Socio-Economic Status and Names

Relationships in 1880 Male Census Data

Rebecca Vick, University of Minnesota

Minnesota Population Center

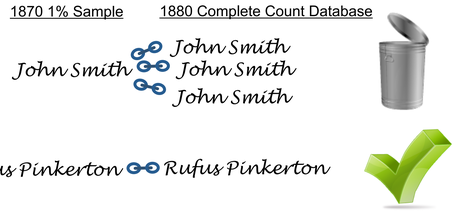
University of Minnesota

Study Question

This research asks if there is a relationship between name commonness and socio-economic status. This question is important because it has implications for linked samples.

When linking historical records across two or more samples, the best way to avoid false links is to exclude records that match to more than one record. This usually occurs with common name like John Smith. Unique names, like Percy Pinkerton, typically have strong similarity to one and only one record, or to no records at all (in the case of mortality, under enumeration, or enumeration error).

Names are a personal identification method, something so basic that perhaps their relationship with other demographic characteristics are assumed to be random or benign. But is that assumption true? It behooves record linkers who use name data to know whether or not this is possible. I use 1880 U.S. Census data and the Duncan socio-economic index measure to examine this question.



Record Linkage

Record linkage : The process of connecting records for the same individual across two or more data sources.

Linked files are uniquely rich in information about individual life change such as migration, occupational mobility and household composition. Historical linked datasets contain new information that could solidify, enlighten or expand our knowledge of social science and demographic history.

Name Processing

- Non-alphabetic characters, titles and non-pertinent characters removed
- Dictionary of standardized names applied to first name data to correct for abbreviations and nicknames.

Table 1. The Twenty most-common First -Last Name Combinations Males Age 30-50, 1880 10% IPUMS U.S. Census Sample

rank	first and last names	frequency	Percent	Cumulative frequency	Cumulative percent
1	john smith	838	.0012369	838	.0012369
2	william smith	747	.0011026	1585	.0023396
3	john brown	497	.0007336	2082	.0030732
4	william johnson	483	.0007129	2565	.0037861
5	james smith	482	.0007115	3047	.0044976
6	john williams	477	.0007041	3524	.0052016
7	john johnson	476	.0007026	4000	.0059043
8	john miller	449	.0006628	4449	.006567
9	george smith	441	.0006509	4890	.0072179
10	william jones	407	.0006008	5297	.0078187
11	john jones	396	.0005845	5693	.0084032
12	william brown	374	.000552	6067	.0089553
13	henry smith	354	.0005225	6421	.0094778
14	john davis	351	.0005181	6772	.0099959
15	charles smith	322	.0004753	7094	.0104712
16	james brown	307	.0004532	7401	.0109243
17	william davis	288	.0004251	7689	.0113494
18	john wilson	287	.0004236	7976	.0117731
19	james johnson	265	.0003912	8241	.0121642
20	george brown	263	.0003882	8504	.0125524

Duncan Socio-Economic Index Score

The Duncan Socio-economic Index score or SEI is the occupational standing measure used to measure socio-economic status for this study.

- Composite measure based upon income, education and prestige associated with 1950 occupations..
- Available in all IPUMS census samples , 1850-2012 for all persons with an occupational response,(i.e. IPUMS OCC1950 code between 000-970)
- See Duncan's 1961 paper "A Socioeconomic Index for All Occupations".

Occupations and their Corresponding SEI Scores
Laborer: 6
Farmer: 14
Carpenter: 19
Telegraph Messenger: 22
Bill Collector: 39
Postmaster: 60
Physician: 92

Analysis

- Data for analysis : 1880 Males aged 30-50 who had an occupation (i.e. SEI>0)
- Name categories range from most (1) to least common based on name frequency
- Initially, eleven categories created by looking for natural breaks, then by making divisions in smaller and smaller frequency increments until reaching names that only occurred once (the data are heavily skewed to the left (nearly 70% of all names occurring 4 times or less, 51% occurring only once)
- Eleven categories were then collapsed into four for easier analysis and interpretation. Groups one and two contain the most common names, and groups three and four the least common

Table 2. Name Commonness in Four Categories: 1880 Males Aged 20-50 with Occupational Responses

Category	Name Occurrences	Mean SEI	Frequency	Percent	Cumulative percent
1 – Most common	>=285	19.41367	7,767	1.18	1.18
2	50-284	19.96422	47,730	7.25	8.43
3	5-49	21.49199	149,566	22.71	31.14
4 – Least common	1-4	21.82056	453,478	68.86	100.0
Total			658,541	100.0	

Mean SEI grows slightly from category one to category four indicating lower socio-economic status for those with common names.

Results

I constructed a regression model predicting SEI with name commonness categories. All four categories have a statistically significant difference in SEI scores from that of category four - least common names.

Table 3. Regression Predicting SEI Using Name Commonness Categories

SEI		Coef.	Std. Err.
Most Common	1	-2.406885*	.224898
	2	-1.856343*	.0945714
	3	-.3285678*	.0586009
Least Common (Constant)	4	21.82056	.0291841

*statistically significant at the p=.05 level.

Conclusion Summary

The results show a statistically significant difference in socio-economic status between those with common and uncommon first and last name combinations. Those with common names tend to have slightly lower status (2.4 points) than those with more unique names. But, the effects would likely be negligible on a final linked sample.

Conclusion

Although statistically significant, the SEI difference between common and uncommon names is very small (2.4 points). It is difficult to determine how this difference might affect a linked sample, but we can use this information to form an opinion. The proportion of working males with very common names is very small. The names deemed most common in this paper comprise a little over 1% of the overall study population. And the majority (69%) have uncommon names. the SEI difference between least common and most common names is 2.4 points- the difference between a carpenter and a telegraph messenger, which I argue is rather meaningless. In my estimation, the threat to the quality of linked samples that stems from throwing out a disproportionate number of common names to avoid false links is likely negligible.