m158-project1-NHANES

Ian Horsburgh and Sara Colando

NHANES Data Description:

We are using data from the National Health and Nutrition Exam Survey (NHANES) database that was collected between 2009-2012 with adjusted weighing. The target population of NHANES is "the non-institutionalized civilian resident population of the United States". Therefore, the observational unit is a civilian resident of the United States (of any age).

Relevant Variables:

Age (quantitative) - Age in years at screening of study participant. All subjects 80 and older were recorded as 80.

Testosterone (quantitative) - Testerone total (ng/dL), recorded for patients 6 and older. Note that no testosterone data for 2009-2010 was recorded.

Physactivedays (numerical discrete) - Number of days in a typical week that partipant does moderate or vigorous-intensity activity. Reported in patients 12 years or older.

Pulse (quantitative) - 60 second pulse rate

HHIncomeMid (numerical) - Total annual gross income for the household in US dollars, derived from the median of each partition. Variable was partitioned into blocks with the smallest one being (0,4999) and the largest block being (100,000 or more).

Gender (categorical) - Gender (sex) of study participant coded as "male" or "female".

SleepTrouble (categorical) - Paritipant has told a doctor or other health professional that they had trouble sleeping. Reported in patients 16 and older. Either "yes" or "no".

Depressed (categorical) - Self reported number of days where participant felt down, depressed or hopeless. Reported in patients 18 and olders with categories of "none", "several", "majority (more than half the days)", or "almostall".

Education (categorical) - Educational level of study participant. Reported for ages 20 or older. Categories to choose from are "8thgrade", "9-11thgrade", "Highschool", "SomeCollege", or "CollegeGrad".

Work (categorical) - Categorizes whether study participant is "working", "not working" or no data was collected.

Summary Statistics

```
## # A tibble: 10 x 5
##    skim_variable  numeric.mean numeric.sd complete_rate numeric.p50
##    <chr>                 <dbl>      <dbl>         <dbl>       <dbl>
##  1 Gender                NA         NA            1            NA
##  2 SleepTrouble          NA         NA            0.777        NA
##  3 Depressed             NA         NA            0.667        NA
##  4 Education             NA         NA            0.722        NA
##  5 Work                  NA         NA            0.777        NA
```

```
##  6 Age                  36.7       22.4          1           36
##  7 HHIncomeMid       57206.     33020.           0.919    50000
##  8 Testosterone        198.       227.           0.413       43.8
##  9 PhysActiveDays        3.74       1.84         0.466        3
## 10 Pulse                73.6       12.2          0.856       72
```

We can see that our data has varying rates of completion with most being over 50% reported, but Testosterone and PhysActiveDays have 41.25% and 46.6% completion rates respectively. As such, it is likely that our statistics for Testorsterone and PhysActiveDays are not fully representative of the population.

Additionally, Testosterone has a extremely high standard deviation (SD) relative to its mean, indicating the data distribution may be skewed. This is corroborated by the mean and median (p.50) being substantially different for Testosterone, and so we expect the distribution to be skew right.

For all the other variables, median (p.50) and mean seems relatively close indicating the data distribution is somewhat symmetrical. However, to confirm this fact, we will have to look at histograms and bar graphs to see if there is a skewness to the data distribution.

Graphs

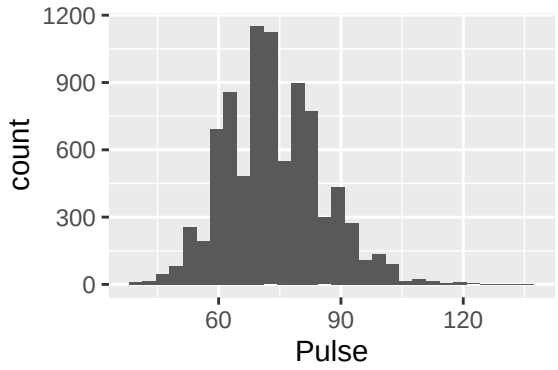

Figure 1: Frequency Distribution of (60-sec) Pulse

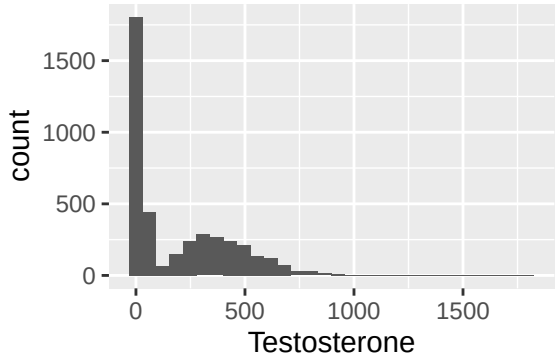

Figure 2: Frequency Distribution for Testosterone Levels (ng/ml)

Commentary on Graphs

The distribution of Pulse is approximately unimodal symmetric and bell-shaped.

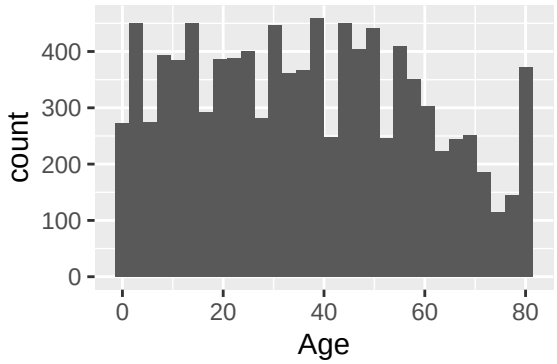

Figure 3: Frequency of Age

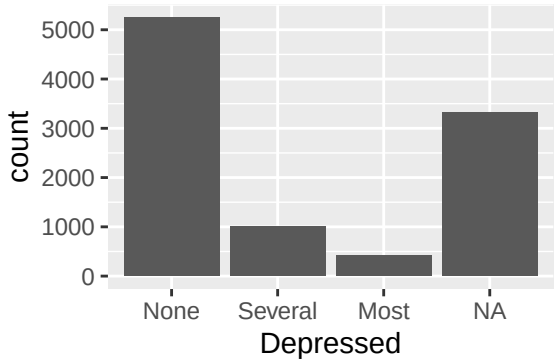

Figure 4: Frequency of Depression (based on proportion of days/week)

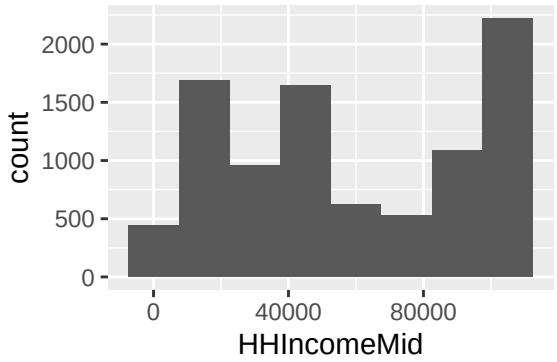

Figure 5: Frequency of Household Income Salaries

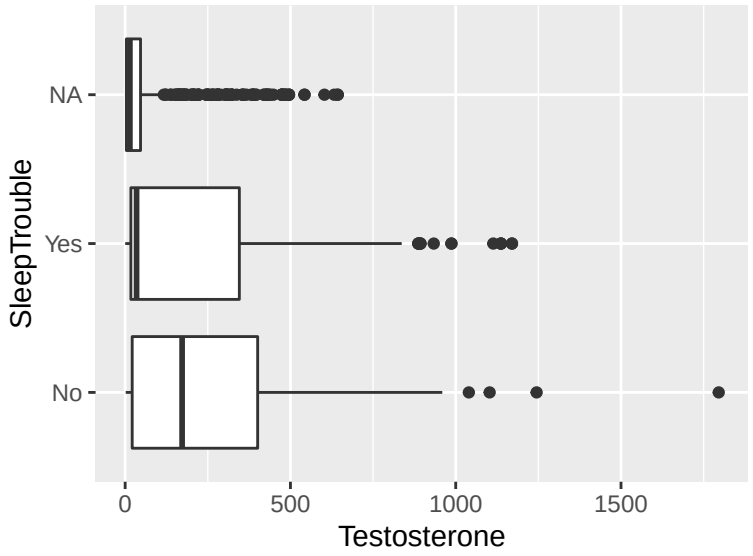

Figure 6: Box Plot of Testosterone Levels (ng/ml) vs. Sleep Troubles

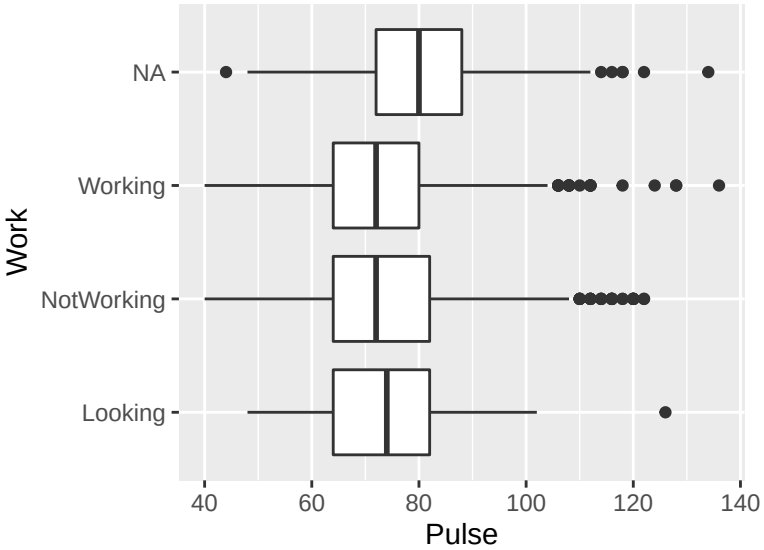

Figure 7: Box Plot of (60 second) Pulse vs. Work

The distribution of testosterone Levels is skew right with the mode being at 0.

Data distribution is relatively uniform wth a slight peak around 30-40 years old and othrs at the edges (10 and younger as well as 80 and above.

The most common response was none with also a high proportion of N/A"

The distribution of Household Income (middle) looks slightly skew left with the mode being 100,000k+. However, median and mean look to be centered around 50-60,000 US dollars.

The median testosterone levels were higher for those who answered No sleep troubles than for those who answered Yes sleep troubles. However, the interquartile range testosterone level for Yes and No sleep troubles looked approximately equal. This shows that there may be a possible relationship between Testosterone levels and Sleep Troubles.

The Median pulse for Working and Not Working look approximately equal whereas the average pulse for Looking for work looks slightly higher. On the other hand, the interquartile range for NotWorking and Looking are about the same whereas the interquartile range for working is smaller. This shows that possibly working status and pulse have a relationship between each other.

Additional Information

We were surprised by the amount of N/A data for some variables – especially for those that seemed like general information (such as working status or education).

The data in the sample was about what we expected – though it was a little surprising to see an uptick in people 80 (or older) in our sample. However, upon reflection, this would make sense, as 80+ adults are probably more likely to come to the doctors office and complete the NHANES survey.

Given the number of observations and the relatively high completion rate for most categories, we do think the sampling data is representative of our population (i.e. US citizens). It will be interesting to see though if the high completion is true for all ages – or just certain age subsets of US citizens.

The NHANES data represents a sample from the population of resident citizens of the US (who are 2 months or older). The specific NHANES data we are using was collection for 2009-2010 and 2011-2012. These results were voluntary and primarily collected via interview and physical examinations. There are some limitations of describing the larger population using NHANES data:

Firstly, many of our relevant variables were only collected within adults. Therefore, there are limitations to how we could apply our sample data to the population of all resident citizens of the US (who are 2 months or older) because we do not have certain variable data collected for non-adults.

Additionally, NHANES' study was voluntary and primarily self-reported or collected by healthcare professionals. Therefore, it is likely that people who feel sicker were overrepresented in our sample. Additionally, it is likely that people who have health insurance (or can afford wellness check-ups etc.) are also overrepresented in our sample.