EDUC 641 Lab: Applied Statistics in Education and Human Services I

Lab 3: 10/23 and 10/24

## Focus on Assignment 2, chi-square tests, and using the filter() function to create new datasets

**Goals:**

1. Chi-square goodness-of-fit test of categorical association
2. Sub-sample comparison using **filter()** function

***What’s NEW?!***

Using the **filter()** function in R.

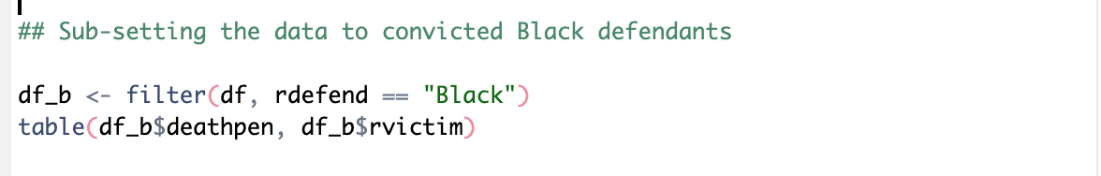
The **filter()** function in R is used to *subset* or filter rows from a dataframe based on specific conditions or selected variable criteria. It's a fundamental function for data manipulation and is commonly used in data analysis and data wrangling tasks. Here's how it works:

**Code:**

**subset\_data <- filter(**your\_data, variable **== “variable specification/condition”)**

* **your\_data** this is *your* dataframe (i.e., what you name cat.csv using ‘<-’ ).
* **Variable specification/condition:** This is a *logical* condition (meaning it has to be in quotes) that specifies which rows you want to keep. Rows that satisfy this logical condition will be included in the filtered result.
  + This is where you select the aspects, or levels, of a variable(s) you want to keep
* The newly created **subset\_data** dataframe will contain only the rows that meet the specified conditions (male or female, etc.).

**Example**



**Notice:** the ' == ’ sign to specify the condition or the criteria that the variable must meet to be

included in the newly subset data. AND that in generating a new table() when calling to your variables using ‘ $ ‘ make sure you’re using your new **subset\_data** dataframe.

**Worksheet**

**Don’t forget to load packages (install.packages()** if you haven’t already**)**

**library(tidyverse)**

1. **Start a new R script. Copy from assignment one or read in data and assign to object/dataframe name**
   1. **your\_data** <- **read.csv**(**here**("data/your\_data.csv")) # just like assignment one

# make sure categorical variables are factored appropriately

1. **Create tables and figures to observe the variables**
   1. **Creating a two-way table**
      1. **your\_table <- table**(your\_data$variable1, your\_data$variable2)

**addmargins**(your\_table) # addmargins() function gives you row and column sums.

* 1. **Proportion table** 
     1. **prop\_table** <- **prop.table**(**table**(your\_data$variable1, your\_data$variable2), margin= 1) # see note

**prop\_table**

**Note:** Use margin = 1 if you want the proportion to add to 1 in each row. Use margin = 2 if you want the proportion to add to 1 in each column. If you don’t give a margin, the proportions across all cells will add to 1.

* 1. **Create one figure representing variable1 and variable2**

Observe categorical variables using code from last weeks lab:

**ggplot(**your\_data, **aes(x =** variable1**, fill =** variable2**)) +**

**geom\_bar(position =** "dodge"**) +**

**labs(title =** "Title of Graph", **x =** "x-axis title", **y =** "Counts",

**fill =** “Custom Category Name for Variable 2”**) +**

**theme\_minimal()**

1. **Run Chi-Square tests**
   1. **chi\_df** <- **chisq.test(**your\_data $variable1, your\_data $variable2**)**

**chi\_df** # code for 3.3

* 1. **chi\_df$expected** # for 3.2
  2. **chi\_df$observed**

1. **Sub-setting data using filter()**
   1. **subset\_data <- filter**(your\_data**,** variable3 **==** "variable3 condition")

# variable 3 the other categorical variable you’re subsetting, i.e., gender

* 1. Use the chi-square test function **chisq.test()** to analyze your variables using this new **subset\_data**