',\Phi'}(\theta, \frac{\pi}{2})\, d\theta \\
&= \pi \int_{\pi/4}^{\pi/2} \frac{1}{4\pi}\cos(\theta)\, d\theta \\
&= \frac{1}{4}\left(1 - \frac{1}{\sqrt{2}}\right) \neq \frac{1}{8}
\end{aligned}$$

by using the Hôspital-Bernoulli rule, and the integral with variable bounds (Analysis II with Buffoni)

After all this, we observe that even if different computations are correct, they will not necessarily give the same result. Several conclusions can be drawn from this. The first and most important one is when conditioning, especially on a set of measure 0, it is necessary to provide the accompanying sub-σ-algebra of the event that is conditioned on. Secondly, we must accept that conditional probability on sets of measure 0 is not uniquely defined.
The Borel-Kolmogorov paradox was formulated at a time when the conceptual foundations of probability theory were not yet entirely clear. In particular, conditional probability was restricted to Bayes' rule, which is a limited concept. It was only in 1933, with the work of Kolmogorov that the abstract conceptual structure of probability theory became clarified. Thus, the notion of conditional expectation, introduced by Kolmogorov, became crucial in probability theory. Lastly, we could say that "conditional probabilities are truly conditional": they depend on conditions, i.e. on a σ-algebra.

II. Bertrand's paradox

Before trying to understand the paradox, let's do a little introduction of it. Joseph Bertrand was a french mathematician that studied the classical interpretation of probability theory. In fact, Bertrand introduced his paradox by showing an example to prove that probabilities may not be well defined if the method that produces the random variable is not clearly defined. He introduced this paradox in his work *Calcul des probabilités (1889)*.

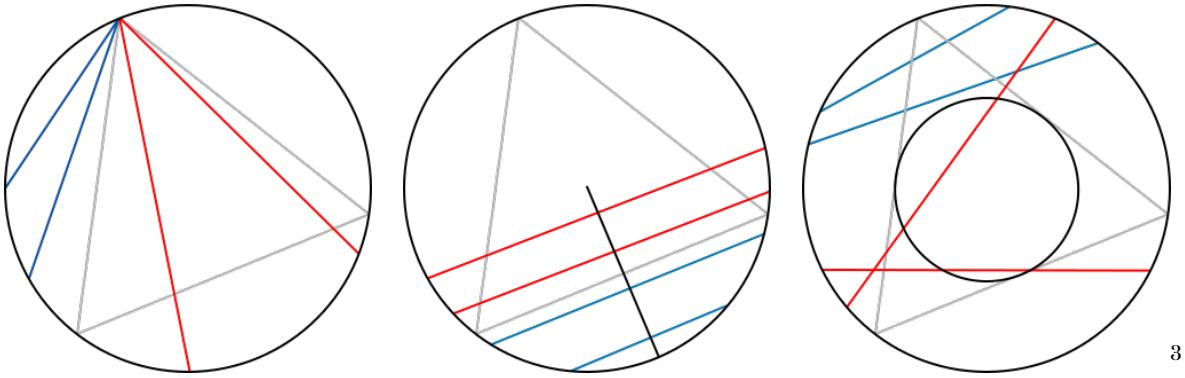
Now that we've done a little historical introduction, let's see the paradox. The Bertrand's paradox goes as follows: we consider an equilateral triangle inscribed in a circle. And we suppose that we choose a random chord of the circle. The question is: what is the probability that the length of the chord is longer than the side of the triangle. To answer this question, we are going to see 3 methods to find the probability. We will see them with a visual approach instead of a mathematical approach.

The first one is called the *method of random endpoints*. The idea of this method is to choose randomly two points on the circle and calculate the distance between these two points. Without loss of generality, we can say that the first point is chosen on one of the vertices of the triangle. Indeed, by a little rotation we have the desired result. Therefore, the only important thing is the position of the second point relative to the first one. Indeed, we observe that if the other chord endpoint lies on the arc on the opposite side of the triangle compared to our first point, the length of the chord is going to be longer than a side of the triangle. Furthermore, the length of the arc is one third of the circumference of the circle, therefore the probability that a random chord is longer than a side of the triangle is $\frac{1}{3}$.

The second method is called the *random radius method*. For this method, we first choose a radius of the circle (by rotation we may assume this radius is perpendicular to one side of the triangle), a point on the radius and construct the chord through this point and perpendicular to the radius. As a matter of fact, the length of the chord is going to be larger than a side of the triangle if the chosen point is nearer the center of the circle than the point where the side of the triangle intersects the radius. We have that the side of the triangle bisects the radius, so that means that the probability of a random chord is longer than a side of the triangle is $\frac{1}{2}$.

Finally, the third method is the *random midpoint*. We use the fact that a chord is fully determined by its midpoint. Chords whose length exceeds the side of the triangle have their midpoint inside a smaller circle with radius equal to $1/2$ that of the given one. But the area of the smaller circle is $1/4$ of the area of the larger circle. Hence the probability that a random chord is longer than a side of the triangle is $\frac{1}{4}$.

Here we have pictures to represent the 3 methods



[3]

So the conclusion of this paradox is that there is actually no unique selection method, so there can not be a unique solution. In fact, those methods proposed by Bertrand correspond to different selection methods and there is no reason to prefer one over another. Actually, it depends on how we look at the problem, because in our case, we don't have a lot of information on how should we find the answer. This is why there is no reason to prefer one over the other two, and hence no right or wrong answer. This shows us how important is, in mathematics, to well define our model, but also the question we are answering. Because, sometimes we make some assumptions just to simplify or because we don't know that much information about our model. That is why we have to remember the assumptions and take them into consideration when we interpret the results.

[3]Wikipedia - Bertrand's paradox (Probability)

References

Borel-Kolmogorov paradox:

1. Alexander Meehan and Snow Zhang, "The Borel-Kolmogorov Paradox Is Your Paradox Too: A Puzzle for Conditional Physical Probability", published online by Cambridge University Press
   https://bit.ly/3HARsvS

2. Wikipedia - Borel–Kolmogorov paradox
   https://bit.ly/3FQPTIT

3. Yonatan Oren, *Conditional Densities and the Borel-Kolmogorov Paradox*[Video], YouTube
   https://bit.ly/3FrUiAI

4. Zalán Gyenis, Gábor Hofer-Szabó, Miklós Rédei, "The Borel-Kolmogorov Paradox and conditional expectations"
   https://bit.ly/3VYKVj6

5. Z.Gyenis, G.Hofer-Szabó, M.Rédei, "Conditioning using conditional expectations: the Borel–Kolmogorov Paradox", March 26 2016.
   https://bit.ly/3Wgu7US

6. Mathijs Kolkhuis Tanke, "Paradoxical results from conditional probability: the importance of the σ-algebra", Sep. 25 2019
   https://bit.ly/3W8qccF

Bertrand's paradox:

1. Alexander Bogomolny, "Bertrand's Paradox"
   https://bit.ly/3HwAVJv

2. "Bertrand paradox (probability)"
   https://bit.ly/3Ypg7tu

3. Numberphile, *Bertrand's Paradox (with 3blue1brown) - Numberphile* [Video], YouTube
   https://bit.ly/3VZGdS2

4. Wikipedia - Bertrand paradox (probability)
   https://bit.ly/2Pxwd4U

5. Paul Keeler, "The Bertrand paradox", June 30 2019
   https://bit.ly/3PtSKus

Definition 4.19 Let $\overline{X} = (X_1, X_2)$ be a random vector with a continuous joint density. Let y be such that the marginal density of X_2 is positive: $f_{X_2}(y) > 0$. Then the conditional law of X_1, given $X_2 = y$ is defined to be continuous r.v with the following density

$$f_{X_1 \mid X_2=y}(x) := \frac{f_{X_1,X_2}(x,y)}{f_{X_2}(y)}$$

Proposition 4.14: If $\phi : \mathbb{R}^n \to \mathbb{R}^n$ is a diffeomorphism and $\overline{X}$ is a random vector with density $f_{\overline{X}}$, then $\phi \circ \overline{X}$ is also a random vector with a density given by :

$$f_{\phi \circ \overline{X}}(\overline{x}) = \frac{1}{|J_\phi(\phi^{-1}(\overline{x}))|} f_{\overline{X}}(\phi^{-1}(\overline{x}))$$

Corollary 4.16: Suppose X, Y are random vectors with density f_X, f_Y. Then $X + Y$ is a random vector with density

$$f_{X+Y}(z) = \int_{\mathbb{R}} f_X(x) f_Y(z - x) dx$$

Lemma 4.12: (Marginal densities). Let $\bar{X} = (X_1, ..., X_n)$ be a random vector with density $f_{\overline{X}}$ such that for every $I_0 \subseteq \{1, ..., n\}$ the function $f_{I_0^c}(x')$ obtained by fixing all the co-ordinates in I_0 is Riemmann-integrable. Then the marginal laws $\mathbb{P}_{I_0}$ obtained by projecting on the co-ordinates contained in I_0 admits a density given by integrating out all the components in $\{1, ..., n\} \backslash I_0$.

Analysis II: Let $-\infty \leq \alpha < \beta \leq \infty, E \in \mathbb{R}^n$ a non-empty open set, and $a, b \in C^1(E)$ such that

$$\mathrm{Im}(a), \mathrm{Im}(b) \subset (\alpha, \beta)$$

and $f : (\alpha, \beta) \times E \to \mathbb{R}$ a continuous function, with $\frac{\partial f}{\partial x_i} : (\alpha, \beta) \times E \to \mathbb{R}$ also continuous functions. Then $g \in C^1(E)$ and

$$\frac{\partial g}{\partial x_i}(x) = \frac{\partial b}{\partial x_i}(x) f(b(x), x) - \frac{\partial a}{\partial x_i}(x) f(a(x), x) + \int_{a(x)}^{b(x)} \frac{\partial f}{\partial x_i}(t, x) dt$$