**Memorandum**

To: NTRHD Intern

From: Dr. Brad Cannell, Director, NTRHD

Re: Descriptive Analysis I

**Overview:**

In today’s lab we will practice performing an interpreting descriptive analysis on categorical and numerical variables using R. Examples 2 and 3 are taken from the [Daniels Biostatistics textbook](https://www.wiley.com/en-us/Biostatistics%3A+A+Foundation+for+Analysis+in+the+Health+Sciences%2C+11th+Edition-p-9781119496571). Sometimes calculating statistics by hand, and then checking them with statistical software, can help you develop a better intuition for interpreting output.

**Example 1. Descriptive analysis of 2010 Census data**

The following data comes from the [United States Census Bureau website](https://data.census.gov/cedsci/).

"AL", "South", 4779736, 657792, 37.9,

"AK", "West", 710231, 54938, 33.8,

"AZ", "West", 6392017, 881831, 35.9,

"AR", "South", 2915918, 419981, 37.4,

"CA", "West", 37253956, 4246514, 35.2,

"CO", "West", 5029196, 549625, 36.1,

"CT", "NE", 3574097, 506559, 40.0,

"DE", "South", 897934, 129277, 38.8,

"FL", "South", 18801310, 3259602, 40.7,

"GA", "South", 9687653, 1032035, 35.3,

"HI", "West", 1360301, 195138, 38.6,

"ID", "West", 1567582, 194668, 34.6,

"IL", "NCntrl", 12830632, 1609213, 36.6,

"IN", "NCntrl", 6483802, 6483802, 37.0,

"IA", "NCntrl", 3046355, 452888, 38.1,

"KS", "NCntrl", 2853118, 376116, 36.0,

"KY", "South", 4339367, 578227, 38.1,

"LA", "South", 4533372, 557857, 35.8,

"ME", "NE", 1328361, 211080, 42.7,

"MD", "South", 5773552, 707642, 38.0,

"MA", "NE", 6547629, 902724, 39.1,

"MI", "NCntrl", 9883640, 1361530, 38.9,

"MN", "NCntrl", 5303925, 683121, 37.4,

"MS", "South", 2967297, 380407, 36.0,

"MO", "NCntrl", 5988927, 838294, 37.9,

"MT", "West", 989415, 146742, 39.8,

"NE", "NCntrl", 1826341, 246677, 36.2,

"NV", "West", 2700551, 324359, 36.3,

"NH", "NE", 1316470, 178268, 41.1,

"NJ", "NE", 8791894, 1185993, 39.0,

"NM", "West", 2059179, 272255, 36.7,

"NY", "NE", 19378102, 2617943, 38.0,

"NC", "South", 9535483, 1234079, 37.4,

"ND", "NCntrl", 672591, 97477, 37.0,

"OH", "NCntrl", 11536504, 1622015, 38.8,

"OK", "South", 3751351, 506714, 36.2,

"OR", "West", 3831074, 533533, 38.4,

"PA", "NE", 12702379, 1959307, 40.1,

"RI", "NE", 1052567, 151881, 39.4,

"SC", "South", 4625364, 631874, 37.9,

"SD", "NCntrl", 814180, 116581, 36.9,

"TN", "South", 6346105, 853462, 38.0,

"TX", "South", 25145561, 2601886, 33.6,

"UT", "West", 2763885, 249462, 29.2,

"VT", "NE", 625741, 91078, 41.5,

"VA", "South", 8001024, 976937, 37.5,

"WA", "West", 6724540, 827677, 37.3,

"WV", "South", 1852994, 297404, 41.3,

"WI", "NCntrl", 5686986, 777314, 38.5,

"WY", "West", 563626, 70090, 36.8

**Task 1.** Please create a data frame in R from this data.

* Name the data frame **census**
* When creating the **census** data frame, please use the following column names (you don’t need to type the definitions anywhere. They are just written below for your benefit):
  + **state**: Abbreviated state name.
  + **region**: Region the state is located in.
  + **pop**: The total population of the state.
  + **pop65**: The population of people in the state who are age 65 or older.
  + **medage**: The median age of the state.

**Task 2.** Create a new factor variable for each of the categorical variables in the **census** data frame (i.e., **state** and **region\_f**). Please use the **\_f** naming convention when you create these new columns.

**Task 3.** View the structure of the data frame you created above using the str() or dplyr::glimpse() functions.

**Task 4.** Use R to calculate the frequencies for the **region\_f** variable.

**Task 5.** Create a bar graph depicting the number of states in each region.

**Task 6.** Reorder the **region\_f** factor vector from the region that contains the fewest number of states to the region that contains the greatest number of states. Then, create another bar graph depicting the number of states in each region. The region that contains the fewest number of states should be the farthest left bar on the graph and the region containing the greatest number of states should be the farthest right bar on the graph.

**Questions**

**1** When you viewed the structure of the **census** data frame above, how many columns were there?

**2** Which region contains the largest number of states?

**3** How many states are in the North Central AND Northeast region?

**Example 2. Numerical descriptions of numerical variables**

Porcellini et al. studied 13 HIV-positive patients who were treated with highly active antiretroviral therapy (HAART) for at least 6 months. The CD4 T cell counts at baseline for the 13 participants are listed below (Daniel, 2005).

230 205 313 207 227 245 173 58 103 181 105 301 169

**Task 7.** Use this data to create a numerical vector in R.

**Task 8.** Use R to calculate the mean, median, mode, standard deviation, minimum value, and maximum value of this vector.

**Questions**

**4** What is the mean CD4 T cell count for these 13 participants (rounded to the nearest tenth)?

**5** What is the median CD4 T cell count for these 13 participants?

**6** T/F? There is no mode CD4 T cell count for these 13 participants.

**7** Which measure of central tendency is best to report based on your results?

**Example 3. Numerical and graphical descriptions of numerical variables**

Thilothammal et al. designed a study to determine the efficacy of BCG (bacillus Calmette-

Guérin) vaccine in preventing tuberculosis meningitis. Among the data collected on each subject

was a measure of nutritional status (actual weight expressed as a percentage of expected weight

for actual height). The nutritional status values of the 107 cases studied are listed below (Daniel,

2005).

73.3 54.6 82.4 76.5 72.2 73.6 74.0 80.5 71.0 56.8 80.6 100.0 79.6 67.3 50.4 66.0 83.0 72.3 55.7

64.1 66.3 50.9 71.0 76.5 99.6 79.3 76.9 96.0 64.8 74.0 72.6 80.7 109.0 68.6 73.8 74.0 72.7 65.9

73.3 84.4 73.2 70.0 72.8 73.6 70.0 77.4 76.4 66.3 50.5 72.0 97.5 130.0 68.1 86.4 70.0 73.0 59.7

89.6 76.9 74.6 67.7 91.9 55.0 90.9 70.5 88.2 70.5 74.0 55.5 80.0 76.9 78.1 63.4 58.8 92.3 100.0

84.0 71.4 84.6 123.7 93.7 76.9 79.6 45.6 92.5 65.6 61.3 64.5 72.7 77.5 76.9 80.2 76.9 88.7 78.1

60.6 59.0 84.7 78.2 72.4 68.3 67.5 76.9 82.6 85.4 65.7 65.9

**Task 9.** Use this data to create a numerical vector in R.

**Task 10.** Use R to calculate the mean, median, mode, standard deviation, minimum value, and maximum value of this vector.

**Task 11.** Create a histogram of these values (hint: you *must* pass a data frame to ggplot(). It will not plot a standalone vector).

**Questions**

**8** What is the standard deviation of these participants’ nutritional status (rounded to the nearest tenth)?

**9** What argument in the geom\_histogram() function can you use to adjust the number of bins in your histogram?

**Example 4. Analyze some class survey data**

The data below was collected from students in a class. It contains eight variables: **id**, **height**, **weight**, **male** (coded as 1 if the student is male and 0 if the student is female), **bach5300** (coded as 1 if the student took the course BACH5300 and 0 if the student did not take BACH5300), **bios5300** (coded as 1 if the student took the course BIOS5300 and 0 if the student did not take BIOS5300), **epid5300** (coded as 1 if the student took the course EPID5300 and 0 if the student did not take EPID5300), and **gpa**. The periods in the data represent missing values.

**1 170 185 1 1 1 1 3.6**

**2 175 162 1 1 1 1 3.7**

**3 231 180 1 1 1 1 3.8**

**4 189 190 1 1 0 1 3.8**

**5 164 175 1 1 0 1 .**

**6 178 178 1 1 0 0 3.78**

**7 . 192 1 1 1 0 3.87**

**8 184 178 1 1 1 0 3.99**

**9 186 169 1 1 1 0 3.98**

**10 174 130 1 1 0 0 4**

**11 165 140 0 1 0 1 2.8**

**12 155 125 0 1 1 . 3.56**

**13 158 126 0 1 1 1 .**

**14 156 138 0 1 1 1 2.9**

**15 168 116 0 1 1 1 3.5**

**16 145 114 0 1 0 1 3.4**

**17 158 135 0 1 1 0 3.3**

**18 110 141 0 1 0 0 3.8**

**19 153 137 0 1 0 0 3.4**

**20 165 129 0 1 0 0 3.6**

**Task 12.** Please create a data frame in R from this data.

**Task 13.** Use R to calculate the number of missing values, mean, median, minimum value, and maximum value of all of the numeric vectors in this data frame (i.e., **height**, **weight**, and **gpa**).

**Task 14.** Create a boxplot for each of the numeric vectors in this data frame (i.e., **height**, **weight**, and **gpa**).

**Questions**

**10** Which numeric vectors in this data frame (i.e., **height**, **weight**, and **gpa**) have a missing value in at least one observation?

**11** How many non-missing values are there for the variable **gpa**?

**12** How many outlying values are there for the variable **weight**?

**13** Which variable has the greatest number of outlying values?

**Optional**: Please feel free to leave any comments below about the usefulness of this lab. Which parts were helpful? What could I do to improve it? What is still unclear?

**References:**

United States Census Bureau. (2014). American Fact Finder [Data file]. Retrieved from (<http://factfinder2.census.gov>

Daniel, W. (2005). *Biostatistics: A Foundation for Analysis in the Health Sciences* (8th ed.). Hoboken, NJ: John Wiley & Sons Inc.