Universal statistics of selected values

MATTEO SMERLAK\textsuperscript{1} and AHMED YOUSSEF\textsuperscript{2}

\textsuperscript{1} Perimeter Institute for Theoretical Physics - 31 Caroline St. N., Waterloo ON N2L 2Y5, Canada
\textsuperscript{2} LD - Research - Pappelallee 78/79, 10437 Berlin, Germany

received 4 January 2017; accepted in final form 7 April 2017
published online 2 May 2017

PACS 02.50.-r – Probability theory, stochastic processes, and statistics
PACS 87.10.-e – General theory and mathematical aspects
PACS 87.23.Kg – Dynamics of evolution

Abstract – Selection, the tendency of some traits to become more frequent than others under the influence of some (natural or artificial) agency, is a key component of Darwinian evolution and countless other natural and social phenomena. Yet a general theory of selection, analogous to the Fisher-Tippett-Gnedenko theory of extreme events, is lacking. Here we introduce a probabilistic definition of selection and show that selected values are attracted to a universal family of limiting distributions which generalize the log-normal distribution. The universality classes and scaling exponents are determined by the tail thickness of the random variable under selection. Our results provide a possible explanation for skewed distributions observed in diverse contexts where selection plays a key role, from molecular biology to agriculture and sport.

Copyright © EPLA, 2017

Introduction. – In a posthumous manuscript [1], the population geneticist Price noted that “selection has been studied mainly in genetics, but of course there is much more to selection than just genetical selection”. He gave examples of selection processes relevant to psychology, chemistry, archeology, linguistics, history, economics and epistemology, and remarked that “despite the pervading importance of selection in science and life, there has been no abstraction and generalization from genetical selection to obtain a general selection theory.”

Price stressed two key features of the theory to be developed. First, selection should be studied as a mathematical transformation, irrespectively of the (natural or artificial) agency responsible for that transformation. Second, selection theory should encompass both “subset selection”, wherein a subset is picked out from a set according to some criterion, and “Darwinian selection”, i.e., dominance through differential reproduction. If such a general concept could be formulated mathematically, he thought, it would have an impact comparable to Shannon’s formal theory of communication [2].

Whether or not the analogy is apt, there is a clear need for a general theory of selection. In biology, identifying signatures of natural selection (in particular at the genotypic level [3]) is a fundamental problem with important applications, for instance in the context of cancer research [4]. A general theory would also be useful for the development of selection-based search methods throughout the sciences, including genetic algorithms [5] in computer science or SELEX protocols [6] in pharmacology. It would also provide a conceptual framework for the current widespread interest in analytics in sport, education, academia and other competitive fields where selection plays a key role. In spite of a handful of formal explorations [7–9] —and forty-five years after Price’s comments— selection theory is still “a theory waiting to be born” [1]. (Somewhat paradoxically, sophisticated evolutionary models involving selection and mutations, drift, gene flow, etc. are well developed [10–12].)

The fundamental question selection theory should address was clearly articulated in a recent paper by Boyer et al. on molecular evolution [13]. The authors considered large libraries of randomized biomolecules which they selected on the basis of their affinity for a molecular target of interest. As they noted, “merely counting the number of different individuals provides a poor indication of the potential of a population to satisfy a new selective constraint”. The key problem, then, is how to identify the features of the population which characterize its selective potential. How diverse should it be? How should we measure this “diversity”?

In this paper we explore some of the most basic statistical aspects of the selection process. To this effect we define a selected value as the transformation $S_t W$ of a
non-negative random variable $W$ given by
\[ \mathbb{P}(S_t W = w) \propto w^t \mathbb{P}(W = w), \] for some parameter $t > 0$. This definition is in the spirit of the one proposed by Price [14]: “selection on a set in relation to property $W$ is the act or process of producing a corresponding set in a way such that the amounts of each entity are non-randomly related to the corresponding $W$ values”\(^1\). Furthermore, it carries a natural semi-group structure $(S_t \circ S_t = S_{t+t'})$ from which notions of “weak selection” ($t \to 0$) and “strong selection” ($t \to \infty$) can be defined. Transformation (1) also has an intuitive Darwinian interpretation: if $W$ represents the number of viable offspring of an organism in a heterogenous population (its evolutionary “fitness”), then $S_t W$ describes the change in the distribution of fitness after $t$ generations. Note, however, that (1) is equally consistent with subset selection: the variable $S_t W$ may represent a subset of a population biased towards larger values of $W$, in such a way that an entity with $W = 2w_0$ is $2^t$ more likely to be picked than an entity with $W = w_0$.

**Fisher’s fundamental theorem.** – The best known result concerning the relation between the selective potential of a population and its diversity is Fisher’s “fundamental theorem of natural selection” [15]. In the language of evolutionary theory, Fisher’s theorem states that the rate of growth of a population mean fitness under selection is proportional to its variance in fitness. In our notations this reads
\[ \langle S_{t+1} W \rangle = \langle S_t W \rangle + \text{var}(S_t W) \frac{\langle S_t W \rangle}{\langle S_t W \rangle}, \] (2)
This identity —an easy consequence of (1)— captures a basic aspect of selection dynamics: the larger the variation in fitness at a given time is, the faster evolution proceeds, or “variation is the fuel of evolution” as the catchphrase goes. In the limit where all lineages have the same fitness, $\text{var}(S_t W) = 0$, the mean fitness stops growing and evolution comes to a halt. Fisher was impressed by the generality of eq. (2) and compared it to the second law of thermodynamics [15]. Later it was realized that various complications (such as mutations, frequency dependence or finite-size effects) limit the relevance of Fisher’s theorem for biological evolution [16]. More importantly for our purpose, eq. (2) does not predict the behavior of $\langle S_t W \rangle$ as a function of $t$ and $W$, a shortcoming sometimes referred to as “dynamic insufficiency” [17–19]. Figure 1 plots $\langle S_t W \rangle$ for three different ancestral distributions with equal mean and variance: the divergence of the trajectories illustrates that neither $\langle W \rangle$ nor $\text{var}(W)$ are good predictors of $S_t W$ beyond the short-term or weak selection regime $t \simeq 0$. (The same is true of Price’s generalization of Fisher’s theorem, the famous “covariance equation” [20–22].) To make progress, a different approach is needed. As we now show, the key is to focus not on the moments of $W$, but rather on its tail structure\(^2\).

**Assumptions and further definitions.** – We assume that the variable $W$ has an (absolutely) continuous density $p(w)$, i.e., we exclude discrete variables and small population sizes. Next we distinguish two cases:

- **Positive selection.** The variable $W$ has unbounded support $\Sigma$, viz. $\sup \Sigma = \infty$.
- **Negative selection.** The variable $W$ has a finite right end-point $\sup \Sigma \equiv w_* < \infty$.

These two cases are idealizations: in practice, positive selection occurs when $\langle W \rangle \ll w_*$, while negative selection corresponds to $\langle W \rangle \approx w_*$. In evolutionary terms we can think of these idealizations as capturing, respectively, the dynamics of rapid adaptation and of evolutionary stasis. Crossovers between these two regimes are possible, as explained below.

The rate at which the tail function
\[ T(w) \equiv \int_{w_*}^{\sup \Sigma} p(w')dw' \] (3)
goes to zero when $w \to \sup \Sigma$ measures the thickness of the tail of $W$. How exactly this rate should be defined requires some further distinctions:

- **Positive selection.** For unbounded variables we distinguish between light and heavy tails. We say that

\(^1\)We can write this $\mathbb{P}(S_t W = w) \propto f(w)\mathbb{P}(W = w)$ for an arbitrary function $f(w)$. If this function is one-to-one and monotone we can reduce it to $f(w) = w$ by a suitable change of variable.

\(^2\)This was hinted at in [13], but our conclusions are different.
Table 1: Asymptotic location, scale and shape of selected values. Here \( \beta' \equiv \beta/(\beta - 1) \) is the exponent conjugate to \( \beta \) and \( B' \) is a positive constant which can be expressed in terms of \( B \) and \( \beta \) [23].

|                          | \( \mu_t \) | \( \sigma_t \) | \( \xi \) |
|--------------------------|-------------|-------------|---------|
| Positive selection       | \( \ln t/\alpha \) | \( 1/\sqrt{\alpha t} \) | \( \infty \) |
| (heavy tail with index \( \alpha > 0 \)) |             |             |         |
| Positive selection       | \( B'/\beta' t^{\beta'-1} \) | \( \sqrt{B'/\beta' (\beta - 1)^{2}} t^{\beta'-2} \) | \( \infty \) |
| (heavy tail with index \( \beta > 1 \)) |             |             |         |
| Negative selection       | \( \ln w_+ - \gamma/t \) | \( \sqrt{\gamma/t} \) | \( \gamma \) |
| (with index \( \gamma > 0 \)) |             |             |         |

\( W \) has a **light tail** with index \( \alpha > 0 \) if\(^3\)
\[ -\ln T(w) \sim Aw^\alpha, \text{ for some } A > 0 \]  
(4)

and a **heavy tail** with index \( \beta > 1 \) if
\[ -\ln T(w) \sim B(\ln w)^\beta, \text{ for some } B > 0. \]  
(5)

- **Negative selection.** We say that a variable with finite right end-point \( w_+ \) has a **short tail** with index \( \gamma > 0 \) if
\[ T(w) \sim C(w_+ - w)^\gamma, \text{ for some } C > 0. \]  
(6)

Note that not every distribution satisfies these assumptions. Power-law distributions, in particular, have bounded support but do not fall in the classes (4) and (5). We exclude them because they blow up at finite \( t \) under the selection dynamics (1).

Finally, we define the location and scale of a non-negative random variable \( V \) by \( \mu(V) \equiv \langle \ln V \rangle \) and \( \sigma(V) \equiv \langle (\ln V)^2 - \mu(V)^2 \rangle^{1/2} \), respectively; in particular we denote \( \mu_t \equiv \mu(S_t W) \) and \( \sigma_t \equiv \sigma(S_t W) \).

**Main result.** Based on these definitions we can now state our main result: for every variable \( V \) satisfying one of the conditions (4)–(6) above, there exists a number \( 0 < \xi \leq \infty \) and a family of distributions \( \Pi(\mu, \sigma, \xi) \) with mean \( \mu \) and scale \( \sigma \) such that \( S_t W \) approaches \( \Pi(\mu_t, \sigma_t, \xi) \) asymptotically, e.g., in the sense that the relative entropy (Kullback-Leibler divergence) between \( S_t W \) and \( \Pi(\mu_t, \sigma_t, \xi) \) goes to zero as \( t \to \infty \):
\[ D[S_t W \| \Pi(\mu_t, \sigma_t, \xi)] \to 0. \]  
(7)

We denote by \( S_t W \approx_{t \to \infty} \Pi(\mu_t, \sigma_t, \xi) \) this asymptotic relation.

Explicitly, the limiting distributions \( \Pi(\mu_t, \sigma_t, \xi) \) have support \([0, e^{\mu_t + \sigma_t \sqrt{\xi}}]\) and density
\[
\pi(w; \mu, \sigma, \xi) = \frac{e^{-\mu/\sqrt{\xi}/(\sigma - \xi/\sqrt{\xi})}}{\Gamma(\xi)\sigma^\xi} \times \frac{\xi^{\xi/2}}{\sigma^{\xi/2}} (\mu + \sigma \sqrt{\xi} - \ln w)^{-\xi/2}. \]  
(8)

Moreover, the shape parameter \( \xi \) is given by \( \xi = \gamma \) for negative selection and \( \xi = \infty \) for positive selection; in the latter case \( \Pi(\mu, \sigma, \infty) \) is just a log-normal distribution, as can be seen by taking the limit \( \xi \to \infty \) in (8). The behavior of \( \mu_t \) and \( \sigma_t \) as a function of \( t \) is given in table 1.

\footnote{This condition can be generalized in terms of the notion of regularly varying function [20].}

\footnote{We denote by \( \ln D \) the distribution of a variable whose logarithm has distribution \( D \).
}

![Fig. 2: (Colour online) A truncated variable \( W \) (here a unit-mean exponential truncated at \( w_+ = 30 \)) is attracted to the “positive selection” attractor (\( \gamma = \infty \)) as long as \( \langle S_t W \rangle \ll w_+ \), at which point it crosses over to its eventual “negative selection” attractor (\( \gamma = 1 \)). Here we represent this crossover in terms of the skewness and kurtosis of \( \ln S_t W \) (blue line); in this plane the limiting distributions \( \Pi(\mu, \sigma, \gamma) \) form a half-parabola (thick line).}

**Sketch of proof.** By the invariance of relative entropy under reparametrization, eq. (7) is equivalent to the statement that \( X_t \equiv \langle \ln S_t W - \mu_t \rangle/\sigma_t \) converges to one of a one-parameter family of limits \( \Pi(0, 1, \xi) \) (see footnote\(^4\)). We prove this by computing the cumulants \( K^{(p)} \) of \( X_t \) for \( p \geq 3 \) in the limit \( t \to \infty \). To this aim we introduce the cumulant-generating function (cgf) \( \psi_t(\nu) \) of \( \ln S_t W \), such that \( \mu_t = \psi_t'(0), \sigma_t^2 = \psi_t''(0) \), and \( K^{(p)} = \psi_t^{(p)}(0)/\psi_t''(0)^{p/2} \).

Evaluating \( \psi_t(\nu) \) from its initial condition \( \psi(\nu) \) (the cgf of the parent variable \( \ln S_0 W = \ln W \)) is a simple problem once we realize that the selection equation (1) is equivalent to the transport flow
\[ \psi_t(\nu) = \psi(\nu + t) - \psi(t). \]  
(9)

From this observation it follows that the derivatives of \( \psi_t(\nu) \) evaluated at \( \nu = 0 \) are just the derivatives of \( \psi(\nu) \)
evaluated at \( \nu = t \). The large-\( t \) behavior of \( S_t W \) is therefore completely controlled by the large-\( \nu \) behavior of \( \psi (\nu) \), which, in turn, is completely determined by the tail behavior of \( p(w) \):

\[
\psi (\nu) = \ln \int_0^\infty w^\nu p(w) \, dw.
\]

(10)

We can estimate the large-\( \nu \) behavior of \( \psi (\nu) \) above using the approximations (4), (5) and (6) in (10). We treat each case separately.

Positive selection (light tails). When \( W \) has a light tail with index \( \alpha \), Stirling’s formula for (10) gives

\[
\psi (\nu) \sim \nu \ln \nu \alpha.
\]

(11)

For \( \rho \geq 3 \) the cumulant \( K_t^{(p)} \) goes to 0 like \( t^{1-p/2} \) when \( t \to \infty \). The unique distribution with vanishing cumulants is the Gaussian, hence \( W \) is asymptotically log-normal. But since a log-normal distribution with vanishing scale is itself Gaussian, we obtain that

\[
S_t W \approx \mathcal{N} \left( Ct^{1/\alpha}, (at)^{-1/2} \right),
\]

(12)

where \( \mathcal{N}(m, s) \) represents a Gaussian with mean \( m \) and standard deviation \( s \). Note the emergence of the dynamical scaling law \( t^{1/\alpha} \) for the “speed of evolution” \( \langle S_t W \rangle \) under positive selection (fig. 1).

Positive selection (heavy tails). For heavy-tailed distributions we invoke Kasahara’s Tauberian theorem [23,24] to estimate

\[
\psi (\nu) \sim B' \nu^{\beta'},
\]

(13)

where \( \beta' = \beta/(\beta - 1) \) is the exponent conjugate to \( \beta \) and \( B' \) is a positive constant which can be expressed in terms of \( B \) and \( \beta \). From this it follows that \( \psi^{(p)}(t) \) scales like \( t^{\beta' - p} \), and, therefore, \( K_t^{(p)} \) goes to zero like \( t^{\beta'(1-p/2)} \) for all \( p \geq 0 \). This implies that \( W \) is again asymptotically log-normal as \( t \to \infty \). For \( \beta \leq 2 \) we obtain a genuine log-normal distribution (denoted \( \ln \mathcal{N}(\mu, \sigma) \) with location \( \mu \) and scale \( \sigma \)), namely

\[
S_t W \approx \ln \mathcal{N} \left( B' \beta' t^{\beta'-1}, \sqrt{B' \beta' (\beta' - 1)} t^{\beta'-2} \right),
\]

(14)

while for \( \beta > 2 \) the distribution reduces to the Gaussian

\[
S_t W \approx \mathcal{N} \left( e^{B' \beta' t^{\beta'-1} + o(t^{\beta'-1})}, \sqrt{B' \beta' (\beta' - 1)} t^{\beta'-2} \right).
\]

(15)

In this regime the mean \( \langle S_t W \rangle \) grows super-exponentially with \( t \) — an explosive form of selection dynamics fuelled by large amounts of initial variation.
Table 2: Abbreviations: NBA: National Basketball Association; CDN: Canadian Diary Network; ESPN: Entertainment Sport Programming Network; LPI: Lifetime Performance Index; PER: Player Efficiency Rating; ML: maximum likelihood; LLH: log-likelihood. The p-value for log-likelihood ratio tests against the Weibull distributions is computed using the procedure described in appendix C of [25].

| Dataset                      | Source                 | Selected trait w | N         | Skewness | ML (μ, σ, γ)                  | LLH ratio test |
|------------------------------|------------------------|------------------|-----------|----------|------------------------------|----------------|
| Randomized antibodies        | Ref. [13]              | Selectivity      | 6,159     | 8.36     | (−8.89, 0.43, 8.45 × 10⁵)    | p = 9.59 × 10⁻²³ |
| NBA sires (2008/2016)        | CDN LPI                | 10,033           | 0.51      | (7.32, 0.31, 27.07)          | p = 1.32 × 10⁻²⁸ |
| NBA players (2016)            | NBA Height             | 450              | −0.40     | (2.62, 0.31, 28.11)          | p = 0.69       |
| NBA players (2003–2016)      | ESPN Hollinger PER     | 4,141            | 0.57      | (0.70, 0.04, 7.59)           | p = 5.66 × 10⁻⁹ |

**Negative selection.** When W is bounded we have by Laplace’s method for (10)  
ψ(ν) ∼ (ln wₐ)ν − γ ln ν  
(16)

from which we compute \( \lim_{t \to \infty} K_1^{(p)} = \gamma^{1-p/2}(-1)^p (p - 1)! \). These are the cumulants of a flipped gamma distribution. Exponentiating back to \( S_tW \) we obtain  
\[ S_tW \approx \Pi(\ln wₐ - \gamma/t, \sqrt{\gamma}/t, \gamma). \]  
(17)

**Convergence rates.** Like the location \( \mu \) and scale \( \sigma \), the rate of convergence of \( S_tW \) to its limiting shape (measured by the relative entropy \( D_t \equiv D[S_tW || \Pi(\mu, \sigma, \xi)] \)) depends on the tail of \( W \). For positive selection we find \( D_t = O((K_1^{(3)})^2) \), giving a rate of convergence \( O(t^{-1}) \) for light tails and \( O(t^{-\beta}) \) for heavy tails. For negative selection, assuming  
\[ \frac{T(w)}{C(wₐ - w)} - 1 \approx Q(wₐ - w)^q, \]  
(18)

for some \( q > 0 \), we compute \( D_t = O(t^{-2q}) \).

**Crossovers and finite-size effects.** In some cases, the evolution of \( S_tW \) as \( t \) increases can display a crossover between the limiting types for positive and negative selection. This arises, e.g., when \( W \) has a truncated distribution, such as a truncated exponential distribution \( p(w) \propto \theta(w₊ - w)e^{-\lambda w} \) with \( \lambda w₊ \gg 1 \). In that case \( S_tW \) approaches a (log-)normal distribution \( \Pi(\mu, \sigma, \infty) \) as \( t \) increases, until \( S_tW \) becomes comparable to the upper endpoint \( w₊ \), at which point \( \mu_t(w) \) shifts to the negative-selection attractor \( \Pi(\mu, \sigma, 1) \). We can illustrate this behavior by plotting the skewness and kurtosis of \( \ln S_tW \) as a function of \( t \) (fig. 2). In this representation the universal family \( \Pi(\mu, \sigma, \gamma) \) corresponds to a half-parabola where all selected values end in the limit \( t \to \infty \).

It also worth emphasizing that the above results hold in the infinite population limit. For a finite population with size \( N \) the scale parameter \( \mu_t \) is bounded by  
\[ \mu_t(N) \approx p^{-1}(1/N). \]  
(19)

For a thin-tailed distribution with index \( \alpha \) this gives \( \mu_t(N) = \mathcal{O}(\ln^{1/\alpha} N) \). When \( \mu_t \) reaches this value, the granularity of \( W \) in the tail becomes dominant, \( \mu_t \) plateaus, and our limit theorems are no longer relevant.

**Datasets.** We compared our predictions to four natural candidates for empirical selected values (table 2): the performance index (LPI) of commercial sires (selected by dairy farmers), the height and player efficiency rating (PER) of NBA players (selected by team coaches), and the selectivity of randomized antibodies with respect to a molecular target (selected by the experimental apparatus of Boyer et al.). As shown in fig. 3, the universal family \( \Pi(\mu, \sigma, \gamma) \) is a good fit to the empirical distributions, all of which are non-Gaussian (\( p < 10^{-18} \) or less; Pearson \( \chi^2 \) test). Moreover, alternative fits with the three-parameter Weibull distribution always perform worse, significantly so in three cases (table 2). We conclude that selection is a plausible explanation for the observed skewness of these variables.

**Conclusion.** We have shown that selected values have universal properties: they are attracted to a parametric family whose location, scale and shape are solely determined by the tail of the variable being selected. A parallel can be established between these results and the Fisher-Tippett-Gnedenko theorem of extreme value theory [26]. Indeed, the extremality condition \( M_n = \max\{X_1, \ldots, X_n\} \) can also be viewed as representing an alternative form of “selection”, in which the maximum \( M_n \) is picked out from the population \( \{X_1, \ldots, X_n\} \). This analogy is commonly made in the genetic algorithms literature, with selected and extreme values referred to as “proportionate” and “tournament” selection, respectively [27].

Another analogy is with the Lifshitz-Slyozov-Wagner (LSW) theory of Ostwald ripening [28,29]: just like selected and extreme values, the size of particles in a coarsening solution follows a universal distribution characterized by the tail behavior of a suitable probability distribution. This analogy is best seen by rewriting the selection
equation (1) in terms of the density $\pi_t(x)$ of $\ln S_t W$, as

$$\partial_t \pi_t(x) = \left( x - \int y \pi_t(y) dy \right) \pi_t(x).$$

(20)

This integro-differential equation is rather similar to the LSW equation. It was shown in ref. [30] that the LSW equation has the structure of a gradient flow, and in particular has a Lyapunov functional. Whether or not a similar structure can be constructed for the selection equation is an interesting open problem.

***

MS thanks Felix Otto for pointing out the analogy with Ostwald ripening. We are grateful to an anonymous referee for helpful criticisms. Research at the Perimeter Institute is supported in part by the Government of Canada through Industry Canada and by the Province of Ontario through the Ministry of Research and Innovation.

REFERENCES

[1] Price G. R., J. Theor. Biol., 175 (1995) 389.
[2] Shannon C. E., Bell Syst. Tech. J., 27 (1948) 379.
[3] Nielsen R., Annu. Rev. Genet., 39 (2005) 197.
[4] Huang S., Proc. Natl. Acad. Sci. U.S.A., 113 (2016) 3482.
[5] Price G. R., J. Theor. Biol., 175 (1995) 389.
[6] Fisher R., The Genetical Theory of Natural Selection (The Clarendon Press) 1930.
[7] Orr H. A., Nat. Rev. Genet., 10 (2009) 531.
[8] Lewontin R. C., The Genetic Basis of Evolutionary Change (Columbia University Press) 1974.
[9] Barton N. H. and Turelli M., Genet. Res., 49 (1987) 157.
[10] Frank S. A., Evolution, 51 (1997) 1712.
[11] Li C. C., Nature, 214 (1967) 505.
[12] Robertson A., Anim. Prod., 8 (1966) 95.
[13] Price G. R., Nature, 227 (1970) 520.
[14] Bingham N. H., Goldie C. M. and Teugels J. L., Regular Variation (Cambridge University Press, Cambridge) 1989.
[15] Kasahara Y., J. Math. Kyoto Univ., 18 (1978) 209.
[16] Clauset A., Shalizi C. R. and Newman M. E. J., SIAM Rev., 51 (2009) 661.
[17] de Haan L. and Ferreira A., Extreme Value Theory, An Introduction (Springer Science & Business Media, New York, NY) 2007.
[18] Blickle T. and Thiele L., Evol. Comput., 4 (2007) 361.
[19] Lifshitz I. M. and Slyozov V. V., J. Phys. Chem. Solids, 19 (1961) 35.
[20] Wagner C., Z. Elektrochem., 65 (1961) 581.
[21] Niethammer B., Commun. Math. Sci., 2 suppl. issue (2004) 85.