Simulation as a tool for formalising null hypotheses in cognitive science research

Aislinn Keogh (aislinn.keogh@ed.ac.uk) and **Elizabeth Pankratz (e.c.pankratz@ed.ac.uk)**
Centre for Language Evolution, University of Edinburgh, UK

Abstract

The default null hypothesis in typical statistical modelling software is that a parameter's value is equal to zero. However, this may not always correspond to the actual conditions that would hold if the effect of interest did not exist. In two case studies based on recent research in cognitive science and linguistics, we illustrate how data simulation can shed light on unspoken, sometimes even incorrect, assumptions about what the null hypothesis is. In particular, we consider information-theoretic measures of how learners regularise linguistic variability, where the null condition is not always equal to zero change, and an investigation of a cognitive bias for skewed distributions based on the assumption that, without such a bias, distributions would always remain uniform. All in all, simulating null conditions not only improves each researcher's understanding of their own analysis and results, but also contributes to the practice of "open theory". Formalising one's assumptions is, in itself, an important contribution to the scientific community.

Keywords: computational modelling; null hypothesis; open theory; random sampling; regularisation; information theory; cognitive bias; skewed distribution

Introduction

The established modern scientific process has at its core the practice of falsification (Popper, 1963), at least in the frequentist framework of null hypothesis testing (Fisher, 1925; Neyman & Pearson, 1928a, 1928b, 1933). In this framework, researchers use statistical tests to indicate how likely their observed data would be under the null hypothesis—that is to say, under the assumption that the effect of interest does not exist. If the data is sufficiently unlikely, then the null hypothesis may be rejected. But is the null hypothesis that we reject actually informative about our research question? Or put differently: does the actual null hypothesis for a given analysis align with what researchers assume it to be?

The null hypothesis is generally thought of as a claim of "no effect" or "no difference". But by simulating the outcomes that would emerge under those conditions, researchers may discover that the actual null hypothesis is quite different from what is typically assumed by statistical modelling software (e.g. that a parameter is equal to zero).

In this paper, we discuss two case studies based on recent work on language and cognition. Using these case studies, we illustrate how null hypotheses can be simulated and how this process can add crucial nuance to the analysis workflow. We intentionally draw these examples from different areas of enquiry to demonstrate the wide applicability of this approach and to showcase different options for carrying out the simulation procedure.

In the first case study, we consider the analysis of regularisation, a well-studied process whereby language users produce output that is less variable (on some dimension) than their input (Hudson Kam & Newport, 2005, 2009; Reali & Griffiths, 2009; Ferdinand et al., 2019; Smith & Wonnacott, 2010). The null hypothesis in regularisation experiments is that participants are probability matching: producing variants in proportion to their frequency in the input. We use simulation to illustrate that, under common information-theoretical measures of regularisation, probability matching does not always correspond to zero change between input and output (Samara et al., 2017; Ferdinand et al., 2019; Smith & Wonnacott, 2010; Keogh et al., 2024).

The second case study illustrates how a simple exemplar-based "urn model" (Spike et al., 2017) can be used to represent and simulate people's knowledge and use of linguistic items. We use an urn model to simulate plausible null hypotheses of experiments conducted by Shufaniya and Arnon (2022). Their research question is whether humans have a cognitive bias in favour of skewed frequency distributions. The experiments investigate whether uniform frequency distributions become skewed when reproduced by a single participant, and whether this tendency is amplified across multiple "generations" in an iterated chain (Kirby et al., 2008, 2014). The implicit null hypothesis is that, if there is no cognitive bias, frequency distributions should remain uniform. Our simulation indicates that this assumed null hypothesis does not hold and that the data presented by Shufaniya and Arnon should not be taken to support a cognitive bias for skew.

We now discuss each of these case studies in turn.

Case study 1: Change in information-theoretic measures of regularisation

When people are exposed to data that exhibits probabilistic or inconsistent variation and then asked to reproduce the data themselves, their reproductions often "smooth out" the inconsistencies that appeared in their input. In the case of linguistic variation, participants who learn a language where, for example, nouns appear randomly with different determiners will sometimes *regularise* the use of those determiners in their own output—either by producing one form to the exclusion of others (Hudson Kam & Newport, 2005, 2009; Schwab et al., 2018), or by specialising different forms for different contexts (Smith & Wonnacott, 2010; Samara et al., 2017).

The effect of interest in these studies is when participants produce certain forms with *higher* probability than predicted by the input. Therefore, the null hypothesis (and, in fact, the outcome that is often argued to be most common in adults) is that participants are *probability matching*: that is, the probability of each determiner in their output mirrors the input probabilities. There are several ways of comparing participants' output with the input to determine whether we can reject the null hypothesis of probability matching; here, we

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

consider a widely-used approach in which probability distributions over possible forms are used to compute information-theoretic measures.

Quantifying regularisation

In recent work, regularisation has increasingly been quantified using two information-theoretic measures: Shannon entropy and mutual information (Shannon, 1948; Samara et al., 2017; Ferdinand et al., 2019; Smith & Wonnacott, 2010; Keogh et al., 2024). Compared to a traditional frequency-based analysis of regularisation (Hudson Kam & Newport, 2005, 2009; Austin et al., 2022) in which the dependent variable is the change in the frequency of a certain form (usually the form that was most frequent in the input), the information-theoretic approach is sensitive to more subtle changes and makes fewer assumptions about which forms are becoming more frequent.

This approach takes as its basis a probability distribution over the possible variants in the language (e.g. the probability distribution over two possible determiners for a particular noun). The variability in that distribution is reflected in its Shannon entropy (H), the expected number of bits that would be required to encode an event in that distribution. A uniform probability distribution is highly unpredictable, hence highly variable, and has high entropy; a skewed probability distribution is much more predictable and therefore less variable, and has lower entropy.

The second measure, mutual information (I), captures the extent to which variability in a language is conditioned by some context. Languages that are highly variable overall, but vary in a predictable way (e.g. certain determiners are only ever used with certain nouns), have higher mutual information.

Figure 1 shows the relationship between these two measures and different probability distributions for four example languages that exhibit variation in determiner usage across a small set of nouns.[1]

One way to identify regularisation behaviour is to compare participants' output to their input using these measures. Specifically, a *decrease* in entropy corresponds to the situation where certain forms become more frequent at the expense of others. An *increase* in mutual information corresponds to the situation where variation is maintained at a global level but regularised in individual contexts.[2] But what can we take as evidence of a *significant* change in these measures? In standard statistical software, a parameter is considered "statistically significant" if an equal or more extreme value is sufficiently unlikely to be observed, assuming that

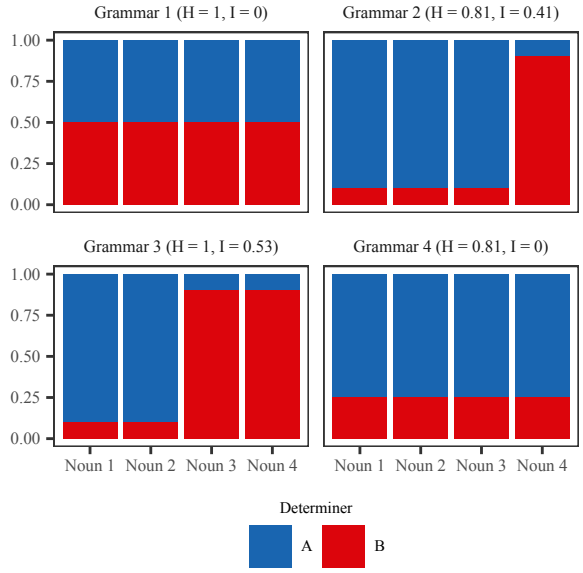

Figure 1: Example probability distributions of determiners over nouns, and their corresponding entropy (H) and mutual information (I) values. With only two determiners, as here, both measures have an upper bound of 1. Their lower bound is always 0. More skewed distributions overall correspond to lower entropy; lexical conditioning of variation (even probabilistic conditioning) corresponds to higher mutual information.

the parameter's true value is zero. However, a number of researchers have observed that zero change in entropy or mutual information is, in fact, *not* the expected outcome under the null hypothesis of probability matching (Samara et al., 2017; Ferdinand et al., 2019; Smith & Wonnacott, 2010; Keogh et al., 2024). Here, we describe in detail the simulation technique used by Keogh et al. and show how this approach can be used to derive an appropriate null hypothesis based on the properties of an arbitrary regularisation experiment.

Simulating a population of probability matchers

To determine the range of possible outcomes that can be encompassed by probability-matching behaviour, we simulate a population of participants who produce each determiner with probability proportional to its frequency in their input, taking each of the languages in Figure 1 as a possible input. Importantly, since different nouns correspond to different frequency distributions in some of these languages, this process requires us to specify two assumptions about probability matching.

First, we must decide whether we think participants are sampling from the probability distribution over determiners on a noun-by-noun basis, that is, in a context-specific way, or aggregating over the language as a whole. This decision has substantial implications for the possible outcomes under the null hypothesis of probability matching. To illustrate this, consider the case of a participant trained on a de-

[1] Two seemingly important cases are not illustrated: a language with no variation at all ($H = I = 0$), and one where individual nouns are completely consistent ($H = I > 0$). These turn out not to be interesting cases for simulation under certain assumptions about probability matching, which we describe in the next section.

[2] The other possible changes—an increase in entropy and a decrease in mutual information—do not reflect regularisation behaviour, so we do not discuss them here.

terministic grammar (unlike the ones in Figure 1) in which Nouns 1 and 2 always appear with determiner A, and Nouns 3 and 4 always appear with determiner B. If such a participant produced each determiner with probability corresponding to its overall frequency, we might not expect entropy to change much: each determiner would appear roughly 50% of the time in their output. However, we would expect mutual information to decrease considerably, since each noun would become more variable: on average, this strategy would produce an output language resembling Grammar 1 (50/50 split between determiners for every noun). By contrast, if the participant probability-matched on a noun-by-noun basis, we would expect exactly zero change in both measures, since random sampling with $p = 1$ (that is, deterministic production of a single variant) cannot distort the input at all: each noun would always be produced with the only determiner it had been observed with in the input.

The second assumption about probability matching we have to specify is whether we think participants encode the exact frequencies of their input, or whether there is some noise in this encoding. This is an important decision because, again, if some form appeared with $p = 1$ in the input and we use that exact value for random sampling, nothing will change in the output. If we add some noise and instead sample with, for example, $p = 0.95$, it would be possible for participants to produce a noun–determiner pairing they had never seen (although this would be rare).

Here, we simulate the case of context-specific probability matching, and make the simplifying assumption that participants encode the input frequencies exactly. Because regularisation experiments primarily focus on how learners treat unpredictable variation, and because perfect encoding of deterministic grammars will never lead to producing any variation at all, we simulate only input languages in which noun–determiner mappings are probabilistic.

Our input languages only contain two forms, so we can consider the generating process for the output data as sampling from a binomial distribution. We arbitrarily assign determiners to the two outcomes: A as 'success', and B as 'failure'. We use the `rbinom()` function in R (R Core Team, 2024) to generate data from 30 participants who each produce eight observations per noun (these numbers match the quantities in our experiment in Keogh et al., 2024; they should reflect the quantities in the experiment to be analysed.) The probability of 'success' on each trial is given by the frequency of determiner A with that noun in the input. For example, for all nouns in Grammar 1, $p = 0.5$; in Grammar 2, $p = 0.9$ for Nouns 1–3 and $p = 0.1$ for Noun 4. We then calculate the entropy and mutual information of each participant's output language and compare it to the entropy and mutual information of their input. We take the mean change in each measure, aggregated over participants, as the overall outcome of that experiment. We repeat this process 10,000 times to generate a distribution of expected experiment means for each measure under the null hypothesis of probability matching.

Results and analysis

Figure 2 shows the expected change in entropy (left) and mutual information (right) for each of the languages described in Figure 1 when all participants are probability matching. It is clear that entropy is far more likely to go down than up, while on average, mutual information *always* increases.

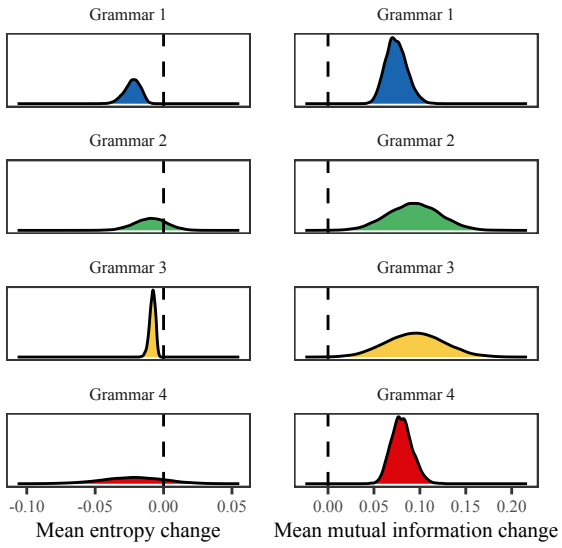

Figure 2: Expected change in entropy (left) and mutual information (right) between input (the four languages described in Figure 1) and output under the null hypothesis of probability matching. In almost all cases, probability matching gives rise to a non-zero change in one or both measures, a result that would be taken as evidence of regularisation by most standard statistical modelling software.

There are cases (Grammars 2 and 4) where zero entropy change is within the highest-density 95% interval of the null distribution; in these cases, the default null hypothesis assumed by standard statistical software would be a reasonable one. However, this is not always the case: in Grammar 1 particularly, zero is clearly well above the upper tail of the null distribution, meaning that we should always expect entropy to decrease given this kind of input.

One important factor contributing to this observation is that entropy and mutual information are bounded: with two variants, as we use here, the maximum value for both measures is 1. The minimum, regardless of the number of variants, is always 0. Since Grammars 1 and 3 have $H = 1$ in the input, entropy can therefore *never* increase. The same constraint holds in the other direction: for example, since $I = 0$ for Grammars 1 and 4, a participant could never produce output with lower mutual information than their input. In other words, if either measure is at its minimum or maximum value in the input language, then the measure can only change in one direction. A similar observation was made by Smith and Wonnacott (2010), who pointed out that participants trained on a lan-

guage exhibiting maximal entropy will always reduce entropy if they deviate at all from their input, which will necessarily correspond to a "significant" result in a standard linear model. This observation is also important because a 95% confidence interval around the mean change will, by nature, not cross zero. It is therefore erroneous, given maximum or minimum input measures, to observe that the 95% confidence interval of the mean change does not cross zero and therefore reject the null hypothesis that participants are probability-matching. If participants probability-match under the assumptions given here, we would in fact expect that the 95% confidence interval does not cross zero.

In sum, these simulations have shown that, when converting probabilities to information-theoretic measures, it is usually reasonable to assume that these measures will show non-zero change between participants' input data and the output they produce, even when participants are probability matching (the process which is typically understood as the absence of a regularisation effect). The exact specification of the null hypothesis depends both on the assumptions researchers make about how probabilities are encoded in learning, and on the properties of the input.

Case study 2: An urn model with no cognitive bias for skewed distributions

Iterated learning experiments (Griffiths & Kalish, 2007; Kirby et al., 2008; Reali & Griffiths, 2009; Kirby et al., 2014) can help shed light on the cognitive biases that give rise to language universals. One of the most striking commonalities between languages is that word frequencies follow a skewed, power-law distribution (Zipf, 1932, 1949), such that a word's frequency is inversely proportional to its rank. Using a series of iterated storytelling experiments, Shufaniya and Arnon (2022) argue that this feature of language arises from a cognitive bias in favour of skewed distributions, a preference which is amplified over the course of iterated learning.

In these experiments, participants are told a story that contains six novel words integrated into an English text. In the first participant's input, all novel words have the same (i.e. uniform) frequency. The participant must then re-tell this story, and that version of the story is shown to another participant. This process is repeated several times, with each participant's output serving as the input to the next participant, to create a transmission chain of 10 "generations". The authors count how often each participant uses each word and compute the Shannon entropy of the resulting distribution. They find that entropy decreases "significantly" over generations, indicating that the frequency distributions became increasingly skewed across transmission. The implied null hypothesis is that, if there is no cognitive bias for skew, the frequency distributions should remain uniform and thus entropy should not decrease.

Here, we replicate Studies 2 and 3 from Shufaniya and Arnon (2022) using a simple cognitive model. Crucially, the agents in our model have no cognitive bias for skew: they produce words by randomly sampling from their memory of their input. The original studies manipulate whether or not participants are required to use all six novel words in their retelling (the two conditions in Study 2), and whether those novel words must be recalled from memory (Study 2) or are provided during the story's re-telling (Study 3). We chose a model architecture that allows us to simulate all of these conditions.

Simulation using an urn model

Traditionally, an urn model is a representation of an agent as a collection of meaning–signal mappings (Spike et al., 2017; Keogh et al., 2024). Each meaning category has an urn, and every time a signal is received with that meaning, a token of the given signal is added to that urn. A token may be "forgotten", in which case it is removed from the urn. To produce a signal, a token is sampled at random from all tokens in the given urn.

What makes an urn model an appropriate choice to simulate the data from Shufaniya and Arnon (2022)? It is a very simple model that still incorporates all the capacities that the original experiments manipulate. Firstly, it is exemplar-based, so it easily allows us to find the frequency distributions of the exemplars it has stored. Secondly, it learns from input, and it produces output stochastically. And finally, it can have memory limitations.

In our simulation, since the novel words have no associated meanings, we simply place them all into a single urn that represents the category "novel word". In Generation 1, an agent receives as input all six novel words, each appearing eight times, like in the original experiment. The agent then produces 48 tokens by repeatedly sampling a single token from the urn of exemplars they have stored. This amounts to random sampling proportional to a token's frequency in memory. The 48 tokens an agent produces serve as input to a new agent in Generation 2, and so on, until Generation 10.

For the simulations in which type reduction is not permitted, we began each agent's output with one token of each of the six types, and then for the remaining 42, sampled tokens randomly from the urn as usual. For the simulations corresponding to the experiment in which participants must recall the novel words from memory (Study 2), we set the agents' memory limit at 24 tokens: less than the number of tokens they receive as input. By limiting agents' memory in this way, we incorporate Shufaniya and Arnon's observation that participants sometimes failed to remember the original words they received. Every time an agent who has already stored 24 exemplars receives a new token as input, they "forget" an old one so that the new one can be stored, effectively overwriting their previous memory. In our model, it is a random token that is overwritten each time.

This model formalises the null hypothesis of a stochastic outcome with no explicit cognitive bias in favour of skewed distributions. Each agent produces tokens randomly, proportional to their frequency in memory. We simulated 1,000

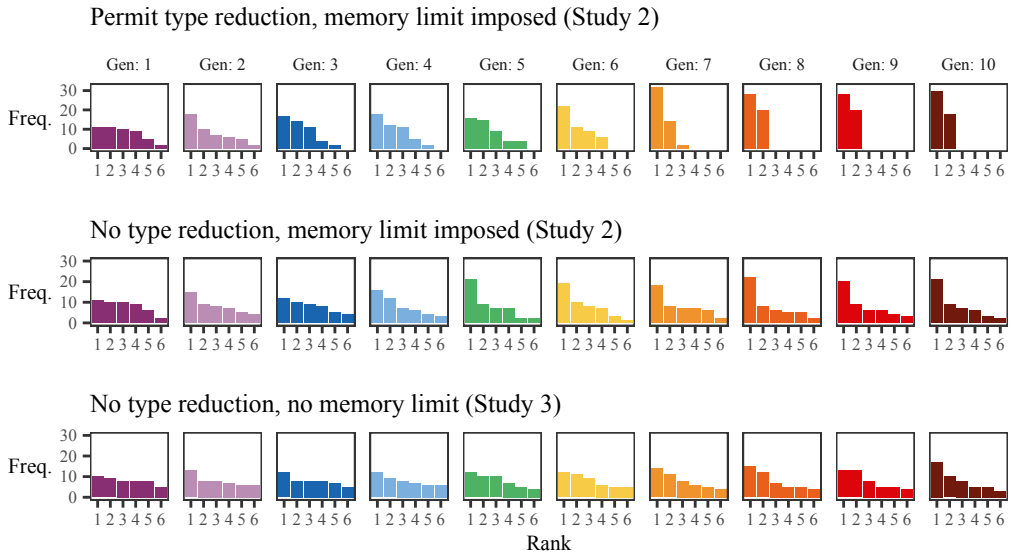

Figure 3: Example frequency distributions from a single simulated transmission chain from each of the three simulated datasets. Regardless of type reduction and memory limit, distributions become increasingly skewed through transmission.

chains of 10 generations each for each combination of variables tested by Shufaniya and Arnon's Studies 2 and 3.

Results and analysis

Example frequency distributions over the six novel words are illustrated in Figure 3. As in the original paper, we computed the entropy of the resulting frequency distributions at each generation for each chain. Figure 4 shows that entropy decreases across generations in all three simulated datasets. Qualitatively, these results closely resemble the plots shown in Shufaniya and Arnon (2022).

We analysed the simulated data by applying the same analyses as the original paper: one linear mixed effects model fit to the data in each panel of Figure 4. Each model predicts entropy as a function of generation (centered), with random intercepts by chain and random slopes over centered generation by chain (Shufaniya & Arnon, 2022, 66). The models were fit in R (R Core Team, 2024) using the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015).

The parameter of interest is the slope coefficient, which estimates how much entropy changes with an increase of one generation. Table 1 shows the original coefficient estimates and standard errors for this parameter reported in Shufaniya and Arnon (2022), along with the ones from our simulated data. The coefficient estimates are strikingly similar, nearly identical.[3]

The claim in Shufaniya and Arnon (2022) is that the null hypothesis can be rejected: entropy decreases significantly across generations, which is a sufficiently unlikely outcome under the assumed null hypothesis that frequency distributions will remain uniform, and that entropy will not change. But our simulations have shown that this null hypothesis is not accurate. If we follow the common understanding that the null hypothesis represents random chance, we would in fact expect the frequency distributions over six novel words to become increasingly skewed over time. Admittedly, it is not likely that participants in Shufaniya and Arnon's storytelling task are randomly sampling from a distribution of tokens. It is more likely that there are other pressures contributing to an increase in skew, such as certain objects showing up frequently due to the typical properties of narratives (i.e. stories are *about* something, and that something is probably more frequent than other, more peripheral, things). Nonetheless, this simulation has shown that it is premature to conclude that there is a cognitive bias in favour of skewed distributions without further experimental work and an analysis that takes the tendency we observe here into account.

Discussion and conclusion

With these two case studies, we have illustrated that assuming a null hypothesis of zero change can lead to inaccurate conclusions. Specifically, in regularisation experiments, the null hypothesis of probability matching does not necessarily correspond to zero change in the dependent variable between input and output (Samara et al., 2017; Ferdinand et al., 2019; Smith & Wonnacott, 2010; Keogh et al., 2024). Rather, the expected outcome of such a strategy depends both on one's assumptions about probability matching and on the properties of the input. And in an iterated storytelling experiment (Shufaniya & Arnon, 2022), the implicit null hypothesis that frequency distributions should retain the form they started

[3]The smaller standard errors produced by our models are because, where Shufaniya and Arnon analysed data from five chains, we generated data from one thousand.

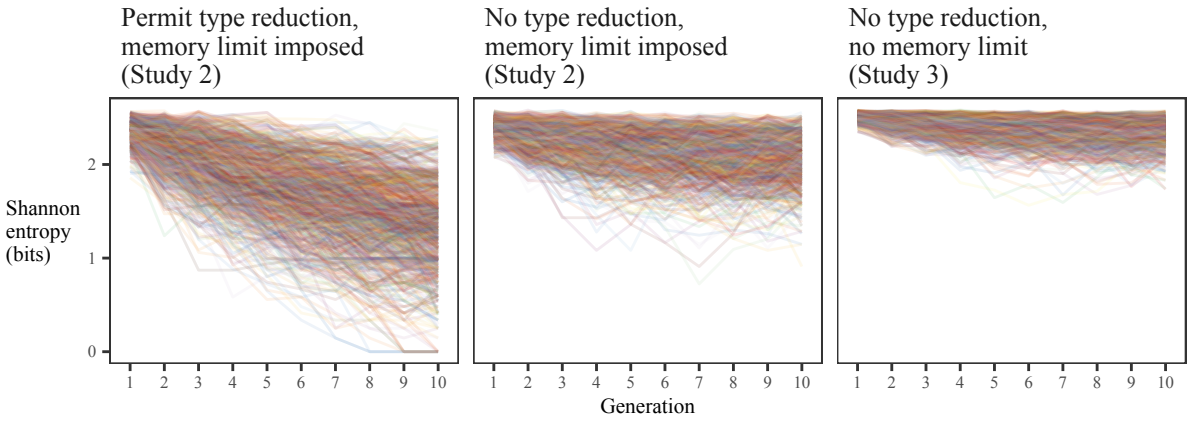

Figure 4: The Shannon entropy of the frequency distribution produced by each simulated chain (N = 1000) at every generation. Entropy decreases most dramatically when type reduction is permitted, but even when it is not, decreases are observed.

	Shufaniya and Arnon (2022)		**Urn model**	
Study 2 (permit type reduction, memory limit)	β = –0.11	SE = 0.013	β = –0.11	SE = 0.002
Study 2 (no type reduction, memory limit)	β = –0.02	SE = 0.01	β = –0.03	SE = 0.001
Study 3 (no type reduction, no memory limit)	β = –0.02	SE = 0.01	β = –0.02	SE = 0.0005

Table 1: Slope coefficient estimates (β) and standard errors (SE) from linear mixed effects models estimating how the Shannon entropy of a frequency distribution changes when increasing the generation by one. The urn model, sampling randomly, shows the same effects observed in the original experiments.

with is not accurate; random sampling alone gives rise to increasingly skewed distributions through transmission.

Simulating null hypotheses leads in several ways to more robust statistical practice. In both the case studies we present here, the results of our simulations would make us more conservative about viewing an outcome as extreme enough to reject the null hypothesis. In this way, we can be even more vigilant about avoiding p-hacking (Head et al., 2015), and we reduce the risk of incorrectly rejecting—or incorrectly failing to reject—an unsuitable null hypothesis. Of course, there may be cases where the default null hypothesis of zero (change in some measure, or difference between groups) turns out to be appropriate. Even in such cases, we would argue that the simulation process is a valuable one to give researchers greater confidence in their analysis.

Readers will note that we stop short of proposing a single alternative analysis technique that can be adopted whenever a more traditional method may make the wrong assumptions. This is because the simulation approach we advocate can be applied so widely that there is no one-size-fits-all analysis to suit every experiment and every null hypothesis. Rather, we would argue that every researcher is the best person to identify an appropriate solution for their specific context.

Nonetheless, our general approach aligns with the growing call for the practice of "open theory" (Guest & Martin, 2021; Lakens & DeBruine, 2021). Simulating null hypotheses is one example of how researchers can "explicitly document ... what their theory assumes" (Guest & Martin, 2021, 2). Not only does this help researchers to better understand their own practice, it also enables the wider research community to interpret findings in light of the researcher's assumptions.

In conclusion, simulating null hypotheses allows researchers to develop a deeper understanding of their research question and analysis, and ultimately to better understand and more accurately interpret the results of their research.

Acknowledgments

AK and EP both gratefully acknowledge funding from the Economic and Social Research Council (ES/P000681/1). EP is also supported by the Social Sciences and Humanities Research Council of Canada (752-2021-0366).

References

Austin, A. C., Schuler, K. D., Furlong, S., & Newport, E. L. (2022). Learning a Language from Inconsistent Input: Regularization in Child and Adult Learners. *Language Learning and Development*, *18*(3), 249–277. doi: 10.1080/15475441.2021.1954927

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Ferdinand, V., Kirby, S., & Smith, K. (2019). The cognitive roots of regularization in language. *Cognition*, *184*, 53–68. doi: 10.1016/j.cognition.2018.12.002

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*(3), 441–480. doi: 10.1080/15326900701326576

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*. doi: 10.1177/1745691620970585

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLOS Biology*, *13*(3), 1-15. doi: 10.1371/journal.pbio.1002106

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*(2), 151–195. doi: 10.1080/15475441.2005.9684215

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, *59*(1), 30–66. doi: 10.1016/j.cogpsych.2009.01.001

Keogh, A., Kirby, S., & Culbertson, J. (2024). Predictability and variation in language are differentially affected by learning and production. *Cognitive Science*, *48*(4), e13435. doi: 10.1111/cogs.13435

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686. doi: 10.1073/pnas.0707835105

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114. doi: 10.1016/j.conb.2014.07.014

Lakens, D., & DeBruine, L. M. (2021). Improving transparency, falsifiability, and rigor by making hypothesis tests machine-readable. *Advances in Methods and Practices in Psychological Science*, *4*(2). doi: 10.1177/2515245920970949

Neyman, J., & Pearson, E. S. (1928a). On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I. *Biometrika*, *20A*(1/2), 175–240.

Neyman, J., & Pearson, E. S. (1928b). On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part II. *Biometrika*, *20A*(3/4), 263–294.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *231*(694-706), 289–337.

Popper, K. (1963). Science as falsification. In *Conjectures and Refutations* (pp. 33–39). London: Routledge and Keagan Paul.

R Core Team. (2024). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317–328. doi: 10.1016/j.cognition.2009.02.012

Samara, A., Smith, K., Brown, H., & Wonnacott, E. (2017). Acquiring variation in an artificial language: Children and adults are sensitive to socially conditioned linguistic variation. *Cognitive Psychology*, *94*, 85–114. doi: 10.1016/j.cogpsych.2017.02.004

Schwab, J. F., Lew-Williams, C., & Goldberg, A. (2018). When regularization gets it wrong: children over-simplify language input only in production. *Journal of child language*, *45*(5), 1054–1072. doi: 10.1017/S0305000918000041

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. doi: 10.1002/J.1538-7305.1948.TB01338.X

Shufaniya, A., & Arnon, I. (2022). A cognitive bias for Zipfian distributions? Uniform distributions become more skewed via cultural transmission. *Journal of Language Evolution*, *7*(1), 59–80. doi: 10.1093/jole/lzac005

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*(3), 444–449. doi: 10.1016/j.cognition.2010.06.004

Spike, M., Stadler, K., Kirby, S., & Smith, K. (2017, April). Minimal Requirements for the Emergence of Learned Signaling. *Cognitive Science*, *41*(3), 623–658. doi: 10.1111/cogs.12351

Zipf, G. K. (1932). *Selected studies of the principle of relative frequency in language*. Cambridge, MA and London, England: Harvard University Press. doi: doi:10.4159/harvard.9780674434929

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, Mass.: Addison-Wesley Press, Inc.