philosophies

MDPI

Article

To Be or to Have Been Lucky, That Is the Question

Antony Lesage [1] and Jean-Marc Victor [2,*]

[1] Physico-Chimie des Électrolytes et Nano-Systèmes Interfaciaux, PHENIX, Sorbonne Université, CNRS, F-75005 Paris, France; antony.lesage@sorbonne-universite.fr
[2] Physique Théorique de la Matière Condensée, LPTMC, Sorbonne Université, CNRS, F-75005 Paris, France
* Correspondence: victor@lptmc.jussieu.fr

Abstract: Is it possible to measure the dispersion of ex ante chances (i.e., chances "before the event") among people, be it gambling, health, or social opportunities? We explore this question and provide some tools, including a statistical test, to evidence the actual dispersion of ex ante chances in various areas, with a focus on chronic diseases. Using the principle of maximum entropy, we derive the distribution of the risk of becoming ill in the global population as well as in the population of affected people. We find that affected people are either at very low risk, like the overwhelming majority of the population, but still were unlucky to become ill, or are at extremely high risk and were bound to become ill.

Keywords: ex ante chances; dispersion of chances; chronic diseases; gambling; statistical test; twin studies; principle of maximum entropy

Citation: Lesage, A.; Victor, J.-M. To Be or to Have Been Lucky, That Is the Question. *Philosophies* **2021**, *6*, 57. https://doi.org/10.3390/philosophies6030057

Academic Editors: Fabien Paillusson and Marcin J. Schroeder

Received: 21 April 2021
Accepted: 6 July 2021
Published: 9 July 2021

1. Introduction

"That evening he was lucky": what do we mean by this? It is even weirder when we say: "the luck turned". Does this mean that we could be visited by fortune? Or that some people are luckier than others on certain days? Of course, we cannot rule out the fact that some people may bias the chances of success simply by cheating. Yet, is there any way to assess the dispersion of chances among gamblers (or just the fraction of cheaters)?

This kind of question is part of the field of probability calculus, which aims at determining the relative likelihoods of events (for a nice historical introduction to probability theory, see [1]). Probability calculus started during the summer of 1654 with the correspondence between Pascal and Fermat precisely on the elementary problems of gambling [2]. Symmetry arguments are at the heart of this calculus: for example, for an unbiased coin, the two results—heads or tails—are a priori equivalent and therefore, have the same probability of occurrence of 1/2. This is why it is not anecdotal that Pascal wanted to give his treatise the "astonishing" title "Geometry of Chance". Another illustration of the power of symmetry arguments is the tour de force of Maxwell who managed to calculate the velocity distribution of particles in idealized gases [3]. At the time when he derived what is since called the Maxwell–Boltzmann distribution, there was no possibility to measure this distribution. It was almost 60 years before Otto Stern could achieve the first experimental verification of this distribution [4], around the same time when he confirmed with Walther Gerlach the existence of the electron spin [5], for which he won the Nobel Prize in 1944. The agreement between theoretical and experimental distributions was surprisingly good. Since its invention in the middle of the 17th century, probability calculus has accompanied most if not all new fields of science, especially since the beginning of the 20th century with the burst of genetics and quantum physics up to the most recent developments of quantum cognition [6], not to mention its countless applications in finance and economy.

In probability theory, events are usually associated with random variables that are measurable. For example, in the heads or tails game, heads may be associated with 1 and tails with 0. Then, for a given number N of draws, one can count the number of times the

heads are flipped. This number k is between 0 and N and the ratio k/N is the frequency of the heads. If the coin is unbiased, this frequency fluctuates around $1/2$ when the game (N draws for each game) is played many times. Importantly, the frequency is observed ex post, i.e., after the game is played; the mean frequency is used as a measure of probability of getting a head. This is the usual way of assessing probabilities in the frequentist perspective of statistics. Remember that assessing probabilities for anticipating the outcome of future events is the very purpose of statistics. However, it is not always possible to deduce probabilities from frequency measurements. For example, suppose that each coin is tossed only once. Can we still assess the dispersion of chances among gamblers?

Dispersion of chances is far from being limited to gamblers. Disease risk is another area where people may be and actually are unequal for genetic or environmental reasons. In this case, the result of a "draw" is whether or not you have a disease D. The "game" is then limited to one "draw" per person. Of course, the mean probability to become ill can still be observed. Yet, can we assess the *dispersion* of disease risks? Then, if so, how can we? As a last emblematic example, we mention social opportunities. Measuring inequality of opportunity is a crucial issue with considerable political stakes, though it is extremely difficult to assess. On this last point, we postpone the in-depth study of the measure of unequal opportunities to a further work.

In all these examples, be it gambling, disease, or social opportunity, the ex ante chances are themselves random variables that cannot be deduced from frequency measurements nor be induced by symmetry arguments. They are hidden variables. Nevertheless, we argue here that the probability distribution function (pdf) of the ex ante chances can be assessed and we propose some tools to (i) first *test* the existence of some dispersion of chances in the population; (ii) then, *infer* the pdf of the ex ante chances; and (iii) explore more specifically the relevance of those tools to and their consequences in the field of chronic diseases, i.e., diseases that occur at various ages and persist throughout life [7]. Importantly, we do not assume any hypothetical functional form for the pdf of chances and then infer its parameters by Bayesian inference as is usually carried out. Here, we first test the inequality of chances in the population, then infer the functional form of the pdf by means of the principle of maximum entropy.

2. A Simple Draw Is Not Enough

Let us first assume that there is a sample of n people tossing a coin and that each of them has a probability p_i to win (hence, $1 - p_i$ to lose). In an unbiased game, all the p_i are identical and equal to $1/2$. Imagine that some gamblers are luckier, others less fortunate—hence, some p_i are greater than $1/2$, others less than $1/2$. This means that the p_i are random variables that are drawn from a probability distribution $f(p)$ that is different from $\delta(p - 1/2)$, where δ is the Dirac delta function. Let Φ and Σ^2 be the mean and variance of $f(p)$. Let us assume now that each individual plays N times. The result of each draw j of the individual i is a random variable X_i^j, either 1 in cases of success or 0 in cases of failure. This is a Bernoulli process: for each i, the random variables X_i^j are i.i.d. (independent, identically distributed, i.e., the probability of success p_i is the same for the N draws of i). Let us define $S_i = \sum_{j=1}^{N} X_i^j$ as the score over N draws. It is the number of times the individual i has won. Given the risk p_i, S_i is a random variable that follows a binomial distribution $B(N, p_i)$. The mean and the variance of S_i for a given risk p_i are

$$E_N[S_i|p_i] = Np_i$$
$$Var_N[S_i|p_i] = Np_i(1 - p_i)$$

Once every individual has played N times, we obtain an estimation of the distribution of the n random variables S_i as a histogram over the $N + 1$ values $k = 0, 1, 2, \ldots, N$. These random variables S_i are independent but *non-identically* distributed, as the p_i are different from one individual to another.

Just as the p_i are drawn from the distribution $f(p)$, the S_i are the realizations of a random variable S (which takes the $N+1$ discrete values $k = 0,\ 1,\ 2, \ldots, N$). The underlying distribution is no longer only on the random variable S, but on the joint probability of S and p. Thus, the marginal probability distribution function of S is given as follows:

$$\forall k = 0,\ 1,\ \ldots, N\ P_N[S=k] = E[P_N[S=k|p]] = E\left[C_N^k p^k (1-p)^{N-k}\right] = \int_0^1 dp\ f(p) C_N^k p^k (1-p)^{N-k} \quad (1)$$

where $E[\cdot]$ is expected value of $\cdot$ with the probability distribution of p, $f(p)$; and $C_N^k = \frac{N!}{k!(N-k)!}$ is the binomial coefficient "N choose k", i.e., the number of k-combinations of N. The mean of S is

$$E_N[S] = E[E_N[S|p]] = E\left[\sum_{k=0}^{N} k C_N^k p^k (1-p)^{N-k}\right] = E[Np]$$

where $E_N[\cdot]$ is the expected value of $\cdot$ with the probability distribution of S, P_N(S); and $E_N[S|p]$ is the conditional expected value of S for a given underlying probability p, i.e., the Bernoulli distribution. Since Φ is the mean of the distribution $f(p)$,

$$E_N[S] = NE[p] = N\Phi \quad (2)$$

and the variance of S is

$$Var_N[S] = E_N\left[S^2\right] - E_N[S]^2$$

where

$$E_N\left[S^2\right] = E\left[E_N\left[S^2\middle|p\right]\right] = E\left[\sum_{k=0}^{N} k^2 C_N^k p^k (1-p)^{N-k}\right] = E\left[Np(1-p) + (Np)^2\right]$$

hence

$$E_N\left[S^2\right] = N\left(E[p] - E\left[p^2\right]\right) + N^2 E\left[p^2\right]$$

and

$$Var_N[S] = N\left(E[p] - E\left[p^2\right]\right) + N^2\left(E\left[p^2\right] - E[p]^2\right)$$

Now, we recall the first two moments of $f(p)$, given its mean Φ and its variance Σ^2

$$E[p] = \Phi$$
$$E\left[p^2\right] = \Sigma^2 + \Phi^2$$

so that

$$Var_N[S] = N\left(\Phi(1-\Phi) - \Sigma^2\right) + N^2\Sigma^2 = N\Phi(1-\Phi) + N(N-1)\Sigma^2 \quad (3)$$

Equation (3) gives the variance of the score S as a function of the variance Σ^2 of $f(p)$. In the following for the sake of clarity, we will refer to Σ^2 as the *dispersion of chances*.

Note that within the limit $N \to \infty$, the probability distribution of the random variable S/N converges to the distribution $f(p)$.

We simulated two populations of n gamblers each drawing N times. Both populations have the same mean chance of gain $\Phi = 1/2$. However, in the first population, the chance distribution is

$$f_0(p) = \delta\left(p - \frac{1}{2}\right)$$

where there is no dispersion of chances, i.e., $\Sigma^2 = 0$. In contrast, in the second population, the chance distribution is

$$f_1(p) = \frac{1}{2}[\delta(p) + \delta(1-p)]$$

where the dispersion is maximal, i.e., $\Sigma^2 = 1/4$. The histograms are plotted in Figure 1 for a number n of gamblers ranging from 10 to 100 and a number N of draws ranging from 1 to 4. Equation (3) shows that if $N = 1$, the variance $Var_1[S] = \Phi(1-\Phi)$ does not depend on the dispersion of chances Σ^2. As a matter of fact, when $N = 1$, the gains are either 0 or 1 so that the histogram of gains has only two bins, one at 0, the other at 1. The mean of gains is Φ and the variance is $\Phi(1-\Phi)$. Neither the mean nor the variance depends on the dispersion of chances Σ^2. Moreover, according to Equation (1), the histogram of gains itself depends only on the mean of the distribution $f(p)$:

$$\begin{aligned} P_1[S=0] &= E[1-p] = 1-\Phi \\ P_1[S=1] &= E[p] = \Phi \end{aligned} \tag{4}$$

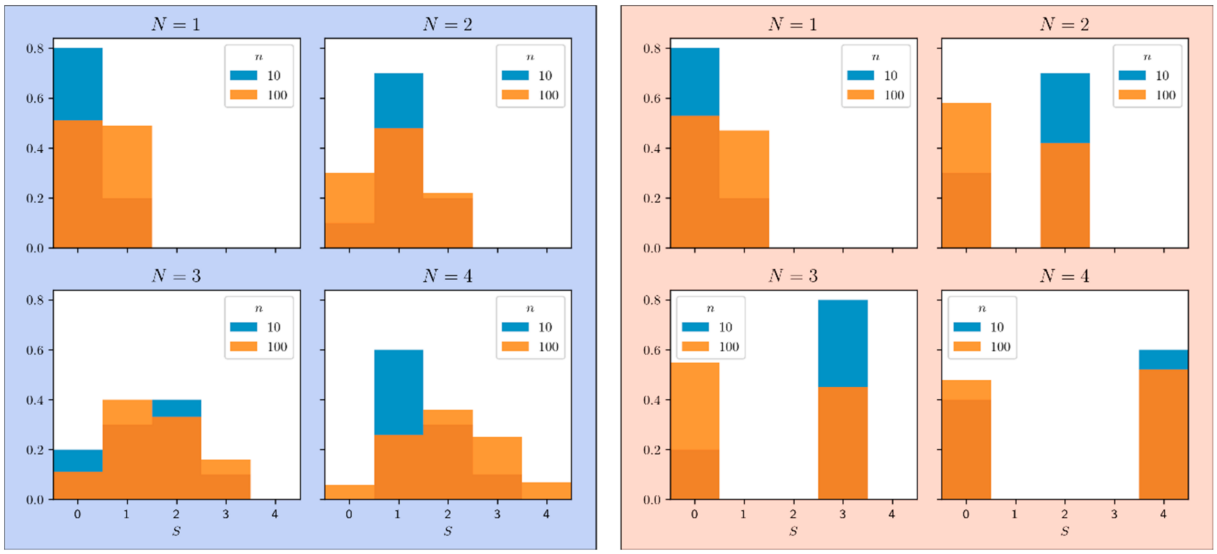

Figure 1. Simulated histograms of gains S for two distributions f_0 and f_1: on the left-hand side $f_0(p) = \delta(p-1/2)$ and on the right-hand side $f_1(p) = \frac{1}{2}[\delta(p) + \delta(1-p)]$. In each case, the histogram of success is plotted for increasing values of the number N of draws ($N = 1, 2, 3, 4$) and for two numbers n of gamblers: $n = 10$ (in blue) and 100 (in orange). For $N = 1$, note that the histogram for f_0 is similar to the histogram for f_1 and both histograms converge to the same limit as n goes to infinity. On the contrary, for each $N \geq 2$, the histograms for f_0 and f_1 diverge as n increases.

The histogram of gains, therefore, cannot provide information on the dispersion of chances, as shown in Figure 1 for the case $N = 1$ where the histograms for f_0 and f_1 are indistinguishable. This means that *a simple draw is not enough* to extract the variance of $f(p)$ from the histogram of gains; multiple draws are necessary, though are they sufficient?

3. A Statistical Test of the Dispersion of Chances

We then note in Figure 1 that the histogram of gains for two draws ($N = 2$) has three bins, one at 0, the second at 1 and the third at 2, with the following values:

$$P_2[S=0] = E\left[(1-p)^2\right] = (1-\Phi)^2 + \Sigma^2$$

$$P_2[S=1] = E[2p(1-p)] = 2\Phi(1-\Phi) - 2\Sigma^2 \tag{5}$$

$$P_2[S=2] = E\left[p^2\right] = \Phi^2 + \Sigma^2 \tag{6}$$

Hence, the histogram of gains now depends on (and only on) both the mean and the variance of $f(p)$. Note that Equation (5) shows that $\Sigma^2 \leq \Phi(1-\Phi)$, since $P_2[S=1] \geq 0$; moreover, $\Phi(1-\Phi)$ is maximal when $\Phi = 1/2$. For three or more draws, we could also have access to higher order moments of $f(p)$. Nevertheless, the minimum condition for the presence of a probability dispersion is that the variance of $f(p)$ is non-zero. We

therefore propose to design a statistical test that will be able to discriminate between both following hypotheses:

Null hypothesis H_0: everybody has the same probability Φ of gain. This means that $f_0(p) = \delta(p - \Phi)$ whose mean is $E[p] = \Phi$ and dispersion $\Sigma^2 = 0$.

Alternative hypothesis H_1: f has the same mean Φ but there is some dispersion of chances among the population, so that some people are luckier than others; hence, f has a non-zero dispersion Σ^2.

According to H_0, the mean of N draws is Φ and the variance is $N\Phi(1-\Phi)$, whereas according to H_1, the mean of N draws is also Φ but the variance is $N\left(\Phi(1-\Phi) - \Sigma^2\right) + N^2\Sigma^2$. Hence, if *the variance* $Var_N[S]$ grows *linearly* with N, then all individuals have the same probability p of success. If, on the contrary, $Var_N[S]$ grows *quadratically* with N, then not all individuals have the same chance of success. We can, therefore, rephrase our hypothesis test as the following alternative based on the dependence of the variance $Var_N[S]$ on the number N of draws:

Null hypothesis H_0: the variance $Var_N[S]$ grows linearly with N.

Alternative hypothesis H_1: the variance $Var_N[S]$ grows quadratically with N.

Figure 2 shows how the variance of S varies as a function of the number of draws N for two typical distributions of mean $1/2$: f_1 with zero dispersion and f_2 with maximum dispersion $1/4$. The distribution f_1 (resp. f_2) illustrates the case of a variance growing linearly (resp. quadratically) with N.

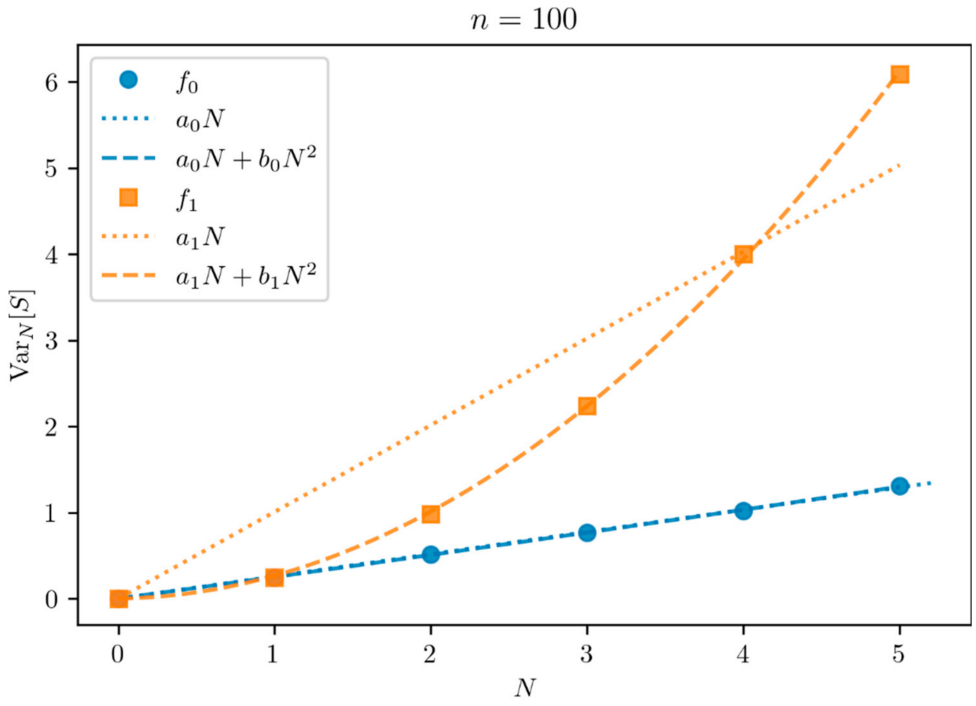

Figure 2. Linear regression fits $Var_N[S]$ for f_0 (blue dotted line), with $a_0 = 0.251 \pm 0.005$ in agreement with Equation (3) when $\Sigma^2 = 0$. Moreover, a_0 agrees with the expected value $\Phi(1-\Phi) = 1/4$. The quadratic fit (blue dashed line) yields an equivalent result. At odds with f_0, the linear regression does not fit $Var_N[S]$ for f_1 (orange dotted line), whereas the quadratic fit (orange dashed line) is excellent, with: $a_1 = 0.01 \pm 0.01$ and $b_1 = 0.244 \pm 0.006$. Here, b_1 agrees with the expected value $\Sigma^2 = 1/4$ and a_1 with the expected value $\Phi(1-\Phi) - \Sigma^2 = 0$.

A relevant statistical test is needed to discriminate between the two hypotheses H_0 and H_1, or at least to reject the null hypothesis H_0. Moreover, in the remainder of this paper, we are particularly interested in the case $N = 2$. It is then necessary to reformulate our hypotheses because it becomes difficult to discriminate the quadratic behavior from the linear behavior with only three points. Therefore, we rephrase our hypothesis test, based on the fact that the number of draws is limited to $N = 2$:

Null hypothesis H_0: the variance of S reads $Var_2[S] = 2\Phi(1-\Phi)$, i.e., $\Sigma^2 = 0$.

Alternative hypothesis H_1: the variance of S reads $Var_2[S] = 2\Phi(1-\Phi) + 2\Sigma^2$ with $\Sigma^2 > 0$.

To estimate the variance of S from a sample of n individuals, the unbiased variance estimator is used:

$$V_n = \frac{1}{n-1}\sum_{i=1}^{n}\left(S_i - \overline{S}\right)^2$$

where $\overline{S}$ is the mean estimator

$$\overline{S} = \frac{1}{n}\sum_{i=1}^{n} S_i$$

The estimation of the variance of S, V_n, from a sample of finite size n is subject to statistical fluctuations. Thus, our hypotheses become:

Null hypothesis H_0: $V_n - 2\Phi(1-\Phi)$ is compatible with 0 considering the error bars, i.e., the standard deviation of V_n.

Alternative hypothesis H_1: $V_n - 2\Phi(1-\Phi) = 2\Sigma^2 > 0$.

The variance of V_n reads (see Appendix A)

$$Var_{\Sigma^2}[V_n] = \frac{2n}{(n-1)^2}\left[\Psi(1-2\Psi) + 7(1-4\Psi)\Sigma^2 - 2\Sigma^4\right] + \frac{8}{(n-1)^2}\left(\Psi + \Sigma^2\right)^2 \quad (7)$$

where $\Psi = \Phi(1-\Phi)$. Its asymptotic expression for $n \gg 1$ reads

$$Var_{\Sigma^2}[V_n] \sim \frac{2}{n}\left[\Psi(1-2\Psi) + 7(1-4\Psi)\Sigma^2 - 2\Sigma^4\right] \quad (8)$$

Figure 3 compares the expression of the variance $Var_{\Sigma^2}[V_n]$ (black dashed line) obtained in Equation (7) and its asymptotic expression (grey dashed line) in Equation (8) with simulations (blue dots) and shows good agreement.

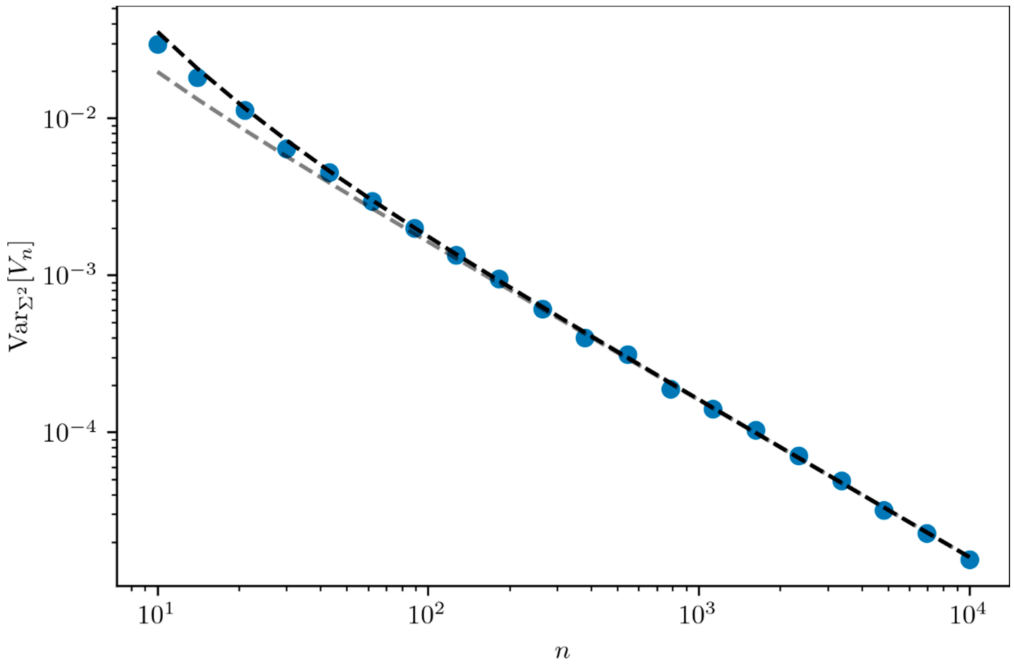

Figure 3. Evolution of the variance of V_n for $N = 2$ as a function of the number n of players. The blue dots are simulated with $\Phi = 0.5$ and $\Sigma^2 = 0.15$. The black dashed line corresponds to the variance of V_n according to Equation (7). The grey dashed line corresponds to the leading-order term in $1/n$ of the expected variance in Equation (8).

It can be noted that the distribution of V_n tends towards a normal distribution $\mathcal{N}(E_{\Sigma^2}[V_n], Var_{\Sigma^2}[V_n])$ of mean $E_{\Sigma^2}[V_n] = Var_2[S]$ and variance $Var_{\Sigma^2}[V_n]$. Now, we wish to estimate the probability of having obtained a value as high as V_n under the null hypothe-

sis H_0, i.e., the p-value. Since V_n follows a normal distribution, the p-value can be expressed as follows

$$p\text{-value} = \frac{1}{2}\left(1 - \mathrm{erf}\left(\frac{z}{\sqrt{2}}\right)\right) = \frac{1}{2}\mathrm{erfc}\left(\frac{z}{\sqrt{2}}\right) \tag{9}$$

where erf and erfc are, respectively, the error function and the complementary error function. By posing $E_0[V_n]$ and $Var_0[V_n]$ as the mean and the variance of V_n under the null hypothesis H_0, i.e., $\Sigma^2 = 0$, we have

$$z = \frac{V_n - E_0[V_n]}{\sqrt{Var_0[V_n]}} = \frac{V_n - 2\Phi(1-\Phi)}{\sqrt{Var_0[V_n]}}$$

Within the limit of large sample sizes $n \gg 1$, one can write using, again, $\Psi = \Phi(1-\Phi)$:

$$z \sim \sqrt{n}\frac{V_n - 2\Psi}{\sqrt{2\Psi(1-2\Psi)}}$$

In the context of Figure 2 restricted to the case $N = 2$ and $n = 100$, the estimated variance V_n for the distribution f_1 (of mean $1/2$ and dispersion $1/4$) leads to $z = 20$, i.e., a p-value of 10^{-87}. This allows us to reject the null hypothesis in this case.

4. Dispersion of Disease Risks for Twins

Inequality in disease risk is a major public health issue [8,9]. Of course, part of this inequality is known to depend on genetic and environmental factors. At the turn of the 2000s, a new approach called genome wide association studies (GWAS) was designed to characterize the genetic predisposition to a chronic disease [10]. GWAS are supposed to find in particular the genes involved in a given disease, and among these genes, the variants most at risk, i.e., the DNA sequences of a given gene that are more represented in the people affected by the disease. Such variants characterize the genetic predisposition to the disease. The mean frequency that an individual will become ill in a given population, specified by genetic and environmental factors, can then be measured. As usual, this frequency can be used as a measure of the probability of becoming ill. However, can we assess the dispersion of disease risk, if only it exists, in this specific population? More generally, is there any way to assess the dispersion of risk in a more objective manner, without any a priori assumption on presumed risk factors? Here comes into play a providential help from the existence of twins. Identical twins, also called monozygotic (MZ) twins, have the same genome, shared the same fetal environment and, generally, share the same living conditions. Therefore, they are most likely to also share the same probability of becoming ill, whatever the disease. Identical twins are, therefore, like a player betting twice. This is much related to the gambling question addressed above for $N = 2$ (two draws). Indeed, as both twins have the same probability p of having disease D, the status—healthy or ill—of each of the two twins is equivalent, respectively, to the outcome—loss or gain—of each of the two draws by one and the same gambler. In this situation, probability p is called a *risk*. Let $f(p)$ be the probability distribution function of the risk of having disease D in the population. We define the random variable S as above, i.e., $S = 0$ if both twins are healthy, $S = 1$ if only one of the two twins is ill and $S = 2$ if both twins are ill. The mean Φ and variance of S are given by Equations (2) and (3), respectively, hence for $N = 2$

$$E_2[S] = 2\Phi \tag{10}$$

$$Var_2[S] = 2\Phi(1-\Phi) + 2\Sigma^2$$

Then, if V_n is significantly greater than $\overline{S}\left(1 - \overline{S}/2\right)$, which amounts to carrying out the hypothesis test presented in the above section, we can conclude that there is some dispersion of the disease risk. As we will see below, the dispersion is in fact unusually large. However, before that, let us calculate the twin concordance rate of the disease D. In

genetics, the twin concordance rate is the probability τ that a twin is affected given that his/her co-twin is affected:

$$\tau = P[X_2 = 1|X_1 = 1] = \frac{P[X_1 = 1, X_2 = 1]}{P[X_1 = 1]} = \frac{P_2[S = 2]}{P[X_1 = 1, X_2 = 1] + P[X_1 = 1, X_2 = 0]}$$

hence

$$\tau = \frac{P_2[S = 2]}{P_2[S = 2] + \frac{1}{2}P_2[S = 1]} = \frac{2P_2[S = 2]}{2P_2[S = 2] + P_2[S = 1]}$$

Note that τ is equal to the probandwise concordance rate, which is known to best assess the twin concordance rate [11].

Using Equations (4) and (6), we can also reformulate the concordance rate of twins in terms of the moments of the distribution $f(p)$:

$$\tau = \frac{P[X_1=1,X_2=1]}{P[X_1=1]} = \frac{P_2[S=2]}{P_1[S=1]} = \frac{E[p^2]}{E[p]} \tag{11}$$

Note we can generalize the concordance rate for a N-tuple:

$$\tau_N = \frac{P_N[S = N]}{P_1[S = 1]} = \frac{E[p^N]}{E[p]}$$

Using Equations (6) and (10), we obtain

$$\tau = \frac{\Phi^2 + \Sigma^2}{\Phi}$$

Thus, the relative risk $RR = \tau/\Phi$ is equal to

$$RR = \frac{E[p^2]}{E[p]^2} = \frac{\Phi^2+\Sigma^2}{\Phi^2} = 1 + \frac{\Sigma^2}{\Phi^2} \tag{12}$$

The twin concordance rate can also be computed using the probability density function $f_a(p)$ restricted to the population of *affected* people. Let $f(X,\ p)$ be the joint probability of an individual to have a risk $p \in [0, 1]$ and to be in the state $X \in \{0, 1\}$. According to Bayes' theorem, also known as the theorem of the probability of causes since it was independently rediscovered by Laplace [12], we write

$$f(X,\ p) = f(p|X)P(X) = f(X|p)f(p)$$

hence

$$f(p|X) = \frac{f(X|p)f(p)}{P(X)}$$

Then, $f(p|X = 1)$ is the distribution of the risk p in the population of affected people

$$f(p|X = 1) = f_a(p)$$

Now, by definition, we have

$$f(X = 1|p) = p$$

and by noting that $P[X = 1] = P_1[S = 1]$, we also have

$$P[X = 1] = E[f(X = 1|p)] = E[p]$$

This leads to the following expression of the risk distribution function among affected people

$$f_a(p) = \frac{pf(p)}{E[p]}$$

Note that $f_a(p)$ is the so-called "size-biased law" of the risk p of becoming ill. Size-biased laws are found in many contexts, notably rare events [13], Poisson point processes [14] or familial risk of disease [15].

The mean risk in the affected population is then

$$E_a[p] = \int_0^1 p f_a(p) dp = \frac{\int_0^1 p^2 f(p) dp}{E[p]} = \frac{E[p^2]}{E[p]}$$

where $E_a[\cdot]$ is the expected value of $\cdot$ among affected people, with the probability distribution $f_a(p)$. Using Equation (11), we obtain

$$E_a[p] = \tau$$

which proves that the mean risk in the affected population is equal to the twin concordance rate.

We proceed now to evaluate the functional form of the distribution $f(p)$. Using the prevalence and the twin concordance rate of the disease D, we have access to, and only to, the mean Φ and dispersion Σ^2 of $f(p)$. The principle of maximum entropy then provides us with the least arbitrary distribution [16,17]. Dowson and Wragg proved [18] that in the class P of absolutely continuous probability distributions on $[0,1]$ with given first and second moments (i.e., given mean and variance), there exists a distribution in P which maximizes the entropy

$$H[f] = -\int_0^1 f(p) \ln f(p) dp \tag{13}$$

and the corresponding density function $f(p)$ on $[0,1]$ is a truncated normal distribution $f(p; m, s, 0, 1)$, which may be either bell-shaped or U-type. Dowson and Wragg show that when $\Phi \ll 1$ and $\Sigma > \Phi$, which is usual for most if not all chronic diseases (unpublished results), the distribution $f(p; m, s, 0, 1)$ is U-type (see Appendix B). This distribution, which will be simply denoted $f(p; m, s)$ in the following, can then be written

$$f(p; m, s) = \frac{1}{sZ}\sqrt{\frac{2}{\pi}} \exp\left(\frac{(p-m)^2}{2s^2}\right)$$

with

$$Z = \text{erfi}\left(\frac{m}{s\sqrt{2}}\right) + \text{erfi}\left(\frac{1-m}{s\sqrt{2}}\right)$$

The imaginary error function $\text{erfi}(x)$ can be expressed using the Dawson function $D(x)$

$$\text{erfi}(x) = \frac{2}{\sqrt{\pi}} e^{x^2} D(x)$$

Therefore, $f(p; m, s)$ can finally be written

$$f(p; m, s) = \frac{1}{\sqrt{2}s} \frac{\exp\left(\frac{p^2-2mp}{2s^2}\right)}{D\left(\frac{m}{s\sqrt{2}}\right) + e^{\frac{1-2m}{2s^2}} D\left(\frac{1-m}{s\sqrt{2}}\right)}$$

It is straightforward to express Φ and Σ^2 in terms of the parameters m and s:

$$\Phi = m - \frac{\sqrt{2}}{2} s \frac{1 - e^{\frac{1-2m}{2s^2}}}{D\left(\frac{m}{s\sqrt{2}}\right) + e^{\frac{1-2m}{2s^2}} D\left(\frac{1-m}{s\sqrt{2}}\right)} \tag{14}$$

$$\Sigma^2 = -s^2\left\{1 - \frac{1}{s}\frac{\sqrt{2}}{2}\frac{m+(1-m)e^{\frac{1-2m}{2s^2}}}{D\left(\frac{m}{s\sqrt{2}}\right)+e^{\frac{1-2m}{2s^2}}D\left(\frac{1-m}{s\sqrt{2}}\right)} + \frac{1}{2}\frac{\left(1-e^{\frac{1-2m}{2s^2}}\right)^2}{\left[D\left(\frac{m}{s\sqrt{2}}\right)+e^{\frac{1-2m}{2s^2}}D\left(\frac{1-m}{s\sqrt{2}}\right)\right]^2}\right\} \quad (15)$$

Inverting this system of equations to obtain the risk distribution function of the disease D in terms of Φ and Σ^2 is a bit trickier and requires a numerical solver. In the next section, we show the outcome of this general formalism for one specific chronic disease, namely Crohn's disease.

5. Application to Crohn's Disease (CD)

Crohn's disease (CD) is one of the most well-documented chronic diseases, particularly in the field of genetics [19]. Its prevalence Φ and twin concordance rate τ are [20]:

$$\Phi \cong 0.0025$$
$$\tau \cong 0.385$$

Then, the twin relative risk is

$$RR \cong 154$$

hence

$$\Sigma^2 = \Phi^2(RR-1) \cong 0.00096$$
$$\Sigma \cong 0.031$$

which means that

$$\frac{\Sigma}{\Phi} \cong 12 \quad (16)$$

The dispersion of the risk of being affected is, therefore, huge for CD.

It is now necessary to calculate the p-value according to Equation (9) in order to be able to reject (or not) our null hypothesis H_0. To do this, we first need to estimate the number of twin pairs n that remains unknown in the Swedish study [20]. Nevertheless, the number of twin pairs with at least one affected twin is known and equal to $n_1 + n_2 = 31.5$, where $n_1 = 24$ and $n_2 = 7.5$ are the number of discordant and concordant twin pairs, respectively [20]. We can reconstruct the sample size n that would have been needed to obtain n_1 and n_2, with probabilities $P_2[S=1]$ and $P_2[S=2]$:

$$P_2[S=1] + P_2[S=2] = \frac{n_1+n_2}{n}$$

By using Equations (5) and (6), we obtain the following sample size

$$n = \frac{n_1+n_2}{1-(1-\Phi)^2-\Sigma^2} \cong 7809$$

Equation (9) is used by calculating z within the limit of large sample sizes $n \gg 1$. This results in $z \approx 2.4$, which allows us to reject the null hypothesis H_0 with the p-value $\approx 8 \cdot 10^{-3}$.

It is then legitimate to calculate the parameters m and s of the truncated normal distribution $f(p;m,s,0,1)$, which maximizes the entropy $H[f]$ given the mean Φ and the dispersion Σ^2. Solving the system of Equations (14) and (15) for $\Phi = 0.0025$ and $\Sigma = 0.031$ gives

$$m \approx 0.505$$
$$s \approx 0.0278$$

Both probability distribution functions $f(p;m,s)$ and $f_a(p;m,s) = pf(p;m,s)/\Phi$ for CD are plotted in Figure 4a and zoomed in in Figure 4b. Quite remarkably, the probability density function $f_a(p;m,s)$ in the population of affected people has two narrow peaks, one close to $p=0$ and the other one close to $p=1$. This means that there are two quite separate categories of people who become ill: in the left peak (close to $p=0$), people are at very low risk, but still *have been unlucky* to become ill, whereas in the right peak (close

to $p = 1$), people are at extremely high risk, hence *are unlucky* a priori, and indeed, were bound to become ill. Not having any luck (to become ill *because* of high risk) or to have been unlucky (to become ill *despite* low risk), that is the question!

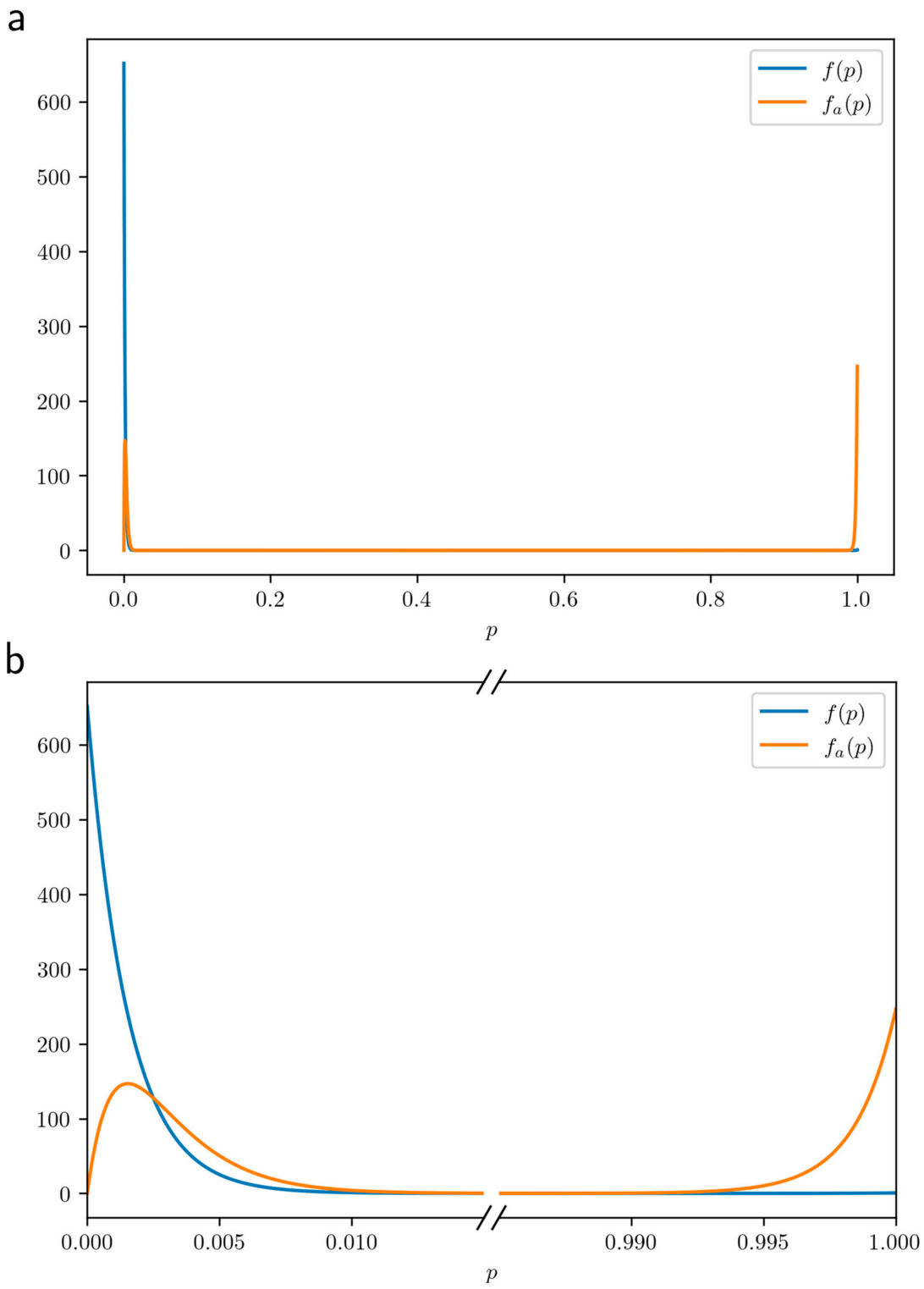

Figure 4. (**a**) CD risk distribution function $f(p; m, s)$ among the population (in blue) is narrow peaked at $p = 0$. The risk distribution function $f_a = pf(p; m, s)\Phi$ among affected people (in orange) has two narrow peaks. (**b**) Zoom in the vicinity of both peaks $p = 0$ and $p = 1$. Concordant twins (almost) all belong to the right peak (at $p = 1$) whereas discordant twins (almost) all belong to the left peak (at $p = 0$).

Finally, we note that concordant twins are very likely to be in the right peak, whereas discordant twins are in the left one. Indeed, when two MZ twins have their common risk p in the left peak, their probability of being concordant is extremely low, of the order of the mean of p^2 restricted to the left peak of $f_a(p)$, which is of the order of 10^{-5}. On the contrary, when two MZ twins have their common risk p in the right peak, their probability to be concordant is extremely high, of the order of 0.997. Interestingly enough, the fraction of people in the right peak (area under the curve) is 38.52%, quite similar to the (probandwise) twin concordance rate of 38.65% [20]. This strongly suggests that concordant twins for a given disease both have a strong predisposition for this disease, whereas discordant twins both have no particular predisposition.

6. Conclusions

Assessing the inequality of chances in a given population is a critical problem that has several issues, notably health and social opportunity. Starting with the simple heads or tails game, we have shown that, although hidden variables such as ex ante chances of gamblers (possibly cheating) cannot be assessed, their *distribution* can actually be assessed whenever multiple draws are available. For this purpose, we have proposed a hypothesis test to evidence the inequality of chances in a given population, then infer the functional form of the probability distribution function of the ex ante chances by means of the principle of maximum entropy, which gives the least arbitrary distribution given the mean and variance of the probability distribution function.

We applied this methodology to chronic diseases and found that the distribution of the risk of becoming ill is usually a U-type truncated normal distribution. We have computed the parameters of this U-type distribution in the case of Crohn's disease using the prevalence and the twin concordance rate of this pathology. Moreover, we have found that the risk distribution function among affected people is bimodal with two narrow peaks, one corresponding to people with no liable risk factor and the other one to people genetically or environmentally destined to become ill. An interesting consequence is that concordant twins for a given disease both have a strong predisposition for that disease, while discordant twins both have no particular predisposition.

One should still not over-interpret the results, as they still rely only on estimates of the prevalence and the twin concordance rate of the disease. It can be thought of as the best possible interpretation in terms of distribution, based on the available information. Nevertheless, maximizing the entropy of the risk distribution function leads to significantly different conclusions than more arbitrary distributions such as, for example, beta-distributions [21].

Twins provide a unique means to play twice at the lottery of diseases. Of course, twins are all the more relevant to assess ex ante chances as they share the same environmental factors. In the same vein, "social twins" or more generally "social clones" would be of great help in assessing inequality of opportunities. However, controlling the environment of such social clones would be rather challenging as the issue of choice comes into play, which may change people's lives with the same opportunities. Assessing the inequality of opportunities is, therefore, one of the most delicate, almost completely open, issues.

Pascal could never complete his treatise "Geometry of Chance". This never-ending treatise is still being written, as evidenced in this Special Issue.

Author Contributions: Conceptualization, J.-M.V.; methodology, A.L., J.-M.V.; writing—original draft preparation, A.L., J.-M.V.; writing—review and editing, A.L., J.-M.V. Both authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This work has benefited from fruitful discussions with Anne Dumay, Jean-Pierre Hugot and Alberto Romagnoni. We thank Bastien Mallein and Laurent Tournier for their careful reading of the manuscript and helpful comments.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Computing the Variance of V_n

To estimate the variance of S from a sample of n individuals, the unbiased variance estimator is used:

$$V_n = \frac{1}{n-1}\sum_{i=1}^{n}\left(S_i - \overline{S}\right)^2$$

where $\overline{S}$ is the mean estimator

$$\overline{S} = \frac{1}{n}\sum_{i=1}^{n} S_i$$

We first recall the following properties of $\overline{S}$:

$$\begin{aligned} E\left[\overline{S}\right] &= E[S] \\ Var\left[\overline{S}\right] &= \frac{Var[S]}{n} \end{aligned}$$

By posing $E\left[\overline{S}\right] = E[S] = m$, we can write

$$V_n = \frac{1}{n-1}\sum_{i=1}^{n}(S_i - m)^2 - \frac{n}{n-1}\left(\overline{S} - m\right)^2$$

hence

$$Var[V_n] = \frac{n}{(n-1)^2}Var\left[(S-m)^2\right] + \frac{n^2}{(n-1)^2}Var\left[\left(\overline{S}-m\right)^2\right] - \frac{2n}{(n-1)^2}Cov\left[\sum_{i=1}^{n}(S_i-m)^2, \left(\overline{S}-m\right)^2\right]$$

To simplify the calculation, we consider sufficiently large samples (typically $n > 30$) so that the distribution of $\overline{S}$ tends towards the normal distribution $\mathcal{N}\left(E\left[\overline{S}\right], Var\left[\overline{S}\right]\right)$ of mean $E\left[\overline{S}\right] = m$ and variance $Var\left[\overline{S}\right] = Var[S]/n$, according to the central limit theorem. The variable $\overline{S}$ is then independent of S_i, which has, as an immediate effect, a null covariance $Cov\left[\sum_{i=1}^{n}(S_i-m)^2, \left(\overline{S}-m\right)^2\right] = 0$. Then, $Var\left[(S-m)^2\right]$ and $Var\left[\left(\overline{S}-m\right)^2\right]$ remain to be determined. Let us start with the latter, which is simpler.

$$Var\left[\left(\overline{S}-m\right)^2\right] = E\left[\left(\overline{S}-m\right)^4\right] - E\left[\left(\overline{S}-m\right)^2\right]^2$$

with

$$E\left[\left(\overline{S}-m\right)^2\right] = Var\left[\overline{S}\right] = \frac{Var[S]}{n}$$

and

$$E\left[\left(\overline{S}-m\right)^4\right] = E\left[\overline{S}^4\right] - 4E\left[\overline{S}^3\right]m + 6E\left[\overline{S}^2\right]m^2 - 3m^4$$

Since $\overline{S}$ is a Gaussian variable, its moments read

$$E\left[\overline{S}^2\right] = m^2 + Var\left[\overline{S}\right]$$

$$E\left[\overline{S}^3\right] = m\left(m^2 + 3\,Var\left[\overline{S}\right]\right)$$

$$E\left[\overline{S}^4\right] = m^4 + 6m^2\,Var\left[\overline{S}\right] + 3\,Var\left[\overline{S}\right]^2$$

All the terms in m cancel each other out, hence

$$Var\left[\left(\overline{S}-m\right)^2\right] = 2\,Var\left[\overline{S}\right]^2 = 2\left(\frac{Var[S]}{n}\right)^2$$

Now all that remains is to determine $Var\left[(S-m)^2\right]$. This term requires expressing the moments of S as a function of the moments (up to the 4th order) of the distribution f. First, let us start by explicating the variance.

$$Var\left[(S-m)^2\right] = E\left[(S-m)^4\right] - E\left[(S-m)^2\right]^2$$

with

$$E\left[(S-m)^2\right] = E\left[S^2\right] - m^2$$

and

$$E\left[(S-m)^4\right] = E\left[S^4\right] - 4E\left[S^3\right]m + 6E\left[S^2\right]m^2 - 3m^4$$

Then, we calculate the ℓ-th moments of S (for $\ell = 2,\ 3,\ 4$)

$$E\left[S^\ell\right] = E_p\left[E_S\left[S^\ell\middle|p\right]\right] = E_p\left[\sum_{k=0}^{N} k^\ell C_N^k p^k (1-p)^{N-k}\right]$$

$$E\left[S^2\right] = NE[p] + N(N-1)E\left[p^2\right]$$

$$E\left[S^3\right] = NE[p] + 3N(N-1)E\left[p^2\right] + N(N-1)(N-2)E\left[p^3\right]$$

$$E\left[S^4\right] = NE[p] + 7N(N-1)E\left[p^2\right] + 6N(N-1)(N-2)E\left[p^3\right] + N(N-1)(N-2)(N-3)E\left[p^4\right]$$

We also have the variance of S expressed with the moments of p:

$$Var[S] = NE[p](1-NE[p]) + N(N-1)E\left[p^2\right]$$

In general, we need to know the higher order moments of the distribution f if we want to go further. However, we are only interested here in the case $N = 2$, where some welcome simplifications arise. It turns out the higher order moments of the distribution f do not contribute to the moments of S.

$$\begin{aligned} E\left[S^2\right] &= 2E[p] + 2E\left[p^2\right] \\ E\left[S^3\right] &= 2E[p] + 6E\left[p^2\right] \\ E\left[S^4\right] &= 2E[p] + 14E\left[p^2\right] \end{aligned}$$

hence,

$$Var\left[\left(\overline{S}-m\right)^2\right] = \frac{8}{n^2}\left(E[p](1-2E[p]) + E\left[p^2\right]\right)^2$$

and

$$Var\left[(S-m)^2\right] = 2E[p](2E[p]-1)(4E[p]-1)^2 + 4E\left[p^2\right]^2 + 2\left(7 - 28E[p] + 32E[p]^2\right)E\left[p^2\right]$$

It is further simplified by using $E[p] = \Phi$ and $E\left[p^2\right] = \Phi^2 + \Sigma^2$.

$$Var\left[\left(\overline{S}-m\right)^2\right] = \frac{8}{n^2}\left(\Phi(1-\Phi) + \Sigma^2\right)^2 \quad Var\left[(S-m)^2\right] = 2\Phi(1-\Phi)\left(1 - 2\Phi + 2\Phi^2\right) + 14(2\Phi-1)^2\Sigma^2 - 4\Sigma^4$$

Then, we obtain the following expression

$$Var[V_n] = \frac{2n}{(n-1)^2}\Big(\Phi(1-\Phi)\Big(1-2\Phi+2\Phi^2\Big)+7(2\Phi-1)^2\Sigma^2-2\Sigma^4\Big)+\frac{8}{(n-1)^2}\Big(\Phi(1-\Phi)+\Sigma^2\Big)^2$$

Finally, we can simplify further by posing $\Psi = \Phi(1-\Phi)$:

$$Var[V_n] = \frac{2n}{(n-1)^2}\big(\Psi(1-2\Psi)+7(1-4\Psi)\Sigma^2-2\Sigma^4\big)+\frac{8}{(n-1)^2}\big(\Psi+\Sigma^2\big)^2$$

Appendix B. The Truncated Normal Distribution $f(p;m,s,0,1)$ Is U-Type When $\Phi \ll 1$ and $\Sigma > \Phi$

The prevalence Φ of chronic diseases is most generally of the order of 10^{-3} and the relative risk RR of MZ twins is then of the order of 100. Thus, according to Equation (12), Σ/Φ is of the order of 10. As an example, $RR \cong 12$ for Crohn's disease (see Equation (16)). Therefore, $\Phi \ll 1$ and $\Sigma > \Phi$ is the rule for chronic diseases.

Dowson and Wragg [18] show that the truncated normal distribution $f(p)$ that maximizes the entropy $H[f]$ (see Equation (13)) with given mean $\mu_1 = \Phi$ and second moment $\mu_2 = \Phi^2 + \Sigma^2$ is U-type when μ_1 and μ_2 are above the arc $O\hat{M}A$ (see Figure 1 and text below in [18]). This dividing curve separates U-type from bell-shaped distributions. On this curve, the distribution $f(p)$ that maximizes the entropy $H[f]$ is no longer a truncated normal distribution but becomes a truncated *exponential* distribution (the arc $O\hat{M}A$ is the set of points (μ_1, μ_2) whose coordinates are the first two moments of truncated exponential distributions on $[0,1]$). A truncated exponential distribution on $[0,1]$ can be written

$$f_{exp}(p) = \frac{\lambda}{1-e^{-\lambda}}e^{-\lambda p}$$

with $\lambda \in \,]-\infty, +\infty[$. On the dividing curve $O\hat{M}A$, the first and second moments of $f_{exp}(p)$ are given by

$$m_1 = \frac{1}{\lambda} - \frac{1}{e^{\lambda}-1} \tag{A1}$$

$$m_2 = \frac{2}{\lambda^2} - \Big(1+\frac{2}{\lambda}\Big)\frac{1}{e^{\lambda}-1} \tag{A2}$$

It is easily seen that $0 < m_1 < 1/2$ when $\lambda \in \,]0, +\infty[$ and $1/2 < m_1 < 1$ when $\lambda \in \,]-\infty, 0[$. The limiting case $\lambda \to 0$ corresponds to $m_1 = 1/2$.

The truncated normal distribution $f(p)$ that maximizes the entropy $H[f]$ with given mean $\mu_1 = \Phi$ and second moment $\mu_2 = \Phi^2 + \Sigma^2$ is U-type when μ_1 and μ_2 are above the arc $O\hat{M}A$, i.e., $\mu_2 > m_2$ for $\mu_1 = m_1$. Now, when $m_1 = \Phi \ll 1$, Equation (A1) gives $\lambda \gg 1$ so that $\lambda \sim 1/m_1$. Then, Equation (A2) gives $m_2 \sim 2/\lambda^2$, hence $m_2 \sim 2m_1^2$, i.e., $m_2 \sim 2\Phi^2$. Therefore, $f(p)$ is U-type if $\Phi^2 + \Sigma^2 > 2\Phi^2$, i.e., $\Sigma > \Phi$.

References

1. Debnath, L.; Basu, K. A short history of probability theory and its applications. *Int. J. Math. Educ. Sci. Technol.* **2014**, *46*, 13–39. [CrossRef]
2. Godfroy-Génin, A.-S. Pascal: The geometry of chance. *Math. Sci. Hum.* **2000**, *150*, 7–39.
3. Maxwell, J.C. Illustrations of the Dynamical Theory of Gases. Part I. On the motions and collisions of perfectly elastic spheres. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1860**, *19*, 19–32. [CrossRef]
4. Stern, O. Eine direkte Messung der thermischen Molekulargeschwindigkeit. *Z. Phys.* **1920**, *2*, 49. [CrossRef]
5. Gerlach, W.; Stern, O. Moment Das magnetische des Silberatoms. *Z. Phys.* **1922**, *9*, 353–355. [CrossRef]
6. Bruza, P.D.; Wang, Z.; Busemeyer, J.R. Quantum cognition: A new theoretical approach to psychology. *Trends Cogn. Sci.* **2015**, *19*, 383–393. [CrossRef]
7. Corbett, S.; Courtiol, A.; Lummaa, V.; Moorad, J.; Stearns, S. The transition to modernity and chronic disease: Mismatch and natural selection. *Nat. Rev. Genet.* **2018**, *19*, 419–430. [CrossRef]
8. Arcaya, M.C.; Arcaya, A.L.; Subramanian, S.V. Inequalities in health: Definitions, concepts, and theories. *Glob. Health Action* **2015**, *8*, 27106. [CrossRef]
9. Gomes, M.G.M.; Oliveira, J.F.; Bertolde, A.; Ayabina, D.; Nguyen, T.A.; Maciel, E.L.; Duarte, R.; Nguyen, B.H.; Shete, P.B.; Lienhardt, C. Introducing risk inequality metrics in tuberculosis policy development. *Nat. Commun.* **2019**, *10*, 2480. [CrossRef]

10. Tam, V.; Patel, N.; Turcotte, M.; Bossé, Y.; Paré, G.; Meyre, D. Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **2019**, *20*, 467–484. [CrossRef]
11. McGue, M. When assessing twin concordance, use the probandwise not the pairwise rate. *Schizophr. Bull.* **1992**, *18*, 171–176. [CrossRef]
12. Stigler, S.M. Memoir on the Probability of the Causes of Events, Pierre Simon Laplace. *Statist. Sci.* **1986**, *1*, 364–378. [CrossRef]
13. Patil, G.P.; Rao, C.R. Weighted distributions and size-based sampling with applications to wildlife populations and human families. *Biometrics* **1978**, *34*, 179–189. [CrossRef]
14. Mihael, P.; Jim, P.; Marc, Y. Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Relat. Fields* **1992**, *92*, 21–39.
15. Davidov, O.; Zelen, M. Referent sampling, family history and relative risk: The role of length-biased sampling. *Biostatistics* **2001**, *2*, 173–181. [CrossRef]
16. Pressé, S.; Ghosh, K.; Lee, J.; Dill, K.A. Principles of maximum entropy and maximum caliber in statistical physics. *Rev. Mod. Phys.* **2013**, *85*, 1115–1141. [CrossRef]
17. Cofré, R.; Herzog, R.; Corcoran, D.; Rosas, F.E. A Comparison of the Maximum Entropy Principle Across Biological Spatial Scales. *Entropy* **2019**, *21*, 1009. [CrossRef]
18. Dowson, D.C.; Wragg, A. Maximum-Entropy Distributions Having Prescribed First and Second Moments. *IEEE Trans. Inf. Theory* **1973**, *19*, 689–693. [CrossRef]
19. Verstockt, B.; Smith, K.G.C.; Lee, J.C. Genome-wide association studies in Crohn's disease: Past, present and future. *Clin. Transl. Immunol.* **2018**, *7*, e1001. [CrossRef]
20. Halfvarson, J. Genetics in twins with Crohn's disease: Less pronounced than previously believed? *Inflamm. Bowel Dis.* **2011**, *17*, 6–12. [CrossRef]
21. Stensrud, M.J.; Valberg, M. Inequality in genetic cancer risk suggests bad genes rather than bad luck. *Nat. Commun.* **2017**, *8*, 1165. [CrossRef] [PubMed]