**BS723 - Spring 2022**

**Project 1**

**Background**

Starting in 1997, the federal government mandated the separation of the “Asian or Pacific Islander” race category into two separate categories:

(1) *Asian* – “A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam.” (1)

(2) *Native Hawaiian or Other Pacific Islander* – “A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands. The term “Native Hawaiian” does not include individuals who are native to the State of Hawaii by virtue of being born there.” (1)

According to 2010 census data, only 0.4% of the U.S. population identifies as Native Hawaiian or Pacific Islander (NHPI) or NHPI in combination with another race. (2) Due to its small size, researchers conducting federal surveys have struggled to obtain sufficient sample sizes to estimate health statistics for the NHPI population.

In 2014, the National Center for Health Statistics conducted the first federal study of NHPI civilians residing in the U.S. with the development of the Native Hawaiian and Pacific Islander National Health Interview Survey (NHPI NHIS). (4) The survey was modelled after the National Health Interview Survey, which has been administered to the general U.S. population regularly for the past 50 years.

For this project you will use subset of the publicly-available data from the 2014 NHPI NHIS survey to explore sex differences for several measures.

1. To determine differences in risk factors among participants in the 2014 NHPI NHIS survey compared to overall USA measures.

2. To examine the association between sex and *prevalent events* among participants in the 2014 NHPI NHIS survey.

A description of the variables in the dataset “**proj1\_2022 sas7bdat**” can be found on pages 2-3. Any missing data is denoted by a period (.) for numeric variables and by a blank space (“ “ ) for character variables.

**References**

(1) Executive Office of the President, Office of Management and Budget, Office of Information and Regulatory Affairs, *Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity*,1997, Washington, DC,OMB Publications Office.

(<https://www.federalregister.gov/documents/1997/10/30/97-28653/revisions-to-the-standards-for-the-classification-of-federal-data-on-race-and-ethnicity>)

(2) Galinsky AM, Zelaya CE, Simile C, Barnes PM. Health conditions and behaviors of Native Hawaiian and Pacific Islander persons in the United States, 2014. National Center for Health Statistics. Vital Health Stat 3(40). 2017.

(3) <https://www.nimhd.nih.gov/about/overview/>

(4) <https://www.cdc.gov/nchs/nhis/nhpi.html>

(5) Kessler, R. C., Barker, P. R, Colpe, L. J., Epstein, J. F., Gfroerer, J. C., Hiripi, E., Howes, M.J., Normand, S.L., Manderscheid, R. W., Walters, E. E., Zaslavsky, A. M. (2003). Screening for serious mental illness in the general population. Arch Gen Psychiatry, 60(2), 184-9.

| **Variable Name** | **Description** | **Coding** |
| --- | --- | --- |
| ID | Participant ID | Numeric, 1-2590 |
| AGE | Participant age | 18-74 = 18-74 years  75 = 75+years |
| SEX | Participant sex | 1 =Male  2 =Female |
| MAR\_STAT | Marital status | 1=Married  2=Widowed  3=Divorced or separated  4=Never married  5=Living with a partner  9=Unknown |
| HEIGHT | Height in inches | Numeric, with special values:  96=Not available  97=Refused  98=Not ascertained  99=Don't know |
| WEIGHT | Weight (pounds) | Numeric, with special values:  996=Not available"  997=Refused"  998=Not ascertained"  999=Don't know" |
| HTN | Hypertension? | 1=Yes  2=No  7=Refused  8=Not ascertained  9=Don't know |
| HYBPLEV | Blood pressure level | 1=Not told  2=High  3=Normal  4=Low  5=Borderline  7=Refused  8=Not ascertained  9=Don't know" |
| HTNMED | Hypertension medicine? | 1=Yes  2=No  7=Refused  8=Not ascertained  9=Don't know |
| MI | Heart attack? | 1=Yes  2=No  7=Refused  8=Not ascertained  9=Don't know |
| SMKSTAT2 | Smoking Status: | 1=Current every day smoker  2=Current some day smoker  3=Former smoker  4=Never smoker  5=Smoker, current status unknown  9=Unknown if ever smoked |
| COPD | COPD (Chronic obstructive pulmonary disease)? | 1=Yes  2=No  7=Refused  8=Not ascertained  9=Don't know |
| CANCER | Cancer? | 1=Yes  2=No  7=Refused  8=Not ascertained  9=Don't know |
| DIABETES | Diabetes? | 1=Yes  2=No  3=Borderline  7=Refused  8=Not ascertained  9=Don't know |

**Part 1: Data Preparation (35 pts)**

1. Create a library called “proj1” that points to the directory containing the permanent dataset **proj1.proj1\_2022.sas7bdat**.
2. Using the code book above, in a single data step, create a temporary dataset called “*nhpi*” with the following changes:
   1. For all variables, recode any values of “Refused”, “Not Ascertained” or “Don’t know” to missing (.). [applies to **HTN**, **HTNMED**, **MI**, **COPD**, **CANCER** and **DIABETES**]
   2. For **HEIGHT** and **WEIGHT**, convert all special values to missing (.).
   3. Use **HEIGHT** and **WEIGHT** to create a **BMI** variable, where
3. Create labels for all variables using the “Description” in the table above. Create a label for the BMI variable that your created also.
4. Using the **SMKSTAT2** variable, create a new **SMOKING** variable so that:
   1. 1, 2 or 5 are coded as “current”
   2. 3 is coded as “former”
   3. 4 is coded as “never”
   4. 9 is coded as “unknown”
5. For **HTN**, **HTNMED**, **MI**, **COPD**, and **CANCER**, assign a format that categorizes it as 1 being “YES”, 2 is “NO” and all others as “Unknown”. Format **DIABETES** similarly, buta value of 3 should have a “Borderline” format value. For **SEX**, format 1 as “Male” and 2 as “Female”
6. Recode the **MAR\_STAT** variable to be:
   1. 0 for “Never married”
   2. 1 for “Widowed”, “divorced”, or “separated”
   3. 2 for “Married” or “living with a partner”
   4. “Unknown marital status” should be set to missing (.)

**\*\*\*See SAS Code at end of this document on page 8\*\*\***

**Part 2: Making a table 1 (25 pts)**

Most research papers have a first table that describes the cohort using descriptive statistics

* For continuous measures, provide means and standard deviations for each variable [mean (SD)]and provide a p-value for the test of significant differences between men and women for that variable
* For categorical variables, provide frequencies and percentages for each category [frequency (percent)] and a p-value for the test of significant differences among the categories between men and women for that variable
* **Note**: missing or unknown values or categories should not be included in significance tests, but should be displayed in the table as a category with frequencies and percentages.
* Based on the table one that you created, which risk factors are significantly different when comparing men and women?

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Men (n=1231)** | **Women (n=1359)** | **P-value** |
| AGE (years)\* | 46.49 (16.06) | 47.87 (15.70) | 0.0276 |
| MAR\_STAT\*\*\*  Married/living with partner  Widowed, separated, or divorced  Never married | 824 (66.95)  131 (10.64)  276 (22.42) | 843 (62.03)  273 (20.09)  241 (17.73) | <0.0001 |
| HEIGHT (in)\*\*\* | 69.31 (2.77) | 63.89 (2.49) | <0.0001 |
| WEIGHT (lbs)\*\*\* | 198.95 (35.81) | 166.67 (37.21) | <0.0001 |
| BMI (kg/m2) | 29.12 (4.85) | 28.70 (6.15) | 0.0697 |
| HTN  No  Yes  Unknown | 773 (62.79%)  456 (37.04%)  2 (0.16%) | 867 (63.80%)  488 (35.91%)  4 (0.29%) | 0.5660 |
| HTNMED  No  Yes  Unknown | 85 (6.90%)  370 (30.06%)  776 (63.04%) | 84 (6.18%)  404 (29.73%)  871 (64.09%) | 0.5569 |
| MI\*\*  No  Yes  Unknown | 1179 (95.78%)  51 (4.14%)  1 (0.08%) | 1328 (97.72%)  28 (2.06%)  3 (0.22%) | 0.0021 |
| SMOKING\*\*  Current  Former  Never  Unknown | 244 (19.82%)  322 (26.16%)  659 (53.53%)  6 (0.49%) | 215 (15.82%)  324 (23.84%)  815 (59.97%)  5 (0.37%) | 0.0069 |
| COPD  No  Yes  Unknown | 1199 (97.40%)  28 (2.27%)  4 (0.32%) | 1322 (97.28%)  31 (2.28%)  6 (0.44%) | 0.9875 |
| CANCER\*\*\*  No  Yes  Unknown | 1175 (95.45%)  55 (4.47%)  1 (0.08%) | 1255 (92.35%)  103 (7.58%)  1 (0.07%) | 0.0010 |
| DIABETES  No  Borderline  Yes  Unknown | 997 (80.99%)  40 (3.25%)  193 (15.68%)  1 (0.08%) | 1114 (81.97%)  55 (4.05%)  188 (13.83%)  2 (0.15%) | 0.2605 |

**\*** p ≤ 0.05 **\*\*** P ≤ 0.01 **\*\*\*** p ≤ 0.001

The risk factors that are significantly different at the 0.05 level when comparing men and women were: age, marital status, height, weight, myocardial infarction status, smoking status, and cancer status.

**Part 3: Study Questions (40 pts)**

1. The table below shows the mean **BMI** value for the entire US in 2014 by **SEX**. ***Within each*** **SEX**, perform the appropriate test to determine if the mean **BMI** in the NHPI population is different from the overall US BMI. Report the null and alternative hypotheses, name of test, test statistic, p-value, and your conclusion for males and females.

|  |  |
| --- | --- |
| **Sex** | **Mean BMI (All USA)** |
| **M** | 27.80 |
| **F** | 27.51 |

The null hypothesis is that the mean BMI among NHPI males in our sample is equal to the mean BMI of males in the U.S. of 27.80 kg/m2. The alternative hypothesis is that the mean BMI among NHPI males in our sample is not equal to the mean BMI of males in the U.S. of 27.80 kg/m2. Similarly, for females our null hypothesis is that the mean BMI among NHPI females in our sample is equal to the mean BMI of females in the U.S. of 27.51 kg/m2. The alternative hypothesis is that the mean BMI among NHPI females in our sample is not equal to the mean BMI of females in the U.S. of 27.51 kg/m2.

We will conduct a one sample t-test of means for both males and females at the 0.05 level. In our sample, the mean BMI for males was 29.1224 kg/m2 with a 95% confidence interval of (28.8347, 29.4100). We calculated a t-statistic of 9.02 with 1093 degrees of freedom, and p<0.0001, therefore we reject the null hypothesis. With such a low p-value (<0.0001) there is evidence to suggest that the average BMI of NHPI males within our sample is significantly different and higher than the U.S. male mean BMI and is not equal to 27.80 kg/m2. Our sample mean BMI for NHPI females is 28.7048 kg/m2 with a 95% confidence interval of (28.3569, 29.0526). We calculated a t-statistic of 6.74 with 1202 degrees of freedom, and p<0.0001, and we reject the null hypothesis. With such a low p-value (<0.0001) there is evidence to suggest that the average BMI of females within our sample is significantly different and higher than the U.S. female mean BMI and is not equal to 27.51 kg/m2.

1. The investigators were also interested in whether there were significant differences in **BMI** by **SEX** within the NHPI population. Report the null and alternative hypotheses, name of test, test statistic, p-value, the appropriate measure of effect (+ 95% confidence interval), and your conclusion.

The null hypothesis is that the mean BMI of NHPI males in our sample is equal to the mean BMI of NHPI females in our sample. The alternative hypothesis is that the mean BMI of NHPI males in our sample is not equal and is different from the mean BMI of NHPI females in our sample.

The mean BMI of the NHPI males in our sample is 29.1224 kg/m2 (sd=4.8496) compared to a mean BMI of 28.7048 kg/m2 (sd=6.1493) among our NHPI females. The mean BMI of NHPI males is 0.4176 kg/m2 lower than the mean BMI of NHPI females. The 95% confidence interval of this difference is (-0.0336, 0.8687).

First we performed an F-test of equal variances. There was significant evidence (F=1.61, df=(1202,1903), p<0.0001) that the variances are not equal and are different from one another. We therefore used a two-sample t-test of means assuming unequal variances to test whether the mean BMI of NHPI males in our sample was different from the mean BMI of NHPI females in our sample. The resulting t-statistic is 1.82 with 2250.7 degrees of freedom and p=0.0697.

We therefore fail to reject the null that the mean BMIs of males and females in our sample are equal at the 0.05 level because our p-value was greater than our alpha level of 0.05. There is not a statistically significant difference (p=0.0697) between the mean BMI of NHPI males in our sample (29.1224 kg/m2) compared to the mean BMI of NHPI females in our sample (28.7048 kg/m2).

1. Another question of interest was whether there was an association between **SEX** and **CANCER** in subjects who are current smokers. Report the null and alternative hypotheses, name of test, test statistic, p-value, the appropriate measure of effect (+ 95% confidence interval), and your conclusion.

The null hypothesis is that the risk of cancer is the same across males who are current smokers and females who are current smokers (RR=1). The alternative hypothesis is that the risk of cancer is different between males who are current smokers and females who are current smokers (RR≠1).

We will conduct a chi-square test of independence at the 0.05 level. The proportion of NHPI male current smokers in our sample who have cancer was 0.0246, while the proportion of NHPI female current smokers in our sample who have cancer was 0.0558. The relative risk ratio was 0.4406 (95% CI=0.1682, 1.1537), meaning that male current smokers had 0.4406 times the risk (56% reduced risk) of having cancer compared to female current smokers in our sample.

We calculated a chi-square test statistic of 2.9573 with 1 degree of freedom, and p=0.0855, therefore we fail to reject the null hypothesis at the 0.05 level that the risk of cancer is the same across NHPI male and female current smokers in our sample and that the relative risk ratio was equal to 1. There is not significant evidence (p=0.0855) that the relative risk ratio is not equal to 1 nor is there significant evidence that the risk of cancer is not the same between male current smokers and female current smokers in our sample.

**SAS Code:**

libname proj1 "/home/u60739998/BS 723/Project 1";

proc format;

value htnf 1="Yes" 2="No" .="Unknown";

value htnmedf 1="Yes" 2="No" .="Unknown";

value mif 1="Yes" 2="No" .="Unknown";

value copdf 1="Yes" 2="No" .="Unknown";

value cancerf 1="Yes" 2="No" .="Unknown";

value diabetesf 1="Yes" 2="No" 3="Borderline" .="Unknown";

value sexf 1="Male" 2="Female";

run;

data nhpi;

set proj1.proj1\_2022;

if htn in(7, 8, 9) then htn=.;

if htnmed in(7, 8, 9) then htnmed=.;

if mi in(7, 8, 9) then mi=.;

if copd in(7, 8, 9) then copd=.;

if cancer in(7, 8, 9) then cancer=.;

if diabetes in(7, 8, 9) then diabetes=.;

if height in(96, 97, 98, 99) then height=.;

if weight in(996, 997, 998, 999) then weight=.;

BMI = (weight\*703)/(height\*\*2);

label id="Participant ID" age="Participant age" sex="Participant sex" mar\_stat="Marital status"

height="Height in inches" weight="Weight (pounds)" htn="Hypertension?" hybplev="Blood pressure level"

htnmed="Hypertension medicine?" mi="Heart attack?" smkstat2="Smoking status:"

copd="COPD (Chronic obstructive pulmonary disease)?" cancer="Cancer?" diabetes="Diabetes?"

bmi="Body mass index (kg/m2)";

if smkstat2 in(1, 2, 5) then SMOKING="Current";

else if smkstat2= 3 then SMOKING="Former";

else if smkstat2= 4 then SMOKING="Never";

else if smkstat2= 9 then SMOKING="Unknown";

format htn htnf. htnmed htnmedf. mi mif. copd copdf. cancer cancerf. diabetes diabetesf. sex sexf.;

if mar\_stat=4 then mar\_stat=0;

else if mar\_stat in(2,3) then mar\_stat=1;

else if mar\_stat in(1,5) then mar\_stat=2;

else if mar\_stat=9 then mar\_stat=.;

run;

/\* table data for quantitative variables \*/

proc means data=nhpi;

var age height weight bmi;

where sex=1;

run;

proc means data=nhpi;

var age height weight bmi;

where sex=2;

run;

/\* get p-values comparing males to females for quant. variables \*/

proc ttest data=nhpi;

class sex;

var age height weight bmi;

run;

/\* table data for qualitative variables \*/

proc freq data=nhpi;

tables sex\*mar\_stat sex\*htn sex\*htnmed sex\*mi sex\*smoking sex\*copd sex\*cancer sex\*diabetes / missing;

run;

/\* run chi-sq test for qual. variables \*/

proc freq data=nhpi;

tables sex\*mar\_stat sex\*htn sex\*htnmed sex\*mi sex\*smoking sex\*copd sex\*cancer sex\*diabetes / expected chisq fisher nocol nopercent;

run;

/\* statistical testing \*/

data nhpi2;

set nhpi;

run;

proc sort data=nhpi2;

by sex;

run;

/\* one sample t-test of means, Question 1 \*/

proc ttest data=nhpi2 h0=27.80 alpha=0.05;

by sex;

var bmi;

run;

proc ttest data=nhpi2 h0=27.51 alpha=0.05;

by sex;

var bmi;

run;

/\* two sample t test of means/independent samples t-test, Question 2 \*/

proc ttest data=nhpi;

class sex;

var bmi;

run;

/\* Question 3 \*/

data nhpi3;

set nhpi;

run;

proc sort data=nhpi3;

by smoking;

run;

proc freq data=nhpi3;

by smoking;

tables sex\*cancer / expected measures chisq fisher nocol nopercent;

run;