p1_write_up

Name: Yue Liu

NetID: yl992

Question 1

Step 5

My Prevalence Calculations:

SEX	Count_Yes	Count_No	Sex_Prevalence
1	13897844	92318176	0.13084508344409818
2	15541109	139022282	0.10054844746515687

_AGEG5YR	Count_Yes	Count_No	Age_Prevalence
12	2381834	10936121	0.1788438239955008
1	126534	12838634	0.009759534161069104
13	2006829	10520506	0.16019600337980902
6	1207538	14253176	0.07810363738699261
3	324417	11762560	0.026840209921802614
5	471310	10404063	0.04333736415293526
9	5255536	33823796	0.13448377264995215
4	431783	11882862	0.03506256168976044
8	3853245	28253679	0.1200128981524359
7	2096156	20594390	0.09238014810220961
10	5912198	32542342	0.15374512346266528
11	5007216	21963347	0.18565485637062898
14	217055	1650690	0.11621233091241041
2	147302	9914292	0.014640026222485225

_IMPRACE	Count_Yes	Count_No	Race_Prevalence
1	28228377	222261446	0.11269271007469234
6	27471	277151	0.09018061728962452
3	21863	357149	0.05768418941880468
5	469740	4828186	0.0886648850889952
4	22782	73714	0.23609268777980436
2	668720	3542812	0.15878307466261685

Actual Prevalence:

I found this CDC to be a good source: Prevalence of Both Diagnosed and Undiagnosed Diabetes | Diabetes | CDC

	Calculated Prevalence (%)	Actual Prevalence (%)
Male	13.1	15.4 (13.5–17.5)
Female	10.1	14.1 (11.8–16.7)

The calculated sex prevalence is quite close to the actual prevalence.

	Calculated Prevalence (%)	Actual Prevalence (%)
White, non-Hispanic	11.3	13.6 (11.4–16.2)
Black, non-Hispanic	15.9	17.4 (15.2–19.8)
Asian, non-Hispanic	5.8	16.7 (14.0–19.8)
Hispanic	8.9	15.5 (13.8–17.3)

The calculated race prevalence for white and black is close to the actual prevalence, but Asian and Hispanic people are not.

	Calculated Prevalence (%)	Actual Prevalence (%)
18–44	7.4	4.8 (4.0–5.9)
45–64	42	18.9 (16.1–22.1)
⩾65	67	29.2 (26.4–32.1)

The calculated age prevalence for age group 18-44 is close to the actual prevalence, but the prevalence for age group 45-64 and >65 is not. (7.4 vs 4.8, 42 vs 18.9, 67 vs 29.2)

How to improve:

Each row in the BRFSS dataset does not necessarily correspond to a single person in the U.S and joining BRFSS and nhis could potentially double count too. Both of these could contribute to making the data more imbalanced, since the majority groups are more likely to be counted more than once than minority group. This can be proved by the comparison above, as

calculated prevalence for minority groups Asian and Hispanic is lower than the actual prevalence. Additionally, the actual prevalence has different categories from what I have, which requires me to sum up categories. As it is in the case for age, this introduces more variance and larger margin of error for my final result.

A primary key, which could be the ssn or some other form ID that unique identifies a person, can greatly improve the accuracy of my calculation by eliminating the potential problem and multi-counting.

Thanks for grading. Have a nice day :)