Name:

Stats in practice #1

Describing Data with Graphs

# The Dataset

The CDI dataset contains demographic information for the 440 most populated counties in the United States. Data for a *random sample* of n=100 of these counties can be found on the course webpage in a file called cdi\_sample.csv. For each county in the dataset, there are 16 variables, many of which we will be using throughout the semester.

## Load the data into R

### Open Rstudio and open a new “script”

### Load the dataset into R by typing:

### cdi <- read.csv(‘http://math.unm.edu/~knrumsey/cdi\_sample.csv’)

### To see the 16 variables contained in this dataset type:

### names(cdi)

### For today’s assignment, we will focus on two variables: State and Percentage of adult population with a HS Diploma.

# Variable: STATE

## Thinking about the variable

### Is State a quantitative or categorical variable?

### This dataset consists of a random sample of 100 counties from the 440 *most populated* counties in the US. Which US states do you expect to be the most represented in this dataset?

## Creating a Barplot

### First, let’s create a variable in R by typing:

### state <- cdi$State

### Create a table of the data by typing:

### table(state)

### According to your table, which state is the most represented in this dataset? Is this what you expected?

### Now we can create our first Barplot by typing:

### barplot(table(state), las=2, cex.names=0.7)

### *Notes: the “las=2” argument puts the state labels vertical instead of horizontally, and “cex.names=0.7” changes the font size. You can also add the argument “main=”title here” to add a title to the plot. (:*

### Let’s make a new barplot now with a slight change.

### barplot(sort(table(state)), las=2, cex.names=0.7)

### Which Barplot is easier to read and interpret? **Copy this plot and paste it below**.

# Variable: Percentage of Population with HS Diploma

## Thinking about the variable

### Is Percentage of adult diploma with a HS Diploma a quantitative or categorical variable?

### Let’s speculate for a minute. Roughly, what do you think the average of this variable to be? Do you think the distribution will be symmetric or skewed? What do you think the min and max to be? (No wrong answers here)

### Which other variables in the dataset might be related to percent of the adult population with a HS diploma.

## Creating a Histogram

### First, let’s create a variable in R by typing:

### percentHS <- cdi$PercentHSGraduates

### Find the average of this variable by typing:

### mean(percentHS)

### Is this close to what you expected?

### We will learn more about this next week, but you can summarize the data in R by typing:

### summary(percentHS)

### What is the min and max of this variable? Is it close to what you expected?

### Now we can create our first Histogram by typing:

### hist(percentHS)

### Does the distribution look symmetric or skewed?

### The “breaks” argument to the hist function tells R how many bars you want (*it’s actually more complicated than that, but roughly true).* Make a “sky-scraper” histogram by typing:

### hist(percentHS, breaks=4)

### Also make a “pancake” histogram by changing “breaks” to 50. Explain what is wrong with both histograms.

### Create a final histogram by choosing a value of “breaks” that you think best shows the distribution. Try to add a title and x-axis label to this plot using the arguments “main” and “xlab”. Optional: In this hist function, set “freq=FALSE”. What does this do? **Copy this plot and paste it below**.