SAS Analytics Guide:

How to estimate frequency

Data version: Controlled Tier *All of Us* Curated Data Repository (CDR) v7 (2022Q4R11)
Analysis tool: SAS Studio
Authors:

Joyonna Gamble-George, Ph.D., MHA (✉)
Yale University School of Public Health

Please note: This guide aims to demonstrate the utilization of **PROC FREQ** using SAS Studio. However, it is important to note that this guide is not comprehensive and does not encompass all facets of the scientific process which researchers are required to undertake. Specifically, it only does not delve into data cleaning and verification, assumption validation, model diagnostics, potential follow-up analyses, or any other possible approaches for estimating prevalence using **PROC FREQ**.

Table of Contents

Introduction

The objective of the study described in this guide is to estimate the prevalence of negative social determinants of health among adult patients that experienced a life-threatening medical event, episode, or accident (i.e., heart attack, stroke, severe allergic reaction, asthma attack, or traumatic injuries) or a terminal disease (i.e., HIV/AIDS, Alzheimer's disease, organ failure, or advanced cancer) and understand its association with mood impairments, substance misuse, and healthcare utilization.

To begin our discussion, we first define a few terms used throughout this guide:

- ***Descriptive statistics:*** Describes the characteristics of a dataset (e.g., sex at birth, race, ethnicity) in terms of averages, its spread, and the shape it produces.
- ***PROC FREQ:*** A SAS procedure used for frequency analysis. It is used to calculate the frequency distribution and summary statistics for categorical variables.

Description of the dataset

For this example, we will use the following criteria to define the case cohort for this analysis:

- Any adult participant (aged ≥18 years).
- At least one occurrence of a diagnosis code for myocardial infarction (i.e., heart attack) or a cerebrovascular accident (i.e., stroke) in their electronic health records (EHR).

Categorical variables:

- Gender identity
- Race
- Ethnicity
- Age
- Sex at birth

Statistical analysis procedures

Estimating prevalence and extracting categorical variables

Step 1: Create a cohort.
Step 2: Create a dataset.
Step 3: Import the dataset into SAS Studio.
Step 4: Create a SAS program for descriptive analysis.
Step 5: Estimate lifetime prevalence.

- You can use the **PROC FREQ** command to analyze the frequency distribution of categorical variables.
- For example, if you want to analyze the frequency distribution of the variable "Gender." The output of **PROC FREQ** will display the frequency distribution of the variable "Gender," along with summary statistics such as percentages and cumulative percentages.
- You can also use **PROC FREQ** to analyze the frequency distribution of multiple variables at once. For example, if you want to analyze the frequency distribution of “Ethnicity,” “Gender,” “Race,” and “Sex at birth,” you can use the following code.

Example code:

```
proc freq data=combined;
tables ethnicity*group / chisq;
tables gender*group / chisq;
tables race*group / chisq;
tables sex_at_birth*group / chisq;
run;
```

Example results:

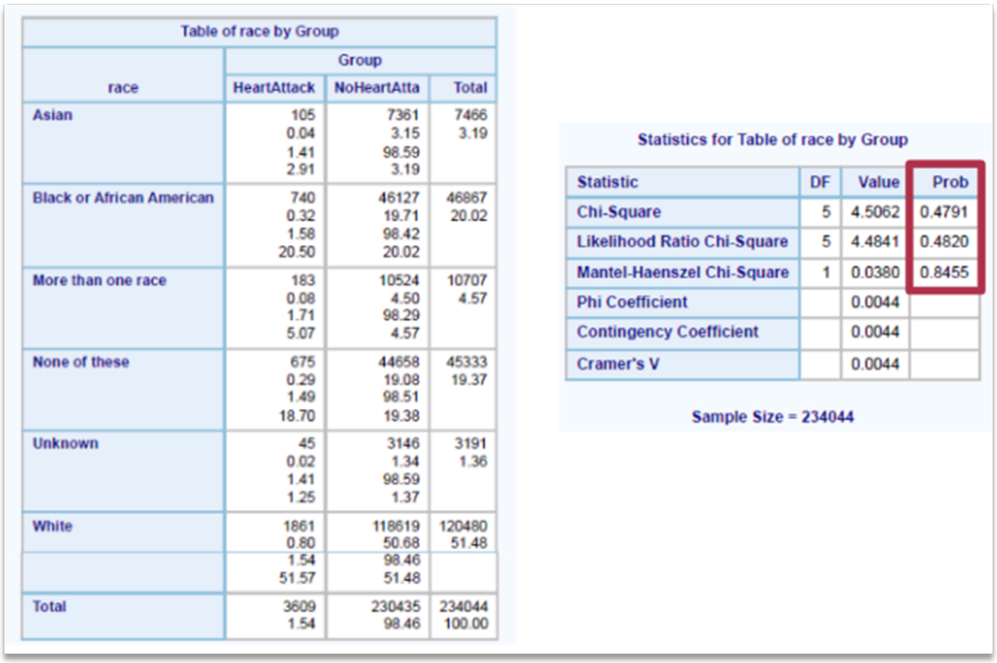


Table of race by Group			
	Group		
race	HeartAttack	NoHeartAtta	Total
Asian	105 0.04 1.41 2.91	7361 3.15 98.59 3.19	7466 3.19
Black or African American	740 0.32 1.58 20.50	46127 19.71 98.42 20.02	46867 20.02
More than one race	183 0.08 1.71 5.07	10524 4.50 98.29 4.57	10707 4.57
None of these	675 0.29 1.49 18.70	44658 19.08 98.51 19.38	45333 19.37
Unknown	45 0.02 1.41 1.25	3146 1.34 98.59 1.37	3191 1.36
White	1861 0.80 1.54 51.57	118619 50.68 98.46 51.48	120480 51.48
Total	3609 1.54	230435 98.46	234044 100.00

Statistics for Table of race by Group

Statistic	DF	Value	Prob
Chi-Square	5	4.5062	0.4791
Likelihood Ratio Chi-Square	5	4.4841	0.4820
Mantel-Haenszel Chi-Square	1	0.0380	0.8455
Phi Coefficient		0.0044	
Contingency Coefficient		0.0044	
Cramer's V		0.0044	

Sample Size = 234044

Calculating lifetime prevalence

Lifetime prevalence is calculated by the number of existing cases divided by the total population of participants with any EHR data.

Step 1: Count the number of heart attack cases.

Example code:

```
proc SQL;
select count(*) into :heartattack_cases;
from heartattack;
quit;
```

Step 2: Count the total number of participants in both datasets.

Example code:

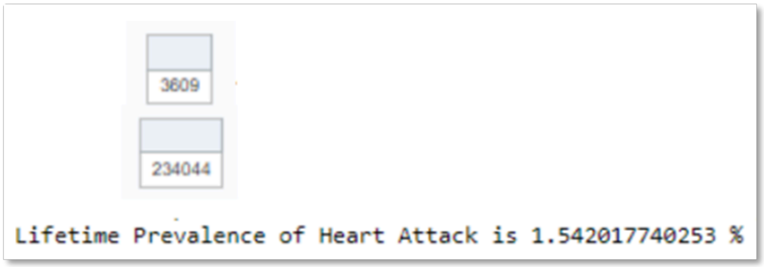
```
proc SQL;
select count(*) into :total_population;
from combined;
quit;
```

Step 3: Calculate the prevalence.

Example code:

```
%let prevalence =%sysevalf(&heartattack_cases / &total_population *100);
%put Lifetime Prevalence of Heart Attack is &prevalence %;
run;
```

Example results:



```
3609
234044
Lifetime Prevalence of Heart Attack is 1.542017740253 %
```

Additional resources

For additional information about using SAS Studio in the Researcher Workbench, explore the following articles: Exploring All of Us data using SAS Studio and How to run SAS in the Researcher Workbench.

Updated April 9, 2024