Case Study: SAS BRFSS 10 Categorical Table 1

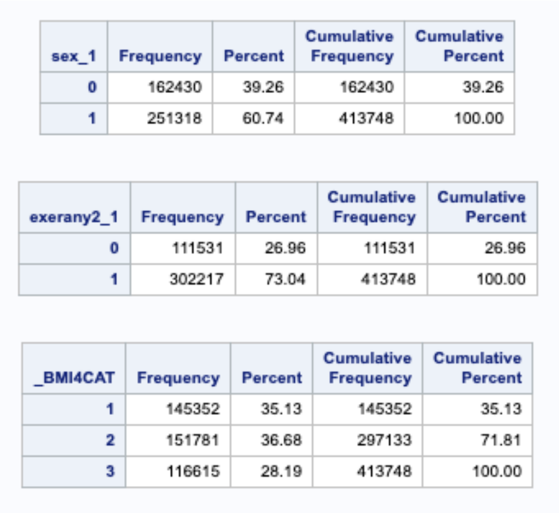
This case study will reintroduce you to categorical variables and how you may use statistical procedures to investigate these types of data, display them in a "Table 1", and develop a results section describing your Table 1. Use data from the Behavioral Health Needs Assessment Survey from 2010 to complete this case study.

It is expected that all students have taken CITI training prior to conducting these secondary data analysis case studies. Please go to http://ohrc.nu.edu/, visit the tab "IRB", and read about what an IRB is and the role of the National University IRB. Then please make sure to visit the CITI link and take the training.
Next, go to the NU Health Science Research Center http://ohrc.nu.edu/ and go into "Data Sets". From there go to "**Source**: government agency or foundation" then "Federal (National)", and then "Centers for Disease Control and Prevention [CDC]". From there, you will see the link for **Behavioral Risk Factor Surveillance System (BRFSS)**.
From this link find 2010 data and download the SAS zipped file and the codebook.
The objective of this analysis is to investigate the association between **diabetes** and **BMI** after controlling for **exercise** and **gender**. The outcome variable is **diabetes** and the variable of interest (exposure) is **BMI**.

Conduct a **complete case analysis** for this objective following these guidelines:

1. Use the raw variable categorization of BMI (_BMI4CAT)
2. Categorize gender (SEX) into a two-level variable (male=0, female=1) where male is category 1 of the raw variable and female is category 2
3. Categorize diabetes (DIABETE2) into a two-level variable (no=0, yes=1) where yes is category 1 of the raw variable and no is category 3
4. Categorize exercise (EXERANY2) into a two-level variable (no=0, yes=1) where yes is category 1 of the raw variable and no is category 2
5. For the complete case analysis, restrict your sample based on the following conditions:
    a. 18<=AGE<=99
    b. SEX: raw categories 1 and 2
    c. DIABETE2: raw categories 1 and 3
    d. EXERANY2: raw categories 1 and 2
    e. Education (EDUCA): raw categories 1-6
    f. _BMI4CAT: raw categories 1,2 and 3
    g. General health (GENHLTH): raw categories 1-5

1) (10 pts) Using PROC FREQ, show the simple frequency tables for gender, exercise, and BMI



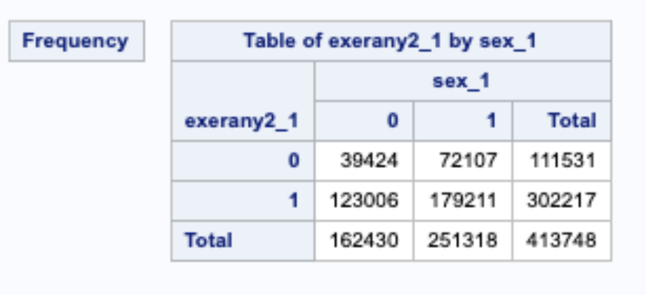
sex_1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	162430	39.26	162430	39.26
1	251318	60.74	413748	100.00

exerany2_1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	111531	26.96	111531	26.96
1	302217	73.04	413748	100.00

_BMI4CAT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	145352	35.13	145352	35.13
2	151781	36.68	297133	71.81
3	116615	28.19	413748	100.00

2) (20 pts) Using PROC FREQ, create a 2x2 contingency table for exercise by gender.

a. Show the PROC FREQ output **which shows only the counts in each cell**



Frequency

Table of exerany2_1 by sex_1

	sex_1		
exerany2_1	0	1	Total
0	39424	72107	111531
1	123006	179211	302217
Total	162430	251318	413748

b. Show your hand calculation of the chi-square statistic for testing whether there is an association between gender and exercise.

Proportion (P)

$P_{yes} = 302217 / 413748 = 0.7304374$
$P_{no} = 111531 / 413748 = 0.2695626$
$P_{male} = 162430 / 413748 = 0.392582$
$P_{female} = 251318 / 413748 = 0.607418$

$P_{yes+male} = (0.7304374)(0.392582) = 0.2867565$
$P_{yes+female} = (0.7304374)(0.607418) = 0.4436808$
$P_{no+male} = (0.2695626)(0.392582) = 0.1058254$
$P_{no+female} = (0.2695626)(0.607418) = 0.1637372$

Expected Count (E)

$E_{yes+male} = (0.2867565)(413748) = 118644.94$
$E_{yes+female} = (0.4436808)(413748) = 183572.06$
$E_{no+male} = (0.1058254)(413748) = 43785.058$
$E_{no+female} = (0.1637372)(413748) = 67745.942$

Observed (O)

$O_{yes+male} = 123006$
$O_{yes+female} = 179211$
$O_{no+male} = 39424$
$O_{no+female} = 72107$

$$X^2 = \sum_{ij} \frac{(O_{ij} - Eij)^2}{E_{ij}}$$

	yes+male	yes+female	no+male	no+female
O	123006	179211	39424	72107
E	118644.942	183572.058	43785.0584	67745.9416
O - E	4361.05837	-4361.0584	-4361.0584	4361.05837
(O - E)^2	19018830.1	19018830.1	19018830.1	19018830.1
(O - E)^2 / E	160.300387	103.604166	434.368043	280.737556
Σ	**979.010153**			
χ^2	**979.010153**			

c. Based on your calculated chi-square value, is there an association between gender and exercise? Explain

i. What is your null hypothesis?

My null hypothesis is that there is no association between Body Mass Index (BMI) after controlling for exercise and gender.

ii. What significance level are use assuming?

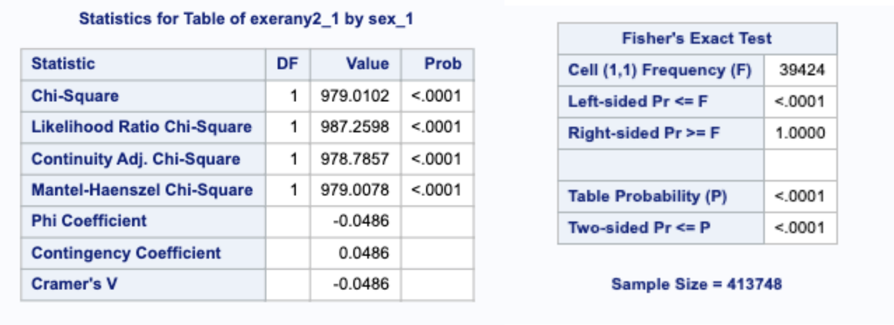
0.05

iii. What is the critical chi-square value?

3.841

d. Show that your (1) hand calculated chi-square statistic, and (2) conclusion on the presence of an association matches that produced by PROC FREQ. Include the relevant PROC FREQ output in your answer.

My hand calculated chi-squared statistic is the same value as the PROC FREQ chi-squared output (979.0102). The critical chi-square value is 3.841 because we're using a significance level of 0.05 with 1 degree of freedom. Since the 979.0102 > 3.841, this shows there is an association between exercise and gender and we can reject the null hypothesis.

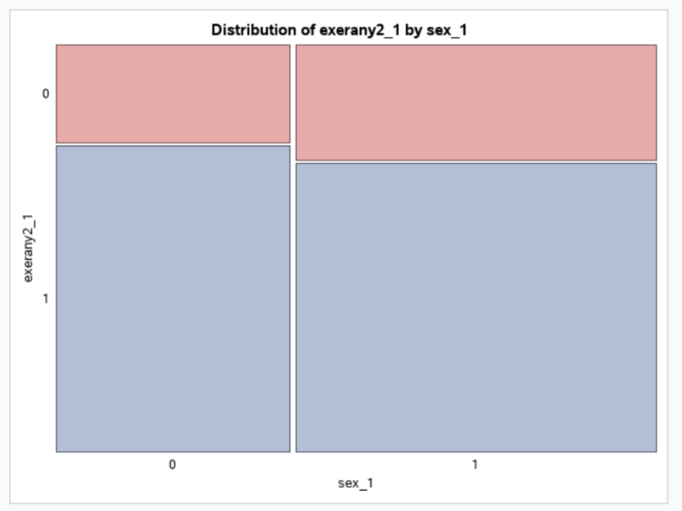


Statistics for Table of exerany2_1 by sex_1

Statistic	DF	Value	Prob
Chi-Square	1	979.0102	<.0001
Likelihood Ratio Chi-Square	1	987.2598	<.0001
Continuity Adj. Chi-Square	1	978.7857	<.0001
Mantel-Haenszel Chi-Square	1	979.0078	<.0001
Phi Coefficient		-0.0486	
Contingency Coefficient		0.0486	
Cramer's V		-0.0486	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	39424
Left-sided Pr <= F	<.0001
Right-sided Pr >= F	1.0000
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Sample Size = 413748



3) (30 pts) Create your “Table 1” for this objective. You can use this table template (copy and paste into Excel):

Table 1. Characteristics of 413,748 BRFSS 2010 participants by BMI category.

	Population		Normal		Overweight		Obese		
Variable	N	%	n	%	n	%	n	%	p value *
Gender									
Male	162430	39.3%	43061	29.6%	72605	47.8%	46764	40.1%	<.0001
Female	251318	60.7%	102291	70.4%	79176	52.2%	69851	59.9%	
Exercise									
Yes	302217	73.0%	113204	77.9%	114407	75.4%	74606	64.0%	<.0001
No	111531	27.0%	32148	22.1%	37374	24.6%	42009	36.0%	

* p values based on Pearson chi-square test of association

4) (40 pts) Write the results section for this “Table 1”

Of the 451,075 BRFSS 2010 participants, 413,748 (92%), had complete data for the objective. The demographic characteristics of this population are compared in Table 1. There were proportionately more females than males (60.7% vs. 39.3%, respectively) in the population. Females that had a “Normal” BMI (70.4%) had the highest ratio of the population (p<0.0001). There were proportionately more exercisers than non-exercisers (73.0% vs. 27.0%, respectively) in the population. Out of the participants that exercised, the highest ratio had a “Normal” BMI (77.9%) (p<0.0001).

Extra Credit (10 pts)

The calculation of a chi-square statistics makes use of an “expected value”. Using the exercise by gender contingency table, give an intuitive explanation of how the expected value is derived.

The expected count is the proportion for each variable combination (yes+male, yes+female, no+male, no+female) with respect to the entire population (413748).

The proportion for each variable needs to be calculated first by taking the total counts for each variable (yes, no, male, female) and dividing it by the entire population (413748). Next, we need to calculate each variable combination (yes+male, yes+female, no+male, no+female) by multiplying its corresponding proportion together. Since we now have the proportion for each variable combination, we can finally calculate the expected count by multiplying it by the entire population (413748).

The further the observed values are from the expected values, the more likely that there really is a significant difference and there is an association between the variables.