The Stata Journal (2024)
24, Number 3, pp. 535–545 DOI: 10.1177/1536867X241276115

Stata tip 156: Concentration and diversity measures using egen

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

1 The problem

A common desire arising in many fields is to summarize the inequality or unevenness of a set of observed proportions or fractions $p_1, \ldots, p_S$ that quantify the abundance of S distinct categories. Each such proportion lies between 0 and 1 inclusive, and their sum should be 1. Those proportions can be presented directly as data or arise from summarizing counts or measurements presented as more detailed data. Consider two toy examples and the jargon of two fields. Suppose S is 5. A set of observed proportions of 1, 0, 0, 0, and 0 shows maximum "concentration" (a term common in economics), whereas one of 0.2, 0.2, 0.2, 0.2, and 0.2 shows maximum "diversity" (a term common in ecology). Other jargon could be cited here and may be familiar to you.

The notation S for the number of categories mirrors ecological convention whereby S evokes the number of species. Naturally, nothing stops use of the same notation for applications to different taxonomic levels (for example, genera or families) and indeed to categorizations that have no relation to biological taxonomy.

Even this brief epitome raises a bundle of linked questions, including quite how to summarize such a set, what important detail is lost by any such summary, and what guiding theory is available upstream and what applications lie downstream of any descriptive exercise. Various comments, and especially various references, in Cox (2005, 2022) apply here.

Community-contributed commands can easily be found to help, but the focus of this tip is quite different: to underline how the `egen` command in particular can be an invaluable workhorse for calculations.

It is important (or at least attractive) to many Stata users to be independent of community-contributed commands. As we will see, the emphasis here is on producing new variables, which themselves are often needed for further analyses. Concentration, diversity, or whatever else you call it can variously be an outcome you are trying to explain or a predictor you might include in some model. Any command that produces only tabulations of results may be of little or no help in that regard. Being able to produce results step by step may increase understanding of how measures are defined. Seeing examples of how `egen` and other commands may be used should help you to appreciate how to combine different Stata tools to reach some desired end. A simple

rule of thumb is that whatever is defined in your literature by a few lines of algebra, or even one line, should often be computable with a few lines of Stata.

2 Chosen measures: repeat rate and entropy

Two measures will be used as examples here, although some of their relatives will also be mentioned. If you have come here because of the tip title, you are very likely to be familiar with the main ideas, which may quite possibly be under different names. Some small details in this section may nevertheless be novel to you.

The first is here called "repeat rate", a term that seems to go back to A. M. Turing and was often used by Good (for example, 1953, 1965). For $p_1, \ldots, p_S$, repeat rate is

$$\sum_{s=1}^{S} p_s^2 =: R$$

One interpretation of repeat rate arises by imagining that we randomly sample pairs of individuals. The repeat rate is the probability that each pair contains individuals that belong to the same category. Here "individuals" could mean, say, individual people, animals, or plants; or small parts of a continuum, say, atoms or quanta of income or wealth or points within areas of land. This interpretation helps explain another excellent name, "match probability" (MacKay 2003).

For observed proportions 1, 0, 0, 0, and 0, R is $1^2 + 4 \times 0^2$ or 1, and for 0.2, 0.2, 0.2, 0.2, and 0.2, R is 5×0.04 or 0.2. Thus, R measures concentration, unevenness, or inequality. Its complement $1 - R$ and its reciprocal $1/R$ measure diversity, evenness, or equality. Note that any proportions of 0 do not affect the value of R, which is determined by positive proportions alone.

Repeat rate is often named for various people, with the intent of honoring predecessors but also with varying historical accuracy. Additional names associated with R or its relatives include C. Gini, W. F. Friedman (the cryptographer, not the economist), G. U. Yule, E. H. Simpson (for whom Simpson's paradox is named), A. O. Hirschman, O. C. Herfindahl, J. H. Greenberg, and P. M. Blau.

The second measure, "entropy", has more clear-cut antecedents in the work of Shannon in information theory, which, in turn, owed much to precedents in physics and engineering. For the original articles and much more, see Sloane and Wyner (1993). For more on Shannon, his work, and its context, see Gleick (2011) and Soni and Goodman (2017). MacKay (2003) is one excellent entry into the more technical literature. The monographs of Theil (1967, 1972) are lucid and well illustrated. Leinster (2021) embraces various mathematical perspectives while also being inspired by empirical applications. Here I introduce entropy by

$$\sum_{s=1}^{S} p_s \ln(1/p_s) =: H$$

A more common version is the algebraically equivalent $-\sum p_s \ln p_s$. I prefer the first version, which makes it a little more obvious that H is a weighted average over $\ln(1/p_s)$. Using a different base of logarithm, say, $\log_2$ or $\log_{10}$, may seem congenial, conventional, or compelling, with no issue raised beyond the need to be explicit and consistent. A deeper question is why $\ln(1/p)$ is an appropriate scale in the first place, which hinges on the pleasant and helpful consequences of that choice, such as entropy being additive for independent variables.

What happens when any p_s is 0? Any queasiness over working with $0 \ln(1/0)$ fades on plotting $p \ln(1/p)$ as a function of p, which indicates that $p \ln(1/p)$ tends to zero as p does. The point can be established more rigorously, but it implies a practical rule that $0 \ln(1/0)$ is always to be taken as 0. We may need to override Stata's inclination to return $\ln(1/0)$ as missing. The principle that zero proportions do not affect H is thus parallel to the same principle for R.

For observed proportions 1, 0, 0, 0, and 0, H is $1 \ln 1 + 4 \times 0 \ln(1/0)$ or 0, and for 0.2, 0.2, 0.2, 0.2, and 0.2, H is $5 \times 0.2 \ln 5$ or $\ln 5 \approx 1.609$. Thus, H measures diversity, evenness, or equality.

Further literature references could be multiplied indefinitely, but here is one personal favorite. The text of Schmitt (1969) is unusual among introductory treatments in mentioning both repeat rate and entropy. That appears to have been the only publication of Samuel Arthur Schmitt (1926–1978), as a delicate side effect of his working in the intelligence community. Yet it still stands as an original and stimulating perspective on statistics from a Bayesian point of view.

3 An easy sandbox example

As a first sandbox, we use the data displayed in figure 1, which incidentally was produced using the community-contributed `tabplot` command (Cox 2016).

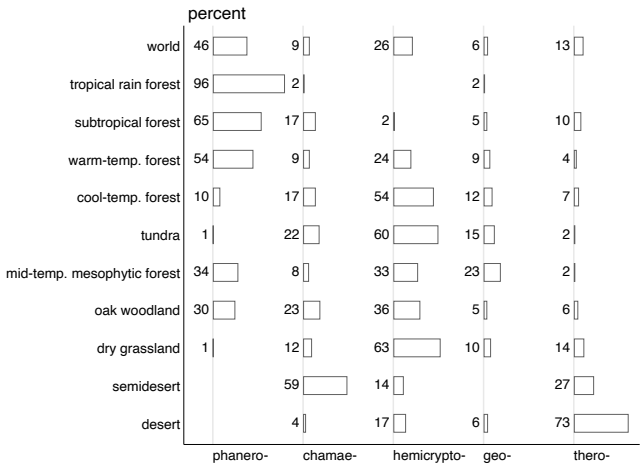

Figure 1. Abundance of various plant life-forms in various ecological communities. Data from Whittaker (1975, 63–64). All life-form names have suffix "phytes" (for example, "geo-" indicates "geophytes").

For various broadly defined communities, we see the percentages of plants that belong to particular categories called life-forms. The dataset is available with the media for this issue. Definitions of each category are given in the `notes` to the dataset and will not all be repeated here. As one example, geophytes are perennial herbs with their perennating tissues underground. Percentages are rounded to the nearest integer: a trivial side effect is that the total for one category is 99%.

The data come from an outstanding and still valuable text (Whittaker 1975, 63–64). For more on Robert H. Whittaker (1920–1980), see Westman and Peet (1982). Despite its age, the article of Smith (1913) serves well as a concise example of early work in this style, initiated by Christen C. Raunkiær (1860–1938), and of characterizing vegetation composition quantitatively.

This is an easy example to start with because the data are presented in long layout, with repeated observations for each life-form in each community. The data are presented as percentages, which are almost what we need.

An extra reason for using this example is that some values of zero are explicit in the data. Whatever code we write should be able to handle such a convention cleanly. Conversely, the convention of `tabplot` used for figure 1 is that zero values are matched by holes in the display. If you like, a bar could be drawn, but it would be of height zero and so hard to see. `tabplot` will not show numeric text for zero values either.

After reading in the data, the first use of `egen` is to calculate proportions directly.

```
. use lifeforms
. egen p = pc(percent), by(id1) prop
```

If we had not known about that `egen` function, `pc()`, we could have gotten proportions by first using another, the `total()` function, and then dividing. The `total()` function can now be used directly to get the measures we want. Three decimal places are more than enough for most displays.

```
. egen R = total(p^2), by(id1)
. egen H = total(p * ln(1/p)), by(id1)
. format R H %4.3f
```

A careful check shows that although `0 * ln(1/0)` will be returned as missing by Stata, missing values are ignored by the `total()` function, producing the effect we need in calculating entropy H.

Now each summary measure is repeated for each observation to which it refers. For many purposes, you need only one observation to be used. This is the role in life of the `egen` function `tag()`:

```
. egen tag = tag(id1)
```

In general, any group might occur one or more times in a dataset, so there are only two possible general rules: tagging the first or tagging the last, the last being the same as the first for a group of one. `tag()` always tags the first observation in the current sort order of the dataset, but that choice should not bite. The whole point is that `tag()` should be used only if all values in a group are identical and you want to use only one. `if tag` is idiomatic because the values produced are only ever 1 or 0, and never missing. `tabdisp` is an often overlooked command with essentially the same result here.

```
. list id1 R H if tag, noobs sep(0)
```

id1	R	H
world	0.308	1.358
tropical rain forest	0.922	0.196
subtropical forest	0.474	1.040
warm-temp. forest	0.367	1.237
cool-temp. forest	0.350	1.305
tundra	0.431	1.048
mid-temp. mesophytic forest	0.284	1.351
oak woodland	0.279	1.386
dry grassland	0.441	1.097
semidesert	0.441	0.940
desert	0.567	0.829

```
. tabdisp id1, cellvar(R H)
```

(*output omitted*)

Note that we suppressed the display of the `tabdisp` output because it is similar to the `list` output.

4 A more challenging example

`nlswork.dta`, downloadable with Stata, contains data from the US National Longitudinal Survey of Young Women, 14–24 years old in 1968. Each observation is of an individual woman, so we must first calculate proportions relevant to the problem of concern, here taken to be racial diversity within each industrial code. The `count()` function of `egen` can help here, especially in ignoring missing values as desired, but for simplicity, we just segregate observations with missing values using an indicator variable. After that, we use `generate` directly and count observations at different levels.

```
. webuse nlswork, clear
(National Longitudinal Survey of Young Women, 14-24 years old in 1968)
. generate good = !missing(ind_code, race)
. bysort good ind_code race: generate freq = _N
. bysort good ind_code: generate total = _N
. generate p = freq / total if good
(341 missing values generated)
```

Next it is essential that we use each proportion just once. The tagging technique used in the previous section is one good way to do that. Notice that using the `tag` variable as a multiplier selects values we want to use and ignores the others. Any term multiplied by 1 enters a total as itself, while any term multiplied by 0 yields 0 and thus does not affect a total.

Then we can look at results as previously.

```
. egen tag = tag(ind_code race)
. egen R = total(p^2 * tag), by(ind_code)
. egen H = total(p * ln(1/p) * tag), by(ind_code)
. format R H %4.3f
. tabdisp ind_code if good, cellvar(R H)
```

Industry of employment	R	H
1	0.568	0.678
2	0.714	0.461
3	0.683	0.587
4	0.529	0.692
5	0.606	0.627
6	0.646	0.569
7	0.685	0.530
8	0.588	0.666
9	0.500	0.718
10	0.669	0.589
11	0.587	0.648
12	0.558	0.676

There are naturally other ways to do that which may appeal, depending partly on what else you want to do. You could first `contract` the dataset to one of the frequencies. Or you could use frames.

5 Dealing with wide layout

Data may not arrive in an ideal long layout, or you may prefer a wide layout for some reason. To use the technique of sections 3 and 4, you need a long layout, which is not fatal to (moderately) easy calculation of these measures.

The term "layout" I owe to Clyde Schechter's postings on Statalist. It has an advantage of being less overloaded than more common terms such as "structure" or "format".

Naturally, you could `reshape long`, which may be a good idea on several grounds. Otherwise, this section first produces a wide version of the life-forms data and then shows how you might proceed on wide data. For more on working rowwise, see Cox (2009, 2020).

```
. use lifeforms, clear
. reshape wide percent, i(id1) j(id2)
(j = 1 2 3 4 5)
Data                               Long   ->   Wide
-----------------------------------------------------------------------------
Number of observations               55   ->   11
Number of variables                   4   ->   7
j variable (5 values)               id2   ->   (dropped)
xij variables:
                                 percent   ->   percent1 percent2 ... percent5
-----------------------------------------------------------------------------

. describe
Contains data
 Observations:            11
    Variables:             7
                                                (_dta has notes)
-------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------
id1             byte    %31.0g     id1_short

percent1        float   %9.0g                 1 percent
percent2        float   %9.0g                 2 percent
percent3        float   %9.0g                 3 percent
percent4        float   %9.0g                 4 percent
percent5        float   %9.0g                 5 percent
community       str31   %31s
-------------------------------------------------------------------------------
Sorted by: id1
     Note: Dataset has changed since last saved.
```

A different `egen` function, `rowtotal()`, does what you would guess from its name.

```
. egen total = rowtotal(percent*)
```

We are going to loop over the variables holding percentages for each category. It is vital to ensure that any categories with zero entries are handled correctly in the calculation of H. We can do that by just ignoring them.

```
. generate R = 0
. generate H = 0
. quietly forvalues j = 1/5 {
  2.  replace R = R + (percent`j'/total)^2
  3.  replace H = H + (percent`j'/total) * ln(total/percent`j') if percent`j' > 0
  4. }
. format R H %4.3f
. tabdisp id1, cellvar(R H)
```

id1	R	H
world	0.308	1.358
tropical rain forest	0.922	0.196
subtropical forest	0.474	1.040
warm-temp. forest	0.367	1.237
cool-temp. forest	0.350	1.305
tundra	0.431	1.048
mid-temp. mesophytic forest	0.284	1.351
oak woodland	0.279	1.386
dry grassland	0.441	1.097
semidesert	0.441	0.940
desert	0.567	0.829

The results are naturally the same as before.

6 Variants as other variables

We can push beyond the details of section 2 to explore some variants of R and H. While interesting in their own right, they also serve to illustrate that generating results as variables places us close to getting related results as other variables.

Note first what is always true with equal probabilities of S categories:

$$R = \sum_{s=1}^{S} p_s^2 = S(1/S^2) = 1/S$$

and

$$H = \sum_{s=1}^{S} p_s \ln(1/p_s) = S(1/S) \ln S = \ln S$$

Hence, $1/R$ and $\exp(H)$ have interpretations as "numbers equivalents", measuring an equivalent number of equally abundant categories, even though only exceptionally will that number be an integer. It may also be helpful that $1/R$ and $\exp(H)$ are now measured on the same scale.

It follows immediately that you can `generate` such results directly as new variables.

7 Conclusion

`egen` is your friend for concentration and diversity calculations. Basic `egen` functions such as `pc()`, `total()`, `rowtotal()`, and `tag()` are invaluable in calculating group-wise measures such as repeat rate and entropy. Getting such results as new variables maximizes flexibility for later tabulation, graphics, and modeling work.

We have drawn short of showing how weights might appear in some datasets. In essence, that complication can be surmounted by inserting weights in expressions for the numerator and denominator of probability.

Furthermore, the virtue of `egen` in allowing concise code comes with a small price: it can be a little slow. If you are doing this repeatedly for large datasets, you may need more efficient code, but that is a different story.

References

Cox, N. J. 2005. Speaking Stata: The protean quantile plot. *Stata Journal* 5: 442–460. https://doi.org/10.1177/1536867X0500500312.

———. 2009. Speaking Stata: Rowwise. *Stata Journal* 9: 137–157. https://doi.org/10.1177/1536867X0900900107.

———. 2016. Speaking Stata: Multiple bar charts in table form. *Stata Journal* 16: 491–510. https://doi.org/10.1177/1536867X1601600214.

———. 2020. Speaking Stata: More ways for rowwise. *Stata Journal* 20: 481–488. https://doi.org/10.1177/1536867X20931007.

———. 2022. Speaking Stata: The largest five—a tale of tail values. *Stata Journal* 22: 446–459. https://doi.org/10.1177/1536867X221106436.

Gleick, J. 2011. *The Information: A History, A Theory, A Flood*. New York: Pantheon.

Good, I. J. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika* 40: 237–264. https://doi.org/10.2307/2333344.

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

Leinster, T. 2021. *Entropy and Diversity: The Axiomatic Approach*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108963558.

MacKay, D. J. C. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.

Schmitt, S. A. 1969. *Measuring Uncertainty: An Elementary Introduction to Bayesian Statistics*. Reading, MA: Addison-Wesley.

Sloane, N. J. A., and A. D. Wyner, eds. 1993. *Claude Elwood Shannon: Collected Papers*. New York: Institute of Electrical and Electronics Engineers.

Smith, W. G. 1913. Raunkiaer's "life-forms" and statistical methods. *Journal of Ecology* 1: 16–26. https://doi.org/10.2307/2255456.

Soni, J., and R. Goodman. 2017. *A Mind at Play: How Claude Shannon Invented the Information Age*. New York: Simon and Schuster.

Theil, H. 1967. *Economics and Information Theory*. Amsterdam: North-Holland.

———. 1972. *Statistical Decomposition Analysis: With Applications in the Social and Administrative Sciences*. Amsterdam: North-Holland.

Westman, W. E., and R. K. Peet. 1982. Robert H. Whittaker (1920–1980): The man and his work. *Vegetatio* 48: 97–122. https://doi.org/10.1007/BF00726879.

Whittaker, R. H. 1975. *Communities and Ecosystems*. New York: Macmillan.