Name: _____ SID: _____

Collaborators: _____

Week 1 Problem Set

R Workbook
PHW142

You must put your name and SID at the top of the page.

Using the WCGS Dataset

The using R introductory materials in the Course Overview and the R Tutorial have already given you some experience creating new variables and making boxplots, histograms, and bar charts.

There is an R Markdown file to get you started with this assignment.

One of the goals of the original study was to see if there is evidence of a connection between so-called Type A behavior, "characterized by excessive competitiveness and aggression " – road rage is one manifestation – and coronary heart disease.

The variable behpat0 in the dataset is a four-category code for Type A behavior:

1 Type A

2 some Type A

3 a mix of Type A and Type B

4 Type B

Preliminaries and Navigation Aids

We are using the same dataset for this assignment as for the practice problems, rather than having you get used to a new one. Please work through the R Tutorial first, then access the R markdown (.Rmd) file in the R Lab Workbook Tab in Week 1.

`w1-ps-code.Rmd`

Run each code chunk in the R Markdown file and save the results from the Console pane and the graphs into this PDF as you make them.

For the last parts of the assignment, you will need to **copy and paste some of the code chunks into the R Markdown file and edit them to accomplish the tasks we are asking you to do.**

The R Markdown file has comments to help you figure out what the functions do.

Since this assignment is being graded for completion, you may discuss it with other students and ask us detailed questions in office hours.

Highlights from the R Markdown File

- Finding average systolic blood pressure at baseline (`sbp0`) for smokers and non-smokers
- Summarizing data with the `dplyr::summarize` function, and base R subsetting.
  - Code is shown two ways–subsetting with `dplyr` and with base R. You do not have to do both.
- Creating a factor variable for 4-level behavior type variable, `behpat0`.
- Creating the usual BMI categories using the cut function
- Creating a two-way table for behavioral pattern and CHD
  - This will be a model for the other tables the assignment asks you to make
- Creating side-by-side bar charts for CHD by behavior pattern category
- Producing summary statistics and side-by-side histograms for BMI

Additional Analyses

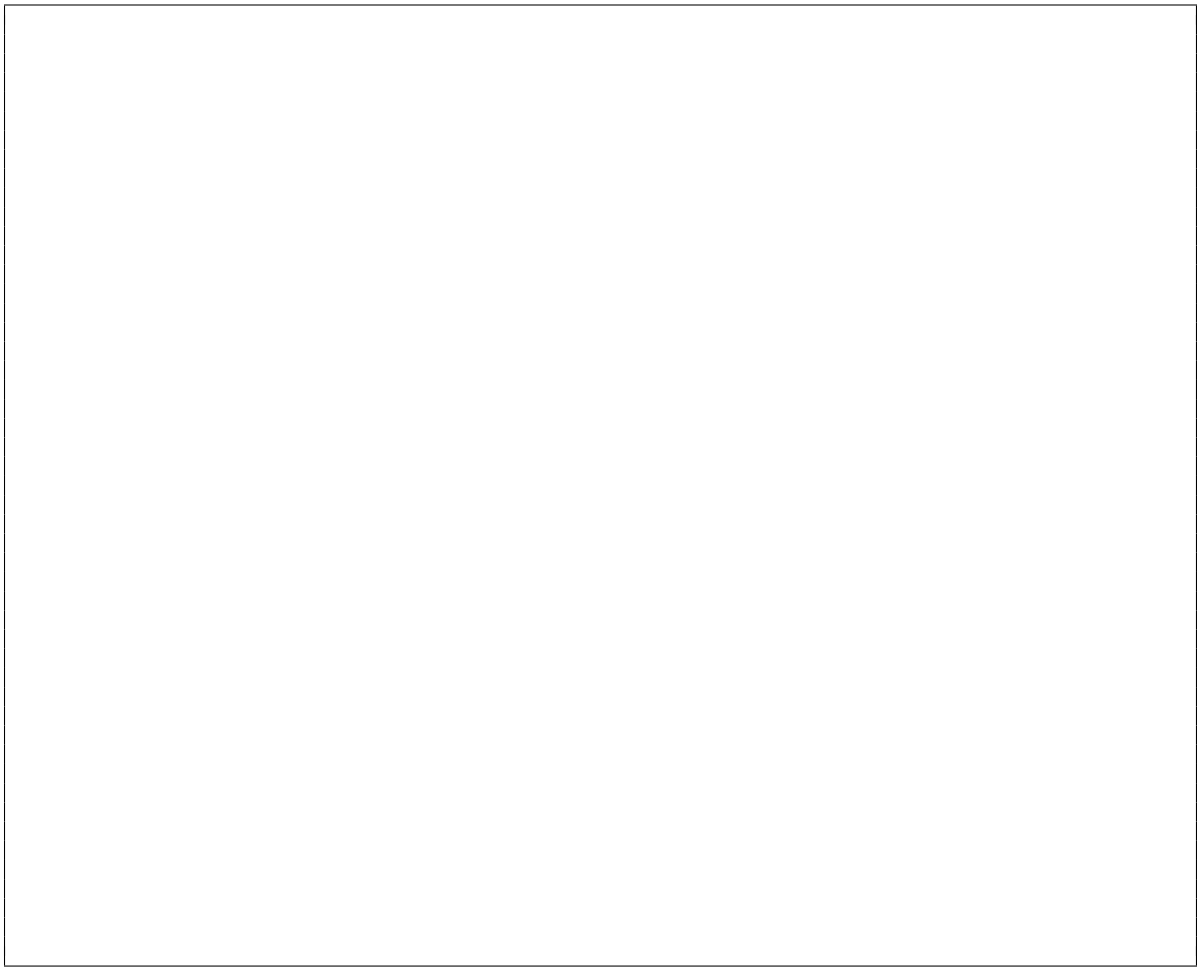
- Two-way tables of percentages
  - binary high blood pressure at baseline (`highsbp0`) and CHD
  - BMI category at baseline and CHD
  - smoking status at baseline (`smoker0`) and CHD

  For each of these tables put the `chd69` variable second in the list, the `prop.table` function option 1 gives row percents
- Get the 5-number summary for weight at baseline (`weight0`)
- Make a histogram for weight at baseline (`weight0`)
  - You will find it helpful to get the 5 number summary to choose the range
- Make a boxplot for the weight at baseline
  - You will need the 5-number summary to answer the questions about the boxplot

Week 1 R Problem Set

Copy and paste your graphs and numericalsummaries into this document in the spaces under the questions. Think of this as practice for preparing your problem set solutions in a way that makes it easy for us to find and appreciate your work.

1. Average systolic blood pressure at baseline for smokers and non-smokers



2. Proportions of coronary heart disease (CHD) for all 4 behavioral pattern levels

3. Side-by-side bar chart for CHD by behavioral pattern category

4. Proportion of CHD for high blood pressure at baseline

5. Proportions of CHD for each BMI category

6. Proportions of CHD for smokers and non-smokers at baseline

7. 5-number summary for weight at baseline

8. Histogram for weight at baseline

9. Box plot for weight at baseline

10. Calculate the IQR for the weight at baseline.

11. Find the values of the upper and lower fences.

12. Using the values of the fences and your numerical summary, explain why there are observations plotted with isolated circles in the boxplot.

13. Compare the average and the median of the weights. How different are they?

    Explain, based on the graphs, and the sample size.

14. Use the histogram, the box plot, and the numerical summary to decide if the distribution of weights is symmetric or skewed, and briefly explain your reasoning.