EXTRAORDINARILY CORRUPT OR STATISTICALLY COMMONPLACE? REPRODUCIBILITY CRISES MAY STEM FROM A LACK OF UNDERSTANDING OF OUTCOME PROBABILITIES.

DRAFT

Caetano Souto-Maior[1,2‡]
[1]BCAM – Basque Center for Applied Mathematics, Bilbao, Spain
[2]Laboratory of Systems Genetics, National Heart Lung and Blood Institute,
National Institutes of Health, Bethesda, MD, United States of America
‡ csouto@bcamath.org

June 6, 2022

ABSTRACT

Failure to consistently reproduce experimental results, i.e. failure to reliably identify or quantify an effect — often dubbed a 'reproducibility crisis' when referring to a large number of studies in a given field — has become a serious concern in many communities and is widely believed to be caused by (i) lack of systematic methodological description, poor experimental practice, or outright fraud. On the other hand, it is common knowledge of the scientific practice that (ii) replicate experiments — even when performed in the same lab, by the same experimenter — will rarely show complete quantitative agreement between them. The presence of the widely believed (i) and commonplace (ii) explanations are not mutually exclusive, but they are incompatible as justifications for irreproducibility. Invoking the former implies an anomaly, a crisis, while the latter is statistically expected and therefore amenable to quantification.

Interpreting two or more studies as conflicting is often a reduction to a mechanicist view where a ground truth exists that must be observed with every properly performed experiment; a slightly less naive view (at best) is a frequentist view where statistical tests must confidently identify a true effect (i.e. a single parameter value) as significant almost always (i.e. an arbitrary proportion of 95% of times). A broader view, however, may consider that the effect can only be observed as a probability distribution; individual experiments are, therefore, not expected to differ only by sampling and power to identify a significant effect, but by variation at the level of the parameter value itself — i.e. it is accepted that there are sources of variation that cannot be controlled with infinite precision, for instance in the environment and from the experimenter, or it is acknowledged that there may be unknown, uncontrolled factors that will introduce biases. Quantitatively, that perspective is consistent with a Bayesian hierarchical formulation, where the effect (commonly called the group-level) parameters are under a hyperprior and above individual experiment parameters.

Put another way, the Bayesian hierarchical view allows reconciliation between seemingly discordant results by interpreting each experiment as a sample itself of a (group- or system-level) distribution, which in turn sets the range and probability of expected outcomes for new individual experiments. As a corollary, a large number of replicates will increase the confidence not only in the expected value but also in the deviation for it. Thus, "validating" an experiment does not mean getting the same number every time, but establishing the range and likelihood of well-performed experiments. Conversely, once an experiment has been extensively replicated, the effect distribution is informative of how much each repetition deviates from expectation, whether they are actually extreme — and potentially contain anomalies or misconduct — or if they are probabilistically not surprising. This formulation has profound consequences for assessments and claims on reproducibility.

1 Background: reproducibility crises in experimental research

It has traditionally been assumed that a scientific finding, once found, must be true; scientific tradition had good reasons for the assumption: the researcher presenting it has the appropriate expertise, the experiments are well designed and performed, and the methodological description is detailed enough for anyone to reproduce the results. That is if the need ever arises to try and reproduce the work, otherwise there would normally be no reason to doubt it – if a natural process exists, it can be observed anytime, by anyone, using the appropriate tools and just enough expertise on the topic [Henderson, 2020]. Although this is not logical proof of the converse statement, reproducibility of a scientific result is (or adds to) evidence of the *truth* of a finding (as opposed to an artifact of tools or of the experimenter); it is expected to follow in throughly-performed experiments without further questioning. Given that assumption, a number of studies not being replicable is worrisome, and irreproducibility of the majority of the work in a field is a bonafide crisis [Errington et al., 2021, Klein et al., 2022]; in fact it is good reason to question the body of scientific knowledge itself.

In the past decades to centuries many traditions were forced to change in scientific practice [Beer and Lewis, 1963], from being an amateur endeavor performed over an undefined amount of time and communicated on printed paper to being a professional, public funding-bound activity instantly available worldwide upon publication, to mention only a few changes. Nevertheless, historical contingencies still guides much of modern research, and requiring and depositing the same degree of societal trust in "science" today as it was a century ago is not warranted – research is now performed under very different constraints, in addition to having both broader and deeper importance to societal development. In other words, the work performed by the extensive network of research institutions and almost nine million professionals [Schneegans et al., 2021], that forms the basis of a technology-oriented (and possibly obsessed) society, must be stringently verified if we are to justify the expense and and reliance of its results.

In that light, reproducibility of results or lack thereof has come to epitomize the very trust in the robustness of science and in scientists themselves, and there are reasons for concern about the quality of current research. Being a profession with merit and career prospects largely based on publications and especially publication-derived metrics, under anything less than ideal conditions it is likely that deleterious incentives will arise. Academia has become increasingly competitive [Carson et al., 2013], with publication record requirements becoming extremely high and often unrealistic or unreasonable; therefore, striving for solid, carefully described, and replicated work is often relegated to the background in favor of flashy results or subjective novelty [Begley et al., 2015] – e.g. an interestingly unexpected "story" is more likely to be published in a "glam" journal and subsequently cited [Chu and Evans, 2021]. Academia offers precarious positions – sometimes compared to a *ponzi scheme* – which entails long-term income loss without guarantee of a stable position later on; it relies on archaic structures that place a disproportionate amount of power in the hands of supervisors leading to lack of career freedom at best and abuse and harassment at worst, with mental health issues and suicide rates at much higher rates than most professions. Academia is often said to be utterly "broken", and a corollary is that a system that is irremediably rotten cannot but produce mostly (or a high proportion of) false or unreliable findings [Lydersen and Langaas, 2021]. I do not dispute any of those claims – their effects are likely deleterious – and agree that most of them need to be urgently addressed; nevertheless, I argue that they are not the only reason for apparently inconsistent results, and not being easily quantifiable cannot be formally (but should be systemically) addressed.

It is worth noting that there have been counterclaims stating there is no evidence of deterioration in scientific knowledge, although they are mostly subjective or use the limited, coarse data available. That is not the counterpoint I wish to make, instead it is this: in most or all experimental research (especially in the life sciences) it is fully expected that any repetition will *not* yield the exact same results, in fact they may yield very different ones even when performed with the same materials, by the same experienced researcher, and only a few weeks or days apart. Differences in repeated measurements can be partly explained by measurement error and uncontrolled or uncontrollable environmental factors, and are statistically incorporated into analyses as implicit test assumptions or explicit variance parameters; therefore, they are formally accounted for within the scope of the experimental design. It is possible to try and control or reduce to some degree this variation, but impossible to eliminate it completely – that is also the reason why sample sizes should be as large as possible. This quantitative reality is well-accepted and understood by all (or at least most) experimental scientists in the context of individual measurements, but discrepancies between entire replicate experiments are seen as qualitative – they do not fit that basic statistical framework *as is*, and may be labeled as instances of *irreproducibility*.

I argue here that an apparent qualitative discrepancy can arise from an inadequate interpretation of the structure of variance, and is in fact quantitative and amenable to existing statistical treatments, not unlike that of individual measurements with error. In the next sections I demonstrate how simplistic assumptions can lead to that picture, how to correctly interpret replicate experiments in a general (and intuitive) way, and the implications for integrating information from replicate experiments.

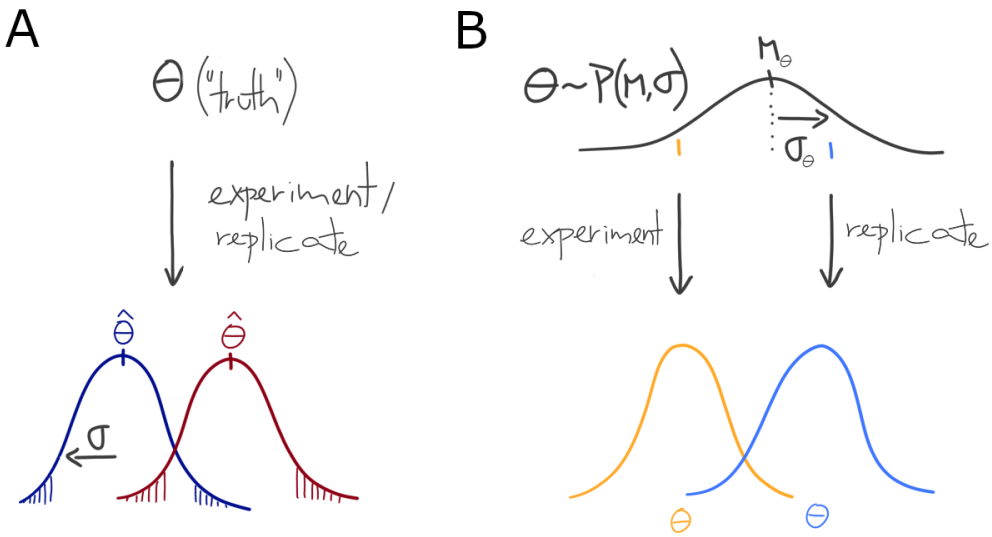

Figure 1: Under a naive view of experimentation and replication there is a true effect value θ (i.e. the *parameter*), and different experiments will produce different *point estimates* for the parameter only due to sampling. (A). A more general view allows for variation in the parameter: each experiment is a draw from a distribution $P(\theta|\mu, \sigma)$, i.e. the effect of interest is interpreted as a system-level distribution (black) that applies to all experiments – individual results (orange, blue) are expected to vary even under perfect sampling (B).

2 Naive views of experimentation and reproducibility in the natural sciences: "The *truth* is out there"

The most naive, deterministic view of an experiment is that it should always lead to the exact same outcome, the same exact measurement value – i.e. there is a true value θ that can be observed with infinite or arbitrary precision. That will essentially never happen except for deterministic computer simulations (and even then only up to high-but-finite machine precision), since even the narrowest physical observations will still show some intrinsic variation and high-precision equipment is also always associated to some measurement error. As Leonelli [2018] points out, replicability/reproducibility have different meanings as well as usefulness in systems with different expected variation; here I will focus on experimental research, particularly biological systems, where there is clearly noticeable variation between subjects of a sample within any experiment (there may be extreme cases with barely any variation, like a highly-toxic compound at a very high dose, but those are both rare and of little significance).

Under this basic statistical "the-truth-is-out-there" picture, there is still a *true parameter* θ, but it can only be observed with some variance σ^2 due to sampling – although it is rarely the case that observations are direct sampling of the parameter (Fig. 1A), any statistical test or model can be formalized to have θ represent the relevant quantity conceptualized here. Given an experimental design with sample size N it should be possible to identify an *effect* of a certain magnitude at a chosen significance level (e.g. the infamous $\alpha = 0.05$).

An example of this formulation is a simple linear regression, $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i + \cdots + \beta_n x_i + \varepsilon_i$ where y_i denotes the value of the i^{th} observation and random noise ε_i is added to each as a draw from a normal distribution. The parameter $\theta = \boldsymbol{\beta}$ can be written as a vector of coefficients $\boldsymbol{\beta} = [\beta_0\ \beta_1\ \beta_2\ \cdots\ \beta_n]$, which yields a compact linear algebra notation with X as the design matrix for the experiment and vector of random noise $\boldsymbol{\varepsilon}$:

$$\begin{aligned} \mathbf{y} &= X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \varepsilon &\sim \mathcal{N}(0, \sigma) \end{aligned} \tag{1}$$

It is also common practice to repeat an experiment and pool the data, but in practice the pooling is only done if the new data "improves" the result, which is believed to be by virtue of increased sample size: if the repetitions are qualitatively "similar" (e.g. the effect has same sign and subjectively the magnitude is considered similar) no issue is taken, but if individually they are "different" (e.g. one is significant and the other is not, or they have opposite signs) they are considered inconsistent (Figure 2A) – in any case, no two results can be formally reconciled this way without some *post hoc* analysis, only subjectively be classified as "agreeing" or "disagreeing".

Here a series of practices that are at least questionable are unfortunately also common: apply different statistical tests and use those where all experiments turn out significant, analyze all combinations of subset pooling and use those that turn out significant, discard supposed outliers (or whole replicates) without justification, and others; I will not deal here with misuse of basic statistical methodology intentionally or by neglect of good practice.

The assumption when analyzing two or more repetitions of the same experiment is that the parameters generating the data are unchanged because this is the *truth*; if that holds, generating more samples by repetition cannot affect θ, it simply increases N and with it the power obtain significance when testing for an effect. This assumption is convenient and allows repetition to feed into a simple calculation of power to identify an effect by appealing to the central limit theorem, which states that the standard error of the mean will decrease with sample size as $s \approx \sigma/\sqrt{N}$. Nevertheless, this assumption may also be wrong and, as mentioned above, with no way to reconcile discrepant results the only explanation is irreproducibility by sloppiness, incompetence, fraud (or at best, bad luck). Adopting a less narrow assumption, however, allows other explanations. There have been calls, for instance, for the use on formal methods to address replicability by conditioning on the original result and accepting that replication rates vary across fields and types of experiments [Devezer et al., 2020]; this moves away form the naive view by implicitly abandoning the notion that the same result must be observed each time. In the next section I will instead try and make this explicit using familiar statistical constructs.

3 The truth in uncertain: incorporating variation into biological effects

A more general picture relaxes the assumption of a true, fixed θ for all experiments performed, past or future, instead allowing it to vary between experiments. Some of the formal features of this paradigm are intuitively familiar to any experimenter: individual data points cannot be observed separately at arbitrary times or places an then analyzed together, they must be observed within the same experimental context – measurements produced in different contexts (e.g. by different experimenters, different days of the week, with different equipment, reagents, or protocols) usually are not directly comparable. Similarly, it is universally accepted in the life sciences that treatment and control groups need to be part of the same experiment, performing a measurement on a treatment group and assessing controls at a different occasion will almost certainly produce artifactual results.

Replicates, on the other hand, are by design and necessity performed in different experimental contexts – whether it is repetition by the same experimenter at a later time, or an attempt to reproduce the results by a different laboratory. It is also intuitively known that replicates will not produce the same results, but it is believed that the "broad features" must be conserved between replicates (tough this is usually loosely defined) – in sum: variation is expected because not everything can be controlled, but there must be some uniformity since the same biological mechanisms should be at play [Glass, 2014].

The simplest formalization of this would be a mixture, or compound distribution: at the experiment level, deviation τ is expected around an average θ_j, but that average itself is distributed according to another distribution (Fig. 1B), i.e. $\theta \sim P(\mu, \sigma^2)$.

$$\begin{aligned} \theta &\sim P(\mu, \sigma^2) \\ x &\sim Q(\theta, \tau^2) \\ f(x) &= \int_{-\infty}^{\infty} q(x|\theta)p(\theta)d\theta \end{aligned}$$

where f is the probability density function and the fixed parameters τ, μ, σ are omitted for clarity.

Staying within a frequentist framework this leads to *random effects* formulations [Borenstein et al., 2010]. That is, in addition to random error ε_i for individual observations, further error γ_j is associated to each replicate (i.e. the former is different for each observation i, the latter is the same for observations within the same replicate j, but different between replicates, $\gamma_j \neq \gamma_k \iff j \neq k$). Under the guise of linear regression, for instance, this leads to variable-slopes and/or variable-intercept models; each replicate can have their own slope/intercept because the noise added to the parameters is specific to each repetition; however, the fixed linear coefficients β_l themselves are common to all of them. Formally this amounts to including a vector $\boldsymbol{\gamma} \sim \mathcal{N}(0, G)$ in equation 1 and a design matrix Z that encode the replicate-specific errors, which for convenience is assumed to be gaussian and do not change θ on average:

$$\begin{aligned} \mathbf{y} &= X\boldsymbol{\beta} + Z\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \\ \varepsilon &\sim \mathcal{N}(0, \sigma) \\ \gamma &\sim \mathcal{N}(0, G) \end{aligned} \tag{2}$$

As a corollary, seemingly discrepant results are easily explained: having slopes with opposite signs, or different significance can be explained by the random effects formalized as gaussian noise.

While the frequentist description of equation 2 is convenient in many ways – with Ordinary Least Squares as a starting point, it can be straightforwardly formalized using compact linear algebra notation to give a closed-form solution – an equivalent and arguably more intuitive description is that of hierarchical Bayesian inference. Under this formalism the parameters that generated the observations for each experiment can be seen as draws from a distribution – the former can be dubbed individual- or replicate-level effects and the latter group- or system-level effects that is common to all replicates. In the context of linear models, individual-level effects are equivalent to the varying slopes/intercepts of random effects models, except the process of obtaining them is somewhat reversed: they arise directly from a distribution as opposed to starting with a single value and adding noise to it through the vector $\boldsymbol{\gamma}$. If the distribution of parameters is gaussian, the formulations are equivalent; however, the bayesian approach is general and admits any system-level distribution, not just gaussian – in addition to being more intuitive and visual (Fig. 1).

The hierarchical paradigm readily incorporates the intuitively-understood phenomena mentioned above – e.g. the parameters that generate data are different between replicates due to uncontrolled or uncontrollable variables, but they are not arbitrary or unrelated, they are instances of a common distribution with well-defined mean and finite variance. More importantly, the system-level distribution integrates data from multiple repetitions, and conversely allows specific replicates to be assessed as "typical" or "extreme" according to the probabilities given by the common distribution (Figure 2B). As mentioned above, a naive fixed-effects approach not only cannot account for this variation between settings, but replication cannot be integrated into a single informative quantity. At a basic level it will cause misinterpretation of repeated experiments within a lab, at a larger scale it will generate false expectations for reproducibility of results.

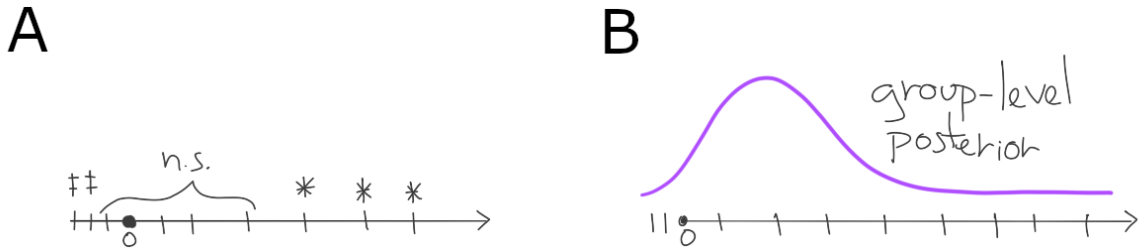

Figure 2: Replication (ticks along the axis) under a significance testing framework will result in a string of significant($*$, or in the opposite direction ‡) and nonsignificant ($n.s.$) outcomes which cannot be quantitatively assessed (A). Under a hierarchical model replication contributes increasingly to the confidence in the group-level distribution, and individual experiments can be assessed against the probabilities given by this distribution (B).

4 A brief case study

Analysis of Differential Expression (DE) of genes relies on measuring counts of RNA transcripts (or possibly other gene products, like protein). These are quantified using technologies like the direct sequencing of RNA molecules (*RNA-seq*), though formerly they used other methods like *microarrays* containing a large number of probes that would anneal to specific sequences [Lowe et al., 2017]. Linear models are then applied to compare those counts between different experimental groups and assess significance of the differences.

Expression data is known to be noisy and, being a fairly complicated and expensive method, sample sizes are constrained by effort and cost. Entire experiments often consist of a handful of samples (although each sample produces readouts for thousands of genes, this does little for statistical power). Given the complexity of the systems and characteristics of the methodology it is common to see wide variation in the readouts obtained by different experiments on the same system – once again, it is expected that despite changing values the general *trends* (also and again loosely defined, e.g. expression in one group is consistently greater than the other). It is also common for that not to be observed. Many analysis packages implement a linear or Generalized Linear Model (GLM) for that purpose, but they normally do not allow for random effects, and therefore require either that replicates are pooled as one, or that they are analyzed separately with independent results obtained for each of them.

In [Souto-Maior et al., 2021] a large artificial selection plus RNA-seq experiment was conducted, with three populations (one of them being a control population), thirteen time points, which was replicated twice weeks apart. The resulting data set was quite large compared to typical RNA-seq experiments, with a total of 312 samples; furthermore sequencing

of samples along generation allows clear visualization (Fig. 3) and inference of the linear trend along time (as opposed to the difference between treatments at an initial and final time point, or even of a single time point).

Nevertheless, the expanded design does not solve the issue of replication, and in some cases the generation trend is opposite between replicates (the slope is positive for one replicate and negative for the other for the same treatment group, see for instance Figure 2 of that reference). A hierarchical GLM allows each replicate to have their own slope, while the group-level distribution accounts for the range of possible values of the linear coefficients – this is appropriate in this system because selective breeding creates phenotypes by combination and recombination of any genes that may have an effect, and it is not expected that the same genes will be selected to the same extent every time the artificial selection procedure is replicated.

Replicates have inbuilt sources of variation that are not controlled; therefore, assuming parameters do not change between experiments is a procedure very likely to produce artifactual results. Acknowledging this variation through a system-level distribution of parameters, out of which individual replicates are drawn allows statements about the system to be made, instead of about individual – and potentially contradictory – replicates. Conversely, the system-level distribution may have large variance, in which case statements about its parameters cannot be very precise – increasing the number of replicates, R, will increase this precision, just like increasing number of observations, N, will increase within-experiment precision.

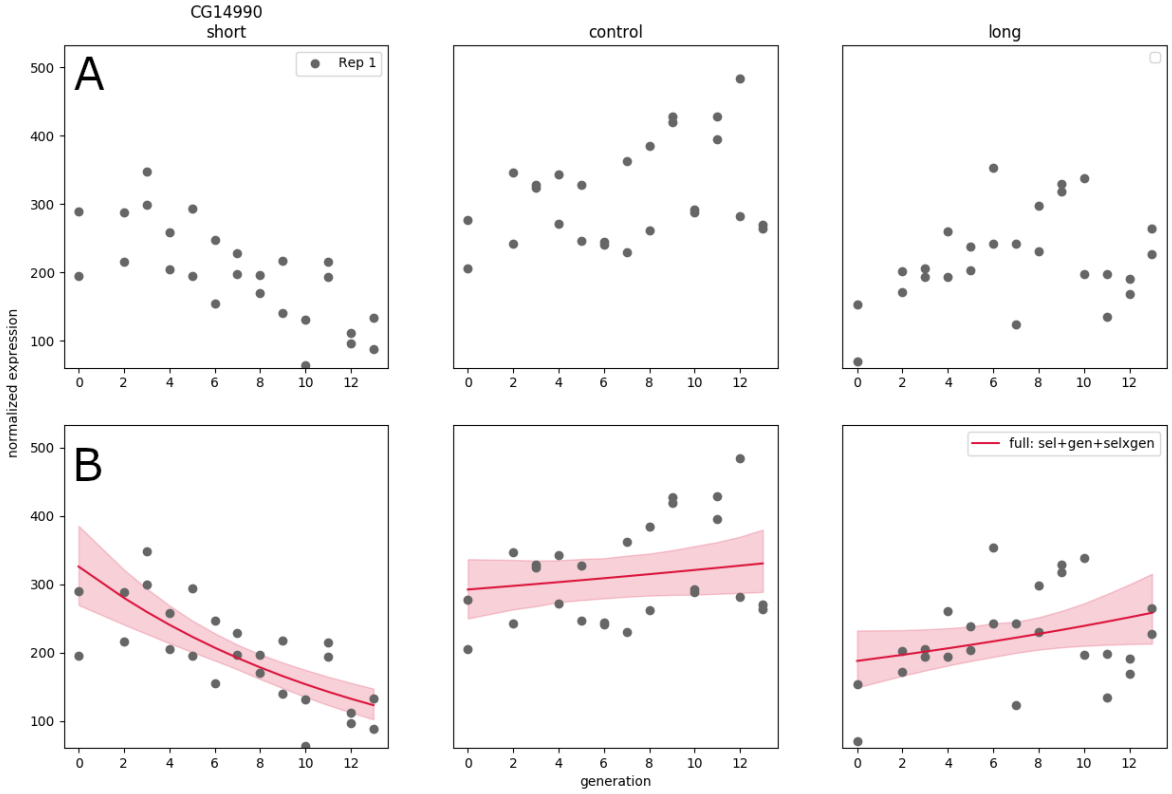

Figure 3: RNA expression data for a single replicate under different treatments and controls as a function of generation in artificial selection experiment (A). (Generalized) Linear Model fit to that data showing trends in expression as a function of time (B).

5 Implications and perspectives

Reproducibility is essential to science given the expectation of uniformity in the laws of nature; nevertheless, these same laws of nature are probabilistic at all scales, whether due to properties intrinsic to the system or epistemological constraints. It cannot be and it is not expected that all replication will yield the same result, after all, under the slightest uncertainty "the same" is a subjective label.

There was a time when statistics was not a part of mainstream research, but at this day and age no serious scientist would analyze individual observations separately, without a statistical framework to account for their distribution. Drawing conclusions about an experiment from a single observation is not informative about the population, and neither is making statements about a system from a single experiment – the absence of proper sampling in both cases only gives the illusion of confidence. Like variation between individual observations, variation between replicates must be formally accounted for and integrated to improve our expectation of system parameters – the two levels of estimates conceptualized here (observation and replicate) immediately suggest a hierarchical structure. As a consequence, this formalization marks the transition between a paradigm where replicates agree (or disagree) qualitatively – and interpretation becomes increasingly difficult as data accumulates – to one where all replication contributes to a better understanding of the system, and even opposite results increase our confidence on the expected variation in a system.

The implications of this formulation go beyond the interpretation of replicates of entire experiments. Given a working model, any observations of the system (provided that the proper metadata is recorded, e.g. time) will contribute to a better precision of the system-level distribution. Conversely, arbitrarily large data sets can be put together by replicating whole or parts of experiments or even disjoint observations (for instance time points beyond the original range), all of which can be performed under different settings. While in principle several small, disparate experiments can be integrated to create very large data sets using this framework, large and well-performed experiments will still be valuable, since they will make greater contributions to the system-level distribution (by virtue of having more data points and of having better precision, respectively). On the other hand, it is possible to improve on large experiments by adding smaller experiments to the same analysis, instead of having to repeat a larger experiment.

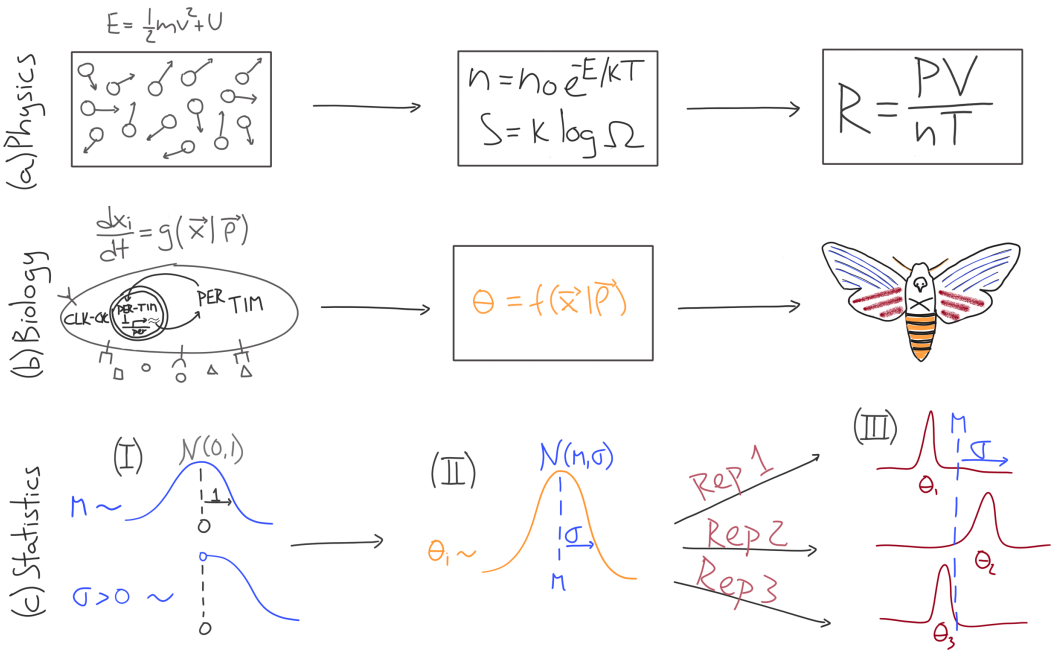

Figure 4: Robust higher-level phenomena emerge from integrating lower level interactions in (a) physical systems, where laws and universal constants are precisely defined by system state, and (b) biological systems, where intrinsic features are captured by (c) a Bayesian hierarchical structure: hyperpriors (I) provide generic information to *a priori* (II) distributions $\mathcal{N}(0, \sigma)$ on model parameters θ_i; the latter distribution reconciles replicate- specific estimates from disparate experimental settings (III) with deviation σ from expectation μ.

Another application of this framework is in obtaining confidence in estimates where some parameters are conserved but others are expected to be completely independent between settings. That is the case for instance of estimates of the reproductive number (R_0) of SARS-CoV-2; this quantity depends both on intrinsic biological parameters (like infectivity of each viral particle) as well as behavioral (contact rates between people), public health policy (vaccination coverage, use of nonpharmaceutical interventions), and possibly geographical (humidity, temperature) variables. Inference of R_0 will therefore yield wildly different values if done at different times or places; even if the biological component can be disentangled from the rest it is more than likely that estimates will vary between settings. The usual result would then be a series of estimates with little information about which one is "correct" or "better" nor how to compare them. Under a hierarchical framework all data sets could be fit under the same or different models with the common parameter being

under a common distribution which would allow estimates to be compared according to their likelihood (Fig. 2) and how much they deviate from the mean.

Finally, an analogy between this statistical framework and natural laws, where observable quantities are not directly informative about the underlying processes that generate them is illustrated in Fig. 4. It is necessary to account for the relevant structure to interpret the observations in a unified way. In physics variation can sometimes be ignored, and through statistical mechanics macroscopic laws arise from deterministic microscopic interactions. In biology randomness must be properly described: phenotypes are a high-level, noisy observation of system with specific parameters, but with variation due to uncontrolled environmental and experimental factors. Integrating multiple observations under a suitable modeling and statistical framework allows proper interpretation of the the underlying mechanisms and improves confidence in their inference.

6 Acknowledgments

I would like to acknowledge Susan Harbison and María-Xosé Rodríguez Álvarez for the support and feedback, as well as members of the Applied Statistics research line at BCAM for additional comments.

References

J. J. Beer and W. D. Lewis. Aspects of the professionalization of science. *Daedalus*, pages 764–784, 1963.

C. G. Begley, A. M. Buchan, and U. Dirnagl. Robust research: Institutions must do their part for reproducibility. *Nature*, 525(7567):25–27, 2015.

M. Borenstein, L. V. Hedges, J. P. Higgins, and H. R. Rothstein. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research synthesis methods*, 1(2):97–111, 2010.

L. Carson, C. Bartneck, and K. Voges. Over-competitiveness in academia: A literature review. *Disruptive science and technology*, 1(4):183–190, 2013.

J. S. Chu and J. A. Evans. Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences*, 118(41), 2021.

B. Devezer, D. J. Navarro, J. Vandekerckhove, and E. Ozge Buzbas. The case for formal methodology in scientific reform. *Royal Society open science*, 8(3):200805, 2020. Publisher: The Royal Society.

T. M. Errington, M. Mathur, C. K. Soderberg, A. Denis, N. Perfito, E. Iorns, and B. A. Nosek. Investigating the replicability of preclinical cancer biology. *eLife*, 10:e71601, Dec. 2021. ISSN 2050-084X. doi:10.7554/eLife.71601. URL `https://doi.org/10.7554/eLife.71601`. Publisher: eLife Sciences Publications, Ltd.

D. J. Glass. *Experimental design for biologists*. Number QH323. 5 G52. Cold Spring Harbor Laboratory Press Cold Spring Harbour, NY, USA, 2014.

L. Henderson. The Problem of Induction. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2020 edition, 2020.

R. A. Klein, M. Vianello, F. Hasselman, B. G. Adams, J. Adams, Reginald B, S. Alper, D. Vega, M. Aveyard, J. Axt, M. T. Babalola, and et al. Many labs 2: Investigating variation in replicability across sample and setting, Feb 2022. URL `osf.io/8cd4r`.

S. Leonelli. Rethinking reproducibility as a criterion for research quality. In *Including a symposium on Mary Morgan: curiosity, imagination, and surprise*. Emerald Publishing Limited, 2018.

R. Lowe, N. Shirley, M. Bleackley, S. Dolan, and T. Shafee. Transcriptomics technologies. *PLoS computational biology*, 13:e1005457, 2017.

S. Lydersen and M. Langaas. What proportion of published research findings are false? *Tidsskrift for Den norske legeforening*, 2021.

S. Schneegans, J. Lewis, and T. Straza. *UNESCO Science Report: The race against time for smarter development*, volume 2021. UNESCO Publishing, 2021.

C. Souto-Maior, Y. Lin, Y. L. Serrano Negron, and S. T. Harbison. Multiple shifts in gene network interactions shape phenotypes of drosophila melanogaster selected for long and short night sleep duration. page 2021.07.11.451943, 2021.