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

https://github.com/WendyOlsen/entropyOrdinalData2024

We acknowledge the Asian Barometers data for India for 2019.

Our open-source code is on Github.

A 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

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

ONE COULD CREATE CUMULATIVE CODINGS ("ENCODING")

Figure 1: LIKERT SCALE, DISTINCT DISCRETIZATION

Option1	Option2	Option3	Option4	Option5
1	0	0	0	0
0	1	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	0	1
0	0	0	0	1
0	0	0	0	1
Etc. n Rows				

It is a sparse matrix.

Edu 2_1	Edu 2_2	Edu 2_3	Edu 2_4	Edu 2-5
1	0	0	0	0
1	1	0	0	0
1	1	0	0	0
1	1	1	0	0
1	1	1	1	0
1	1	1	1	0
1	1	1	1	0
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
Etc.				

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).

* 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 uninformativeness of a 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)

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 [4]

The *H* estimate is a biased estimate

although the ML estimate θ_k^{ML} is not biased.

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

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

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 [4] Hausser-Strimmer (2009).

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

DATA AND METHODS USED HERE

METHOD 1: ENTROPY ESTIMATION

(COMPARE THE DISTINCT CODING WITH THE CUMULATIVE CODING)

METHOD 2: REGRESSION ESTIMATES WITH A VARIETY OF ORDINAL VARIABLES

METHOD 3: SIMULATION AND M.S.E.

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.

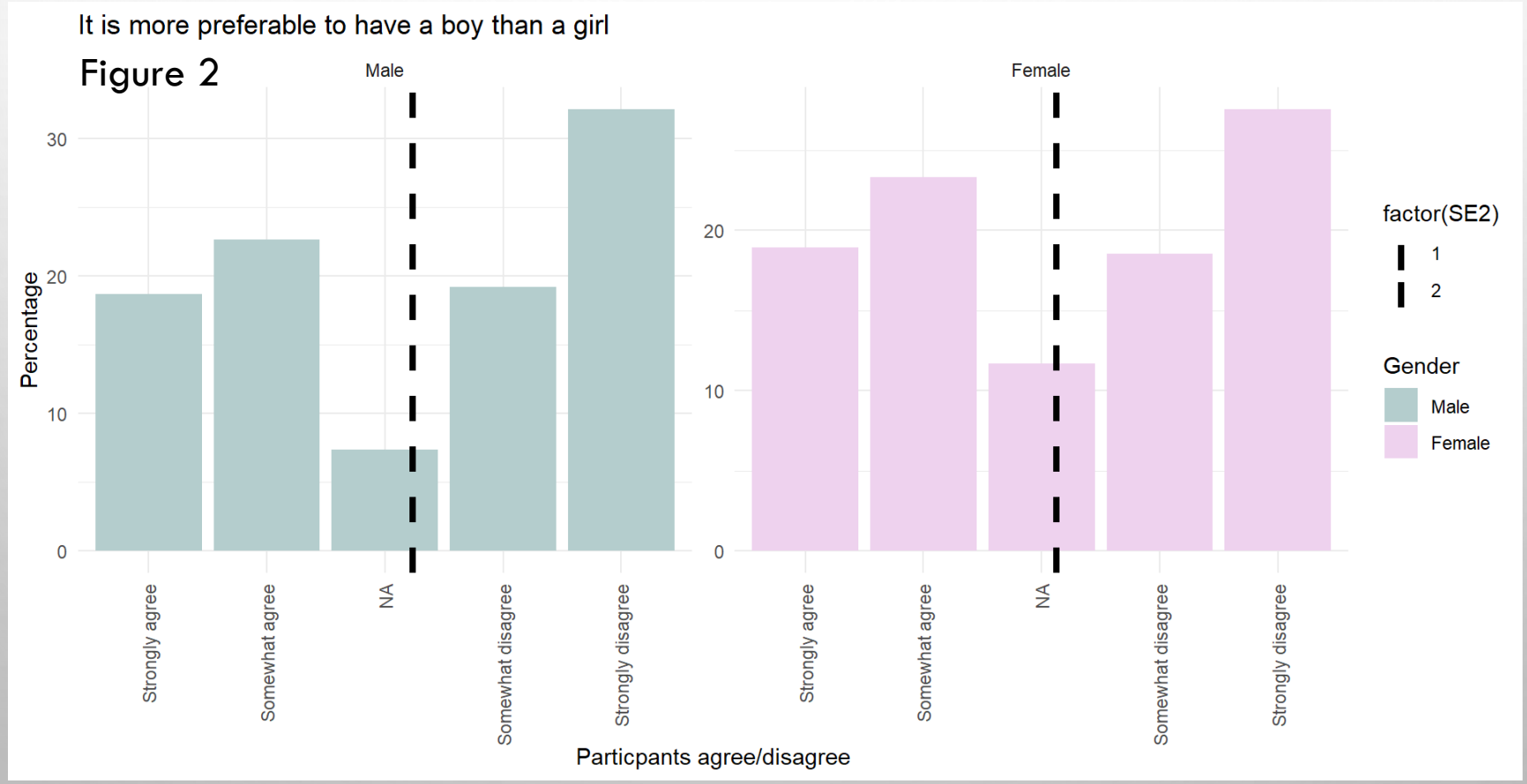
Our example uses the 2019 Asian Barometers - India

LIKERT SCALE – ASIAN BAROMETERS

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

THE ENTROPY MEASURES DEVIATIONS FROM UNIFORM.

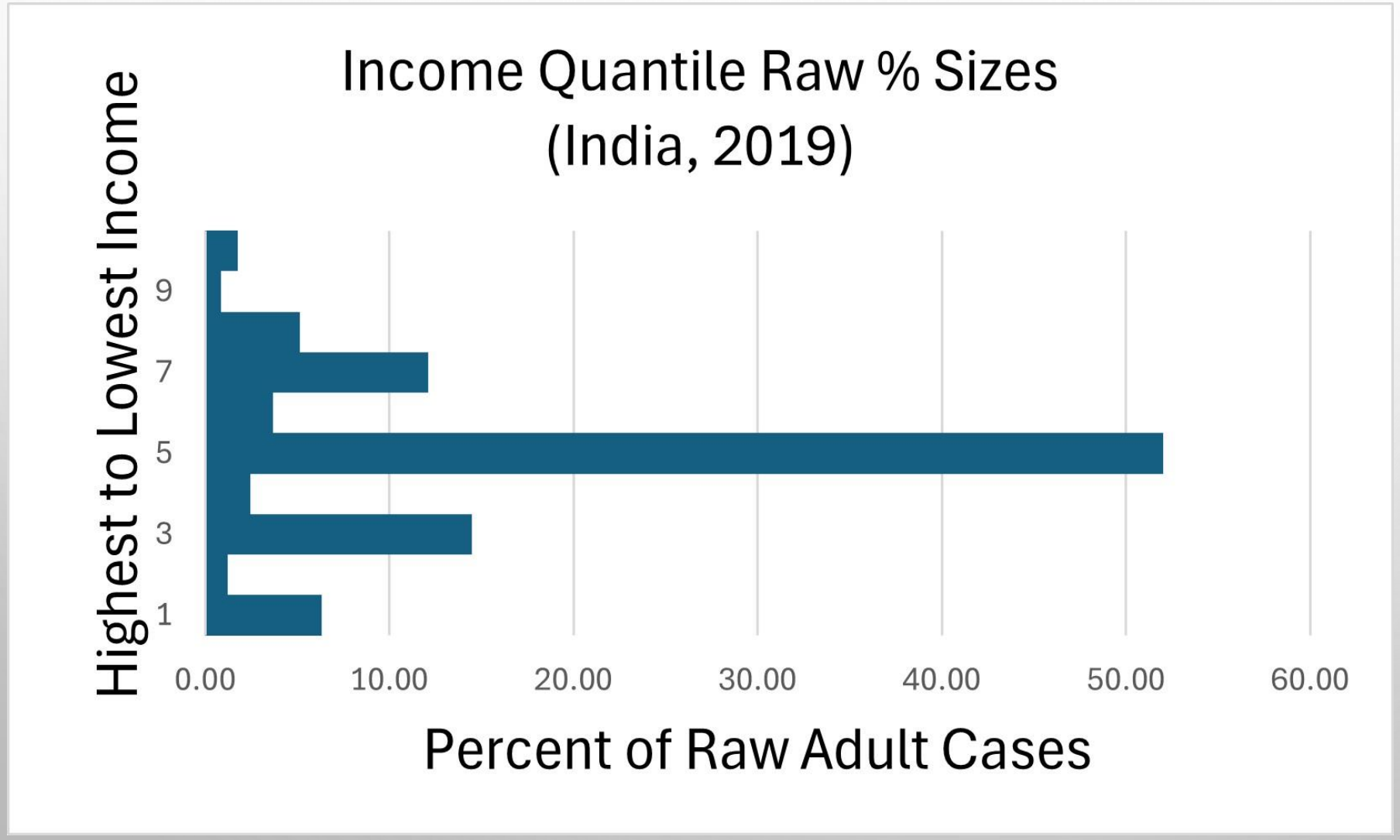
Boy-Preferring



Not Boy-preferring

MANCHESTER 1824
The University of Manchester

SAMPLE IS WIDELY DISTRIBUTED; COVERS 19 STATES



"CUMULATIVE ORDINAL" CODING FOR EDUCATION REPRESENTS THE REALITY OF COMBINING PRIMARY, SECONDARY, AND LATER SCHOOLING.

- *THE ONTIC NATURE OF THE THING TO WHICH WE REFER*
  - *DISTINCT ORDINAL VS. CUMULATIVE ORDINAL HAVE DIFFERENT ENTROPY.*
  - *RELATIVE STATISTICAL ENTROPY MEASURES THIS.*

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

MANCHESTER 1824
The University of Manchester

- & AFTER REGRESSION USE $\textbf{AIC} = -2\log(\hat{L}) + 2K$

WHERE L IS THE LIKELIHOOD AND K IS THE NUM. OF REGRESSION PARAMETERS, EQ. 6

THE RESULTS (ENTROPY TESTS)

Methods Used	Sample:	Overall Test; One Input Vector Giving q Binary Columns	If Multiple Vectors:
First entropy measures.	N=5,318 Adults only 19 states of India	Relative entropy differed by 4-5% between the group of binaries for the distinct vs cumulative coding. Cumulative coding's entropy was lower. This means information was greater.	The dataframe entropy depends in part on mutual entropy. The comparative results were switched around.

SAMPLE OF THE ENTROPY RESULTS

		Cumulative Encoding			Distinct Encoding (Standard Scheme)		
Entropy of single variables		H		RSI	H		RSI
	Education	1.51	More inform ative.	1.51/1.6094= 0.938 normalised H	1.57		1.57/1.6094= 0.975 normalised H

Validated twice.
First in R, comparing Shannon entropy from R entropy package with raw calculations.

Second, in Stata, using H and RSI formulae.

THE RESULTS (SIMULATION TEST ON EDUCATION)

Methods Used	Sample:	H for Discrete Education	H for Cumulative Education	Relative Entropy Tests
Simulation Repeat 1000 samples with replacement from 5,318 Used a Multinomial distribution	N=5,318	The 95% interval for H, the entropy, around the mean 1.52885, was: {1.5384, 1.5171} This range is about 2% of the raw H value in nats. H's MSE was 0.0003.	H mean estimate was 1.51251 95% Interval: {1.5071, 1.5173} This range is about 1% H's MSE was 0.000007	We divide H by the different constants. RSI distinct = 0.950 MSE 0.0001 RSI cumul =.940 MSE 0.000002

THE RESULTS (REGRESSIONS)

Methods Used	Sample:	Overall Test:	Results:
Ordered probit. Cumulative coding vs. distinct coding.	N=5,318	• Income by Edu* • Educ by Age, base 18-25 • Income by Age • Likert 1 'prefer a boy' by income • Likert 1 by educ • Likert 1 by age.	We compare AIC using the Δdf as a ChiSquared.
		Ran 6 sets of 2 regressions	No change in Degrees of Freedom; no difference in fit.
Compare the AIC	N is same, but p rises to q and differs.	AIC test is for the distinct coding of X and for the cumulative coding.	So AIC could be the same. And it is.

*SAMPLE REGRESSION RESULT: INCOME QUANTILE BY AGE DECILE MODEL 3A BY 3B

Number of observations	3A, N=5318, distinct coding	3B, N=5318, cumulative coding
Akaike Information Criterion	17181.78	17181.78
Bayes Information Criterion	17300.20	17300.20
Log likelihood	-8572.89	-8572.89

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 ONTOLOGY;

& CONSIDER REFERENT

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

FUTURE RESEARCH

CHEMICAL SCIENCES

- A SYSTEM HAS WASTES. ENTROPY OF WASTE FLOWS VS. METAL INGOTS…
- CHEMISTRY, ENVIRONMENTAL SCIENCE, PHYSICS, MEDICINE & RADIOGRAPHY USE THE RSI MEASURES FOR A RANGE OF MEASUREMENTS.

ENTROPY IN A SOCIAL-DATA-SCIENCE ANALYSIS SUCH AS 'ASSOCIATION RULES'

- STAGE 1 APPLY PHILOSOPHICAL KNOWLEDGE TO THE INPUTS
  - DISCRETIZE THE ORDINAL INPUT
  - IF CUMUL-RANKED, IT'S 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

[1] Borsboom, Mellenbergh, and van Heerden (2003) The Theoretical Status of Latent Variables, *Psychological Review*.

[2] 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.

SOFTWARE PACKAGE ENTROPY IN R

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

REFERENCES, CONT.

[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

ADDENDA

CLEVER TRICK FOR ONE-HOT ENCODING

(BY ZIYANG ZHOU)

```
# TO CREATE AN EMPTY DATA FRAME FOR TWO INPUT VARIABLES FOR ONE-HOT ENCODING
EDU2_ENCODED <- DATA.FRAME(MATRIX(0, NROW = NROW(DF_EDU), NCOL = LENGTH(UNIQUE(DF_EDU$EDU))))
# ORDER UNIQUE VALUES NUMERICALLY
ORDERED_UNIQUE_VALUES <- SORT(UNIQUE(EDU))
# SET COLUMN NAMES BASED ON UNIQUE VALUES IN 'DF_Q63'
COLNAMES(EDU2_ENCODED) <- PASTE("EDU_", ORDERED_UNIQUE_VALUES, SEP = "")

# TRAVERSE THROUGH THE 'DF_Q63' COLUMN AND FILL IN ONE-HOT ENCODING
FOR (I IN 1:NROW(DF_EDU)) {
VALUE <- DF_EDU$EDU[I]
WHILE(VALUE>0){
EDU2_ENCODED[I, PASTE("EDU_", AS.CHARACTER(VALUE), SEP = "")] <- 1
VALUE = VALUE-1
}
}
```

PROBLEM WITH THE ENTROPY OF CUMULATIVELY ENCODED ORDINAL VARIABLE

- THE NUMBER OF CASES AFFECTS THE RSI.
- COLUMN 1 IS ALL 1'S (A CONSTANT)
- THEREFORE, IT DROPS OUT OF H, AS LN(1)=0.
- THE N IS COUNTED AS AN EMPIRICALLY SPECIFIC SUM, IN ADDITION TO COLUMN 1, OF THE POSITIVE RESPONSES CUMULATED IN VECTORS 2-5. EG
- 15,884 IN THE CASE OF 5,318 RESPONDENTS FOR EDUCATION-CUMULATIVE.
- ITS FORMULA IS $(X_1+2X_2+3X_3+4X_4+5X_5)$

WHERE X_K IS THE COUNT OF THE CELLS FOR EDUC==K.