SEPT. 3-5, 2024 – ROYAL STATISTICAL SOCIETY- - BRIGHTON

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

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

2024

We acknowledge the Asian Barometers data for India for 2019 – see References for details.

Our open-source code is on Github.

RESEARCH QUESTION

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

DATASET ASIAN BAROMETERS INDIA 2019, N=5318 ADULTS IN 19 STATES

Going beyond data science to social data science

* 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), $x_1, x_2 \ldots, x_n$
each outcome having probability, $p_1, p_2 \ldots, p_n$, 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)

LIKERT SCALES ARE DISTINCT-ORDINAL

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

How we measure gender norms – use attitude and opinion questions – make a factor to smooth out the variations and gain the central tendency (Kim, et al., 2022)-*WES journal*. *Here, we use one Likert scale at a time as our ordinal input*.

Note: In the 2019 Asian Barometers - India

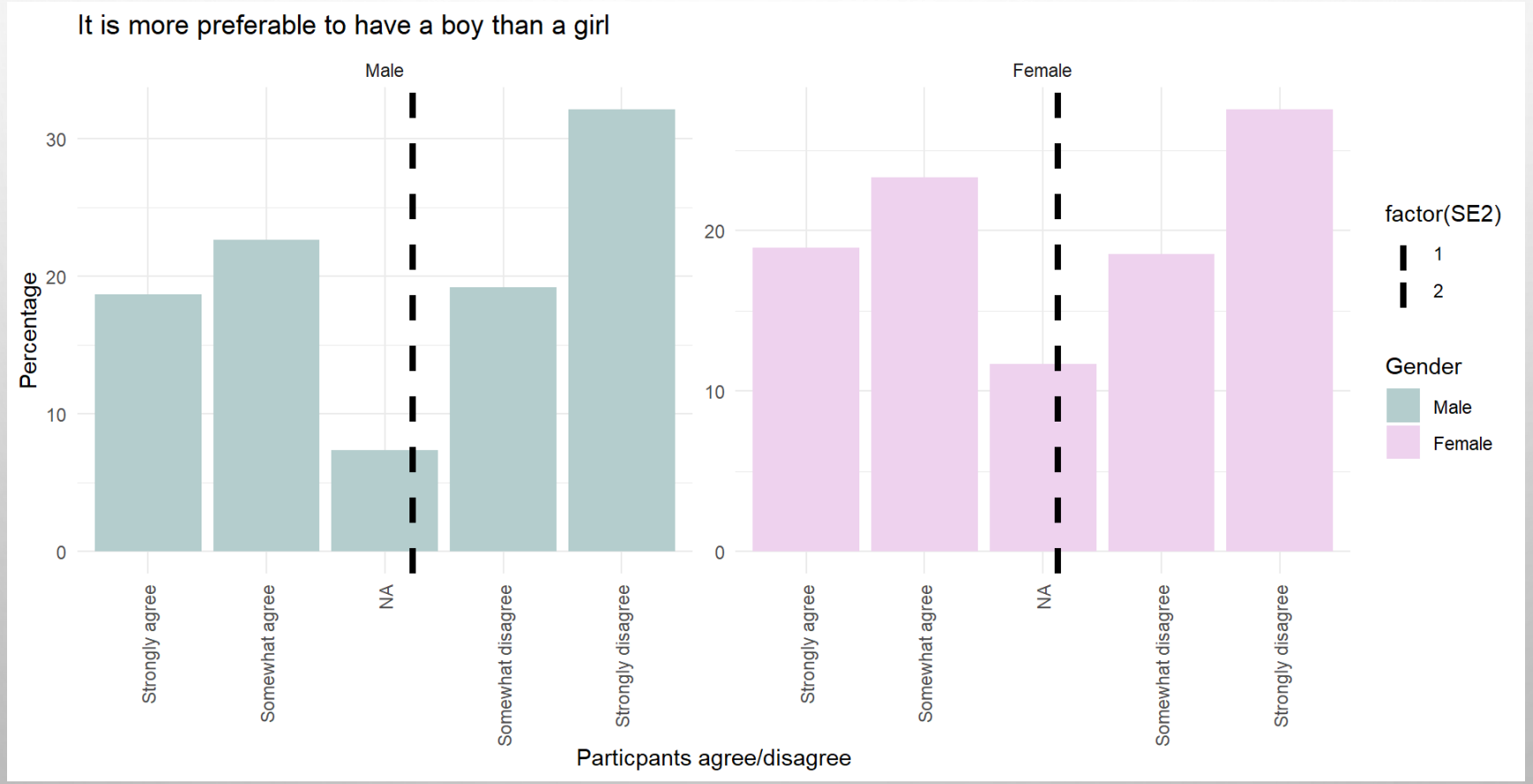
Sexism was embedded in questions ←→ desirability bias of a patriarchal gender norm

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

EDUCATION IS ALSO ORDINAL. CUMULATIVE ORDINAL REFLECTS THE REALITY OF COMBINING PRIMARY, SECONDARY, AND OPTIONAL LATER SCHOOLING.

- *ONTIC NATURE OF THE THING TO WHICH WE REFER*
  - *DISTINCT ORDINAL*
  - *CUMULATIVE ORDINAL*
  - *THE DEVIATION OF THE TWO MEASURES FOR EDUCATION IS EMPIRICALLY DIFFERENT:*
    - *-SINGLE VARIABLE, CUMULATIVE CODING HAS MORE ENTROPY. (LESS INFORMATIVE)*
    - *-MULTIPLE VARIABLES, CUMULATIVE EDUCATION IN GROUP, THERE IS LESS ENTROPY! (MORE INFORMATIVE)*
    - *DO NOT DECIDE PURELY ON DATA-SCIENCE GROUNDS, BUT ON ONTOLOGY GROUNDS.*

THE RESULTS (ENTROPY AND REGRESSION TESTS)

Methods Used	Sample:	Overall Test:	Results:	If multiple vectors:
First entropy measures. Second regressions. Ordered probit. Cumulative coding vs. distinct coding.	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.

SUMMARY AND POINTERS FORWARD

SURVEY OF THE MAIN POINTS TODAY

- ENTROPY IS SLIGHTLY DIFFERENT FOR CUMULATIVE ORDINAL VS DISTINCT ORDINAL VARIABLES
- SUPERVISION IS NEEDED

POINTERS TO HOW TO CARRY OUT SUPERVISION

- STAGE 1 EXAMINE THE ONTIC STATUS OF EACH.
- WHAT IS ITS REFERENCE OBJECT: PROCESS?
- OR AN ENTITY
- AND DOES THAT IMPLY CUMULATION OR DISTINCT-ORDERED STATUS?

- **STAGE 2** DISCRETIZE AS USUAL

- STAGE 3 THEN RE-GROUP THE VARIABLE TO GET THE WHOLE PICTURE

REFERENCES

ENTROPY

Open Source Code –thanks to Ziyang Zhou - for Entropy Calculations – uses one-hot encoding. github.com/WendyOlsen/entropyRSS2024

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

[2] Watts and Crow (2022), The Shannon Entropy of a Histogram, *arXiv* : 2210.02848.

[3] 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

[4] 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