The Stata Journal (2009)
9, Number 2, pp. 306–314

Speaking Stata: I. J. Good and quasi-Bayes smoothing of categorical frequencies

Nicholas J. Cox
Department of Geography
Durham University
Durham, UK
n.j.cox@durham.ac.uk

Abstract. I. J. Good (1916–2009) was a prolific scientist who contributed to many fields, mostly from a Bayesian standpoint. This column explains his idea of quasi-Bayes (a.k.a. pseudo-Bayes) estimation or smoothing of categorical frequencies in a contingency table, which is especially useful as a way of dealing with awkward sampling or random zeros. It shows how the method can be implemented, almost calculator-style, using a combination of Stata and Mata. Convenience commands `qsbayesi` and `qsbayes` are also introduced.

Keywords: st0168, qsbayesi, qsbayes, categorical data, contingency tables, Mata, pseudo-Bayes, quasi-Bayes, random zeros, sampling zeros, smoothing

1 I. J. Good, polymathic statistician

I. J. Good (1916–2009) was a prolific scientist who contributed to many fields, including mathematics, cryptography, computer science, philosophy of science, physics, and statistics. Much of his work took a Bayesian viewpoint. His life started as Isidore Jacob Gudak, the son of Polish immigrants in London, and ended as Irving John (Jack) Good, a retired academic in Virginia. His career took him back and forth between government work (much of it still classified) in Britain and the United States, and academic study and positions at Cambridge, Manchester, Oxford, and Virginia Tech. Many details have emerged, however, of his statistical contributions to codebreaking during World War II in work with Alan M. Turing and others (e.g., Good [1979]). The interview conducted by Banks (2008) gives much more detail and a fine sense of Good's wit, in every sense of that word. Obituaries include those by van der Vat (2009) and Banks (2009). The collections in Good (1962, 1983) and Banks and Smith (2008) anthologize some of Good's work, but he published other books and several hundred papers and he wrote many more, so no collection is at all comprehensive. The titles of two of these books, *Good Thinking: The Foundations of Probability and its Applications* and *The Good Book: Thirty Years of Comments, Conjectures, and Conclusions by I. J. Good*, indicate a wordplay that is pervasive. Among many other things, Good was a master at being serious without being solemn.

2 Quasi-Bayes smoothing

In March 1972, I bought a used copy of Good's short monograph *The Estimation of Probabilities: An Essay on Modern Bayesian Methods* (1965) in a bookshop just a short distance away from where he had studied mathematics some 30 years earlier. I appear to have paid 20 pence, much less than a dollar or a Euro. I can echo the testimony of Fienberg (2008) that the monograph long ago paid for itself in repeated reading. This column focuses on just one idea in that book and how to implement it in Stata. That idea was picked up by Fienberg, Holland, and Sutherland in various papers in the 1970s (e.g., Fienberg and Holland [1970, 1972, 1973] and Sutherland, Holland, and Fienberg [1975]) and was discussed systematically and clearly by Bishop, Fienberg, and Holland (1975) under the heading of pseudo-Bayes estimation. Good (2008) himself preferred a heading of quasi-Bayes, so that preference is honored here. See survey works on categorical data analysis, such as Agresti (2002), for the wider context.

The main problem is very simple. You have some sample frequencies for two or more categories and wish to estimate the underlying probabilities. The difficulties with taking empirical probabilities as they come are easily underlined by thinking about those categories for which you have observed zero frequencies, even though it is clear that such categories could have been observed. These zeros may be called sampling or random zeros. (In contrast, zeros for categories that could not be populated even in principle, such as pregnant males, may be called structural or fixed zeros. We will not focus on those further, beyond explaining later that they can be accommodated easily in the method under consideration.)

The story underlying sampling zeros we take to be just that your sample failed to catch such categories, most obviously because they are relatively uncommon, and so by chance did not fall into your net. Problems of poor sampling design, such as when a telephone-based survey did not catch people without telephone access, are beyond the scope of this method to fix.

Sampling or random zeros are awkward for virtually any exploratory or inferential purpose. Most simply, their definition implies that because we do not believe them, it is contradictory to take them literally. In addition to being problematic for standard approaches, they prove awkward in plotting frequencies or probabilities on logarithmic or logit scales, which is often a very good idea (e.g., Cox [2004, 2008]).

The best way to get to the underlying positive probabilities is not obvious. Several recipes are likely to occur to statistically minded researchers, most commonly in terms of fitting one or more appropriate models. Even if that appears to be the main line of attack, having other methods in the toolbox to provide checks should also appeal to you. In particular, although a model provides a kind of smoothing by fitting, it may not be the best kind of smoothing if you want to keep fairly close to the data.

Dealing with zeros ad hoc is one possibility, and dodges and fudges of some kind are often recommended, such as adding 1/2 to observed sampling zeros, with or without adjustment to observed positive frequencies. As Good (1965, 56) himself said, "Real life is both complicated and short, and we make no mockery of honest adhockery."

But many statisticians and scientists would prefer a more systematic approach. The approach followed here treats the problem as one of flattening or shrinking or smoothing the observed frequencies toward those implied by a set of prior probabilities. The word *prior* indicates a Bayesian flavor to the method, but it is a flavor that should be acceptable even to those skeptical or squeamish about anything Bayesian. Smoothing categorical data is territory less visited than smoothing with respect to one or more coordinates—say, coordinates defined by time, space, or predictors—but several ideas have been proposed. Simonoff (1996) is one gateway to the literature.

Imagine a vector of observed frequencies $n_i, i = 1, \ldots, I$, with total frequency N and a corresponding vector of prior probabilities, q_i, with total probability 1. The quasi-Bayes recipe produces smoothed estimates of frequencies $N\widehat{p}_i$, where

$$\widehat{p}_i = \frac{N}{N+K}\frac{n_i}{N} + \frac{K}{N+K}q_i$$

and shrinkage is tuned by the constant

$$K = \frac{N^2 - \sum_{i=1}^{I} n_i^2}{\sum_{i=1}^{I} n_i - Nq_i}$$

Otherwise put, smoothed frequencies $\widehat{n}_i$ are given by

$$\widehat{n}_i = N\left(\frac{1}{N+K}n_i + \frac{K}{N+K}q_i\right)$$

Simply, and expectably, smoothed frequencies are chosen to be a compromise between those observed and those expected from prior probabilities. These estimates minimize the total mean squared error between estimated and estimand probabilities. The only issue is identifying suitable prior probabilities. Once they are identified, the task is a straightforward calculation.

The setup implied by the notation of a single vector is more general than it may seem at first. The idea extends to frequencies that researchers regard as structured in contingency tables that are two-dimensional or higher. The twist is merely that the prior probabilities may then arise from a more complicated calculation involving different predictor levels, possibly interactions, and so forth. If there is some ordering of categories, that too may affect the prior probabilities. All is at the researcher's discretion. A simple analogue is a chi-squared test, for which any generating process should be mirrored in calculation of expected frequencies. That is implicit rather than explicit in the notation.

As usual, it may be a good idea to consider various possible vectors of prior probabilities. The most conservative or agnostic option of uniform probabilities, $1/I$, typically produces only modest smoothing, contrary perhaps to some intuitions.

The modest generality of the formulation also extends to handling structural or fixed zeros directly. Any such formulation would have both $n_i = 0$ and $q_i = 0$ so that, correspondingly, $\widehat{p}_i = 0$. So zero observed frequencies would never be smoothed upward if the zeros are entirely credible.

3 Stata implementation

You might want to apply quasi-Bayes smoothing informally in a calculator style. One possibility is that you already have a contingency table of counts, say, from a report or published paper, which you should then put into a Stata or Mata matrix. Another possibility is that you have read the raw data in Stata, and so you can use `tabulate` or some similar command to produce such a table. The `matcell()` option of `tabulate` is especially handy for getting hold of a matrix of frequencies.

For example, having read in Stata's auto data, we can look at a table of two categorical variables, whether various cars in the United States were foreign and their repair record in 1978. There is no evident structural reason why repair records 1 and 2 were impossible for foreign cars, so we interpret the observed zeros as indicating small underlying probabilities.

```
. sysuse auto, clear
(1978 Automobile Data)
. tabulate for rep78, matcell(freq)
```

	Repair Record 1978					
Car type	1	2	3	4	5	Total
Domestic	2	8	27	9	2	48
Foreign	0	0	3	9	9	21
Total	2	8	30	18	11	69

You could stay in Stata and use its matrix and scalar functions to proceed further. A better option is to start Mata:

```
. mata :
```

The following sequence of Mata commands is not as short as it could be, but it shows a sequence of operations in simple steps. One possible prior is just uniform across both row and column categories. Once we are done, we send the matrix back to Stata.

```
: freq = st_matrix("freq")
: prior = J(2, 5, 1/10)
: N = sum(freq)
: diff = freq - N * prior
: K = (N * N - sum(freq :* freq)) / sum(diff :* diff)
: qb = N * ((freq :/ (N + K)) + (K / (N + K)) * prior)
: st_matrix("qb", qb)
: end
```

If you are new to Mata, the most exotic syntax here is likely to be `J(2, 5, 1/10)` and the operators `:*` and `:/`. The former creates a matrix with 2 rows, 5 columns, and elements all 1/10 = 0.1, and the latter stipulate elementwise multiplication and division. Although the initial formulation was in terms of a vector of frequencies, we can apply the same recipe to a matrix.

Once back in Stata, we can look at the frequencies. `matrix list` by default will give more decimal places than you care about, so the `format()` option may be needed.

```
. matrix list qb
qb[2,5]
           c1         c2         c3         c4         c5
r1  2.4175474  7.9062649  25.287203  8.8210511  2.4175474
r2  .58797493  .58797493  3.3323337  8.8210511  8.8210511
. matrix list qb, format(%3.2f)
qb[2,5]
       c1     c2     c3     c4     c5
r1   2.42   7.91  25.29   8.82   2.42
r2   0.59   0.59   3.33   8.82   8.82
```

As earlier signaled, this uniform prior has only a moderate smoothing effect. The zeros are smoothed to 0.59, and the peak frequency of 27 is smoothed to 25.29.

For this two-dimensional table, there is arguably a more natural default: the prior probabilities implied by a model in which the two categorical variables are independent. That then respects the far-from-uniform marginal distributions.

All the information needed is already at hand. Re-enter Mata, and see that getting the expected frequencies and expected probabilities under an independence model is possible in one line:

```
: rowsum(freq) * colsum(freq) / sum(freq)
                1            2            3            4            5
  1   1.391304348   5.565217391   20.86956522   12.52173913   7.652173913
  2   .6086956522   2.434782609   9.130434783    5.47826087   3.347826087

: rowsum(freq) * colsum(freq) / sum(freq)^2
                1            2            3            4            5
  1   .0201638311   .0806553245   .3024574669   .1814744802   .1109010712
  2   .0088216761   .0352867045   .1323251418   .0793950851   .0485192187
```

Textbooks typically explain how to calculate expected frequencies under independence for a two-way table. You take the product of row frequencies and column frequencies and divide the result by the grand total. In Mata, and indeed in any matrix language, we need to put the terms in precisely that order to set up the calculation conformably. Here the row sums are a 2×1 matrix or column vector and the column sums are a 1×5 matrix or row vector, and so their (matrix) product is a 2×5 matrix. The sum of frequencies and its square are both scalars: here Mata uses `/` to mean division of the elements of a matrix by a scalar.

Now we just need to repeat the quasi-Bayes calculation for that different prior:

```
: priori = rowsum(freq) * colsum(freq) / sum(freq)^2
: diffi = freq - N * priori
: Ki = (N * N - sum(freq :* freq)) / sum(diffi :* diffi)
: qbi = N * ((freq :/ (N + Ki)) + (Ki / (N + Ki)) * priori)
: st_matrix("qbi", qbi)
: end
```

Let's look at the results. The first set of smoothed frequencies, although crudely derived, were not at all absurd, but these are better.

```
. matrix list qbi, format(%3.2f)
qbi[2,5]
       c1     c2     c3     c4     c5
r1   1.86   7.43  25.57   9.82   3.32
r2   0.14   0.57   4.43   8.18   7.68
```

Note in particular how the two observed zeros are smoothed to different estimated frequencies when marginal information is taken into account. Let's emphasize once more that no other information has yet been considered, such as whether there is an interaction or the ordered nature of the repair record, but the story will be left at this point.

4 More complicated situations

The main message so far is that quasi-Bayes smoothing is easy in principle and easy in practice. Anyone interested in the idea is likely to want to apply it in much more complicated situations. Here are some brief suggestions on Stata strategy and tactics.

To work with other setups, you may need to move beyond `tabulate` and deal directly with appropriate modeling commands. A different data structure is then advisable. The `contract` command ([D] **contract**) produces a dataset with cell frequencies as a key variable. Note carefully the key options `zero` and `nomiss`. `zero`, which you will almost certainly need, ensures that cells with zero frequencies are included explicitly in the new dataset. `nomiss`, which you will probably want, discards cells for missing categories.

With this new dataset, you can then use an appropriate modeling command, choosing between `poisson` or `glm` largely as a matter of taste or convenience.

5 Canned commands

Two new commands have been written to allow quasi-Bayes smoothing to be carried out more conveniently and are published formally with this column. `qsbayesi` is intended as an immediate command, with the twist that *immediate* here means working with Stata matrices. `qsbayes` is for the more traditional situation in which observed frequencies populate the values of a variable. This might be the result of a previous `contract`, as mentioned in the previous section, or the frequencies might have been entered directly.

5.1 Syntax for qsbayesi

`qsbayesi` *freq_matrix* [*prior_matrix*] [, `prob` `format(`*format*`)`]

Description

`qsbayesi` takes a matrix of frequencies, *freq_matrix*, and shrinks or smooths it toward a set of frequencies implied by prior probabilities. This will have the effect of replacing sampling zeros with positive estimates whenever the priors are positive. These estimates minimize the total mean squared error between estimated and estimand probabilities.

If *prior_matrix* is specified, it must be the same shape as *freq_matrix* and sum to 1. If *prior_matrix* is not specified, it is taken to be a matrix of equal probabilities.

Options

`prob` specifies that probabilities rather than estimated frequencies be shown.

`format(`*format*`)` controls the format with which matrix output is printed.

5.2 Syntax for qsbayes

`qsbayes` *datavar* [*priorvar*] [*if*] [*in*] [, `by(`*rowvar* [*colvar* [*layervar*]]`)`
`generate(`*newvar*`)` `prob` *tabdisp_options*]

Description

`qsbayes` takes *datavar*, which should be a set of frequencies, and shrinks or smooths it toward a set of frequencies implied by prior probabilities. This will have the effect of replacing sampling zeros with positive estimates whenever the priors are positive. These estimates minimize the total mean squared error between estimated and estimand probabilities.

If *priorvar* is specified, it must sum to 1 for the data used. If *priorvar* is not specified, it is taken to be a set of equal probabilities.

Options

`by(`*rowvar* [*colvar* [*layervar*]]`)` indicates that *datavar* refers to a table with rows (and columns, if specified [and layers, if specified]) indexed by the variable(s) named, which will structure a display of cell estimates using `tabdisp`. If `by()` is not specified, cell estimates will be displayed according to observation numbers.

`generate(`*newvar*`)` generates a new variable containing results.

`prob` specifies that probabilities rather than estimated frequencies be shown (and kept, if desired).

tabdisp_options are options of `tabdisp`. The default includes `center`.

6 Conclusion

Although it has not entered the mainstream of categorical data analysis, quasi-Bayes smoothing is easy to understand and implement and is flexible according to researchers' ideas and needs. This column has shown how to implement this method with a combination of Stata and Mata and has introduced two new commands that may be used according to circumstance. Those interested further should look both at the original account in Good (1965) and at the systematic discussion in Bishop, Fienberg, and Holland (1975), who provide numerous details and examples going far beyond the introduction here. A neatly choreographed example in which two social mobility tables are used to smooth each other in various ways is especially entertaining and thought provoking.

7 References

Agresti, A. 2002. *Categorical Data Analysis*. 2nd ed. Hoboken, NJ: Wiley.

Banks, D. 2008. A conversation with I. J. Good. In *The Good Book: Thirty Years of Comments, Conjectures, and Conclusions by I. J. Good*, ed. D. Banks and E. P. Smith, 13–34. Houston, TX: Rice University Press. Article originally published in 1996, *Statistical Science* 11: 1–19.

———. 2009. Obituary: I. J. Good 1916–2009. *IMS Bulletin* 38(5): 11–12. http://bulletin.imstat.org/pdf/38/5.

Banks, D., and E. P. Smith. 2008. *The Good Book: Thirty Years of Comments, Conjectures, and Conclusions by I. J. Good*. Houston, TX: Rice University Press.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press. Reissue, New York: Springer, 2007.

Cox, N. J. 2004. Speaking Stata: Graphing categorical and compositional data. *Stata Journal* 4: 190–215.

———. 2008. Stata tip 59: Plotting on any transformed scale. *Stata Journal* 8: 142–145.

Fienberg, S. E. 2008. I. J. Good—An appreciation. In *The Good Book: Thirty Years of Comments, Conjectures, and Conclusions by I. J. Good*, ed. D. Banks and E. P. Smith, 1–4. Houston, TX: Rice University Press.

Fienberg, S. E., and P. W. Holland. 1970. Methods for eliminating zero counts in contingency tables. In *Random Counts in Scientific Work, Volume 1: Random Counts*

in Models and Structures, ed. G. P. Patil, 233–260. University Park, PA: Pennsylvania State University Press.

———. 1972. On the choice of flattening constants for estimating multinomial probabilities. *Journal of Multivariate Analysis* 2: 127–134.

———. 1973. Simultaneous estimation of multinomial cell probabilities. *Journal of the American Statistical Association* 68: 683–691.

Good, I. J., ed. 1962. *The Scientist Speculates: An Anthology of Partly-Baked Ideas*. London: Heinemann.

Good, I. J. 1965. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. Cambridge, MA: MIT Press.

———. 1979. Studies in the history of probability and statistics. XXXVII: A. M. Turing's statistical work in World War II. *Biometrika* 66: 393–396.

———. 1983. *Good Thinking: The Foundations of Probability and Its Applications*. Minneapolis, MN: University of Minnesota Press.

———. 2008. Terminology for various kinds of Bayesian methods. In *The Good Book: Thirty Years of Comments, Conjectures, and Conclusions by I. J. Good*, ed. D. Banks and E. P. Smith, 76–78. Houston, TX: Rice University Press. Article originally published in 1980, *Journal of Statistical Computation and Simulation* 11: 309–313.

Simonoff, J. S. 1996. *Smoothing Methods in Statistics*. New York: Springer.

Sutherland, M., P. W. Holland, and S. E. Fienberg. 1975. Combining Bayes and frequency approaches to estimate a multinomial parameter. In *Studies in Bayesian Econometrics and Statistics: In Honor of Leonard J. Savage*, ed. S. E. Fienberg and A. Zellner, 585–617. Amsterdam: North-Holland.

van der Vat, D. 2009. Jack Good. *Guardian*, April 29. http://www.guardian.co.uk/science/2009/apr/29/jack-good-codebreaker-obituary.

About the author

Nicholas Cox is a statistically minded geographer at Durham University. He contributes talks, postings, FAQs, and programs to the Stata user community. He has also coauthored 15 commands in official Stata. He was an author of several inserts in the *Stata Technical Bulletin* and is an editor of the *Stata Journal*.