mathematics

MDPI

Article

Bertrand's Paradox Resolution and Its Implications for the Bing–Fisher Problem

Richard A. Chechile

Psychology and Cognitive and Brain Science, Tufts University, Medford, MA 02155, USA; richard.chechile@tufts.edu

Abstract: Bertrand's paradox is a problem in geometric probability that has resisted resolution for more than one hundred years. Bertrand provided three seemingly reasonable solutions to his problem — hence the paradox. Bertrand's paradox has also been influential in philosophical debates about frequentist versus Bayesian approaches to statistical inference. In this paper, the paradox is resolved (1) by the clarification of the primary variate upon which the principle of maximum entropy is employed and (2) by imposing constraints, based on a mathematical analysis, on the random process for any subsequent nonlinear transformation to a secondary variable. These steps result in a unique solution to Bertrand's problem, and this solution differs from the classic answers that Bertrand proposed. It is shown that the solutions proposed by Bertrand and others reflected sampling processes that are not purely random. It is also shown that the same two steps result in the resolution of the Bing–Fisher problem, which has to do with the selection of a consistent prior for Bayesian inference. The resolution of Bertrand's paradox and the Bing–Fisher problem rebuts philosophical arguments against the Bayesian approach to statistical inference, which were based on those two ostensible problems.

Keywords: Bertrand's paradox; Bing–Fisher problem; philosophical theories of probability; non-informative Bayesian prior; Jeffreys prior

MSC: 28D20; 46N30; 49K45; 60G99; 62F15

Citation: Chechile, R.A. Bertrand's Paradox Resolution and Its Implications for the Bing–Fisher Problem. *Mathematics* **2023**, *11*, 3282. https://doi.org/10.3390/math11153282

Academic Editor: Velizar Pavlov

Received: 20 June 2023
Revised: 19 July 2023
Accepted: 24 July 2023
Published: 26 July 2023

1. Introduction

Probability theory has been fraught with some seemingly simple problems that have confounded even experts. Bertrand's paradox is one of these vexing problems [1]. The context for this problem is shown in Figure 1. Bertrand (1822–1900) asked the reader to consider an equilateral triangle that is inscribed in a circle of radius r.

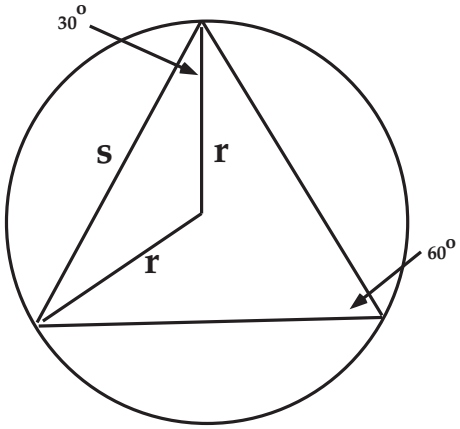

Figure 1. Figure underlying the Bertrand paradox problem.

Bertrand explored the probability of a randomly drawn chord having a length greater than the length of the side of the inscribed equilateral triangle. From the general cosine law applied to the smaller 30-30-120 triangle, shown in Figure 1, the results conclude that $s^2 = 2r^2 - 2r^2\cos(120^\circ) = 3r^2$. Without loss of generality, we can set $r = 1$, which results in ascertaining that the length of the equilateral triangle is $s = \sqrt{3}$. Bertrand's question thus reduces to finding $P(L > \sqrt{3})$ where L is the length of a *randomly* sampled chord for the circle of radius 1. Bertrand provided three different answers to their question.

Currently, Bertrand's paradox remains an illustration to many theorists about how the generation of random chords from a circle is an improperly posed problem since the three methods delineated by Bertrand result in different distributions for the length of the chords. For example, Mosteller in his famous book on fifty challenging problems in probability, produced a Bertrand-like problem [2]. His problem number 25 states,

> If a chord is selected at random on a fixed circle, what is the probability that its length exceeds the radius of the circle? ([2], p. 7).

Like Bertrand, Mosteller provided three analyses as possible solutions. He stated,

> Until the expression 'at random' is made more specific, the question does not have a definite answer [2], p. 7.

While Mosteller's point is clearly correct, the paradox has greater importance than simply an illustration of an under-specified problem. Bertrand's conflicting solutions were designed to illustrate his dissatisfaction with the Bayes and Laplace use of a probability distribution to represent an unknown parameter that can have any continuous value [3,4]. A key example of the Bayes/Laplace approach is the use of a uniform prior distribution for the unknown binomial rate parameter [5]. The Laplace justification for this prior distribution was based on the principle of insufficient reason (i.e., no reason to prefer any one value over any other value) [5]. Yet at the core of the reaction against the Bayes/Laplace analysis of an unknown population parameter is the concern that the parameter is a constant rather than a random variable. To circumvent the use of probability for parameters, Ellis introduced the relative frequency definition of probability in an effort to make probability an objective quantity [6]. Bertrand was also troubled about parameters having a probability distribution, but Bertrand's argument against the Bayes/Laplace approach was the demonstration of the ostensible paradox [3,4].

Many critics of the Bayesian approach to statistical inference have evoked Bertrand's paradox as a rationale. For example, von Mises, who was a leading and vigorous critic of the Bayes/Laplace use of the uniform distribution for the binomial rate parameter, stated that

> The attempts to justify, in various ways, the assumption of equally likely cases or, more generally, of a uniform distribution by having recourse to principles of symmetry or of indifference fails definitely in the treatment of the problems first considered by Bertrand, and later given the name of 'Bertrand's Paradox' by Poincaré [7], p. 77.

Fisher was also unhappy with the Bayes/Laplace approach to the inverse probability problem [8,9]. Fisher provided an argument against Laplace's analysis for the binomial rate parameter [10]. Fisher's reason for rejecting the Bayes/Laplace approach involved the issue of the change in the distributional shape of the prior when there is a reparametrization of variables. Fisher was not the first writer to make the point that the uniform prior for the binomial rate parameter was altered by a transformation of variables; this idea dates back to an earlier paper by Bing [11,12]. But Fisher was perhaps the most well-known statistician in the twentieth century, so his rejection of the Bayes/Laplace approach was very influential in stopping the development of Bayesian statistics for more than a decade [4]. Eventually, papers by Ramsey and by de Finetti [13,14] helped to reintroduce Bayesian theory, but by that time most researchers were exclusively working within the frequentist framework.

Today Bayesian inference is a widely used method in statistics, but there are some who still regard Bertrand's paradox as an unresolved problem that has philosophical implications for the method of statistical inference (e.g., [15]).

In this paper, a case is advanced that both Bertrand's paradox and the Bing–Fisher problem have a common set of mathematical issues that can be resolved. In both cases, there is a primary continuous variable that drives the problem, and there are nonlinear transformations that connect the primary variable to secondary variables. Achieving a maximum entropy distribution for the primary variable requires that the secondary variables be sampled with special nonuniform distributions that can be found from an examination of the Jacobian.

In terms of the organization of the paper, Bertrand's solutions to the problem are briefly reviewed and critiqued in Section 2. In Section 2, a case is made that stresses the importance of assessing stochastic processes for sampling random chords in terms of the entropy of the chord length distribution. It is also shown in Section 2 that all three of Bertrand's solutions do not represent a maximumly random process over the set of chord lengths. In Section 3, a maximum entropy stochastic process on chord length is identified. Moreover, a procedure is provided that establishes how to randomly sample other geometric properties, which are nonlinearly related to chord length, so as to be consistent with the maximum entropy distribution for chord length. In Section 4, the same method used to resolve the Bertrand paradox is shown to answer the Bing–Fisher problem. In Section 5, there is a discussion of the implications for resolving the Bertrand paradox and the Bing–Fisher problem. A case is made in the discussion that the resolution of Bertrand's paradox and the Bing–Fisher problem is important for refuting frequentist arguments against the Bayesian use of a probability distribution for representing an unknown population parameter.

2. Bertrand's Solutions Re-Examined

2.1. Reviewing Bertrand's Solutions

A random chord should have a random slope, and the chords need to be generated by a random process. Bertrand supplied three ways for generating a random chord. These solutions can be labeled the *radial method*, the *angular-separation method*, and the *within-disk method*. Figure 2 defines the key geometric features for describing each of the three Bertrand solutions.

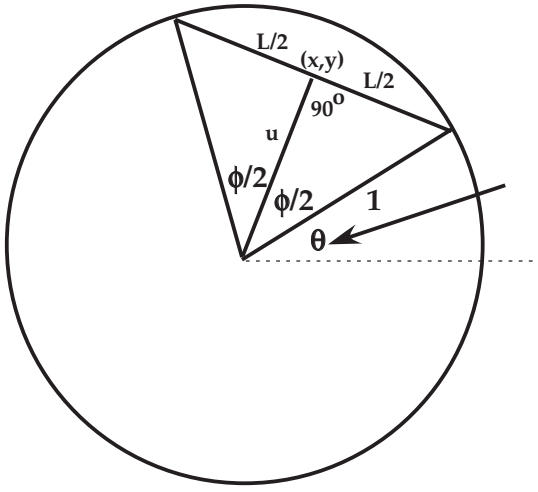

Figure 2. Figure for each of the three Bertrand solutions.

For the *radial method*, the first step is to generate a random angle $\theta + \frac{\phi}{2}$ for the slope of a line by sampling a value from the uniform distribution on the $[0, 2\pi]$ interval. The second step is to obtain a random length u for a line that connects the center of the circle with the midpoint of a chord. The value for u is drawn from a uniform distribution over the $[0, 1]$

interval. The chord is at a right angle to the radial line. Please note that $u^2 + \frac{L^2}{4} = 1$, so $u = \sqrt{1 - \frac{L^2}{4}}$. It follows that

$$P(L_a \leq L \leq L_b) = \int_{\sqrt{1-\frac{L_b^2}{4}}}^{\sqrt{1-\frac{L_a^2}{4}}} du = \sqrt{1 - \frac{L_a^2}{4}} - \sqrt{1 - \frac{L_b^2}{4}}\,. \quad (1)$$

$$P(L \leq L_b) = 1 - \sqrt{1 - \frac{L_b^2}{4}}\,. \quad (2)$$

Setting $L_b = \sqrt{3}$ in (2) results in $P(L \leq \sqrt{3}) = \frac{1}{2}$; thus $P(L > \sqrt{3}) = \frac{1}{2}$, which is the radial method answer to Bertrand's problem. Please note that the first step of this method satisfies the rotational symmetry property of a random chord, but it does not affect the length of the chord. It is the second step of randomly sampling the length of the radial segment to the chord midpoint that determines the chord length.

For the *angular-separation method*, two random angles are sampled. The first angle θ is drawn from the uniform distribution over the $[0, 2\pi]$ interval for a radial line between the circle center and the first point for the chord. The second random angle is ϕ, which is sampled from a uniform distribution over the $[0, \pi]$ interval. The radial line at $\theta + \phi$ establishes the second point for the chord. Please note that in Figure 2, the line of length u evenly divides the angle ϕ and thereby creates two congruent triangles. Thus, it follows that $\frac{L}{2} = \sin\frac{\phi}{2}$. Note if $\phi = 0$, then $L = 0$, and if $\phi = \pi$, then $L = 2$. For this reason, the random angle ϕ is sampled from the uniform distribution over the $[0, \pi]$ interval. It follows that

$$P(L \leq L_b) = \frac{2\,\sin^{-1}\frac{L_b}{2}}{\pi}\,. \quad (3)$$

By setting $L_b = \sqrt{3}$ into Equation (3) results in $P(L \leq \sqrt{3}) = \frac{2}{3}$. It thus follows that the answer to Bertrand's problem for the angular-separation method is $P(L > \sqrt{3}) = 1 - P(L \leq \sqrt{3}) = \frac{1}{3}$.

For the first two methods, there are two steps for generating a random chord, but for the *within-disk method* there is a single-step process of sampling over a two-dimensional surface. This method is an idealization of throwing a random dart at the disk with radius 1. All points within the disk have the same probability of being sampled. For any point (x, y) sampled within the disk, there is a radial distance u to the center of the circle where $u^2 = x^2 + y^2$. Any sampled point is considered the midpoint of the chord that has the slope of the tangent line to the circle of radius u at point (x, y). From the right triangle shown in Figure 2, it is clear that the link between the distance u and the length of the chord is $u^2 = 1 - \frac{L^2}{4}$. Moreover, all points (x, y) on the circle of radius u will also result in a chord of the same length; these chords will differ in their slope. All sampled points within the smaller disk of radius u result in chords of greater lengths. Thus, Bertrand argued that

$$P(L \geq L_b) = \frac{\pi\,u^2}{\pi\,1^2} = u^2\,. \quad (4)$$

$$= 1 - \frac{L_b^2}{4}\,. \quad (5)$$

By setting $L_b = \sqrt{3}$ in Equation (5), we find that $P(L \geq \sqrt{3}) = 1 - \frac{3}{4} = \frac{1}{4}$, which is the *within-disk* answer to Bertrand's problem.

2.2. Why All Three of Bertrand's Solutions Are Flawed

There is a need to clarify a semantic issue in the statement of Bertrand's problem. Namely, what does the phrase *a random chord of a circle* mean? One might argue that any

chord that is the result of a stochastic process is a random chord. Such a definition would yield a *random variable*, but this weak sense of the word *random* is not satisfactory, because there is an infinite number of stochastic processes that can be defined to yield a probability distribution of chord lengths. Besides the three processes that were described in the above subsection, consider the following construction shown in Figure 3 as the context for a novel chord-sampling method, which is called the *beta-sampling method*.

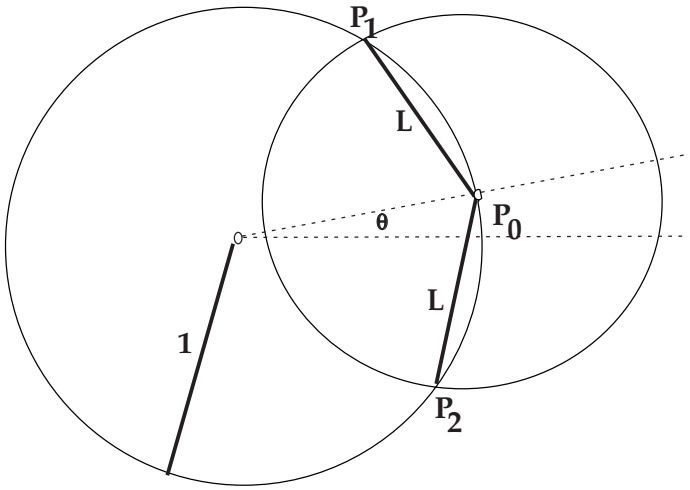

Figure 3. Figure underlying beta-sampling method.

There are four steps in the beta-sampling method. Step 1 is to sample a starting point on the circumference of the unit circle by generating a random angle θ on the $[0, 2\pi]$ interval. This point is labeled P_0, in Figure 3, and it is at one end of the chord. Step 2 is to sample a random L equal to 2 times a random value sampled from a beta distribution. The beta distribution has the probability density shown in Equation (6), and it is dependent on two shape parameters a and b that must be positive [16].

$$f(x) = \begin{cases} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} & 0 \leq x \leq 1, a > 0, b > 0, \\ 0 & \text{elsewhere.} \end{cases} \tag{6}$$

If $a = b = 1$, the beta distribution is the same as the uniform distribution on the $[0, 1]$ interval. When $a > b$, the beta is negatively skewed, and when $a < b$, the distribution is positively skewed. Since the sampled value from the beta distribution is a score on the $[0, 1]$ interval, it follows that L is a random length on the $[0, 2]$ interval. Step 3 is to construct a circle with radius L that has point P_0 as its center. Except for the two extreme cases of either $L = 0$ or $L = 2$, there are two points of intersection between the two circles. These are marked P_1 and P_2. Step 4 is to accept at random based on the flip of a fair coin either the $(P_0\ P_1)$ chord or the $(P_0\ P_2)$ chord. In the extreme case of $L = 0$ in Step 2, the second circle collapses to the point P_0. In the other extreme case when $L = 2$ in Step 2, the second circle only intersects the unit circle at the point that corresponds to the angle of $\theta = \theta_0 + \pi$ where θ_0 is the sampled angle in Step 1.

The beta-sampling method generates chords that satisfy symmetry in their slopes, and they are produced by a random process. Since the values for the a and b parameters can be any positive values, it is also clear that there are an unlimited number of distributions for chord lengths that are possible with the weak definition of the phrase *a random chord of a circle*. This example makes clear that a more restrictive definition of *randomness* is required. In fact Bertrand's three "solutions" used a uniform distribution over different geometric properties; i.e., the chord midpoint distance from the center for the *radial method*, the angle ϕ for the *angular-separation method*, and the points within the unit disk for the *within-disk method*. It seems clear from Bertrand's three proposed solutions that Bertrand intended to invoke the principle of insufficient reason to argue for a uniform distribution, *but he did*

not use a uniform distribution over chord lengths. Because there is a nonlinear transformation required from these other geometric properties to chord length, all three of Bertrand's solutions did not result in a uniform distribution over chord lengths.

The deviation from a uniform distribution also makes the distribution of chords less uncertain or more informative in a Shannon information sense. Shannon defined information in regard to the reduction of an entropy measure [17,18]. Given a discrete set of outcomes with probabilities $\{p_i\}$, for $i = 1, \cdots, N$, Shannon entropy H is

$$H = -\sum_{i=1}^{N} p_i \log_2 p_i = \frac{-1}{\ln 2} \sum_{i=1}^{N} p_i \ln p_i \,. \tag{7}$$

If there is certainty as exemplified by a distribution where one outcome has the probability of 1 and all other outcomes have a probability of 0, then $H = \frac{-\ln 1}{\ln 2} = 0$. When there is uncertainty, the entropy measure is positive. Shannon [17] also had a continuous form for entropy where for a finite interval I, entropy was defined as $-\int_I f(x) \log f(x) dx$ where $f(x)$ is the density function. For the continuous case, Shannon proved, using the calculus of variations, that for a finite interval the continuous uniform distribution had maximum entropy. However, for many problems entropy in the continuous form results in difficult integrals to compute. Consequently, in this paper, the discrete form of entropy is used. The discrete entropy for the maximum entropy (or most uncertain) distribution is one where each of N outcomes is equally probable, which results in an entropy value of $H_{max} = \frac{\ln N}{\ln 2}$. Theorem 1 is a formal statement of this well-known fact.

Theorem 1. *Given a set of N outcomes that are mutually exclusive and exhaustive with probabilities* $\{p_i\}$ *for* $i = 1, \cdots, N$ *for* $N \geq 2$ *that correspond to intervals of equal width for a continuous variable x on a finite support interval, the Shannon entropy measure from Equation (7) is at its maximum when each* $p_i = \frac{1}{N}$*, which results in an entropy value of* $H_{max} = \frac{\ln N}{\ln 2}$.

There are several proofs available; my version of a simple proof is provided in Appendix A for the reader's convenience. It is important in Theorem 1 to stipulate that the support for x is a finite interval such as the interval $[0, x_{max}]$ where x_{max} is finite. The variance of the uniform distribution is $\frac{1}{12}(x_{max} - x_{min})^2$, and it diverges for an unbounded range. Shannon showed that the condition of maximum entropy for an unbounded support interval, which nonetheless has a *finite variance*, implies that x has a Gaussian distribution [17]. However, in the current paper, the support for chord length is bounded; it is twice the radius of the circle. In general, any random variable on the finite support of $x_{max} - x_{min}$ can be segmented into arbitrarily small intervals of width $\frac{x_{max} - x_{min}}{N}$ with the choice of a sufficiently large integer N. When entropy is maximized, the probability for each of the intervals is equal.

The maximum entropy condition from Theorem 1 will occur when the continuous variable has a uniform distribution over the support interval. However, one can argue for a specific N that there can be other continuous distributions that also result in $p_i = \frac{1}{N}$ for $i = 1, \cdots, N$. For example, the periodic function $f(x) = \frac{\pi}{2} |\sin \pi N x|$ also results in $p_i = \frac{1}{N}$ for $i = 1, \cdots, N$. However, this function would not satisfy this condition for all integers $N \geq 2$, whereas the uniform distribution $f(x) = \frac{1}{x_{max} - x_{min}}$ would satisfy the condition that $p_i = \frac{1}{N}$, $i = 1, \cdots, N$ for *all* $N \geq 2$ *for dividing the support interval*. For N equal-width partitions, we can compute the standardized uncertainty of the distribution as the quantity

$$H_* \quad = \quad \frac{-\sum_{i=1}^{N} p_i \ln p_i}{\ln N} \,, \tag{8}$$

where H_* is the uncertainty of the distribution measured on a $[0, 1]$ scale. Using $N = 20$, the range of possible chord lengths is compartmentalized into intervals of size 0.1 (i.e., $[0, 0.1]$, $(0.1, 0.2], \cdots, (1.9, 2]$). The choice of N for this calculation was set to 20 only for obtaining an approximate estimate of H_* for the three methods that Bertrand proposed. The H_*

values for the radial, angular-separation, and within-disk methods are, respectively, 0.802, 0.940, and 0.936. These values show that none of the three proposed solutions by Bertrand are totally random because the chord generation procedures induce some restrictions on the distribution of chord lengths, making them somewhat more predictable. Moreover, each of the three Bertrand proposed solutions results in uneven wagers. Consider the proposition $H_0 : 0 < L \leq 1$ versus the proposition $H_1 : 1 < L \leq 2$. The uniform distribution on L is a fair bet where the odds are 1 : 1. However, the odds ratio for the radial, angular-separation, and within-disk methods are, respectively, 1:6.4641, 1:2, and 1:3. These long odds against H_0 strongly illustrate how lopsided these three methods are for generating chords. Thus, all three solutions proposed by Bertrand are *informative* sampling schemes in the sense that they result in a systematic deviation from even odds between H_0 and H_1. Please note that the term *informative* is used here because the H_0:H_1 odds are not even and because an effort is made to avoid using the word *biased* since the phrase *statistical bias* has a longstanding technical meaning in statistics (e.g., [19]).

Although the major point of this section concerns the degree of randomness of the three Bertrand methods, the within-disk method has an additional problem that disqualifies it as a valid method for random chord generation. For the radial and the angular-separation methods, there is an initial step of generating a random angle, which establishes the particular radial line segment for the radial method or establishes the initial point on the circumference of the circle for the angular-separation method. The second step for those two methods is either sampling a random midpoint for the radial method or sampling a random angle ϕ for the angular-separation method. The result of those two random processes results in a *single chord*. The probability for the length of the chord is computed over the possibilities based on the second random process. The value for the first step has no bearing on the length of the chord. In contrast to those methods, the within-disk method computes effectively those two random processes in a single step. Unlike the other two methods, this procedure computes the probability of the chord length over all possible radial segments rather than the single radial segment where the dart landed. But, *all the other possible radial segments are irrelevant after the dart lands on one particular radial segment*. Just like the specific value for the angle in the first step is irrelevant for the probability of the particular chord length for the radial method and the angular-separation method, the other possible radial segments are also irrelevant for the within-disk method. Hence the step of computing the area within the inner circle with a radius of u is an error. The area of the inner circle is the union of all possible radial segments (all the points on the circumference of the inner circle) plus the union of all possible longer chords in any orientation. What should be computed instead is the linear length of the radial segment sampled. In essence, the within-disk method is a disguised version of the radial method. Thus, the within-disk procedure must be rejected outright as a valid stochastic process. Shackel [15] also argued that the within-disk method should be rejected as a legitimate stochastic process for generating a single chord. Yet there are still two different Bertrand procedures that generate chords, but both of these methods are not sampling procedures that are purely random.

3. Resolving Bertrand's Paradox

3.1. Resolving Bertrand's Paradox via the Maximum Entropy Distribution for Chord Lengths

In light of the critique of Bertrand's solutions, there is a clear way to randomly sample chords that is purely random. The beta-sampling method where $a = b = 1$ results in a *uniform distribution of chord lengths over the* $[0, 2]$ *interval*. This distribution has maximum entropy for chord lengths. The cumulative distribution is $P(L \leq L_b) = \frac{L_b}{2}$, and the probability density function is $f(L) = \frac{1}{2}$ for $L \in [0, 2]$. Thus, the probability that the chord has a length less than the length of an inscribed equilateral triangle is $P(L \leq \sqrt{3}) = \frac{\sqrt{3}}{2}$. Consequently the unique answer to Bertrand's problem for the probability that a random chord has a length that is greater than the length of a side of the equilateral triangle is $P(L > \sqrt{3}) = 1 - \frac{\sqrt{3}}{2} \approx 0.1339746$. In general, for a randomly sampled chord that has maximum entropy

$$P(L_a \le L \le L_b) = \int_{L_a}^{L_b} \frac{1}{2}\, dL = \frac{1}{2}(L_b - L_a)\,. \tag{9}$$

If chords are randomly sampled, such that the lengths have a uniform distribution as shown above, then how should the radial distance to the chord midpoints u and the angular separation ϕ be sampled? In general, the rule for transforming an integration from an x variate to a v variate is

$$A = \int_a^b f(x)\, dx \quad = \quad \int_{h(a)}^{h(b)} f(g(v))g'(v)dv\,, \tag{10}$$

where $x = g(v)$, $v = h(x)$, and where $g'(v)$ is the Jacobian of the transformation from the differential of dx to the differential dv. Thus, for the transformation from the dL differential to the du differential, we have $L = g(u) = 2\sqrt{1-u^2}$, $u = h(L) = \sqrt{1-\frac{L^2}{4}}$, and $g'(u) = \frac{-2u}{\sqrt{1-u^2}}$. To re-express Equation (9) in terms of the metric of the radial distance to the chord midpoint yields $P(L_a \le L \le L_b)$ as

$$\int_{L_a}^{L_b} \frac{1}{2}\, dL \quad = \quad \int_{\sqrt{1-\frac{L_a^2}{4}}}^{\sqrt{1-\frac{L_b^2}{4}}} \frac{1}{2}\left(\frac{-2u}{\sqrt{1-u^2}}\right) du\,. \tag{11}$$

$$\frac{1}{2}(L_b - L_a) \quad = \quad \int_{\sqrt{1-\frac{L_b^2}{4}}}^{\sqrt{1-\frac{L_a^2}{4}}} \left(\frac{u}{\sqrt{1-u^2}}\right) du\,. \tag{12}$$

Please note that the integrand of the right-hand-side of Equation (12) is the u-space probability density function that is consistent with the uniform distribution over chord lengths. With a probability density function $f(u) = \frac{u}{\sqrt{1-u^2}}$, the radial method generates chords such that $P(L > \sqrt{3}) = 1 - \frac{\sqrt{3}}{2}$. Thus, if the appropriate density function is used for sampling u values, then the same answer to Bertrand's problem is found as for the beta-sampling method where $a = b = 1$, which results in a uniform distribution of chord lengths over the [0, 2] interval.

Let us re-examine the angular-separation method so that it is consistent with the random selection of chords where the density function is $f(L) = \frac{1}{2}$ for $L \in [0, 2]$. The Jacobian to transform the differential from dL to $d\phi$ is $g'(\phi) = \cos\frac{\phi}{2}$ with $L = g(\phi) = 2\sin\frac{\phi}{2}$ and $\phi = h(L) = 2\sin^{-1}\frac{L}{2}$. Thus, it follows that $P(L_a \le L \le L_b)$ is

$$\int_{L_a}^{L_b} \frac{1}{2}\, dL \quad = \quad \int_{2\sin^{-1}\frac{L_a}{2}}^{2\sin^{-1}\frac{L_b}{2}} \left(\frac{1}{2}\right) \cos\frac{\phi}{2}\, d\phi\,. \tag{13}$$

Simplifying Equation (13) by the linear transformation from $d\phi$ to $d\alpha$ where $\alpha = \frac{\phi}{2}$ yields

$$P(L_a \le L \le L_b) \quad = \quad \int_{\sin^{-1}\frac{L_a}{2}}^{\sin^{-1}\frac{L_b}{2}} \cos\alpha\, d\alpha\,. \tag{14}$$

$$\quad = \quad \sin\sin^{-1}\frac{L_b}{2} - \sin\sin^{-1}\frac{L_a}{2}\,. \tag{15}$$

$$\quad = \quad \frac{1}{2}(L_b - L_a)\,. \tag{16}$$

From Equation (14) it is clear that the probability density in α space is $\cos\alpha$. With this density function, the angular-separation method yields the answer to Bertrand's problem as $1 - \frac{\sqrt{3}}{2}$.

3.2. The Importance of a Dominant Metric Representation

Nonlinear transformations of metrics are at the core of Bertrand's paradox. In general, in Equation (10) there is a Jacobian for the transformation when the integration over one variate is transformed nonlinearly to an integration over a different variate. It is only for a *linear transformation* such as $x = cy = g(y)$, which has a Jacobian of $g'(y) = c$, that the integrand keeps the same functional shape. Thus, a probability density function for one variate will be nonlinearly changed when there is a nonlinear transformation. We have seen that there are three important variates in the analyses in Section 3.1—the chord length L, the radial distance to chord midpoint u, and the angular separation between the endpoints of the chord ϕ. In one of Bertrand's analyses, he used a uniform distribution over u to obtain the distribution over chord lengths L that is not uniform. In fact, there is a strong likelihood of longer chords. Alternatively, he used a uniform distribution over ϕ to obtain a different informative distribution for chord lengths, which also had a preference for longer chords. Had Bertrand used a uniform distribution over u, and examined the resulting distribution for the angular separation ϕ, then the density function for ϕ would be proportional to $\sin\frac{\phi}{2}$, which is not uniform. Consequently, it is not possible to have a uniform distribution on any of the three variates and to also have a uniform distribution for any of the other two variates. Note this mathematical fact is due to *properties of integration, so it is more general than a special problem with probability*. This mathematical fact is also well-known in statistics.

Given the nonlinear relationship between any pair of the three variates, the question arises: why choose any one of the variates as the one for a maximum entropy representation? It is argued here that the problem statement dictates the dominant framework. Bertrand's problem is about the length of the chords. The problem deals with the comparison between a *random chord length* and the *length of the side of an equilateral triangle*. The problem is not about *a random angular separation of a chord* in comparison to the *angle of an equilateral triangle*. If the problem were to be changed to be one of comparing random angles, then there would be a different probability answer. Thus, the chord length is the appropriate basis for sampling a random chord due to the problem statement. The problem statement dictates a preferential or dominant variable to use for the maximum entropy distribution. Moreover, the mathematical measure of a chord is its length, so it is not surprising that Bertrand framed the question in terms of the *length* of a random chord and the *length* of a side of the inscribed triangle.

3.3. Some Simulation Examples

Further insights about Bertrand's problem can be obtained from some simulations. The first simulation illustrates the density displays for chord length for three methods. The stochastic processes are (1) uniform sampling of the radial length to the chord midpoint, (2) uniform sampling of the angular separation between the chord endpoints, and (3) uniform sampling of the chord length itself. The code for an R function that implements each of these three methods is provided in Appendix B.1. For each sampling method, 100,000 samples for chord length were obtained. See Figure 4 for the resulting probability density estimates on chord length for these three methods. These plots clearly show that the new method advanced in this paper results in a flat distribution over chord lengths whereas the two Bertrand methods do not, as was pointed out earlier in Section 2.2.

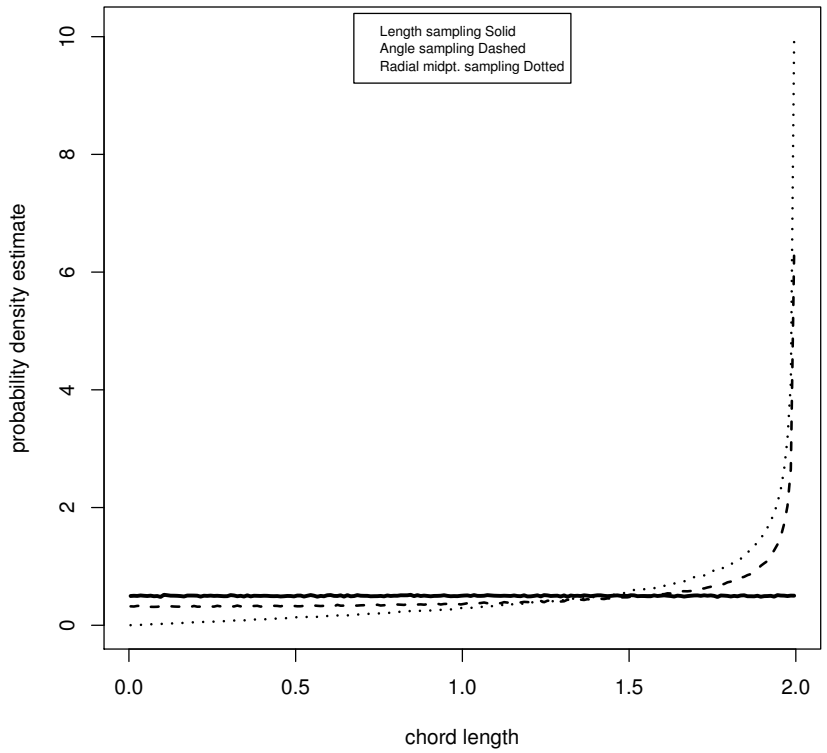

Figure 4. Figure shows the probability density estimates for the three chord-sampling methods. The probability density estimates were obtained from 100,000 random samples for each method. The solid line has maximum entropy for chord length, the dashed line has maximum entropy for angular separation, and the dotted line has maximum entropy for radial distance to chord midpoint.

The next set of simulations is designed to illustrate the outcome of drawing random chords for the new method. The initial point (x_0, y_0) for each chord is sampled from a uniform distribution on $[0, 2\pi]$. The corresponding endpoint for each chord is one of the two intersection points between the original circle and the circle that has a center at (x_0, y_0) and has a radius of L. The value of L is randomly sampled from a uniform distribution over $[0, 2r]$. In general, $x_0 = r\cos\theta$ and $y_0 = r\sin\theta$ where r is the radius of the original circle, and θ is a randomly sampled angle for the initial point for the chord. The other endpoint for the chord of length L is the point (x_1, y_1) where $x_1 = r\cos\alpha$, and $y_1 = r\sin\alpha$. Since $(x_1 - x_0)^2 + (y_1 - y_0)^2 = L^2$, it follows after some algebra that

$$\begin{aligned} x_0x_1 + y_0y_1 &= r^2 - \frac{L^2}{2}, \\ \cos\theta\cos\alpha + \sin\theta\sin\alpha &= 1 - \frac{L^2}{2r^2}, \\ \cos\theta\cos\alpha + \sin\theta\sin\alpha &= \cos(\theta - \alpha) = \cos(\alpha - \theta), \\ \therefore\ \alpha &= \theta \pm \cos^{-1}\left(1 - \frac{L^2}{2r^2}\right). \end{aligned}$$

The two possible values for the angle α correspond to the two possible intersection points between the two circles as illustrated in Figure 3. In the simulation, one of these two values is independently selected by a virtual flip of a fair coin. The software for simulating n random chords with this procedure is provided in Appendix B.2.

Figure 5 provides the results from six simulations where the circle has radius $r = 1$, and the chord length L is uniformly distributed on $[0, 2]$. The number of chords for the six panels are: $n = 20 \cdot 2^i$ where $i = 1, 2, \cdots, 6$. The plots show that the region near the origin of the disk has relatively fewer chord segments compared to the outer disk region.

There are two reasons for this inequality in the density for filling the disk with random chords. First, chords begin and end on the circle boundary, so the outer region must on average have a higher density than the inner region. Second, the inner region can only be traversed if the chord length is large. The method for uniform sampling of length, unlike the other two alternative sampling methods shown in Figure 4, does not overly sample for longer chords. Yet the issue raised by Bertrand's problem is not about the density of chord segments in various regions of the disk, so this salient visual pattern is not relevant.

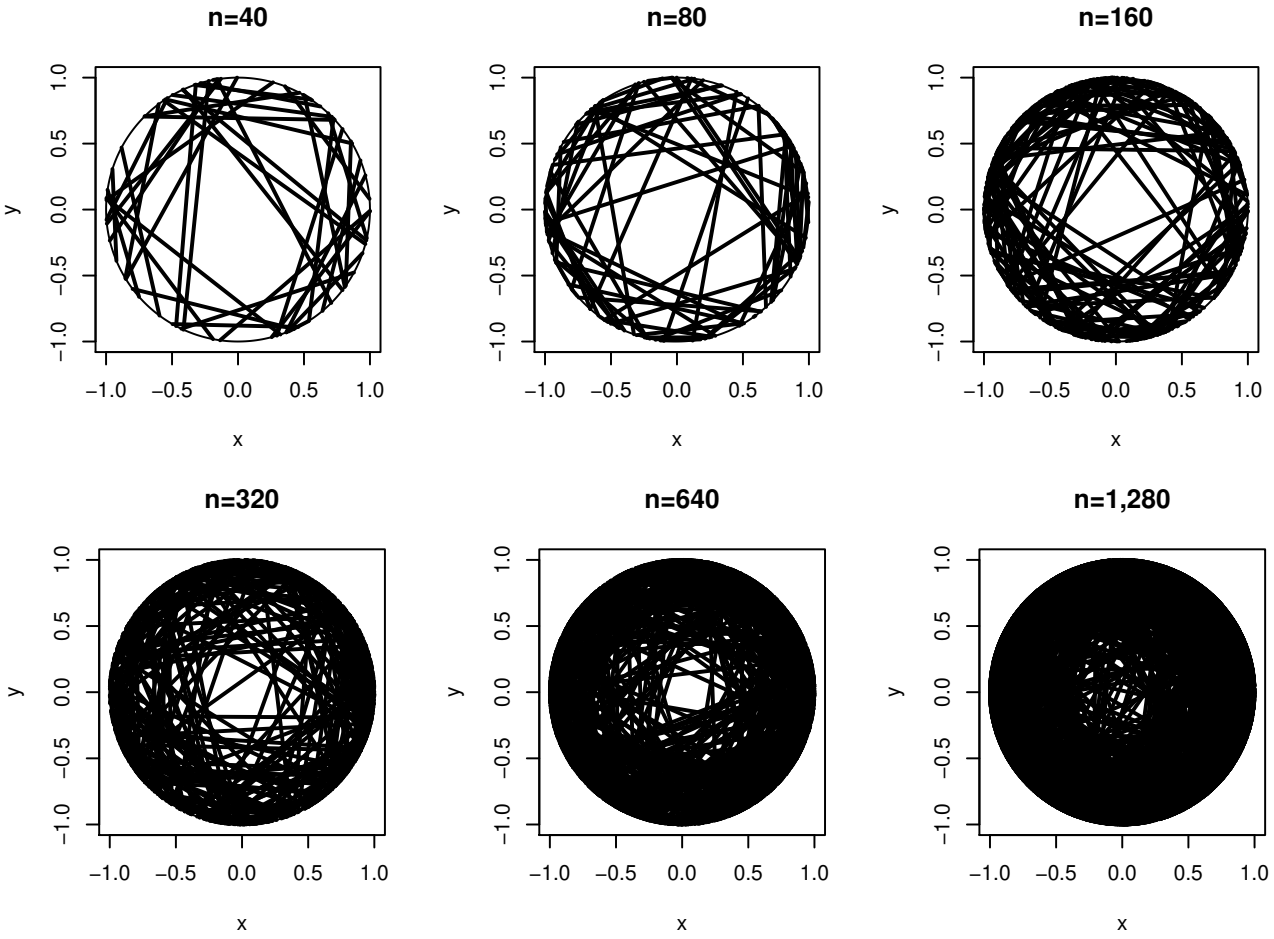

Figure 5. Figure shows the results for six sets of random chords where the chord length is drawn from a uniform distribution on the $[0, 2]$ interval and where the initial point for the chord sampled from a uniform distribution on the $[0, 2\pi]$ interval. The number of chords per panel n are 40, 80, 160, 320, 640, and 1280.

Bertrand's problem has also been simulated with Poisson-line stochastic processes (e.g., [20]). For such applications, there is a Poisson random process that generates the number of lines. If the number of lines from the Poisson process yields the value of n, then n lines are generated via any particular method for randomly sampling chords. In the software cited above, the radial method is used for producing random chords. However, it should be noted that Bertrand's problem is not about the number of chords. The number of chords can be simply one. Bertrand's question remains, regardless of the number of attempts to draw random chords. It is known for the simulations used for Figure 5 that all the chords were sampled via a stochastic process where the probability that the length of a sampled chord exceeding $\sqrt{3}$ is equal to $1 - \frac{\sqrt{3}}{2}$. We do not need to generate a large sample of chords to ascertain that result. However, if one were to ask a different question about the random chords, then Monte Carlo simulations might be needed. For example, suppose we were interested in the mean and variance of the largest chord length for a set of n random chords

produced by a particular chord generation method. This question is a function of n, and the rate of convergence to an asymptotic distribution is not generally known. This type of question has attracted some interest (e.g., [21]), but it is a very different problem from the one raised by Bertrand.

In light of the discussion of the number of chords produced by the uniform length-sampling method, we can pose a different Bertrand-type question. Namely, what is the fewest number of random chords sampled on a unit circle such that the probability is 0.95 or greater that at least one chord is longer than $\sqrt{3}$? Given that each of the n chords has a probability of $\frac{\sqrt{3}}{2}$ for having a length less than $\sqrt{3}$, it then follows that the probability that at least one chord exceeds $\sqrt{3}$ is $1 - (\frac{\sqrt{3}}{2})^n$. The resulting answer to the question is $n = 21$ because when $n = 21$ the probability is 0.9512 that at least one chord is longer than $\sqrt{3}$; whereas when $n = 20$ the probability is 0.9437. To confirm this answer, a Monte Carlo study of two million trials was examined with $n = 21$ and another two million trials with $n = 20$. The proportion of the Monte Carlo samples where the largest chord in the set exceeded $\sqrt{3}$ was 0.9513 for $n = 21$, but it was 0.9439 for $n = 20$.

3.4. Bertrand's Problem in a Historical Context

As noted previously, Bertrand had an agenda for his problem. He was a critic of the use of a probability distribution for representing an unknown parameter in a Bayesian analysis when there is an uncountably infinite number of possible outcome states [3,4]. Bertrand did not argue directly against the Bayesian approach when there were a finite number of states such as computing the probability as to which of two possible bags of colored marbles was chosen *at random*. In this case, the assigning of a prior probability of $\frac{1}{2}$ to each bag based on the principle of insufficient reason seemed rational. But he felt that the generalization of that principle was problematic for cases, such as the estimation for a biased coin, where there is an uncountable infinite number of possible values for the unknown rate parameter [3]. As noted by Jaynes [22],

> Since Bertrand proposed it in 1889, this problem has been cited to generations of students to demonstrate that Laplace's "principle of indifference" contains logical inconsistencies (p. 478).

In the time since Bertrand's analysis, it is surprising that the simple answer to Bertrand's problem, which is based on the maximum entropy principle as applied to the mathematical measure of chords, has somehow eluded discovery heretofore. Many scholars instead debated the relative merits of the three solutions provided by Bertrand (e.g., [2,7,15,22–27]). Opinions about Bertrand's three solutions varied among these theorists. Von Mises [7] agreed with Bertrand, and used the paradox as an argument against the Bayes/Laplace use of probability for a population parameter. Other writers saw the problem as being ambiguous about the stochastic process of selecting chords [2,25]. Once a method is selected, then there was a probability that the chord was larger than the side of the inscribed triangle. In essence, these writers agreed with Bertrand without drawing conclusions about the principle of indifference or maximum entropy. Gyenis and Rédei [27] did not challenge any of Bertrand's solutions, but they questioned whether Bertrand's paradox met their standard for a philosophical paradox. As they stated in a philosophical journal,

> The interpretation proposed here should make clear that Bertrand's Paradox cannot be "resolved" — not because it is an unresolvable, genuine paradox but because there is nothing to be resolved: the "paradox" simply states a provable, non-trivial mathematical fact, a fact which is perfectly in line both with the correct intuition about how probability theory should be used to model phenomena and how probability theory is in fact applied in the sciences. (p. 350).

Regardless of the definition of a paradox, the problem that Bertrand identified calls for clarification. What is a random chord, and what is the probability that a random chord

of the unit circle has a length greater than $\sqrt{3}$? These are fair questions, and it would be troubling if these questions did not yield a single answer.

Other writers have argued based on symmetry and invariance principles that Bertrand's radial method is the correct solution [22–24,26,28]. For example, Jaynes treated the Bertrand problem in the context of another problem in geometric probability in which the plane is superimposed with a random set of lines. If a line does not intersect with the circle, then it does not count. But if a line does intersect with the circle, then the chord is the distance between the two intersection points. Jaynes argued that the distribution of chord lengths should be invariant with the translation of the circle in the plane, i.e., if the circle is moved in the background field of lines, then the answer to Bertrand's problem should not change. However, philosopher Louis Marinoff correctly pointed out that by imposing the translational invariance requirement is changing the original Bertrand problem [29]. Translational symmetry does not mean that the distribution of chords meets the randomness requirement. Translational invariance just means that the preference for producing long chords is stable in regard to the movement of the circle in a field of lines.

Jaynes, who was a theoretical physicist, also argued that the radial method is correct because of an empirical experiment of tossing straws [22]. Jaynes argued if the straw missed the circle, then it was disregarded, but if it overlapped the circle, then that determines a chord. He claimed the distribution of chord lengths was consistent with the radial method distribution function. Based on tossing 128 straws, Jaynes reported that a chi-squared statistical test of the null hypothesis was consistent with the distribution shown in Equation (2). This experiment is not convincing for many reasons. First, the goodness-of-fit test assumed the radial method distribution as the null hypothesis, and any frequentist statistical test cannot *prove* the hypothesis that was assumed in the first place. Second, the analysis by Porto and associates [30] demonstrated that the length of straws relative to the radius of the circle dramatically influences the answer to Bertrand's question. Furthermore, other physicists [31] argued that there can be many different stochastic processes for tossing straws that have a different answer to Bertrand's problem. These physicists further suggest that the solution to Bertrand's problem is to compute the arithmetic mean of the distinctly different answers to Bertrand's problem. However, it is more reasonable to reject the whole idea of trying to answer Bertrand's problem with any experiment where the outcomes can be widely variable. Bertrand's problem is about a mathematical operation of constructing chords. It is not about an actual stochastic process that is occurring in nature. Real stochastic processes do occur in nature, but these processes do not necessarily reflect pure randomness.

Holbrook and Kim [32] described a physics thought experiment that does not require actual data to obtain a probability. Their thought experiment consisted of arranging a circular cloud-chamber detector perpendicular to the direction of cosmic rays. The chord is the path of the ray through the detector. For an ideal detector of radius 1, the chord path is greater than $\sqrt{3}$ when it is within $\pm\frac{1}{2}$ of the center. This thought experiment is implementing the same operations considered when the radial method was discussed previously, and it arrived at the same answer of $\frac{1}{2}$ as was found for the radial method. Consequently, the feature of cosmic rays and the cloud chamber are not needed to arrive at the probability of Bertrand's question. Thus, the Holbrook and Kim thought experiment is not convincing. Moreover, Ardakani and Wulff [33] described a different novel method for generating chords, and this method resulted in the answer of $\frac{1}{3}$ to Bertrand's problem—a value that is consistent with Bertrand's angular-separation method. Consequently, neither the Holbrook-Kim method nor the Ardakani-Wulff method resolves Bertrand's paradox. Bertrand already showed that different stochastic processes result in different answers to their problem.

Kaushik [34] developed a stochastic process that resulted in the conclusion that $P(L > \sqrt{3}) = \frac{1}{4}$. Recall that $\frac{1}{4}$ was the answer for Bertrand's within-disk method, which was previously rejected for reasons discussed in Section 2.2, and it was also rejected as a valid stochastic process by Shackel [15]. However, the arguments raised against the within-

disk method do not apply to the process proposed by Kaushik, so this stochastic procedure should be examined more carefully. With the Kaushik procedure, each point along the diameter of the unit circle is uniformly sampled. For simplicity, let the sampled point be a distance t from the point $(1, 0)$ along the path to $(-1, 0)$. At the point $(1 - t, 0)$, a perpendicular line is drawn to the circle. The chord endpoints are $(1, 0)$ and $(1 - t, \sqrt{2t - t^2})$. Kaushik points out that for each value for t, there is a corresponding chord of length $\sqrt{2t}$. Conversely, for each possible chord of length L, there is a corresponding distance t. But the problem with the Kaushik solution is that t is sampled from a uniform distribution on the $[0, 2]$ interval, which results in an *informative distribution for chord length*. This fact is demonstrated by examining the odds ratio between the hypotheses of $L \leq 1$ and $1 < L \leq 2$, which is 1 to 3 instead of 1 to 1. To obtain the flat chord length density function $f(L) = \frac{1}{2}$ for $L \in [0, 2]$ and given the general transformation formula from Equation (10), we should set $L = g(t) = \sqrt{2t}$ and $t = h(L) = \frac{L^2}{2}$. Thus, the Jacobian is $g'(t) = \frac{1}{\sqrt{2t}}$, and it follows from Equation (10) that

$$P(L \leq L_b) = \int_0^{\frac{L_b^2}{2}} \frac{1}{2\sqrt{2t}} dt \,, \tag{17}$$

$$= \frac{L_b}{2} \,. \tag{18}$$

Thus, $P(L \leq \sqrt{3}) = \frac{\sqrt{3}}{2}$, so $P(L > \sqrt{3}) = 1 - \frac{\sqrt{3}}{2}$, which is the correct answer advanced in this paper to Bertrand's problem. From Equation (17) the effective density function for t should be $\frac{1}{2\sqrt{2t}}$. However, if $f(t) = \frac{1}{2}$, then $P(L \leq \sqrt{3}) = \frac{3}{4}$, which results in the incorrect answer of $\frac{1}{4}$ to Bertrand's problem. The Kaushik analysis, like the other Bertrand methods, fails to employ the proper density function that is consistent with the uniform distribution for chord length. Instead, Kaushik sampled from a uniform distribution on a secondary geometric feature that is nonlinearly linked to chord length.

While most papers dealt with the classic solutions discussed by Bertrand, several investigators proposed alternative stochastic processes for generating chords that result in different answers to Bertrand's problem [29,35,36]. Chiu and Larson [35] discussed five alternative answers to Bertrand's problem, but none of these alternatives had a density function that was uniform for chord length. Jevremovic and Obradovic [36] used Monte Carlo simulations to evaluate three alternative methods to ascertain the probability that random chords exceeded $\sqrt{3}$. However, none of these three methods came remotely close to the value of $1 - \frac{\sqrt{3}}{2} \approx 0.13397$, which occurs when the density function on chord length is uniform. The smallest value found by these investigators from their Monte Carlo simulations was about 0.61, so these procedures also had a strong preference towards generating long chords. These papers illustrate that there are many more than the original three stochastic processes examined by Bertrand for generating an informative distribution of chords. In each case, the researchers imposed a uniform distribution on a geometric variable that is nonlinearly linked to chord length.

Marinoff [29] also entertained several solutions to Bertrand's problem. One of the solutions had the answer of $1 - \frac{\sqrt{3}}{2}$, which is the same value arrived at in this paper. However, Marinoff did not argue that this random process had any special status. Marinoff also did not discuss the uniqueness of this solution in terms of the maximum entropy for chord length nor did he discuss the role of the Jacobian as the underlying reason for why Bertrand's paradox occurs. Moreover, there has not been a subsequent discussion of this stochastic process in the literature since it was published in 1994.

In light of this review of the literature on Bertrand's paradox, it is interesting to observe that this problem is deceptive, and it has resisted resolution for a long time. What appears to mislead most theorists is the importance of the Jacobian for Bertrand's problem. As noted above with the Kaushik stochastic process, there is clearly a one-to-one linkage between any possible chord of length L and the corresponding distance t that produced that length.

Both t and L span the interval of $[0, 2]$. Yet despite the one-to-one correspondence between the variates, which both have an uncountably infinite number of possible values, it is still necessary to account for the inequality of the dL and dt differentials. The error was using a maximum entropy, uniform distribution for the t variable rather than for the chord length L. It is fine to generate chords with the Kaushik stochastic process provided that the density function for sampling t is linked via the Jacobian to the uniform distribution of chord length as shown in Equation (17). There is one correct answer to Bertrand's problem, but it was apparently not clear to either Bertrand or others why a purely random chord generation process must generate chords that have a uniform distribution over the possible lengths for the chord. Yet Bertrand was correct to stress the important difference between sample spaces that have a finite number of elementary outcomes and sample spaces that have an infinite number of elementary outcomes. The differential and the Jacobian are only used for continuous variables. For example, suppose the sample space for a random variable n consists of the integers $\{1, \cdots, 100\}$, and we are interested in a nonlinear transformation such as $m = n^2$. The n and m sample spaces are finite, and probabilistic statements do not involve calculus and the Jacobian. A uniform distribution for n is the discrete uniform over the integers $\{1, \cdots, 100\}$, and the corresponding distribution for m is the discrete uniform distribution over the 100 squares $\{1, 4, \cdots, 10{,}000\}$. The probability of a value less than k in the sample space for n is equal to the probability of a value less than k^2 in the sample space for m.

4. The Bing–Fisher Problem

The inverse probability inference method introduced by Bayes [37] and later by Laplace [5] troubled many nineteenth and early twentieth-century theorists. It seemed clear that the unknown parameter for the binomial was a constant, so it could not be represented by a probability distribution. By the mid-twentieth century, after the development of information theory [17] and with the development of the Kolmogorov axioms of probability [38], a more flexible use of probability was more acceptable for some theorists (e.g., [13,14,39]). But before those advances, there was an incentive to make probability an objective construction that did not depend on human judgment or knowledge. The relative frequency theory was the dominant alternative to the Bayes/Laplace use of a probability distribution to represent unknowns [6]. Yet the arguments against the Bayes/Laplace approach was not a debate about Bayes' theorem itself, but rather it was about the use of a probability distribution for parameters when making a statistical inference. Bertrand's paradox was one such justification for rejecting the Bayes/Laplace approach, despite the fact that Bertrand's problem was not about an unknown population parameter [7]. Bing is regarded as the first to criticize the Bayes/Laplace approach because the distribution for the unknown population parameter is altered by a re-parametrization of the stochastic process [11,12]. After Fisher raised this same point, the use of Bayesian statistics vanished for more than a decade [4,10].

The same issues raised in this paper with the re-analysis of Bertrand's problem also apply to rebut the Bing–Fisher argument. Just like the Bertrand reanalysis, there is a preferred direct representation for the binomial model, and there are alternative representations that involve a nonlinear transformation. For Bertrand's problem, the chord length was the preferred representation because length is the mathematical measure of chords. The preferred variable for the binomial is the population proportion for one of the two categories, which we can denote as ϕ. The binomial likelihood function given N trials with n_1 number of successes and $n_2 = N - n_1$ number of failures is

$$P(n_1, n_2|\phi) = \frac{N!}{n_1!\, n_2} \phi^{n_1} (1 - \phi)^{n_2}\,. \tag{19}$$

Later we will consider two different nonlinear formulations of the binomial model, but these alternatives are indirect representations because the probabilities for the transformed variates are still driven by the value for ϕ. In terms of the preferred binomial model, it is

well known that a prior distribution from the beta functional family shown in Equation (6) combines with the binomial likelihood from (19) to result via Bayes' theorem in a posterior distribution that is another member of the beta functional family [5,37]. In terms of the binomial ϕ parameter, the uniform distribution is a beta distribution where the shape parameters, which can be denoted as a' and b', are equal to 1. The posterior distribution is the beta distribution where the two shape parameters are $a = a' + n_1$ and $b = b' + n_2$.

The ϕ representation in Equation (19) is the preferred system in a similar sense as the chord length was the direct representation for Bertrand's problem. For this system, we can have a distribution that maximizes Shannon information (i.e., the uniform prior for ϕ on the $[0, 1]$ interval). This prior distribution has maximum entropy (i.e., H_* from (8) is 1). A nonlinear transformation of the parameter will alter the distribution. To see the effect of transformations, let us consider two alternative representations of the binomial model. First let us consider the transformation $\phi = \sin^2 \theta$ with $\theta \in [0, \frac{\pi}{2}]$. The θ parameter cannot have a uniform distribution and still be consistent with the ϕ parameter being uniform. The Jacobian to change the integration variable from $d\phi$ to $d\theta$ is $2 \sin \theta \cos \theta$. As with the analysis of Bertrand's problem, it is recommended that we use as the prior for θ the above Jacobian of $2 \sin \theta \cos \theta$. Such a prior is not a low information prior, but it is linked via the variable transformation back to the preferred parametrization of ϕ that does have maximum entropy. So, an investigator who is using the θ parametrization, should adopt a prior of $2 \sin \theta \cos \theta$ to preserve $H_* = 1$. If the researcher were to instead use a uniform prior for θ on $[0, \frac{\pi}{2}]$, then H_* would have the suboptimal value of about 0.953.

For a second example, let us consider the parametrization of $\phi = \frac{\psi}{1+\psi}$ where ψ is a non-negative real number. The ψ parameter is the odds ratio $\frac{\phi}{1-\phi}$. Since ψ is on an infinite support interval, ψ cannot have a uniform distribution. The Jacobian for the $d\phi$ to the $d\psi$ integration is $\frac{1}{(1+\psi)^2}$. Consequently, a researcher, who chooses to use the ψ-space representation, should use the prior of $\frac{1}{(1+\psi)^2}$ for ψ to be consistent with the preferred maximum entropy ϕ parametrization.

Thus, one creditable response to the Bing–Fisher challenge to the Bayesian program of statistical inference is to not transform the direct stochastic representation and then impose a flat prior for the alternative variate. *If one chooses to transform the variate, then employ a prior in that system that can remain consistent with the direct representation*.

Historically a different rebuttal to the Bing–Fisher argument was advanced [39]. Jeffreys developed a method to employ a low information prior to the different parametrizations of a statistical model. This prior is proportional to the square root of the determinant of the Fisher information matrix. Using the same rule for constructing the Jeffreys prior on the two different alternative parameterizations results in posterior distributions for the two systems that are compatible by the usual rules for transforming variables of integration. This remarkable discovery was important to reinvigorate interest in the Bayesian approach. Nonetheless, it should be noted that nonlinear transformations of the Jeffreys prior still alters the distribution, but it does so in a way that can enable a reconnection to the other parametrizations by the rules for the change of variables. The Jeffreys rule for the binomial model in Equation (19) is a beta distribution where the shape parameters are $a' = b' = 0.5$. This prior is U shaped, and it diverges at the two endpoints of 0 and 1. The binomial Jeffreys prior is not a maximum entropy prior, and it has an H_* value of approximately 0.9668 [40]. See Figure 6 for a graph of the Jeffreys prior for the ψ representation along with the corresponding prior for ψ that is consistent with a uniform distribution for ϕ.

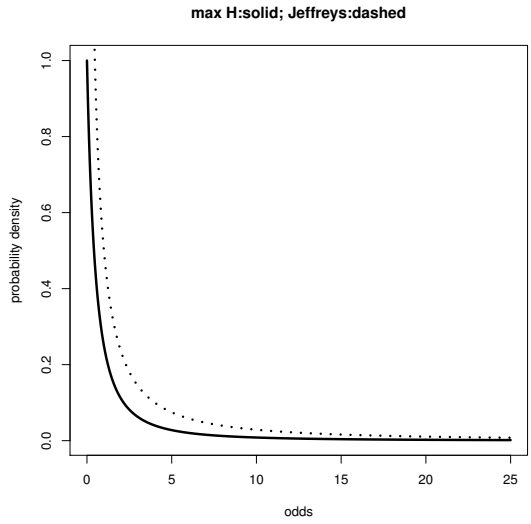

Figure 6. A comparison between the Jeffreys prior (dashed line) for $\psi = \frac{\phi}{1-\phi}$ and the prior consistent with a uniform prior for ϕ (solid line labeled max H), which has a ψ density function of $f(\psi) = (1+\psi)^{-2}$. The Jeffreys density is $f(\psi) = \frac{1}{(1+\psi)\sqrt{\psi}}$. The plot is limited to the interval $(0, 25)$.

5. Discussion

Generally, mathematical objects, such as vectors, matrixes, and groups have been used in many ways in science without generating a vigorous and long-lasting philosophical debate. Exceptions to this observation are probability theory and number theory. The development of thought about integers, rational numbers, irrational numbers, zero, negative numbers, transcendental numbers, complex numbers, and hypercomplex numbers had a slow and contentious evolution. Similarly, there have been contentious debates about the meaning of probability. While there is a consensus on the axioms of probability, there remains a debate about what can have a probability representation and what cannot. All are agreed that the likelihood for data outcomes given a population parameter can have a probability representation. But investigators following the Bayes/Laplace tradition also allowed states of the world or the degree of our knowledge about a researchable question to have a probability representation. However, many other theorists felt that this application of probability is inappropriate because the unknown parameter is a fixed quantity, so it cannot be treated as if it was a random variable. Bayesian researchers answered this concern by pointing out that the prior distribution follows the rules of probability theory, so why not use probability along with Bayes' theorem to obtain a posterior distribution. The prior and posterior distributions are a quantification of the degree of knowledge or information about the parameter of interest. Eventually, with enough data, the variance of the posterior distribution will approach zero in the limit, but at any point in time, the posterior is a reflection of our current knowledge about the parameter. Moreover, no special rules are needed to do point estimation, interval estimation, and tests of hypotheses. All of these important aspects of statistical inference are achieved by applying conventional probability theory on the posterior distribution. Meta-analysis can be performed using the posterior distribution from an earlier study as the prior distribution for a later study.

Critics of the Bayes/Laplace tradition did not question Bayes' theorem itself. Instead, they demonstrated that a given distribution could not be constructed in a consistent fashion. Bertrand's paradox was designed to show that a uniform distribution for generating random chords resulted in different outcomes. This result was troubling because Bertrand's problem was in the area called geometric probability where a likelihood function was not involved. All the pertinent information was contained within the geometric structure. Later the Bing–Fisher concern, which deals with the conflicting results that occurred with nonlinear transformations of the parameters, was also used to reject the Bayes/Laplace approach. To the critics, the resolution to these paradoxical results was to disallow parameters to have a probability representation in the first place. The consequence of these issues was

effective in steering statistical practice by 1922 to universally adopt a relative frequency view for probability. The goal was to make probability strictly objective. The cost of this philosophical choice in the application of probability theory was that special methods were required for point estimation, interval estimation, and decision-making. These methods are ad hoc in the sense that they did not emerge from the simple application of probability theory. Bayesian theorists have highlighted problems and paradoxes that can occur when using frequentist methods. One example is the stopping-rule problem where frequentist methods yield different conclusions from a test of statistical significance depending on the rules for terminating the experiment. This occurs because frequentists compute the likelihood of the observed data as well as the likelihood of the unobserved outcomes that are more extreme. However, in Bayesian statistics, the only likelihood computed, after an experiment is conducted, is the likelihood for the observed data [41,42]. It is outside the scope of this paper to elaborate on any of the problems with frequentist procedures, but Chechile provided a specific example of the stopping-rule problem as well as other problematic issues with frequentist methods [40].

Bayesian statistics now has a growing number of users, but the frequentist tradition is still the dominant statistical approach employed in most research areas. Even in the twenty-first century, Bertrand's paradox was on the books as an unresolved problem. This paper has resolved Bertrand's paradox by showing there is a best answer to the problem. In the process of resolving Bertrand's problem, it was also shown that the same approach provided an answer to the Bing–Fisher concern about Bayesian statistics being sensitive to a re-parametrization of the model. Thus, there is now no principled argument against the use of the uniform distribution as a representation for a parameter over a *finite support interval*. In fact, there is an advantage to this prior as a low-information prior because it represents maximum entropy.

Funding: This research received no external funding.

Data Availability Statement: The data plotted in Figures 4 and 5 are random outcomes based on implementing the software in respectively Appendixes B.1 and B.2.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Proof of Theorem 1

Proof. For $i = 1, \cdots N-1$, let us denote $p_i = \bar{p} + \epsilon_i$ where $\bar{p} = \frac{1}{N}$, which is the mean for the distribution and where ϵ_i is the ith deviation from the mean for outcome i. Thus, it follows that $p_N = 1 - (N-1)\bar{p} - \sum_{i=1}^{N-1} \epsilon_i$. The maximum entropy distribution can be found by setting $\frac{\partial H}{\partial \epsilon_i} = 0$ for each ϵ_i, $i = 1, \cdots, N-1$ and solving for the values of ϵ_i. For any integer $i \in [1, \cdots, N-1]$ we find that the condition $\frac{\partial H}{\partial \epsilon_i} = 0$ results in

$$\ln \frac{\bar{p} + \epsilon_i}{1 - (N-1)\bar{p} - \sum_{i=1}^{N-1} \epsilon_i} = 0\,, \tag{A1}$$

$$\bar{p} + \epsilon_i = 1 - (N-1)\bar{p} - \sum_{i=1}^{N-1} \epsilon_i\,, \tag{A2}$$

$$\epsilon_i = 1 - N\bar{p} - \sum_{i=1}^{N-1} \epsilon_i\,. \tag{A3}$$

Since $\bar{p} = \frac{1}{N}$, Equation (A3) results in the solution that

$$\epsilon_i = -\sum_{i=1}^{N-1} \epsilon_i \equiv A\,, \tag{A4}$$

$$\sum_{i=1}^{N-1} \epsilon_i = (N-1)A\,, \tag{A5}$$

$$-A = (N-1)A\,, \tag{A6}$$

$$0 = NA\,. \tag{A7}$$

Since $N \geq 2$, it follows that $A = 0$. Thus, from Equation (A4) it follows that $\epsilon_i = 0$ for $i = 1, \cdots, N-1$. Consequently $p_i = \frac{1}{N}$ for $i = 1, \cdots N-1$. Finally $p_N = 1 - \frac{N-1}{N} = \frac{1}{N}$. So, the maximum entropy condition results in a uniform distribution for the p_i values where each outcome has the probability of $\frac{1}{N}$. This outcome results in the maximum entropy value of $H = \frac{\ln N}{\ln 2}$. □

Appendix B. Software for Figures 4 and 5

Appendix B.1. Software for Figure 4

The following code is for an R function that creates four output vectors. These vectors are for the three stochastic processes for generating random chords discussed in association with Figure 4. The vector **xm** consists of 200 midpoints on the [0, 2] interval (i.e., 0.005, 0.015, $\cdots$, 1.995). The vectors **pl**, **pA**, and **pu** correspond to the empirical probability density estimated based on 100,000 Monte Carlo random samples. To implement this function the user needs to load the code into the R workspace and then create an R output object from the command **A <- three_random_chord_processes()**. The user can see the four output vector values via commands such as **A$pl**.

```
three_random_chord_processes<-function(pl = rep(0,200),
pa=rep(0,200),pu=rep(0,200),samples=1000000){
# software for generating random chords three ways
x=seq(0,2,.01)
rL=runif(samples,0,2)
rA=runif(samples,0,pi)
ru=runif(samples,0,1)
LA=2*sin(.5*rA)
Lu=2*((1-ru^2)^.5)
Fl=rep(0,200)
FA=rep(0,200)
Fu=rep(0,200)
for (i in 2:201){
Fl[i-1]=(sum(rL<x[i]))/samples
FA[i-1]=(sum(LA<x[i]))/samples
Fu[i-1]=(sum(Lu<x[i]))/samples
}
pl=rep(0,200)
pA=rep(0,200)
pu=rep(0,200)
pl[1]=Fl[1]
pA[1]=FA[1]
pu[1]=Fu[1]
for (i in 2:200){
pl[i]=Fl[i]-Fl[i-1]
pA[i]=FA[i]-FA[i-1]
pu[i]=Fu[i]-Fu[i-1]
```

```
}
pl=100*pl
pA=100*pA
pu=100*pu
xm=rep(0,200)
for (i in 1:200){
xm[i]=x[i]+.005}
outlist<- list(xm=xm,pl=pl,pA=pA,pu=pu)
# outlist is the list for output vectors
# xm values are the midpts. of chord length intervals
# pl values are prob. density estimates for the intervals
# that are based on a uniform over chord length
# pA are prob. density estimates for angular method
# pu are prob. density estimates for radial method
}
```

Appendix B.2. R Function Used for Figure 5

The following R function plots n chords for a circle of radius r.

```
max_entropy_chords<- function(n, r=1){
# Function plots n chords for a circle of radius r
# via the maximum entropy method for chord length.
theta=runif(n,0,2*pi)
x0=r*cos(theta); y0=r*sin(theta)
# x0 and y0 are paired vectors of length n for the coordinates
# for the initial points for the chords.
# These points are uniformly distributed over the circumference
# of the circle.
L=runif(n,0,2*r)
# L is a vector of length n for the random chord lengths that
# are sampled from a uniform distribution on [0,2r].
alpha=rep(0,n)
alphaplus=theta+acos(1-((L^2)/(2*r*r)))
alphaminus=theta-acos(1-((L^2)/(2*r*r)))
for (i in 1:n){
clip=runif(1,0,1)
if (clip<=.5){alpha[i]=alphaminus[i]} else {alpha[i]=alphaplus[i]}
}
# There are two possible intersections between the two circles.
# One circle with center (0,0) and radius r and the other
# circle with center (x0,y0) and radius L. The angle for the
# intersection point is either theta+arccos(1-((L*L/(2*r*r))) or
# theta-arccos(1-((L*L/(2*r*r))). For each chord one of the two
# solutions is randomly sampled with a 50/50 chance. These radial
# angles are stored in the alpha vector. Thus, there are n alpha
# values that correspond to the n values sampled for theta, and L.
x1=r*cos(alpha)
y1=r*sin(alpha)
# The points (x1,y1) are the set of for second chord endpoints.
# There are thus n line segments for the n first points (x0,y0)
# that are paired with their corresponding second points (x1,y1).
## The following code plots the circle and the n chords.
cirang=seq(0,2*pi,length=300)
xc=r*cos(cirang)
yc=r*sin(cirang)
```

```
par(pty="s")
plot(xc,yc,type="l",xlab="x",ylab="y",lwd=1)
segments(x0,y0,x1,y1,lwd=2)
}
```

References

1. Bertrand, J. *Calcul des Probabilités*; Gauthier-Villars: Paris, France, 1889.
2. Mosteller, F. *Fifty Challenging Problems in Probability with Solutions*; Dover: New York, NY, USA, 1965.
3. Galavotti, M.C. The Interpretation of probability: Still an open issue? *Philosophies* **2017**, *2*, 20. [CrossRef]
4. Zabell, S.R.A. Fisher on the history of inverse probability. *Stat. Sci.* **1989**, *4*, 247–256. [CrossRef]
5. Laplace, P.S. Mémoire sur la probabilité des causes par les événements. *Oeuvres Complétes* **1774**, *8*, 5–24.
6. Ellis, R.L. On the foundations of the theory of probability. *Trans. Camb. Philos. Soc.* **1842**, *8*, 1–6.
7. von Mises, R. *Probability, Statistics and Truth*; Dover: New York, NY, USA, 1957.
8. Aldrich, J. Fisher and the making of maximum likelihood. *Stat. Sci.* **1997**, *12*, 162–176.
9. Fienberg, S. When did Bayesian inference become "Bayesian"? *Bayesian Anal.* **2006**, *1*, 1–40. [CrossRef]
10. Fisher, R.A. On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc.* **1922**, *222*, 309–368.
11. Bing, F. Om aposteriorisk Sandsynlighed. *Mat. Tidsskr.* **1879**, *3*, 1–22.
12. Hald, A. *A History of Mathematical Statistics from 1759 to 1930*; Wiley: New York, NY, USA, 1998.
13. Ramsey, E.P. Truth and probability. In *The Foundations of Mathematics and Other Logical Essays*; Braithwaite, Ed.; Trübner & Co.: London, UK, 1931; pp. 156–198.
14. de Finetti, B. La prévision se lois logiques, ses sources subjectives. *Ann. Henri Poincare* **1937**, *7*, 1–68.
15. Shackel, N. Bertrand's paradox and the principle of indifference. *Philos. Sci.* **2007**, *74*, 150–175. [CrossRef]
16. Johnson N.L.; Kotz, S. *Continuous Univariate Distributions*; Wiley: New York, NY, USA, 1970; Volume 2.
17. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 623–656. [CrossRef]
18. Shannon, C.E.; Weaver, W.W. *The Mathematical Theory of Communication*; Univ. Illinois Press: Champaign, IL, USA, 1949.
19. Lehmann, E.L. A general concept of unbiasedness. *Ann. Math. Stat.* **1957**, *22*, 587–592. [CrossRef]
20. Keeler, P. Software. Available online: https://hpaulkeeler.com (accessed on 12 July 2023).
21. Vidovič, Z. Random chord in a circle and Bertrand's paradox: New generation method, extreme behavior and length moments. *Bull. Korean Math. Soc.* **2021**, *2*, 433–444.
22. Jaynes, E.T. The well-posed problem. *Found. Phys.* **1973**, *3*, 477–492. [CrossRef]
23. Borel, E. *Eléments de la Théorie des Probabilités*; Herman et Fils: Paris, France, 1909.
24. Poincaré, H. *Calcul des Probabilités*; Gauthier-Villars: Paris, France, 1912.
25. Kac, M. More on randomness. *Am. Sci.* **1984**, *72*, 282–283.
26. van Fraassen, B. *Laws and Symmetry*; Clarendon Press: Oxford, UK, 1989.
27. Gyenis, Z.; Rédei, M. Defusing Bertrand's paradox. *Br. J. Philos. Sci.* **2012**, *66*, 1–18. [CrossRef]
28. Faigle, U. Is there a 'correct' solution to Bertrand's paradox. *Int. J. Math. Educ. Sci. Technol.* **1979**, *10*, 121–124. [CrossRef]
29. Marinoff, L. A resolution of Bertrand's paradox. *Philos. Sci.* **1994**, *61*, 1–24. [CrossRef]
30. Porto, P.D.; Crosignani, B.; Ciattoni, A.; Liu, H.C. Bertrand's paradox: A physical way out along the lines of Buffon's needle throwing experiment. *Eur. J. Phys.* **2011**, *32*, 819–825. [CrossRef]
31. Aerts, D.; Sassoli de Bianchi, M. *Universal Measurement: How to Free Three Birds in One Move*; World Scientific: Singapore, 2017.
32. Holbrook, J.; Kim, S.S. Bertrand's paradox revisited. *Math. Intell.* **2000**, *22*, 16–19. [CrossRef]
33. Ardakani, M.K.; Wulff, S.S. An extended problem to Bertrand's paradox. *Discuss. Math. Probab. Stat.* **2014**, *34*, 23–34. [CrossRef]
34. Kaushik, P. A new solution of Bertrand's paradox. *Theory Probab. Appl.* **2022**, *67*, 158–160. [CrossRef]
35. Chiu, S.S.; Larson, R.C. Bertrand's paradox revisited: More lessons about that ambiguous word, random. *J. Ind. Syst. Eng.* **2009**, *1*, 1–26.
36. Jevremovic, V.; Obradovic, M. Bertrand's paradox: Is there anything else? *Qual. Quant.* **2012**, *46*, 1709–1714. [CrossRef]
37. Bayes, T. An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc.* **1764**, *53*, 370–418. [CrossRef]
38. Kolmogorov, A.N. *Grundbegriffe der Wahrcheinlichkeitsrechnung*; Springer: Berlin/Heidelberg, Germany, 1933.
39. Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proc. R. Soc. A* **1946**, *186*, 453–461.
40. Chechile, R.A. *Bayesian Statistics for Experimental Scientists: A General Introduction Using Distribution-Free Methods*; MIT Press: Cambridge, MA, USA, 2020.
41. Lindley, D.V.; Phillips, L.D. Inference for a Bernoulli process (a Bayesian view). *Am. Stat.* **1976**, *30*, 112–119.
42. Berger, J.O.; Wolpert, R.L. *The Likelihood Principle*; Institute of Mathematical Statistics: Hayward, CA, USA, 1988.