Example of Fellegi-Sunter Match Probabilities

Lars Vilhuber

November 1, 2017

Setup

From the INFO7470 lecture, we have

$$R \equiv \frac{Pr[\gamma_r | ab_r \in M]}{Pr[\gamma_r | ab_r \in U]}$$

In the lecture, John asks you to consider the case when the single variable under consideration is "sex":

$$R \equiv \frac{Pr[\gamma_r^1 | ab_r \in M]}{Pr[\gamma_r^1 | ab_r \in U]}$$

So assume that the binary variable $a_1 = b_1 = \texttt{sex}$ in this dataset is coded `m` for male and `f` for female, and that due to entry errors it may be miscoded in A about **1%** of the time, but is recorded with **100%** correctly in B. Both datasets are drawn from and representative of the general U.S. population in 2010 (see f.i. Age and Sex Composition: 2010).

Question 1

What is $Pr[\gamma_r^1 | ab_r \in M]$?

Answer

When the two records are from the set of true matches, there is still a chance that the sex variable is miscoded on one of the two source records. Thus, up to **1%**, the error rate in A, regardless of sex, $Pr[\gamma_r^1 | ab_r \in M] = 0.99$.

Question 2

What is $Pr[\gamma_r^1 | ab_r \in U]$?

Answer:

In 2010, 50.9% of the U.S. population was female (p_f). So the likelihood that records in the set of non-matches have the same sex is approximately equal to the population percentage, if the samples are large enough. If $sex = \texttt{m}$, then that probability is $p_m = \mathbf{49.1}$, and if $sex = \texttt{f}$, that probability is $p_f = \mathbf{49.1}$. Thus

$$\begin{aligned} Pr[\gamma_r^1 | ab_r \in U] &= Pr[\gamma_r^1 | ab_r \in U, a = \text{“}m\text{”}] Pr[a = \text{“}m\text{”}] \\ &\quad + Pr[\gamma_r^1 | ab_r \in U, a = \text{“}f\text{”}] Pr[a = \text{“}f\text{”}] \\ &= p_m^2 + p_f^2 \end{aligned} \tag{1}$$

$= 0.500162$

Question 3

What then is w_r?

Answer

$w_r = log_2(R^*) = log_2(\ 0.99\ /\ 0.500162\ ) \approx log_2(\ 1.979\ ) \approx 0.985.$