Article

To be or to have Been Lucky, That is the Question

A. Lesage and J-M. Victor *

[1] Sorbonne Université, CNRS, Physique théorique de la matière condensée, LPTMC, F-75005 Paris, France.
[2] Sorbonne Université, CNRS, Physico-chimie des électrolytes et nano-systèmes interfaciaux, PHENIX, F-75005 Paris, France. antony.lesage@sorbonne-universite.fr
* 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 to become 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.

1. Introduction

"That evening he was lucky": what do we mean by this? And 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. But 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. The probability calculus started during summer 1654 with the correspondence between Pascal and Fermat precisely on elementary problems of gambling. 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 $1/2$. 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. At the time when he derived what is called since 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 [1], around the same time when he confirmed with Walther Gerlach the existence of the electron spin, for which he won the Nobel Prize in 1944. The agreement between theoretical and experimental distributions was surprisingly good.

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, then the mean frequency is used as a measure of the 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. But can we assess the *dispersion* of disease risks? And 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*. 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 done. 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 case of success or 0 in case 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 $S_i = \sum_{j=1}^{N} X_i^j$ the score over N draws. It is the number of times the individual i has won. 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 p_i are

$$E[S_i|p_i] = Np_i$$

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

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 $k=0,1,2,\dots,N$. 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 $k=0,1,2,\dots,N$). 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:

$$\forall k = 0, 1, \dots, N \;\; P_N[S = k] = E_p\big[P_N[S = k|p]\big] = E_p[C_N^k p^k (1-p)^{N-k}] = \int_0^1 dp f(p) C_N^k p^k (1-p)^{N-k} \tag{1}$$

where $E_p[\cdot]$ is expected value of $\cdot$ with the probability distribution of p, $f(p)$. The mean of S is

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

where $E_S[\cdot]$ is the expected value of $\cdot$ with the probability distribution of S, $P_N(\mathrm{S})$ and $E_S[S|p]$ 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 $f(p)$,

$$E_S[S] = NE_p[p] = N\Phi \tag{2}$$

and the variance of S is

$$Var(S) = E_S[S^2] - E_S[S]^2$$

where

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

hence

$$E_S[S^2] = N\big(E_p[p] - E_p[p^2]\big) + N^2 E_p[p^2]$$

and

$$Var(S) = N\big(E_p[p] - E_p[p^2]\big) + N^2 E_p[p^2] - N^2 E_p[p]^2$$

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

$$E_p[p] = \Phi$$

$$E_p[p^2] = \Sigma^2 + \Phi^2$$

so that

$$Var(S) = N\Phi(1-\Phi) + N(N-1)\Sigma^2 = N(\Phi(1-\Phi) - \Sigma^2) + N^2\Sigma^2 \tag{3}$$

Note that within the limit $N \to \infty$, the probability distribution function of the reduced variable $x = k/N$ (where $k = 0,\ 1,\ 2, \dots, N$) converges to the distribution $f(p)$.

Equation (3) shows that, if $N = 1$, the variance $Var(S) = \Phi(1-\Phi)$ *does not depend on the variance* Σ^2 of $f(p)$. 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 variance Σ^2 of $f(p)$. Moreover, according to equation (1), the histogram of gains itself depends only on the mean of the distribution $f(p)$:

$$P_1[S = 0] = E_p[1-p] = 1 - \Phi$$

$$P_1[S = 1] = E_p[p] = \Phi$$

The histogram of gains cannot therefore provide information on the dispersion of chances. For example, the two following distributions:

$$f_1(p) = \delta\left(p - \frac{1}{2}\right)$$

and

$$f_2(p) = \frac{1}{2}[\delta(p) + \delta(1-p)] \tag{4}$$

have the same mean $\Phi = 1/2$, hence result in the same histograms (Figure 1 for $N = 1$). However, the variance of f_1 is null whereas the variance of f_2 is $1/4$. (Note that $1/4$ is the maximal variance that a probability distribution $f(p)$ can take). 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?

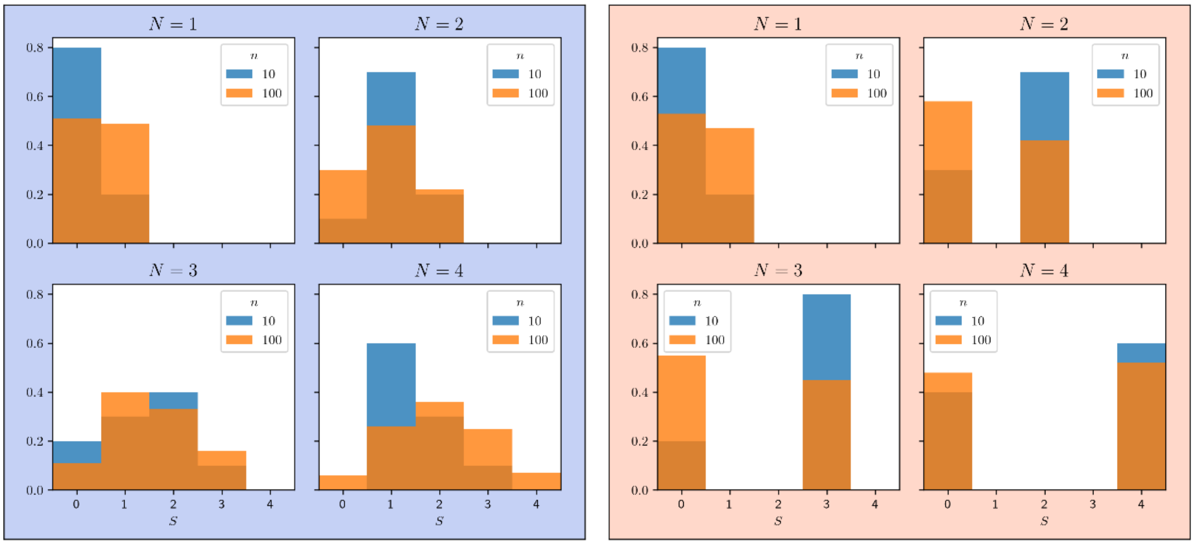

Figure 1. Histograms of gains S for two distributions f_1 and f_2: on the left-hand side $f_1(p) = \delta(p-1/2)$ and on the right-hand side $f_2(p) = \frac{1}{2}[\delta(p) + \delta(1-p)]$. In each case, the histogram of success is plotted for increasing values of the number N of draws ($N = 1{,}2{,}3{,}4$) and for two numbers n of gamblers: $n = 10$ (in blue) and 100 (in orange). For $N = 1$ note that the histogram for f_1 is similar to the histogram for f_2 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_1 and f_2 diverge as n increases.

3. A statistical test of the dispersion of chances.

We then note on 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:

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

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

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

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:

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

ii. 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 variance 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 $N(\Phi(1-\Phi)-\Sigma^2)+N^2\Sigma^2$. Hence if *the variance* $Var(S)$ grows *linearly* with N, then all individuals have the same probability p of success. If on the contrary $Var(S)$ grows *quadratically* with N then not all individuals have the same chance of success. We can therefore rephrase our hypothesis test in the following alternative based on the dependence of the variance $Var(S)$ on the number N of draws:

i. Null hypothesis H_0: *the variance* $Var(S)$ grows *linearly* with N;
ii. Alternative hypothesis H_1: *the variance* $Var(S)$ grows *quadratically* with N.

Figure 2 plots the variance of the two distributions f_1 and f_2 as a function of the number N of draws for $n = 100$ gamblers.

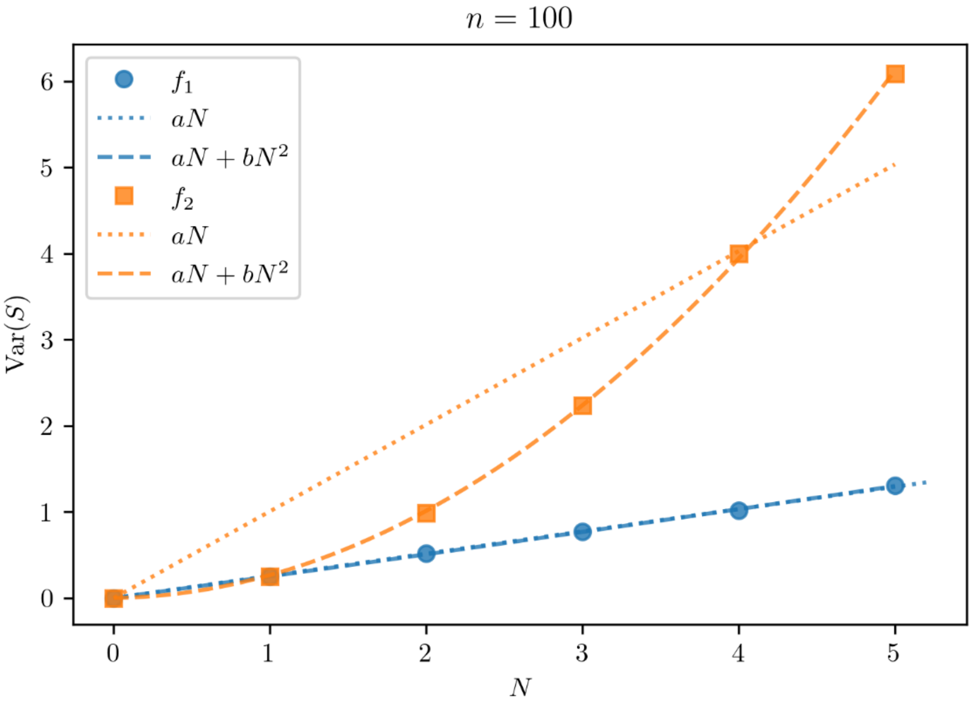

Figure 2. Linear regression fits $Var(S)$ for f_1, with $a = 0.251 \pm 0.005$ in agreement with equation (3) when $\Sigma^2 = 0$. Moreover, a agrees with the expected value $\Phi(1-\Phi) = 1/4$. At odds with f_1, the linear regression does not fit $Var(S)$ for f_2 whereas the quadratic fit is excellent, with: $a = 0.244 \pm 0.006$ and $b = 0.01 \pm 0.01$. Here a agrees with the expected value $\Sigma^2 = 1/4$ and b 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 more 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 $N = 2$:

i. Null hypothesis H_0: the variance of S reads $Var_0[S] = 2\Phi(1-\Phi)$, i.e. $\Sigma^2 = 0$;

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

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

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

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

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

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:

i. 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;

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

The variance of V_n is (see Appendix 1)

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

where $\Psi = \Phi(1-\Phi)$.

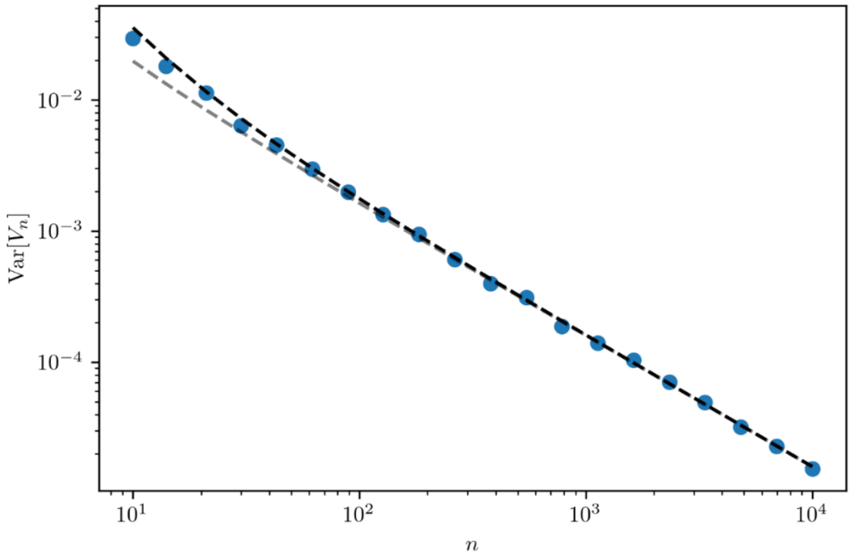

Figure 3. Evolution of the variance V_n of S for $N = 2$ as a function of the number n of players. The blue dots are simulated with $\Phi = 0.5$ and $\Sigma^2 = 0.15$. The black dashed line corresponds to the expected variance according to the equation (7). The grey dashed line corresponds to the leading-order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

It can be noted that the distribution of V_n tends towards a normal distribution $\mathcal{N}(E[V_n], Var[V_n])$ of mean $E[V_n] = Var[S]$ and variance $Var[V_n]$. Now we wish to estimate the probability of having obtained a value as high as V_n under the null hypothesis H_0, i.e. the p-value. Since V_n follows a normal distribution, the p-value can be expressed as follows

$$\text{p-value} = \frac{1}{2}\left(1 - \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right)\right) = \frac{1}{2}\operatorname{erfc}\left(\frac{z}{\sqrt{2}}\right) \tag{8}$$

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 $\Psi = \Phi(1-\Phi)$:

$$z \sim \frac{\sqrt{2n}\,\Sigma^2}{\sqrt{\Psi(1-2\Psi)}}$$

4. Dispersion of disease risks for twins.

Inequality in disease risk is a major public health issue. Of course, part of this inequality is known to depend on genetic and environmental factors. The mean frequency that an individual will become ill in a given population, specified by genetic and environmental factors, can be measured. And, as usual, this frequency can be used as a measure of the probability to become ill. But can we assess the dispersion of disease risk, if only it exists, in this specific population? And 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 twins, have the same genome, shared the same fetal environment and generally share the same living conditions. So that they are most likely to share also the same probability to become 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 to have 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 to have 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 Σ^2 of S are given by equations (2) and (3) respectively, hence for $N = 2$

$$E[S] = 2\Phi \tag{9}$$

$$Var(S) = 2\Phi(1-\Phi) + 2\Sigma^2$$

Then if V_n is significantly greater than $\bar{S}(1 - \bar{S}/2)$, which amounts to carry 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. But 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:

$$\tau = P(X_2 = 1|X_1 = 1) = \frac{P(X_1 = 1, X_2 = 1)}{P(X_1 = 1)} = \frac{P(S = 2)}{P(X_1 = 1, X_2 = 1) + P(X_1 = 1, X_2 = 0)}$$

hence

$$\tau = \frac{P(S=2)}{P(S=2) + \frac{1}{2}P(S=1)}$$

Note that τ is equal to the probandwise concordance rate, which is known to best assess the twin concordance rate [2].

Using equations (2) and (6), we can also reformulate the concordance rate of twins in terms of the moments of the distribution $f(p)$:

$$\tau = \frac{2P_2[S=2]}{E_S[S]} = \frac{E_p[p^2]}{E_p[p]}$$

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

$$\tau_N = \frac{NP_N[S=N]}{E_S[S]} = \frac{E_p[p^N]}{E_p[p]}$$

Using equations (6) and (9) we get

$$\tau = \frac{\Phi^2 + \Sigma^2}{\Phi} \tag{10}$$

so that the relative risk $RR = \tau/\Phi$ is equal to

$$RR = \frac{\Phi^2 + \Sigma^2}{\Phi^2} = 1 + \frac{\Sigma^2}{\Phi^2} \tag{11}$$

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 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

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

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

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

Note that $f_a(p)$ is the so-called "size-biased law" of the risk p to become ill. Size-biased laws are found in many contexts, notably rare events [3], Poisson point processes [4] or familial risk of disease [5].

The mean risk in the affected population is then

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

where $E_a[\cdot]$ is the expected value of $\cdot$ among affected people, with the probability distribution $f_a(p)$. Using again equations (6) we get

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

which proves that the twin concordance rate (10) is equal to the mean risk in the affected population (11)

$$\tau = E_a[p] \tag{13}$$

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 standard deviation Σ of $f(p)$. The principle of maximum entropy then provides us with the least arbitrary distribution [6]. Dowson and Wragg proved [7] 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

$$H(f) = -\int_0^1 f(p) \ln f(p)\, dp \tag{14}$$

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 (concave) or U-type (convex). 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 2). 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

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

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

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

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

$$f(p; m, s) = \frac{1}{\sqrt{2}s} \frac{\exp\left(\frac{p^2 - 2mp}{2s^2}\right)}{D\left(\frac{m}{s\sqrt{2}}\right) + e^{\frac{1-2m}{2s^2}} D\left(\frac{1-m}{s\sqrt{2}}\right)} \tag{15}$$

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

$$\Phi = m - \frac{\sqrt{2}}{2} s \frac{1 - e^{\frac{1-2m}{2s^2}}}{D\left(\frac{m}{s\sqrt{2}}\right) + e^{\frac{1-2m}{2s^2}} D\left(\frac{1-m}{s\sqrt{2}}\right)} \tag{16}$$

$$\Sigma^2 = -s^2 \left\{ 1 - \frac{1}{s}\frac{\sqrt{2}}{2} \frac{m + (1-m)e^{\frac{1-2m}{2s^2}}}{D\left(\frac{m}{s\sqrt{2}}\right) + e^{\frac{1-2m}{2s^2}} D\left(\frac{1-m}{s\sqrt{2}}\right)} + \frac{1}{2} \frac{\left(1 - e^{\frac{1-2m}{2s^2}}\right)^2}{\left[D\left(\frac{m}{s\sqrt{2}}\right) + e^{\frac{1-2m}{2s^2}} D\left(\frac{1-m}{s\sqrt{2}}\right)\right]^2} \right\} \tag{17}$$

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

5. Application to Crohn disease (CD).

Crohn disease (CD) is one of the most well documented chronic disease (ref). Its prevalence Φ and twin concordance rate τ are [8]:

$$\Phi \cong 0.0025$$

$$\tau = 0.385$$

Then the twin relative risk is

$$RR \cong 154 \tag{18}$$

hence

$$\Sigma^2 = \Phi^2(RR - 1) \cong 0.00096 \tag{19}$$

$$\Sigma \cong 0.031$$

which means that

$$\frac{\Sigma}{\Phi} \cong 12 \tag{20}$$

The dispersion of the risk to be affected is therefore huge for CD. It is also true for most chronic diseases (unpublished results).

It is now necessary to calculate the p-value according to equation (8) 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 [8]. 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 [8]. 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 get the following sample size

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

Equation (8) 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 standard deviation Σ. Solving the system of equations (16-17) 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 4. 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!

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 to be 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% [6]. 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.

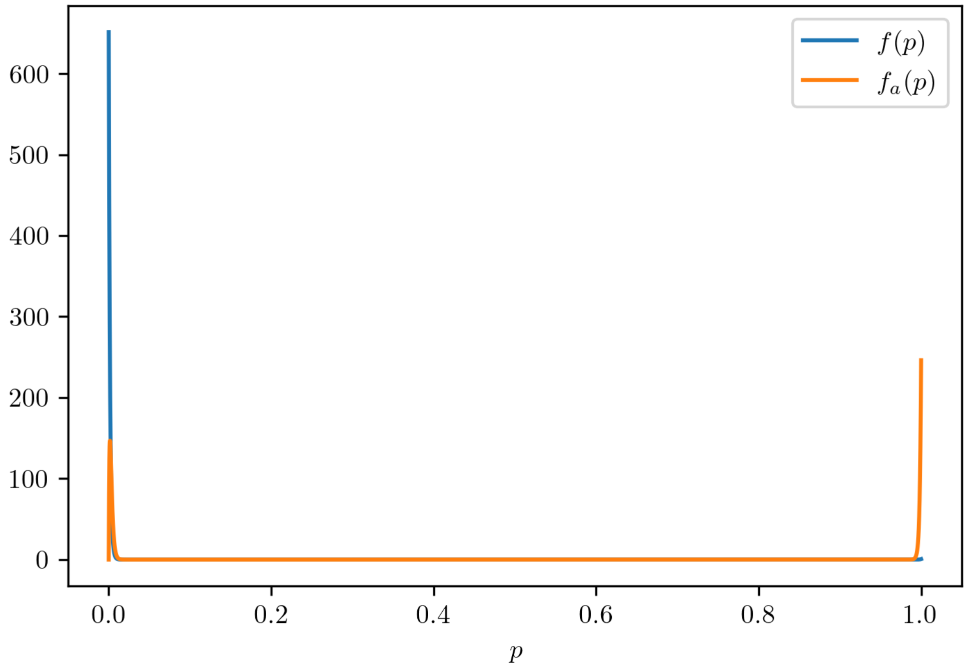

Figure 4. CD risk distribution function $f(p;m,s)$ among the population (in blue) is narrow peaked at $p = 0$. The risk distribution function $f_a = pf(p;m,s)/\Phi$ among affected people (in orange) has two narrow peaks, one close to 0, the other one close to 1. The look of both peaks is given in Figure 5.

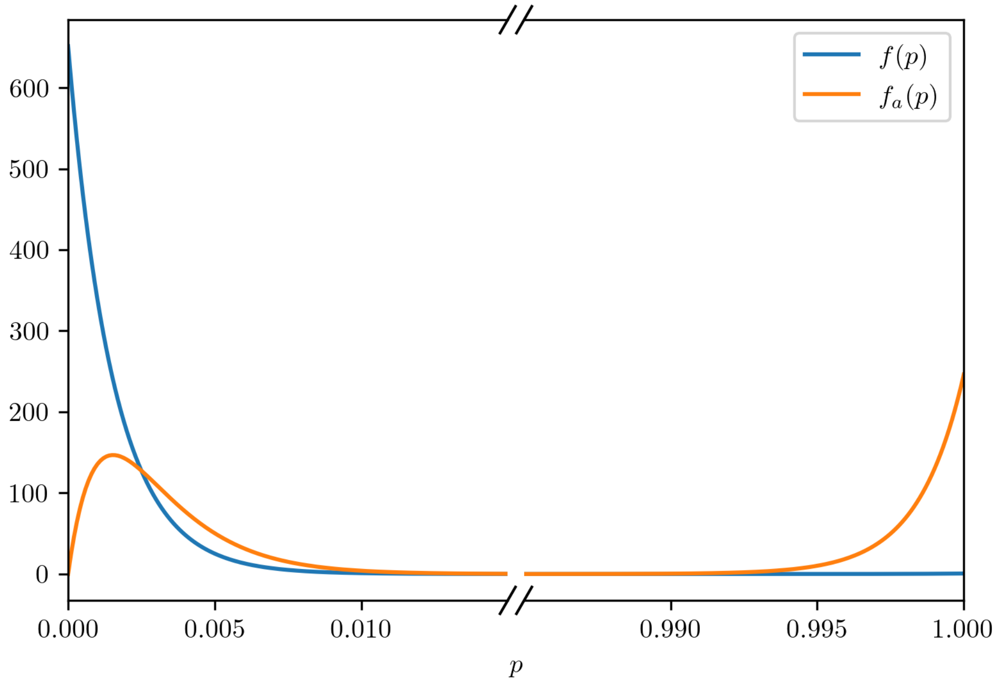

Figure 5. In blue: look of $f(p; m, s)$ in the vicinity of 0. In orange: look of both peaks of $f_a(p)$ in 0 and 1 respectively. Concordant twins (almost) all belong to the right peak (at $p = 1$) whereas discordant twins (almost) all belong to the left peak (at $p = 0$).

6. Conclusion.

Assessing 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 be actually assessed whenever multiple draws are available. For this purpose, we have proposed a hypothesis test to evidence 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 to become ill is usually a U-type truncated normal distribution. We have computed the parameters of this U-type distribution in the case of Crohn disease using the prevalence and the twin concordance rate of this pathology. We have moreover 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.

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.

Since its invention in the middle of the 17th century, the 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 [9], not to mention the countless applications to finance and economy.

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

Funding

This research received no external funding.

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 1. Computing the variance of V_n.

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

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

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

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

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

$$E[\bar{S}] = E[S]$$

$$Var(\bar{S}) = \frac{Var(S)}{n}$$

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

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

hence

$$Var[V_n] = \frac{n}{(n-1)^2}Var[(S-m)^2] + \frac{n^2}{(n-1)^2}Var[(\bar{S}-m)^2]$$

$Var[(S-m)^2]$ and $Var[(\bar{S}-m)^2]$ remain to be determined. Let us start with the latter, which is simpler.

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

with

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

and

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

Now if we consider samples of size $n \geq 30$, according to the central limit theorem, the distribution of $\bar{S}$ tends towards the normal distribution $\mathcal{N}(E[\bar{S}], Var[\bar{S}])$ of mean $E[\bar{S}] = m$ and variance $Var[\bar{S}] = Var[S]/n$. The moments of $\bar{S}$ are written then

$$E[\bar{S}^2] = m^2 + Var(\bar{S})$$

$$E[\bar{S}^3] = m\big(m^2 + 3\,Var(\bar{S})\big)$$

$$E[\bar{S}^4] = m^4 + 6m^2\,Var(\bar{S}) + 3\,Var(\bar{S})^2$$

All the terms in m cancel each other out, hence

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

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

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

with

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

and

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

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

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

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

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

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

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

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

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.

$$E[S^2] = 2E[p] + 2E[p^2]$$

$$E[S^3] = 2E[p] + 6E[p^2]$$

$$E[S^4] = 2E[p] + 14E[p^2]$$

Hence,

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

and

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

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

$$Var[(\bar{S}-m)^2] = \frac{8}{n^2}(\Phi(1-\Phi)+\Sigma^2)^2$$

$$Var[(S-m)^2] = 2\Phi(1-\Phi)(1-2\Phi+2\Phi^2) + 14(2\Phi-1)^2\Sigma^2 - 4\Sigma^4$$

Then, we obtain the following expression

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

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

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

Appendix 2. 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. So, according to equation (11), Σ/Φ is of the order of 10. As an example, $RR \cong 12$ for Crohn disease (see equation 20). Therefore $\Phi \ll 1$ and $\Sigma > \Phi$ is the rule for chronic diseases.

Dowson and Wragg [7] show that the truncated normal distribution $f(p)$ that maximizes the entropy $H(f)$ (see equation (14)) 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 OMA (See Figure 1 and text below in [7]). 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 OMA 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 OMA, the first and second moments of $f_{exp}(p)$ are given by

$$m_1 = \frac{1}{\lambda} - \frac{1}{e^{\lambda}-1} \qquad (A1)$$

$$m_2 = \frac{2}{\lambda^2} - \left(1+\frac{2}{\lambda}\right)\frac{1}{e^{\lambda}-1} \qquad (A2)$$

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 OMA, 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] O. Stern (1920) Zeits. f. Physik 2, 49.

[2] M. McGue (1992) When assessing twin concordance, use the probandwise not the pairwise rate. Schizophrenia Bulletin 1992, 18, 171-176.

[3] Patil GP, Rao CR. (1978) Weighted distributions and size-based sampling with applications to wildlife populations and human families. Biometrics 34, 179–189.

[4] Perman, Mihael, Pitman, Jim, and Yor, Marc (1992) Size-biased sampling of Poisson point processes and excursions. Probability Theory and Related Fields 92, 21-39.

[5] Davidov O, Zelen M. (2001) Referent sampling, family history and relative risk: the role of length-biased sampling. Biostatistics 2, 173–181.

[6] S. Pressé et al (2013) Principles of maximum entropy and maximum caliber in statistical physics. Reviews of Modern Physics 85(3) 1115-1141.

[7] D.C. Dowson and A. Wragg (1973) Maximum-Entropy Distributions Having Prescribed First and Second Moments. IEEE Transactions on Information Theory Volume 19, Issue 5, 689 – 693.

[8] Halfvarson, J. (2011). Genetics in twins with Crohn’s disease: Less pronounced than previously believed?: Inflammatory Bowel Diseases 17, 6–12.

[9] P. D. Bruza, Z. Wang, J. R. Busemeyer (2015) Quantum cognition: a new theoretical approach to psychology. Trends in Cognitive Sciences July 2015, Vol. 19, No. 7.