# APPENDIX:

# Modeling the Relationship Between Surname and Hispanicity in Different Demographic Contexts

# Bernard Grofman School of Social Sciences, University of California Irvine

# [BGTravel@uci.edu](mailto:BGTravel@uci.edu)

# April 12, 2015

## **Five propositions about surname matching**

Let

hi = the number of Hispanics among those with the ith name

nonhi = the number of non-Hispanics among those with the ith name

pi = the number of people with the ith name

N = total number of distinct names in the data set

H = total number of Hispanics in the data set

nonH = total number of non-Hispanics in the data set

prob(Hispanic|name i) = The proportion of individuals with a given name who self-identify as Hispanic/of Spanish heritage

prob(name i |Hispanic) = the proportion of Hispanics who have a given name (in the national sample of non-Hispanics)

prob(non-Hispanic|name i) = The proportion of individuals with a given name who self-identify as non-Hispanic

prob(name i |non-Hispanic) = the proportion of non-Hispanics who have a given name (in the national sample of non-Hispanics)

prob(Hispanic) = 1 – prob (non-Hispanic) = the proportion of Hispanic /those of Spanish heritage in the sample

 the cumulative mean proportion Hispanic among the names arrayed from most to least Hispanic, for the range from the first to the nth name.



 the cumulative mean proportion non-Hispanic among the names arrayed from most to least Hispanic, for the range from the first to the nth name.



Proposition 1: If, for each surname, in any sample, the surname’s share of total Hispanic population, prob (name i |Hispanic), and its share of total non-Hispanic population, prob (name i |non-Hispanic), is treated as a random sample from the corresponding national name distributions within each of the two groups, then the proportion of individuals with a given surname who self-identify as Hispanic, prob(Hispanic|name i), is not a constant, but is a function of the Hispanic proportion (and thus also of the non-Hispanic proportion) of the sample we are looking at. In particular,

prob(Hispanic|name i) =

prob(name i |Hispanic) \* prob(Hispanic)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

prob(name i |Hispanic) \* prob(Hispanic) + prob(name i |non-Hispanic) \* prob(non-Hispanic)

Proof: The result is simply a restatement of Bayes Theorem. The basis of Bayes Theorem is the identity

prob(Hispanic|name i) \* prob(name i) = prob(name i |Hispanic) \* prob(Hispanic)

From this identity we derive the equation

prob(Hispanic|name i) = (prob(name i |Hispanic) \* prob(Hispanic)) / prob(name i)

Now, we can use a further identity, namely

p(A) = p(A|B)p(B) + p(A|not B)p(not B)

to show, after some straightforward algebra, that

prob(Hispanic|name i) =

prob(name i |Hispanic) \* prob(Hispanic)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (1)

prob(name i |Hispanic) \* prob(Hispanic) + prob(name i |non-Hispanic) \* prob(non-Hispanic)

But, from Eq. (1), we can see that prob(Hispanic|name i) depends both on the underlying conditional probabilities, prob (name i |Hispanic) and prob (name i |non-Hispanic), which under the given assumptions, for a large enough sample, we may take to be essentially fixed, while the further parameter,

prob(Hispanic) = 1- prob(non-Hispanic),

is context dependent. Thus, prob(Hispanic|name i) varies with the Hispanic proportion in the sample. *q.e.d.*

Proposition 2: If we array names from most Hispanic to least Hispanic and we treat the first s names as 100% Hispanic and the remaining names (from the (s+1)th to the Nth) as non-Hispanic, then the value of s such that the names classified as Hispanic yield the true Hispanic population is given by s such that



Proof: If we array names from most Hispanic to least Hispanic and we treat the first s names as 100% Hispanic and the remaining names (from the (s+1)th to the Nth) as non-Hispanic, then we are positing that the total Hispanic population is given by , but

 =  + =  +  = H.

*q.e.d.*

In other words, to maximize the accuracy of our [0,1] classifications of names as either Hispanic or not Hispanic we wish to set the number of Type I errors (false positives) equal to the number of Type II errors (false negatives).

Proposition 3: If, for each surname, its share of total Hispanic population and its share of total non-Hispanic population is treated as fixed, then the number of (most Hispanic) names we would need to use to equalize the number of Type I and Type II errors increases with the proportion Hispanic in the total population.

Proof: The proof of this proposition is quite straightforward. For any given cutoff point, as we increase the proportion Hispanic in the sample, the number of Type I errors (false positives) above that cutoff declines, since we are reducing the share of non-Hispanics in the population. Thus, the number of non-Hispanics in each surname will also go down since we are assuming that the proportion of non-Hispanics coming from any given surname is fixed. Similarly, for that same cutoff point, as we increase the proportion Hispanic in the sample, the number of Type II errors (false negatives) below that cutoff increases, since we are increasing the share of Hispanics in the population. Thus, the number of Hispanics in each surname will also go up since we are assuming that the proportion of Hispanics coming from any given surname is fixed. But, if we have reduced Type I error to the right of the former cutoff and increased Type II error in the other direction, then to equalize the two now requires us to increase the number of names we count as 100% Hispanic, i.e., lower the threshold.[[1]](#footnote-1) *q.e.d.*

In the next proposition we offer an alternative way to think about what needs to be equalized to maximize the predictive success of our choice of name threshold.

Proposition 4: If we array names from most Hispanic to least Hispanic and we treat the first s names as 100% Hispanic and the remaining names (from the (s+1)th to the Nth) as non-Hispanic, then the value of s such that thenames classified as Hispanic yield the true Hispanic population is given by s such that the (cumulative) average Hispanic share of the population among the names from the first to the sth name equals the proportion of the total Hispanic population found among those names, i.e.,

 = 

Proof: Once we set up this proposition in mathematical notation, the result become obvious, since we have the same numerator on both sides and the denominators are equal by assumption of our choice of s*. q.e.d.*

The intuitive meaning of this proposition is less clear than that of either of our other three propositions, but in the later empirical section we will be able to give Proposition 4 an enlightening (and perhaps surprising) empirical content.

Proposition 5: Consider any two surnames, say A and B, that have the property that they differ from one another both in the proportion of all those who claim Hispanic heritage who have each of the two surnames and in the proportion of all those who do not claim Hispanic heritage who have each of the two surnames. If we assume that any given population of Hispanics is a close to random sample from the national population of Hispanics in terms of surnames and any population of non-Hispanics is a close to random sample from the national population of non-Hispanics in terms of surnames, and we know the shares of the national Hispanic and national non-Hispanic population, respectively, that each surname constitutes, by finding the ratio of those in a given population who have surname A to those who have surname B, we can directly infer the Hispanic proportion of that population.

Proof: This proposition follows directly from the law of conditional probability and from Bayes Theorem. We start with

prob(Hispanic|name A) \* prob(name A) = prob(name A |Hispanic) \* prob(Hispanic)

From this identity we derive the equation

prob(name A) = (prob(name A |Hispanic)\* prob(Hispanic)/ prob(Hispanic|name A)

Similarly,

prob(name B) = (prob(name B |Hispanic)\* prob(Hispanic)/ prob(Hispanic|name B)

Dividing these two equations we obtain the ratio

prob(name A)/ prob(name B) =

­­­­­­­­­­­­­­­ prob(name A |Hispanic)\* prob(Hispanic) / prob(Hispanic|name A)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (2)

prob(name B |Hispanic)\* prob(Hispanic)/ prob(Hispanic|name B)

Therefore,

prob(name A)/ prob(name B) =

prob(name A |Hispanic) / prob(Hispanic|name A)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (2)’

prob(name B |Hispanic)/ prob(Hispanic|name B)

since one of the terms in Eq. (2) is found in both numerator and denominator and may be cancelled out.

Moving terms from numerator to denominator, we may write Eq. (2)’ as Eq. (2)’’ below.

prob(name A)/ prob(name B) =

prob(name A |Hispanic) \* prob(Hispanic|name B)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (2)’’

prob(name B |Hispanic)\*prob(Hispanic|name A)

Now, we can twice substitute the identity of Bayes Theorem, Eq. (1) into Eq. (2), to eliminate two of the conditional probabilities in that equation. We obtain, after some algebra, Eq. (3).

prob(name A)/ prob(name B) =

prob(name B |Hispanic) \* prob(Hispanic) \* prob(name A |Hispanic)

/(prob(name A |Hispanic) \* prob(Hispanic) + prob(name A |non-Hispanic) \* prob(non-Hispanic))

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (3)

prob(name A |Hispanic) \* prob(Hispanic) \* prob (name B |Hispanic)

/(prob(name B |Hispanic) \* prob(Hispanic) + prob(name B |non-Hispanic) \* prob(non-Hispanic))

Which, in turn, after cancellation, simplifies to

prob(name A)/ prob(name B) =

prob(name A |Hispanic) \* prob(Hispanic) + prob(name A |non-Hispanic) \* prob(non-Hispanic)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (3)’

prob(name B |Hispanic) \* prob(Hispanic) + prob(name B |non-Hispanic) \* prob(non-Hispanic)

But, since we may take prob(name A |Hispanic), prob(name A |non-Hispanic), prob(name B |Hispanic), and prob(name B |non-Hispanic) to be essentially known parameters (from the national sample), and since

prob(Hispanic) = 1 – prob(non-Hispanic)

= the proportion of Hispanic /those of Spanish heritage in the sample,

once we know the actual ratio of those with surname A to those with surname B in our sample, under the above assumptions, by plugging in the other four known (subject only to sampling error) parameter values into Eq. (3), after straightforward simple algebra, we can directly calculate the Hispanic proportion in the sample which, of course, is what we want to find. *q.e.d.*

1. Note that this result does not necessarily go through were we to array names not according to their percentage Hispanic but according to what proportion of all Hispanics are found with that name. The latter takes into account how common the name is, while the former does not. [↑](#footnote-ref-1)