Identifiability, log-linear models, and Observed/Expected (response to Stanton & Stanton, 2022)

Colin Wilson, Johns Hopkins University, `colin.wilson@jhu.edu`

In commenting on Wilson and Obdeyn (2009), the squib by Stanton and Stanton (2022) makes several errors of interpretation and practice. It mischaracterizes a brief argument given in Wilson and Obdeyn (2009) as incomplete or obviously false, fails to observe basic principles of statistical inference, overlooks the role of regularization and model comparison in the original paper, and blurs the distinction between grammatical hypotheses and statistical quantities. The positive contribution of Stanton and Stanton (2022) concerning Observed/Expected is a simple algebraic identity.

These notes provide some background on how log-linear models can be coherently used to formalize and compare grammatical hypotheses, with many references to comprehensive textbooks and other introductory material. The Appendix gives a proof of general conditions, including the specific case considered by Wilson and Obdeyn (2009), under which Observed/Expected exaggerates the strength of OCP constraints on combinations of similar or identical segments that are independently lower in probability. The on-line `supplementary material` contains R code for deriving O/E values, fitting and comparing simple log-linear models, and performing other relevant computations mentioned here.[*]

1. Parametric statistical models

A parametric statistical model specifies a probability distribution using a function $p(x;\theta)$, where x is a random variable and θ is a finite vector of real-valued parameters.[1] The simplest example is the Bernoulli distribution, **Bernoulli**$(x;\pi)$, where $x \in [0,1]$ and the single parameter $\pi \in [0,1]$ is the probability that x is equal to 1. Another common example is the univariate Gaussian or normal distribution, **Normal**$(x;\mu,\sigma)$, where $x \in \mathbb{R}$ and the two parameters are the mean μ and the standard deviation $\sigma > 0$. The log-linear distributions studied here generalize the Bernoulli distribution to multiple outcomes that have internal structure. These models have the form $\log p(x;\lambda) = \left[\sum_{k=1}^{K} \lambda_k \cdot f_k(x)\right] - \log Z$, where each $f_k(x)$ is fixed a real-valued

[*] **Acknowledgments.** Thanks to Gaja Jarosz for questions about an early version of Wilson and Obdeyn (2009) that motivated the proof in the Appendix. This research was supported by NSF grant BCS-1941593.

[1] When the set of possible values of the random variable x is discrete or countably infinite, as is the case for the Bernoulli and log-linear distributions, $p(x;\theta)$ is a probability mass function. When the set of possible values is continuous (non-countably infinite), as for the Normal distribution, $p(x;\theta)$ is a probability density function. The distinction between mass and density functions is not needed here, so the probability distribution of a model will be identified with $p(x;\theta)$. For many technical details throughout, see the early chapters of Wasserman (2004); Gelman and Hill (2006); Bishop (2006); Gelman et al. (2013); Murphy (2014); Kruschke (2015), or other textbooks on statistical inference and machine learning.

function, each λ_k is a parameter, and Z is a normalizing constant (not a parameter) that ensures the probability distribution sums to one over all of the possible outcomes.[2]

1.1 Identifiability and maximum-likelihood estimation

A parametric statistical model is *identifiable* if different parameter values necessarily specify different probability distributions: that is, if the relation between θ and $p(x;\theta)$ is one-to-one.[3] To illustrate, consider a data set of real numbers that average to 42. Researcher A analyzes the data using a normal distribution with the typical parameterization $\mathbf{Normal}(x;\mu,\sigma)$ and obtains the single estimate $\mu = 42$. Researcher B analyzes the same data with the parameterization $\mathbf{Normal}(x;\mu_1+\mu_2,\sigma)$ and concludes that there are in fact infinitely many possible estimates: any μ_1 and μ_2 that add up to 42 specify the same probability distribution and are equivalent with respect to the model and the data. Researcher C analyzes the data with $\mathbf{Normal}(x;\mu_1\cdot\mu_2,\sigma)$ and also concludes that there are infinitely many possible estimates, though different ones.

The first model is identifiable, and the sample average is the *maximum-likelihood* (ML) *estimator* of the mean parameter μ.[4] Under very general conditions, known as *regularity conditions*, maximum-likelihood estimators exist and are *consistent*: they become arbitrarily close to the true population parameters — under the assumption that the model is correct — as the size of the data set increases. One of the regularity conditions is that the model must be identifiable. This condition does not hold for the other two models, therefore ML estimates do not strictly speaking exist for their parameters and consistency is out of the question. By including redundant parameters, more than are needed to specify the probability distribution, researchers B and C have created unnecessary confusion and their conclusions differ from those of A and from each other for spurious reasons.

Identifiability is a basic concept of parametric statistical modeling, and has counterparts in linguistic methodology. Linguistic theories and analyses quite generally aim for parsimony and eschew redundant rules, parameters, or constraints. Even when trivial duplication has been avoided, particular data sets can underdetermine the ordering of rules, setting of parameters, or ranking/weighting of constraints. Indeterminices of this sort are well-known, clearly flagged in the literature, and unproblematic for many research goals (e.g., Hayes et al., 2003; Bane and Riggle, 2008) In some cases, they motivate efforts to further simplify the theory or to collect disambiguating data.

[2] The Bernoulli distribution can also be written as a single-parameter log-linear model, $\log p(x=1) = \lambda - Z$. This is the form that is fit by the function `glm(er)` with `family = binomial(link = "logit")` in R (R Core Team, 2013) or by the Stan funtion `bernoulli_logit` (Carpenter et al., 2017), which use $p(x=1) = \exp(\lambda)/(\exp(\lambda)+1) = 1/(1+\exp(-\lambda))$.

[3] The terminology surrounding identifiability varies somewhat across sources. A more granular approach defines a particular parameter of a model as identifiable if changing its value necessarily changes the distribution, and an entire model (or parameter vector) as identifiable if each of its parameters has this property. McCullagh and Nelder (1989) talk of *aliasing* instead of identifiability, and make a distinction between parameters and estimators that is more subtle than needed here. These various notions are ‘internal’, pertaining to the functional definition of a model. There are related ‘external’ issues that arise from collinearity of predictors (e.g. Gelman and Hill, 2006; Baayen, 2008) and zero counts or other aspects of the data sample (e.g. Fienberg and Rinaldo, 2012). Fienberg and Rinaldo (2007) review the history of log-linear models, highlighting the large body of research on determining when ML parameter estimates exist.

[4] Myung (2003) is an introductory level tutorial on ML estimation.

1.2 A nonidentifiable log-linear model

In parametric statistical modeling, fitting a nonidentifiable model to data by maximizing the likelihood function — and then comparing the resulting coefficients for equality to hypothetical population values, reporting z-scores and p-values for them, and so forth — is an unforced error. This error is made in Stanton and Stanton (2022) [SS], see Table 2 (p. 6) and Table 3 (p. 11) of that paper with surrounding discussion; it was not made or suggested in Wilson and Obdeyn (2009) [WO].

Both SS and WO consider log-linear models for categorical data, such as 3×3 tables of consonant cooccurrence counts. The particular model used by SS, but not by WO, has the free parameters shown in Table 1 (compare to the design matrix in Table 1, p. 5 of SS).

	p	t	k
p	$\lambda_{p1} + \lambda_{p2} + \lambda_{p1p2}$	$\lambda_{p1} + \lambda_{t2} + \lambda_{p1t2}$	$\lambda_{p1} + \lambda_{k2} + \lambda_{p1k2}$
t	$\lambda_{t1} + \lambda_{p2} + \lambda_{t1p2}$	$\lambda_{t1} + \lambda_{t2} + \lambda_{t1t2}$	$\lambda_{t1} + \lambda_{k2} + \lambda_{t1k2}$
k	$\lambda_{k1} + \lambda_{p2} + \lambda_{k1p2}$	$\lambda_{k1} + \lambda_{t2} + \lambda_{k1t2}$	$\lambda_{k1} + \lambda_{k2} + \lambda_{k1k2}$

Table 1

Each parameter λ_α in the table can be thought of as the weight of a soft grammatical constraint in the probabilistic version of Harmonic Grammar (HG; Legendre et al., 1990a,b). For example, λ_{k1k2} can be regarded as the weight of the HG constraint that is violated by coocurrence of two instances of /k/ (strict OCP) or any two dorsal consonants (OCP-Place). The log-linear model with these parameters defines the probability of combination x_1y_2 as $\log p(x_1y_2;\lambda) = \lambda_{x1} + \lambda_{y2} + \lambda_{x1y2} - \log Z$. Equivalently, the probability of x_1y_2 is $p(x_1y_2;\lambda) = p^*(x_1y_2;\lambda)/Z$, where $p^*(x_1y_2) = \exp(\lambda_{x1} + \lambda_{y2} + \lambda_{x1y2})$ and Z is the sum of $p^*(x_1y_2;\lambda)$ over all of the possible combinations (i.e., over all cells in the table).[5]

If each λ_α is mapped to $\theta_\alpha = \exp(\lambda_\alpha)$, the probability distribution defined by this model can also be written as follows (with Z implicit and defined as before):

	p	t	k
p	$\theta_{p1} \cdot \theta_{p2} \cdot \theta_{p1p2}$	$\theta_{p1} \cdot \theta_{t2} \cdot \theta_{p1t2}$	$\theta_{p1} \cdot \theta_{k2} \cdot \theta_{p1k2}$
t	$\theta_{t1} \cdot \theta_{p2} \cdot \theta_{t1p2}$	$\theta_{t1} \cdot \theta_{t2} \cdot \theta_{t1t2}$	$\theta_{t1} \cdot \theta_{k2} \cdot \theta_{t1k2}$
k	$\theta_{k1} \cdot \theta_{p2} \cdot \theta_{k1p2}$	$\theta_{k1} \cdot \theta_{t2} \cdot \theta_{k1t2}$	$\theta_{k1} \cdot \theta_{k2} \cdot \theta_{k1k2}$

Table 2

In either form, the model that SS work with is highly over-parameterized and nonidentifiable.[6] We could remove all of the θ_{x1} and θ_{y2} terms (that is, set them equal to one) and define the same probability distribution using only the θ_{x1y2} parameters (actually only eight of them are needed to perfectly fit any data set, with no intercept). Or we could fix the parameters on the diagonal to whatever values we like and adjust some of the others as needed by the data. There are infinitely many equivalent possibilities, and no reason to dwell on their differences.

[5] See Goldwater and Johnson (2003) on the relationship between HG (which they refer to as Optimality Theory) and log-linear models. This relationship was already present in the founding work on Harmony Theory (Smolensky, 1986), a general theory of cognition of which Harmonic Grammar is the leading application.

[6] As it was entered into R the model in fact contained one more parameter, an 'intercept' that is constant across all cells.

Maximum-likelihood estimates do not technically exist for this model, and no interpretation or significance should be assigned to the coefficients that result from fitting it to data. Indeed, software such as R will refuse to provide estimates for the entire vector of parameters and instead drop some of them prior to fitting (witness the NAs and absurdly small coefficients in Tables 2 and 3 of SS). Which parameters are dropped depends on the order of terms in the equation for fitting (i.e., the order of columns in the design matrix). Of course this order has no statistical, let alone grammatical, importance; and the resulting simplified model is still 'fully saturated', able to perfectly fit any possible data set. Entering a nonidentifiable model into `glm` and letting R sort out a minimal parameterization is bad practice, to be avoided except when the resulting coefficients are irrelevant for the purpose of the analysis.

Nonidentifiable models like those in Table 1 / Table 2 can be avoided in one of several standard ways. First, the model can be augmented with *regularization* terms that usually prefer either small parameter values or few non-zero values. This approach is typical of machine learning, where interest lies in the test performance of the model as a whole, not in particular values of its parameters, and there is concern about overfitting to the training data. Regularization is a general method for reducing the set of possible estimates (ideally to one) for a nonidentifiable model.[7] Second, and closely related, it is possible to place a prior distribution on the parameters and replace ML estimation with Bayesian inference. While one can find particular values that jointly maximize the likelihood and the prior (the *maximum a-posteriori* or MAP estimates), Bayesian inference more naturally focuses on the entire posterior distribution on the parameters rather than specific numbers. This is the approach taken in WO, as discussed in section 2 below.[8]

Third, as is common in frequentist statistics, we can place *identifiability constraints* on the parameters. The constraints may take the form of fixing certain parameters to 0 ('treatment' coding) or of requiring certain parameters to sum to zero ('sum', or 'effect', or 'contrast' coding). The issue of (non)identifiability of log-linear models, and the method of imposing identifiability constraints, is covered in introductory textbooks (e.g. Agresti, 2002, 2019; Wickens, 1989; Wasserman, 2004). If the constraints on the parameters are insufficient to make the model identifiable, there can be no discussion of whether ML estimation has "succeeded" or "failed" at finding "correct" parameter values — just as there could not be a debate about whether 42 should be broken down into $40+2$ or $21+21$. And there is little reason to demonstrate nonidentifiability for special cases, as section 3.3 of SS does for a 3×3 table, when much more general results have been shown formally and in easily accessible sources.

1.3 An identifiable log-linear model

In the short section that SS comment on, WO considered the hypothetical distribution in Table 3. Here the probability of each consonant combination is determined by the product of a first consonant (row) parameter θ_{x1}, a second consonant (column) parameter θ_{x2}, and a parameter θ_{OCP} that is $1/2$ if the two consonants have the same place of articulation (and implicitly 1 otherwise).

[7] See any of the machine learning texbooks cited earlier or, for a more sophisticated treatment, chapters 13 and 14 of Smolensky et al. (1996).

[8] Both regularization and prior distributions are ways of formalizing domain-general or language-specific inductive learning biases (e.g., Goldwater and Johnson, 2003; Wilson, 2006; Culbertson et al., 2013; Cotterell et al., 2015; White, 2017).

	p 1/3	t 1/2	k 1/6
p 1/3	1/3·1/3·**1/2**	1/3·1/2	1/3·1/6
t 1/2	1/2·1/3	1/2·1/2·**1/2**	1/2·1/6
k 1/6	1/6·1/3	1/6·1/2	1/6·1/6·**1/2**

Table 3

The particular values assigned to these parameters are not relevant, WO gave them only for the purpose of definite illustration. Instead, the table is of interest because of its qualitative structure. It formalizes the restrictive grammatical hypothesis that, while consonants of different places of articulation may vary in their underlying probabilities, the strength of the OCP or OCP-Place constraint is exactly the same across labial, coronal, and dorsal combinations. WO showed that, for the most probable sample of counts from this hypothetical distribution, the O/E values for the cells on the diagonal are not the same:

	p	t	k
p	**0.58**	1.29	1.05
t	1.29	**0.72**	1.17
k	1.05	1.17	**0.48**

Table 4

If smaller O/E values on diagonal cells are taken to indicate stronger application of the OCP, here they will misleadingly provide evidence against the restrictive hypothesis even though it is true by construction. In particular, the O/E values incorrectly suggest that the OCP applies more strongly to consonants that are independently lower in probability. The Appendix proves fairly general conditions under which O/E values distort OCP strengths when interpreted in this way.

* * *

Is the log-linear model suggested by Table 3 identifiable? Does the answer matter at all for the point made in WO? The model has many fewer free parameters than the one assumed by SS. The off-diagonal factors θ_{x1y2} ($x_1 \neq y_2$) are all fixed to one (i.e., according to this hypothesis there are no grammatical constraints on combinations of consonants of different places) and the diagonal factors θ_{x1x2} are all equal to θ_{OCP}. Notice further that the row parameters θ_{x1} sum to one, as do the column parameters θ_{x2}. The model *with these constraints imposed* is identifiable. (This can be verified by checking that the columns of the corresponding design matrix, unlike those of Table 1 in SS, are linearly independent.)

The sum-to-one identifiability constraints are conceptually appealing because $(\theta_{p1}, \theta_{t1}, \theta_{k1})$ can be immediately interpreted as a probability distribution over consonants in the first position of a combination — the distribution that would hold if, counterfactually, the OCP were inert (i.e., if $\theta_{OCP} = 1$). The same applies to $(\theta_{p2}, \theta_{t2}, \theta_{k2})$ for the second member of a combination. It follows that the table would be automatically normalized (i.e., all of the cell probabilities would add up to one) absent any OCP interference.

Any log-linear model with a full set of row and column parameters can be made to satisfy the sum-to-one constraints by normalizing each dimension separately. Clearly, it does not make sense to ask which of the row or column parameters should be fit to data and which should be determined from the other two, just as it is pointless to ponder whether $42-2=40$ or $42-40=2$.

The model with these identifiability constraints is not, however, the easiest one to specify and fit in software such as R. It is easier to work with an alternative parameter vector θ' in which $\theta'_{x1}=1$ (resp. $\theta'_{x2}=1$) for some choice of x_1 (resp. x_2). This is equivalent to the constraint that $\lambda'_{x1}=0$ (resp. $\lambda'_{x2}=0$). Therefore, the model can be fit by simply omitting the corresponding columns of the design matrix. In the spirit of coronal unmarkedness (though the choice is irrelevant), we can can set $\theta'_{t1}=\theta'_{t2}=1$, then multiply the other θ'_{x1} and θ'_{x2} terms by 2 (i.e., divide them by $\theta_{t1}=\theta_{t2}=1/2$). This yields a table that specifies exactly the same probability distribution using the minimum number of θ'_α parameters that are not equal to one (i.e., the minimum number of non-zero λ'_α parameters):

	p 2/3	t	k 1/3
p 2/3	2/3·2/3·**1/2**	2/3	2/3·1/3
t	2/3	**1/2**	1/3
k 1/3	1/3·2/3	1/3	1/3·1/3·**1/2**

Table 5

R code is provided in the supplementary material to specify the constrained design matrix of this model and perform ML estimation of its λ' parameters given idealized data. To recover the θ' parameters, set $\theta'=\exp(\lambda')$. To recover the θ parameters of Table 3, normalize the θ'_{x1} estimates and, separately, the θ'_{x2} estimates. There are unique ML estimates for each set of identifiability constraints, but again no basis for arguing about which constraints are the "correct" ones.

2. Log-linear model comparison

As correctly noted in section 2.2 of SS, WO did not attempt to show that ML estimation of any log-linear model would recover underlying population parameters such as those in Table 3 / Table 5. This could not be expected for the over-parameterized model that SS adopt. It would be possible to show, any sampling issues aside, for models with fewer parameters or identifiability constraints. However, recovery of specific parameter values is not very central to the goal of comparing grammatical hypotheses that differ in restrictiveness (or simplicity, parsimony). This was the main project of WO, which demonstrated how it could be carried out rigorously with log-linear models.

Consider how we could compare two hypotheses about the OCP given an observed sample of consonant cooccurrence frequencies. According to the more restrictive hypothesis, the OCP has the same strength across all places of articulation (as in Table 3, but focus on the structure of the parameters not their numerical values). According to the less restrictive hypothesis, the OCP may be stronger for some places than others. These hypotheses can be formalized as two log-linear models — identifiable or not — that differ minimally in their parameterizations; see Table 6 and Table 7 below.

Common methods for quantitatively comparing models such as these use the *likelihood-ratio test* for nested models (LRT), AIC values, or BIC values.[9] To apply these methods, we need the maximum possible

[9] See McCullagh and Nelder (1989, Appendix A), Wickens (1989, section 4.5), Casella and Berger (2002, chapter 8), Wasser-

	p	t	k
p	$\lambda_{p1} + \lambda_{p2} + \lambda_{OCP}$	$\lambda_{p1} + \lambda_{t2}$	$\lambda_{p1} + \lambda_{k2}$
t	$\lambda_{t1} + \lambda_{p2}$	$\lambda_{t1} + \lambda_{t2} + \lambda_{OCP}$	$\lambda_{t1} + \lambda_{k2}$
k	$\lambda_{k1} + \lambda_{p2}$	$\lambda_{k1} + \lambda_{t2}$	$\lambda_{k1} + \lambda_{k2} + \lambda_{OCP}$

Table 6

	p	t	k
p	$\lambda_{p1} + \lambda_{p2} + \lambda_{OCP\text{-}Lab}$	$\lambda_{p1} + \lambda_{t2}$	$\lambda_{p1} + \lambda_{k2}$
t	$\lambda_{t1} + \lambda_{p2}$	$\lambda_{t1} + \lambda_{t2} + \lambda_{OCP\text{-}Cor}$	$\lambda_{t1} + \lambda_{k2}$
k	$\lambda_{k1} + \lambda_{p2}$	$\lambda_{k1} + \lambda_{t2}$	$\lambda_{k1} + \lambda_{k2} + \lambda_{OCP\text{-}Dor}$

Table 7

value of the likelihood function for each model given the data. However, the methods do not use the resulting parameter estimates or require them to be unique. They compare models solely on the basis of their maximum likelihood values (not their ML estimates), the number of free parameters that they contain, and (in the case of BIC) the number of data points in the sample. For nonidentifiable models, *the* ML parameter estimates do not exist, but *some* parameter estimates that achieve maximum likelihood do (under appropriate regularity conditions) and any such values will do. Concerns about identifying underlying population values are not relevant here, what matters is the grammatical structure formalized by each hypothesis and its restrictiveness relative to alternatives.

WO adopted a different method of model comparison, one based on the *Laplace approximation* to the posterior distribution $p(M|D)$, where M is a parametric statistical model and D is a data set.[10] This method explicitly does not seek particular parameters values (ML are otherwise). Instead, it approximately integrates over all possible values given the likelihood function and a prior distribution on the parameters. The prior distribution used by WO has the form of a normal distribution with a mean of zero and the same standard deviation for each parameter.[11]

The supplementary material includes code to compute likelihood-ratio tests, AIC, BIC, and Laplace approximation values for these two modesl. When applied to the most probable sample from the distribution of Table 6, these model comparison methods uniformly favor the more restrictive hypothesis (in this case also the true one) in which the OCP has the same strength for all places of articulation. When applied to actual samples of cooccurrence counts, for which the underlying population is of course unknown, the methods provide conceptually sound and quantitative evidence that can help distinguish among alternative grammatical theories.

The original mistake of SS, from which the others follow, is a failure to appreciate the importance of model comparison in WO and applications of parametric statistical models quite generally. SS refers in several places to “the” log-linear model, meaning the one in Table 1 that is incorrectly attributed to WO (see

man (2004, pp. 164 and section 13.6), among many other sources. The fact that the first model is nested within the second can be made explicit by rewriting the OCP parameters of the latter, for example as $\lambda_{OCP\text{-}Cor} = \lambda'_{OCP}$, $\lambda_{OCP\text{-}Lab} = \lambda'_{OCP} + \lambda'_{OCP\text{-}Lab}$, and $\lambda_{OCP\text{-}Dor} = \lambda'_{OCP} + \lambda'_{OCP\text{-}Dor}$.

[10] For introductions to the Laplace approximation, see for example MacKay (2004, p. 301) or Bishop (2006, pp. 214-217).

[11] A lower value of the standard deviation hyperparameter corresponds to a stronger prior preference for parameter values that are close to zero. This prior corresponds to ℓ_2 regularization and is very widely used.

SS, pp. 2, 15). But log-linear models are a family, not an individual. Members of the family differ in their restrictiveness and simplicity, just as systems of grammatical rules, parameters, or constraints differ. Model comparison is a rigorous way of determining the degree to which the data is consistent with a restrictive hypothesis or provides evidence for a richer, more permissive alternative.

3. On defending O/E

SS criticizes the argument that it attributes to WO as "lacking rigor" (p. 9), "weak" (p. 13), and "flawed" (p. 15). But WO neither asserted nor needed to assert that there is a unique way of fitting the log-linear model in Table 1 / Table 2 to cooccurrence counts. We hope that noone would try to make this argument, which any textbook on the subject will reject in passing and which plainly defies common sense. There could not possibly be a unique way of specifying 9 cell probabilities with 15 (really, any more than 8) free parameters. Like a hypothetical researcher who attempts to fit a mean with two parameters instead of one, SS has invented a quandry that is both unnecessary and irrelevant.

What about the defense of O/E that SS offers? The positive result in its section 3.1, showing that O/E values are a possible "solution" to the nonidentifiable model assumed there, follows from basic algebra. SS does not provide the general form of the result, instead discussing a particular example, so it is given here.

Suppose we have a contingency table with cell probabilities $p_{x1y2} = \theta_{x1} \cdot \theta_{y1} \cdot \theta_{x1y2}$ (where $1/Z$ has been absorbed into, say, θ_{x1y2}). Ignoring issues of fractional counts, we can map each p_{x1y2} to its most probable observed value through multiplication by an integer n (the total number of observations in a hypothetical data set). By definition, the Observed/Expected value for a cell in this data sample is:

$$\frac{n \cdot p_{x1y2}}{n \cdot p_{x1+} \cdot p_{+y2}} = \frac{p_{x1y2}}{p_{x1+} \cdot p_{+y2}}$$

where p_{x1+} is the sum of the probabilities in row x_1 (equivalently, for the most probable sample, the sum of the counts in that row divided by n) and p_{+y2} is the corresponding sum for column y_2.

In the absence of any identifiability or other constraints on the parameters, we are free to set each θ_{x1y2} equal to the O/E value for x_1y_2. If we then set θ_{x1} equal to p_{x1+} and each θ_{y2} equal to p_{+y2}, the original probabilities are recovered exactly:

$$\begin{aligned} p_{x1y2} &= \theta_{x1y2} \cdot \theta_{x1} \cdot \theta_{y2} \\ &= \frac{p_{x1y2}}{p_{x1+} \cdot p_{+y2}} \cdot p_{x1+} \cdot p_{+y2} \\ &= \frac{p_{x1y2}}{\cancel{p_{x1+}} \cdot \cancel{p_{+y2}}} \cdot \cancel{p_{x1+}} \cdot \cancel{p_{+y2}} \end{aligned}$$

In this exercise, we have divided p_{x1y2} by two marginal terms and then multiplied by the same two terms. We could do the same with any product of terms that are constant for a given row or column. For example, we could set $\theta_{x1y2} = p_{x1y2}/p_{x1+}$ (forward transitional probability) and multiply by $\theta_{x1} = p_{x1+}$ and $\theta_{y1} = 1$, or define $\theta_{x1y2} = p_{x1y2}/p_{+y2}$ (backward transitional probability) and multiply by $\theta_{x1} = 1$ and $\theta_{y1} = p_{+y2}$.

It is true that there are infinitely many solutions of this type. But it is difficult to imagine a statistical, grammatical, or other argument that would be advanced by canceling out freely chosen terms in this way.[12]

* * *

Other sections of SS attempt to defend O/E on different grounds. It is useful to consider two of the relevant remarks together:

> (1) "The O/E statistic is a simple method that can be done on pen and paper, and its results are easy to interpret. ... The loglinear model, by contrast, is harder to implement and interpret, because doing either of these requires a certain degree of statistical sophistication. Just given these differences, one might prefer O/E: it is easier to implement and accessible to a wider range of linguists" (p. 2).
>
> (2) "[I]f we are to prefer the loglinear model over O/E, it needs to be shown that the way humans generalize better-resembles a matrix discovered by the loglinear model than it does a matrix discovered by O/E. One way of achieving this goal would be to compare the matrices discovered by these two methods with aspects of speakers' grammars discovered through experimental tasks" (pp. 15-16).

Ignore here the reference to "the" log-linear model, a misconception dealt with above. Point (1) suggests that O/E might be preferred because many linguists lack expertise in statistical modeling (if this is indeed true). Point (2) suggests that experimental data — a very broad category — might someday provide evidence that humans generalize in a way that accords with O/E values better than some log-linear alternative.

The problem with this logic is that showing (2) would require exactly the expertise claimed to be in short supply by (1). Given some experimental data, how would we rigorously compare two hypotheses about the underlying generalizations that humans form? Not by writing down the data alongside the predictions of each hypothesis on paper (even assuming that we could determine the predictions without computational implementation). Sound practice in quantitative and experimental linguistics involves formulating each hypothesis and associated linking assumptions as a statistical model, and comparing the hypotheses in light of the data using methods such as those of the previous section (e.g., AIC, BIC, etc.). O/E values may be readily calculated by hand, but they are definitely not "easy to interpret" with respect to human competence or performance. They are useless at best, and quite plausibly misleading, in the absence of statistical analysis.

As a quantity — one of infinitely many values that could be computed from a frequency table — O/E solicits no defense. However, O/E ratios cannot be taken at face value as measuring the strength of the OCP or other constraints on cooccurrence. Furthermore, O/E values are disconnected from the large statistical literature on properly formulating and quantatively comparing alternative hypotheses. Researchers who are interested in developing restrictive grammatical theories and evaluating them against data (whether lexical or experimental) will find that O/E has little to offer, in contrast to the flexibility and precision of log-linear modeling. There are many textbooks and other beginner-friendly resources that provide the background

[12] SS does not mention restrictions on the infinitely many "solutions" of the nonidentifiable model, but they exist. For example, we cannot in general set θ_{x1y2} equal to $p_{x1y2}/(p_{x1+} + p_{+y2})$, which contains a sum rather than a product of marginal terms.

needed to construct and evaluate such models. As the field becomes increasingly interdisciplinary and quantitatively sophisticated, familiarity with basic concepts of statistical modeling will be necessary to responsibly conduct and comment on research.[13]

Appendix

The Observed/Expected value of a combination x_1y_2 is a ratio. The numerator is the observed frequency of x_1y_2 in a sample of data. The denominator is an estimate of the frequency with which x_1y_2 would be expected to occur, under the assumption that there are no restrictions on combinations: that is, under the assumption that the first and second members of a combination are independent random variables. This assumption is false, of course, whenever combinations of identical (or similar) consonants are penalized by the OCP or OCP-Place constraint.

Recall the log-linear model of Table 3 / Table 5, which has a single OCP constraint that applies with equal strength across all places of articulation, and no other constraints on combinations. For the most likely sample of data from the probability distribution defined by this model, and ignoring rounding of fractional counts, the O/E value for any combination x_1x_2 of same-place elements is:

$$\begin{aligned}\frac{n \cdot p_{x1x2}}{n \cdot p_{x1+} \cdot p_{+x2}} &= \frac{p_{x1x2}}{p_{x1+} \cdot p_{+x2}}\\ &= \frac{\theta_{x1} \cdot \theta_{x2} \cdot \theta_{OCP}}{Z \cdot p_{x1+} \cdot p_{+x2}}\\ &= \frac{1}{Z} \cdot \theta_{OCP} \cdot \left(\frac{\theta_{x1}}{p_{x1+}}\right) \cdot \left(\frac{\theta_{x2}}{p_{+x2}}\right)\end{aligned}$$

where p_{x1+} is the sum of the probabilities of all combinations with x_1 as the first member, and p_{+x2} is the sum of the probabilities of all combinations with x_2 as the second member. We can ignore the factor $1/Z$, which is always constant for all combinations and will cancel out in the proof below.

If each of the parenthesized terms is equal to one, the O/E ratio will be proportional to θ_{OCP} — more importantly, it will be the same for all of the combinations that violate the OCP. Otherwise, the strength of the OCP as measured by O/E will be 'deflated' (higher values) by parenthesized terms greater than one and 'inflated' (lower values) by terms smaller than one. The degree of deflation or inflation can easily vary across combinations, distorting the underlying unity with which the OCP actually applies, by hypothesis.

In particular, suppose that (i) x_1x_2 and y_1y_2 are two combinations that violate the OCP, (ii) $\theta_{OCP} = \theta_{x1x2} = \theta_{y1y2} < 1$, (iii) there are no constraints other than the OCP on combinations, (iv) $\theta_{x1} > \theta_{y1}$, and (v) $\theta_{x2} > \theta_{y2}$. Assumption (ii) states that OCP-violating combinations are dispreferred (i.e., less probable than would be expected under independent combination). Assumptions (iv) and (v) correspond to the situation in which one place of articulation, such as coronal, is positionally more probable than another. (Actually it

[13] The tutorial and notes on log-linear modeling of Ferraro and Eisner (2013) are highly recommended (`www.cs.jhu.edu/~jason/tutorials/loglin/formulas.pdf`), as are similar notes by Michael Collins (`www.cs.columbia.edu/~mcollins/loglinear.pdf` and Mark Johnson (`web.science.mq.edu.au/~mjohnson/papers/Johnson12IntroMaxEnt.pdf`). None of these discusses identifiability or model comparison, however, for which see many of the other references cited here.

will suffice for only one of the inequalities to be strict.) It follows that, in the most probable data sample, the O/E value of x_1x_2 will be larger than that of y_1y_2.
Proof. The first parenthesized term above simplifies to:

$$\frac{\theta_{x1}}{p_{x1+}} = \frac{\theta_{x1}}{\sum_\beta p_{x1\beta2}} = \frac{\theta_{x1}}{(1/Z)\cdot\sum_\beta p^*_{x1\beta2}} = \frac{\theta_{x1}}{(1/Z)\cdot\sum_\beta \theta_{x1}\cdot\theta_{\beta2}\cdot\theta_{x1\beta2}} = \frac{Z}{\sum_\beta \theta_{\beta2}\cdot\theta_{x1\beta2}}$$

and similarly for the second parenthesized term, $(\theta_{x2}/p_{+x2}) = Z/(\sum_\alpha \theta_{\alpha1}\cdot\theta_{\alpha1x2})$. Under the conditions assumed above, we can now establish the following inequality:

$$\begin{array}{rcl}
\dfrac{\theta_{x1}}{p_{x1+}} & > & \dfrac{\theta_{y1}}{p_{y1+}} \\
\dfrac{Z}{\sum_\beta \theta_{\beta2}\cdot\theta_{x1\beta2}} & > & \dfrac{Z}{\sum_\beta \theta_{\beta2}\cdot\theta_{y1\beta2}} \\
\sum\limits_\beta \theta_{\beta2}\cdot\theta_{y1\beta2} & > & \sum\limits_\beta \theta_{\beta2}\cdot\theta_{x1\beta2} \\
\theta_{x2}\cdot 1+\theta_{y2}\cdot\theta_{OCP}+C & > & \theta_{x2}\cdot\theta_{OCP}+\theta_{y2}\cdot 1+C \\
\theta_{x2}-\theta_{x2}\cdot\theta_{OCP} & > & \theta_{y2}-\theta_{y2}\cdot\theta_{OCP} \\
\theta_{x2}\cdot(1-\theta_{OCP}) & > & \theta_{y2}\cdot(1-\theta_{OCP})
\end{array}$$

where C is a constant determined by combinations that do not contain x_2 or y_2, and the last line follows from assumptions (ii) and (v). A parallel computation for the second parenthesized term, using assumptions (ii) and (iv), establishes that $\theta_{x2}/p_{+x2} > \theta_{y2}/p_{+y2}$.

Plugging these results back into the equation for O/E given the hypothesized model, we have finally:

$$\theta_{OCP}\cdot\left(\frac{\theta_{x1}}{p_{x1+}}\right)\cdot\left(\frac{\theta_{x2}}{p_{+x2}}\right) \quad > \quad \theta_{OCP}\cdot\left(\frac{\theta_{y1}}{p_{y1+}}\right)\cdot\left(\frac{\theta_{y2}}{p_{+y2}}\right)$$

That is, O/E values overestimate the strength of the OCP for combinations that are independently less probable (because their first and second members have smaller row and column parameters).

It is unsurprising that O/E can provide a distorted estimate of OCP strengths, as its denominator assumes complete independence when by hypothesis there is some interaction between members of a combination. The proof above establishes one set of conditions under which it does so in a systematic way and for a ‘perfect’ sample. Further work could generalize the current result to an expectation over possible samples, and identify other conditions of distortion.[14]

[14] It is likely that results along these lines can be found in previous research on log-linear modeling of contingency tables, but I have not found them yet and any pointers are welcome.

References

Agresti, A. (2002). *Categorical Data Analysis*. Hoboken, NJ: John Wiley & Sons.

Agresti, A. (2019). *An Introduction to Categorical Data Analysis*. Hoboken, NJ: John Wiley & Sons.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Bane, M. and J. Riggle (2008, June). Three correlates of the typological frequency of quantity-insensitive stress systems. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, Columbus, Ohio, pp. 29–38. Association for Computational Linguistics.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). Stan: A probabilistic programming language. *Journal of Statistical Software 76*(1), 1–32.

Casella, G. and R. L. Berger (2002). *Statistical Inference*. Pacific Grove, CA: Brooks/Cole.

Cotterell, R., N. Peng, and J. Eisner (2015). Modeling word forms using latent underlying morphs and phonology. *Transactions of the Association for Computational Linguistics 3*, 433–447.

Culbertson, J., P. Smolensky, and C. Wilson (2013). Cognitive Biases, Linguistic Universals, and Constraint-Based Grammar Learning. *Topics in Cognitive Science 5*(3), 392–424.

Ferraro, F. and J. Eisner (2013, August). A virtual manipulative for learning log-linear models. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*, Sofia, Bulgaria, pp. 66–76. Association for Computational Linguistics.

Fienberg, S. E. and A. Rinaldo (2007). Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *Journal of Statistical Planning and Inference 137*(11), 3430–3445.

Fienberg, S. E. and A. Rinaldo (2012). Maximum likelhood estimation in log-linear models. *The Annals of Statistics 40*(2), 996–1023.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. Hoboken, NJ: CRC Press.

Gelman, A. and J. Hill (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Goldwater, S. and M. Johnson (2003). Learning OT constraint rankings using a maximum entropy model. In J. Spenader, A. Eriksson, and Ö. Dahl (Eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, pp. 111–120. Stockholm: Stockholm University.

Hayes, B., B. Tesar, and K. Zuraw (2003). OTSoft 2.5. software package, http://www.linguistics.ucla.edu/people/hayes/otsoft/.

Kruschke, J. K. (2015, January). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Burlington, MA: Academic Press / Elsevier.

Legendre, G., Y. Miyata, and P. Smolensky (1990a). Harmonic Grammar — A formal multi-level connectionist theory of linguistic well-formedness: An application. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pp. 884–891. Hillsdale, NJ: Lawrence Erlbaum.

Legendre, G., Y. Miyata, and P. Smolensky (1990b). Harmonic Grammar — A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pp. 388–395. Hillsdale, NJ: Lawrence Erlbaum.

MacKay, D. J. C. (2004). *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*. Boca Raton, FL: Routledge.

Murphy, K. (2014). *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology 47*(1), 90–100.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of {H}armony {T}heory. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Volume 1, pp. 194–281. Cambridge, MA: MIT Press.

Smolensky, P., M. C. Mozer, and D. E. Rumelhart (Eds.) (1996). *Mathematical Perspectives on Neural Networks*. New York: Psychology Press.

Stanton, J. and J. F. Stanton (2022). In Defense of O/E. lingbuzz/006391.

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. New York: Springer.

White, J. (2017). Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a P-map bias. *Language 93*(1), 1–36.

Wickens, T. D. (1989). *Multiway Contingency Tables Analysis for the Social Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science 30*(5), 945–982.

Wilson, C. and M. Obdeyn (2009). Simplifying subsidiary theory: Statistical evidence from Arabic, Muna, Shona, and Wargamay. Ms., Johns Hopkins University.