Report_4

Your Name

9 December 2019

Instructions:

1. Update the read.csv command so that R knows where to find the data on your computer.
2. Set the random number seed (see below) to a unique value.
3. Replace all the italic text with your own discussion (this is a focus of this report).
4. After you have completed the first confidence interval and z-test examples, repeat the analysis for the other cases described below.

Introduction

The NHANES data set consists of four data files: adult.csv, youth.csv, lab.csv and exam.csv. The adult file contains information about subjects who were over 17 years old. The youth file contains information about younger subjects. The lab and exam files contain additional data about both adults and youth.

Reading the data into R

For this report, we only need to read the exam.csv file into R.

```
exam = read.csv("../nhanes/exam.csv", header=TRUE)
```

The size of this data set is

```
dim(exam)
```

```
## [1] 31311  2368
```

There are 31311 rows and 2368 columns in the exam file.

Setting a Random Number Seed.

Change the 10 in the following command to some other positive whole number. Pick a number that is different from that of other students. This value is used to randomly select a subject from the data set.

```
set.seed(10)
```

Confidence Intervals

First extract the data that we will use for the confidence interval and z-test examples:

```
x1 <- data.frame(SEQN=exam$SEQN, HSSEX=exam$HSSEX, HSAGEIR=exam$HSAGEIR,
                 PEP8=exam$PEP8, PEPMNK1R=exam$PEPMNK1R)
x2 <- na.omit(x1)
x3 <- x2[x2$PEP8 != 8,  ]
x4 <- x3[x3$PEPMNK1R != 888, ]
```

The size of this data set is

```
dim(x4)
```

```
## [1] 22993     5
```

There are 22993 rows and 5 columns in the exam file.

Confidence Interval on Age (sample size = 25)

What is the average age of the subjects? We can use a confidence interval to get an estimate from a small sample.

```
n = 25
N = dim(x4)[1]
sample_rows = sample(1:N, n)
sample_data = x4[sample_rows, ]
```

Now compute the mean and sd of the ages in the sample:

```
m = mean(sample_data$HSAGEIR)
m
```

```
## [1] 29.76
```

```
s = sd(sample_data$HSAGEIR)
s
```

```
## [1] 20.51682
```

Compute the standard error (for the average):

```
se = s / sqrt(n)
se
```

```
## [1] 4.103364
```

The 95% confidence interval on the true average age is from 21.5532711 to 37.9667289. The true average age of all subjects is 36.5149393.

Confidence Interval on Age (sample size = 100)

Repeat the previous example using $n = 100$.

z-Tests

If the sample size is large, we use the z-test to test the null hypothesis that the mean of a sample differs from a hypothesized mean simply due to chance.

Blood Pressure of a Sample of Male Subjects

We can find the average systolic blood pressure, take a sample of the male subjects, then test the hypothesis that the average blood pressure for the males is equal to the population average.

Overall, the average K1, systolic blood pressure of all subjects was:

```
m = mean(x4$PEPMNK1R)
m
```

```
## [1] 118.4289
```

Extract the male subjects from the data set:

```
men = x4[x4$HSSEX == 1, ]
```

Randomly select a sample of the male subjects.

```
n = 100
N = dim(men)[1]
men_sample_rows = sample(1:N, n)
men_sample = men[men_sample_rows, ]
```

Compute the mean, standard deviation and standard deviation (for the average) of blood pressure measurements for this sample:

```
x = mean(men_sample$PEPMNK1R)
s = sd(men_sample$PEPMNK1R)
se = s / sqrt(n)
```

Based on this sample, test the hypothesis that the blood pressure of the male subjects was equal to the blood pressure for all subjects. The null hypothesis is that the difference between the overall average and the sample average is due to chance. The alternative hypothesis is that the difference is not due to chance.

```
z = (x - m) / se
z
```

```
## [1] 1.967438
```

By the Central Limit Theorem, the sample mean is approximately a normal distribution. So we can compute the p-value as follows.

```
1 - pnorm(z)
```

```
## [1] 0.02456636
```

Since this p-value is small, it supports the alternative hypothesis. The average male blood pressure seems to differ (be higher than) the overall average.

Blood Pressure of a Sample of Female Subjects

Repeat the above example using just the female subjects. Start with the line women = x4[. . .]

Blood Pressure of a Sample of Subjects of a Specified Age

Repeat the above example using just the subjects of a specified age. Start with the line subjects = x4[...]