SEPT. 9-11, 2024 –BRITISH SOCIETY

FOR POPULATION STUDIES

BATH, UK

ENTROPY OF ORDINAL INPUTS IN A SOCIAL DATA SCIENCE CONTEXT: ONTIC AND STATISTICAL OPTIONS

BY WENDY OLSEN & ZIYANG ZHOU

MANCHESTER 1824
The University of Manchester

https://github.com/WendyOlsen/entropyOrdinalData2024

We acknowledge the Asian Barometers data for India for 2019.
Our open-source code is on Github.

TYPICAL RESEARCH QUESTION

WHAT FACTORS EXPLAIN OUTCOMES OR ASSOCIATIONS, WHEN SOME VARIABLES ARE ORDINAL?

THIS PAPER'S QUESTIONS:

WHAT IS THE APPROPRIATE ONTIC WAY TO DEAL WITH ORDINAL INFORMATION AT STAGE 1 OF A PROJECT?

& WHAT IMPACT DOES CUMULATIVE CODING HAVE ON RESULTS?

MANCHESTER 1824
The University of Manchester

Going beyond data science to social data science

REVIEW OF LITERATURE

RETRODUCTION FROM DATA TO AN ORDINAL OR CARDINAL REALITY

MANCHESTER 1824

The University of Manchester

REVIEW OF LITERATURE

TWO POSSIBLE ENCODINGS

IN R WE USED ONE-HOT ENCODING TO GAIN CUMULATIVE CODING

(10 PAGES)

IN STATA IT IS JUST 20 LINES OF CODE FOR EACH 10 ORDERED LEVELS

MANCHESTER
1824
The University of Manchester

GAPS IN THE LITERATURE

DATA SCIENCE – NO STUDIES OF CUMULATIVE ENCODING

NATURAL SCIENCE – NEEDS ENTROPY MEASURES TO SUIT ORDINAL INPUTS.

SOCIAL SCIENCE – USES SEM, MCA, ETC. (VERY GOOD).

MANCHESTER 1824
The University of Manchester

* SUPERVISED LEARNING: AIM FOR EXPLAINING SOME OUTCOMES.
* UNSUPERVISED LEARNING: AIM FOR DISCERNING ASSOCIATIONS, WITHOUT LOSING THE ORDINAL STATUS OF INPUT SIGNALS.

Entropy is a measure of the uninformativeness of any data set.[1,2] A vector has entropy.

Ordinal variables' entropy can be measured if we discretize them.

For a signalling event, X, with n possible values (outcomes), $x_1, x_2 \dots, x_n$
each outcome having probability, $p_1, p_2 \dots, p_n$, the entropy of X, denoted $\mathrm{H}(X)$, is given by

$$\mathrm{H}(X) = -\sum_{i=1}^{n} p_i \ln p_i$$

Our manual calculations matched the R package[4] perfectly (12 digits accuracy). (see Github code)

github.com/WendyOlsen/entropyRSS2024 (Z Zhou & WO)

STATISTICAL METHODS TO USE THE DISCRETIZED ORDINAL SIGNALS

Start with a model of a distribution

Multinomial distribution or an ordered distribution of levels

Regularize and shrink

Hausser & Strimmer, 2009, 2022 (R package entropy)

(see Github code and notice discretization routines in R base, arules, etc.)

github.com/WendyOlsen/entropyRSS2024 (Z Zhou & WO)

STATISTICAL METHODS IN HAUSSER-STRIMMER

The *H* estimate is a biased estimate

although the ML estimate $\theta_{\mathrm{k}}^{\mathrm{ML}}$ is not biased.

$\widehat{H_k^{shr}}$ = - $\sum_{k=1}^{q} \theta_k^{shr} * \ln\left(\theta_k^{shr}\right)$ measured in nats Eq. 2

(shr = shrinkage estimate, Hausser-Strimmer, 2009: 1473)

We want a standard error for entropy.

The lambda parameter averages two models:

$\theta_k^{shr} = \lambda t_k + (1-\lambda)\theta_k^{ML}$ Eq. 3

The mean-squared error (MSE) of *H* is used by Hausser-Strimmer (2009).

It is feasible, as James-Stein estimator equivalent to a Bayesian estimator.

DATA AND METHODS USED HERE

METHOD 1: ENTROPY ESTIMATION (EXACT MATCH TO THE ENTROPY PACKAGE JAMES-STEIN ESTIMATES)

METHOD 2: REGRESSION ESTIMATES WITH A VARIETY OF ORDINAL VARIABLES

METHOD 3: SIMULATION AND MSE

LIKERT SCALES ARE DISTINCT-ORDINAL

The entity is an attitude. Each Attitude is **distinctive**.
The ontology of attitudes is unlike that of education.

See paper with references in our Github.

See our earlier publications on gender norms

Note: In the 2019 Asian Barometers - India

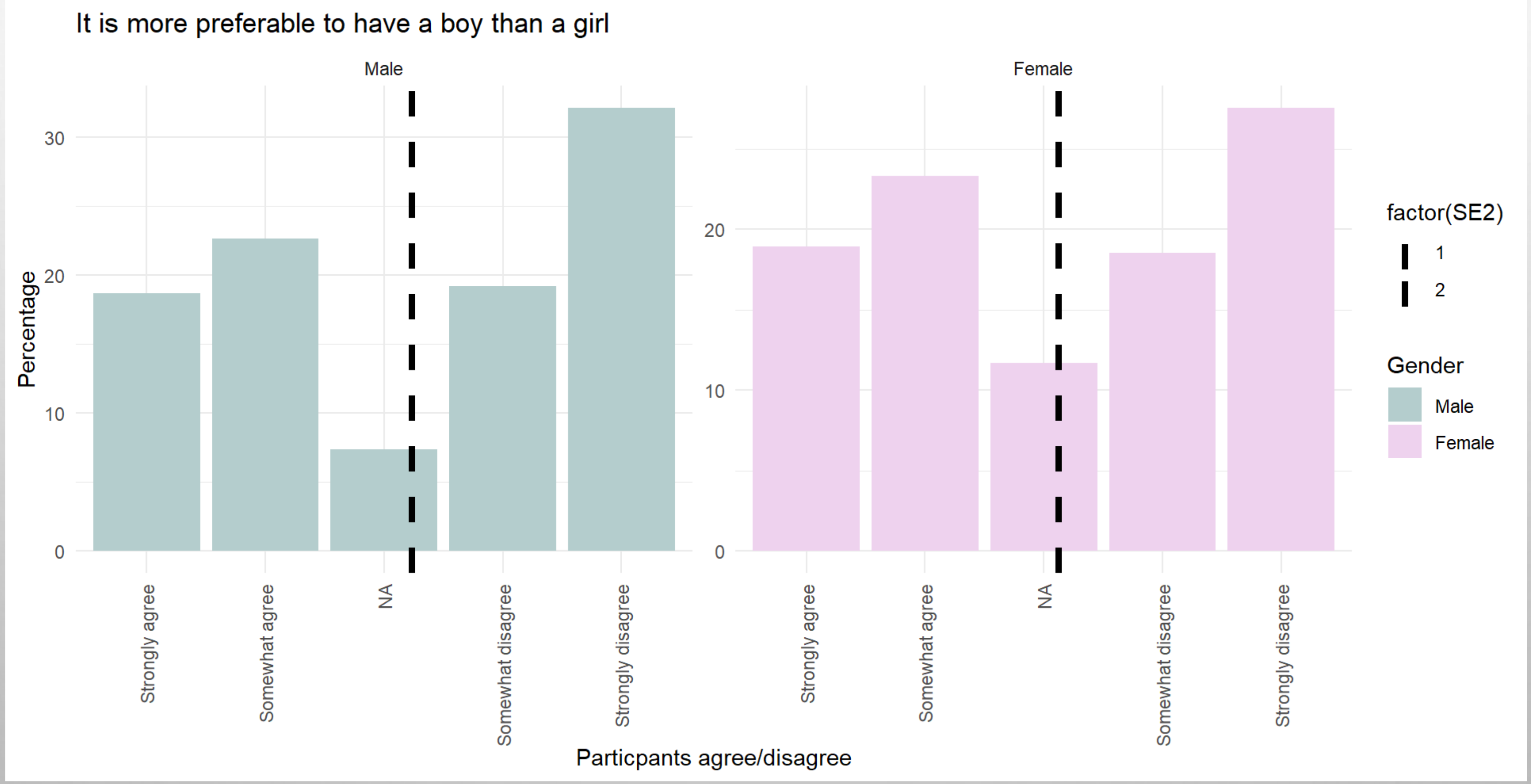
Sexism was embedded in questions ←→ desirability bias of a patriarchal gender norm

LIKERT SCALE – ASIAN BAROMETERS – FIGURE 1

DOES THE RESPONDENT PREFER A BOY OR A GIRL, IF JUST 1 CHILD IS TO BE BORN?

THE ENTROPY MEASURES DEVIATIONS FROM UNIFORM.

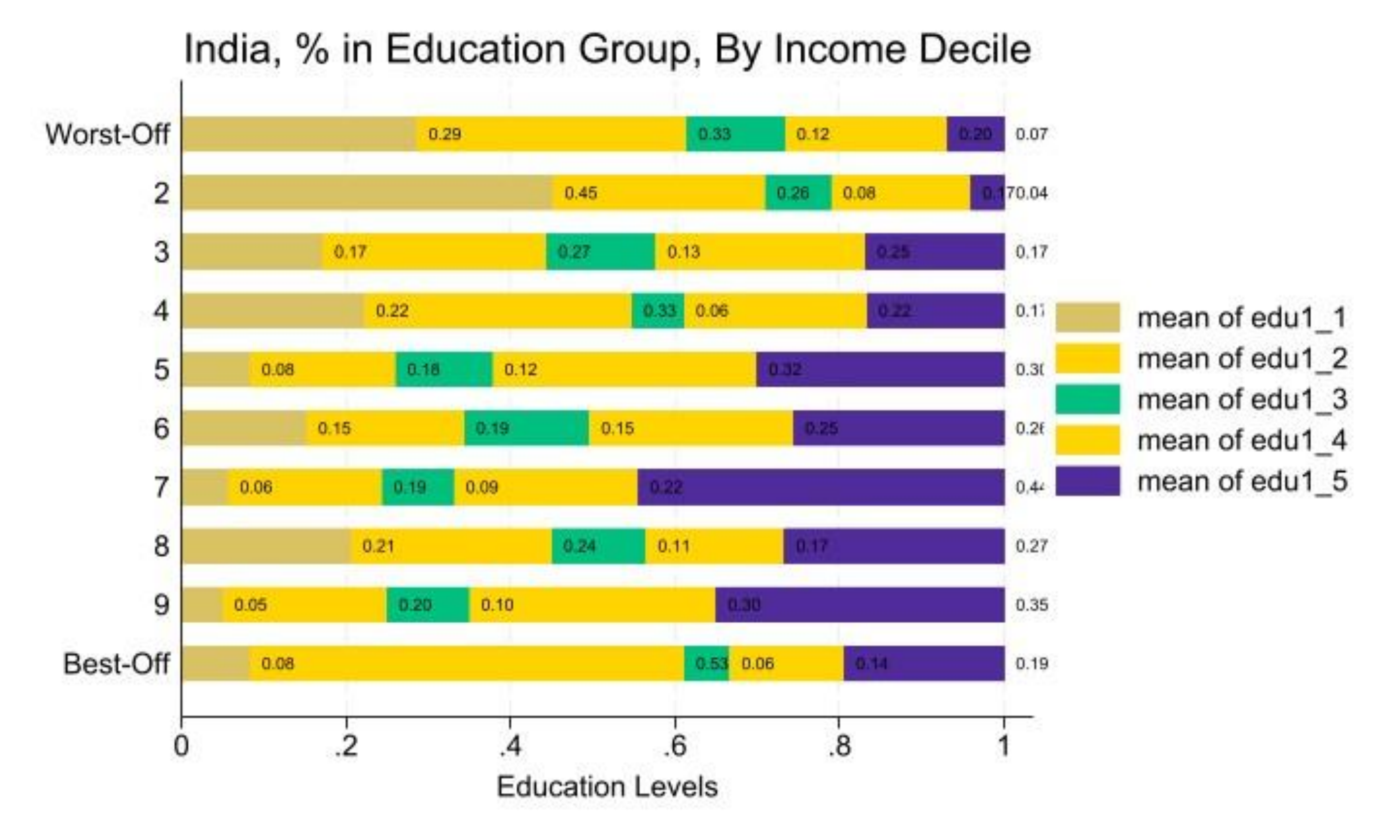
Boy-Preferring



Not Boy-preferring

MANCHESTER 1824
The University of Manchester

SAMPLE WIDELY DISTRIBUTED IN 19 STATES



MANCHESTER
1824
The University of Manchester

"CUMULATIVE ORDINAL" REFLECTS THE REALITY OF COMBINING PRIMARY, SECONDARY, AND OPTIONAL LATER SCHOOLING.

- *ONTIC NATURE OF THE THING TO WHICH WE REFER*
  - *DISTINCT ORDINAL VS. CUMULATIVE ORDINAL*

RSE = $\frac{H}{H_{max}}$ EQ. 5

  - *THE DEVIATION OF THE TWO MEASURES FOR EDUCATION IS EMPIRICALLY DIFFERENT.*
  - AIC = -$2\mathrm{LOG}(\hat{L})$ + 2K EQ. 6
  - *THE REGRESSION RESULTS ALSO DIFFERED.*

MANCHESTER 1824
The University of Manchester

THE RESULTS (ENTROPY AND REGRESSION TESTS)

Methods Used	Sample:	Overall Test:	Results:	If multiple vectors:
First entropy measures.	N=5,318 Adults only 19 states of India	Entropy differed by <5% between the group of binaries for the distinct vs cumulative coding. Distinct was lower. But in groups of variables, this result switched.	IF ONE VECTOR: Higher entropy implied less informative education data. Cumulative was slightly less informative.	The dataframe entropy depends in part on mutual entropy. The results were switched. Regression results also ambiguous.

MANCHESTER
1824
The University of Manchester

THE RESULTS (SIMULATION TEST ON EDUCATION)

Methods Used	Sample:	H for Discrete Education	H for Cumulative Education	
Simulation Repeat 1000 samples with replacement from 5,318 Multinomial distribution	N=5,318 Adults only 19 states of India	The 95% interval for H, the entropy, was {1.5384, 1.5171} This range is about 2% of the raw H value in nats. Its MSE is 0.027.	NOW: PREVIOUSLY: Cumulative was slightly less informative.	

THE RESULTS (REGRESSION TESTS)

Methods Used	Sample:	Overall Test:	Results:	
Second regressions. Ordered probit. Cumulative coding vs. distinct coding.	N=5,318 Adults only 19 states of India			
		Ran 3 sets of 3 regressions		
Compare the AIC	N is same, but p rises to q and differs.		We compare AIC using the Δdf as a ChiSquared.	

SUMMARY AND POINTERS FORWARD

SURVEY OF THE MAIN POINTS TODAY

- ENTROPY IS SLIGHTLY DIFFERENT FOR CUMULATIVE ORDINAL VS DISTINCT ORDINAL VARIABLES.
- & CAN INTRODUCE RANKED LEVELS. APPLY CHEBYSHEV'S INEQUALITY. MULTIPLE TIMES (SUM OF INDEP. R.V.S)

+OBV. SUPERVISION IS NEEDED.

POINTERS TO HOW TO CARRY OUT SUPERVISION

- STAGE 1 ONTIC

& CONSIDER REFERENT

- **STAGE 2** DISCRETIZE
- STAGE 3 RE-GROUP

FUTURE RESEARCH

ENTROPY OF WASTE FLOWS VS. METAL INGOTS

- ENTROPY IS DIFFERENT FOR CUMULATIVE ORDINAL VS DISTINCT ORDINAL VARIABLES
- CHEMISTRY, PHYSICS, MEDICAL & RADIOGRAPHY CAN USE THE SOLUTIONS

ENTROPY IN A MULTI-STAGE ANALYSIS

- STAGE 1 APPLY PHILOSOPHICAL KNOWLEDGE TO DATA SCIENCE
  - ORDINALISE AND CARDINALIZE THE INPUT DATA
  - RANK 1< RANK 2 < RANK 3
  - THIS IS NOT A MULTINOMIAL DISTRIBUTION
- **STAGE 2** DISCRETIZE AFTER ENCODING IN A NOVEL WAY
- STAGE 3 THEN RE-GROUP THE VARIABLE TO GET THE WHOLE PICTURE

REFERENCES 1

ENTROPY

Open Source Code –thanks to Ziyang Zhou - for Entropy Calculations – uses one-hot encoding. github.com/WendyOlsen/entropyOrdinalData2024

Borsboom, Mellenbergh, and van Heerden (2003) The Theoretical Status of Latent Variables, *Psychological Review,* DOI 10.1037/0033-295X.

Watts, S., & Crow, L. (2019), Big variates — visualising and identifying key variables in a multivariate world, *Nuclear Instruments and Methods in Physics Research Section A*, 940, 441-447. https://doi.org/10.1016/j.nima.2019.06.060

SOFTWARE PACKAGE ENTROPY IN R

- HAUSSER, JEAN, AND KORBINIAN STRIMMER (2022), *PACKAGE 'ENTROPY' (SIC)*, OCTOBER 13. CRAN REPOSITORY, HTTPS://STRIMMERLAB.GITHUB.IO/SOFTWARE/ENTROPY/.
- SEE WEB-PAGE *ESTIMATION OF ENTROPY, MUTUAL INFORMATION AND RELATED QUANTITIES,* HTTPS://STRIMMERLAB.GITHUB.IO/, ACCESSED SEPTEMBER 2024.

REFERENCES 2

HAUSSER, JEAN, AND KORBINIAN STRIMMER (2009) ENTROPY INFERENCE AND THE JAMES-STEIN ESTIMATOR, WITH APPLICATION TO NONLINEAR GENE ASSOCIATION NETWORKS, *JOURNAL OF MACHINE LEARNING RESEARCH*, 10, 1469-1484. URL HTTPS://JMLR.CSAIL.MIT.EDU/PAPERS/V10/HAUSSER09A.HTML, ACCESSED AUG. 2024.

FURTHER INFORMATION SOURCE

- *SOURCE: _ASIAN BAROMETER PROJECT (2018-2021), INDIA, URL HTTPS://WWW.LOKNITI.ORG/PAGE/ACCESSING-DATA AND_ HTTPS://WWW.ASIANBAROMETER.ORG/DATAR?PAGE=D10, AVAILABLE FOR ACADEMIC PURPOSES ONLY ON AN OPEN ACCESS BASIS. WRITE TO THE DATA PROVIDERS PERSONALLY TO GET ACCESS [ONLINE DATASET], (ACCESSED AUG 2024; SCROLL DOWN TO THE BOTTOM TO SEE THE FORM WHICH YOU WILL FILL IN.)
- _ACKNOWLEDGEMENT:_
- _DATA ANALYZED IN THIS ARTICLE WERE COLLECTED BY THE ASIAN BAROMETER PROJECT (2018-2021), CO-DIRECTED BY PROFESSORS YUN-HAN CHU AND RECEIVED FUNDING FROM THE NATIONAL SCIENCE AND TECHNOLOGY COUNCIL, ACADEMIA SINICA AND NATIONAL TAIWAN UNIVERSITY. THE ASIAN BAROMETER PROJECT OFFICE (WWW.ASIANBAROMETER.ORG) IS SOLELY RESPONSIBLE FOR DATA DISTRIBUTION. THE AUTHOR(S) APPRECIATE THE ASSISTANCE IN PROVIDING DATA BY THE INSTITUTES AND INDIVIDUALS AFOREMENTIONED. THE VIEWS EXPRESSED HEREIN ARE THE AUTHORS' OWN._

- DOCUMENTATION OF THE DATASET FOR INDIA
- THE TECHNICAL REPORT WILL ARRIVE INSIDE THE DATASET ZIP FILE, AFTER YOU REGISTER FOR THE DATA.
- IF IN DOUBT, CONTACT EMAIL: ASIANBAROMETER@NTU.EDU.TW