Graphics for ordinal outcomes or predictors

Nicholas J. Cox

Department of Geography

Durham University

Ordinal data are common

Ordered or ordinal variables are common in many fields
— and a leading data type in some.

Example: opinion **grades** from strongly disagree to strongly agree.

Graphics obvious … or obscure?

Graphics for ordinal data may appear to range from obvious (bar charts) to more powerful but obscure (mosaic plots? correspondence analysis?) .

This problem receives little direct attention in

Friendly, M. 2000. *Visualizing Categorical Data*. Cary, NC: SAS Institute.

Friendly, M. and Meyer, D. 2016. *Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data*. Boca Raton, FL: CRC Press.

This presentation

This presentation surveys various graphics commands made public through the *Stata Journal* (*SJ*) or SSC

— principally friendlier and more flexible bar charts and dedicated distribution or quantile plots.

Specific commands include `tabplot`, `floatplot`, `qplot`, and `distplot`.

Detailed code will be posted after the presentation as a `do`-file.

Something old, something new

Talk at the London meeting 2001
https://www.stata.com/meeting/7uk/cox1.pdf

Partial write-up 2004 *Stata Journal* 4(2):190--215
https://www.stata-journal.com/article.html?article=gr0004

...

`floatplot` posted on SSC 2021

Rule 1!

I think that rule 1 for the statistician is *examine the data*.

Irving John Good

(1916—2009)

Examples of grades

Repair record 1978 `rep78` in the auto dataset ordered 1 to 5

Nahuatl (Aztec language) adjectives of hotness of peppers: coco, cocopatic, cocopetz-patic, cocopetztic, copetzquauitl, cocopalatic

Whitewater difficulty: Easy, Novice, Intermediate, Advanced, Expert, Extreme (a rapid classification…)

Many other examples: Lord (1995); Strachan and Moseley (2017)

Solution 1: Stacked bar chart

Use `graph bar` or `graph hbar` or `catplot` (SSC) for convenience

+ Well understood and frequently used plot type

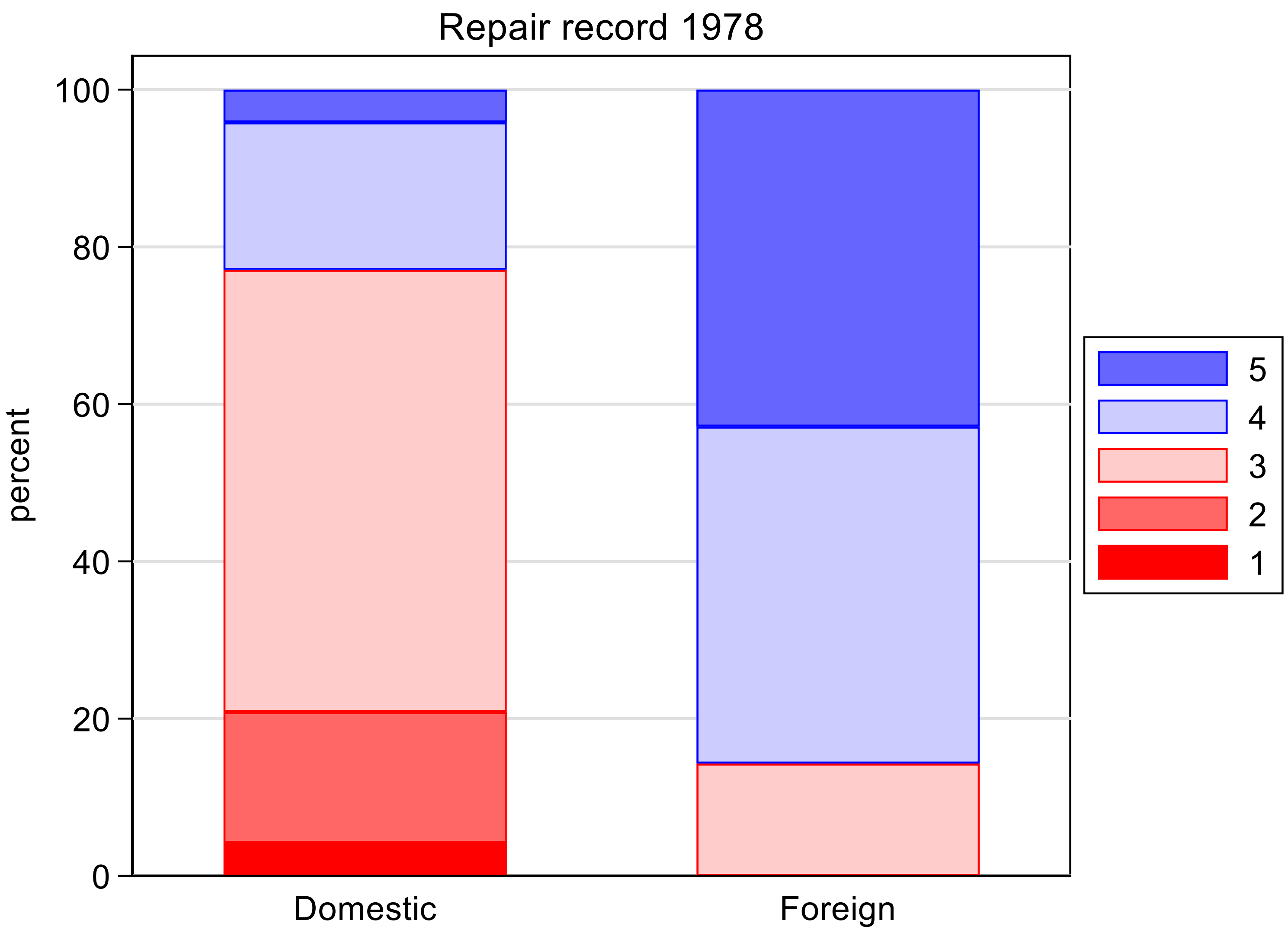
+ Respects ordering of categories

+ Suitable colour scheme possible

– Hard to see absent (zero) or rare categories

– Dominated by adding up to 100%

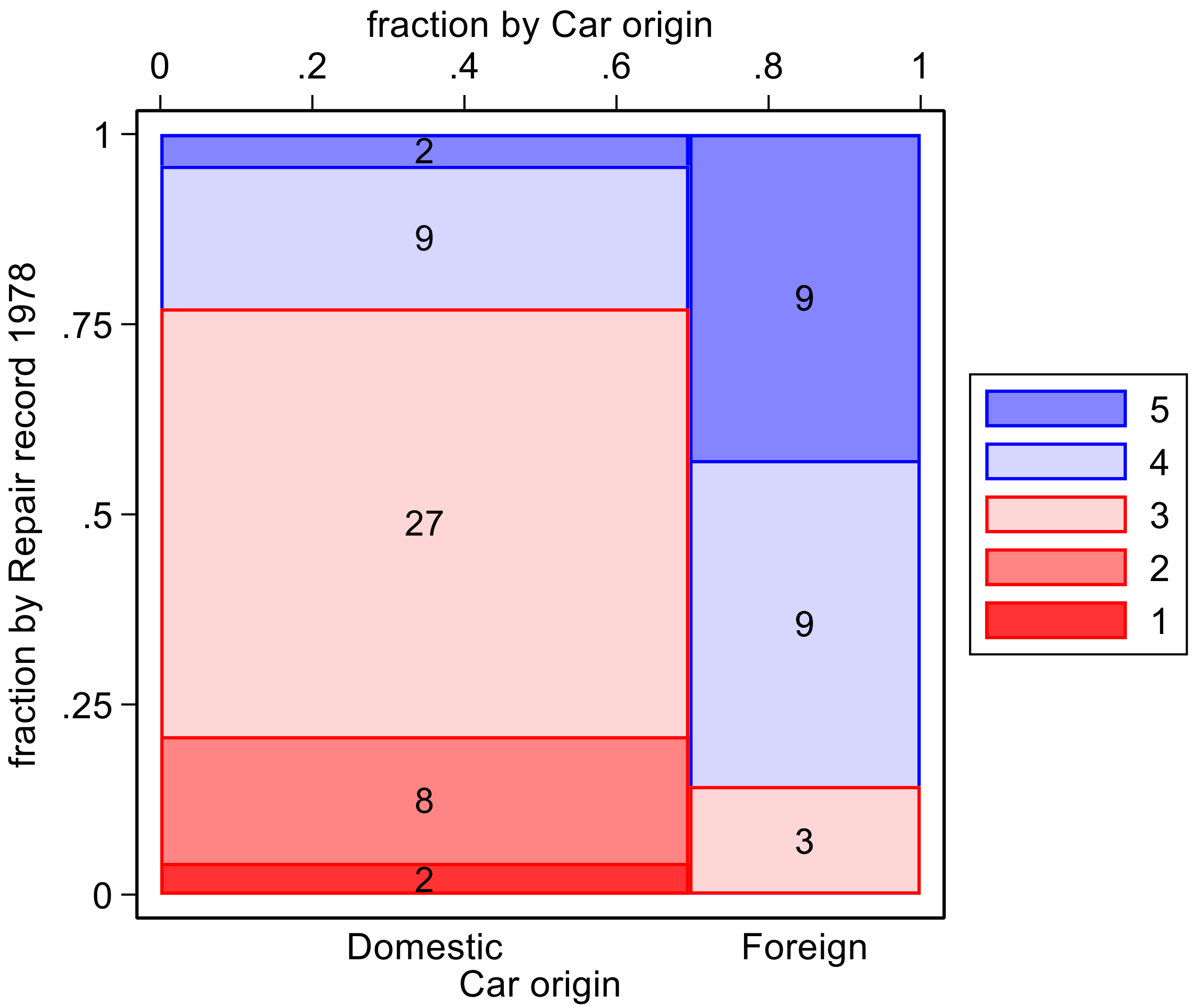
– Hard to annotate with numeric detail


Repair record 1978
percent
100
80
60
40
20
0
Domestic
Foreign
5
4
3
2
1

Solution 2: Mosaic or spine plot

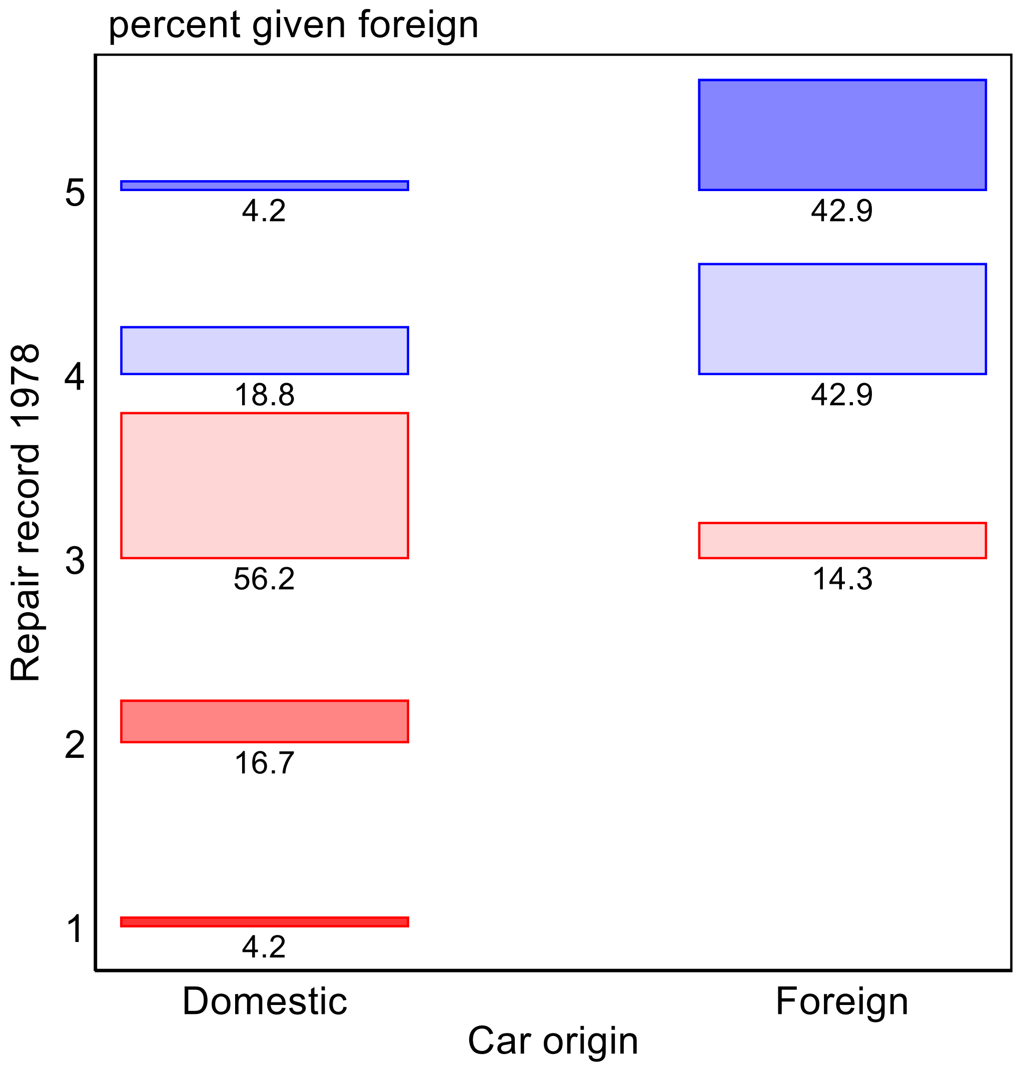
Use `spineplot` (*Stata Journal* 8: 105–121 and 16: 521–522)

Better for seeing two-way distribution and numeric annotation?


fraction by Car origin
0
.2
.4
.6
.8
1
fraction by Repair record 1978
1
.75
.5
.25
0
2
9
27
8
2
9
9
3
Domestic
Foreign
Car origin
5
4
3
2
1

Solution 3: Two-way bar chart

Use `tabplot` (*Stata Journal* 4: 190–215; 12: 549–561; 16: 491–510; 17: 779; 20: 757–758)


percent given foreign
Repair record 1978
5
4
3
2
1
4.2
18.8
56.2
16.7
4.2
42.9
42.9
14.3
Domestic
Foreign
Car origin

Antarctic Peninsula example

The more rounded the stones on a beach, the longer it has been exposed as such.

Bentley, M.J., Hodgson, D.A., Smith, J.A. and Cox, N.J. 2005. Relative sea level curves for the South Shetland Islands and Marguerite Bay, Antarctic Peninsula. *Quaternary Science Reviews* 24: 1203–1216.

Powers scale of roundness

Powers, M.C. 1953.

A new roundness scale for sedimentary particles.

Journal of Sedimentary Petrology 23: 117–119.

suggested an ordered scale and gave example photographs as well as discussing its relation to earlier measurements of roundness.

Very angular

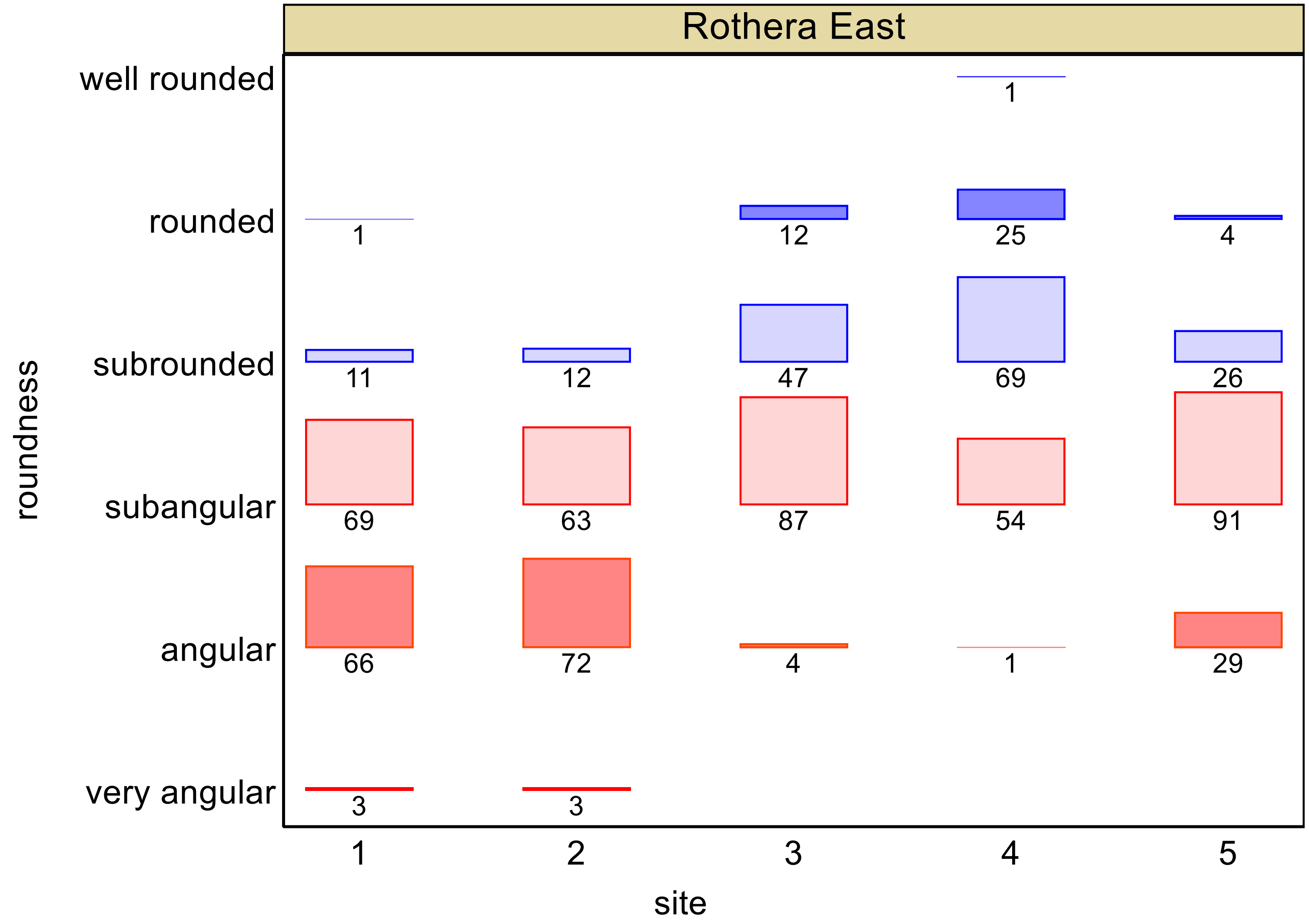
Angular

Subangular

Subrounded

Rounded

Well rounded


Rothera East
roundness
well rounded
rounded
subrounded
subangular
angular
very angular
1
1
12
25
4
11
12
47
69
26
69
63
87
54
91
66
72
4
1
29
3
3
1
2
3
4
5
site

Problem: Ordering the other axis

The horizontal axis for the previous graph used arbitrary site identifiers.

Here, and often, we need to think up a better choice of order.

Howard Wainer mocked alphabetical order: Alabama first! Austria first!
Scots will add: Aberdeen first?

Solution 4: Order the other axis

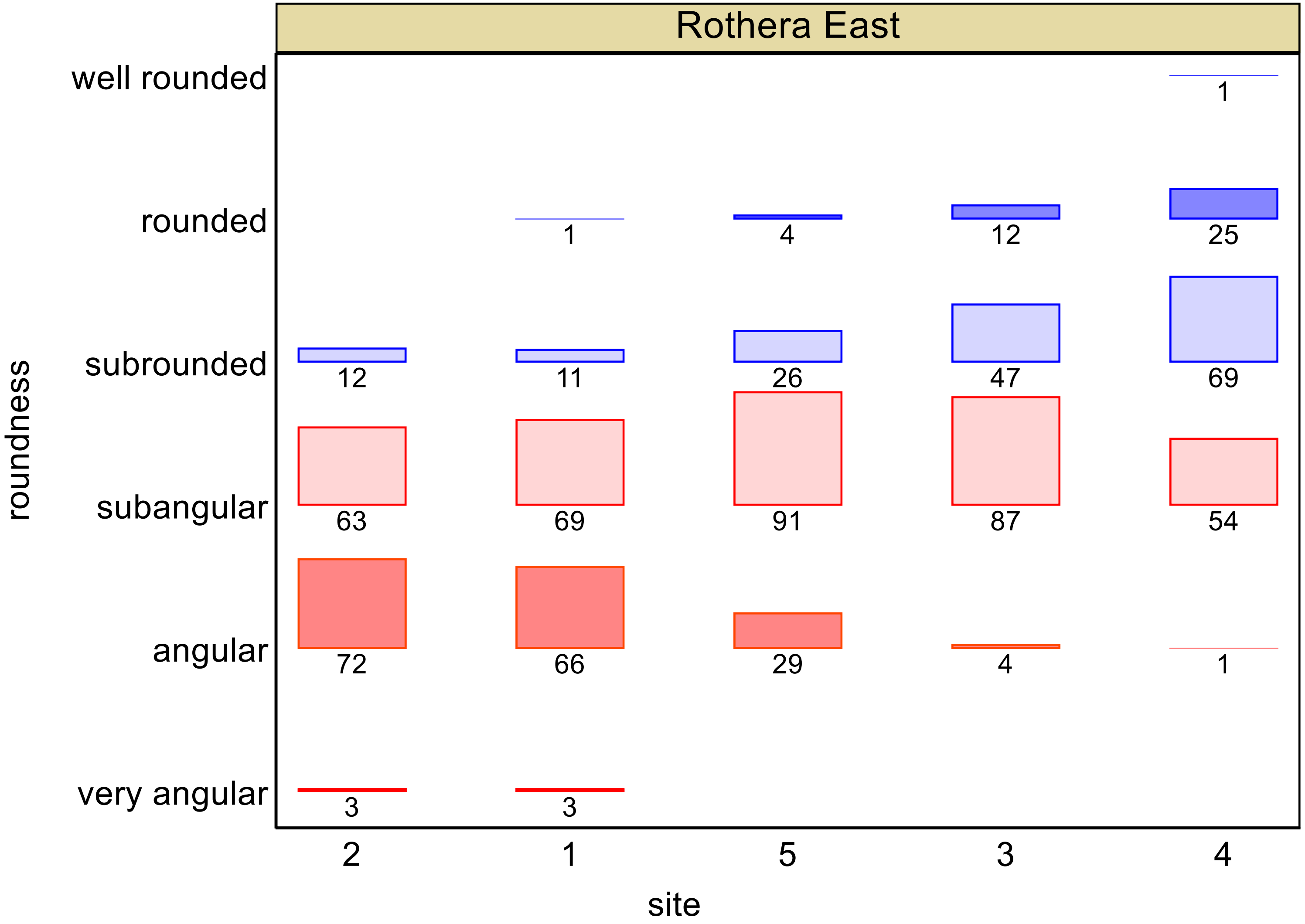
For a detailed discussion, including the `myaxis` command, see

https://www.statalist.org/forums/forum/general-stata-discussion/general/1598767-myaxis-available-from-ssc-reorder-categorical-variables-especially-for-later-table-or-graph-use

Stata Journal 21(3): in press (2021)

Despite what measurement scale zealots say, ordering by mean of the ordinal outcome often works well.

There are other choices, e.g. fraction of one of more categories.


Rothera East
roundness
well rounded
rounded
subrounded
subangular
angular
very angular
site
2
1
5
3
4
1
1
4
12
25
12
11
26
47
69
63
69
91
87
54
72
66
29
4
1
3
3

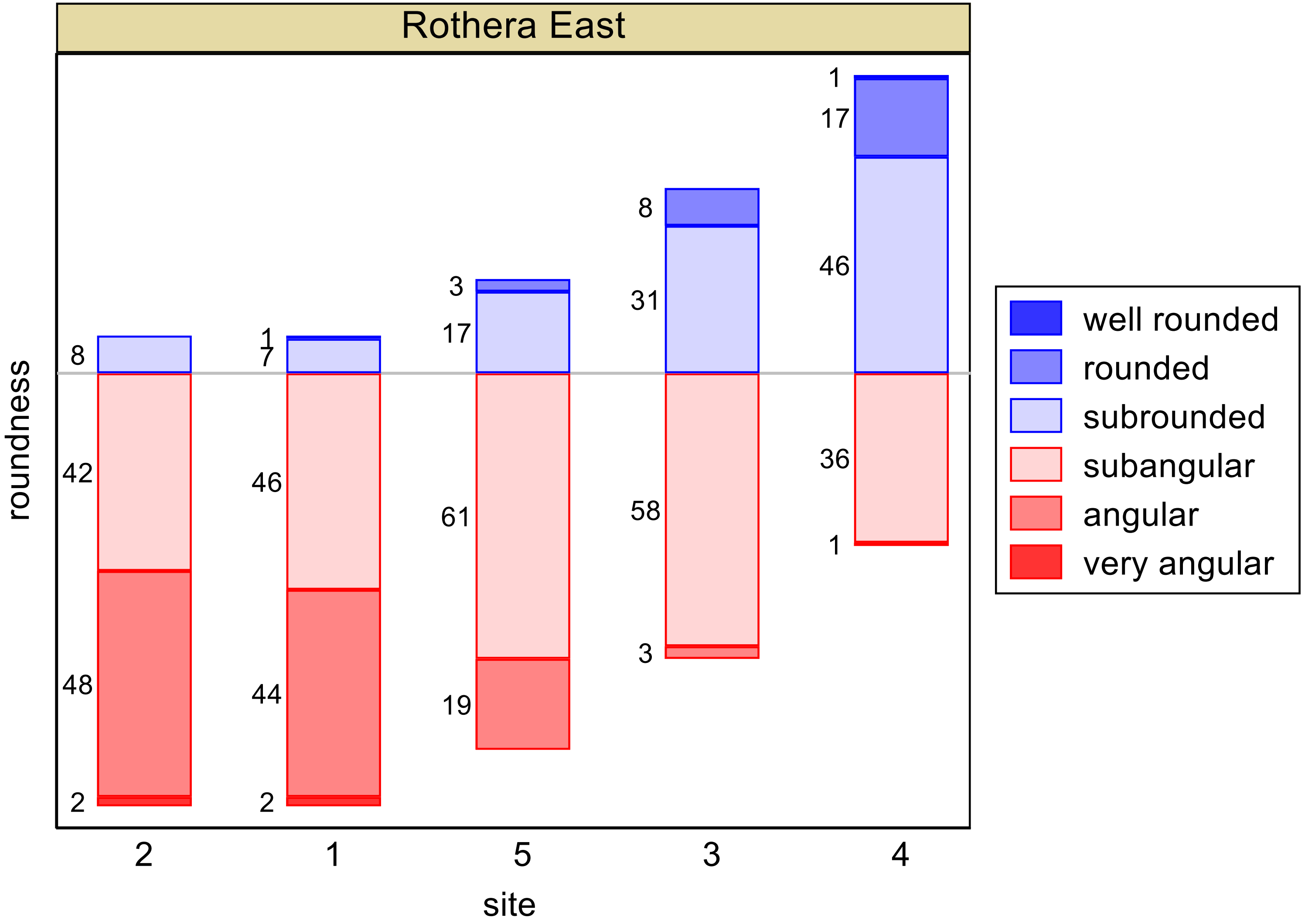
Solution 5: Sliding or floating bar chart

Anchor distribution on or between middle categories

2003 solution: `slideplot` (SSC), a wrapper for `graph bar` etc.
2021 solution: `floatplot` (SSC), a wrapper for `twoway rbar`

Robbins and Heiberger call these *diverging stacked bar charts*.

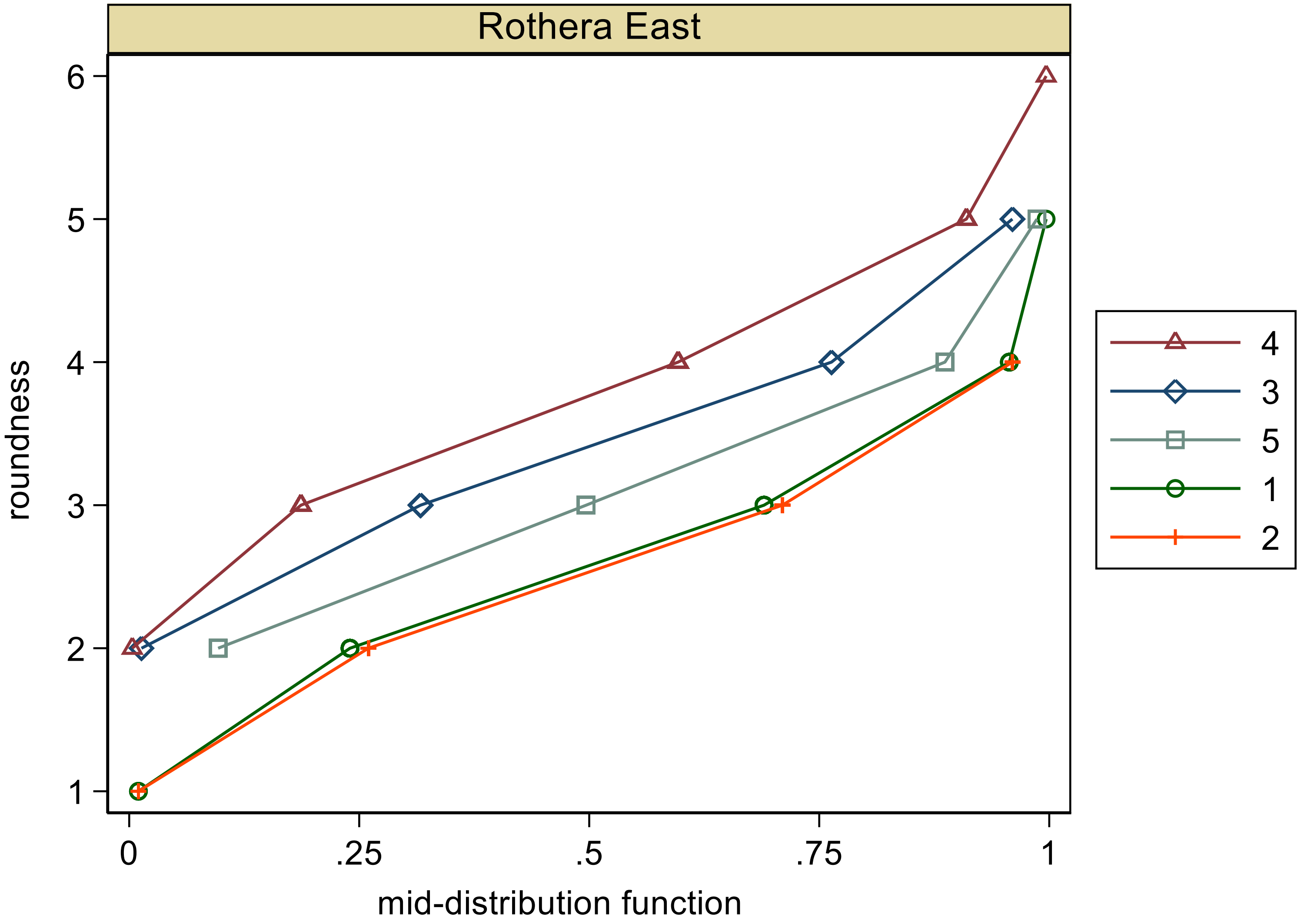
`floatplot`, unlike `slideplot`, allows the bar for a middle category (e.g. 3 on a 5-point scale) to straddle the outcome axis)

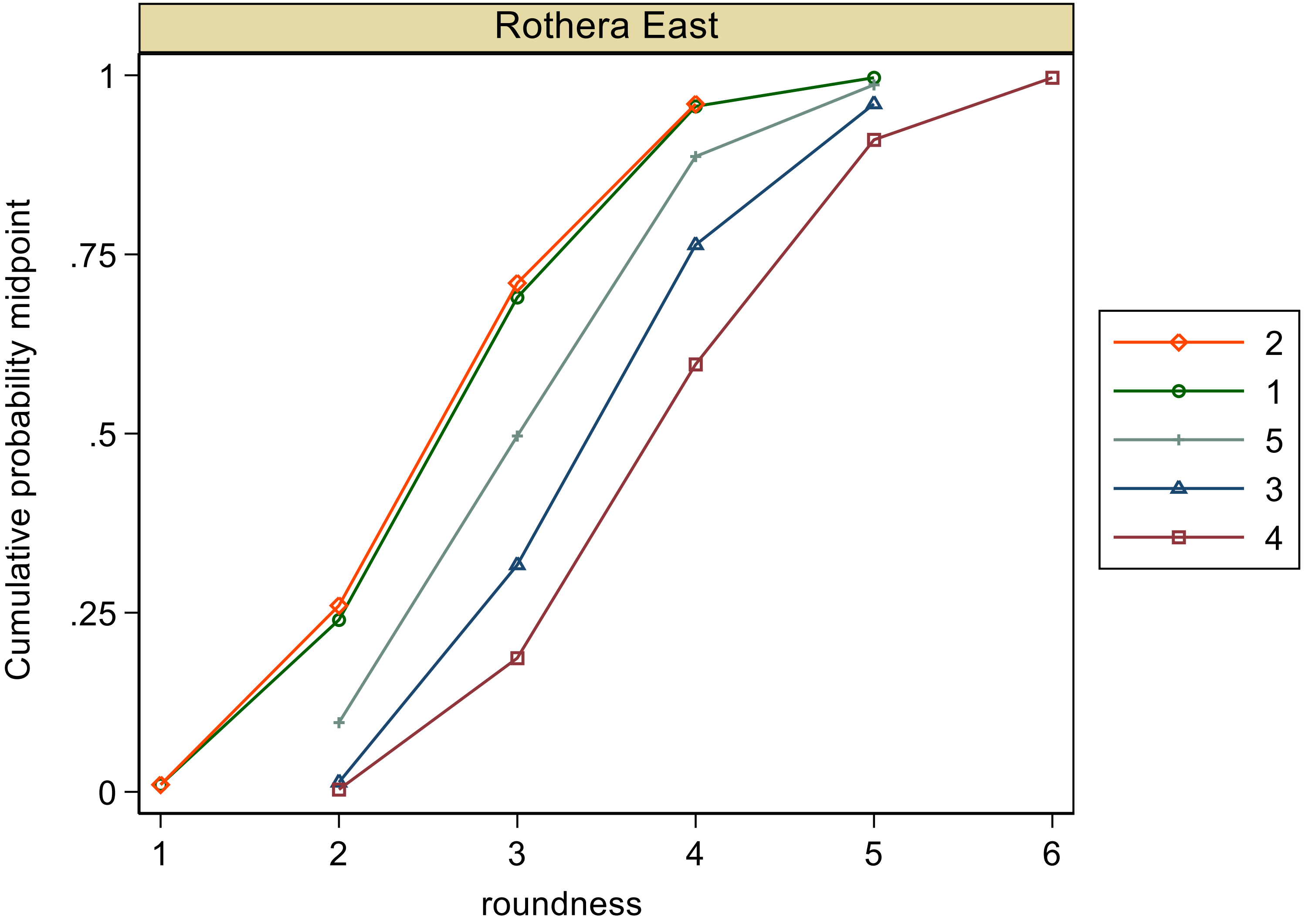

Rothera East
roundness
site
2
1
5
3
4
8
42
48
2
1
7
46
44
2
3
17
61
19
8
31
58
3
1
17
46
36
1
well rounded
rounded
subrounded
subangular
angular
very angular

Solution 6: Plot cumulative probabilities

With ordinal scales, cumulative probabilities carry all the information.
First twist: plot midpoints of each bin on cumulative probability scale.
Plotting $P(Y < y)$ means that the lowest cumulative probability is 0.
Plotting $P(Y \leq y)$ means that the highest cumulative probability is 1.
Plotting their mean reduces either problem.

Ridit	Bross (1958)
Split fraction below	Tukey (1977, 496-497)
Mid-distribution function	Parzen (1993, 3295)
Grade function	Haberman (1996, 240-241).


Rothera East
roundness
6
5
4
3
2
1
0
.25
.5
.75
1
mid-distribution function
4
3
5
1
2


Rothera East
Cumulative probability midpoint
1
.75
.5
.25
0
1
2
3
4
5
6
roundness
2
1
5
3
4

`qplot` for quantile plots

Stata Technical Bulletin 51: 16–18 (1999) … *SJ* 5: 442–460 (2005) … *SJ* 19: 748 (2019)

`distplot` for (empirical) (cumulative) distribution (function) plots

Stata Technical Bulletin 51: 12–16 (1999) … *SJ* 19: 260 (2019)

What is the outcome or response?

A principle for graphics is to plot the outcome or response on the vertical or y axis.

It is honoured even when plotting histograms or probability density or survival functions.

With ordinal outcomes, what precisely is the response?

Solution 7: Transform the probability scale

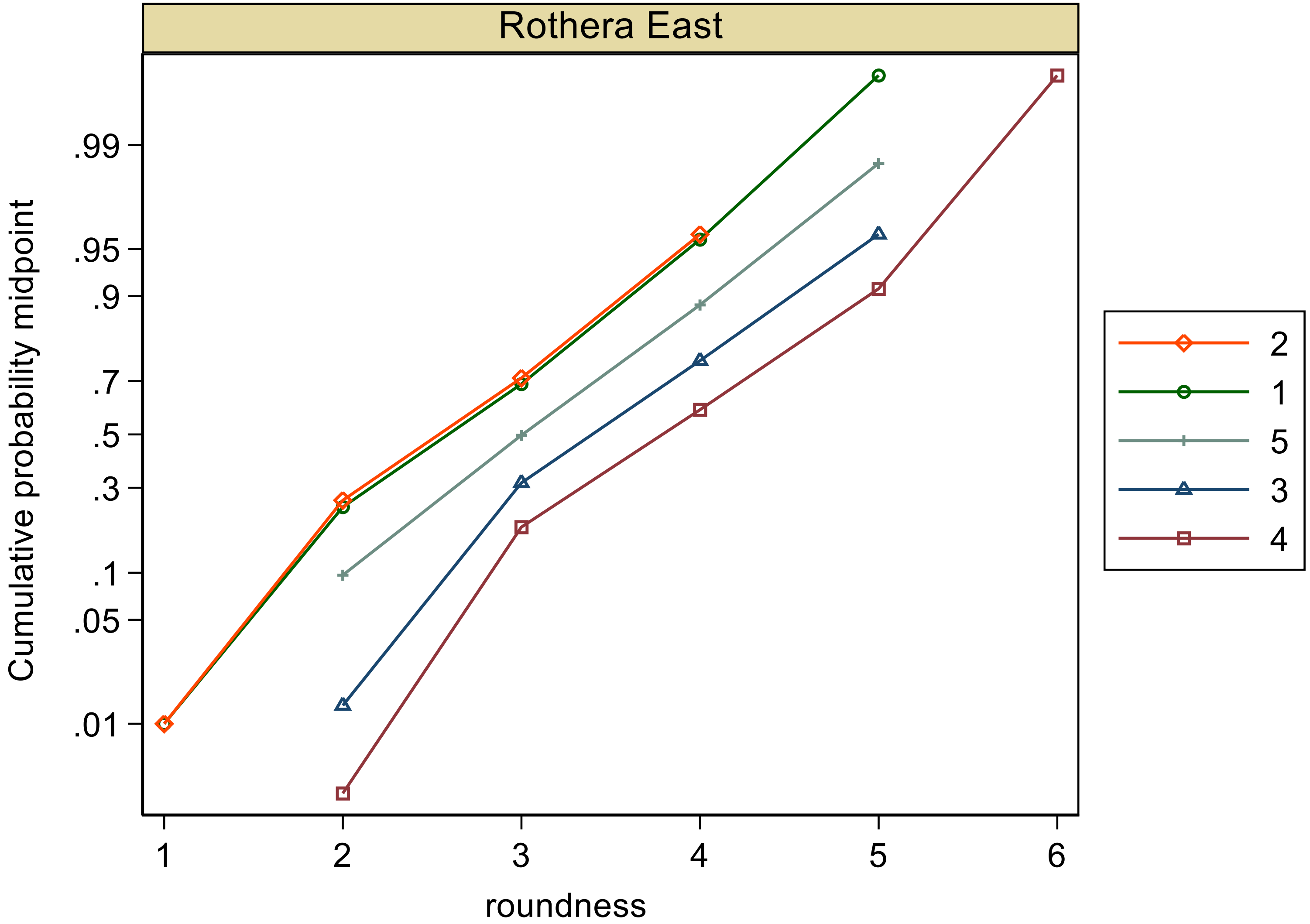
In practice, cumulative distributions are often S-shaped: intermediate categories are more common than extreme categories.

Is this nature, society or classifiers?

Empirically, we might benefit from a transformation such as logit.

However, if you use a transformed scale, do show probabilities as labels.

See `mylabels` (SSC) or *Stata Journal* 8: 142–145 (2008)


Rothera East
Cumulative probability midpoint
.99
.95
.9
.7
.5
.3
.1
.05
.01
1
2
3
4
5
6
roundness
2
1
5
3
4

Seriation

In archaeology and various environmental sciences, there is a problem of **seriation**, at its simplest finding the best ordering of rows and columns given a data matrix.

Example: put archaeological sites in approximate date order according to which artifacts have been found where.

How to get numeric scores

Logistic scores: let the probabilities indicate latent scores

Mosteller, F. and Tukey, J.W. 1977. *Data Analysis and Regression*. Reading, MA: Addison-Wesley.

Correspondence analysis is much better established.

Graphs force us to note the unexpected; nothing could be more important.

John Wilder Tukey

1915–2000

Using the data to guide the data analysis is almost as dangerous as not doing so.

Frank E. Harrell Jr

1951-

That's all Folks!

All graphs use Stata scheme `s1color`, which I strongly recommend as a lazy but good default.

This font is Georgia.

`This font is Lucida Console.`

Sources of quotations

Williams, J.S. 1971. Two nonstandard methods of inference for single parameter distributions. In Godambe, V.P. and Sprott, D.A. (Eds) *Foundations of Statistical Inference*. Toronto: Holt, Rinehart and Winston, 314—329 [Comment by I.J. Good, 326—327; quotation is on 326]

Tukey, J.W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley. p.157.

Harrell, F.E. 2015. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Cham: Springer. p.ix

References on grade schemes

Lord, J. 1995. *Sizes: The Illustrated Encyclopedia*. New York: HarperCollins.

Strachan, J. and Moseley, J. 2017. *The Order of Things: How Hierarchies Help Us Make Sense of the World*. London: Robinson.

Ridits, or roses by any other name

Bross, I.D.J. 1958. How to use ridit analysis. *Biometrics* 14: 18–38.

Haberman, S.J. 1996. *Advanced Statistics Volume I: Description of Populations*. New York: Springer.

Parzen, E. 1993. Change PP plot and continuous sample quantile function. *Communications in Statistics –Theory and Methods* 22: 3287–3304.

Tukey, J.W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.