U.S. Racial Stratification Through the Lens of Utility Burdens

An Indices of Informational Association Analysis of Complex Systems

Kurt Semm [*1]

[1]Barnard College, Columbia University

Senior Economist at The Institute for New Economic Thinking

Abstract:

This article adopts joint informational association measures in piecewise order to analyze disparities in utility costs or the percentage of income a household spends on their water, electricity, and gas bills due to an individual's race, gender, and ethnicity in the US. Using Census Bureau data between 2008 and 2019, the research quantifies association measures for more than 12 million households. It finds that white men's heads of households consistently pay less for utilities (lower utility burdens) than white women, black men and women, non-white Hispanic men and women, and American Indian men and women. White men pay less for utilities if they are in the bottom 90% of the income distribution. This relative utility burden converges if black, non-white Hispanic, and American Indian households are in the top 10% of the income distribution. For this reason, the analysis uses joint piecewise information association measures to quantify the relative differences in the amount of information social scientists can infer about an individual household utility burden based on their social identity or the stratified grouping due to race, ethnicity, and sex. These measures verify that scientists can infer more informational value from households with social identities in racially stratified groups than their white counterparts. While providing a unique case study of informational association measures, this study advances the US reparations debate by analyzing the inherent racial disparities in consuming basic goods and not only on a household's income. Finally, these results illuminate how market-based economies create persistent utility burden inequalities.

*Email: ksemm@barnard.edu

***Keywords*:** *Entropy, Joint Information Association Measures, Racial Stratification, Utility Burdens, Inequality*

Introduction

What impact do race and gender have on what an individual pays for utilities (water, gas, electricity) in the United States? Commonly, in economics, a dominant method is evolutionary game theory to analyze how individuals maximize or satisfy their preferences, or strategies in the case of bounded rationality, within the constraints of cultural norms and values in certain areas. This paper takes a different approach by analyzing individuals within the complex systems they inhabit. To follow a Political Economy framework, researchers must select a method that addresses the complexities inherent in household utility payments. This study uses Shannon entropy as an alternative measure of inequality to evaluate disparities in utility burden income due to geography, race, and sex ([1]) [1]. Furthermore, the analysis examines how income inequality and informational associations impact the utility burden of individual households. The study introduces an alternative method of measuring income inequality and demonstrates systemic discrimination in utility payments based on social identity and relative income. It builds upon traditional methods of assessing inequalities in a capitalist system regarding natural resources and the consumption of basic goods. The argument examines discrimination in utility burdens through the lens of social identity and income decile.

The paper describes the use of Census Bureau data to construct the analysis. The third section, Method, explains the application of Shannon entropy and information association indices, comparing these to traditional measures of income inequality such as GINI and Theil. Section four analyzes the calculated metrics and figures that measure income and utility burden inequality. Section five evaluates the limitations and advantages of using information theory and Shannon entropy to measure inequality. Finally, section six suggests future directions for employing informational associations to extend the theoretical model and the policy implications to overcome racial discrimination in utility bills across the United States.

[1]This research was supported by my work at Barnard College, The New School for Social Research, and the Institute for New Economic Thinking. The authors disclose no financial relationships or other potential conflicts of interest related to this study. Any errors or omissions are the sole responsibility of the authors. The views expressed are those of the authors and do not necessarily reflect the views of the supporting organizations. For further information on research dissemination and disclosures, the authors adhered to the NBER research dissemination disclosure policy.

1 Data

The sample is a survey from the Census Bureau Household Data (IPUMS) between 2008 and 2019. Out of the 51 million total households, it excludes farms, adjusted utility burden, and income for inflation, and it does not include imputed utility burdens (utility burden included in a household's rent). Also, the analysis adopts a method for adjusting the household size for income and utility burdens by adjusting the number of rooms in the household, suggested by dividing the PEW Research Center. Before diving into the data, we must point out our Annual utility burden measures. Then, the research places each individual, or head of household, in different group income deciles and Education attainment and combines race and sex to form an "Identity." After removing households in the top two percent of the income and utility burdens distribution and NAs created in the Identity variable, the total number of observations for the case study is 12,717,383. For further details, please find the attached R code in Appendices A and B.

2 Mutual Information and Its Probabilistic Content

2.1 Setting up the Inequality Measures

Investigating the likelihood that everyone pays the same utility burden regardless of identity or income decile. While it is clear that not everyone would pay the same amount for water per year, based on usage, house size, and season, I argue (with several others cite) that water is a human right. Therefore, if we want to find the social cost of water, we should examine what people are paying for water. However, we want to examine the impact of Identity and household income on what a person pays for water. Similar to entropy, I adopt mutual information, interactions, and decompositions from (Dos Santos et al. 2022).

The luxury of mutual information is it provides broad, nonlinear measures of phenomenological association between the sets of individual degrees of freedom within the joint frequency distribution of observed individuals and their possible values. To offer more clarity, allow me to define and write out the expression mathematically. First, consider two mutually exclusive sets of individual degrees of freedom within a system: the quantity we want to examine Y, and the other characteristics or covariates X. Further, we can define the mutual information amongst Y and X as the difference between entropy in the marginal distribution of Y and the average entropy of distributions of Y conditional on the values of X, or:

2.1.1 Definition

$$I(Y,X) = H(Y) - H(Y|X) = H(Y) - \sum_{x} f_x H(Y|x) \tag{1}$$

Another way we could say mutual information is the average reduction in heterogeneity in individual values of Y, given that we consider what their distribution across all individuals is to consider what their distribution is if they were organized into groupings defined by each value x takes. Using mutual information, we reduce the observer's uncertainty about the system's micro-configurations as they move from observing the marginal distribution fy to the conditional distributions fy|x. We measure how much we learn in informational units about the individual values of y as we learn about their measure of x. Hence, the information measured is the information the two sets of degrees of freedom share.

Next, we should note the MI has a minimum value of zero if the conditional distributions fy|x has the entropy as the population, marginal distribution of Y. Then, MI increases as the conditional distributions fy|x are less heterogeneous than the population, marginal distribution of Y. Also, MI's possible maximum value of H(Y) is when those conditional distributions entropy is zero, or in situations where H(Y) $\leq$ H(X). In the previous case, the heterogeneity of the individual values in Y associates with the heterogeneity of in the values of X, meaning the observer has no uncertainty about the latter in reference to the the former.

2.1.2 A Non-Parametric Measure of Association

Equation 2 illustrates that this method of mutual information provides broad, nonlinear measures of phenomenological association between the sets of individual degrees of freedom.

$$I(Y,X) = H(Y) + H(X) - H(Y,X) \tag{2}$$

Given this definition and equation 1, it is now possible to capture the micro-level multiplicity and entropy in equation one by re-writing it as:

$$I(Y,X) = \frac{1}{N} log \frac{\mathbb{W}_{fY}\mathbb{W}_{fX}}{\mathbb{W}_{fY,x}} \tag{3}$$

Equation four captures the extent of how a functioning system creates the interdependencies between the individual values of Y and X. If, for the system to function the two sets of degrees of freedom are independent, then the number of micro-configurations supporting their joint distribution is the prod-

uct $\mathbb{W}_{fY}$ $\mathbb{W}_{fX}$ of the number of micro-configurations supporting their marginal distributions. Meaning, that for each $\mathbb{W}_{fY}$ sustaining f_y, there are $\mathbb{W}_{fX}$ possible micro-configurations over values of x in the system available. The extent to which the functioning of the system establishes a relationship between the measures of y and x, then we expect is fewer than $\mathbb{W}_{fX}$ possible micro-configurations over values of x compatible with any of the $\mathbb{W}_{fX}$ possible micro-configurations over values of y. Finally, the result, $\mathbb{W}_{fY}$,x is smaller than $\mathbb{W}_{fY}$ $\mathbb{W}_{fX}$. We account for the shortfall presented above because MI quantifies the shortfall as informational units.

3 Information, Interactions, and Decomposition's

3.1 Entropy as Degeneracy

$$I(Y,Y) = \frac{1}{N}\sum_{i} logP(\sigma(y_i), f_y, N, 1) \tag{4}$$

$$I(Y,Y) = H(Y) = -\sum_{y} f_y log f_y = -\frac{1}{N}\sum_{y} log f_y^{n_y} = -\frac{1}{N} log\pi(f_Y, f_Y, s, N) \tag{5}$$

3.2 The Conditional, Joint, and Total Information

$$I(Y,Z|X) = H(Y|X) - H(Y|Z,X) = \sum_{x} f_x I(Y,Z|x) \tag{6}$$

$$I(Y,X,Z) = I(Y,X) + I(Y,Z|X) \tag{7}$$

3.3 The Interaction Information

$$I(Y,X,Z) = I(Y,X) - I(Y,X|Z) = I(Y,X) + I(Y,Z) - I(Y,X,Z) \tag{8}$$

$$I(Y,X,Z) = \frac{1}{N} log\mathbb{P}(f_{Y|X,Z} \| f_{Y|Z}) - log\mathbb{P}(f_{Y|X} \| f_Y) \tag{9}$$

$$\mathcal{I}(Y,\mathcal{V}) \equiv I(Y,\mathcal{V}_1,...,\mathcal{V}_{k-1}) - I(Y,\mathcal{V}_1,...,\mathcal{V}_{k-1}|\mathcal{V}_k) \tag{10}$$

3.4 Decomposition's of Information

$$I(Y) = I(Y,X) + I(Y|X) = I(Y,X) + \sum_{x} f_x I(Y|x) \tag{11}$$

$$I(Y,X) = I(Y,X,Z) + I(Y,X|Z) = I(Y,X,Z) + \sum_{z} f_z I(Y,X|z) \tag{12}$$

$$\mathcal{I}(Y) = \mathcal{I}(Y|\mathcal{V}_1) + \mathcal{I}(Y,\mathcal{V}_1|\mathcal{V}_2) + ... + \mathcal{I}(Y,\mathcal{V}_1,...,\mathcal{V}_{k-1}|\mathcal{V}_k) + \mathcal{I}(Y,\mathcal{V}) \tag{13}$$

4 Informational Association Measures on Utility Burdens: Empirical Evidence

4.1 Observable Economic Characteristics

Table 1: Informational Measures and Utility Burdens

Fig.	**Characteristics**	**Indices**	**Main Results**
1	Education	$\alpha(Y \parallel \chi_{educ})$	College education enhances variability in utility burdens, while individuals with lower educational attainment, particularly high-school dropouts, experience more consistent and predictable utility burdens.
2	Age	$\alpha(Y \parallel \chi_{age})$	The influence of age on utility burdens has become significantly more pronounced for individuals in their TBD.
3	Education, Age	$\alpha(Y \parallel \chi_{educ} \parallel \chi_{age})$	Higher education levels amplify diversity in returns from experience, with these returns peaking later in life for the more educated. Variations in returns across different educational groups continue to increase over time.
4	Gender	$I(X_{gender} \mid X_{educ})$	Across all educational achievement groups, being female strongly predicts utility burdens, while men experience greater diversity in utility burden differentiation.
5	Race/ethnicity, Gender	$I(X_{ident} \mid X_{educ})$, $I(X_{ident} \mid X_{age})$	Social identity consistently plays a significant role in predicting utility burdens for Black, American Indian, and Hispanic households. In contrast, White households show greater variation in utility burdens based on factors unrelated to social identity.
6	Race/ethnicity, Gender	$I(X_{ident} \mid X_{educ})$, $I(X_{ident} \mid X_{age})$	This observation aligns with observation 5 but provides additional detail on how gender and race contribute incrementally to differentiation in utility burdens.
7	Compound social identity	$I(X_{ident} \mid X_{educ})$, $I(X_{ident} \mid X_{age})$	Social identity significantly influences utility burdens for women from historically marginalized groups, whereas men from these groups have less variation compared to white men. Over time, the impact of social identity on utility burdens for white women has diminished.
8	Compound social identity and education	$m(Y; \chi_{id}; \chi_{ed})$	Informational synergies emerge between education, identity, and utility burdens for Black and white women with low levels of education, as well as for white individuals with high levels of education.
9	Compound social identity, education, relative utility burden	$m(Y; \chi_{id}; \chi_{ed})$, $S_{ed,id}$	As education and utility burdens increase, educational attainment becomes a weaker predictor of utility burdens for women. In contrast, white men exhibit the opposite trend, with educational attainment becoming a stronger predictor.

5 Discussion

The research reveals several significant and nuanced insights derived from the information measures, shedding light on how a person's social identity and income shape their relationship with water in a capitalist system. The figures demonstrate that utility burdens for white men show minimal informational variation, indicating limited predictability about what they pay for water. This suggests greater flexibility for white men in terms of water, electricity, and gas pricing. In stark contrast, utility burdens for American Indian, Black, and Hispanic women exhibit substantially higher informational variation, reflecting a stronger correlation between their social identities, income levels, and water costs. This disparity underscores the pronounced role of social identity and economic stratification in shaping utility burdens, offering deeper analytical perspectives on systemic inequities.

Conclusion

6 Appendix

6.1 A

Variable	**Code (Value)**	**Meaning**	**Reasoning**
FARM	1	Not a farm	We want to see individuals first, people who do not farm.
RELATE	1	Head of Household	Respondent is the head of the household.
SINK	2	Sink	Control for material conditions.
PLUMPING	20	With Complete Plumbing	Control for material conditions.
SHOWER	4	Has Shower	Control for material conditions.
HOTWATER	4	Hot and cold piped water	Control for material conditions.
FOODSTMP	1	Does not receive food stamps	Control for material conditions.
Inflation Measure		BLS.gov	(Actual Value/Index value)*100. Real not nominal Income, costs, wages.

Household Size	Pew Research Methodology	Adjusted income = Income / (number of people in HH)^n, where $n = 0.5$. Have to adjust data for number of people in the household.

6.2 B R Code for Analysis

Listing 1: R Code for Utility Burden Analysis

```
#### Packages ####
library(ipumsr)
library(dplyr, warn.conflicts = FALSE)
library(survey)
library(srvyr)
library(plyr)
library(dineq)
library(data.table)
library(tidyverse)
library(tibble)
library(reshape)
library(reshape2)
library(readr)
library(readxl)
library(ggplot2)
library(gridExtra)
library(scales)
library(knitr)
library(kableExtra)
library(frequency)
library(purrr)
library(stargazer)
library(sf)
library(stats)
library(spatstat.geom)
library(hrbrthemes)
```

```
library(viridis)
library(formattable)

setwd("/Users/kurtsemm/Documents/Information_Theory/Utility_Burdens")
# Note that you can pass in the loaded DDI into the `read_ipums_micro()`
cps_ddi <- read_ipums_ddi("usa_00024.xml")
cps_data <- read_ipums_micro(cps_ddi, data_file = "usa_00024.dat.gz", verbose =
    FALSE)



## Clean data so water cost and income are greater than one.
## FARM equal to 1 means the Household is NOT a farm and we assume household  ##
     has complete plumbing. We leave household type alone for right now,
## also shower does not exist for the years we are looking at.
## Using constant PUMA instead of non-constant PUMA.

## Controlling for Material Conditions ##

cps_data%>% select(-SERIAL, -PERNUM)
cps_data <- cps_data %>%
  #select_if(is.labelled) %>%
  filter(COSTWATR >= 1,
         HHINCOME >= 1,
         FARM == 1,
         PLUMBING == 20,
         SINK == 2,
         SHOWER == 4,
         HOTWATER ==4,
         FOODSTMP == 1,
         RELATE == 1,
         !is.na(COUNTYFIP)) #COUNTYFIP goes back to 2005
#filter(SINK == 2) # We lose years 2001-2008, not included in survey before than
    .

```

```
library(readxl)
CPI <- read_excel("CPI.xlsx")
cpi <- CPI %>%
  select(YEAR, Annual)

#Combine cpi with our dataset & adjust for inflation
cps_data %>% group_by(YEAR)
cps_usa <- cps_data %>% left_join(cpi)

cps_usa <- cps_usa %>%
  mutate(HHINCOME = (HHINCOME/Annual)*100,
         COSTWATR = (COSTWATR/Annual)*100,
         COSTELEC = (COSTELEC/Annual)*100,
         COSTGAS =  (COSTGAS/Annual)*100)

# Next, we have to adjust income and water costs for number of people in
    household, number of rooms, number of bedrooms,
# Using https://www.pewresearch.org/social-trends/2011/10/03/appendix-b-
    adjusting-household-income-for-household-size/ as a reference.

n = 0.5

#Adjust for number of individuals in household.

cps_usa <- cps_usa %>%
  mutate(HHINCOME = (HHINCOME/(PERNUM^n)),
         COSTWATR = (COSTWATR/(PERNUM^n)),
         COSTELEC = (COSTELEC/(PERNUM^n)),
         COSTGAS =  (COSTGAS/(PERNUM^n)))

cps_usa<- cps_usa %>%
  mutate(
    util_burden = ((COSTWATR+COSTELEC+COSTGAS)/HHINCOME),
    elec_burden = (COSTELEC/HHINCOME),
    gas_burden = (COSTGAS/HHINCOME),
```

```
92      water_burden = (COSTWATR/HHINCOME))
93  #### Create upper bound for histogram #####
94
95  #Utility Burden
96  util_pctile <- cps_usa %>%
97    group_by(YEAR) %>%
98    do(util_98 = reldist::wtd.quantile(.$util_burden, q = 0.98, weight = .$HHWT))
          %>%
99    mutate(util_98 = unlist(util_98)) %>%
100   as.data.frame()
101
102 #Water Burden
103 water_pctile <- cps_usa %>%
104   group_by(YEAR) %>%
105   do(water_98 = reldist::wtd.quantile(.$water_burden, q = 0.98, weight = .$HHWT)
          ) %>%
106   mutate(water_98 = unlist(water_98)) %>%
107   as.data.frame()
108
109 #Electricity Burden
110 elec_pctile <- cps_usa %>%
111   group_by(YEAR) %>%
112   do(elec_98 = reldist::wtd.quantile(.$elec_burden, q = 0.98, weight = .$HHWT))
          %>%
113   mutate(elec_98 = unlist(elec_98)) %>%
114   as.data.frame()
115
116 #Gas Burden
117 gas_pctile <- cps_usa %>%
118   group_by(YEAR) %>%
119   do(gas_98 = reldist::wtd.quantile(.$gas_burden, q = 0.98, weight = .$HHWT))
          %>%
120   mutate(gas_98 = unlist(gas_98)) %>%
121   as.data.frame()
122
```

```
123
124
125 ## Join columns pctile's to cps_usa and water_pctile.KS-8/14.
126 cps_usa <- cps_usa %>%
127   left_join(util_pctile) %>%
128   left_join(water_pctile) %>%
129   left_join(elec_pctile) %>%
130   left_join(gas_pctile)
131
132 util_list<-c("util_burden", "elec_burden", "gas_burden", "water_burden")
133
134 #### Head of Household social identity ####
135 # Age Level
136 cps_usa <- cps_usa %>%
137   mutate(AGE = as.numeric(AGE)) %>%
138   mutate(AGE_GRP = case_when(
139     AGE >= 1 &  AGE <=  9 ~ "09 and Less",
140     AGE >  9 &  AGE <= 19 ~ "19 and Less",
141     AGE >  19 & AGE <= 29 ~ "29 and Less",
142     AGE >  29 & AGE <= 39 ~ "39 and Less",
143     AGE >  39 & AGE <= 49 ~ "49 and Less",
144     AGE >  49 & AGE <= 59 ~ "59 and Less",
145     AGE >  59 & AGE <= 69 ~ "69 and Less",
146     AGE >  69 & AGE <= 79 ~ "79 and Less",
147     AGE >  79 & AGE <= 89 ~ "89 and Less",
148     AGE >= 89 ~ "89 & More",
149     TRUE ~ NA_character_))
150
151 # Education Level
152 cps_usa <- cps_usa %>%
153   mutate(EDUC = as_factor(EDUC),
154          year = as.factor(YEAR)) %>%
155   mutate(EDUC4 = factor(case_when(
156     EDUC %in% c("N/A or no schooling", "Nursery school to grade 4",
157                 "Grade 5, 6, 7, or 8", "Grade 9", "Grade 10", "Grade 11")
```

```
158     ~ "A: Less that HS",
159     EDUC == "Grade 12" ~ "B: HS Finished",
160     EDUC %in% c("1 year of college", "2 years of college", "3 years of college")
        ~ "C: Some College",
161     EDUC %in% c("4 years of college", "5+ of college") ~ "D: College Grad",
162     TRUE ~ NA_character_)))
163
164 # To create our Identity variable we need to create a new column using mutate()
165 # Identity by Race
166 cps_usa <- cps_usa %>%
167   mutate(RACE = as_factor(RACE),
168          HISPAN = as_factor(lbl_na_if(HISPAN, ~.val == 9)),
169          SEX = as_factor(SEX)) %>%
170   mutate(RACE_ETHD = case_when(
171     HISPAN %in% c("Mexican", "Puerto Rican", "Cuban", "Other") ~ "hispanic",
172     HISPAN == "Not Hispanic" & RACE == "White" ~ "white",
173     HISPAN == "Not Hispanic" & RACE == "Black/African American" ~ "black",
174     HISPAN == "Not Hispanic" & RACE == "Other Asian or Pacific Islander" ~ "
        asian",
175     HISPAN == "Not Hispanic" & RACE == "American Indian or Alaska Native" ~ "
        american indian",
176     HISPAN == "Not Hispanic" & RACE %in% c("Other race, nec", "non-Hispanic") ~
        "other")) %>%
177   mutate(RACE_ETH = factor(
178     case_when(RACE_ETHD %in% c("other") ~ "other",
179               RACE_ETHD %in% c("hispanic", "white", "black", "american indian",
                  "asian") ~ RACE_ETHD)),
180     RACE_ETHD = factor(RACE_ETHD)) %>%
181   mutate(IDENTITY = factor(case_when(RACE_ETH == "white" & SEX == "Male" ~ "
        White Men",
182                                      RACE_ETH == "white" & SEX == "Female" ~ "
                                           White Women",
183                                      RACE_ETH == "black" & SEX == "Male" ~ "
                                           Black Men",
```

```
                                    RACE_ETH == "black" & SEX == "Female" ~ "
                                        Black Women",
                                    RACE_ETH == "hispanic" & SEX == "Male" ~ "
                                        Hispanic Men",
                                    RACE_ETH == "hispanic" & SEX == "Female" ~
                                        "Hispanic Women",
                                    RACE_ETH == "american indian" & SEX ==  "
                                        Male" ~ "American Indian Men",
                                    RACE_ETH == "american indian" & SEX == "
                                        Female" ~ "American Indian Women",
                                    RACE_ETH == "asian" & SEX == "Male" ~ "
                                        Asian Men",
                                    RACE_ETH == "asian" & SEX == "Female" ~ "
                                        Asian Women",
                                    RACE_ETH == "other" & SEX == "Male" ~ "
                                        Other Men",
                                    RACE_ETH == "other" & SEX == "Female" ~ "
                                        Other Women",
                                    TRUE ~ NA_character_)))
summary(cps_usa$YEAR)
#2008-2019


sample<-cps_usa %>%
  mutate(util_pct = case_when(util_burden >= util_98 ~ 1, T ~ 0),
         water_pct = case_when(water_burden >= water_98 ~ 1, T ~ 0),
         elec_pct = case_when(elec_burden >= elec_98 ~ 1, T ~ 0),
         gas_pct = case_when(gas_burden >=gas_98 ~ 1, T ~ 0)) %>%
  filter(util_pct == 0,
         water_pct == 0,
         elec_pct == 0,
         gas_pct == 0)

summary(sample$util_burden)
saveRDS(sample, "sample_data.rds")
```

```
211
212 for (i in 2008:2019){
213   sample %>%
214     filter(year==i) %>%
215     ggplot(., aes(x=util_burden, fill=SEX)) +
216     geom_histogram(color="#e9ecef", alpha=0.6, position = 'identity', binwidth =
          1) +
217     scale_fill_manual(values=c("darkgreen", "blue")) +
218     theme_ipsum() +
219     labs(fill="")
220   ggsave(histogram, file = paste(i,"util_burden.pdf"))
221 }
222
223 majority_util<-sample %>%
224   filter(YEAR==2019)
225
226 ggplot(majority_util, aes(x = util_burden, fill = SEX)) +
227   geom_density(alpha = 0.6) +
228   scale_fill_manual(values = c("darkgreen", "blue")) +
229   theme_minimal() +
230   labs(x = "Util␣Burden", y = "Density", title = "Distribution␣of␣Util␣Burden")
231
232 #### Entropy Measure Functions ####
233 ent <- function(x){ ### Equation 1
234   -sum(ifelse(x > 0, x * log(x), 0))
235 }
236
237 H_X <- function(D){
238   ent(prop.table(apply(D, MARGIN = 1, FUN = sum)))
239 } ##Equation 3
240
241 H_z <- function(D){
242   X_cond_Z = apply(D, MARGIN = c(1,3), FUN = sum)
243   f_X_cond_Z = prop.table(X_cond_Z, margin=2)
244   apply(f_X_cond_Z, MARGIN = 2, FUN = ent)
```

```
}

H_y <- function(D){ ### Equation 2
  X_cond_Y = apply(D, MARGIN = c(1,2), FUN = sum)
  f_X_cond_Y = prop.table(X_cond_Y, margin=2)
  apply(f_X_cond_Y, MARGIN = 2, FUN = ent)
}

H_z_y <- function(D){ ### Equation 5
  X_cond_Z_Y = prop.table(D, margin=c(2,3))
  apply(X_cond_Z_Y, MARGIN = c(2,3), FUN = ent)
}

histogrammer2 <- function(var, top, weight) {
  breaks_flex=seq.int(from=0, to=top, length.out=34)
  hist=spatstat.geom::whist(var, breaks=breaks_flex, weights=weight)
  return(data.frame(freq=hist,freq=breaks_flex[-1]-diff(breaks_flex)/2))
}
#### Create Histogrammer ####

freqr <- function(var, top, weight, k = 34) { # takes the variable to be binned,
    the upper limit, and the survey weight
  breaks_flex = seq.int(from=0, to=top, length.out = k)
  as.numeric(spatstat.geom::whist(var, breaks = breaks_flex, weights=weight)) #
    returns the weighted frequencies per bin
}

midr <- function(var, top, weight, k = 34) { # takes the variable to be binned,
   the upper limit, and the survey weight
  breaks_flex = seq.int(from=0, to=top, length.out = k)
  breaks_flex[-1]-diff(breaks_flex)/2 # returns the bin mid-points
}

Nr <- function(var, top, k = 34) { # takes the variable to be binned, the upper
   limit
```

```
  breaks_flex = seq.int(from=0, to=top, length.out = k)
  hist(var, breaks = breaks_flex, plot = FALSE)$counts # returns the sample size
       per bin
}

freqr_result <- freqr(var = sample$util_burden, top = sample$util_98,
                      weight = sample$HHWT)

set.seed(0308)
util_xyz_histogram <- sample %>%
  filter(!is.na(EDUC4),
         !is.na(IDENTITY)) %>%
  dplyr::group_by(year, IDENTITY, EDUC4, .drop=FALSE) %>%
  dplyr::reframe(
    freq = freqr(util_burden, util_98, HHWT),
    mid_burden = midr(util_burden, util_98, HHWT),
    N = Nr(util_burden, util_98)
  ) %>%
  mutate(rel_freq = freq/sum(freq)) %>%
  filter(freq > 0) %>%
  mutate(Sex=factor(case_when(
    IDENTITY %in% c("Black␣Men",
                    "Hispanic␣Men",
                    "Asian␣Men",
                    "American␣Indian␣Men",
                    "Other␣Men",
                    "White␣Men")~"Men",
    IDENTITY %in% c("Black␣Women",
                    "Hispanic␣Women",
                    "Asian␣Women",
                    "American␣Indian␣Women",
                    "Other␣Women",
                    "White␣Women") ~ "Women")),
    Ethnicity=factor(case_when(
      IDENTITY %in% c("Black␣Men","Black␣Women") ~ "Black",
```

```
310       IDENTITY %in% c("Hispanic␣Men","Hispanic␣Women") ~ "Hispanic",
311       IDENTITY %in% c("Asian␣Men","Asian␣Women") ~ "Asian",
312       IDENTITY %in% c("American␣Indian␣Men","American␣Indian␣Women") ~
313         "American␣Indian",
314       IDENTITY %in% c("Other␣Men","Other␣Women") ~ "Other",
315       IDENTITY %in% c("White␣Men","White␣Women") ~ "White")))
316
317 #### Create Information Measures ####
318 A_X_z <- function(d, x, y, z){ # Informational association of Z and X, Equation
      6
319   form <- deparse(substitute(x~y~z))
320   D = reshape2::acast(d, formula = form, value.var = "freq",
321                       fun.aggregate = length)
322   ret = 1 - H_z(D)/H_X(D)
323   ret2 = data.frame(A_X_z = ret, row.names = NULL)
324   ret2$temp = as.character(names(ret))
325   names(ret2)[2] <- deparse(substitute(z))
326   return(ret2)
327 }
328
329 A_X_y <- function(d, x, y, z){  # Informational association of Y and X Equation
      6
330   form <- deparse(substitute(x~y~z))
331   D = reshape2::acast(d, formula = form, value.var = "freq",
332                       fun.aggregate = length)
333   ret = 1 - H_y(D)/H_X(D)
334   ret2 = data.frame(A_X_y = ret, row.names = NULL)
335   ret2$temp = as.character(names(ret))
336   names(ret2)[2] <- deparse(substitute(y))
337   return(ret2)
338 }
339
340
341 I_z_y <-  function(d, x, y, z){# Incremental info assoc of Z, X when Y is known
342   form <- deparse(substitute(x~y~z))
```

```
343   D = reshape2::acast(d, formula = form, value.var = "freq",
344                       fun.aggregate = length)
345   ret = (H_y(D)-H_z_y(D))/H_X(D)
346   ret2 = data.frame(ret, row.names = NULL)
347   ret2$temp = as.character(rownames(ret))
348   ret3 = gather(ret2, key = temp2, value = "I_z_y", -temp)
349   names(ret3)[1] <- deparse(substitute(y));
350   names(ret3)[2] <- deparse(substitute(z))
351   return(ret3)
352 } ### Equation
353
354 I_y_z <-  function(d, x, y, z){# Incremental info assoc of Y, X when Z is known
355   form <- deparse(substitute(x~y~z))
356   D = reshape2::acast(d, formula = form, value.var = "freq",
357                       fun.aggregate = length)
358   ret = t((H_z(D)-t(H_z_y(D)))/H_X(D))
359   ret2 = data.frame(ret, row.names = NULL)
360   ret2$temp = as.character(rownames(ret))
361   ret3 = gather(ret2, key = temp2, value = "I_y_z", -temp)
362   names(ret3)[1] <- deparse(substitute(y));
363   names(ret3)[2] <- deparse(substitute(z))
364   return(ret3)
365 }
366
367 M_x_y_z <-  function(d, x, y, z){ # Interaction information of Y, Z, for X
368   form <- deparse(substitute(x~y~z))
369   D = reshape2::acast(d, formula = form, value.var = "freq",
370                       fun.aggregate = length)
371   ret = ret = (H_X(D) - H_y(D) - t(H_z(D)-t(H_z_y(D))) ) / H_X(D)
372   ret2 = data.frame(ret, row.names = NULL)
373   ret2$temp = as.character(rownames(ret))
374   ret3 = gather(ret2, key = temp2, value = "M_x_y_z", -temp)
375   names(ret3)[1] <- deparse(substitute(y));
376   names(ret3)[2] <- deparse(substitute(z))
377   return(ret3)
```

```
}
```

References

[1] Paulo L. dos Santos and Noé Wiener. "Indices of Informational Association and Analysis of Complex Socio-Economic Systems". In: *Entropy* 21.4 (2019). ISSN: 1099-4300. DOI: 10.3390/e21040367. URL: https://www.mdpi.com/1099-4300/21/4/367.