**HW1**

**Table 1. Demographic distributions among those who have consumed any alcohol in the past year at baseline (alc\_bin = 1) compared to those who have not consumed any alcohol in the past year at baseline (alc\_bin = 0)**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable  N = 4136 | Among those that did not consume alcohol in previous year at baseline  n = 3374 | Among those that did consume alcohol in previous year at baseline  n = 749 | p-value |
| *Variable\** | *n (%)* | *n (%)* |  |
| Gender  Male  Female | 1629 (48.28%)  1740 (51.57%) | 415 (55.41%)  329 (43.93%) | 0.0002\*\* |
| Race  Black only  White only  Hispanic only  Other | 1475 (43.72%)  445 (13.19%)  950 (28.16%)  494 (14.64%) | 303 (40.45%)  93 (12.42%)  239 (31.91%)  112 (14.95%) | 0.3071 |
| Family Structure  Mother and Father Together  Other | 1792 (53.11%)  1562 (46.30%) | 353 (47.13%)  393 (52.47%) | <0.0001\*\* |
| Eligible for Free/Reduce Price Lunch  Yes  No  Don’t Know | 2345 (69.50%)  673 (19.95%)  344 (10.20%) | 505 (67.42%)  178 (23.77%)  63 (8.41%) | <0.0001\*\* |

\*Row Percentages may not add up to 100%, depending on the number of individuals missing information for the defined categories

\*\*Statistically significant assuming 95% level of confidence (α = 0.05)

To generate p-values within table 1 and to determine whether there is a statistically significant difference in the distribution of each demographic variable (gender, race, family structure, and eligibility for free or reduced-price lunches) between those who have used alcohol in the past year at baseline and those who have not used alcohol in the past year at baseline, I used a **chi-square test**. The explanatory variables, gender (= 22.15, p = 0.0002 < 0.05), family structure (= 114.73, p = <0.0001 < 0.05), and eligibility for free/reduced-price lunches (= 86.67, p = <0.0001 < 0.05), all seemed to be associated with past year alcohol use at the 95% level of confidence. However, race **did not** exhibit this association at the 95% level of confidence (= 9.43, p = 0.3071 > 0.05).

From this data, it seems that male students, as compared to female students, are more likely to consume alcohol in the past year. Students with a family structure other than one where their mother and father are still together, as compared to those whose mother and father are still together, are more likely to consume alcohol in the past year. Also, those eligible for free/reduced-price lunches, as compared to those who are not eligible or do not know their eligibility status, are more likely to consume alcohol in the past year. Race was not found to be statistically significantly associated with alcohol use in the past year, though the data seemed to imply that students in the “black only” category *may be* more likely to consume alcohol in the past year.

To make the greatest public health impact, it is important to design interventions that target the groups of individuals who are much more susceptible to adolescent use of alcohol. Ideally, we would reach the entire population with our designed intervention, but this approach is not often feasible given the limited resources available for public health interventions. Therefore, we must target the groups that are most likely to be affected by adolescent alcohol use to have the greatest impact. Our sample statistics indicate that it would be most efficient and effective to target male students and students with family structures other than one where their mother and father are still together with adolescent alcohol use reduction interventions. It would also be important to target students who are eligible for free/reduced-price lunches.

Prior to making any programmatic decisions, I would like to know whether the sample surveyed is representative/generalizable to the larger population of students that would be targeted by the program. For example, if the data is not generalizable, it would not be wise to base programmatic decisions off this data alone. Additionally, I would want to know the ability of each of the statistically significant explanatory variables (gender, family structure, and eligibility for free/reduced-price lunches) to predict alcohol use in the past year. To assess this, my next steps as a data analyst would include modeling the available data using multiple logistic regression (rather than linear regression since the outcome variable is categorical). I could then use the results from multiple logistic regression modeling to determine how the odds of alcohol use in the past year changes for a certain category of each explanatory variable compared to a set reference category. (For example, we can use multiple logistic regression to determine the odds of alcohol use in the past year for male students as compared to female students, by using PROC LOGISTIC). This would allow me to determine whether targeting of male students, students whose mother and father are not together, and/or students who are eligible for free/reduced-price lunches would be worth the allocated effort and resources prior to program implementation. It would also allow me to somewhat understand the amount of change I can expect to see due to the implemented program, though attributable risk may be a better measure for assessing this.

**SAS Code**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Course: Data Analysis - EPID 5314 \*

\* Assignment: Homework 1 \*

\* Due Date: 9/1/2021 \*

\* Programmer(s): Jessie Ausman \*

\* Program Name: HW1 \*

\* Save Program/Log/Output: C:\Users\jessa\Desktop\EPID 5314\Homework1\*

\* Save Data Files: C:\Users\jessa\Desktop\EPID 5314\PNC Data File\PNC Datasets \*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Read in Datasets

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

Libname PNC "C:\Users\jessa\Desktop\EPID 5314\PNC Data File\PNC Datasets";

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Create Temporary Datasets

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

**Data** wave1;

set pnc.pnc02;

**run**;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Create Categorical Var for Alcohol Use in Last Year

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

**Proc** **freq** data=wave1;

table alcyear1;

**run**;

**data** wave1\_2;

set wave1;

if alcyear1 in (**2**,**3**,**4**,**5**,**6**,**7**) then alc\_bin1 = **1**;

if alcyear1 eq **1** then alc\_bin1 = **0**;

if missing(alcyear1) then alc\_bin1 = **.**;

else;

**run**;

**proc** **freq** data=wave1\_2;

table alc\_bin1;

**run**;

**data** wave1\_3;

set wave1\_2;

**run**;

**proc** **freq** data=wave1\_3;

table alc\_bin1;

**run**;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Create Var for "other race" (a1, d1, o1)

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

**proc** **freq** data=wave1\_3;

table race1;

**run**;

**data** wave1\_4;

set wave1\_3;

if race1 = **2** then race\_black = **1**;

if race1 in (**1**,**3**,**4**,**5**,**6**) then race\_black = **0**;

if missing(race1) then race\_black = **.**;

else;

**run**;

**proc** **freq** data=wave1\_4;

table race\_black;

**run**;

**data** wave1\_5;

set wave1\_4;

if race1 = **5** then race\_white = **1**;

if race1 in (**1**,**2**,**3**,**4**,**6**) then race\_white = **0**;

if missing(race1) then race\_white = **.**;

else;

**run**;

**proc** **freq** data=wave1\_5;

table race\_white;

**run**;

**data** wave1\_6;

set wave1\_5;

if race1 = **3** then race\_his = **1**;

if race1 in (**1**,**2**,**4**,**5**,**6**) then race\_his = **0**;

if missing(race1) then race\_his = **.**;

else;

**run**;

**proc** **freq** data=wave1\_6;

table race\_his;

**run**;

**data** wave1\_7;

set wave1\_6;

if race\_black = **0** or race\_white = **0** or race\_his = **0** then race\_other = **1**;

if missing(race\_black) then race\_other = **.**;

if race\_black = **1** or race\_white = **1** or race\_his = **1** then race\_other = **0**;

else;

**run**;

**proc** **freq** data=wave1\_7;

table race\_other;

**run**;

**data** wave1\_8;

set wave1\_7;

if race\_black = **1** then race\_tot = **1**;

if race\_white = **1** then race\_tot = **2**;

if race\_his = **1** then race\_tot = **3**;

if race\_other = **1** then race\_tot = **4**;

if missing(race1) then race\_tot = **.**;

else;

**run**;

**proc** **freq** data=wave1\_8;

table race\_tot;

**run**;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Create New Var for "family structure" 1 vs. (2,3,4,5,6,7,8)

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

**proc** **freq** data=wave1\_8;

table family1;

**run**;

**data** wave1\_9;

set wave1\_8;

if family1 = **1** then fam\_together = **1**;

if missing(family1) then fam\_together = **.**;

if family1 in (**2**,**3**,**4**,**5**,**6**,**7**,**8**,**9**) then fam\_together = **0**;

else;

**run**;

**proc** **freq** data=wave1\_9;

table fam\_together;

**run**;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Create PROC FREQ Tables

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*Table for Alcohol use and Gender;

**Proc** **freq** data=wave1\_9;

tables alc\_bin1 \* gender1/ missing chisq;

**run**;

\*Table for Alcohol use and Race;

**Proc** **freq** data=wave1\_9;

tables alc\_bin1 \* race\_tot/ missing chisq;

**run**;

\*Table for Alcohol Use and Family Structure;

**Proc** **freq** data=wave1\_9;

tables alc\_bin1 \* fam\_together/ missing chisq;

**run**;

\*Table for Alcohol Use and Free Lunch;

**Proc** **freq** data=wave1\_9;

tables alc\_bin1 \* lunch1 / missing chisq;

**run**;