Lab 2 Homework: Summarizing Data

Background

Alzheimer's Disease (AD) is a serious neurologic disorder that affects an estimated 5.3 million Americans; it is the most common cause of dementia among the elderly. Characterized by a progressive cognitive decline, AD has been notoriously difficult to diagnose due to symptom-overlap with other mental disorders; until recently, AD could only be confirmed posthumously. Researchers are investigating identifying biomarkers in interdisciplinary research involving biostatisticians, neuroscientists, geneticists, and clinicians, among other experts.

DX	Alzheimer's disease diagnosis AD - Alzheimer's disease MCI - Mild cognitive impairment Normal - Normal cognitive function
AGE	Age (years)
APOE4	Type of APOE4 variant (genetics) 0 - No copies of the ApoE4 allele 1 - One copy of the ApoE4 allele 2 - Two copies of the ApoE4 allele
GENDER	Patient gender
MMSE	Mini Mental State Exam (score out of 30, lower scores indicate more cognitive impairment)
adas	Alzheimer's Disease Assessment Scale (larger scores indicate greater dysfuction)
WholeBrain	Brain volume (mm^3)

The Data

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a longitudinal study that began in 2005, and is designed to track AD biomarkers, identify at-risk patients, and evaluate the efficacy of novel treatments. The study consists of healthy individiduals (the control group) as well as adults with early Alzheimer's Disease (AD). More details can be found on the website http://adni.loni.usc.edu/about/. The data set provided is ADNI.csv, which contains information on 276 subjects and 7 variables. The mini mental state exam is a 30 question assessment commonly used to assess cognitive impairment. The Alzheimer's Disease Assessment Scale is a more comprehensive measure of cognitive impairment. The apolipoprotein E (APOE) gene, on chromosome 19, has variants associated with high risk of AD. In this lab, your goal is to provide summary statistics and visualizations of the data.

Practice

1. Import the `ADNI.txt` data set. How many observations are there?

2. For each variable in the data set, describe what the variable represents in reality (categorical or numerical), as well as its variable type in R (using the str command). The first one has been filled in for you.

Description	`DX`	`AGE`	`APOE4`	`GENDER`	`MMSE`	`adas`	`WholeBrain`
In reality	categorical						
In R	character						

3. Notice the discrepancy with the APOE4 variable between what it represents in reality and its R variable type. Recode this variable to a factor. Which APOE4 genetic variant is the least common?

4. Which graph is appropriate to visualize the distribution of AGE? Produce this graph and describe the distribution.

5. Let's examine the relationship between Alzheimer's diagnosis (DX) and the two cognitive impairment tests (MMSE and adas).
   a) Which graph is appropriate to visualize the relationship between Alzheimer's diagnosis (DX) and the cognitive impairment test MMSE? What about DX and adas?
   b) Produce these graphs and describe the relationship between diagnosis and each cognitive test.
   c) Which cognitive test identifies potential outliers?

6. Identify descriptive statistics for the overall characteristics of the study participants, and for each diagnosis group.
   a) The WholeBrain volume measured in mm3 is on a large scale. Convert this measurement to a smaller scale. Create a new variable that represents the whole brain volume in mm3 divided by 100,000.
   b) Fill in the following table with descriptive statistics for the overall characteristics of the study participants, and for each diagnosis group. Numerical variables should be summarized as mean +/- standard deviation, and categorical variables have been summarized as n (%). The n represents the count, and the percent represents the percent of that column with the attribute of interest. Some cells have been filled in for you to help you get started. Round all summary statistics to one decimal. (Hint: all percents should use the column sample size as the denominator.)

	Overall	Diagnosis Group

		AD	MCI	Normal
	n = 276	*n* =	*n* =	*n* = 94
Age			72.9 +/- 7.3	
Gender (male)	153(55.4%)			
Brain volume x10^5 mm^3	10.2 +/- 1.1			
APOE4				
No copies				62 (66.0%)
One copy				28 (29.8%)
Two copies				4 (4.3%)

7. Fill in the blanks or circle the correct answer in following paragraph which utilizes descriptive statistics from the table to summarize characteristics of the ADNI study participants. Note that the first two sentences describe general characteristics of the overall study participants and the last two sentences compare two groups of subjects.

The ADNI study has _________ participants. The average age is ________ years and ________% are male. The Alzheimer's group has a ***lower/higher*** *average brain volume than the Normal group (_______ vs ________mm^3). Patients with Alzheimer's diagnosis have a* ***lower/higher*** *prevalence of two copies of the APOE4 allele compared to normal diagnosis patients (_____% vs 4.3%).*