Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2020 doi:10.20944/preprints202012.0425.v1

To be or to have been lucky, that is the question

A. Lesage[1] and J-M. Victor[2]

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.

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 can't 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)? We propose here some tools to tackle this question and we explore their relevance to and consequences in the field of chronic diseases.

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 loose). In an unbiased game, all the p_i are identical and equal to $\frac{1}{2}$. Imagine that some gamblers are luckier, others less fortunate, hence some p_i are $> \frac{1}{2}$, others are $< \frac{1}{2}$. This means that the p_i are random variables that are drawn from a probability distribution $f(p)$ that is different from $\delta\left(p - \frac{1}{2}\right)$. 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. 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 $S_i = \sum_{j=1}^{N} X_i^j$. S_i is a random variable that follows a binomial distribution $B(N, p_i)$. S_i is the number of times the individual i has won. The mean of S_i and its variance are

$$< S_i >= Np_i \quad (1)$$

$$var(S_i) = Np_i(1 - p_i) \quad (2)$$

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.

[1] Sorbonne Université, CNRS, Physico-chimie des électrolytes et nano-systèmes interfaciaux, PHENIX, F-75005 Paris, France.
[2] Sorbonne Université, CNRS, Physique théorique de la matière condensée, LPTMC, F-75005 Paris, France.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2020 doi:10.20944/preprints202012.0425.v1

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 probability distribution function of S is given as follows:

$$\forall k = 0, 1, \dots N \quad Prob(S = k) = \int_0^1 dp f(p)\, C_N^k p^k (1-p)^{N-k} \tag{3}$$

The mean of S is

$$\langle S \rangle = \sum_{k=0}^{N} k Prob(S = k) = \int_0^1 dp f(p) \sum_{k=0}^{N} k C_N^k p^k (1-p)^{N-k} = \int_0^1 dp f(p)\, Np = N\langle p \rangle$$

hence

$$\langle S \rangle = N\Phi \tag{4}$$

and its variance is

$$var(S) = \langle S^2 \rangle - \langle S \rangle^2$$

where

$$\langle S^2 \rangle = \sum_{k=0}^{N} k^2 Prob(S = k) = \int_0^1 dp f(p) \sum_{k=0}^{N} k^2 C_N^k p^k (1-p)^{N-k}$$

$$= \int_0^1 dp f(p)\, [Np(1-p) + (Np)^2]$$

hence

$$\langle S^2 \rangle = N(\langle p \rangle - \langle p^2 \rangle) + N^2 \langle p^2 \rangle$$

and

$$var(S) = N(\langle p \rangle - \langle p^2 \rangle) + N^2 \langle p^2 \rangle - N^2 \langle p \rangle^2$$

Now

$$\langle p \rangle = \Phi$$

and

$$\langle p^2 \rangle = \Sigma^2 + \Phi^2$$

so that

$$var(S) = N(\Phi(1-\Phi) - \Sigma^2) + N^2\Sigma^2 \tag{5}$$

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

Equation (5) 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 (3), the histogram of gains itself depends only on the mean of the distribution $f(p)$:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2020 doi:10.20944/preprints202012.0425.v1

$$Prob(S=0)=\int_0^1 dpf(p)(1-p)=1-\Phi \tag{6}$$

$$Prob(S=1)=\int_0^1 dpf(p)p=\Phi \tag{7}$$

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

$$f_1(p)=\delta\left(p-\frac{1}{2}\right) \tag{8}$$

and

$$f_2(p)=\frac{1}{2}[\delta(p)+\delta(p-1)] \tag{9}$$

have the same mean $\Phi=\frac{1}{2}$, hence result in the same histograms. However the variance of f_1 is null whereas the variance of f_2 is $\frac{1}{4}$. (Note that $\frac{1}{4}$ 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?

A statistical test of the dispersion of chances.

We first note that the histogram of gains for two draws has three bins, one at 0, the second at 1 and the third at 2, with the following values:

$$Prob(S=0)=\int_0^1 dpf(p)(1-p)^2=\langle(1-p)^2\rangle=(1-\Phi)^2+\Sigma^2 \tag{10}$$

$$Prob(S=1)=2\int_0^1 dpf(p)p(1-p)=2\Phi(1-\Phi)-2\Sigma^2 \tag{11}$$

$$Prob(S=2)=\int_0^1 dpf(p)p^2=\Phi^2+\Sigma^2 \tag{12}$$

Hence the histogram of gains now depends on (and only on) both the mean and the variance of $f(p)$. 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 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 $\langle f\rangle=\Phi$ and variance $\Sigma^2=0$;

(ii) Alternative hypothesis H_1: f has the same mean Φ but some people are luckier than the others so that 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2020 doi:10.20944/preprints202012.0425.v1

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.

Then the Ramsey Regression Equation Specification Error Test (RESET) is a relevant tool to conclude. To be more specific, when $N \geq 2$, one has to calculate the F statistic given by

$$F = \frac{\left(RSS_L - RSS_Q\right)}{\left(\frac{RSS_Q}{N-1}\right)} \tag{13}$$

where RSS_L (resp. RSS_Q) is the residual sum of squares of the linear (resp. quadratic) regression. Under the null hypothesis H_0, the F statistic has an F-distribution with $(1, N-1)$ degrees of freedom. H_0 is rejected if the value of F calculated from the data is greater than the critical value of the F-distribution for some fixed false-rejection probability (usually 0,01).

Dispersion of disease risks for twins.

Dispersion of chances is far from being limited to gamblers. Disease risks are another area where people may be and actually are unequal. Here a providential help comes from the existence of twins. Identical twins, also called monozygotic twins, have the same genome, shared the same fetal environment and generally the same living conditions, so that they are most likely to share the same probability to become ill, whatever the disease. Let us now calculate the concordance rate of some disease D for monozygotic twins. This is much related to the gambling question addressed above. Indeed, as both twins have the same probability p to have disease D, the status - ill or healthy - of each of the two twins is equivalent 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 A and B be the two monozygotic twins. We note A_d (resp. B_d) the event "A has disease D" (resp. "B has disease D") and let $f(p)$ be the probability distribution function of the risk to have disease D in the overall population. We define the random variable S as above, i.e. $X = 0$ if both twins are healthy, $X = 1$ if only one of the two twins is ill and $X = 2$ if both twins are ill. The mean Φ and variance Σ^2 of S are given by equations (4) and (5) respectively, hence for $N = 2$

$$\langle S \rangle = 2\Phi \tag{14}$$

$$var(S) = 2(\Phi(1-\Phi) - \Sigma^2) + 4\Sigma^2 = 2\Phi(1-\Phi) + 2\Sigma^2 \tag{15}$$

Then if $var(S)$ is significantly greater than $\langle S \rangle \left(1 - \frac{\langle S \rangle}{2}\right)$, which amounts to carry out the hypothesis test presented in the above section, we can conclude that there is some dispersion of the risk of disease. As we will see below the dispersion is in fact unusually large. But before that, let us calculate the concordance rate of disease D for monozygotic twins. In genetics, concordance is the probability that a pair of individuals will both have a certain characteristic, given that one of the pair has the characteristic. Concordance rate in a population is best

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2020 doi:10.20944/preprints202012.0425.v1

assessed by the probandwise rate [1], namely the conditional probability of B to have disease D, knowing that A has disease D, i.e. $P(B_d|A_d)$.
Of course, for two unrelated persons

$$P(B_d|A_d) = \frac{P(A_d \text{ and } B_d)}{P(A_d)} = \frac{P(A_d)P(B_d)}{P(A_d)} = P(B_d) = p \tag{16}$$

and then

$$\langle P(B_d|A_d)\rangle = \langle p\rangle = \Phi \tag{17}$$

Now for identical twins, equation (16) is still true, then a naive averaging of both sides of equation (16) would necessary lead to equation (17). What is wrong ? As a matter of fact, the conditioning on A_d is not insignificant: in formula (17) it is indeed necessary to average p by taking into account the fact that A_d is realized, thus by using the probability density function $f(p|A_d)$ and not $f(p)$:

$$\langle P(B_d|A_d)\rangle = \int_0^1 pf(p|A_d)dp \tag{18}$$

The probability density function $f(p|A_d)$ can be obtained according to Bayes formula

$$f(p|A_d)\,P(A_d) = f(p)P(A_d|p) \tag{19}$$

Now

$$P(A_d|p) = p$$

and

$$P(A_d) = \langle p\rangle$$

so that

$$f(p|A_d) = \frac{pf(p)}{\langle p\rangle} \tag{20}$$

then

$$\langle P(B_d|A_d)\rangle = \int_0^1 pf(p|A_d)dp = \int_0^1 p\frac{pf(p)}{\langle p\rangle}dp = \frac{1}{\langle p\rangle}\int_0^1 p^2 f(p)dp = \frac{\langle p^2\rangle}{\langle p\rangle} \tag{21}$$

Another way to understand the difference between equation (21) fo twins and equation (17) for unrelated persons is to note that, for unrelated persons we average over the whole population whereas for identical twins we have to average over *affected* people only. And $f(p|A_d)$ is nothing but the probability distribution of the risk to have disease D in the population of *affected* people.
We finally get the relative risk

$$RR = \frac{\langle P(B_d|A_d)\rangle}{\langle P(B_d)\rangle} = \frac{\langle p^2\rangle}{\langle p\rangle^2} \tag{22}$$

which can be expressed as a function of the mean Φ and variance σ^2 of $f(p)$:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2020 doi:10.20944/preprints202012.0425.v1

$$RR = \frac{\langle p^2 \rangle}{\langle p \rangle^2} = \frac{\Sigma^2 + \Phi^2}{\Phi^2} = 1 + \frac{\Sigma^2}{\Phi^2} \tag{23}$$

For Crohn disease:

$$\Phi \cong 0.0025 \tag{24}$$

and

$$RR \cong 100 \tag{25}$$

hence

$$\Sigma^2 = \Phi^2(RR - 1) \cong 0.00062 \tag{26}$$

$$\Sigma \cong 0.025 \tag{27}$$

Which means that

$$\Sigma \cong 10\Phi \tag{28}$$

For Crohn disease, but this is also true for most chronic diseases, the dispersion of the risk to be affected is therefore huge indeed. The distribution $f(p)$ may then be approximated by a beta-distribution:

$$f(p; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \tag{29}$$

The parameters α and β can be expressed as a function of Φ and Σ^2:

$$\alpha = \Phi\left[\frac{\Phi(1-\Phi)}{\Sigma^2} - 1\right] \tag{30}$$

$$\beta = (1-\Phi)\left[\frac{\Phi(1-\Phi)}{\Sigma^2} - 1\right] \tag{31}$$

For Crohn disease we get:

$$\alpha \cong 0.0076 \tag{32}$$

$$\beta \cong 3 \tag{33}$$

Therefore the mean risk in the affected population is

$$\Phi_a = \frac{\int_0^1 p[pf(p)]dp}{\int_0^1 [pf(p)]dp} = \frac{\langle p^2 \rangle}{\langle p \rangle} = \frac{\Sigma^2 + \Phi^2}{\Phi} \tag{34}$$

As $\Sigma \cong 10\Phi$ the mean risk in the affected population is

$$\Phi_a \cong 100\Phi \tag{35}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2020 doi:10.20944/preprints202012.0425.v1

and the relative risk of affected people as compared to controls is

$$\frac{\Phi_a}{\Phi} \cong 100 \tag{36}$$

which means that affected people are much more predisposed to the disease than controls. Note that $\frac{\Phi_a}{\Phi}$ is equal to the relative risk RR computed above for twins.
At this stage we remark that affected people did not really have bad luck to become ill but actually had a large predisposition to become ill!

A modular approach of chronic diseases.
Suppose now that a disease D is the result of the impairment of K redundant functions, so that people become ill when and only when the whole set of K functions is impaired [2]. In other words, as long as at least one function is performed, there is no disease. This kind of modeling has been developed recently to fit the incidence of various chronic diseases as a function of the age of onset [3]. We assume that these redundant functions are independent, i.e. the network of functions is organized in a modular way. To be more specific, each function is achieved as a module of the global physiological network (including most notably the regulatory gene network, the metabolic network and the cell signaling network) and modules operate in an independent way. The probability p to become ill is thus the product of the K probabilities p_i that each redundant function i is impaired. Importantly, an impaired function of module i means that this module is *permissive for* disease D, but this does not mean that module i is not functioning at all. On the contrary, we suggest that module i can then be *protective against* another disease.
For the sake of simplicity, we assume that all the K functions have the same probability distribution function $g(p)$. Now we can relate the mean Φ and variance Σ^2 of $f(p)$ to the mean φ and variance σ^2 of $g(p)$:

$$\Phi = \varphi^K \tag{37}$$

$$\Sigma^2 = (\varphi^2 + \sigma^2)^K - \varphi^{2K} \tag{38}$$

It has been shown in [3] that the number of redundant functions is $K \cong 10$ for various « not so rare » chronic diseases. By this we mean chronic diseases that have a typical prevalence, i.e. the fraction of the population that is affected by the disease, between 0.5 per 1000 and 5 per 1000. We recall indeed that in Europe, a disease is considered to be rare when it affects less than 1 person per 2000. Hence the mean risk for « not so rare » chronic diseases is $\Phi \cong 10^{-3}$. Equations (37) and (38) then give the mean φ and variance σ^2 of the risk distribution function $g(p)$ for one module:

$$\log(\varphi) = \frac{\log(\Phi)}{K} \cong -0.3 \tag{39}$$

hence

$$\varphi \cong \frac{1}{2} \tag{40}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 December 2020 doi:10.20944/preprints202012.0425.v1

and according to equation (28)

$$100\varphi^{2K} \cong (\varphi^2 + \sigma^2)^K - \varphi^{2K} \quad (41)$$

hence

$$\sigma^2 \cong \left(100^{\frac{1}{K}} - 1\right)\varphi^2 \quad (42)$$

As $K \cong 10$ we finally get

$$\sigma \cong 0.4 \quad (43)$$

Remember that σ cannot exceed 0.5, which is the maximum standard deviation that a probability distribution function $f(p)$ can take: in this extreme case, $f(p)$ is equal to $f_2(p)$ as given above in equation (9), which is the sum of two symmetric Dirac delta functions, one in 0 (protective state against disease D), the other one in 1 (permissive state for disease D). We recall that a protective state against some disease D may be permissive for another disease D' and conversely a permissive state for disease D may be protective against D'. Equation (43) shows that the risk distribution function $g(p)$ for one module has a very wide dispersion, yet it is not reduced to two deterministic behaviors as is $f_2(p)$. It is neither reduced to one completely random behavior as is $f_1(p)$ which mimics the *bet-hedging* strategy of some plants in alternate environments [4]. We speculate that evolution may have shaped the modules of the physiological network to make them function in a rather deterministic way in stable environments but to be versatile enough so as to allow individuals and their progeny to adapt to rapidly changing environments.

Conclusion.

The calculation of probabilities started during summer 1654 with the correspondence between Pascal and Fermat on elementary problems of gambling. Since then, it has become one of the most abundant fields of mathematics, accompanying most if not all new fields of science, from quantum physics to recent developments in quantum cognition [5], not to mention the countless applications to finance and economy. On this last point, we postpone the in-depth study of the measure of unequal opportunities to a further work.

Pascal could never complete his treatise to which he wanted to give the « astonishing » title "Geometry of Chance". This never-ending treatise is still being written, as evidenced in this special issue.

References

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

[2] G. Debret, C. Jung, J-P. Hugot et al (2011) Genetic Susceptibility to a Complex Disease: The Key Role of Functional Redundancy. History and Philosophy of the Life Sciences Vol. 33, issue 4, 497-514.

[3] Victor J-M, Debret G, Lesne A, Pascoe L, Carrivain P, Wainrib G, et al. (2016) Network modeling of Crohn's disease incidence. PLoS ONE 11(6): e0156138.

 doi:10.20944/preprints202012.0425.v1

[4] D. Cohen (1966) Optimizing reproduction in a randomly varying environment. Journal of Theoretical Biology. 12 (1): 119–129.

[5] 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.